Neural Readability Pairwise Ranking for Sentences in Italian Administrative Language

Martina Miliani[1,2] and **Serena Auriemma**[2] and **Fernando Alva-Manchego**[3] and **Alessandro Lenci**[2]
[1] University for Foreigners of Siena
[2] CoLing Lab, Department of Philology, Literature, and Linguistics, University of Pisa
[3] School of Computer Science and Informatics, Cardiff University, UK
martina.miliani@fileli.unipi.it, serena.auriemma@phd.unipi.it
alvamanchegof@cardiff.ac.uk, alessandro.lenci@unipi.it

Abstract

Automatic Readability Assessment aims at assigning a complexity level to a given text, which could help improve the accessibility to information in specific domains, such as the administrative one. In this paper, we investigate the behavior of a Neural Pairwise Ranking Model (NPRM) for sentence-level readability assessment of Italian administrative texts. To deal with data scarcity, we experiment with cross-lingual, cross- and in-domain approaches, and test our models on Admin-It, a new parallel corpus in the Italian administrative language, containing sentences simplified using three different rewriting strategies. We show that NPRMs are effective in zero-shot scenarios (∼0.78 ranking accuracy), especially with ranking pairs containing simplifications produced by overall rewriting at the sentence-level, and that the best results are obtained by adding in-domain data (achieving perfect performance for such sentence pairs). Finally, we investigate where NPRMs failed, showing that the characteristics of the data used for fine-tuning, rather than its size, have a bigger effect on a model's performance.

1 Introduction

Due to its complexity, the style of Italian administrative texts has been defined as "artificial" and "obscure" (Lubello, 2014). During the last decades, Italian institutions have fostered the use of a plain language in writing official acts and communications (Fortis, 2005). However, the readability of Italian administrative texts still remains an issue (Cortelazzo, 2021), and measuring their complexity can help institutions improve information accessibility, and guarantee a substantive equality of citizens (Vedovelli and De Mauro, 1999).

One way to tackle this problem is with technologies for Automatic Readability Assessment (ARA) that predict the complexity of texts (Collins-Thompson, 2014). This task has been widely investigated in the educational domain, usually classifying texts according to school grade levels or international frameworks for language proficiency. Currently, most models for ARA are based on neural networks (Vajjala, 2022), which are trained in a supervised fashion by fine-tuning pre-trained language models (Imperial, 2021; Martinc et al., 2021; Lee and Vajjala, 2022). However, this approach could require large amounts of monolingual in-domain data, which is limited in specific sectorial languages like the one used in Italian administrative texts, for which the available resources are quite scarce (Tonelli et al., 2016; Brunato, 2015).

In this paper, we tackle the data scarcity issue in two ways. First, we introduce Admin-It (Sec. 3), a parallel corpus in the Italian administrative language with sentences that were simplified following three different styles of rewriting. Then, we repurpose Lee and Vajjala (2022)'s Neural Pairwise Ranking Model (NPRM) to rank sentences (instead of documents) from the Italian administrative language (Sec. 4), because that model obtained better results than traditional classification and regression approaches in zero-shot cross-lingual set-ups.

We evaluate the performance of NPRMs on Admin-It in zero-shot settings (Sec. 5), fine-tuning models with data from different languages (i.e., Italian, English and Spanish) and domains (i.e., administrative, educational, and news). We show that, overcoming the limitations of traditional ARA system in cross-domain set-ups (Dell'Orletta et al., 2012; Vajjala, 2022), NPRMs obtain good results in cross-domain and cross-lingual scenarios, especially when ranking sentences simplified via overall rewriting (Sec. 6).

Finally, we conduct a qualitative analysis on the errors made by NPRMs (Sec. 7), and observe how models deal with various kinds of simplification, such as overall rewriting versus the application of single operations of simplification (e.g., lexical substitution, splitting or deleting).

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 849–866
November 20–23, 2022.

To sum up, our main contributions are:

- We create Admin-It, a parallel corpus of sentences for the Italian administrative language containing different simplification styles;[1]
- We prove that the Neural Pairwise Ranking Model is also effective for automatic readability assessment of sentences;
- We experiment with NPRMs in cross-domain and cross-lingual set-ups, analyzing their performances when fine-tuned with data of different languages and domains, and show that they reach good results in zero-shot scenarios;
- We analyze the models' errors according to the styles of simplification applied in different subsections of Admin-It.

While ARA is normally a document-level task, we tackle it at the sentence level due to the characteristics of the datasets available in Italian (Tonelli et al., 2016) and the administrative domain (Scarton et al., 2018), which mainly contain aligned sentences (see details in Sec. 5.1). Also, a sentence-based approach for readability could be more effective in detecting easy and complex to read texts, since complex documents may also contain easy-to-read sentences (Dell'Orletta et al., 2014; Todirascu et al., 2016; Howcroft and Demberg, 2017).

2 Related Work

Early ARA techniques consisted in the so-called "readability formulae". Such formulae were created for educational purposes and mainly considered shallow text features, like word and sentence length or lists of common words (Lively and Pressey, 1923; Flesch, 1948; Kincaid et al., 1975).

However, longer words and sentences are not necessarily complex, and these formulae have been proved to be unreliable (Si and Callan, 2001; Petersen and Ostendorf, 2009; Feng et al., 2009). In addition, traditional readability formulae should not be applied to fragments with less than 100 words, making them unsuitable to assess the readability of sentences, which is usually considered more difficult than predicting readability of documents (Dell'Orletta et al., 2011; François, 2015).

NLP and Machine Learning fostered the emergence of "AI readability" systems (François, 2015), leading to the creation of new techniques for both supervised and unsupervised approaches (Vajjala, 2022). Traditional supervised techniques model ARA as classification (Schwarm and Ostendorf, 2005; Vajjala and Meurers, 2012), regression (Heilman et al., 2008), or ranking (Ma et al., 2012; Vajjala and Meurers, 2014) tasks, exploiting a wide range of linguistic features, at a lexical (Chen and Meurers, 2018), syntactic (Schwarm and Ostendorf, 2005; Kate et al., 2010), and discourse level (Graesser et al., 2004; Barzilay and Lapata, 2008; Pitler and Nenkova, 2008). More recent systems are based on neural networks (Nadeem and Ostendorf, 2018; Martinc et al., 2021; Imperial, 2021), exploiting contextual embeddings like *BERT* (Devlin et al., 2019) to encode large quantity of linguistic knowledge. However, such models still need to be fine-tuned to be applied in downstream tasks. For some languages and domains, like Italian administrative texts, this is not possible since there is not enough available data for a full supervised approach. For this reason, we adopted a cross-lingual approach and created our own resource for the Italian administrative language (i.e., Admin-It).

Recently, Lee and Vajjala (2022) used neural models to address ARA as a ranking task. Their Neural Pairwise Ranking Model (NPRM) ranks a group of documents by their readability, regardless of its size (i.e., the number of reading levels). Their NPRM obtained better results than classification and regression approaches for texts in English, Spanish and French, in both monolingual and zero-shot cross-lingual set-ups. As such, we decided to exploit this architecture but for ranking sentences. Furthermore, while Lee and Vajjala (2022) found that the NPRM struggles in a cross-domain setting, they did not deeply analyzed the behaviour of the model when dealing with data whose domains are wide apart (e.g., news and bureaucratic domains). In contrast, we study the impact on performances given both by the datasets used for fine-tuning the NPRM and by the specific kind of simplification applied to the sentences being ranked.

3 Admin-It

Given the paucity of data in the Italian administrative language for sentence readability and simplification, we decided to build **Admin-It**, a parallel corpus of Italian administrative language. The corpus comprises 736 sentence pairs corresponding to two readability levels: original and simplified. We organized the corpus in three subsets according to the different nature of the applied simplification:

Operations (Admin-It$_{OP}$): 588 pairs of sen-

[1] https://github.com/Unipisa/admin-It

tences from the subsection of the Simpitiki corpus (Tonelli et al., 2016) related to the administrative domain. These pairs contain manual simplifications produced by rewriting original sentences using single operations, such as split, reorder, merge, lexical substitutions, among others. The authors report that most simplifications in this dataset involve lexical transformations at word (single terms) and phrase (e.g., multiword expressions) levels, whereas the merging operation is never applied.

Rewritten Sents (Admin-It$_{RS}$): New 100 pairs of original-simplified sentences. The original sentences were selected from websites of Italian municipalities,[2] and from the longest phrases from the PaWaC Corpus (Passaro and Lenci, 2015). We manually rewrote the sentences simplifying them both at lexical and syntactic levels. Our simplification criteria were based on the *Thirty rules for good administrative writing* by Cortelazzo (2021) and by considering the typical traits of the administrative language (Brunato et al., 2015). For example, some particularly frequent simplification operations are: the replacement of verbal phrases formed by verb + noun with the corresponding simple verbs (e.g., from *apporre la firma* [append a signature] to *firmare* [sign]; from *effettuare un pagamento* [make a payment] to *pagare* [pay]) and the transformation of nouns in verbs, since nominalization is a typical trait of administrative language that affects its degree of readability. In addition, uncommon nouns and verbs were replaced by synonyms present in the Basic Italian Vocabulary (De Mauro, 2000), which contains the most frequent terms of contemporary Italian. An exception were the technical terms of the administrative language or its subsectors (e.g., *catasto* [real estate registry]; *deroga*[waive]; *referendum abrogativo* [abrogative referendum]). At the syntactic level, the number of subordinate clauses and parenthetical expressions was reduced, favoring coordination and shorter sentences.

Rewritten Docs (Admin-It$_{RD}$): 48 pairs of sentences selected from administrative texts, which were collected and simplified by Cortelazzo (1998; 1999) and made publicly available.[3] This resource contains pairs of original-simplified documents rewritten according to linguistic simplification and communicative effectiveness criteria. We manually aligned selected sentences by choosing from the documents only those sentences in which the simplified version had the same informative/semantic content as the original "complex" sentence, without applying any further manipulation.

[2] http://www.semilchattadino.it
[3] http://www.cortmic.eu

Dataset	*# pairs*	*Lev Dist.*	*Char Length*
Admin-It	736	49.6 ± 92.5	238.7 ± 139.4
– Admin-It$_{OP}$	588	13.6 ± 18.7	204.2 ± 90.6
– Admin-It$_{RS}$	100	202.1 ± 122.7	425.5 ± 204.6
– Admin-It$_{RD}$	48	172.3 ± 127.0	271.3 ± 148.1

Table 1: Some statistics of Admin-It and its subsets: number of sentence pairs, Levenshtein distance between original and simplified sentences, and length in characters of orignal and simplified sentences.

In order to make Admin-It publicly available, we masked potentially sensitive data mentioned in the sentences, such as bank account numbers, addresses, licence numbers, phones and emails. Table 1 reports some quantitative information about the corpus. Admin-It$_{RS}$ has the highest average length of all subsets since, by design, it contains simplifications for very long sentences. Furthermore, both Admin-It$_{RS}$ and Admin-It$_{RD}$ register high Levenshtein distances since these two subsets were simplified through overall rewriting, whereas in Admin-It$_{OP}$, one single simplification operation per sentence was applied. Examples of sentence pairs can be found in Appendix A (Table 6).

4 Neural Pairwise Ranking for Sentences

In this section, we briefly describe the Neural Pairwise Ranking Model (NPRM) of Lee and Vajjala (2022) that ranks documents according to their readability, and then explain how we apply it to rank original-simplified sentence pairs.

NPRM for Documents. The model's input is composed of a list of (v, r) tuples, such as $X = [(v_i, r_i), ..., (v_n, r_n)]$, where v_i is the vector representation of a document and r_i is its readability score. By analyzing all permutations of pairs of documents in the list, the model aims at maximizing the probability that $r_i > r_j$, i.e., that the readability score of a document is higher than the score assigned to the other document in the pair, so that the predicted scores p^1_{ij}, p^2_{ij} correspond to $p^1_{ij} = P(r_i > r_j|v_i, v_j)$ and $p^2_{ij} = 1 - P(r_i > r_j|v_i, v_j)$. The NPRM is parametrized as $NPRM = softmax(\psi(f(v_i, v_j)))$, where f is a *BERT* model and ψ is a fully connected layer. The adopted loss function is the Pairwise Logistic Loss (Han et al., 2020).

NPRM for Sentences. In our setting, the input text is sentences instead of documents. Even though the NPRM can rank an arbitrary number of texts in each list of tuples, due to the characteristics of our data, we rank sentences in only two readability levels: complex and simple. Therefore, the input is now a list of two tuples with the vector representations of the original (s_o) and simplified (s_s) versions of the same sentence, and their readabilities. That is $X_i = [(s_{o_i}, r_{o_i}), (s_{s_i}, r_{s_i})]$. No further changes were made to the original model.

To validate our adaptation of the model, we examined the performance of the NPRM for ranking sentences in a monolingual setting for English. We fine-tuned it on the OSE corpus (see Sec. 5.1) via 5-Fold cross validation with `bert-base-uncased`. The resulted ranking accuracy was quite high (0.96) and close to the one obtained by Lee and Vajjala (2022) for the document-level setup in the same corpus (0.98). This supports using NPRMs for ranking sentences.[4]

5 Experimental Settings

We adapted the released code of Lee and Vajjala (2022)[5] for our sentence-level task, but retained their parameter settings during the fine-tuning of the NPRMs and the training of the baselines. Models were trained and fine-tuned on an Nvidia GPU TITAN RTX .

5.1 Datasets

We fine-tuned our models using data in three languages (English, Spanish and Italian) and three domains (news, administrative and educational). As a pre-processing step, for all datasets, we filtered out instances where the original and simplified sentences were identical.[6]

OneStopEnglish (*OSE*): Contains 189 articles from the British newspaper The Guardian that were rewritten by teachers into three readability levels (elementary, intermediate, and advanced) for learners of English as a second language (Vajjala and Lučić, 2018). It has a total of 567 documents. We used the sentence-aligned version of the corpus that contains 5,994 sentence pairs.

NewsEla English (*NewsEn*): Contains news articles in English that were rewritten by professional editors from Newsela (an educational company) in up to four readability levels (Xu et al., 2015). We used the automatic and manual sentence alignments released by Jiang et al. (2020). After our filtering, we obtained 488,390 pairs.

NewsEla Spanish (*NewsEs*): Contains translations into Spanish of the original articles in the NewsEla corpus, which were then manually simplified into different levels of linguistic proficiency, with a total of 1,221 documents. We used the automatic sentence alignments released by Palmero Aprosio et al. (2019). After our filtering, the dataset contains 52,048 pairs of sentences.

Simpitiki/Wikipedia (*Simpitiki$_W$*): Introduced in Tonelli et al. (2016), this corpus includes 575 pairs of original-simplified sentences extracted from Italian Wikipedia edits and manually annotated with simplification operation types, following the annotation scheme proposed by Brunato et al. (2015). Beyond our standard filtering, we also removed 7 pairs with the token "$[\cdots]$" to avoid sentences containing discontinued portions of text. This resulted in 568 pairs of sentences.

SimPA: This is an English sentence-level simplification corpus in the administrative domain (Scarton et al., 2018). It contains 5,500 pairs of sentences: 3,300 with lexical-only simplifications; 1,100 with syntactic simplifications applied after lexical simplification; and 1,100 with lexical and syntactic simplifications applied at the same time. After our filtering, we obtained 4,637 pairs.

5.2 Baselines

Similarly to Lee and Vajjala (2022), we used SVMRank as baseline, a non-neural ranker that uses the difference between features extracted from the sentence pairs as input to an SVM. We trained two baseline models that differ on the input features. Baseline$_L$ considers the sole sentence length in characters,[7] whereas Baseline$_E$ exploits sentence embeddings extracted from *BERT*, using them as a training feature for the SVMRank model.

For what concerns Baseline$_L$, we decided to focus on sentence length to mimic the behaviour of traditional readability formulae, and because it is a raw text feature that we could easily extract and compare between corpora of different languages. In addition, such baseline assigns a ranking even in cases of ties (see how we handled this in the evalu-

[4] See Appendix B for more details on these preliminary experiments on English in in- and cross-domain settings.

[5] https://github.com/jlee118/NPRM/

[6] See some statistics of this corpora in Appendix A.

[7] We did not use the sentence length in tokens to avoid having the same length for the original and simplified versions of a sentence, since many simplifications in Admin-It$_{OP}$ only consist of lexical substitutions at the word level.

ation step in Sec. 5.3). Finally, Baseline$_L$ models were trained following different combinations of data, similar to our NPRMs.

With regards to Baseline$_E$, the sentence embeddings are obtained from an Italian *BERT* model that we call *BertIta*[8], following the code shared by Imperial (2021), who used mean pooling to extract such representations.[9] We trained this SVMRank on Simpitiki$_W$, described in Sec. 5.1.

5.3 Evaluation metrics

Our models are evaluated in terms of Ranking Accuracy (RA), that is the percentage of pairs ranked correctly. We used the implementation provided by Lee and Vajjala (2022), but changed the way it handles ties. More specifically, if the model assigns the same rank to both elements of a pair (i.e., it cannot decide which sentence is simpler), we score it as incorrect. This is because in Admin-It (our test set), simplified sentences should be easier to understand than their original counterparts, reducing the possibility of valid ties. This also prevents overestimating the performance of our length-based baseline. Furthermore, while Lee and Vajjala (2022) suggest using multiple ranking metrics for evaluation (e.g., normalized discounted cumulative gain), we only compute RA in our experiments. The advantage of the other metrics is their ability to handle rankings among several elements and ties in more sophisticated ways. However, our setting is simpler, only comparing two sentences at the time and evaluating ties as errors. Therefore, we decided to base our evaluation only on RA.

5.4 Statistical Significance Testing

To assess if differences in scores between pairs of models are statistically significant, we used a non-parametric statistical hypothesis test, McNemar's Test (McNemar, 1947). We used this test since our models are evaluated using RA, which is computed over a dichotomous variable: when a pair of sentences is ranked correctly 1 is assigned to that pair, 0 otherwise.[10] A p-value lower than 0.05 will indicate that the difference between the scores is statistically significant.

[8] https://huggingface.co/dbmdz/bert-base-italian-uncased

[9] He used the sentence-transformers library by Reimers and Gurevych (2019).

[10] We computed McNemar's Test by adapting the code shared by Lee and Vajjala (2022).

6 Results and Discussion

We describe different **zero-shot experiments**, fine-tuning our models on combinations of monolingual, cross-lingual, in-domain and cross-domain data, and always using Admin-It for testing. While the NPRMs showed variations in performance depending on the fine-tuning setting (as will be explained below), that was not the case for Baseline$_L$, perhaps due to the simplicity of the features extracted, i.e., the length of sentences expressed in characters. For this reason, in Table 2, we do not state what training data was used for such baseline, since the scores are the same for all cases.

6.1 Monolingual and Cross-domain

We first fine-tuned our models with only Italian data, but not from the administrative domain. Our models were fine-tuned on Simpitiki$_W$, with the NPRM exploiting *BertIta*. As shown in Table 2, the NPRM got a lower RA score than both the baselines, a difference that, as shown in Figure 1, is also statistically significant for the overall Admin-It ($p<0.01$ with Baseline$_E$, and $p<0.001$ with Baseline$_L$)[11]. This could be a consequence of the small size of Simpitiki$_W$, which has less than 600 pairs of sentences. And this also may explain why Baseline$_E$, trained on a such corpus, reaches lower performances than Baseline$_L$.

Replacing *BertIta* with *mBERT*,[12] the multilingual version of *BERT*, resulted in higher scores for the NPRM, which are significantly different for the whole Admin-It ($p<0.001$), Admin-It$_{OP}$ ($p<0.001$), and Admin-It$_{RS}$ ($p<0.01$). This is probably due to the large quantity of data used to train *mBERT*. However, such model overpasses Baseline$_L$ only on Admin-It$_{OP}$, which contains simplifications with the same style as Simpitiki$_W$ (i.e., each sentence was simplified by applying only one operation). In contrast, the NPRM fails when simplifications involve a multi-operation rewriting process, as is the case in Admin-It$_{RS}$ and Admin-It$_{RD}$. However, the differences in scores between this model and Baseline$_L$ are not statistically significant.

6.2 Cross-lingual and In-domain

We now experiment with adding in-domain data for fine-tuning (i.e., from administrative texts), but

[11] The heatmaps of the subsets of Admin-It and tables with the numeric values are reported in Appendix E.

[12] https://huggingface.co/bert-base-multilingual-uncased

Test	Baseline$_L$	Baseline$_E$	NPRM (*BertIta*)	NPRM (*mBERT*)		
			Simpitiki$_W$	Simpitiki$_W$	SimPA	Simpitiki$_W$+SimPA
Admin-It	0.640	0.588	0.519	0.660	**0.719**	0.716
– Admin-It$_{OP}$	0.594	0.558	0.502	0.638	**0.685**	0.677
– Admin-It$_{RS}$	0.840	0.740	0.570	0.790	**0.940**	0.930
– Admin-It$_{RD}$	**0.792**	0.646	0.625	0.667	0.667	0.750

Table 2: Ranking accuracies obtained by the baselines and two NPRMs (with different base pre-trained language models) when fine-tuned on Simpitiki/Wikipedia (Simpitiki$_W$) and/or SimPA, and tested on Admin-It.

Test	*OSE*	*NewsEn*	*NewsEs*	*OSE+NewsEs*	*OSE+NewsEn+NewsEs*
Admin-It	0.777	0.765	0.760	**0.785**	0.783
– Admin-It$_{OP}$	0.745	0.731	0.716	0.743	**0.748**
– Admin-It$_{RS}$	0.970	0.960	0.970	0.980	**0.990**
– Admin-It$_{RD}$	0.771	0.771	0.854	**0.896**	0.771
Test	*OSE+S.*	*NewsEn+S.*	*NewsEs+S.*	*OSE+NewsEs+S.*	*OSE+NewsEn+NewsEs+S.*
Admin-It	0.787	0.784	0.791	**0.803**	0.766
– Admin-It$_{OP}$	0.747	0.760	0.762	**0.767**	0.736
– Admin-It$_{RS}$	**1.000**	0.970	0.980	0.980	0.990
– Admin-It$_{RD}$	0.833	0.688	0.750	**0.875**	0.667

Table 3: Ranking accuracy achieved by NPRM (*mBERT*) fine-tuned with OSE, NewsEla English, NewsEla Spanish and their combinations. In the lower part of the table also SimPA (*S.*) was added for fine-tuning. In bold the best result for each table section, whereas the best result for each subset of Admin-It is underlined.

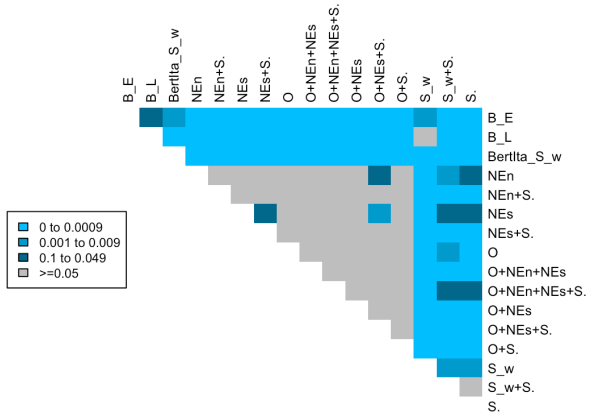

Figure 1: The heatmap shows the p-values obtained with McNemar's Test for pairs of models on the overall Admin-It. Grey cells represent a p-value equal or higher than 0.05. We tested the performances of Baseline$_L$ (B$_L$), Baseline$_E$ (B$_E$), NewsEn (NEn), NewsEs (NEs), SimPA (S.), Simpitiki$_W$ (S$_W$), OSE (O), and their combinations.

not in the same language. In this case, we trained Baseline$_L$ and fine-tuned a *mBERT*-based NPRM on SimPA.

As shown in Table 2, when fine-tuned only on SimPA, the NPRM already surpasses Baseline$_L$ (trained on Simpitiki$_W$ or SimPA) for Admin-It$_{OP}$ ($p<0.001$) and Admin-It$_{RS}$ ($p<0.05$). Adding Simpitiki$_W$ to SimPA to fine-tune the NPRM did not result in better performance. Rather, the RA scores on Admin-It$_{OP}$ and Admin-It$_{RS}$ are lower than those obtained by fine-tuning only on SimPA, although neither for the whole Admin-It nor for its subsets the difference in scores is statistically significant. The decreasing of the performances could be due to the lower quality of Simpitiki$_W$ simplifications, which were semi-automatically collected from users' edits on Wikipedia. On Admin-It$_{RD}$, however, even though not significantly, the performance improved when fine-tuning on both datasets, but still remains lower than Baseline$_L$.

6.3 Cross-lingual and Cross-domain

We proceed to fine-tune our models using out-of-domain data (i.e., news) in other languages (i.e., English and Spanish). In particular, models are fine-tuned on OSE, NewsEn and NewsEs. Results are reported in Table 3 (upper half).

Despite OSE being smaller than NewsEn and NewsEs, the NPRM fine-tuned on it reached better overall results than when fine-tuned on the other datasets. In particular, even if the differences are not significant, that NPRM achieved a higher RA in Admin-It$_{OP}$ and comparable scores in Admin-It$_{RS}$. On the other hand, the NPRM fine-tuned on NewsEs obtained a sensible improvement in RA for Admin-It$_{RD}$, even surpassing Baseline$_L$,

although not significantly. The best result for this subset (and on Admin-It overall) is obtained by combining OSE and NewsEs. Adding NewsEs could have helped because Spanish is more similar to Italian than English, both belonging to the same family of Romance languages and therefore sharing similar morphosyntactic structures (Banfi, 2003). The results obtained by OSE and NewsEs on the whole Admin-It are significantly different from both the baselines, SimPA, Simpitiki$_W$ (with *BertIta* and *mBERT*), and the combination of SimPA and Simpitiki$_W$ ($p<0.001$). With regards to Admin-It$_{RD}$, a statistical significance is observed when comparing the model to Baseline$_E$ ($p<0.01$), SimPA ($p<0.01$), and Simpitiki$_W$ ($p<0.01$ with *BertIta* and $p<0.001$ with *mBERT*). A p-value lower than 0.05 is observed when compared with NewsEn, and with Simpitiki$_W$ and SimPA combined. The lack of significance with Baseline$_L$ may be due to the small size of this subset.

Finally, combining all three datasets allowed an NPRM to obtain the best results in Admin-It$_{OP}$ and Admin-It$_{RS}$ in this setting. On both subsets, there are significant differences with both the baselines and the NPRMs fine-tuned only on Simpitiki$_W$ ($p<0.001$). When compared to SimPA and to the combination of SimPA and Simpitiki$_W$, the significance is reached only on Admin-It$_{OP}$ ($p<0.01$).

We also experimented with pairwise combinations of the three datasets without substantial improvements (see Appendix C for more scores of these experiments).

6.4 Cross-lingual and In-domain

We now experiment with adding in-domain data to the previous setting, even if it is in another language. That is, models are now fine-tuned on OSE, NewsEn, NewsEs and SimPA.

As shown in Table 3 (bottom half), adding in-domain data always lead to an improvement in the overall scores, although it is statistically significant only when SimPA is added to NewsEs ($p<0.05$). The only exception to such an improvement is the NPRM fine-tuned on the combination of NewsEn, NewsEs, and OSE. This could reveal that the size of the dataset used for fine-tuning is less relevant under certain conditions. In fact, the highest improvement is for the NPRM fine-tuned on OSE, NewsEs, and SimPA. This appears to be the best model for overall Admin-It and Admin-It$_{OP}$, whereas mixing OSE and SimPA allows the NPRM to reach a perfect RA on Admin-It$_{RS}$. A possible explanation for such high score is that Admin-It$_{RS}$ contains sentences simplified on several linguistic levels. Therefore, the original and simplified versions of a sentence are very different from one another (as shown by the high average Levenshtein distance in Table 1), possibly making it easier for the NPRM to rank them. Regarding the statistical significance, none of these results are significantly different from the scores obtained by the other models implemented in this setting. Finally, even though adding SimPA contributes to improving the RAs, the NPRMs already obtained high scores without using any in-domain data at all. We also experimented with adding Simpitiki$_W$ to the dataset combinations in this setting. However, in line to what we observed in Sec. 6.2, it did not result in further improvements in overall RA (see Appendix C for an overview of such scores).

7 Analysis

We analyze where the NPRMs failed when ranking sentence pairs from Admin-It$_{RD}$ and Admin-It$_{OP}$. We focus on these two subsets of Admin-It given the high results already obtained on Admin-It$_{RS}$.

7.1 Admin-It$_{RD}$

NPRMs reached the highest RAs in this subset (0.896) when fine-tuned on OSE+NewsEs, OSE+NewsEs+Simpitiki$_W$, or OSE+NewsEn+Simpitiki$_W$. We analyze the errors made by the first model since it also achieved the highest RA (0.785) on the overall dataset among those models. This NPRM failed to rank five out of 48 sentence pairs in Admin-It$_{RD}$.

In some cases, given the same semantic content, punctuation could have affected the scoring because commas split the sentences in various parenthetical expressions (see the first example in Table 4). However, when a sentence contains terms, structures, or formulaic expressions typical of the Italian administrative language, the model ranks the pair correctly regardless of the punctuation, and even in the presence of a higher number of parenthetical expressions in the simplified sentence.

In another case, a sentence was classified as complex when information was added to clarify some implicit information. As shown in the second example in Table 4, to provide such information, the annotator added some deverbal nouns (e.g., *predisposizione* [provision], *posizionamento* [positioning]),

Original: *Si prega inoltre di informare questo Ufficio dell'evasione della pratica mediante il modulo allegato o anche telefonicamente (0001112), affinché la stessa non venga tenuta in sospeso.* [Please also inform this Office of the processing of your file by means of the enclosed form or by telephone (0001112), so that it is not held in abeyance.] **Simplified:** *Per poter archiviare la pratica, chiediamo cortesemente di restituirci il modulo allegato, anche via fax, o di inviarci un messaggio di posta elettronica.* [In order to be able to file the papers, we kindly ask you to return the attached form to us, also by fax, or send us an e-mail.]
Original: *L'Ufficio Anagrafe del Comune provvederà d'ufficio alle conseguenti variazioni nel registro della popolazione residente; alla messa in opera delle nuove targhe sull'edificio provvederanno direttamente gli Uffici comunali competenti. Si comunica inoltre che la suddetta variazione viene segnalata direttamente da questo ufficio ai seguenti enti: ENEL, SIT s.p.a. e Servizio Postale.* [The Registry Office of the Municipality will provide ex officio for the consequent variations in the register of the resident population; the installation of the new plates on the building will be carried out directly by the competent municipal offices. Please also note that the above-mentioned variation will be notified directly by this office to the following entities: ENEL, SIT s.p.a. and Postal Service.] **Simplified:** *Il Comune aggiornerà d'ufficio quanto di sua competenza (anagrafe, autorizzazioni, tributi, comunicazioni agli enti pensionistici ed all'Azienda Provinciale per i Servizi Sanitari), installerà la targhetta indicante il numero civico e comunicherà la variazione direttamente all'ENEL, alla SIT S.p.A. e all'Ente Poste Italiane.* [The municipality will update ex officio all matters within its jurisdiction (registry office, authorisations, tributes, communications to pension authorities and to the Provincial Health Services Agency), install the plaque indicating the house number and communicate the change directly to ENEL, SIT S.p.A. and the Italian Post Office.]

Table 4: Examples of sentence pairs that an NPRM did not rank correctly in Admin-It$_{RD}$. The errors are probably due to the presence of parenthetical expressions (upper half) or due to adding deverbal nouns and in-domain terms (bottom half) in the simplified version of the sentences.

or in-domain terms (e.g., *anagrafe* [civil registry], *tributi* [tributes], *enti pensionistici* [pension authorities], *Azienda Provinciale per i Servizi Sanitari* [Provincial Health Services Agency]), which may have affected the pair ranking. Since sentences in Admin-It$_{RD}$ were manually aligned after simplification was performed at the document level, the annotators could better identify the information needed to be added or made explicit. Probably these sentences underwent more insertions than those in Adminit$_{RS}$. When the simplification is operated directly at the sentence level, in fact, it is more difficult to understand which information to add, since the context is missing.

7.2 Admin-It$_{OP}$

This subset of Admin-It contains sentences from Simpitiki (Tonelli et al., 2016) with annotations of the simplification operations applied to each original sentence. With this information, we computed RA scores for NPRMs (*mBERT*) fine-tuned on different datasets and tested on sentences containing specific simplification operations (Figure 2).[13]

NPRMs were better at ranking sentences involving the Split operation when they were fine-tuned using in-domain data from SimPA. This is because any administrative language is usually characterized by long sentences that are generally split to ease reading. Therefore, SimPA could have provided more training instances containing this operation than the other datasets.

However, despite being in-domain, SimPA does not always help. For example, for sentence pairs containing Reorderings, the NPRM fine-tuned only on SimPA got the lowest RA. This can be explained by the fact that in more than half of the corpus only lexical level simplifications were performed.

As also observed by Tonelli et al. (2016), transformations are the most frequent operations. In particular, they registered a high number of lexical substitutions, probably to replace technical terms and formulaic expressions typical of the administrative language. On sentence pairs with Lexical Substitutions at the word level, the best result is achieved by an NPRM fine-tuned on OSE+NewsEs, whereas for phrase-level substitutions, the highest RA is obtained by fine-tuning with OSE, NewsEs and SimPA. The contribution of OSE to these results may stem from the fact that it is a corpus for people learning English as a second language. Since a high percentage of the vocabulary of the text must be known by learners in order to understand it, OSE may contain several lexical substitutions (Hsueh-Chao and Nation, 2000). For lexical substitutions at the phrase level, instead, formulaic expressions typical of the administrative language may be targeted in the simplification process, so in-domain data from SimPA may be beneficial.

[13] See Appendix D for a tabular visualization of the scores for all the simplification operations.

Figure 2: Each bar plot represents RAs achieved on a single simplification operation in Admin-It$_{OP}$. In brackets the number of sentence pairs simplified with that operation.

NPRMs performed worse on sentences with Insert operations. This is probably because most of the training datasets provided automatically-aligned sentences, and, most likely, pairs containing not overlapping (added) content were filtered out from the data. This could also explain the low scores obtained in Admin-It$_{RD}$, where the annotator applied a more elaborative simplification (Srikanth and Li, 2021), adding details to explicit some information (Sec. 7.1).

We also analyze the scores obtained on sentence pairs with transformations involving verbal features. Here, the NPRM fine-tuned on OSE is the best, also reaching high scores when adding SimPA or NewsEs+SimPA to the data used for fine-tuning. However, using only SimPA results in the lowest scores in this set. This could be explained by the ARA experiments using OSE performed by Vajjala and Lučić (2018). They found that a feature-based model that used char-ngrams performed better than one based on word n-grams. Since the model could better distinguish between complex and simple texts through character rather than stem variations, this could suggest that OSE exemplifies well variations at the morphological level, including verbal inflections. Also, given that for learners of English as second language it could be more difficult to master verbal inflectional morphology, the simplification in this corpus might have often involved verbs.

Despite our best efforts, we cannot easily explain the performance of the NPRMs on sentence pairs with other operations. However, our analysis already offers some insights into how the models behave, serving as a first step for a more comprehensive study to be carried out in future work.

8 Conclusions and Future Work

In this paper, we investigated the behavior of a Neural Pairwise Ranking Model (NPRM) for assessing the readability of sentences from the Italian administrative language in zero-shot settings. To deal with data scarcity in this domain, we built Admin-It, a corpus of original-simplified parallel sentences in the Italian administrative language, containing three different styles of simplifications. This corpus allowed us to prove that NPRMs are effective in cross-domain and cross-lingual zero-shot settings, especially when simplifications were produced over single sentences and at several linguistic levels. We also conduced an error analysis and showed that the characteristics of the data used for fine-tuning rather than its size have an impact on a model's performance. In addition, we determined that simplifications where information was added are poorly handled by the models.

In future work, we plan to analyze how NPRMs perform on sentences with the same simplification style (e.g., Admin-It$_{RS}$) annotated for different degrees of complexity by humans. We also plan to improve Admin-It$_{RS}$ to address the needs of specific targets, such as second language learners, who require the insertion of definitions of technical terms (not provided in the current version). To develop ARA models in this setting, we could leverage the alignments of Srikanth and Li (2021) that focus on elaborative simplifications. Furthermore, we plan to fine-tune models with in-domain data from languages with higher proximity to Italian, e.g., with datasets similar to the one built for Spanish by Morato et al. (2021). Moreover, we would like to apply our models in concrete applications, like evaluation of automatic simplifications. Finally, we aim at extending our approach to other domains and languages besides the administrative one.

References

Emanuele Banfi. 2003. *Lingue d'Europa: elementi di storia e di tipologia linguistica / Emanuele Banfi, Nicola Grandi*. Università Linguistica 482. Carocci, Roma.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Dominique Brunato. 2015. A study on linguistic complexity from a computational linguistics perspective. a corpus-based investigation of italian bureaucratic texts. *Major in Linguistics, University Of Siena*.

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.

Xiaobin Chen and Detmar Meurers. 2018. Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading*, 41(3):486–510.

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.

Michele A. Cortelazzo. 1998. Semplificazione del linguaggio amministrativo. *Quaderni del Comune di Trento. Progetti*, 3.

Michele A. Cortelazzo. 2021. *Il linguaggio amministrativo: principi e pratiche di modernizzazione*. Studi superiori. Carocci.

Michele A. Cortelazzo, Federica Pellegrino, and Matteo Viale. 1999. *Semplificazione del linguaggio amministrativo. Esempi di scrittura per le comunicazioni ai cittadini*. Comune di Padova.

Tullio De Mauro. 2000. Dizionario illustrato della lingua italiana. *Paravia, Torino*.

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.

Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2012. Genre-oriented readability assessment: A case study. In *Proceedings of the Workshop on Speech and Language Processing Tools in Education*, pages 91–98.

Felice Dell'Orletta, Martijn Wieling, Giulia Venturi, Andrea Cimino, and Simonetta Montemagni. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*, pages 4171–4186, Minneapolis, MN, USA.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 229–237.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221.

Daniela Fortis. 2005. Il dovere della chiarezza. quando farsi capire dal cittadino è prescritto da una norma. *RIVISTA ITALIANA DI COMUNICAZIONE PUBBLICA*, 25:82–116.

Thomas François. 2015. When readability meets computational linguistics: a new paradigm in readability. *Revue française de linguistique appliquée*, (2):79–97.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.

Shuguang Han, Xuanhui Wang, Mike Bendersky, and Marc Najork. 2020. Learning-to-rank with bert in tf-ranking. *arXiv preprint arXiv:2004.08476*.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the third workshop on innovative use of NLP for building educational applications*, pages 71–79.

David M Howcroft and Vera Demberg. 2017. Psycholinguistic models of sentence processing improve sentence readability ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 958–968.

Marcella Hu Hsueh-Chao and Paul Nation. 2000. Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1):403–30.

Joseph Marvin Imperial. 2021. BERT embeddings for automatic readability assessment. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 611–618, Held Online. INCOMA Ltd.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. *CoRR*, abs/2005.02324.

Rohit Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond Mooney, Salim

Roukos, and Chris Welty. 2010. Learning to predict readability using diverse linguistic features. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 546–554.

J. Peter Kincaid, Robert P. Fishburne, R L Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. In *Institute for Simulation and Training*.

Justin Lee and Sowmya Vajjala. 2022. A neural pairwise ranking model for readability assessment. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3802–3813, Dublin, Ireland. Association for Computational Linguistics.

Bertha A. Lively and Sidney L. Pressey. 1923. A method for measuring the vocabulary burden of textbooks. *Educational administration and supervision*, 9(7):389–398.

Sergio Lubello. 2014. *Il linguaggio burocratico*. Le bussole. Carocci.

Yi Ma, Eric Fosler-Lussier, and Robert Lofthus. 2012. Ranking-based readability assessment for early primary children's literature. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 548–552.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Jorge Morato, Ana Iglesias, Adrián Campillo, and Sonia Sanchez-Cuadrado. 2021. Automated readability assessment for spanish e-government information. *Journal of Information Systems Engineering and Management*, 6(2):em0137.

Farah Nadeem and Mari Ostendorf. 2018. Estimating linguistic complexity for science texts. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 45–55, New Orleans, Louisiana. Association for Computational Linguistics.

Alessio Palmero Aprosio, Sara Tonelli, Marco Turchi, Matteo Negri, and Di Gangi Mattia. 2019. Neural text simplification in low-resource conditions using weak supervision. In *Workshop on Methods for Optimizing and Evaluating Neural Language Generation (NeuralGen)*, pages 37–44. Association for Computational Linguistics (ACL).

Lucia C. Passaro and Alessandro Lenci. 2015. Extracting terms with extra. In *Proceedings of EUROPHRAS 2015*, pages 188–196, Malaga, Spain. Tradulex.

Sarah E. Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer speech & language*, 23(1):89–106.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 186–195.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 3982–3992.

Carolina Scarton, Gustavo Paetzold, and Lucia Specia. 2018. Simpa: A sentence-level simplification corpus for the public administration domain. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Sarah E. Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL'05)*, pages 523–530.

Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 574–576.

Neha Srikanth and Junyi Jessy Li. 2021. Elaborative simplification: Content addition and explanation generation in text simplification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5123–5137, Online. Association for Computational Linguistics.

Amalia Todirascu, Thomas François, Delphine Bernhard, Núria Gala, and Anne-Laure Ligozat. 2016. Are cohesive features relevant for text readability evaluation? In *26th International Conference on Computational Linguistics (COLING 2016)*, pages 987–997.

Sara Tonelli, Alessio Palmero Aprosio, and Francesca Saltori. 2016. Simpitiki: a simplification corpus for italian. In *CLiC-it/EVALITA*.

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Language Resources and Evaluation Conference*, pages 5366–5377, Marseille, France. European Language Resources Association.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the seventh workshop on building educational applications using NLP*, pages 163–173.

Sowmya Vajjala and Detmar Meurers. 2014. Assessing the relative reading level of sentence pairs for text simplification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 288–297.

Massimo Vedovelli and Tullio De Mauro. 1999. *Dante, il gendarme e la bolletta: la comunicazione pubblica in Italia e la nuova bolletta Enel*. Laterza.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

A Additional Information on the Datasets

Operation	*# operations*
Split	**18**
Reordering	**20**
Merging	**0**
Insert	**27**
Verb	5
Subject	1
Other	21
Delete	**33**
Verb	1
Subject	1
Other	31
Transformation	**490**
Lexical Substitution (word level)	253
Lexical Substitution (phrase level)	184
Anaphoric replacement	3
Noun to Verb	32
Verbal Voice	1
Verbal Features	17
Total	**588**

Table 7: The operations applied in Admin-It$_{OP}$ (Tonelli et al., 2016).

Table 5 presents some quantitative data for the different subsections of Admin-It and the datasets used for fine-tuning the NPRMs. Table 6 shows some pairs of sentences extracted from Admin-It, one for each simplification type. Finally, Table 7 shows all the operations applied in Admin-It$_{OP}$.

B Cross-domain scenario in English

Test set	*OSE* (*BERT*)	*OSE* (*mBERT*)
SimPA	0.625	0.771
SimPA$_{LS}$	0.643	0.793
SimPA$_{SS}$	0.604	0.682
SimPA$_{LS-SS}$	0.599	0.800

Table 8: The ranking accuracy achieved fine-tuning on OSE two different NPRMs: one based on *BERT*, trained only on English texts, and the other one based on *mBERT*, trained on texts in several languages.

We conduced some preliminary experiments on NPRM at the sentence level. Firstly, we fine-tuned and tested the model based on `bert-base-uncased` on in-domain data, i.e., an English news corpus, OSE. Testing it via 5-Fold cross validation, we obtained a quite high ranking accuracy (0.959)[14]. Then, we analyzed the model behavior in a cross-domain scenario on English (see Table 8). We fine-tuned the NPRM on OSE, and tested it on an English administrative corpus, SimPA. Firstly, we used OSE to fine-tune `bert-base-uncased`, the pre-trained base *BERT* model on English. As expected, the domain difference affected the ranking accuracy (0.625). However, the domain shift is much better handled by the model when fine-tuned on a multilingual pre-trained model, even though both training and test set are in English. The total ranking accuracy achieved using `bert-multilingual-base-uncased` is 0.771. The obtained model improved of around 0.14 points in ranking accuracy. Moreover, differently from SimPA$_{LS}$, where only a lexical simplification was applied, for SimPA$_{SS}$ a lower improvement is registered (0.078): the simplified sentences here have been manipulated on both lexical and syntactic levels, and recognizing the simple-to-read sentence results in an easier task. Finally, the highest improvement is registered for SimPA$_{LS-SS}$, where sentence pairs are composed by sentences simplified only at the lexical level and sentences simplified both at the lexical and syntactic levels (0.201).

[14]This experiment is also reported in Sec. 4.

C Additional results

In Table 9 are reported results obtained by adding in-domain data (SimPA), Italian data in the educational domain (Simpitiki$_W$), and both of them, to datasets in the news domain in English and Spanish (OSE, NewsEn, and NewsEs). Some of the results are shown also in Sec. 6, but are reported here to ease a comparison between the models.

D Results for each simplification operation

As described in Section 7.2, we analyzed the results obtained by some of the fine-tuned models on Admin-It$_{OP}$, the Admin-It subset where the original-simplified pairs of sentences are rewritten by applying only one operation. The models selected for this analysis are those fine-tuned on a single corpus (i.e., Simpitiki$_W$, OSE, NewsEn, NewsEs, and SimPA) and the best performing ones (i.e., NewsEn+NewsEs+OSE, OSE+NewsEs, OSE+NewsEs+SimPA, and OSE+SimPA). Results are reported in Table 10 and plotted in Figure 2 (Sec. 7.2).

Dataset	*# pairs*	*Min Lev*	*Avg Lev*	*Max Lev*	*Min Length*	*Avg Lenght*	*Max Lenght*
Admin-It	736	9	49.60	560	23	238.68	951
- Admin-It$_{OP}$	588	1	13.64	199	23	204.24	548
- Admin-It$_{RS}$	100	29	202.12	560	65	425.50	951
- Admin-It$_{RD}$	48	9	172.29	559	37	271.35	820
OSE	5994	1	26.59	194	15	129.34	425
NewsEn	488390	1	83.00	752	2	102.79	798
NewsEs	52048	1	93.28	510	7	134.18	601
Simpitiki$_{W}$	568	2	14.01	99	25	396.33	3646
SimPA	4637	1	34.73	287	8	161.38	463

Table 5: Details about number of pairs, Levenshtein distance, and length in characters concerning the Admin-It corpus and its subsets, and all the other datasets used in our experiments.

Admin-It$_{OP}$
Original: *La perdita del requisito della residenza nel Comune di Trento, comporta la cancellazione della domanda di ammissione al nido e il mancato inserimento della stessa nella graduatoria.* [Loss of the requisite of residency in the Municipality of Trento entails the cancellation of the application for admission to the nursery school and its non-inclusion in the ranking list.]
Simple: *Non avere più la residenza nel Comune di Trento comporta la cancellazione della domanda di ammissione al nido e il mancato inserimento della stessa nella graduatoria.* [If you no longer reside in the Municipality of Trento, your application for admission to the nursery school is cancelled and you are not included in the ranking list.]
Admin-It$_{RS}$
Original: *L'interessato a esercitare il trasporto di animali vivi, equini, bovini, bufalini, ovini, caprini, suini, e degli animali da cortile a mezzo autoveicolo deve presentare all'Ufficio Relazioni con il Pubblico (Urp) del Comune o all'Ufficio Commercio Denuncia inizio attività (Dia) per il trasporto di animali vivi in triplice copia, utilizzando l'apposito modulo scaricabile da questa pagina oppure in distribuzione presso l'Ufficio Commercio e l'Urp, in orario di apertura, allegando la fotocopia del libretto di circolazione.* [Anyone interested in transporting live animals, equines, cattle, buffaloes, sheep, goats, pigs and farmyard animals by motor vehicle must submit a triple copy of the Denuncia inizio attività (Dia) for the transport of live animals to the Public Relations Office (Urp) of the Municipality or to the Trade Office, using the appropriate form that can be downloaded from this page or is distributed at the Trade Office and Urp, during opening hours, enclosing a photocopy of the vehicle registration certificate.]
Simple: *Chi intende trasportare con un'auto o un veicolo animali vivi, come cavalli, buoi, bufali, pecore, capre e maiali (o altri animali da cortile), deve presentare la Denuncia Inizio Attività (Dia) per il trasporto di animali vivi. La Dia deve essere presentata in tre copie all'Ufficio Relazioni con il Pubblico (Urp) del Comune o presso l'Ufficio Commercio. Il modulo è scaricabile da questa pagina, ma è anche distribuito dall'Ufficio Commercio e dall' Urp, durante l'orario di apertura. Insieme al modulo va consegnata una copia del libretto di circolazione.* [Anyone who intends to transport live animals, such as horses, oxen, buffaloes, sheep, goats and pigs (or other farmyard animals) in a car or vehicle must submit a Denuncia Inizio Attività (Dia) for the transport of live animals. The Dia must be submitted in three copies to the Municipality's Public Relations Office (Urp) or to the Trade Office. The form can be downloaded from this page, but is also distributed by the Commerce Office and the Urp, during opening hours. A copy of the vehicle registration certificate must be handed in together with the form.]
Admin-It$_{RD}$
Original: *Al fine di verificare, prima di una eventuale assegnazione, la permanenza dei requisiti previsti dalla legge, si invita la S.V. a contattare con urgenza l'Ufficio Domanda del Settore Edilizia residenziale telefonando al n. 000/1112223 o al n. 000/1112223, oppure presentandosi presso la sede - via S. Martino e Solferino 00 - negli orari di ricevimento al pubblico (lunedì, mercoledì dalle ore 10.00 alle ore 12.00 e giovedì dalle ore 15.15 alle 17.15).* [In order to verify, before a possible assignment, the permanence of the statutory requisites, we kindly ask you to urgently contact the Office for Applications of the Residential Building Sector by phoning 000/1112223 or 000/1112223, or by coming to the office - via S. Martino e Solferino 00 - during the public reception hours (Mondays, Wednesdays from 10.00 to 12.00 and Thursdays from 15.15 to 17.15).]
Simple: *È necessario verificare che lei sia ancora in possesso dei requisiti previsti dalla legge. Per questo la invitiamo a telefonare con urgenza al numero 000 1112223 o allo 000 1112223, oppure a venire all'Ufficio Domanda del Settore Edilizia residenziale, in via S. Martino e Solferino 00 (il lunedì e mercoledì dalle 10 alle 12, o il giovedì dalle 15.15 alle 17.15).* [It is necessary to check that you still meet the legal requirements. For this reason, we invite you to urgently call 000 1112223 or 000 1112223, or come to the Office for Applications of the Residential Building Sector, in via S. Martino e Solferino 00 (on Mondays and Wednesdays from 10 a.m. to 12 noon, or on Thursdays from 3.15 p.m. to 5.15 p.m.).]

Table 6: Examples of pairs of sentences in Admin-It subsets.

Test set	*OSE*	*NewsEn*	*NewsEs*	*OSE+NewsEn*	*OSE+NewsEs*	*OSE+NewsEn+NewsEs*
Admin-It	0.777	0.765	0.760	0.742	0.785	0.783
- Admin-It$_{OP}$	0.745	0.731	0.716	0.699	0.743	0.748
- Admin-It$_{RS}$	0.970	0.960	0.970	0.960	0.980	0.990
- Admin-It$_{RD}$	0.771	0.771	0.854	0.813	**0.896**	0.771
				+SimPA		
Admin-It	0.787	0.784	0.791	0.792	**0.803**	0.766
- Admin-It$_{OP}$	0.747	0.760	0.762	0.760	**0.767**	0.736
- Admin-It$_{RS}$	**1.000**	0.970	0.980	0.970	0.980	0.990
- Admin-It$_{RD}$	0.833	0.688	0.750	0.813	0.875	0.667
				+Simpitiki$_W$		
Admin-It	0.774	0.765	0.724	0.734	0.754	0.753
- Admin-It$_{OP}$	0.741	0.724	0.675	0.682	0.704	0.713
- Admin-It$_{RS}$	0.970	0.980	0.960	0.960	0.980	0.960
- Admin-It$_{RD}$	0.771	0.813	0.833	**0.896**	**0.896**	0.813
				+SimPA & Simpitiki$_W$		
Admin-It	0.764	0.788	0.758	0.750	0.774	0.754
- Admin-It$_{OP}$	0.716	0.752	0.716	0.713	0.733	0.709
- Admin-It$_{RS}$	0.990	0.980	0.970	0.950	0.980	0.990
- Admin-It$_{RD}$	0.875	0.833	0.833	0.792	0.854	0.813

Table 9: The ranking accuracy achieved by NPRMs fine-tuned on OSE, NewsEn, NewsEs and their combinations. The second section shows the results when SimPA is added to the previous setting; in the third, Simpitiki$_W$ was added to the corpora of the first section; in the fourth, both Simpitiki$_W$ and SimPA were added for fine-tuning. In bold the best results achieved for each subsection of Admin-It and for the overall test set.

Operation	*Simpitiki$_W$*	*OSE*	*NewsEn*	*NewsEs*	*SimPA*	*OSE+SimPA*
Split	0.778	0.556	0.667	0.444	1.000	1.000
Reordering	0.500	0.600	0.300	0.700	0.100	0.150
Insert - Verb	0.000	0.000	0.000	0.000	0.000	0.000
Insert - Subject	1.000	1.000	0.000	0.000	1.000	1.000
Insert - Other	0.333	0.238	0.476	0.381	0.048	0.095
Delete - Verb	1.000	1.000	1.000	1.000	1.000	1.000
Delete - Subject	1.000	1.000	0.000	1.000	1.000	1.000
Delete - Other	0.968	0.871	0.774	0.839	0.935	0.871
Lexical Substitution (word level)	0.601	0.802	0.747	0.708	0.688	0.787
Lexical Substitution (phrase level)	0.690	0.783	0.810	0.810	0.777	0.793
Anaphoric replacement	0.333	1.000	0.667	0.667	0.667	1.000
Noun to Verb	0.625	0.500	0.781	0.656	0.625	0.781
Verbal Voice Transformation	0.000	1.000	1.000	1.000	1.000	1.000
Verbal Features Transformation	0.647	0.824	0.647	0.647	0.588	0.706

Operation	*NewsEn+NewsEs+OSE*	*OSE+NewsEs+SimPA*	*NewsEs+OSE*
Split	0.778	0.833	0.444
Reordering	0.450	0.500	0.750
Insert - Verb	0.400	0.000	0.000
Insert - Subject	1.000	1.000	0.000
Insert - Other	0.476	0.190	0.143
Delete - Verb	1.000	1.000	1.000
Delete - Subject	1.000	1.000	1.000
Delete - Other	0.710	0.871	0.871
Lexical Substitution (word level)	0.802	0.798	0.806
Lexical Substitution (phrase level)	0.783	0.826	0.788
Anaphoric replacement	1.000	0.333	0.667
Noun to Verb	0.563	0.719	0.594
Verbal Voice Transformation	1.000	1.000	1.000
Verbal Features Transformation	0.647	0.765	0.647

Table 10: The ranking accuracy achieved on each operation applied in Admin-It$_{OP}$ by NPRMs based on *mBERT* and fine-tuned with OSE, NewsEn and NewsEs, SimPA, Simpitiki$_W$, and their combinations.

E Statistical Significance Testing

In Figure 3, the heatmap shows the p-values computed with McNemar's Test by comparing model's performances on Admin-It$_{OP}$, Admin-It$_{RS}$, and Admin-It$_{RD}$. Numeric values are shown in Table 11 for the overall Admin-It. P-values for Admin-It$_{OP}$ are shown in Table 12, and the p-values computed on Admin-It$_{RS}$ and Admin-It$_{RD}$ are shown in Table 14 and Table 13, respectively.

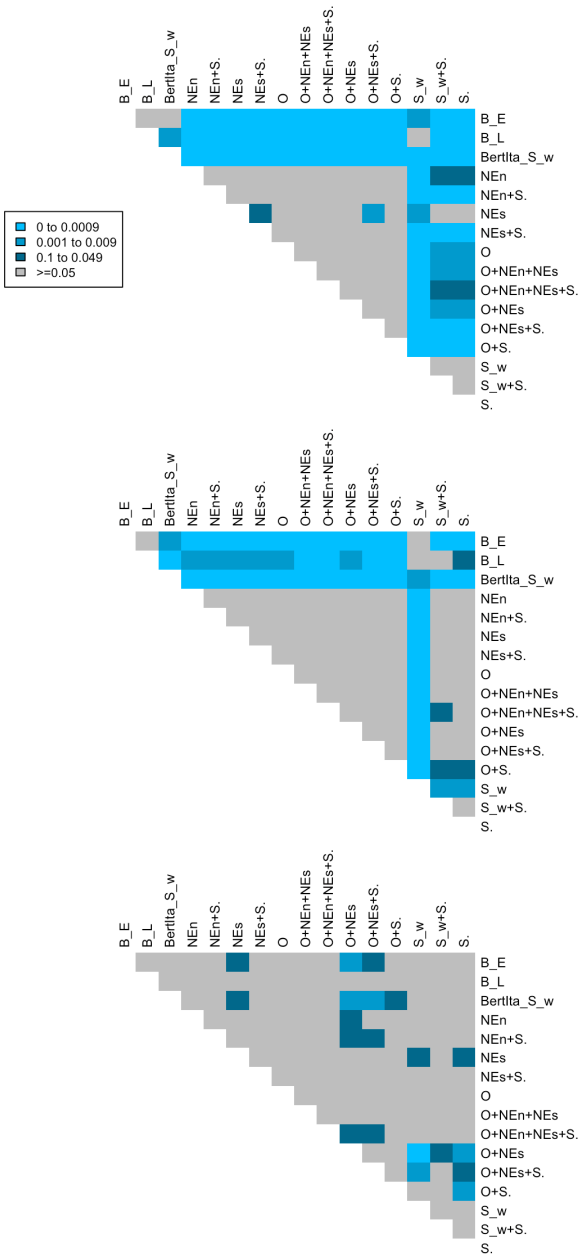

Figure 3: The heatmaps show the p-values obtained with McNemar's Test for pairs of models. From top to bottom: Admin-It$_{OP}$, Admin-It$_{RS}$, and Admin-It$_{RD}$. Grey cells represent a p-value equal or higher than 0.05. We tested the performances of Baseline$_L$ (B$_L$), Baseline$_E$ (B$_E$), NewsEn (NEn), NewsEs (NEs), SimPA (S.), Simpitiki$_W$ (S$_W$), OSE (O), and their combinations.

	B_E	B_L	BertIta$-S_w$	NEn	NEn+S.	NEs	NEs+S.	O
B_E	0							
B_L	<0.05	0						
BertIta$-S_w$	<0.01	<0.001	0					
NEn	<0.001	<0.001	<0.001	0				
NEn+S.	<0.001	<0.001	<0.001	0.207	0			
NEs	<0.001	<0.001	<0.001	0.821	0.22	0		
NEs+S.	<0.001	<0.001	<0.001	0.169	0.754	<0.05	0	
O	<0.001	<0.001	<0.001	0.515	0.75	0.344	0.474	0
O+NEn+NEs	<0.001	<0.001	<0.001	0.294	1	0.207	0.691	0.811
O+NEn+NEs+S.	<0.001	<0.001	<0.001	1	0.309	0.764	0.173	0.617
O+NEs	<0.001	<0.001	<0.001	0.267	1	0.051	0.779	0.642
O+NEs+S.	<0.001	<0.001	<0.001	<0.05	0.319	<0.01	0.417	0.099
O+S.	<0.001	<0.001	<0.001	0.244	0.937	0.143	0.853	0.576
S.	<0.001	<0.001	<0.001	<0.05	<0.001	<0.05	<0.001	<0.001
S_w	<0.01	0.367	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
S_w+S.	<0.001	<0.001	<0.001	<0.01	<0.001	<0.05	<0.001	<0.01

	O+NEn+NEs	O+NEn+NEs+S.	O+NEs	O+NEs+S.	O+S.	S.	S_w	S_w+S.
O+NEn+NEs	0							
O+NEn+NEs+S.	0.266	0						
O+NEs	0.933	0.33	0					
O+NEs+S.	0.251	0.056	0.16	0				
O+S.	0.875	0.29	1	0.299	0			
S.	<0.001	<0.05	<0.001	<0.001	<0.001	0		
S_w	<0.001	<0.001	<0.001	<0.001	<0.001	<0.01	0	
S_w+S.	<0.001	<0.05	<0.001	<0.001	<0.001	0.927	<0.01	0

Table 11: The p-values computed with McNemar's test to compare the performances reached on the whole dataset of Admin-It by Baseline$_L$ (B_L), Baseline$_E$ (B_E), NewsEn (NEn), NewsEs (NEs), SimPA (S.), Simpitiki$_W$ (S_W), OSE (O), and their combinations.

	B_E	B_L	BertIta$-S_w$	NEn	NEn+S.	NEs	NEs+S.	O
B_E	0							
B_L	0.192	0						
BertIta-S_w	0.061	<0.01	0					
NEn	<0.001	<0.001	<0.001	0				
NEn+S.	<0.001	<0.001	<0.001	0.097	0			
NEs	<0.001	<0.001	<0.001	0.526	0.055	0		
NEs+S.	<0.001	<0.001	<0.001	0.168	1	<0.05	0	
O	<0.001	<0.001	<0.001	0.551	0.494	0.184	0.437	0
O+NEn+NEs	<0.001	<0.001	<0.001	0.407	0.589	0.138	0.56	0.934
O+NEn+NEs+S.	<0.001	<0.001	<0.001	0.864	0.251	0.375	0.238	0.761
O+NEs	<0.001	<0.001	<0.001	0.621	0.459	0.094	0.32	1
O+NEs+S.	<0.001	<0.001	<0.001	0.099	0.806	<0.01	0.826	0.237
O+S.	<0.001	<0.001	<0.001	0.512	0.56	0.171	0.439	1
S.	<0.001	<0.001	<0.001	<0.05	<0.001	0.171	<0.001	<0.01
S_w	<0.01	0.073	<0.001	<0.001	<0.001	<0.01	<0.001	<0.001
S_w+S.	<0.001	<0.001	<0.001	<0.05	<0.001	0.11	<0.001	<0.01

	O+NEn+NEs	O+NEn+NEs+S.	O+NEs	O+NEs+S.	O+S.	S.	S_w	S_w+S.
O+NEn+NEs	0							
O+NEn+NEs+S.	0.51	0						
O+NEs	0.859	0.81	0					
O+NEs+S.	0.393	0.182	0.12	0				
O+S.	1	0.693	0.925	0.271	0			
S.	<0.01	<0.05	<0.01	<0.001	<0.001	0		
S_w	<0.001	<0.001	<0.001	<0.001	<0.001	0.05	0	
S_w+S.	<0.01	<0.05	<0.01	<0.001	<0.001	0.693	0.094	0

Table 12: The p-values computed with McNemar's test to compare the performances reached on Admin-It$_{OP}$ by Baseline$_L$ (B_L), Baseline$_E$ (B_E), NewsEn (NEn), NewsEs (NEs), SimPA (S.), Simpitiki$_W$ (S_W), OSE (O), and their combinations.

	B_E	B_L	BertIta$-S_w$	NEn	NEn+S.	NEs	NEs+S.	O
B_E	0							
B_L	0.11	0						
BertIta-S_w	<0.01	<0.001	0					
NEn	<0.001	<0.01	<0.001	0				
NEn+S.	<0.001	<0.01	<0.001	1	0			
NEs	<0.001	<0.01	<0.001	1	1	0		
NEs+S.	<0.001	<0.01	<0.001	0.688	1	1	0	
O	<0.001	<0.01	<0.001	1	1	1	1	0
O+NEn+NEs	<0.001	<0.001	<0.001	0.375	0.625	0.625	1	0.625
O+NEn+NEs+S.	<0.001	<0.001	<0.001	0.375	0.5	0.625	1	0.625
O+NEs	<0.001	<0.01	<0.001	0.688	1	1	1	1
O+NEs+S.	<0.001	<0.001	<0.001	0.688	1	1	1	1
O+S.	<0.001	<0.001	<0.001	0.125	0.25	0.25	0.5	0.25
S.	<0.001	<0.05	<0.001	0.727	0.375	0.453	0.219	0.453
S_w	0.5	0.473	<0.01	<0.001	<0.001	<0.001	<0.001	<0.001
S_w+S.	<0.001	0.093	<0.001	0.508	0.219	0.289	0.125	0.289

	O+NEn+NEs	O+NEn+NEs+S.	O+NEs	O+NEs+S.	O+S.	S.	S_w	S_w+S.
O+NEn+NEs	0							
O+NEn+NEs+S.	1	0						
O+NEs	1	1	0					
O+NEs+S.	1	1	1	0				
O+S.	1	1	0.5	0.5	0			
S.	0.125	0.062	0.289	0.219	<0.05	0		
S_w	<0.001	<0.001	<0.001	<0.001	<0.001	<0.01	0	
S_w+S.	0.07	<0.05	0.18	0.18	<0.05	1	<0.01	0

Table 13: The p-values computed with McNemar's test to compare the performances reached on Admin-It$_{RS}$ by Baseline$_L$ (B_L), Baseline$_E$ (B_E), NewsEn (NEn), NewsEs (NEs), SimPA (S.), Simpitiki$_W$ (S_W), OSE (O), and their combinations.

	B_E	B_L	BertIta$-S_w$	NEn	NEn+S.	NEs	NEs+S.	O
B_E	0							
B_L	0.143	0						
BertIta-S_w	1	0.096	0					
NEn	0.263	1	0.167	0				
NEn+S.	0.832	0.332	0.678	0.344	0			
NEs	<0.05	0.607	<0.05	0.344	0.077	0		
NEs+S.	0.359	0.804	0.21	1	0.549	0.18	0	
O	0.238	1	0.21	1	0.481	0.344	1	0
O+NEn+NEs	0.263	1	0.167	1	0.424	0.344	1	1
O+NEn+NEs+S.	1	0.238	0.824	0.332	1	0.064	0.388	0.359
O+NEs	<0.01	0.267	<0.01	<0.05	<0.05	0.625	0.065	0.109
O+NEs+S.	<0.05	0.424	<0.01	0.062	<0.05	1	0.146	0.227
O+S.	0.064	0.791	<0.05	0.581	0.065	1	0.289	0.581
S.	1	0.238	0.839	0.302	1	<0.05	0.344	0.302
S_w	1	0.21	0.824	0.267	1	<0.05	0.454	0.359
S_w+S.	0.332	0.815	0.263	1	0.607	0.267	1	1

	O+NEn+NEs	O+NEn+NEs+S.	O+NEs	O+NEs+S.	O+S.	S.	S_w	S_w+S.
O+NEn+NEs	0							
O+NEn+NEs+S.	0.267	0						
O+NEs	0.109	<0.05	0					
O+NEs+S.	0.18	<0.05	1	0				
O+S.	0.581	0.057	0.508	0.754	0			
S.	0.302	1	<0.01	<0.05	<0.01	0		
S_w	0.302	1	<0.001	<0.01	0.077	1	0	
S_w+S.	1	0.481	<0.05	0.109	0.388	0.344	0.481	0

Table 14: The p-values computed with McNemar's test to compare the performances reached on Admin-It$_{RD}$ by Baseline$_L$ (B_L), Baseline$_E$ (B_E), NewsEn (NEn), NewsEs (NEs), SimPA (S.), Simpitiki$_W$ (S_W), OSE (O), and their combinations.