Monolingual and Cross-Lingual Acceptability Judgments with the Italian CoLA corpus

Daniela Trotta[†], **Raffaele Guarasci**[‡], **Elisa Leonardelli**[§], **Sara Tonelli**[§]
[†]University of Salerno, Italy, [‡]ICAR-CNR, Naples, Italy
[§]Fondazione Bruno Kessler, Trento, Italy
dtrotta@unisa.it
raffaele.guarasci@icar.cnr.it
{eleonardelli,satonelli}@fbk.eu

Abstract

The development of automated approaches to linguistic acceptability has been greatly fostered by the availability of the English CoLA corpus, which has also been included in the widely used GLUE benchmark. However, this kind of research for languages other than English, as well as the analysis of cross-lingual approaches, has been hindered by the lack of resources with a comparable size in other languages. We have therefore developed the ItaCoLA corpus, containing almost 10,000 sentences with acceptability judgments, which has been created following the same approach and the same steps as the English one. In this paper we describe the corpus creation, we detail its content, and we present the first experiments on this new resource. We compare in-domain and out-of-domain classification, and perform a specific evaluation of nine linguistic phenomena. We also present the first cross-lingual experiments, aimed at assessing whether multilingual transformer-based approaches can benefit from using sentences in two languages during fine-tuning.

1 Introduction

The ability to judge whether a sentence is perceived as natural and well-formed by a native speaker is called acceptability judgment. Despite several open issues concerning methods for collecting and evaluating them (Gibson and Fedorenko, 2013; Sprouse and Almeida, 2010; Linzen, 2019), these judgments have been the most significant source of data in linguistics throughout the history of the discipline (Chomsky, 1965; Schütze, 2016; Dabrowska, 2010).

With the rise of neural language models, several works have tried to assess how much a model can encode linguistic information (Hewitt and Manning, 2019; Manning et al., 2020), ranging from specific phenomena (Marvin and Linzen, 2019; Goldberg, 2019) to a general grammar knowledge (Jawahar et al., 2019; McCoy et al., 2020). Acceptability judgments have proven to be a promising area to test the acquisition of linguistic knowledge by neural language models (Gulordava et al., 2018; Lau et al., 2015). In particular, with the creation of the Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2019) several approaches have been proposed that cast acceptability as a binary classification task and address it by fine-tuning transformer-based models on the corpus (Yang et al., 2019; Warstadt and Bowman, 2019; Raffel et al., 2020). Unfortunately, most classification experiments on acceptability judgments have focused on English, mainly because of the lack of large corpora in other languages. In this work, we therefore describe the creation of a novel corpus of acceptability judgments in Italian, following the methodology used in CoLA for English. We collect 10k sentences extracted from linguistic literature and labelled by experts as acceptable or not. Furthermore, we enrich around 30% of the sentences with additional labels describing nine linguistic phenomena. We also present a set of experiments aimed at testing the performance of a BERT-based classifier on the data and comparing it with results obtained on English. Additionally, cross-lingual experiments using XLM-RoBERTa (Conneau et al., 2020) show the potential of this approach, even if it is outperformed by monolingual models. The main contributions of this work are therefore *i)* the creation and release of the Italian Corpus of Acceptability Judgments (ItaCoLA),[1] that to our knowledge is the largest resource of its kind in a language other than English; *ii)* a set of experiments to assess the performance of BERT-based models on the whole corpus and on specific phenomena. *iii)* a set of experiments using a massive multilingual language model on Italian and English, with the potential to open up novel cross-language

[1]Available at https://github.com/dhfbk/ItaCoLA-dataset

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2929–2940
November 7–11, 2021.

research perspectives.

2 Related Work

2.1 Acceptability corpora

In recent years, studies on automatic assessment of acceptability have become very popular thanks to the release of the CoLa corpus (Warstadt et al., 2019), the first large-scale corpus of English acceptability, containing more thank 10k sentences taken from linguistic literature.

Small acceptability datasets had already been developed before, especially within the still open theoretical debate regarding the status of syntax (Sprouse and Almeida, 2013; Lau et al., 2014) and data collection methods (Culicover and Jackendoff, 2010; Gibson and Fedorenko, 2013). These resources differ in terms of (formal or informal) data collection criteria, sources, evaluation methodology and raters used in the process.

In particular, (Sprouse et al., 2013a), advocating the empirical status of syntax and the reliability of informal collection of acceptability judgments, test a random sample of 300 sentence extracted from the 'Linguistic Inquiry' journal. Annotators were recruited through Amazon Mechanical Turk, which had proven reliable for this type of task (Sprouse, 2011). Evaluation is performed using two experimental methods: magnitude estimation and forced-choice task. These judgments are also compared with ones collected using formal methods (Sprouse et al., 2013b) manually evaluated using a combination of AMT and naive participants without linguistic training.

(Lau et al., 2014) collect a dataset of 600 sentences from the BNC (Consortium et al., 2007), and then introduce infelicities using machine translation to generate sentences of varying level of grammaticality. Judgments have been collected using AMT and applying different evaluation criteria, from binary to gradient. A recent study (Marvin and Linzen, 2019), aimed at evaluating the behaviour of a neural model on specific syntactic phenomena, uses a dataset of sentence pairs automatically built using templates.

Regarding studies on languages other than English, (Linzen and Oseki, 2018) evaluate informal acceptability judgments on Hebrew and Japanese collecting data from several sources ranging from peer-reviewed papers, books and dissertations. A similar study has been conducted in French (Feldhausen and Buchczyk, 2020) and in Chinese (Chen et al., 2020). Both studies use sentences extracted from textbooks. To our knowledge, only for Swedish there is a freely-available corpus whose size is comparable to CoLA and ItaCoLA. The corpus, presented in (Volodina et al., 2021), contains around 9,600 sentences extracted from language learner data.

Concerning Italian, only one dataset has been released to date, in the context of Evalita 2020 evaluation campaign on complexity and acceptability (AcComplIt task) (Brunato et al., 2020). The dataset presents several differences w.r.t. ItaCoLA in terms of size, annotation approach and linguistic phenomena, which we detail in Section 4.2.

2.2 Approaches to acceptability classification

The CoLA corpus was presented together with a number of experiments aimed at assessing the performance of neural networks on a novel binary acceptability task (Warstadt et al., 2019). The best performance was achieved with a pooling classifier and ELMo-style embeddings, yielding 0.341 MCC on in-domain data and 0.281 on the out-of-domain test set. Matthews Correlation Coefficient (MCC) was chosen as an evaluation measure because it is more appropriate than F1 or accuracy for binary classification with unbalanced data (Matthews, 1975). More recently, (Warstadt and Bowman, 2019) extended the classification experiments by comparing a BiLSTM baseline with the performance achieved by transformer encoders such as GPT and BERT. The best approach is obtained by fine-tuning BERT$_{large}$ with a mean MCC of 0.582. Other approaches, instead, focus on unsupervised learning, for example (Lau et al., 2015, 2020) compare different types of language models to infer the probability of a sentence, which is then mapped onto acceptability.

Since CoLA has been included in the GLUE dataset (Wang et al., 2018), a very popular multitask benchmark for English natural language understanding, and an acceptability challenge has been launched on Kaggle,[2] the number of studies dealing with binary acceptability has remarkably increased. Unfortunately, most studies using GLUE report accuracy instead of MCC, making it difficult to identify the best approach. Nevertheless, all top-ranked systems rely on variations of transformer-based models, including ALBERT

[2] https://www.kaggle.com/c/cola-in-domain-open-evaluation/

(Lan et al., 2020) (69.1 Accuracy) and StructBERT (Wang et al., 2020) (69.2 Acc.). More recently, also reformulating acceptability as an entailment task and using smaller language models to few-shot fine-tuning has showed a great potential (Wang et al., 2021), outperforming existing BERT-based approaches (86.4 Acc.).

Concerning acceptability on Italian, a shared task has been organised for the first time at Evalita 2020 Evaluation campaign, proposing a joint classification of complexity and acceptability (Brunato et al., 2020). The dataset, which we use for our out-of-domain evaluation (Section 4.2) was originally created merging data from different psycholinguistic studies, and includes 1,683 sentences with a manually assigned value of acceptability between 1 and 7. Two participants submitted three runs in total. In order to cope with the limited number of training instances, the best performing approach (Sarti, 2020) implemented an ensemble of fine-tuned models to annotate a large corpus of unlabeled text, and leveraged new annotations in a multi-task setting to obtain final predictions over the original test set. The second system (Delmonte, 2020) was rule-based, implementing a set of syntactic and semantic constraints to check to what an extent a sentence can be considered acceptable.

3 ItaCoLA: Italian Corpus of Linguistic Acceptability

In this section we introduce the Italian Corpus of Linguistic Acceptability (ItaCoLA), built with the purpose of representing a large number of linguistic phenomena while distinguishing between acceptable and not acceptable sentences. The methodology of corpus creation and its size are similar to those proposed for the English CoLA in (Warstadt et al., 2019), i.e. we have collected examples from different manuals covering several linguistic phenomena. This fulfills a dual purpose: the size of the corpus allows the application of deep learning approaches to acceptability judgment, while its structure paves the way to cross-language comparative analyses.

Concerning acceptability annotation, for the creation of ItaCoLA we have chosen to keep a Boolean judgment in line with several previous works (Lawrence et al., 2000; Wagner et al., 2009; Linzen et al., 2016). This choice ensures robustness and simplifies classification, while allowing us to keep the original judgments as formulated by an expert (i.e. the authors of the different data sources).

3.1 Sources

ItaCoLA sentences come from various types of linguistic publications covering four decades. Unfortunately, the majority of linguistic textbooks or fundamental theoretical publications in Italian are not available in digital format or are not freely accessible. Therefore, the only viable way to collect data was through manual transcription. Sources include theoretical linguistics textbooks (Graffi and Scalise, 2002; Simone and Masini, 2013) and works that focus on specific phenomena such as idiomatic expressions (Vietri, 2014), locative constructions (D'Agostino, 1983) and verb classification (Jezek, 2003). Overall, we manually copied from a number of sources a total of 10,000 sentences, reporting also the judgment provided by the author, i.e. acceptable or not acceptable. Few examples are listed in Table 1.

3.2 Sentence selection

Following the criteria proposed by (Warstadt et al., 2019), specific choices have been made to exclude some types of sentences from the corpus. This increases data consistency also for future cross-lingual experiments with CoLA. Following sentence types were not included in the dataset:

- Italian translations of sentences, which were originally written in other languages. The syntactic behavior of each language can cause ambiguity in judging the acceptability of translated sentences.
- Isolated phrases without predicative structure or full meaning expression, i.e. noun, prepositional, adjective and adverbial phrases.
- Sentences which are difficult to evaluate without context even by a native speaker. This category includes sentences that are strictly domain-dependent, for instance statements of linguistic rules such as "Una testa lessicale -N assegna Caso al SN che essa regge" (En. *A lexical head -N assigns Case to the SN that it holds*) or sentences extracted from novels, films or newspapers.
- Sentences with an extremely twisted syntax and a very high number of nested subordinates. The latter are often used as borderline

Source	Label	Sentence
Graffi (1994)	0	*Edoardo è tornato nella sua l'anno scorso città. (*Edoardo returned to his last year city*)
Graffi (1994)	1	Ho voglia di salutare Maria (*I want to greet Maria*)
Graffi (1994)	0	*Questa donna mi hanno colpito. (**This woman have impressed me*)
Simone and Masini (2013)	1	Questa donna mi ha colpito. (*This woman has impressed me*)

Table 1: Example sentences from the ItaCoLA dataset. 1 = acceptable, 0 = not acceptable

examples to explain phenomena such as long-distance dependencies or pro-drop, very common in Italian.

Concerning the types of sentences which have been included in the dataset, we can identify some recurring patterns. For example, there is the presence of several minimal pairs, i.e. minimally different sentences contrasting in acceptability (see the sentences in the last two rows of Table 1). Other sentences are short examples created to describe or explain specific phenomena, i.e. "Lucia lavora per studiare" (En. *Lucia works to study*) (Vietri, 2004), or elementary sentences whose syntax matches the canonical SVO (Subject-Verb-Object) order, i.e. "Il poliziotto catturò il ladro" (En. *The policeman caught the thief*). Other common sentence types are those resulting from formal transformation tests. For instance, starting from the elementary sentence "I bambini hanno calpestato le aiuole" (*Children stepped the flowerbeds*), other sentences can be produced by applying deletion, i.e. "I bambini hanno calpestato" (*children stepped*), or pronominalization, i.e. "I bambini le hanno calpestate" (En. *Children stepped on them*). These transformed sentences – besides being in conspicuous number in the corpus – are functional to the purpose of this work, since they are created just to verify whether native speaker intuition is validated by the data.

3.3 Data Cleaning and Refinement

Once the sentences have been selected, some further adjustments have been made at lexical level in order to prevent possible ambiguity and make some outdated examples sound more modern. Also in this case, we follow the same principles used for CoLA. Changes have involved mainly proper nouns and verbs and have been carried out to avoid irrelevant complications due to out-of-vocabulary words:

- Obsolete or uncommon proper nouns and abbreviations of organisations (i.e., Ena, Isa, Lillo, Pat etc.) have been replaced when possible with more common names taken from the lists released by the Italian National Institute of Statistics.[3] According to Vietri (2014) mentions of rare and obsolete named entities in sentences can interfere with acceptability judgments.

- Low-frequency terms, which in most cases pertain to the technical-specialist domain, have been manually simplified using synonyms or broader terms that made them easier to understand without affecting the semantics of the sentence. For instance the sentence "L'artrosi ha anchilosato le mani di Filippo." (En. *The arthrosis has developed ankylosis Filippo's hands*) has been changed to "L'artrosi ha paralizzato le mani di Raffaele." (En. *The arthrosis has paralyzed Raffaele's hands*).

In order to identify low-frequency terms, we lemmatised all sentences using the TINT NLP suite for Italian (Palmero Aprosio and Moretti, 2018), and then associated each lemma with the reference frequency list extracted from the Paisà corpus (Lyding et al., 2014). Words with a frequency < 45 were manually checked and, if possible, replaced with more frequent ones of similar meaning. In total, 130 sentences were modified in this way, while for another 17 sentences a rare word was detected but it was not possible to find a replacement without modifying the meaning of the sentence (or creating a sentence already existing in the dataset). We

[3]The data consulted are updated to 2018 according to the Italian National Institute of Statistics: https://www.istat.it/it/dati-analisi-e-prodotti/contenuti-interattivi/contanomi

Source	N	% acceptable	Topic
D'Agostino (1983)	524	84.2	locative constructions
D'Agostino (1992)	1,364	85.0	discourse analysis
Elia et al. (1981)	2,167	84.8	lexicon and syntactic structures
Elia (1982)	169	79.9	locative adverbs and idioms
Graffi and Scalise (2002)	157	84.1	theoretical linguistics
Graffi (1994)	604	79.5	syntax
Graffi (2008)	122	82.0	generative grammar
Jezek (2003)	817	74.4	verb classification
Simone and Masini (2013)	754	97.7	theoretical linguistics
Vietri (2014)	651	90.0	idiomatic expressions
Vietri (2004)	1,424	85.5	lexicon-grammar approach
Vietri (2017)	970	81.4	anticausative sentences
In-domain	**9,722**	**84.5**	

Table 2: Distribution of ItaCoLA sentences by source. N is the number of sentences from each source. Topic is the main focus of the source, even if other linguistic phenomena can be present as well.

therefore opted for leaving these few sentences in their original form.

An additional check was performed to manually control for typos and transcription errors. We observed that some sentences were present in more than one dataset, usually in case of multiple sources by the same author. Double sentences were thus removed (source was randomly chosen). The final dataset consists of 9,722 sentences from different sources, having each a different percentage of acceptable and not acceptable sentences, with a large prevalence of acceptable instances. An overview of the dataset is reported in Table 2.

4 Monolingual Experiments

The monolingual experiments are aimed at presenting the first classification results on ItaCoLA and at defining standard training, validation and test split, to be used also in future experiments with the corpus. We compare two classifiers: one using LSTM and FastText embeddings, which we consider our baseline, and the other using an Italian version of BERT (Devlin et al., 2019), which we fine-tune using ItaCoLA training dataset. The two classifiers are evaluated in an in-domain and an out-of-domain setting, similar to the evaluation performed on English CoLA. (Warstadt et al., 2019).

For the **in-domain evaluation**, we divide the ItaCoLA corpus into a training, a validation and a test split, including respectively 7,801, 946 and 975 examples. We create the splits so that each source is equally represented in each split and the acceptability/not acceptability ratio is preserved. For the **out-of-domain setting**, training is performed on the same split used for the in-domain experiments. Validation and test, instead, are carried out using the AcComplIt dataset (Brunato et al., 2020). In particular, for validation we use the training set released for the Evalita shared task and for testing we use the official AcComplIt test set. We consider this dataset out-of-domain not only because it comes from different sources compared to ItaCoLA, but also because it was created using crowd-sourcing, i.e. following a completely different approach than ours, which relies on linguistic literature.

Baseline LSTM: As baseline classifier, we implement a bidirectional LSTM with two layers (64 and 32 neurons) and a dropout of 0.3. Each sentence is represented as a sequence of word embeddings, obtained with the Italian model of FastText (Grave et al., 2018) trained on Common Crawl and Wikipedia with size 300.[4] The network is implemented with Keras (Chollet, 2017) (Adam optimizer, learning rate 0.01, loss function: binary crossentropy, 15 epochs). We perform 10 restarts. Reported results represent the mean performance obtained over the restarts.

BERT: Among the Italian BERT-like versions available, we select *Bert-base-italian-xxl-cased*, available on Huggingface.[5] It is a model pretrained on a total general-purpose corpus of 81GB. After randomizing the order of instances in our training set, we fine-tune the model using Py-

[4] https://github.com/facebookresearch/fastText/blob/master/docs/crawl-vectors.md

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

Source	N	% acceptable	Topic
Chesi and Canal (2019)	128	69.5	object clefts
Greco et al. (2020)	515	91.6	copular
Mancini et al. (2018)	320	49.7	subject-verb agreement
Villata et al. (2015)	48	66.7	wh-violations
Chowdhury and Zamparelli (2018)	672	53.6	various from templates
Out-of-domain	1,683	66.0	

Table 3: The content of AcComplIt dataset (Brunato et al., 2020) used for out-of-domain experiments

Torch,[6] with a maximum sequence length of 64, a batch size of 32 for 12 epochs. We perform 10 restarts. Also in this case, reported results are the mean across the repeated classifications.

4.1 In-domain results

Results on the in-domain test set are displayed in Table 4. We report both Matthews Correlation Coefficient (MCC) (Peters et al., 2018), which is the score originally proposed by CoLA authors, and Accuracy. MCC is a measure of correlation for Boolean variables and it is particularly suited when evaluating unbalanced binary classifiers. We report Accuracy as well, which is instead generally used to evaluate acceptability on the GLUE benchmark. Classification performance is in line with the results obtained for English, since Warstadt and Bowman (2019) report MCC = 0.582 (mean of 20 restarts) using BERT$_{large}$ and MCC = 0.320 with the LSTM baseline on in-domain data. In general, these results suggest that neural approaches applied to Italian can work with a performance similar to English, provided that the same amount of training data is available.

4.2 Out-of-domain results

Since acceptability in the AcComplIt dataset used for out-of-domain evaluation is labeled for perceived acceptability on a 7-point Likert scale, we first map these labels to two classes (i.e. acceptable or not) if the average score is $\geq$ 3.5 or below, respectively. We report statistics related to the composition of the dataset and the distribution of acceptable sentences in Table 3

Also in this case classification results are reported in Table 4. Similar to the in-domain data, the BERT-based classifier outperforms the LSTM baseline. However, results are much lower than those reported in the same setting for English, where the best result obtained with a pooling classifier and ELMo-style embeddings is MCC = 0.281. This difference is probably due to a number of factors, including the different approach followed to create the out-of-domain dataset, the fact that we mapped the Likert scale into two classes, and the presence of different linguistic phenomena. Another difference is the average sentence length: while it is 6 tokens in ItaCoLA, sentences in AcComplIt contain on average 10 tokens. Furthermore, in AcComplIt the percentage of not acceptable sentences is higher than in ItaCoLA, i.e. 24% vs. 16% respectively.

5 Analysis of Specific Linguistic Phenomena

Acceptability judgments involve a number of different linguistic phenomena, which we tried to cover as much as possible by selecting different sources for the creation of the dataset. However, in order to fully understand how well classifiers can judge acceptability in the presence of these phenomena, we perform also a fine-grained evaluation focused on specific linguistic constructions.

5.1 Data Annotation

We annotate a subset of the corpus with nine linguistic phenomena. The sentences to be annotated have been selected by manually going through the dataset and extracting examples showing at least one of the phenomena of interest, until around 20% of the overall dataset was annotated. In total 2,088 sentences were annotated, with 2,729 phenomena (1.3 average phenomenon per sentence).

The annotated phenomena can be divided in two macro-groups. The first one contains roughly the same classes proposed for the AcComplit dataset (Brunato et al., 2020), which we use for our out-of-domain evaluation. These classes are reported below as items 1 – 4. The second set of phenomena (items 5 – 9) includes some of the traits annotated in Warstadt and Bowman (2019) for the English language, although it is not always possible to guar-

[6] https://pytorch.org/

Model	In-domain		Out-of-domain	
	Acc.	MCC	Acc.	MCC
LSTM	0.794	0.278 $\pm$ 0.029 (best: 0.334)	0.605	0.147 $\pm$ 0.066 (best: 0.213)
ITA-BERT	0.904	0.603 $\pm$ 0.022 (best: 0.627)	0.683	0.198 $\pm$ 0.036 (best: 0.255)

Table 4: Classification results on the ItaCoLA test set and the out-of-domain AcComplIt test set. Results are the mean of 10 runs $\pm$ StdDev. Best result between parenthesis.

antee perfect equivalence between the syntax of the two languages. We detail them as follows:

1) **Cleft constructions** (136 sentences): Sentences where a constituent has been moved to put it in focus, e.g. "È il toro che Aurora ha preso per le corna e non il bufalo" (En. *It is the bull that Aurora has taken by the horns and not the buffalo.*)

2) **Copular constructions** (855 sentences): Sentences with a copulative verb that joins the subject of the sentence to a noun or an adjective, e.g. "Francesco è un grande oratore" (En. *Francesco is a great speaker.*)

3) **Subject-verb agreement** (406 sentences): Sentences characterized by the presence or lack of subject and verb agreement in gender or number, e.g. "Lorenzo ha detto che Andrea ha parlato con Riccardo" (En. *Lorenzo said that Andrea talked to Riccardo.*)

4) **Wh-islands violations** (53 sentences): Sentences introduced by a Wh- clause presenting correct or wrong syntactic constructions, e.g. "Che libro dice che il professore ha raccomandato di leggere?" (En. *What book does the professor say he recommended you to read?*)

5) **Simple** (365 sentences): Sentences in which only one verb and the mandatory arguments are present, e.g. "Tommaso legge il giornale" (En. *Tommaso reads the newspaper.*)

6) **Question** (177 sentences): Interrogative sentences, e.g. "Chi mi ha colpito?" (En. *Who hit me?*)

7) **Auxiliary** (398 sentences): Sentences containing one of the two auxiliary verbs in Italian, i.e. "essere" (*to be*) or "avere" (*to have*), e.g. "Sono arrivati molti ragazzi" (En. *A lot of guys came in.*)

8) **Bind** (27 sentences): Sentences that contain free pronouns, generally used in Italian to create contrast or focus when used together with the intensifier "stesso" (*itself*), e.g. "Lorenzo allieta se stesso" (En. *Lorenzo cheers himself.*)

9) **Indefinite pronouns** (312 sentences): Sentences containing pronouns that indicate someone or something in a generic and indefinite way, e.g. "Cerco qualcuno con cui parlare" (En. *I'm looking for someone to talk to.*)

5.2 Evaluation

To obtain a better insight into classifier performance on different linguistic phenomena, we evaluate the Italian BERT model also in this setting. To this purpose, we modify the train/test/validation splits: all 2,088 sentences annotated with fine-grained phenomena are used as test set, while the remaining part of the dataset (7,632 sentences) is used for training (6,833 sentences) and validation (800 sentences). We fine-tune *Bert-base-italian-xxl-cased* with the same parameters reported for the previous experiments. Also in this case we perform 10 restarts.

Results are reported in Fig. 1 (left). Overall, we observe a high variability across different phenomena. Some constructions seem to be easier to handle such as Clefts and Subject-Verb Agreement. Surprisingly, Simple sentences do not achieve the highest results despite their linear syntax, which reflects the dominant SVO word order in Italian (Liu, 2010). On English, instead, Warstadt and Bowman (2019) report for this category the best classification results in CoLA. Another evident difference between the two languages is that Copula constructions and Wh-violations are classified poorly in Italian, while Warstadt and Bowman (2019) report for both MCC > 0.50.

Results on Italian are probably influenced by the presence of multiple phenomena in the same sentence. Indeed, 29% of the sentences bears multiple annotations. As regards Simple sentences, we hypothesize that they tend to be wrongly classified

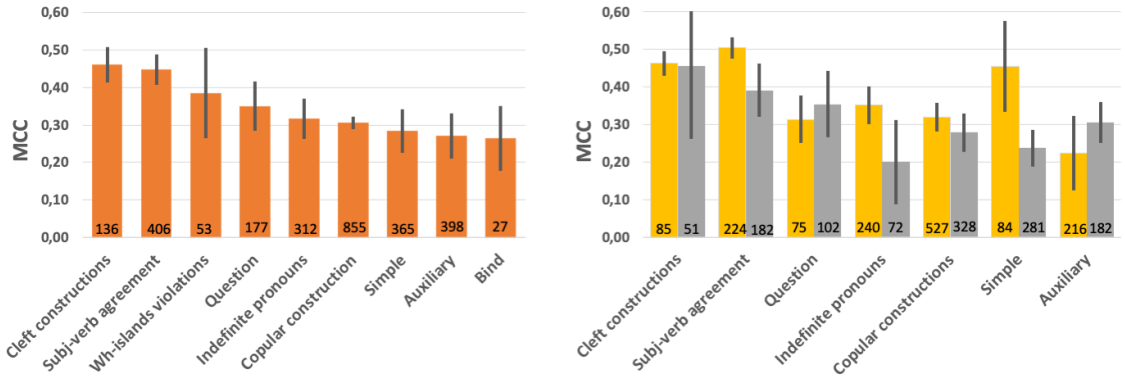

Figure 1: Classification results on a subset of ItaCoLA sentences annotated with different linguistic phenomena. Overall performance (left) and classifier performance distinguishing between sentences showing only one phenomenon (yellow) and multiple ones (grey). The number at the bottom of the bar corresponds to the number of test sentences for each phenomenon. The Bind class has been removed from the right chart because it includes only 27 sentences.

because of the presence of other linguistic phenomena among the ones considered: only 23% of the Simple sentences in our sample have not been annotated with another label.

By re-running classification only on this subset, we observe indeed that performance increases up to 0.455 MCC. The fact that classification of sentences containing only one phenomenon yields better results holds for all categories except for Questions and Auxiliary. We report in Figure 1 (right) a detailed analysis of classification performance distinguishing between sentences with only one label (yellow bars) and with multiple annotated phenomena (grey bars). Interestingly, Wh-islands violation does not appear in the chart on the right because this phenomenon is always accompanied with at least another annotation. MCC on sentences with single labels is on average 0.363 ± 0.021, while it drops to 0.308 ± 0.041 for sentences with multiple annotated phenomena.

6 Cross-lingual Experiments

Given that ItaCoLA has been created following the same principles of English CoLA and that monolingual results on Italian are in line with results obtained on the English dataset using a similar BERT-based approach, we perform a first set of cross-lingual classification experiments, to serve as baseline results for future improvements. We rely on XLM-RoBERTa-base (Conneau et al., 2020), a large multi-lingual language model, trained on 2.5TB of filtered CommonCrawl data. We experiment with different classification settings, which are all evaluated both on ItaCoLA and on CoLA in-domain test sets. This means that, starting from the same multilingual model, we classify English and Italian sentences. We implement the model in Pytorch, using a batch size of 32 and a max sequence length of 64. The learning rate is set to 2e-5, and training goes for 12 epochs. Three restarts are performed for each experiment. The number of restarts was constrained by the fact that evaluation of the English test set was possible only through Kaggle, which limits the number of runs that can be submitted for evaluation. Results are reported in Table 5. We compare three models: one obtained by fine-tuning XLM-RoBERTa with English and Italian training set together, one using only the English training, and one using only Italian sentences. Each model is tested on both languages separately. Results show that in this setting cross-lingual zero-shot learning still performs poorly (MCC = 0.114 both for English and Italian). When training using both languages, results outperform training and testing on the same language, showing the potential of this approach. However, results obtained using XLM-RoBERTa are largely outperformed by the monolingual BERT model (Table 4), confirming the findings already reported in studies on other NLP tasks (Nozza et al., 2020).

7 Conclusions

In this paper we present the Italian Corpus of Linguistic Acceptability, a novel dataset including al-

Training and validation	Test: ItaCoLA		Test: CoLA	
	Acc.	MCC	Acc.*	MCC
ItaCoLA and CoLA	0.88	0.517 ± 0.044 (best: 0.553)	0.82	0.508 ± 0.029 (best:0.535)
only CoLA	0.82	0.114 ± 0.027 (best:0.142)	0.81	0.453 ± 0.04 (best:0.494)
only ItaCoLA	0.86	0.440 ± 0.054 (best: 0.497)	0.76	0.114 ± 0.136 (best:0.211)

Table 5: Monlingual and cross-lingual classification results using XLM-RoBERTa. MCC is the average of three restarts ± StdDev. *For CoLA accuracy is calculated on development set, while MCC on test set via Kaggle because the test set is not available.

most 10k sentences taken from different linguistic resources with a binary annotation of acceptability. The corpus is released in three splits (training, development and test set) so to make replicability and further experiments easier. Part of the dataset has also been manually annotated with 9 linguistic phenomena, enabling a fine-grained evaluation of the classifier performance on specific dimensions. The process to create the corpus has followed as much as possible the one adopted to collect the English CoLA, which has become the *de facto* standard dataset for linguistic acceptability and has greatly fostered the development of automated systems for acceptability judgments. ItaCoLA can represent a first step towards the creation of multilingual benchmarks for acceptability, in line with recent efforts to create massive multilingual resources covering different tasks (Hu et al., 2020).

In the future, we plan to further explore the differences between ItaCoLA and AcComplit (Brunato et al., 2020), the other existing dataset for acceptability in Italian. We will also experiment with the Swedish corpus for acceptability studies presented in Volodina et al. (2021), to check whether the findings in our work, in particular the cross-lingual experiments, hold also for Swedish when paired with English and Italian. Furthermore, we plan to explore classification approaches that yield state-of-the-art results on CoLA. While some of them are not applicable to the new corpus, because of the lack of many massive LMs for Italian, recent studies showed that with smaller language models it should be possible to achieve better results after reformulating NLP tasks as entailment (Wang et al., 2021). We will explore whether this research direction is promising also for acceptability studies for languages with limited resources.

References

Dominique Brunato, Cristiano Chesi, Felice Dell'Orletta, Simonetta Montemagni, Giulia Venturi, and Roberto Zamparelli. 2020. Accompl-it @ EVALITA2020: Overview of the Acceptability & Complexity Evaluation Task for Italian. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Zhong Chen, Yuhang Xu, and Zhiguo Xie. 2020. Assessing introspective linguistic judgments quantitatively: The case of the syntax of chinese. *Journal of East Asian Linguistics*, 29(3):311–336.

Cristiano Chesi and Paolo Canal. 2019. Person Features and Lexical Restrictions in Italian Clefts. *Frontiers in Psychology*, 10:2105.

François Chollet. 2017. *Deep Learning with Python*. Manning.

Noam Chomsky. 1965. Aspects of the Theory of Syntax.

Shammur Absar Chowdhury and Roberto Zamparelli. 2018. RNN simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 133–144, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

BNC Consortium et al. 2007. British national corpus. *Oxford Text Archive Core Collection*.

Peter W Culicover and Ray Jackendoff. 2010. Quantitative methods alone are not enough: Response to gibson and fedorenko. *Trends in Cognitive Sciences*, 6(14):234–235.

Ewa Dabrowska. 2010. Naive v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review*, 27(1).

Emilio D'Agostino. 1983. *Lessico e sintassi delle costruzioni locative: materiali per la didattica dell'italiano*. Liguori.

Emilio D'Agostino. 1992. *Analisi del discorso: metodi descrittivi dell'italiano d'uso*. Loffredo.

Rodolfo Delmonte. 2020. Venses @ AcCompl-It: Computing Complexity vs Acceptability with a Constituent Trigram Model and Semantics. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Annibale Elia. 1982. Avverbi ed espressioni idiomatiche di carattere locativo. *Studi di Grammatica Italiana Firenze*, 11:327–379.

Annibale Elia, Maurizio Martinelli, and Emilio d'Agostino. 1981. *Lessico e strutture sintattiche: introduzione alla sintassi del verbo italiano*. Liguori Napoli.

Ingo Feldhausen and Sebastian Buchczyk. 2020. Testing the reliability of acceptability judgments for subjunctive obviation in French. In *Going romance 2020.*

Edward Gibson and Evelina Fedorenko. 2013. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, 28(1-2):88–124.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *CoRR*, abs/1901.05287.

Giorgio Graffi. 1994. *Le strutture del linguaggio. Sintassi*. Il Mulino, Bologna, Italy.

Giorgio Graffi. 2008. *Che cos'è la grammatica generativa*. Carocci editore, Roma, Italy.

Giorgio Graffi and Sergio Scalise. 2002. *Le lingue e il linguaggio. Introduzione alla linguistica*. Il Mulino, Bologna, Italy.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Matteo Greco, Paolo Lorusso, Cristiano Chesi, and Andrea Moro. 2020. Asymmetries in nominal copular sentences: Psycholinguistic evidence in favor of the raising analysis. *Lingua*, 245:102926.

Kristina Gulordava, Piotr Bojanowski, Édouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. ACL.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. ACL.

Elisabetta Jezek. 2003. *Classi di verbi tra semantica e sintassi*. Edizioni ETS, Pisa, Italy.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jey Han Lau, Carlos Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. 2020. How Furiously Can Colorless Green Ideas Sleep? Sentence Acceptability in Context. *Transactions of the Association for Computational Linguistics*, 8:296–310.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2014. Measuring gradience in speakers' grammaticality judgements. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2015. Unsupervised prediction of acceptability judgements. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1618–1628, Beijing, China. Association for Computational Linguistics.

Steve Lawrence, C Lee Giles, and Sandiway Fong. 2000. Natural language grammatical inference with recurrent neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 12(1):126–140.

Tal Linzen. 2019. What can linguistics and deep learning contribute to each other? response to pater. *Language*, 95(1).

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Tal Linzen and Yohei Oseki. 2018. The reliability of acceptability judgments across languages. *Glossa: a journal of general linguistics*, 3(1).

Haitao Liu. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 120(6):1567–1578. Contrast as an information-structural notion in grammar.

Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli. 2014. The paisa'corpus of italian web texts. In *9th Web as Corpus Workshop (WaC-9)@ EACL 2014*, pages 36–43. EACL (European chapter of the Association for Computational Linguistics).

Simona Mancini, Paolo Canal, and Cristiano Chesi. 2018. The acceptability of person and number agreement/disagreement in Italian: An experimental study.

Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Rebecca Marvin and Tal Linzen. 2019. Targeted syntactic evaluation of language models. *Proceedings of the Society for Computation in Linguistics (SCiL)*, pages 373–374.

B.W. Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451.

R. Thomas McCoy, Robert Frank, and Tal Linzen. 2020. Does syntax need to grow on trees? sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, 8:125–140.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [mask]? making sense of language-specific BERT models. *CoRR*, abs/2003.02912.

Alessio Palmero Aprosio and G Moretti. 2018. Tint 2.0: An all-inclusive suite for nlp in italian. In *Fifth Italian Conference on Computational Linguistics CLiC-it 2018*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Gabriele Sarti. 2020. Umberto-mtsa @ accompl-it: Improving complexity and acceptability prediction with multi-task learning on self-supervised annotations (short paper). In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Carson T. Schütze. 2016. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. University of Chicago Press.

Raffaele Simone and Francesca Masini. 2013. *Nuovi fondamenti di linguistica*. McGraw Hill.

Jon Sprouse. 2011. A validation of amazon mechanical turk for the collection of acceptability judgments in linguistic theory. *Behavior research methods*, 43(1):155–167.

Jon Sprouse and Diogo Almeida. 2010. A quantitative defense of linguistic methodology. *Manuscript submitted for publication*.

Jon Sprouse and Diogo Almeida. 2013. The empirical status of data in syntax: A reply to gibson and fedorenko. *Language and Cognitive Processes*, 28(3):222–228.

Jon Sprouse, Carson Schütze, and Diogo Almeida. 2013a. Assessing the reliability of journal data in syntax: linguistic inquiry 2001–2010. *Lingua*, 134:219–248.

Jon Sprouse, Carson T Schütze, and Diogo Almeida. 2013b. A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. *Lingua*, 134:219–248.

Simonetta Vietri. 2004. *Lessico-grammatica dell'italiano. Metodi, descrizioni e applicazioni*. UTET Università.

Simonetta Vietri. 2014. *Idiomatic constructions in Italian: a lexicon-grammar approach*, volume 31. John Benjamins Publishing Company.

Simonetta Vietri. 2017. *Usi verbali dell'italiano: le frasi anticausative*. Carocci editore.

Sandra Villata, Paolo Canal, Julie Franck, Andrea Moro, and Cristiano Chesi. 2015. Intervention effects in wh-islands: An eye-tracking study. In *Architectures and Mechanisms for Language Processing (AMLaP 2015)*.

Elena Volodina, Yousuf Ali Mohammed, and Julia Klezl. 2021. DaLAJ – a dataset for linguistic acceptability judgments for Swedish. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 28–37, Online. LiU Electronic Press.

Joachim Wagner, Jennifer Foster, Josef van Genabith, et al. 2009. Judging grammaticality: Experiments in sentence classification. *Calico Journal*, 26(3):474–490.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner.

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Jiangnan Xia, Zuyi Bao, Liwei Peng, and Luo Si. 2020. Structbert: Incorporating language structures into pretraining for deep language understanding. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Alex Warstadt and Samuel R. Bowman. 2019. Grammatical analysis of pretrained sentence encoders with acceptability judgments. *CoRR*, abs/1901.03438.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.