EsCoLA: Spanish Corpus of Linguistic Acceptability

Núria Bel[1], **Marta Punsola**[1], **Valle Ruiz-Fernández**[2]
[1]Universitat Pompeu Fabra, [2]Barcelona Supercomputing Center
Barcelona, Spain
{nuria.bel,marta.punsola}@upf.edu, valle.ruizfernandez@bsc.es

Abstract

Acceptability is one of the General Language Understanding Evaluation Benchmark (GLUE) probing tasks proposed to assess the linguistic capabilities acquired by a deep-learning transformer-based language model (LM). In this paper, we introduce the Spanish Corpus of Linguistic Acceptability EsCoLA. EsCoLA has been developed following the example of other linguistic acceptability data sets for English, Italian, Norwegian or Russian, with the aim of having a complete GLUE benchmark for Spanish. EsCoLA consists of 11,174 sentences and their acceptability judgements as found in well-known Spanish reference grammars. Additionally, all sentences have been annotated with the class of linguistic phenomenon the sentence is an example of, also following previous practices. We also provide as task baselines the results of fine-tuning four different language models with this data set and the results of a human annotation experiment. Results are also analyzed and commented to guide future research. EsCoLA is released under a CC-BY 4.0 licence and freely available at https://doi.org/10.34810/data1138.

Keywords: Language Model, Evaluation, Linguistic Acceptability, Corpus, Spanish

1. Introduction

Acceptability judgement is a linguistic task first proposed by the Generative Grammar linguistic theory (Chomsky, 1965). This theory was concerned with discovering the mechanism that could generate all but only the sentences accepted by speakers of a language as possible sentences of their language. Currently, acceptability is one of the standard probing tasks proposed to assess the linguistic capabilities acquired by a deep-learning transformer-based language model (LM). The task consists of fine-tuning a LM to recognize acceptable sentences under the assumption that, only if the representations built by the LM are somehow different for acceptable and for unacceptable sentences, it is possible for a classifier to distinguish them. The Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2018) was the first data set developed to support the task of linguistic acceptability in English, which is part of the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018).

GLUE benchmark was introduced as a tool to evaluate and analyze the performance of language models across a diverse range of existing Natural Language Understanding (NLU) tasks. The initial GLUE consists of nine English understanding tasks selected to cover a broad range of type of tasks, domains, amount of data and difficulties. This set of tasks is intended to challenge a model from different aspects. Although initially all data sets were just in English, parallel data sets are being developed for other languages. In this paper, we present the Spanish Corpus of Linguistic Acceptability EsCoLA, which adds to the effort of having a complete GLUE benchmark for Spanish. To build EsCoLA, we have compiled 11,174 sentences and their acceptability judgements as found in well known Spanish reference grammars. Additionally, all sentences have been annotated with the class of linguistic phenomenon the sentence is an example of, according to a list of fourteen categories. The first thirteen categories are the same as those used by Warstadt and Bowman (2019). As for the fourteenth, it gathers sentences containing specific Spanish phenomena: agreement in nominal constructions, subjunctive mode and tense, spurious preposition for completive clauses ('dequeismo'), subject ellipsis, pronominal cliticization, and 'ser'/'estar' copula selection. EsCoLA aims to be used in conjunction with GLUES[1], the General Language Understanding Evaluation benchmark for Spanish. Currently, GLUES consists of eight tasks and eleven data sets, and, to the best of our knowledge, there is no yet a data set for the linguistic acceptability task.

We also provide as task baselines the results of fine-tuning four different language models with this data set and the results of a human annotation experiment. Results are also analyzed and commented to guide future research.

In this paper, related work is summarized in section 2 and we describe the new resource in section 3. In section, 4 we report how we have used EsCoLA to fine-tune different existing language models for the linguistic acceptability task. Results are presented in section 4.5. Finally, section 5 is devoted to sum up the contributions of the new corpus.

[1]https://github.com/dccuchile/GLUES

LREC-COLING 2024, pages 6268–6277

2. Related Work

The English CoLA data set included in GLUE (Wang et al., 2018) consists of 10k sentences with expert annotations for grammatical acceptability. The objective was to help assessing the linguistic information contained in representations delivered by LMs and used for building classifiers as fine-tuning. To create the CoLA data set, Warstadt et al. (2018) compiled English sentences from 23 theoretical linguistics publications representing a wide array of linguistic phenomena. The corpus was partitioned into training, development and test, in which acceptable sentences are around 70% of the data set. Additionally, the sentences of the CoLA development set, a 10% of the corpus, were annotated for the presence of linguistic phenomena. In Warstadt and Bowman (2019), the original CoLA data set was enlarged for a detailed annotation of 1,043 sentences that were labeled as samples of thirteen major features and 59 minor features.

After CoLA, similar resources have been developed for Swedish, Italian, Norwegian and Russian. DaLAJ (Volodina et al., 2021) is an acceptability data set for Swedish which is made of 9,596 instances: 4,798 pairs of incorrect sentences from the SweLL second language learner corpus (Volodina et al., 2019) and their corresponding correct sentences. The SweLL corpus consists of essays at different levels of proficiency. DaLaJ unacceptability judgments were produced by teachers, assessors, or trained assistants, and sentences were also annotated with information about the error.

ItaCoLA (Trotta et al., 2021), for Italian, follows the original CoLA design. Sources include theoretical linguistics textbooks and works that focus on specific phenomena such as idiomatic expressions, locative constructions and verb classification. It consists of 10k sentences annotated with acceptability binary judgements as originally found in the selected linguistic publications. The percentage of acceptable sentence amounts to 85.4%. A subset of 2,088 sentences is annotated for detailed linguistic phenomena. The annotation includes some of the thirteen categories used by (Warstadt and Bowman, 2019) for English, although there are some differences in the phenomena reported for each of them.

RuCoLA is the Russian Corpus of Linguistic Acceptability (Mikhailov et al., 2022). It was also developed for assessing the linguistic competence of language models within the CoLA paradigm. It consists of 13,4k sentences labeled as acceptable (71.8%) or not (28.2%). RuCoLA combines in-domain sentences manually collected from linguistic literature and out-of-domain sentences produced by different machine translation and paraphrase generation models. Each unacceptable sentence is labeled with four different categories: morphology, syntax, semantics, and hallucinations. Differently to previous corpora, the purpose of the RuCoLA data set is extended towards the evaluation of text generation system with metrics based on acceptability, and their results are not directly comparable to the results of the previous works.

NoCoLA, the Norwegian Corpus of Linguistic Acceptability (Jentoft and Samuel, 2023), consists of two data sets. The source for both NoCoLA data sets was the ASK Corpus, a language learner corpus of Norwegian as a second language (Tenfjord et al., 2006). The first dataset, NoCoLAclass, only encodes acceptability and contains 144,867 sentences, 31.5% of which are grammatically acceptable. The second data set, NoCoLAzero, is a collection of pairs of sentences, of which only one is grammatically acceptable, and follows the data set schema of the Benchmark of Linguistic Minimal Pairs for English, BLiMP (Warstadt et al., 2020). BLiMP is an extension of the first CoLA corpus and contains 67k pairs of ungrammatical and their corresponding grammatical sentences automatically generated via manually-constructed templates that span 12 high-level English phenomena.

SLING, Sino Linguistic Evaluation of Large Language Models (Song et al., 2022), is a corpus of 38k minimal sentence pairs in Mandarin Chinese grouped into 9 high-level linguistic phenomena, many of which are unique to the Chinese language. SLING exploited the Chinese Treebank 9.0 (Nianwen Xue et al., 2016) extracting subtrees from human-validated constituency parses and transforming them with manually designed linguistic templates to create minimal pairs of acceptable-unacceptable sentences, that were, nevertheless, validated by human annotators.

Finally, another similar data set is the one by Hartmann et al. (2021) for Bulgarian and German, which is made of minimal pairs and used to fine-tune a model for particular linguistic probing tasks although different to acceptability.

As for the acceptability task, model performance has been traditionally measured in terms of the Matthews Coefficient Correlation (MCC) (Matthews, 1975) and accuracy. The best performance with the English CoLA corpus was reported in Warstadt and Bowman (2019) after comparing transformer-based language models: GPT and BERT. The best result, with MCC=0.58, was achieved by a BERT-large fine-tuned classifier. Because CoLA is in the GLUE benchmark, posterior better results, around MCC=0.75, have been published in the leaderboard[2] achieved with dif-

[2] https://gluebenchmark.com/leaderboard

ferent architectures. The performance of ItaCoLA with an Ita-BERT is reported to be MCC=0.67 for the in-domain data set (Trotta et al., 2021). The proposed RuCoLA baselines are obtained with six different language models, four monolingual and two multilingual, being ruRoBERTa the best, achieving MCC=0.53 with the in-domain data set.

3. EsCoLA: Spanish Corpus of Linguistic Acceptability

In this section, we describe the Spanish Corpus of Linguistic Acceptability (EsCoLA). The corpus was built following the methodology proposed by the English Corpus of Linguistic Acceptability (Warstadt and Bowman, 2019) as a resource to assess large language models' capabilities of capturing linguistic information. Table 1 shows Spanish EsCoLA compared to the other corpora of linguistic acceptability for different languages we have described above.

data set	lang.	size k	% accep.
CoLA	English	10.6	70.5
DaLAJ	Swedish	9.5	50
ItaCoLA	Italian	9.7	85.4
RusCoLA	Russian	13.4	71.8
NoCoLA	Norwegian	14.4	31.5
EsCoLA	Spanish	11.1	70

Table 1: Comparison of EsCoLA with related binary acceptability corpora for other languages. The language of the data set, the size in thousands, and the percentage of acceptable (acc.) sentences are indicated.

3.1. Partitions

EsCoLA data set is split into two subsets: an in-domain subset, (InDomain) with 10,567 sentences, and an out-of-domain subset (OutDomain) with 607 sentences. The InDomain and OutDomain sentences were collected from different sources to include sentences from different domain specificity, time and purpose to discover overfitting.

3.2. Source

The 10,567 sentences that are in EsCoLA InDomain corpus were extracted from a well-known Spanish reference grammar, *Gramática descriptiva de la lengua española* (GDE) (Demonte and Bosque, 1999). GDE is a compilation of 78 articles from different authors covering the description of a broad list of linguistic phenomena in Spanish that takes into account Spanish regional variants. The 607 sentences in the EsCoLA OutDomain corpus are from three other grammatical description books of Spanish written by prestigious authors and addressed to native but specially to foreign speakers: RAE (2009), Palencia and Aragonés (2007) and Díaz and Yagüe (2019).

3.3. Linguistic Phenomena Annotation

From these sources, we extracted the examples of acceptable and unacceptable sentences as well as the phenomenon each sentence was an example of in the reference grammars. Thus, sentences were first annotated according to the topic of the chapter they were found at, and two experts in syntax revised and discussed the mapping to the thirteen major CoLA categories (Warstadt and Bowman, 2019). We now describe the categories and provide examples of acceptable and unacceptable sentences for each.

1. Simple. Sentences with a verb and a complete mandatory set of subcategorized complements. *El banco perdonará la deuda. Juan cerró las puertas. *Dudo su participación. *Guillermo hace.*
2. Predicative. Copular, small clauses and resultatives. *Balmes es una calle. Juan parece triste. *El diccionario es médico. *Mis amigos estaban gustados.*
3. Adjuncts. Optional modifiers for NPs and VPs and temporal and locative adjuncts. *El alumno estudia con ahínco. En su ensoñación, se imaginaba con mucho dinero. *Amó a Salomé en tres años. *María escucha la radio comiendo su marido.*
4. Argument types. Oblique, prepositional arguments subcategorized by the verb, nouns or adjectives, and expletives. *Esteban sacó partido de la situación. Las ventas se verán afectadas por la crisis. *Mario ha reservado pan a la cena. *Leí un libro para los niños.*
5. Argument alternations, high-arity, passives, including reflexive passives, drop-args and add-args. *Las puertas han sido cerradas. María se depila las pestañas. *Leer la carta es podido por Juan. *Un perro fue muy corrido.*
6. Binding pronouns. *El sol se destruyó a sí mismo. Juan apareció él solo. *Yo he tomado el pulso a mí. *Juan no bebe cuando él trabaja.*
7. Wh-phenomena. Questions and relatives (exclamatives have been excluded). *Me pregunto quién vendrá a estas horas. Todo lingüista que oye un error lingüístico se indigna. *¿Qué grande es tu coche? *Los alumnos que les dimos el premio llegarán más tarde.*

8. Complement clauses, including subjects, arguments of VPs, NPs or APs. *Supongo que es capaz de hacerlo. Eva me comunicó que pensaba dejar a su marido. *Vi el que el coche seguía parado en la acera. *Creo que haber venido.*

9. Auxiliary and modal verbs, negation, polarity and periphrastic verbal constructions. *Juan debe leer mi libro. Luis todavía no ha terminado la tesis. *Hay persona más desgraciada que tú. *Estás debiendo perder mucho dinero.*

10. Infinitival embedded VPs involving referential obligatory phenomena like control, raising, and VP, NP or AP argumental constructions. *María desea plantar rosas en el jardín. Tu amigo es difícil de convencer. *Caminamos hasta el llegar a una ermita. *Es un difícil de solucionar problema.*

11. Complex NPs and APs, including PP arguments. *Resultaba contrario a la libertad de los ciudadanos. El acuerdo es susceptible de revisión. *Se compró un vestido rojo deslumbrante largo. *Ocurrió la exportación azucarera cubana de caña.*

12. S-syntax phenomena. Coordination, subordination and sentence-level adjuncts. *Amelia entró y cerró la puerta. Tanto si vienes como si no, yo iré al cine. *Si te castigará, no vamos. *Es tan alto para que toque el techo con la mano.*

13. Determiners, quantifiers, partitives, and comparative constructions. *Llegaron dos docenas de mujeres. Hemos visto a varios de los alcaldes. *Escribió muchos de artículos. *Comió un cierto helado de menta.*

The linguistic annotation of the EsCoLA sentences is meant to facilitate the detailed analysis of acceptability classifiers both regarding training examples and error analysis. Additionally, for analysis purposes we have created a further fourteenth category that gathers together linguistic phenomena that are characteristic of Spanish. Spanish phenomena included are the following:

- Agreement in nominal constructions. *Aun hervida, la lubina es deliciosa. Me compré unas camisas y una corbata nuevas. *Esta perla de vigilante nocturno que hemos contratado siempre está dormida. *Juan es un traidora.*

- Subjunctive mode and tense. *Seguro que se alegraron cuando regresaron de vacaciones. Las encuestas vaticinan que el número de diabéticos será cada vez mayor. *Que le haces croquetas borrará sus penas. *Vio que era mejor que vendría.*

- Spanish 'de+que'. *Nos advirtieron de que había un poste caído en la carretera. Me informaron de que había mucha gente. *Es fácil de que llueva. *Pienso de que es difícil salir de esa situación.*

- Ellipsis. *Preferimos té. En vez de llevarte mi coche, llévate el de Teresa. *En cuanto Alicia acabe, nos llamará, no su secretaria. *Ana se enoja cada vez que sólo pierde.*

- Cliticization phenomena. *Al ladrón, nos dijeron que la policía lo atrapó. A la casa se le cayó una teja. *Otto se le refería a Lucy. *Al culpable se lo buscó por varios países.*

- 'Ser'/'estar' copula selection. *El diccionario es verde. Está celoso de su mujer. *Es orgulloso de su hijo. *Maggie está la mamá de Gabriel.*

Figure 1 illustrates the distribution of the InDomain and OutDomain subsets per linguistic phenomenon.

3.4. Sentence Selection and Curation

Sentences in the EsCoLA corpus come from a first selection of the chapters of the GDE that described phenomena related to the thirteen categories proposed in Warstadt and Bowman (2019). Then, the source texts were digitalized with an OCR software. We used manual regex patterns to identify the examples (indentation and numbering) in the text which could be potential corpus sentences. Some curation was required to correct the typical OCR errors, or, eventually, to discard some of the extracted sentences. We also identified those source examples that included other notation than the traditional '*' for marking unacceptable sentences. We discarded examples marking dubious acceptability with '?' or other signs, but those examples that included acceptability alternations were taken by creating the two versions: the acceptable and the unacceptable sentence. Finally, in order to reach a 30% of unacceptable cases, the examples that were not full sentences, that is, that contain no main verb, were manually edited to add a neutral verb to convert them into sentences, while keeping the acceptability value. For instance, the example **uvas maduras bastante* ('grapes ripe enough') resulted in **Hay uvas maduras bastante* ('There are grapes ripe enough'). Finally, like in other linguistic acceptability corpora, we manually substituted very low frequency words appearing in the examples

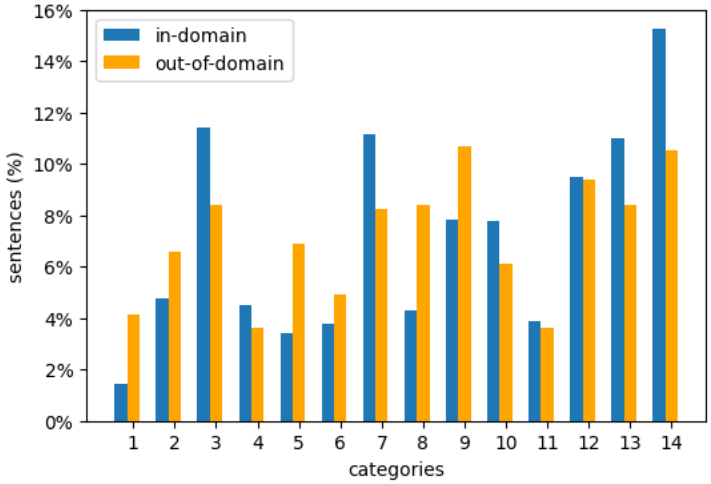

Figure 1: Percentage of sentences per linguistic category in the Indomain and OutDomain EsCoLA data sets. 1: Simple, 2: Predicative, 3: Adjuncts, 4: Argument Types, 5: Argument Alternation, 6: Binding Pronouns, 7: Wh-phenomena, 8: Complement Clauses, 9: Modals, Negation, Periphrasis and Auxiliaries, 10: Infinitive Embedded VPs, 11: Complex NPs and APs, 12: S-syntax, 13: Determiners, Quantifiers, Comparative and Superlative constructions, 14: Spanish Phenomena.

(i.e., wordforms with a frequency below 45 samples in a reference corpus[3]).

4. Experiments

EsCoLA corpus is meant to support an acceptability probing task in Spanish, that is, to classify sentences according to their acceptability. To provide a baseline of this corpus that can be used in future experiments for fair comparison, we performed fine-tuning experiments with different state-of-the-art monolingual and multilingual language models. The experiments were carried out for InDomain and OutDomain data sets as described in section 3.1. To select the language models, we have followed the work by Gutiérrez-Fandiño et al. (2022) and Agerri and Agirre (2023) on evaluating and comparing Spanish language models. We eventually selected for experimentation the models that resulted the best ones in some of the eleven tasks[4] reported in Agerri and Agirre (2023): IXABERTesv2[5], RoBERTa-large-bne[6], XLM-RoBERTa-large[7], and mDeBERTa-v3[8].

IXABERTesv2 and RoBERTa-large-bne are RoBERTa-based models pre-trained with a Masked Language Modelling (MLM) task. XLM-RoBERTa-large and mDeBERTa-v3 are multilingual models. While XLM-RoBERTa-large is a RoBERTa-based model trained with a MLM task, mDeBERTa-v3 incorporates other features like disentangled attention, gradient-disentangled embedding sharing and, instead of being trained with a MLM task, it is trained with a Replaced Token Detection (RTD) task (Clark et al., 2020). More details about the main characteristics of the models are in Table 2.

Model	W	L	H	A	V	P
IXAes	25	12	768	12	50	125
RB-L-bne	135	24	1024	16	50	350
XLM-L	167	24	1024	16	250	550
mDBv3	167	12	768	12	250	198

Table 2: Spanish Language Models as described in Agerri and Agirre (2023). W: training corpus number of words in billions, L: layer size, H: hidden size, A: attention heads, V: vocabulary in thousands, P: number of parameters in millions (Note that we corrected the size of the CC-100 corpus).

Both for the InDomain and OutDomain experiments, the training data is limited to sentences and acceptability labels; all other annotations (i.e. linguistic categories) are not provided to the model so as to mimic human learning. All models are fine-tuned for 5 epochs with a maximum sequence length of 128, a batch size of 64 and a learning

[3]Corpus de Referencia del Español Actual, RAE, https://www.rae.es/banco-de-datos/crea

[4]Tasks were POS tagging, NER, Universal Dependencies, Semantic Text Similarity, Document Classification, Paraphrase Identification, Natural Language Inference, Question Answering and Metaphor Detection.

[5]IXABERTesv2:http://www.deeptext.eus/resources/ixabertes-v2.zip

[6]https://huggingface.co/PlanTL-GOB-ES/roberta-large-bne

[7]https://huggingface.co/xlm-roberta-large

[8]https://huggingface.co/microsoft/mdeberta-v3-base

rate set at 2e-5. Considering that the data set is unbalanced, the loss is computed with weighted cross-entropy.
Following previous works, the performance in our experiments is measured with an accuracy score (acc.) and Matthews Correlation Coefficient (MCC, Matthews, 1975). While accuracy is not very informative in the case of unbalanced data sets, it is broadly acknowledged that MCC is a robust metric that summarizes the classifier performance in a single value, when positive and negative cases are of equal importance. Note that MCC=1 indicates that predictions from the classifier do correlate well with the real class, while MCC=0 means that predictions are random.

4.1. InDomain experiment

The fine-tuning experiments for the InDomain data set were run for the language models IXABERTesv2, RoBERTa-large-bne, XLM-RoBERTa-large and mDeBERTa-v3 as just described. We run a 5-fold cross-validation and the results presented in section 4.4 are averaged. For each round, the InDomain data set is split into 80% for training, 10% for development, and 10% for testing. The data partitions are created so that the original distribution of linguistic categories (see 1 and acceptability labels (70% acceptable, 30% unacceptable) is preserved.

4.2. OutDomain experiment

All the models were also fine-tuned in the out-of-domain setting. We carry out two different experiments: (1) as done in related works, the models are trained on the EsCoLA training subset (the one corresponding to the first fold of the InDomain experiment), while validation and test are performed using the OutDomain data set already mentioned in section 3.1, split into 50% development and 50% test. (2) We also evaluated the performance of the best model using the whole OutDomain data set as test set.

4.3. Human performance

To complete the data, in addition to the judgments provided by the source reference texts, three human experts in linguistics annotated the whole InDomain corpus. These data served to assess human performance as an upper bound for machine performance as in Warstadt et al. (2018). The three annotators were two postgraduate students and one postdoc in linguistics, all native speakers of Spanish. The average MCC among annotators was 0.719, and the average Cohen kappa agreement with the reference was 0.718. There are 888 cases, i.e. an 8.4% of cases, where the majority decision of the annotators contradicts the EsCoLA annotation extracted from the reference book. Note that in English CoLA authors report a 13% of labels that contradict human majority judgements. The disagreement might be caused by problems in the data curation process or by variance due to regional varieties of Spanish or to idiolects.

4.4. Results

In Table 3, we report the average accuracy and MCC scores for the InDomain data set obtained from the 5-fold cross-validation. Also, Figure 2 shows average MCC per model and category. Note that only XLM-RoBERTa showed very high variability (from 0.0 to 0.46 MCC), while for the other ones the minimal and maximal values for MCC were: IXABERTesv2, 0.15-0.39; mDeBERTa-v3, 0.48-0.54, and RoBERTa-large-bne, 0.42-0.47[9].

Model	MCC	acc.
IXABERTesv2	0.29	0.73
RoBERTa-large-bne	0.45	0.77
XLM-RoBERTa-large	0.33	0.74
mDeBERTa-v3	0.52	0.8

Table 3: Average acceptability classification scores MCC and accuracy per model trained and validated on the EsCoLA InDomain data set.

The multilingual mDeBERTa-v3 is the model obtaining the best results among the ones evaluated, followed by RoBERTa-large-bne. In contrast, of note is that the other large model evaluated, XLM-RoBERTa-large, ranks third with lower scores. These results are similar to the smallest model evaluated, IXABERTesv2, which is the one performing the worst.
As for the OutDomain experiment, results are detailed in Table 4. Similarly to the results of the InDomain experiment, for the OutDomain experiment the highest scores are obtained with the multilingual model mDeBERTa-v3, followed by RoBERTa-large-bne and, with worse results, XLM-RoBERTa-large and the smallest model IXABERTesv2.

4.5. Discussion

As described in section 4.4, the best results for linguistic acceptability tasks are obtained with the multilingual model mDeBERTa-v3. This result is in line with the improvements shown by this architecture in other tasks and experiments. DeBERTa-V3-large achieved MCC=0.75 as published in He et al. (2021), which compares with MCC=0.67 achieved with BERT for English.

[9]We also experimented with 5 restarts with different seeds and the models showed a robustness similar to the observed in the cross-validation experiments.

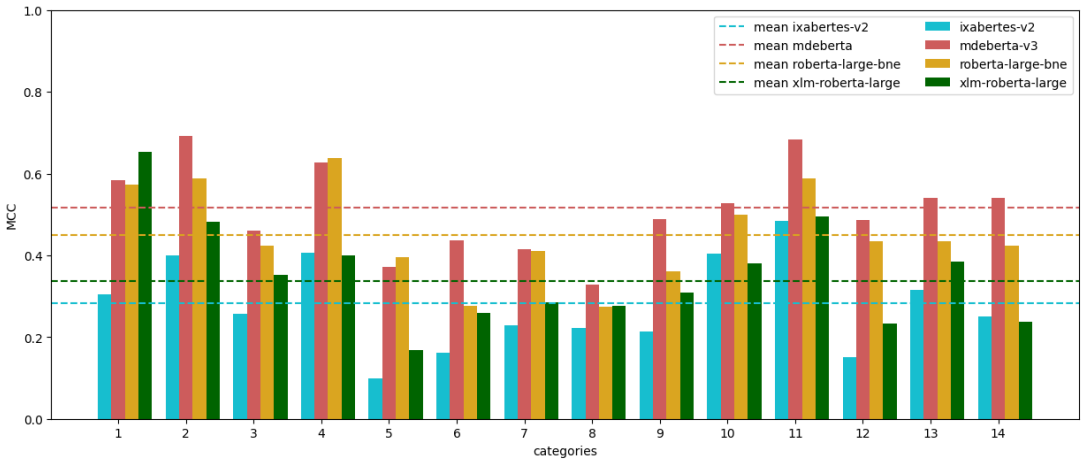

Figure 2: Average MCC per model and linguistic category. 1: Simple, 2: Predicative, 3: Adjuncts, 4: Argument Types, 5: Argument Alternation, 6: Binding Pronouns, 7: Wh-phenomena, 8: Complement Clauses, 9: Modals, Negation, Periphrasis and Auxiliaries, 10: Infinitive Embedded VPs, 11: Complex NPs and APs, 12: S-syntax, 13: Determiners, Quantifiers, Comparative and Superlative constructions, 14: Spanish Phenomena. Dashed lines show average MCC scores.

	exp. 1		exp. 2	
Model	**MCC**	**acc.**	**MCC**	**acc.**
IXABERTesv2	0.03	0.68	0.50	0.80
RoBERTa-large-bne	0.59	0.84	0.51	0.80
XLM-RoBERTa-large	0.20	0.69	0.52	0.81
mDeBERTa-v3	0.68	0.87	0.63	0.85

Table 4: Classification scores MCC and accuracy per model. (1) trained on the EsCoLA InDomain data set, and validated and tested on the OutDomain test set, and (2) trained and validated on the EsCoLA InDomain data set, and tested on the OutDomain whole data set.

The second best-performing model is the monolingual RoBERTa-large-bne, which aligns with expectations when considering, as Agerri and Agirre (2023) mention, its size and the corpora used to train it. However, RoBERTa-large-bne MCC=0.45 is below the results of other acceptability experiments with other languages. Warstadt and Bowman (2019) reported a score of MCC=0.58 for English with a BERT-large model, and Ita-BERT reached MCC=0.67 for Italian (Trotta et al., 2021), although note that the Italian dataset has a smaller number of unacceptable sentences, only 14.6%.

In contrast, the low scores obtained with XLM-RoBERTa-large might be indeed surprising if we take into account that this model resulted the best option for most tasks in Agerri and Agirre (2023), although this model got also bad results in acceptability experiments for Russian (Mikhailov et al., 2022).

Note that, except for the case of IXABERTesv2, surprisingly all our models provided better results in OutDomain experiments than in the InDomain ones. This difference could be due to different factors like the length of the sentences, larger number of very frequent words or the differences in the phenomenon types between both data sets. As for length, the sentences of InDomain data set have, on average, 8.68 tokens, while in the out-of-domain this number drops to 7.93. Warstadt and Bowman (2019) reported an specific experiment for assessing the impact of sentence length in the acceptability task results and reported that performance dropped for longer sentences in a steadily form when longer than 4 tokens, observing more than one MCC point difference for sentences of 7 to 8 tokens. As for frequency, 79% of the words in the InDomain corpus and 83% of the words in the OutDomain corpus are among the 5000 most frequent tokens in Spanish[10]. Finally, the differences in the number of sentences of particular phenomenon types between both data sets are shown in Figure 1.

As for the performance of the models on the specific linguistic phenomena in EsCoLA data set, Figure 2 shows the test MCC scores per linguistic category averaged over the 5 rounds of cross-validation performed in the InDomain setting.

Overall, it can be depicted not only that results among categories are highly variable, but also that there is no existing correlation between the performance of the models regarding a specific phenomenon and the number of sentences for these phenomena in the EsCoLA data set. In other

[10]We again used the reference: https://www.rae.es/banco-de-datos/crea

words, a greater number of training sentences of a specific category seems not to imply better predictions. A good example are simple sentences (1) and sentences with complex nouns and adjective phrases (11): even if they are less than 2% and 4%, respectively in the InDomain data set, the MCC scores are among the highest. In contrast, sentences with adjuncts (3) and wh-phenomena (7) have a greater representation in the data set and, still, obtain lower MCC scores.

Like in English CoLA and in ItaCoLA, questions as included in 7th category are the ones for which the models, including mDeBerta-v3, have more difficulties. Binding pronouns are also a problem for all our models, in line with the difficulties reported for ItaCoLA.

Regarding Spanish-specific phenomena (category 14), Figure 3 shows the averaged MCC scores per phenomenon in this category for a more specific analysis. For instance, note that in ser/estar copula selection (category 14.6), a Spanish-specific phenomenon that is usually difficult for non-native speakers, both mDeBERTav3 and RoBERTa-large-bne get a MCC > 0.5, which is higher than the mean score.

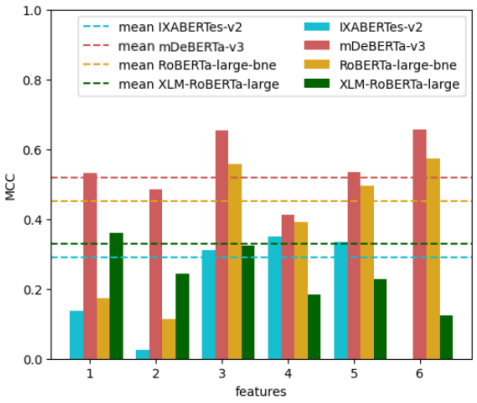

Figure 3: Average MCC per model and type of Spanish feature in category 14. Dashed lines show average MCC scores. 1: Agreement in nominal constructions, 2: Subjunctive mode, 3: Spurious preposition for completive clauses ('dequeismo'), 4: Subject ellipsis, 5: Pronominal cliticization, 6: Ser/estar copula selection.

5. Conclusions

We have described the Spanish Corpus of Linguistic Acceptability EsCoLA, which constitutes the first data set for the acceptability probing task for Spanish. EsCoLA has been developed following the example of other linguistic acceptability data sets for English, Italian, Norwegian or Russian, with the aim of completing the GLUE benchmark for Spanish and therefore promoting the fair evaluation and comparison of existing and future large language models.

The EsCoLA dataset consists of 11,174 sentences and their acceptability judgements as found in well-known Spanish reference grammars. The annotation provided with the data set includes, in addition to the reference acceptability judgements, the linguistic phenomenon the sentence is an example of, and the acceptability judgments of three experts in linguistics for the InDomain partition. EsCoLA data set also includes the task baselines obtained by fine-tuning four different language models and the InDomain and OutDomain partitions. These baselines are also analyzed and commented to guide future research. EsCoLA is released under a CC-BY 4.0 licence[11] and is freely available at https://doi.org/10.34810/data1138.

6. Acknowledgements

This research is part of the LUTEST project, PID2019-104512GB-I00, funded by the MICIU/AEI/ 10.13039/501100011033. BSC participation has been promoted and financed by the Generalitat de Catalunya through the Aina project and by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - NextGenerationEU within the framework of the project ILENIA (2022/TL22/00215337-00215334). We want to thank the collaboration of Yago Soler and Marta García.

7. Ethical considerations and limitations

The data set has been made by copying the examples from published works that are protected by copyright. According to Spanish law, we have respected the copyright because the number of elements taken represents less than 10% of the whole work, and the number of items copied is justified by the aims of the research.

8. Bibliographical References

Yvonne Adesam, Aleksandrs Berdicevskis, and Felix Morger. 2020. *SwedishGLUE–Towards a Swedish Test Set for Evaluating Natural Language Understanding Models*. Research Reports from the Department of Swedish.

Rodrigo Agerri and Eneko Agirre. 2023. Lessons learned from the evaluation of spanish language models. *Procesamiento del Lenguaje Natural*, 70(0):157–170.

[11] http://creativecommons.org/licenses/by/4.0/"

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*, 50 edition. The MIT Press.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. *CoRR*, abs/2003.10555.

Violeta Demonte and Ignacio Bosque. 1999. *Gramática Descriptiva de la lengua española*. Espasa Calpe España.

Lourdes Díaz and Agustín Yagüe. 2019. *Gramática del español como lengua extranjera*. Ediciones marcoELE, revista didáctica español lengua extranjera.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano-Oller, Carlos Rodriguez-Penagos, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, page 39–60.

Mareike Hartmann, de Miryam Lhoneux, Daniel Hershcovich, Yova Kementchedjhieva, Lukas Nielsen, Chen Qiu, and AndersSøgaard. 2021. A multilingual benchmark for probing negation-awareness with minimal pairs. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 244–257, Online. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

Matias Jentoft and David Samuel. 2023. NoCoLA: The Norwegian corpus of linguistic acceptability. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 610–617, Tórshavn, Faroe Islands. University of Tartu Library.

Brian 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.

Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov, and Ekaterina Artemova. 2022. RuCoLA: Russian corpus of linguistic acceptability. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5207–5227, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ramón Palencia and Luis Aragonés. 2007. *Gramática de uso del español C1-C2*. CESMA - S.M.

RAE. 2009. *Nueva Gramática de la lengua española. Morfología y sintaxis*. Espasa Calpe España.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. SLING: Sino linguistic evaluation of large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kari Tenfjord, Paul Meurer, and Knut Hofland. 2006. The ASK corpus - a language learner corpus of Norwegian as a second language. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Daniela Trotta, Raffaele Guarasci, Elisa Leonardelli, and Sara Tonelli. 2021. Monolingual and cross-lingual acceptability judgments with the Italian CoLA corpus. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2929–2940, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Elena Volodina, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg, and Mats Wirén. 2019. The swell language learner corpus: From design to annotation. *Northern European Journal of Language Technology*, 6:67–104.

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.

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.

Alex Warstadt and Samuel R. Bowman. 2019. Linguistic analysis of pretrained sentence encoders with acceptability judgments. *arXiv: Computation and Language*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

9. Language Resource References

Nianwen Xue et al. 2016. *Chinese Treebank 9.0*. Linguistic Data Consortium, LDC2016T13., 9.0.