Procesamiento del Lenguaje Natural, Revista nº 73, septiembre de 2024, pp. 177-190 recibido 03-04-2024 revisado 18-04-2024 aceptado 17-05-2024

CatCoLA, Catalan Corpus of Linguistic Acceptability

CatCoLA, Corpus Catalán de Aceptabilidad Lingüística

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

Abstract: We introduce CatCoLA, the Catalan Corpus of Linguistic Acceptability that will contribute to the Catalan Language Understanding Benchmark (CLUB) to assess and compare the capabilities of language models (LM) trained with texts in Catalan. CatCoLA follows the design of the English CoLA to support the task of classifying sentences as acceptable or not. Because the task is very dependent on the characteristics of particular languages, datasets cannot be translated from one language to another and the availability of these datasets for different languages requires specific developments. CatCoLA consists of 10,443 sentences and their acceptability judgements as found in well-known Catalan 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 dataset and the results of a human annotation experiment. The results are also analyzed and commented to guide future research. CatCoLA is released under a CC BY SA 4.0 licence and freely available at `https://doi.org/10.34810/data1393`.
Keywords: Catalan language, Corpus, Evaluation, Language Model, Linguistic Acceptability

Resumen: Presentamos CatCoLA, el Corpus Catalán de Aceptabilidad Lingüística que contribuirá al Catalan Language Understanding Benchmark (CLUB) con la misión de ayudar a evaluar y comparar las capacidades de los modelos del lenguaje (LM) entrenados con textos en catalán. CatCoLA sigue el diseño del CoLA inglés para la tarea de clasificar oraciones como aceptables o no. Dado que la tarea depende en gran medida de las características de las lenguas particulares, los datos no pueden traducirse de una lengua a otra y la disponibilidad de estos datasets para diferentes lenguas requiere desarrollos específicos. Nuestro corpus consta de 10.443 oraciones y los juicios de aceptabilidad correspondientes, tal y como se han encontrado en gramáticas catalanas de referencia. Además, todas las frases se han anotado con la clase del fenómeno lingüístico del que la frase es ejemplo, también siguiendo prácticas anteriores. También proporcionamos como referencia los resultados de la tarea de cuatro modelos del lenguaje diferentes y los resultados de un experimento de anotación humana. CatCoLA se publica bajo licencia CC BY SA 4.0 y está disponible gratuitamente en `https://doi.org/10.34810/data1393`.
Palabras clave: Aceptabilidad Lingüística, Catalán, Corpus, Evaluación, Modelo del Lenguaje

1 Introduction

We introduce the Catalan Corpus of Linguistic Acceptability (CatCoLA). CatCoLA is meant to provide data for the assessment of the capabilities of Catalan language models (LM) to handle linguistic information. CatCoLA has been developed following the example of Warstadt, Singh, and Bowman (2018), who compiled the first acceptability dataset, which became part of the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018). The GLUE benchmark was the first tool proposed as a common framework to evaluate and analyze the performance of language models. It is made of a diverse range of datasets meant to assess natural language understanding capabilities with tasks intended to challenge models. Although all GLUE datasets were originally in English, datasets for other

ISSN 1135-5948 DOI 10.26342/2024-73-13

languages are being developed, many by just translating the data and the annotations (Ruder et al., 2021). However, due to the highly language-dependent nature of the task, linguistic acceptability datasets cannot be translated and, therefore, their availability for more languages is lower. CatCoLA is intended to be part of the Catalan Language Understanding Benchmark[1] (CLUB) (Armengol-Estapé et al., 2021).

CatCoLA dataset consists of 10,443 sentences and their acceptability judgements. 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 in the original CoLA, while the fourteenth category gathers sentences containing specific Catalan phenomena, such as agreement in nominal constructions, subject ellipsis, 'ser'/'estar' copula selection, constructions with 'hi' pronoun, pronominal cliticization, use of subjunctive verbal mode and tense, consecutio temporum phenomena in subordinate clauses and phrase dislocation phenomena.

We present the corpora of linguistic acceptability available for other languages and related datasets in Section 2. In Section 3, we describe the details of CatCoLA dataset, we list and motivate the linguistic works that have been the sources of the sentences of the corpus, and we sum up the pre-processing tasks performed to curate the data. In Section 4, we describe the different experiments performed with the dataset created and the first baselines results for the acceptability task in Catalan. Results are presented in Section 5 and discussed in Section 6. Finally, the main contributions of our work are summarized in Section 7.

2 Related work

The first dataset designed to evaluate the performance of language models concerning linguistic acceptability was the English Corpus of Linguistic Acceptability (CoLA) (Warstadt, Singh, and Bowman, 2018), included in the GLUE benchmark. The CoLA dataset consists of 10k English sentences extracted from 23 theoretical linguistics publications covering various linguistic phenomena. The corpus was partitioned into training, development and test. Overall, acceptable sentences are around 70% of the dataset. Additionally, the sentences of the CoLA development set, a 10% of the corpus, were further annotated with the linguistic phenomenon they are an example of, following a list of thirteen different categories.

Taking CoLA as a reference, similar resources for more languages are becoming available. ItaCoLA (Trotta et al., 2021), for Italian, was the first dataset that followed the design of the English CoLA. Italian sentence sources are theoretical linguistics textbooks and works, focusing on specific phenomena such as idiomatic expressions, locative constructions and verb classification. ItaCoLA consists of 10k sentences annotated with acceptability binary judgements as originally found in selected linguistic publications. The percentage of acceptable sentence amounts to 85.4%, and 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.

The Spanish Corpus of Linguistic Acceptability (EsCoLA) (Bel, Punsola, and Ruiz-Fernández, 2024), collects about 11k sentences from different grammar reference books and papers. Like CoLA, the ratio of acceptable-unacceptable sentences is 70%-30%. Additionally, all sentences were annotated with the linguistic phenomenon they exemplify from the list of thirteen linguistic categories originally proposed by Warstadt and Bowman (2019) for CoLA. In addition, EsCoLA added a fourteenth class to annotate specific Spanish phenomena, such as agreement in nominal constructions, subject ellipsis, pronominal cliticization, 'ser'/'estar' copulative verb selection, and tense and mood restrictions in subordinate clauses.

Swedish DaLAJ (Volodina, Mohammed, and Klezl, 2021) is made of 9,596 sentences that correspond to 4,798 sentences extracted from SweLL (Volodina et al., 2019), a second language learner corpus, taken as unacceptable samples, plus the corresponding acceptable version. The DaLaJ unacceptability judgments were produced by teachers, assessors, or trained assistants, and sentences were also annotated with information about the error.

[1] `https://club.aina.bsc.es/`

The Russian Corpus of Linguistic Acceptability (RuCoLA) (Mikhailov et al., 2022) consists of 13,4k sentences, of which acceptable sentences amount to 71.8%. 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 labelled with four different categories: morphology, syntax, semantics, and hallucinations. Differently from previous corpora, RuCoLA was used to evaluate a text generation system and its metrics are not directly comparable to the results of other corpus of acceptability.

NoCoLA, the Norwegian corpus of linguistic acceptability (Jentoft and Samuel, 2023), used as source data the ASK Corpus (Tenfjord, Meurer, and Hofland, 2006), a language learner corpus of Norwegian as a second language. It is made of two corpora: the first dataset, NoCoLAclass, with acceptability annotations, contains 144,867 sentences. Note that in this corpus only 31.5% of the sentences are grammatically acceptable. The second dataset, NoCoLAzero, is a collection of pairs of sentences, of which only one is grammatically acceptable, and follows the dataset schema of the Benchmark of Linguistic Minimal Pairs for English, BLiMP (Warstadt et al., 2020). BLiMP is an extension of the first CoLA corpus, which contains 67k pairs of ungrammatical and their corresponding grammatical sentences automatically generated via manually-constructed templates that span 12 high-level English phenomena. Similar datasets are the ones developed by Hartmann et al. (2021) for Bulgarian and German, CLiMP (Xiang et al., 2021) and JBLiMP (Someya and Oseki, 2023). These are made of such minimal pairs and used to fine-tune a model for particular linguistic probing tasks, although different to the linguistic acceptability task.

JCoLA, the Japanese version of CoLA, is made out of 10k sentences from textbooks and handbooks focusing on Japanese syntax (out-domain set), plus a well-known linguistics journal article (out-domain set). A 83.4% and a 82% of the sentences are acceptable in the in-domain and out-domain set, respectively.

The Hungarian CoLA corpus (HuCoLA) is part of the Hungarian Language Understanding Benchmark Kit (HuLU) (Ligeti-Nagy et al., 2024). It consists of 9,944 examples from major linguistic articles manually labelled by 4 annotators, the final label being then agreed on majority decision. Despite being also available, the authors excluded the original sentence labels determined by the authors of the sources. Following the general trend, 78% are considered acceptable.

SLING, which stands for Sino-Linguistic Evaluation of Large Language Models, is a corpus consisting of 38,000 minimal sentence pairs in Mandarin Chinese (Song et al., 2022). These pairs are grouped into nine high-level linguistic phenomena, many of which are unique to the Chinese language. To create SLING, the authors utilized the Chinese Treebank 9.0 (Nianwen Xue et al., 2016). They extracted subtrees from human-validated constituency parses and transformed them with manually designed linguistic templates to create minimal pairs of acceptable and unacceptable sentences. Also, these sentences were validated by human annotators.

CoLA datasets are being used for the acceptability probing task, i.e. fine-tuning an LM to classify sentences as acceptable or not in a particular language. The standard practice is to measure the performance of the classifiers with the Matthews Coefficient Correlation (MCC) (Matthews, 1975) and with accuracy metrics. MCC was chosen because it is considered to be a robust metric that shows in a single value the performance of the classifier for binary values, despite the occasional unbalance of the samples. The best performance in the task with the English CoLA corpus reported in Warstadt and Bowman (2019) was achieved by a BERT-large fine-tuned classifier with MCC=0.58. Because CoLA is in the GLUE benchmark, posterior better results, getting around MCC=0.75, have been published in the leaderboard[2] as achieved with different LM architectures.

As for the other languages, the performance of the Ita-BERT is reported to be MCC=0.67 for the ItaCOLA dataset just described. RuCoLA baselines were obtained with six different LMs, four monolingual and two multilingual ones. RuRoBERTa was the best one, with MCC=0.53 for the in-domain dataset. JCoLA was also used to evaluate

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

several LMs, being Waseda RoBERTa-large the one giving the best results (MCC=0.46). Similarly, PULI BERT-large was the best-performing model (MCC=71.1) among the Hungarian LMs evaluated on HuCoLA. Finally, the Spanish EsCoLA was used to compare four different LMs, including 2 multilingual ones. The multilingual mDeBERTa-v3 performed the best as achieved MCC=0.52 (average value from a five cross-validation experiment).

3 CatCoLA: Catalan Corpus of Linguistic Acceptability

The Catalan Corpus of Linguistic Acceptability (CatCoLA) is built following the methodology proposed by the English Corpus of Linguistic Acceptability (Warstadt, Singh, and Bowman, 2018) to be used to assess Catalan large language models' capabilities of capturing linguistic information. The corpus contains 10,443 sentences annotated as acceptable or not acceptable and classified according to a list of fourteen linguistic classes. CatCoLA size and distribution of annotations are similar to the other corpora of linguistic acceptability as shown in Table 1.

Corpus	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
HuCoLA	Hungarian	9.9	78
JCoLA	Japanese	10	82
EsCoLA	Spanish	11.1	70
CatCoLA	Catalan	10.4	70

Table 1: Comparison of CatCoLA with similar acceptability corpora for other languages. The language of the dataset, the size in thousands, and the percentage of acceptable (accep.) sentences are indicated.

3.1 Corpus composition and sources

CatCoLA is made of two subsets: the in-domain set (InDomain) and the out-of-domain set (OutDomain). The 10,189 sentences in the CatCoLA InDomain set were extracted from the examples illustrating acceptable and unacceptable Catalan sentences from several sources. Most of the examples come from two sources: "Gramàtica del català contemporani" (GCC) (Solà and Rigau, 2002), a prestigious reference Catalan grammar, and the Catalan course for foreign learners of the Consorci per a la Normalització Lingüística (CPNL)[3]. GCC is a compilation of 31 articles from different authors, all of them renowned Catalan linguists, each covering the description of a linguistic phenomenon and with examples taking into account Catalan regional variants and colloquial register. The Catalan course of the CPNL was crawled from its web[4]. To find unacceptable examples to reach the 30% of the corpus and for the linguistic classes used by other acceptability corpora, it was necessary to review other sources, mostly journal articles, which are listed in Appendix A.

As for the 254 sentences of the CatCoLA OutDomain corpus, they were randomly extracted from the ParlaMint Catalan corpus (Pisani, Zevallos, and Bel, 2023), a corpus made of transcribed actual Catalan speakers productions in parliamentary debates[5]. Thus, OutDomain sentences are from a very different linguistic setting with no reference annotations. Acceptability annotations of the OutDomain sentence annotations were made by consensus of two experienced linguists. Moreover, to reach up to 30% of unacceptable sentences, roughly 50 sentences were manually modified to make them unacceptable. Selecting these sentences we aimed to represent, as far as possible, all the linguistic phenomena. For example, the sentence 'Els boscos són més propensos a patir incendis tot l'any' ('Forests are more prone to fires all year round') has been transformed into an unacceptable sentence by changing the preposition 'a' to 'de': 'Els boscos són més propensos de patir incendis tot l'any'. Details of the number of acceptable and unacceptable sentences are given in Table 2.

		Unacceptable	Total
InDomain	train	2404 (29.4%)	8151
	dev	284 (27.8%)	1018
	test	319 (31.2%)	1020
OutDomain	test	76 (29.9%)	254

Table 2: Number of unacceptable and total sentences in CatCoLA corpus per split and domain.

[3] `https://www.cpnl.cat`

[4] All crawled pages were CC BY 4.0 licensed.

[5] The ParlaMint-ES-CT corpus is also CC BY 4.0 licensed.

3.2 Annotation of linguistic classes

For the InDomain set, the linguistic class per sentence was annotated by an expert linguist who took as reference the topic of the paper or the chapter of the grammar the sentence was used as an example. For the OutDomain set, the class was discussed with another linguist to reach a consensus. Here follows the description of the categories as used in the classification exercise, and some examples of acceptable and unacceptable sentences for each class.

1. Simple. Sentences with a verb and a complete mandatory set of subcategorized complements. All arguments are noun phrases and there are no modifiers or adjuncts at any level. Pro-drop romance phenomena are not included. Ex.: L'Anna estudia grec. Les abelles produeixen mel. *Els humans temen. *El ginecòleg va néixer l'infant.

2. Predicative. Copulative sentences, small clauses and resultatives. Ex.: El sector ha esdevingut minoritari. El professor considerà el treball bonníssim. *La Maria és. *En Pere ha quedat intel·ligent.

3. Adjuncts. Sentences showing optional modifiers for NPs and VPs and temporal and locative adjuncts. Ex.: Amb la Maria de directora, l'empresa funciona millor. Avui han acabat els exàmens. *Va trobar el rellotge durant cinc minuts. *La Joana insisteix demanant un augment de sou.

4. Argument types. Oblique and prepositional arguments subcategorized by a verb, a noun or an adjective. Ex.: Tinc dret a una pensió. Prepara el sopar per a les nenes. *S'ha deixat prendre el pèl a uns pocavergonyes. *Has d'enviar aquest paquet en aquell noi.

5. Argument alternations, high-arity, passives, including reflexive passives, drop-args and add-args constructions. Ex.: Aquesta opinió serà defensada. El testament fou impugnat. *La Berta ha endut un llapis. *En Joan s'ha emmalaltit.

6. Binding pronouns. Referential expressions. Ex.: Feu-ho vosaltres mateixes. Cal que ens animem els uns als altres. *La roba s'ha cremat per si sola. *En Toni s'ha empassat un pinyol a si mateix.

7. Wh-phenomena. Questions, reported speech and relatives (exclamatives have been excluded). Ex.: La dona que riu és sueca. En Joan ignorava si calia fer-ho. *Vaig escriure la col a qui se la menja crua. *S'ha quedat al qual no li trauríeu sang si ho provéssiu.

8. Complement clauses, including subjects, arguments of VPs, NPs or APs. Ex.: En Joan vol que li regali un poni. L'humanista sap que l'estupidesa humana no té límits. *Va declarar que on tenia amagats els diners. *Vull que molt sovint.

9. Auxiliary and modal verbs, negation, polarity and periphrastic verbal constructions. Ex.: Continua traient la pols. No tinc cap rellotge. *El Pep compra no pomes. *Tens que fer els deures cada dia.

10. Infinitival embedded VPs involving referential obligatory phenomena like control, raising, and VP, NP or AP argumental constructions. Ex.: He sentit cantar la Maria. Aquest llibre és molt difícil d'entendre. *M'agraden les persones fàcils de començar a satisfer. *El rei admirava el dolçament despertar de la princesa.

11. Complex NPs and APs with prepositional arguments and relational adjectives with obligatory complements. Ex.: En Martí és un home orgullós del seu fill. Aquell jove era propens a la depressió. *Hem instal·lat una solar placa a l'edifici. *L'elecció presidencial dels vocals es va fer a porta tancada.

12. S-syntax phenomena: coordination, subordination and sentence-level adjuncts. Ex.: Encara que protestin, tirarem el projecte endavant. He deixat la feina quan he pogut. *Sabia prou coses com perquè havia estat metge. *Com que eliminades les atletes russes, el campionat va perdre emoció.

13. Determiners, quantifiers, partitives, and comparative constructions. Ex.: He vist un gos tan afectuós com el teu. Tot Catalunya patirà els efectes de la tempesta. *El gosset va néixer poc. *He comprat una llet.

The linguistic annotation of the CatCoLA sentences is meant to facilitate a 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 Catalan. Annotated Catalan phenomena are the following:

14.1 Agreement in nominal constructions. Ex.: Les teves germanes són bones conductores. Les he endevinades totes. *Les coses que s'ha de fer de pressa sempre queden mal feta. *L'euga ha arribat esgotats.

14.2 Ellipsis. Ex.: Compro bombons. Érem a casa. *Suposo que has llegit mateix la carta. *Sou molt treballadors, però no són gens treballadors.

14.3 'Ser'/'estar' copula selection. Ex.: La Maria és professora. L'Anna està embarassada. *En Pere està intel·ligent. *L'euga és cansada.

14.4 Constructions with the 'hi' unvoiced pronoun. Ex.: Hi ha ratolins a les golfes. Em sembla que necessites ulleres perquè no t'hi veus. *Hi seran homes. *Hi ha jo.

14.5 Cliticization phenomena. Ex.: Les teves orquídies li han agradat molt a l'Anna. Vaig donar-los-la ahir. *Lis escriuen una carta. *Al president de la telefònica els hi plouen, els millions.

14.6 Subjunctive mode and tense and consecutio temporum violations. Ex.: En Lluís creia que ens reuniríem demà. M'estava dient que demà ja hauria corregit els exàmens. *Et truca perquè pensis que es preocupi per tu. *En Joan vingui potser demà.

14.7 Dislocation. Ex.: Pa, falta: llet en tenim. En Joan, fa temps que vull veure. *Els llibres va donar als xiquets. *Pagarà la companyia la factura.

3.3 Data Processing

Source texts were either Word .doc files or printed copies that had to be scanned and digitized with OCR software. Once digitized, regular expression patterns were manually developed to identify the lines that included examples. Indentation and numbering were key clues for finding potential corpus sentences. Once a first list of sentences was created, some curation was required to correct typical OCR errors. Later, we looked for '*', which is the standard unacceptability mark in linguistics to label the sentences. We discarded examples marked as of dubious acceptability with '?' or signs other than '*'. However, examples that included acceptability alternations (for instance '(*)', meaning that the example is unacceptable if the text in the parenthesis is in the sentence, but acceptable if it is not) were taken and the two versions, the acceptable and the unacceptable sentence, were created.

To assess the results of the automatic annotation, all sentences in the InDomain set were compared to the human annotation. As detailed in Section 5, results show that only 2.2% of the sentence acceptability contradicted the human judgement.

Note that most of the examples that were not full sentences, i.e. they had no main verb, were discarded. They were only accepted if they were meant to illustrate the unacceptability of sentences that do not contain a modal verb when there should be one. For instance, the sentence '*Recordar que s'ha actualitzat la normativa' ('To remember that the regulation has been updated') is unacceptable because of the lack of a tensed verb, but 'Cal recordar que s'ha actualitzat la normativa' ('It is necessary to remember that the regulation has been updated') is considered acceptable.

Finally, like in other linguistic acceptability corpora, we manually substituted very low-frequency words[6] with synonyms and maintained the diacritical accents as they appear in the sources, even though some of them are no longer normative according to the canonical grammar (IEC, 2016).

4 Experiments

To provide the first baseline results for the acceptability task with the CatCoLA dataset, we have performed experiments with the Catalan RoBERTa-v2 (RoBERTa-ca-v2) (Armengol-Estapé et al., 2021), a monolingual language model based on the transformed-based model RoBERTa (Liu et

[6]Found less than 45 times in the Catalan Timestamped JSI web corpus 2014-2021, with 450M tokens, available at `https://www.sketchengine.eu/`.

al., 2019), which has been already evaluated on the other downstream tasks of the Catalan Language Understanding Evaluation benchmark (CLUB). RoBERTa-ca-v2 was pre-trained on a high-quality Catalan corpus (1.7B tokens) gathered from publicly available corpora and crawlers. The vocabulary of this Catalan model, with a size of 50k tokens, was learnt from scratch using the training set of the corpus compiled, which was tokenized using Byte-Level BPE (Radford et al., 2019).

In addition, in order to compare the results of the previous monolingual model with the ones of a multilingual model with no language-specific tokenization or pre-training, we also fine-tuned XLM-RoBERTa (Conneau et al., 2020). This model is the multilingual version of RoBERTa with a vocabulary of 250k tokens. It was pre-trained on clean CommonCrawl data containing texts in 100 different languages, including Catalan (1,752M tokens) and using Sentence Piece tokenizer (Kudo and Richardson, 2018). More specifically, our experiments were carried out using the base and large sizes of the mentioned models. Table 3 goes into more details about the parameters of the models used for getting the baselines.

Finally, we asked three linguists (two postgraduates and one postdoctoral researcher), all native Catalan speakers, to complete the acceptability task for all the InDomain set, to approach the upper bound to compare the machine performance.

	roberta-ca-v2		**xlm-roberta**	
	base	**large**	**base**	**large**
W	1.7	1.7	167 (1.7)	167 (1.7)
L	12	24	12	24
H	768	1024	768	1024
A	12	16	12	16
V	50	50	250	250
P	110	355	270	550
Tok.	BPE	BPE	SPM	SPM

Table 3: Details on model sizes used as baselines. W: training corpus number of tokens in billions (number of Catalan tokens in parenthesis), L: layer size, H: hidden size, A: attention heads, V: vocabulary in thousands, P: number of parameters in millions.

All four language models were evaluated in an InDomain and and OutDomain setting, as done with the English CoLA (Warstadt and Bowman, 2019). As for the InDomain dataset, we divided the dataset into training (80%), validation (10%) and test (10%) splits, as explained in Section 3, preserving the original 70% acceptable and 30% not-acceptable ratio as well as the linguistic classes distribution (Figure 1). With these partitions, we performed 10 runs with different random seeds. Models were fine-tuned for 5 epochs with a maximum sequence length of 128, a batch size of 64 and a learning rate set at 2e-5. All model implementations, along with the code for fine-tuning and evaluation, are sourced from the Hugging Face's Transformers library[7] (Wolf et al., 2020).

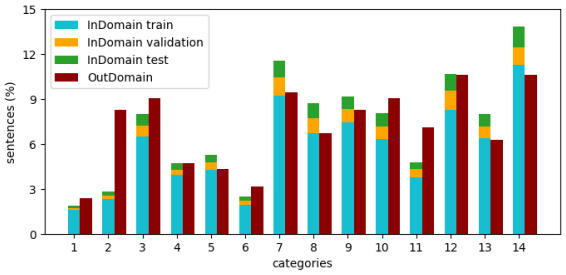

Figure 1: Percentage of samples of annotated linguistic categories in CatCoLA InDomain and OutDomain sets. 1: Simple, 2: Predicative, 3: Adjuncts, 4: Argument types, 5: Argument alternation, 6: Binding pronouns, 7: Wh-phenomena, 8: Complement clauses, 9: Modal verbs, negation, periphrasis and auxiliaries, 10: Infinitive embedded VPs and referential phenomena, 11: Complex NPs and APs, 12: S-syntax, 13: Determiners, quantifiers, comparative and superlative constructions, 14: Catalan phenomena.

In line with prior research, we measured the performance of the models using both accuracy score (acc.) and the Matthews Correlation Coefficient (MCC) (Matthews, 1975). As mentioned in Section 2, accuracy may not offer detailed insights for an imbalanced dataset such as ours. Instead, MCC is widely recognized as a robust metric that effectively computes the model performance when positive and negative cases hold equal significance.

Regarding the OutDomain set, for each model, the run obtaining the best performance in the InDomain setting was then tested on the entire OutDomain set.

[7] https://github.com/huggingface/transformers

5 Results

Table 4 shows Cohen-Kappa and MCC scores for each human annotator compared to the CatCoLA reference. The average of human annotators with the reference was MCC=0.69. There were 225 cases (2.2% of the whole InDomain dataset) for which the three annotator's decision contradicts the label of the reference. Table 4 show the details of the human performance on the task in MCC and Cohen Kappa agreement.

Annotator	Cohen-K	MCC
A1	0.829	0.83
A2	0.594	0.61
A3	0.646	0.64

Table 4: Cohen Kappa and MCC scores obtained by the comparing three different annotators and the reference.

The performance of the language models for both the InDomain and the OutDomain experiments are reported in Table 5. For the InDomain experiments, the results, averaged over the 10 runs, show that the best model is the base version of RoBERTa-base-ca-v2 (MCC=0.52), which is followed by RoBERTa-large-ca-v2 (MCC=0.46). However, of note is that the best run (MCC=0.62) of the large version outperforms the best run of the base one (MCC=0.55) and gives the overall highest performance. On the other hand, the multilingual XLM-RoBERTa base and large versions obtain significantly lower results (MCC=0.26 and MCC=0.05, respectively). Among these, it is again the larger version the one performing the worst, with results close to random.

Regarding the performance of the models on the specific linguistic category, Figure 2 shows the mean MCC scores for the InDomain test dataset. The categories that show lower results are the following: 8, complement clauses; 11, Complex NPs and APs; 12, Sentential Syntax, and 14, the category for all Catalan linguistic phenomena. For this last category, Figure 3 shows the score by phenomena[8].

As for the OutDomain experiment, once again, the monolingual Catalan-BERTa-v2 outperforms the multilingual XLM-RoBERTa. It is noticeable that the larger versions of the models yield better MCC results (MCC=0.59 and MCC=0.11, respectively) than the base ones (MCC=0.40 and MCC=0.10, respectively), which are significantly much higher in the case of the Catalan-BERTa-v2. In fact, the results obtained from testing both large models on the OutDomain set seem to be higher than their corresponding mean classification scores on the InDomain test set. However, we must notice that the OutDomain MCC scores are still lower than the corresponding results for the best run in the InDomain setting, which, as mentioned, is the one used for the OutDomain experiment.

6 Discussion

Both for the InDomain and OutDomain experiments, RoBERTa-ca performs better than XLM-RoBERTa. Note that the problems of a small pre-training dataset are also relevant in multilingual models. However, the good results of the monolingual model with respect to the multilingual one highlights the importance of relying on language-specific models pre-trained with high-quality data, such as RoBERTa-ca, to solve the linguistic acceptability task in Catalan. In fact, Conneau et al. (2020) showed that low-resource languages are under-represented in the vocabulary of these models and that subword tokenizers, trained jointly on multiple languages, tend to over-split the tokens to cover the vocabulary of many languages, which makes it difficult for the language model to learn good quality representations.

Results also show that the best mean MCC scores are obtained with the base version of RoBERTa-ca, whereas the large version shows more variable results. However, it is the RoBERTa-large-ca the one obtaining the highest performance when comparing the results of the best run among the 10 runs performed. As noted before, for the experiment with the OutDomain set, the large version also gives the best results, although it scores a MCC lower than the one obtained by the best run with the InDomain dataset. Therefore, it seems that there might have been some difficulties for the classifier to handle real test sentences after being trained with linguistic examples.

As for the comparison with human annotators, note that the average MCC=0.69

[8]For all models, performance for ellipsis phenomena is 0.

model	InDomain		OutDomain	
	MCC	acc.	MCC	acc.
RoBERTa-base-ca-v2	0.52 ±0.02 (best: 0.55)	0.80 ±0.01 (best: 0.81)	0.40	0.77
RoBERTa-large-ca-v2	0.46 ±0.25 (best: 0.62)	0.79 ±0.06 (best: 0.84)	0.59	0.83
XLM-RoBERTa-base	0.26 ±0.11 (best: 0.35)	0.72 ±0.02 (best: 0.74)	0.10	0.70
XLM-RoBERTa-large	0.05 ±0.11 (best: 0.30)	0.68 ±0.02 (best: 0.72)	0.11	0.70

Table 5: Classification results on the CatCoLA InDomain test set and OutDomain set. For InDomain results, results are the mean of 10 runs ± StdDev, with best result between parenthesis. OutDomain results are obtained from the best run.

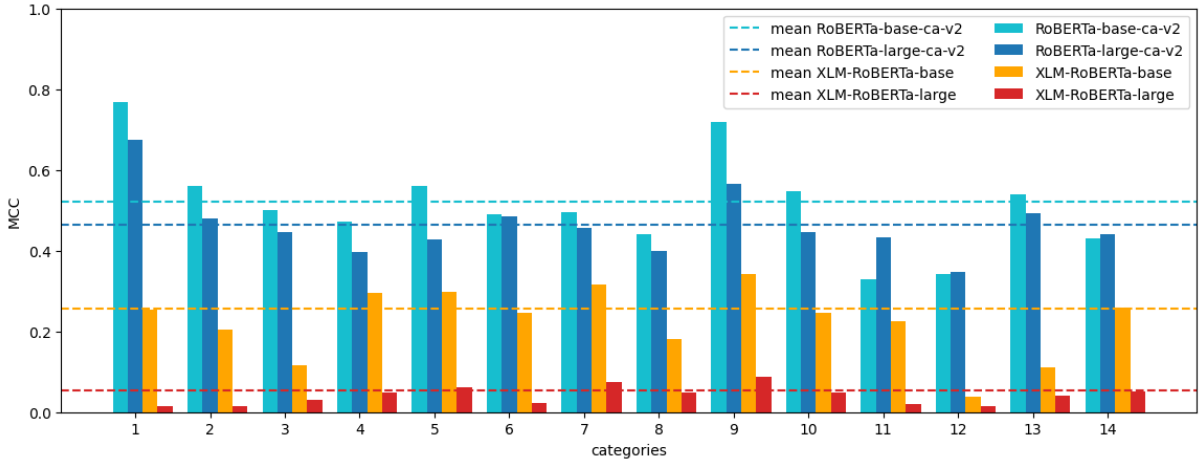

Figure 2: MCC classification results on the CatCoLA InDomain test set per model and linguistic category. Results are the mean of 10 runs. Dashed lines show the mean scores over all categories. Categories are 1: Simple, 2: Predicative, 3: Adjuncts, 4: Argument types, 5: Argument alternation, 6: Binding pronouns, 7: Wh-phenomena, 8: Complement clauses, 9: Modal verbs, negation, periphrasis and auxiliaries, 10: Infinitive embedded VPs and referential phenomena, 11: Complex NPs and APs, 12: S-syntax, 13: Determiners, quantifiers, comparative and superlative constructions, 14: Catalan phenomena.

achieved by them is still higher than the best run of RoBERTa-large-ca. It is also of interest to compare the errors in recognizing unacceptable sentences along the linguistic categories to analyse differences between the fine-tuned classifiers and humans, as shown in Figure 4. Both humans and RoBERTa-base have trouble with sentences of category 10, which contains embedded infinitive verb phrases and displays referential restrictions like 'M'agraden les persones fàcils de començar a satisfer' ('I like people easy to start to please'). Also, RoBERTa-base and human annotators show also problems with category 7, wh-phenomena, which deals with relative and interrogative pronouns that bear graphical accents. These are frequently not correctly written in informal texts, such as the ones that can be found in the crawled data used to pre-train models. The classifiers show very low recall, predicting acceptability for almost all these sentences, while humans made also several errors by tagging as unacceptable sentences that are acceptable according to the reference, showing that there are spelling problems.

Simple sentences (1) and modal verbs, negation and periphrastic constructions (9) are the categories for which RoBERTa performs the best, with an MCC close to 0.8, as shown in Figure 2. Note that none of them are the categories with more train data, as shown in 2. Thus, there seems to be no correlation between the amount of training examples and the performance of the models. For category 9, the presence of the negative adverb 'no', or particular modal verbs could be hints for the classifiers, explaining the good results they obtained also compared to the errors that humans made for this category, as shown in Figure 4.

It is remarkable that the categories for

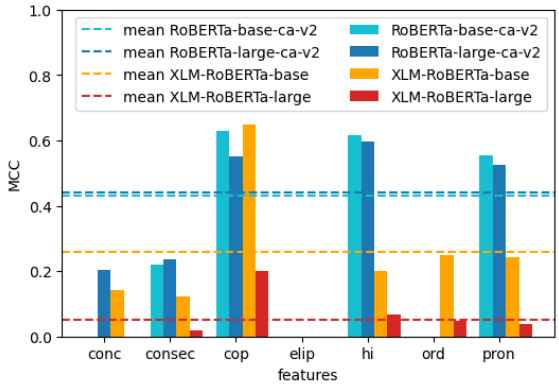

Figure 3: MCC classification results on different Catalan-specific phenomena of CatCoLA InDomain test set. Results are the mean of 10 runs. Dashed lines show the mean scores over all Catalan phenomena (i. e. category 14). conc: agreement, consec: consecutio temporum, cop: copula selection, elip: ellipsis, hi: 'hi' unvoiced pronoun, ord: dislocation, pron: clitization phenomena.

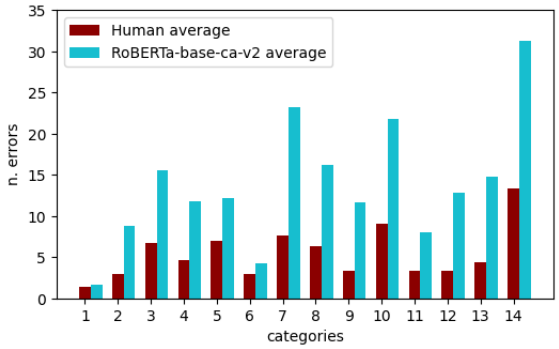

Figure 4: Number (average) of errors per category of unacceptable sentences that humans and RoBERTa-base-ca-v2 classify as acceptable. Categories are 1: Simple, 2: Predicative, 3: Adjuncts, 4: Argument types, 5: Argument alternation, 6: Binding pronouns, 7: Wh-phenomena, 8: Complement clauses, 9: Modal verbs, negation, periphrasis and auxiliaries, 10: Infinitive embedded VPs and referential phenomena, 11: Complex NPs and APs, 12: S-syntax, 13: Determiners, quantifiers, comparative and superlative constructions, 14: Catalan phenomena.

which RoBERTA-base has problems are those with larger sentences. For instance, category 12 contains examples of 10 words per sentence on average, while the total average of the test set is 6.8 words per sentence. In contrast, for categories 10 and 11 the average length is 7.43 words per sentence. Category 10 corresponds to Infinitive embedded VPs and referential phenomena with sentences like 'Volem tornar' ('We want to return'), and category 11, to complex phrases phenomena like in 'És un atleta ample d'espatlles' ('He is a broad-shouldered athlete'). Despite having a very similar average of words per sentence, RoBERTA-base performs above average for category 10, but below for category 11. Thus, for some categories, performance may be correlated to the length of the sentences.

7 *Conclusions*

We have introduced the Catalan Corpus of Linguistic Acceptability CatCoLA, which constitutes the first dataset for the acceptability probing task for Catalan. It follows the example of the datasets for the same task already developed for other languages than English: Spanish, Italian, Norwegian, Swedish, Russian, Hungarian and Japanese, which could not be mere translations as the task is very language-dependent. The creation of CatCoLA wants to promote the fair evaluation and comparison of existing and future large language models that want to handle Catalan, joining the efforts already made to develop the Catalan Language Understanding Benchmark (Armengol-Estapé et al., 2021). CatCoLA dataset consists of 10,443 sentences and their acceptability judgements as found in Catalan reference grammars and linguistic papers. The annotation provided with the dataset also includes annotations of the linguistic category the sentence is an example of. Acceptability judgments of three linguists native speakers of Catalan are also provided for the InDomain subcorpus. In this paper, we have also reported the first task baselines obtained by fine-tuning four different language models for InDomain and OutDomain subcorpus. CatCoLA is released under a CC BY 4.0 licence and is freely available at `https://github.com/nuriabel/LUTEST`.

Acknowledgments

This research is part of the LUTEST project, PID2019-104512GB-I00, funded by the Ministerio de Ciencia, Innovación y Universidades and Agencia Estatal de Investigación (Spain). 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 Mariona Amengual. We are also very grateful to the authors of GCC and other papers who generously sent us their digital files.

References

Armengol-Estapé, J., C. P. Carrino, C. Rodriguez-Penagos, O. de Gibert Bonet, C. Armentano-Oller, A. Gonzalez-Agirre, M. Melero, and M. Villegas. 2021. Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4933–4946, Online, August. Association for Computational Linguistics.

Bel, N., M. Punsola, and V. Ruiz-Fernández. 2024. EsCoLA: Spanish Corpus of Linguistic Acceptability. In *Proceedings of he t2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024))*, Torino, Italy.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Hartmann, M., de Miryam Lhoneux, D. Hershcovich, Y. Kementchedjhieva, L. Nielsen, C. 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, November. Association for Computational Linguistics.

Jentoft, M. and D. 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, May. University of Tartu Library.

Kudo, T. and J. Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Ligeti-Nagy, N., G. Ferenczi, E. Héja, L. J. Laki, N. Vadász, Z. G. Yang, and T. Váradi. 2024. HuLU: Hungarian language understanding benchmark kit. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8360–8371, Torino, Italia, May. ELRA and ICCL.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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

Mikhailov, V., T. Shamardina, M. Ryabinin, A. Pestova, I. Smurov, and E. 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, December. Association for Computational Linguistics.

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

Pisani, M., R. Zevallos, and N. Bel. 2023. Catalan parliamentary plenary session transcriptions from 2015 to 2022. the parlamintcat corpus. *Procesamiento del Lenguaje Natural*, 71(0):125–136.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ruder, S., N. Constant, J. Botha, A. Siddhant, O. Firat, J. Fu, P. Liu, J. Hu, D. Garrette, G. Neubig, and M. Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Solà, J. and G. . Rigau. 2002. *Gramàtica del català contemporani*. Empúries, Barcelona, 4a ed. edition.

Someya, T. and Y. Oseki. 2023. JBLiMP: Japanese benchmark of linguistic minimal pairs. In A. Vlachos and I. Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia, May. Association for Computational Linguistics.

Song, Y., K. Krishna, R. Bhatt, and M. 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, December. Association for Computational Linguistics.

Tenfjord, K., P. Meurer, and K. 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, May. European Language Resources Association (ELRA).

Trotta, D., R. Guarasci, E. Leonardelli, and S. 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, November. Association for Computational Linguistics.

Volodina, E., L. Granstedt, A. Matsson, B. Megyesi, I. Pilán, J. Prentice, D. Rosén, L. Rudebeck, C.-J. Schenström, G. Sundberg, and M. Wirén. 2019. The swell language learner corpus: From design to annotation. *Northern European Journal of Language Technology*, 6:67–104.

Volodina, E., Y. A. Mohammed, and J. 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, May. LiU Electronic Press.

Wang, A., A. Singh, J. Michael, F. Hill, O. Levy, and S. 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, November. Association for Computational Linguistics.

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

Warstadt, A., A. Parrish, H. Liu, A. Mohananey, W. Peng, S.-F. Wang, and S. R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Warstadt, A., A. Singh, and S. R. Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Xiang, B., C. Yang, Y. Li, A. Warstadt, and K. Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In P. Merlo, J. Tiedemann, and R. Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics:*

Main Volume, pages 2784–2790, Online, April. Association for Computational Linguistics.

A List of other sources of InDomain subcorpus

- Abril, Joan (1997). Diccionari pràctic de qüestions gramaticals. Barcelona: Edicions 62. Nova ed., 2010. Barcelona: Educaula.
- Abril, Joan; Riera, Elvira (1997). L'ús dels possessius. Llengua i Ús (Barcelona), n. 10, p. 32-36.
- Abrines, Bartomeu (2011). Els verbs de canvi d'estat en català: la participació en l'alternança causativa. Caplletra (València), n. 50, p. 35-65.
- Albareda, Cristina (2013). La duplicació pronominal en les relatives locatives en català. Zeitschrift für Katalanistik (Friburg, Bochum), v. 26, p. 275-299.
- Ballesta, Joan-Manuel (1987). Algunes consideracions entorn dels verbs copulatius en català. Llengua & Literatura (Barcelona), n. 2, p. 359-375.
- Bartra, Anna (2016). Els components de la passiva. Una perspectiva diacrònica. Caplletra (València), n. 61, p. 295-327.
- Bartra, Anna; Brucart, Josep Maria (1982). Alguns arguments a favor de la categoria sintagma predicatiu. Els Marges (Barcelona), n. 24, p. 91-113.
- Busquets, Joan (2006). Stripping vs. VP ellipsis in Catalan: what is deleted and when? Probus (Dordrecht), v. 18, n. 2, p. 159-187.
- Cuenca Ordinyana, Maria Josep (2006), La connexió i els connectors. Perspectiva oracional i textual. Vic. Eumo Editorial.
- Cuenca Ordinyana, Maria Josep (2012), Sintaxi catalana. Barcelona: Editorial UOC:
- Espinal, M. Teresa (2000). Sobre les expressions lexicalitzades. Els Marges (Barcelona), n. 67, p. 7-31.
- Espinal, M. Teresa (2010). Bare nominals in Catalan and Spanish. Their structure and meaning. Lingua (Amsterdam), v. 120, n. 4, p. 984-1009.
- Generalitat de Catalunya. Departament de Justícia. Curs de llengua catalana. Nivell C.
- Hernanz, M. Lluïsa; Rigau, Gemma (1984). Auxiliaritat i reestructuració. Els Marges (Barcelona), n. 31, p. 29-51.
- Mascaró, Joan et al. (1984) Estudis gramaticals. Universitat Autònoma de Barcelona, p. 109-148.
- Rigau, Gemma (1990). Les propietats d'"agradar": estructura temàtica i comportament sintàctic. Caplletra (València), n. 8, p. 7-19.
- Rigau, Gemma (1993). El comportamiento sintáctico de los predicados existenciales en catalán. Revista de Lenguas y Literaturas Catalana, Gallega y Vasca (Madrid), v. 3, p. 33-53.
- Rigau, Gemma (1994). Les propietats dels verbs pronominals. Els Marges (Barcelona), n. 50, p. 29-39.
- Solà, Jaume (2002). Clitic climbing and null subject languages. Catalan Journal of Linguistics (Bellaterra), v. 1, p. 225-255.
- Solà, Joan; Lloret, Maria-Rosa; Mascaró, Joan; Pérez Saldanya, Manuel (dirs.) (2002). Gramàtica del català contemporani. Amb la col·laboració de Gemma Rigau. 4a ed., definitiva, 2008. Barcelona: Empúries.
- Viana, Amadeu (1990). La sintaxi de la conjugació en català. Caplletra (València), n. 8, p. 81-105.
- Villalba, Xavier (1992). Case, incorporation, and economy: an approach to causative constructions. Catalan Working Papers in Linguistics (Bellaterra), v. 2, p. 345-389.
- Villalba, Xavier (1994a). Clitic climbing in causative constructions. Catalan Working Papers in Linguistics (UAB), v. 3, n. 2, p. 123-152.
- Villalba, Xavier (1994b). Clitics, case checking, and causative constructions. Kansas Working Papers in Linguistics (Kansas), v. 19, n. 1, p. 125-147.
- Villalba, Xavier (2004). Descripció i norma: amb i el canvi de preposició.

Llengua & Literatura (Barcelona), n. 15, p. 257-276.