CatCoLA:
Catalan Corpus of Linguistic Acceptability

Núria Bel[1], Marta Punsola[1], **Valle Ruiz-Fernández**[2]
[1]Universitat Pompeu Fabra
[2]Barcelona Supercomputing Center

What do we mean by *linguistic acceptability*?

	(1a)	(1b)
CA	***L'Anna estudia grec.***	****L'euga ha arribat esgotats.***
ES	*Anna estudia griego.*	**La yegua ha llegado agotados.*
EN	*Anna studies Greek.*	**The mare has arrived exhausted.* (f. sg.) (m. pl.)
	acceptable	**unacceptable**

Linguistic acceptability datasets

Corpus of Linguistic Acceptability (CoLA)

(Warstadt, Singh, and Bowman, 2018)

extracted from theoretical linguistics publications	covering 13 linguistic phenomena	annotated with binary acceptability judgements	divided into: ◦ in-domain set ◦ out-of-domain set

The book was written by John. `acceptable`

Kim persuaded it to rain. `unacceptable`

Linguistic acceptability datasets

corpus	language	n. sents. (k)	acceptable sents. (%)
CoLA (Warstadt, Singh, and Bowman, 2018)	English	10.6	70.5
ItaCoLA (Trotta et al., 2021)	Italian	9.7	85.4
DaLAJ (Volodina, Mohammed, and Klezl, 2021)	Swedish	9.5	50
RuCoLA (Mikhailov et al. 2022)	Russian	13.4	71.8
NoCoLA (Jentoft and Samuel, 2023)	Norwegian	14.4	31.5
JCoLA (Someya, Sugimoto, and Oseki, 2024)	Japanese	10	82
HuCoLA (Ligeti-Nagy et al., 2024)	Hungarian	9.9	78
EsCoLA (Bel, Punsola and Ruiz-Fernández, 2024)	Spanish	11.1	70

Linguistic acceptability datasets

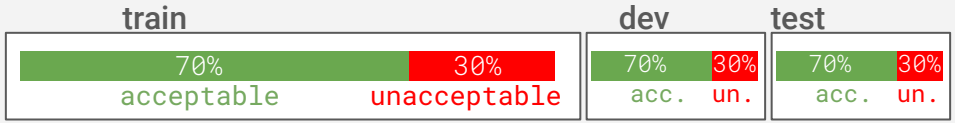
corpus	language	n. sents. (k)	acceptable sents. (%)
CoLA (Warstadt, Singh, and Bowman, 2018)	English	10.6	70.5
ItaCoLA (Trotta et al., 2021)	Italian	9.7	85.4
DaLAJ (Volodina, Mohammed, and Klezl, 2021)	Swedish	9.5	50
RuCoLA (Mikhailov et al. 2022)	Russian	13.4	71.8
NoCoLA (Jentoft and Samuel, 2023)	Norwegian	14.4	31.5
JCoLA (Someya, Sugimoto, and Oseki, 2024)	Japanese	10	82
HuCoLA (Ligeti-Nagy et al., 2024)	Hungarian	9.9	78
EsCoLA (Bel, Punsola and Ruiz-Fernández, 2024)	Spanish	11.1	70
CatCoLA	Catalan	10.4	70

Building the Catalan Corpus of Linguistic Acceptability

CatCoLA

In-domain

10,189 sents.



Sources:

- *Gramàtica del català contemporani* (Solà and Rigau, 2002)
- Catalan course for foreign learners (CPNL)
- Linguistics articles

Out-of-domain

254 sents.

Sources:

- ParlaMint (Pisani, Zevallos, and Bel, 2023)

Building the Catalan Corpus of Linguistic Acceptability

Annotation of linguistic phenomena:

1. Simple
2. Predicative
3. Adjuncts
4. Argument types
5. Argument alternations
6. Binding pronouns
7. Wh-phenomena
8. Complement clauses
9. Auxiliary and modal verbs, polarity, PV constructions
10. Infinitive embedded VPs, referential phenomena
11. Complex NPs and APs
12. S-syntax phenomena
13. Determiners, quantifiers, partitives, comparatives

14. Catalan-specific phenomena
    - Agreement in nominal constructions
    - Ellipsis
    - *Ser/estar* copula selection
    - *Hi* unvoiced pronouns
    - Cliticization phenomena
    - Subjunctive mode
    - Dislocations

(CoLA, Warstadt, Singh, and Bowman, 2018)

Building the Catalan Corpus of Linguistic Acceptability

	(1a)	(1b)
CA	*L'Anna estudia grec.*	**L'euga ha arribat esgotats.*
ES	*Anna estudia griego.*	**La yegua ha llegado agotados.*
EN	*Anna studies Greek.*	**The mare has arrived exhausted.*
		(f. sg.) (m. pl.)
	acceptable	unacceptable
	1. Simple	14. Catalan-specific phenomena
		Agreement in nominal constructions

Linguistic acceptability task: experiments

Human evaluation

Annotation of the
in-domain set by
3 Catalan native linguists.

Model evaluation

Fine-tuning on the in-domain train set

Catalan RoBERTa-v2 (Armengol-Estapé et al., 2021)
XLM-RoBERTa (Conneau et al., 2020)

Linguistic acceptability task: results

Human evaluation

annotator	MCC
A1	0.83
A2	0.61
A3	0.64

avg. **MCC=0.69**

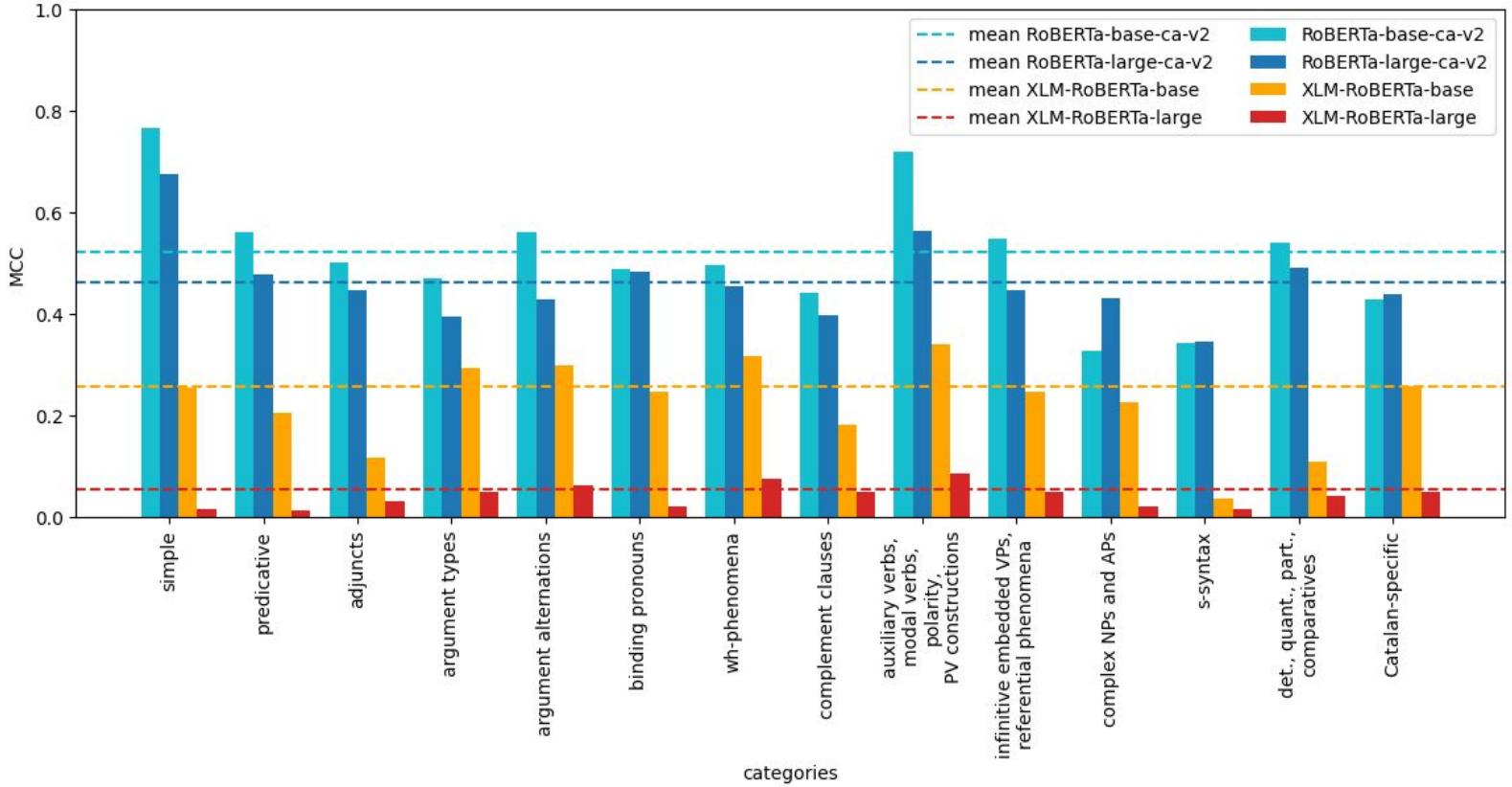
Model evaluation

Fine-tuning on in-domain train set

model	in-domain	out-of-domain
	MCC	MCC
RoBERTa-**base**-ca-v2	0.55 (*0.52*)	0.40
RoBERTa-**large**-ca-v2	0.62 (*0.46*)	0.59
XLM-RoBERTa-**base**	0.35 (*0.26*)	0.10
XLM-RoBERTa-**large**	0.30 (*0.05*)	0.11

best, *avg. of 10 runs*

Linguistic acceptability task: results



CatCoLA:
Catalan Corpus of Linguistic Acceptability

Núria Bel[1], Marta Punsola[1], Valle Ruiz-Fernández[2]
[1]Universitat Pompeu Fabra
[2]Barcelona Supercomputing Center

This research is part of the **LUTEST** project, PID2019-104512GB-I00, a UPF project funded by the MICIU: Ministerio de Ciencia, Innovación y Universidades and Agencia Estatal de Investigación, Spain (10.13039/501100011033).

BSC participation has been promoted and financed by the Government of Catalonia through the **Aina** project and by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project **ILENIA**, with reference 2022/TL22/00215337.