Comprehensive Review of Under-Resourced Languages: How Much Progress Have We Made in Catalan?

Patricia Morales-Hurtado[1,*]

[1]Universitat Oberta de Catalunya (UOC), Rambla del Poblenou 154-156 08018 Barcelona, Spain

Abstract

This article presents a comprehensive review of recent advances in natural language processing (NLP) applied to Catalan as an under-resourced language. Using a methodology based on Boolean searches across multiple databases (between 2020 and 2026), 35 relevant publications were identified. The results show significant progress, particularly in the creation of both monolingual and multilingual datasets, as well as in the development of terminology resources, language models, evaluation benchmarks and other applications. Initiatives such as the AINA Project, aimed at building linguistic and technological infrastructures in Catalan, are also highlighted. Although limitations persist, such as a scarcity of resources and indexing problems, Catalan shows an increasing development in various tasks. It concludes that there is a growing scientific interest in Catalan, and proposes some future lines of research.

Keywords

Catalan, natural language processing, under-resourced languages, comprehensive review

1. Introduction

In recent years, natural language processing (NLP) has made huge strides with the introduction of new technologies and innovative methodologies, such as neural networks and large language models (LLMs). These contributions are often compiled in systematic reviews published regularly in journals and at conferences [1, 2]. In this way, scholars can keep updated on the latest developments in the field and build upon them.

The language most frequently studied within NLP is English, as it is spoken by approximately 1.5 billion people, as recorded in [3], and has become *lingua franca* in science; therefore, it is the one most used as a language of study in research. This means that other languages (which may or may not have a large number of speakers) are under-represented within the domain and do not evolve to meet all research needs in linguistic contexts with limited resources, such as Catalan, Arabic, or Hindi. Although there is a section of the scientific community attempting to address this shortfall [4], there is still much to discover, analyse, create, and explore for minority languages in natural language processing.

On the one hand, another factor that must be taken into account is the lack of datasets, syntactic analysers (parsers), and other similar resources for these languages. In natural language processing, such resources are essential for the successful development of terminology extractors, sentiment analysis, corpus creation, machine translation, and so on. Some difficulties encountered when carrying out these projects include semantic ambiguity, syntactic complexity and dialectal variation [5]. As there are few research groups working in these languages, they are still in the early stages of development in this field.

Furthermore, there is no record of any comprehensive review of advances in natural language processing published exclusively in Catalan. This is not to mean merely that one has been written in Catalan, but rather that it discusses the specific contributions made using Catalan as the language of study. Such reviews were found for other languages, as in [6, 5], but in comparison with English, it is concluded that there is a need to support and encourage more research groups to focus on languages with limited resources as well.

With the aim of continuing along this path, the purpose of this article is to compile the latest developments in NLP using Catalan as the language of study. It should be noted that in this paper we will not be assessing neither the quality nor the performance of the publications analysed. Furthermore, there is a desire to highlight those researchers who are dedicated to promoting Catalan as a language of interest for science.

The article is divided into several sections. In Section 2, we explain the methodology used to gather information on developments in NLP in Catalan. Then, the contribution of the publications are analysed in Section 3 and the results are interpreted in Section 4. Finally, conclusions and outline possible lines of future work are presented.

2. Methodology

In this section, the methodology used to compile the material for analysis is described. Initially, the focus was solely on terminological extraction with Catalan as the language of study or one of the languages of study. However, upon realising that no sufficient results were obtained to carry out the study (in many cases, there were none in the databases we consulted), it was decided to broaden the search to include all research related to NLP in Catalan.

In order to conduct an exhaustive search for publications, the approach was based on the method presented in [7], which uses the Boolean method to find all texts that include "Catalan" in the keywords or in the body of the text, along with one or more of the following concepts: "terminology extraction", "automatic terminology extraction", "automated terminology extraction" or "natural language processing".

Searches were conducted in Catalan, Spanish and English, as these are the languages the author is proficient in and to cover a wider range of sources from which documentation can be extracted. The result is as follows (we have chosen the example written in English, as this is the language of this article):

> "Catalan" AND {"corpora" OR "automatic terminology extraction" OR "generative modelling" OR "transfer learning" OR "language model" OR "machine translation" OR "sentiment analysis" OR "text classification" OR "automatic extraction"}

SEPLN 2026: 42nd International Conference of the Spanish Society for Natural Language Processing, León, Spain, 22-25 September 2026.

*Corresponding author.

pmoraleshu@uoc.edu (P. Morales-Hurtado)

0009-0009-2527-516X (P. Morales-Hurtado)

Table 1
Number of publications sorted by source and language

Source	Number of publications ca	es	en
ACL Anthology	–	0	12
ACM Digital Library	–	–	2
Arxiv	–	–	2
IEEE Xplore Digital Library	–	0	0
Information	–	–	1
Linguamática	1	0	0
Natural Language Engineering	–	–	1
ScienceDirect	–	0	1
Scientific Data	–	–	1
Scopus	–	0	0
Semantic Scholar	0	0	0
Sociedad Española para el Procesamiento del Lenguaje Natural	–	2	7
Springer Nature	–	–	4
Terminàlia	1	0	0
Terminology	–	–	0
Terminology Science and Research	–	–	0
Tradumàtica	0	0	0
Web of Science	–	–	0
Total of publications by language	**2**	**2**	**31**
Total of publications		**35**	

To narrow down the results, ensure the highest possible reliability and ensure they are currently relevant, the exclusion criteria listed below were applied from the outset:

- They must have been published between 2020 and 2026;
- They must be journal articles, book chapters, conference papers and workshop papers, and
- They must be open-access.

Finally, the journals and databases selected were the most representative of the field in the three languages. For the journals, we used quartile (Q1 or Q2) as our metric. Below these quartiles, Catalan journals were added (namely, *Tradumàtica*, *Linguamática* and *Terminàlia*), as the author also wanted to include publications written in Catalan. Table 1 shows the number of publications broken down by source and language, as well as the total by language and overall. Where a language is not available in the search, the symbol – has been used.

3. Analysis of the publications

In this section, we describe the publications identified following a detailed search in accordance with the methodology. For clarity and organisation, we have categorised them under the following headings: terminology repositories, datasets, language models, evaluation, and other applications.

3.1. Terminology repositories

In the field of terminology repositories, we highlight the contributions of [8, 9]. [8] describes MHeTRep, a multilingual corpus in the medical field. Semantic tagging (provided by the main categories of SNOMED-CT) was used for term extraction, and, as a source resource for Catalan, it is based on the translation of the SNOMED-CT dataset. Furthermore, the extraction results were evaluated in two scenarios: intrinsic, in which the relevance of the category is examined according to its frequency across the different datasets, and extrinsic, in which the coverage of other annotated corpora is compared with MHeTRep. In the article, the authors specify that access to the repository will be public, but there is no further information as of today.

In [9], Vàzquez explains the creation of the open-acces digital dictionary *Terminologia de IATE en català* [Terminology of IATE in Catalan] in collaboration with the TERMCAT terminology centre, which currently covers the following subject areas: law, economics and health sciences. IATE (Interactive Terminology for Europe) database, Terminologia Oberta [Open Terminology] portal (published by TERMCAT), Wikipedia and *Diari Oficial de la Generalitat de Catalunya* (DOGC) [Official Gazette of the Government of Catalonia] were used as extraction corpora. The TBXTools and Moses tools were used for terminological extraction and the search for equivalents in parallel corpora. The result is an open-access dictionary containing equivalents in Catalan, Spanish, English and French, together with the corresponding IATE entry code.

3.2. Datasets

Focusing on the datasets, we find the following articles:

- [10] provides 101 datasets, of which 13 are in Catalan or include it among the languages in the datasets;
- [11] details the datasets in Spanish (with some Latin American variants) and languages of the Iberian Peninsula that they have compiled up to the date of publication, including both their own and external contributions that still need to be evaluated;
- [12] highlights the annotation criteria for negation (covering the marker, focus, and scope), particularly gradation, the use of the adverb pas in Catalan, and cases where negation is not expressed even when the negative particle is present;
- [13] is a multilingual corpus covering various domains, whose main feature is the use of syntactic representations based on dependencies and constituents;
- [14] creates a parallel corpus of human-written texts and machine translations to detect the use of artificial intelligence in text generation;
- [15] is the first Catalan dataset to evaluate the linguistic acceptability of large language models;
- Although [16] does not create a new dataset, it uses TeCla [17] to test the effectiveness of models specialised in lexical simplification and complexity prediction;
- [18] extends the CrowS_Pairs dataset to analyse biases in corpora written in Catalan;
- [19] constructs a corpus from news articles with the aim of analysing automatic summaries generated by LLMs;

- [20] compiles the datasets created and under development within the context of the AINA Project;
- [17] explains the creation of three datasets: CaText (based on the DOGC, Catalan General Crawling, Catalan Government Crawling and the Agència Catalana de Notícies [Catalan News Agency] corpus); XQUAD-ca and ViquiQuAD (based on Wikipedia articles), and TeCla (based on news articles from the Agència Catalana de Notícies);
- [21] describes the challenge held at IberLEF 2024, which aimed to investigate the detection and attribution of texts generated by artificial intelligence compared to those written by humans, and
- [22] is a corpus compiled from messages on social media platform X (Twitter at the time of publication) that determines the user's stance on the Catalan independence movement;
- [23] creates six datasets for two tasks: CaSum and ViLaSum for summarisation, and WMT2013-ca, taCon, Cyber-ca and gEnCaTa for machine translation;
- [24] describes CARMEN-I, a corpus of anonymized clinical records from the Hospital Clinic of Barcelona in Catalan and Spanish;
- [25] releases a synthetic paraphrase dataset in 17 languages, created with monolingual data and neural machine translation.

Tables 2 and 3 provide details of the type of task, the name of the dataset and its authors (wherever possible, the original publication is provided); however, in this article we have selected only those in which Catalan is one of the languages or the language of the dataset. Except Parallel Trees and Phrases, all datasets are open-access.

3.3. Language models

In this section, language models found while performing the search will be explained:

- [17] have developed a BERTa model for their study on multilingual models in low-resource languages. It is built on a RoBERTa base model (110 million parameters). Its pre-training objective is solely masked language modelling, unlike the original BERT, which also uses the auxiliary task of predicting the next sentence. In this paper, they conclude that creating monolingual models for minority languages is more effective than other current multilingual models.
- [37] rely on BERTa, amongst other models including Catalan and Galician, to create two Transformers for both languages. Their task is to restore capitalisation and punctuation in the results of automatic speech recognition systems. They are also capable of identifying proper nouns, country names and organisations for the capitalisation restoration process. The results obtained are 90.2 % for Galician and 90.86 % for Catalan. Links are provided to access the resource.
- [23] develops a Catalan BART-base, using the Catalan Textual Corpus for training. It is created to test the summarisation datasets explained in the article (CaSum and VilaSum). It shows a good performance compared to NASCA and mBART. The models and datasets can be accessed via links in the text.
- [38] introduce LOLA, a multilingual large language model trained on more than 160 languages using a sparse Mixture-of-Experts Transformer architecture. It is based on a GPT-style, evaluated against 17 other models in four tasks (QA, reasoning, NLI, and reading comprehension), and demonstrates significant performance improvements compared to the models analysed in the paper.
- [39] show an innovative fully automated pipeline for creating language-specific BERT models from Wikipedia data, WikiBERT. It outperforms multilingual BERTs, therefore, proposes the creation of monolingual models instead.

All language models detailed in this sections are open-access.

3.4. Evaluation

With regard to the evaluation of models in Catalan, [17] have developed the Catalan Language Understanding Benchmark (CLUB). It consists of eight tasks: morphosyntactic tagging (POS), named entity recognition and classification (NERC), semantic text similarity (STS), text classification (TC), and monolingual and multilingual question answering (QA). All were supervised by qualified annotators—that is, translators, philologists, editors or linguists—with experience in language-related tasks and who are native speakers of Catalan. In doing so, the authors provide benchmarks that serve to test and evaluate the capabilities of monolingual and multilingual models in Catalan. The repository can be accessed in the link provided in the article.

3.5. Other applications

In this section, we outline other applications described in the analysed publications. To organise them, we have divided them into the following subcategories:

- **Sentiment analysis**. [40] uses a transition system based on dependency graph parsing to create sentiment graphs. In their evaluation, the model they propose outperforms other graph-based models and does so more quickly. Moreover, [41] make a systematic review of cross-lingual sentiment analysis (CLSA) and explain new perspectives for the field.
- **Text classification**. The use of an approach based on textual entailment is effective for languages such as Catalan, which are characterised by having few datasets [42]. [43] create a model, with an automatic prediction rate of 98 %, that distinguishes between texts in Catalan and Valencian. Furthermore, [44] present ContextMEL, a system that generates models to detect three contextual modifiers: temporality, negation and certainty. [42, 43] have links to their respective resources.
- **Automatic extraction**. [45] describe WikiTailor, an open-access system that extracts corpora from Wikipedia articles. To do this, it explores the website's category graph and performs a breadth-first search of the categories associated with the domain. Meanwhile, [46] presents a new methodology for extracting discourse markers by calculating their information content and grouping them into functional categories.

Table 2
Monolingual datasets organised by task type, name and authors. As mentioned in this section, [16] do not create a new dataset, but instead use TeCla [17]. For this reason, the symbol * has been used

Task	Name of dataset	Authors
Entity Identification and Linking	Catalan Entity Identification and Linking (CEIL)	[20]
Ethics	InToxiCat	[20]
Math Reasoning	MGSM_ca	[11]
Natural language understanding	NLUCat	[20]
Negation	NoNiRes	[12]
Question answering	Catalanqa	[20]
	Coqcat	[20]
	Piqa_ca	[11]
	Siqa_ca	[11]
	VilaQuAD	[17]
	ViquiQuAD	[17]
	XQUAD-ca	[17]
Paraphrasing	Parafraseja	[10, 11]
Pre-training	CaText	[17]
Sentiment analysis	Catalan Structured Sentiment Analysis (CaSSA)	[20]
Simplification and prediction of lexical complexity	*	[16]
Stance and emotion detection	CaSERa	[20, 26]
	CaSET	[20, 26]
	Catalonia_independence	[22]
Summarization	CaBreu	[20]
	CaSum	[23]
	DACSA	[19]
	VilaSum	[23]
Task-orientated conversation	XitXat	[20]
Text classification	TeCla	[17]
Textual entailment	TE-ca	[17]
	WNLI_ca	[20]

- **Machine translation**. On the one hand, [47] test the effectiveness of models trained from scratch and pre-trained models in Catalan-Chinese translation. Using their PTAMT method (translation via a pivot language, in this case Spanish, with multilingual training) and neural machine translation, they improve the quality and increase the lexical diversity of the final output. On the other hand, [48] evaluates how six translation systems handle anaphoric elements: two rule-based, two neural, and two GPT models. Although all have struggled with the task, the GPTs, specifically Gemini, have achieved better scores compared to the rest. Also, [49] explores the possibility of neural machine translation with expert-in-the-loop validation and confirms the enhancement of named entity recognition (NER) systems for under-resource languages, such as Catalan. Its generated corpora and components are publicly accessible.
- **Automatic detection**. [50] establish a framework for detecting messages in Catalan on X (Twitter) and build an automated tool based on a lexical and morphological corpus, whilst [51] propose Relational Embeddings, a method for representing interaction data. Thanks to their behaviour as dense graphs, they have managed to reduce dispersion and integrate stance-related information from different sources.
- **Coreference resolution**. In [52], experiments are conducted using the CorefUD dataset to assess cross-lingual transferability. They conclude that harmonised annotation is extremely useful in smaller training corpora, and that joint multilingual models perform better than monolingual ones.
- **Named Entity Recognition**. [53] introduces NERCat, a fine-tuned version of the GLiNER model, to improve performance of NER in Catalan. The article demonstrates the enhanced applicability of fine-tuned models for low-resource languages.
- **Transfer learning**. [54] detail a new training strategy, Token-Level Filter Training. Its main purpose is effectively filter out unstable token predictions from

Table 3
Multilingual datasets organised by type of task, name (with their ISO language code) and authors

Task	Name of dataset (ISO language code)	Authors
Adaptation	Phrases (ca-CA, ca-VA, es)	[11] [mentioned in the article, but without a bibliographical reference]
Anonymisation	CARMEN-I (ca, es)	[24]
Common-sense reasoning	COPA (ca, es, eu)	[10, 11, 20]
Ethics	CrowS-Pairs (ar, ca, de, en, es, fr, it, mt, zh)	[18]
Linguistic acceptability	CoLA (ca, es, gl)	[15, 27, 28]
Machine-generated text detection	IberAuTexTification (ca, en, es, eu, gl, pt)	[21]
Machine-generated text model attribution	IberAuTexTification (ca, en, es, eu, gl, pt)	[21]
	IBERMAT (ca, es, eu, gl)	[14]
Machine translation	Cyber-ca (ca, es, en)	[23]
	FLORES (196 languages)	[29, 30]
	gEnCaTa (ca, en)	[23]
	taCon (ca, en, es, eu, gl)	[23]
	WMT2013-ca (ca, es)	[23]
Mathematics	mgsm_direct (ca, es, eu, gl)	[11]
Question answering	ARC (ca, eu)	[10, 11]
	OpenBookQA (ca, es)	[10, 11]
	XStoryCloze (ca, gl, pt)	[10, 11]
Paraphrasing	Paracotta (ar, ca, cs, de, en, es, et, fr, hi, id, it, nl, ro, ru, sv, vi, zh)	[25]
	PAWS (ca, es, gl, pt)	[20, 28, 31]
Reading comprehension	Belebele (122 languages)	[32]
Stance detection	MultiStanceCat (ca, es)	[33]
Syntactic tree annotation	Parallel Trees (ca, es, fr, id, is, it, nl, pt, tr, vi)	[13]
Textual entailment	XNLI (ca, es, gl)	[20, 28, 34]
Truthfulness	Truthfulqa (ca, en, es, eu, gl)	[35]
	Veritasqa (ca, es, eu)	[36]

the parent model. They also introduce a hierarchical ranking loss method to help the child model better learn the parent's strategies.

4. Discussion

In this section, an interpretation is provided for the resources presented in the publications. There is a significant effort on the part of research groups to cover all areas within the discipline and to steer it towards a more inclusive and multilingual future. It was observed that the most significant progress in the data analysed lies in the creation of datasets, whether through the translation or expansion into Catalan of existing datasets or through the creation of new monolingual corpora in Catalan. Likewise, other tasks are not lagging behind when compared with majority languages. By this, the intention is to show that Catalan is not at an earlier stage in terms of technological development than other, more widely studied languages.

In addition, the author wants to highlight the efforts put in [20, 55]. These publications outline the objectives and progress made within the AINA project. The aim is to promote the Catalan language in the digital age through artificial intelligence (AI) and language technologies (LT). Its main goal is to provide the language with an infrastructure for the development of AI- and LT-based applications, such as voice assistants, machine translators, etc. It seeks to ensure that the inclusion of Catalan in such applications is profitable for local and global companies in the sector. Also, Catalan speakers can exercise their right to use their mother tongue in the digital world, and it promotes research in this field. As a result, they have published annotated datasets (some of which are mentioned in this article, such as CaText and InToxiCat), training corpora for generative models, three trilingual GPT-3 models, and Common Voice speech technologies. They have also developed machine translation models with lines of research into unsupervised training techniques.

It should be noted that the main limitation we encountered when searching for publications for this study was the indexing in certain databases. Although the Boolean search specified that the text must contain the words described in the methodology, it often returned citations as

valid where one of the authors' surnames was *Catalan* or similar. For this reason, a more effective way of filtering out such results should be found; for example, by adding a function that determines the minimum number of times a term must appear within the text. It is also clear that our search has been complicated by the scarcity of publications and resources that use Catalan as their language of study. Moreover, as one of the criteria was that publications had to be open-access, some advances were not included in this article. In future work, all publications will be considered in order to have a broader and more accurate picture of SOTA advances. We would like to encourage increased research into minority languages in order to democratise advances in NLP across all languages, regardless of their presence in academia, and specifically natural language processing.

5. Conclusions and future work

Natural language processing is a discipline that has advanced and evolved over the decades. With the development of new technologies, it has been possible to achieve unprecedented results, prompting us to reflect on the capabilities of cutting-edge systems and their multidisciplinary applications within NLP, such as artificial intelligence. Even so, these advances are largely monopolised by the most widely studied languages (English, Spanish, Chinese, etc.), whilst minority languages in the field are relegated to the background. As might be expected, publications in these languages are more difficult to access, either because they are scarce or because they do not appear in international academic journals; consequently, it is impossible to know for certain what progress has been made.

In this article, we have compiled 35 papers that use Catalan as one of their languages of study and which have been published over the last six years. Using Boolean search criteria, we have examined relevant and reliable journals and databases (in Catalan, Spanish and English) within the sector, and we have outlined the contributions detailed therein without going into assessments of quality or performance. We have divided the sections into terminology repositories, datasets, language models, evaluation, and other applications. From the data analysed, it appears that the scientific community is interested in the application of NLP to Catalan; however, we cannot extrapolate this to other minority languages, as we have only analysed one. As of today and as far as the author is aware, it is the first comprehensive review conducted for the compilation of research on Catalan within natural language processing.

Finally, future research could focus on the effectiveness and reliability of large language models in under-resourced languages, with the potential development of specific LLMs for these languages. This is a relatively new field with considerable scope for exploration. Efforts could also continue to expand the capabilities of multilingual neural networks to incorporate more minority languages. It would also be interesting to create more annotated datasets in other fields of knowledge to lay the foundations for these languages; to conduct research into machine translation, not only between majority and minority languages but also between minority languages themselves; and to improve speech recognition and transcription to increase their use and dissemination among the general public. In future research, the author will apply the insights gathered in this article to compare terminology extractors based on traditional approaches with those using artificial intelligence.

Acknowledgments

This work was supported by the Industrial Doctorates Plan from the Department of Research and Universities of the Government of Catalonia (reference number: 2025 DI 00112).

References

[1] G. M. di Nunzio, S. Marchesin, G. Silvello, A systematic review of Automatic Term Extraction: What happened in 2022?, Digital Scholarship in the Humanities 38 (2023) i41–i47. doi:10.1093/llc/fqad030.

[2] H. T.-H. Tran, M. Martinc, J. Caporusso, J. Delaunay, A. Doucet, S. Pollak, Recent Advances in Automatic Term Extraction: A Comprehensive Survey, ACM Computing Surveys 58 (2026) 1–35. doi:10.1145/3787584.

[3] D. M. Eberhard, G. F. Simons, C. D. Fennig, Ethnologue: Languages of Africa and Europe, 26th ed., SIL International, Global Publishing, 2023.

[4] A. Al-Thubaity, A Novel Dataset for Arabic Domain Specific Term Extraction and Comparative Evaluation of BERT-Based Models for Arabic Term Extraction, ACM Transactions on Asian and Low-Resource Language Information Processing 24 (2025). doi:10.1145/3748323.

[5] H. Avetisyan, D. Broneske, Large Language Models and Low-Resource Languages: An Examination of Armenian NLP, in: J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, A. A. Krisnadhi (Eds.), Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings), Association for Computational Linguistics, Nusa Dua, Bali, 2023, pp. 199–210. doi:10.18653/v1/2023.findings-ijcnlp.18.

[6] P. Pakray, A. Gelbukh, S. Bandyopadhyay, Natural language processing applications for low-resource languages, Natural Language Processing 31 (2025) 183–197. doi:10.1017/nlp.2024.33.

[7] J. C. Blandón Andrade, C. M. Medina Otálvaro, C. M. Zapata Jaramillo, A. Morales Ríos, Approaches, Tools, Algorithms, and Methods for Automatic Term Extraction: A Systematic Literature Mapping, Journal of Intelligent and Fuzzy Systems (2025). doi:10.1177/18758967251392652.

[8] J. Vivaldi, H. Rodríguez, MHeTRep: A multilingual semantically tagged health terms repository, Natural Language Engineering 29 (2023) 1364–1401. doi:10.1017/S1351324922000055.

[9] M. Vàzquez, Creating Terminological Resources in the Digital Age for Less-resourced Languages, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italy, 2024, pp. 4088–4091.

[10] J. Á. González, I. Borrego Obrador, Á. Romo Herrero, A. M. Sarvazyan, M. Chinea-Ríos, A. Basile, M. Franco-Salvador, IberBench: LLM evaluation on Iberian languages, Computer Speech and Language 96 (2026). doi:10.1016/j.csl.2025.101899.

[11] M. Grandury, J. Aula-Blasco, J. Falcão, C. Fourrier, M. González, G. Martínez, G. Santamaría, R. Agerri, N. Aldama, L. Chiruzzo, J. Conde, H. Gómez, M. Guerrero, G. Ivetta, N. López, F. M. Plaza-del-Arco, M. T. Martín-Valdivia, H. Montoro, C. Muñoz, P. Reviriego, L. Rosado, A. Vaca, M. E. Vallecillo-Rodríguez, J. Vallego, I. Zubiaga, La Leaderboard: A Large Language Model Leaderboard for Spanish Varieties and Languages of Spain and Latin America, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 32482–32524. doi:10.18653/v1/2025.acl-long.1561.

[12] L. Tañá Velasco, M. Nofre Maiz, B. Calvo Figueras, C. Armentano-Oller, NoNiRes: corpus del catalán anotado con negación, Sociedad Española para el Procesamiento del Lenguaje Natural (2023) 39–51. doi:10.26342/2023-71-3.

[13] C. Alzetta, A. Miaschi, F. Dell'Orletta, G. Venturi, S. Montemagni, Parallel Trees: A novel resource with aligned dependency and constituency syntactic representations, Language Resources and Evaluation 59 (2025) 3445–3485. doi:10.1007/s10579-025-09826-3.

[14] A. Picazo-Izquierdo, E. L. Estevanell-Valladares, R. Mitkov, R. Muñoz Guillena, M. Palomar, IBERMAT - Corpus of Human and Machine Translated Multi-Domain Content in Basque, Catalan, Galician and Spanish: Description and Exploitation, Sociedad Española para el Procesamiento del Lenguaje Natural (2025) 337–348. doi:10.26342/2025-75-24.

[15] N. Bel, M. Punsola, V. Ruiz-Fernández, CatCoLA, Catalan Corpus of Linguistic Acceptability, Sociedad Española para el Procesamiento del Lenguaje Natural (2024) 177–190. doi:10.34810/data1393.

[16] H. Saggion, S. Bott, S. Szasz, N. Pérez, S. Calderón, M. Solís, Lexical complexity prediction and lexical simplification for Catalan and Spanish: Resource creation, quality assessment, and ethical considerations, in: M. Shardlow, H. Saggion, F. Alva-Manchego, M. Zampieri, K. North, S. Štajner, R. Stodden (Eds.), Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024), Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 82–94. URL: https://aclanthology.org/2024.tsar-1.9/. doi:10.18653/v1/2024.tsar-1.9.

[17] J. Armengol-Estapé, C. P. Carrino, C. Rodriguez-Penagos, O. de Gibert Bonet, C. Armentano-Oller, A. González-Agirre, M. Melero, M. Villegas, Are Multilingual Models the Best Choice for Moderately Under-Resourced Languages? A Comprehensive Assessment for Catalan, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, online, 2021, pp. 4933–4946. doi:10.18653/v1/2021.findings-acl.437.

[18] K. Fort, L. Alonso Alemany, L. Benotti, J. Bezançon, C. Borg, M. Borg, Y. Chen, F. Ducel, Y. Dupont, G. Ivetta, Z. Li, M. Mieskes, M. Naguib, Y. Qian, M. Radaelli, W. S. Schmeisser-Nieto, E. Raimundo Schulz, T. Saci, S. Saidi, J. Torroba Marchante, S. Xie, S. E. Zanotto, A. Névéol, Your stereotypical mileage may vary: Practical challenges of evaluating biases in multiple languages and cultural contexts, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italy, 2024, pp. 17764–17769. URL: https://aclanthology.org/2024.lrec-main.1545/.

[19] E. Segarra, V. Ahuir, L.-F. Hurtado, J. Á. González, DACSA: A large-scale Dataset for Automatic summarization of Catalan and Spanish newspaper Articles, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, USA, 2022, pp. 5931–5943. doi:10.18653/v1/2022.naacl-main.434.

[20] A. González-Agirre, M. Marimon, C. Rodríguez-Penagós, J. Aula-Blasco, I. Baucells, C. Armentano-Oller, J. Palomar-Giner, B. Kulebi, M. Villegas, Building a Data Infrastructure for a Mid-Resource Language: The Case of Catalan, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italy, 2024, pp. 2556–2566.

[21] A. M. Sarvazyan, J. Á. González, F. Rangel, P. Rosso, M. Franco-Salvador, Overview of IberAuTexTification at IberLEF 2024: Detection and Attribution of Machine-Generated Text on Languages of the Iberian Peninsula, Sociedad Española para el Procesamiento del Lenguaje Natural (2024) 421–434. doi:10.26342/2024-73-32.

[22] E. Zotova, R. Agerri, M. Nuñez, G. Rigau, Multilingual Stance Detection: The Catalonia Independence Corpus, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, J. Mariani, H. Mazo, A. Moreno, J. Okijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 1368–1375.

[23] O. de Gibert, K. Kharitonova, B. Calvo Figueras, J. Armengol-Estapé, M. Melero, Sequence-to-Sequence Resources for Catalan (2022). doi:10.48550/ARXIV.2202.06871.

[24] S. Lima-López, E. Farré-Maduell, L. Gasco, J. Rodríguez-Miret, S. Frid, X. Pastor, X. Borrat, M. Krallinger, A textual dataset of de-identified health records in Spanish and Catalan for medical entity recognition and anonymization, Scientific Data 12 (2025) 1088. doi:10.1038/s41597-025-05320-1.

[25] A. F. Aji, T. N. Fatyanosa, R. E. Prasojo, P. Arthur, S. Fitriany, S. Qonitah, N. Zulfa, T. Santoso, M. Data, ParaCotta: Synthetic multilingual paraphrase corpora from the most diverse translation sample pair, in: K. Hu, J.-B. Kim, C. Zong, E. Chersoni (Eds.), Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation, Association for Computational Lingustics, Shanghai, China, 2021, pp. 533–542. URL: https://aclanthology.org/2021.paclic-1.56/.

[26] B. Figueras, I. Baucells, T. Caselli, Dynamic Stance: Modeling Discussions by Labeling the Interactions, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023,

Association for Computational Linguistics, Singapore, Singapore, 2023, pp. 6503–6515. doi:10.18653/v1/2023.findings-emnlp.432.

[27] N. Bel, M. Punsola, V. Ruiz-Fernández, EsCoLA: Spanish Corpus of Linguistic Acceptability, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italy, 2024, pp. 6268–6277.

[28] A. I. Vladu, I. de-Dios-Flores, C. Magariños, J. E. Ortega, J. R. Pichel, M. Garcia, P. Gamallo, E. Fernández Rei, A. Bugarín, M. González González, S. Barro, X. L. Regueira, Proxecto nós: Artificial intelligence at the service of the galician language, in: Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing: Projects and Demonstrations, volume 3224, A Coruña, Spain, 2022, pp. 26–30.

[29] F. Guzmán, P.-J. Chen, M. Ott, J. Pino, G. Lample, P. Koehn, V. Chaudhary, M. Ranzato, The FLORES Evaluation Datasets for Low-Resource Machine Translation: Nepali–English and Sinhala–English, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 6098–6111. doi:10.18653/v1/D19-1632.

[30] N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, A. Fan, The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation, in: B. Roark, A. Nenkova (Eds.), Transactions of the Association for Computational Linguistics, volume 10, MIT Press, Cambridge, Massachusetts, USA, 2022, pp. 522–538. doi:10.1162/tacl_a_00474.

[31] Y. Yang, Y. Zhang, C. Tar, J. Baldridge, PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3687–3692. doi:10.18653/v1/D19-1382.

[32] L. Bandarkar, D. Liang, B. Muller, M. Artetxe, S. N. Shukla, D. Husa, N. Goyal, A. Krishnan, L. Zettlemoyer, M. Khabsa, The Belebele Benchmark: A Parallel Reading Comprehension Dataset in 122 Language Variants, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 749–775. doi:10.18653/v1/2024.acl-long.44.

[33] M. Taulé, F. Rangel, M. A. Martí, P. Rosso, Overview of the task on multimodal stance detection in tweets on catalan# 1oct referendum, in: P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, J. Carrillo-de-Albornoz (Eds.), Proceedings of the Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), Seville, Spain, 2018, pp. 149–166.

[34] A. Conneau, R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk, V. Stoyanov, XNLI: Evaluating Cross-lingual Sentence Representations, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2475–2485. doi:10.18653/v1/D18-1269.

[35] B. Calvo Figueras, E. Sagarzazu, J. Etxaniz, J. Barnes, P. Gamallo, I. de-Dios-Flores, R. Agerri, Truth Knows No Language: Evaluating Truthfulness Beyond English, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 31204–31218. doi:10.18653/v1/2025.acl-long.1507.

[36] J. Aula-Blasco, J. Falcão, S. Sotelo, S. Paniagua, A. González-Agirre, M. Villegas, VeritasQA: A Truthfulness Benchmark Aimed at Multilingual Transferability, in: O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. di Eugenio, S. Schockaert (Eds.), Proceedings of the 31st International Conference on Computational Linguistics, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2025, pp. 5463–5474.

[37] R. Pan, J. A. García-Díaz, P. J. Vivancos-Vicente, R. Valencia-García, Evaluation of transformer-based models for punctuation and capitalization restoration in Catalan and Galician, Sociedad Española para el Procesamiento del Lenguaje Natural (2023) 27–38. doi:10.26342/2023-70-2.

[38] N. Srivastava, D. Kuchelev, T. Moteu Ngoli, K. Shetty, M. Roeder, H. Zahera, D. Moussallem, A.-C. Ngonga Ngomo, LOLA – an open-source massively multilingual large language model, in: O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, S. Schockaert (Eds.), Proceedings of the 31st International Conference on Computational Linguistics, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2025, pp. 6420–6446. URL: https://aclanthology.org/2025.coling-main.428/.

[39] S. Pyysalo, J. Kanerva, A. Virtanen, F. Ginter, WikiBERT models: Deep transfer learning for many languages, in: S. Dobnik, L. Øvrelid (Eds.), Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), Linköping University Electronic Press, Sweden, Reykjavik, Iceland (Online), 2021, pp. 1–10. URL: https://aclanthology.org/2021.nodalida-main.1/.

[40] D. Fernández-González, Structured sentiment analysis as transition-based dependency graph parsing, Artificial Intelligence Review 59 (2026). doi:10.1007/s10462-025-11463-9.

[41] C. Zhao, M. Wu, X. Yang, W. Zhang, S. Zhang, S. Wang, D. Li, A Systematic Review of Cross-Lingual Sentiment Analysis: Tasks, Strategies, and Prospects, ACM Computing Surveys 56 (2024) 1–37. doi:10.1145/3645106.

[42] I. Baucells de la Peña, B. Calvo Figueras, M. Villegas, O. López de Lacalle, Entailment-based Task Transfer for Catalan Text Classification in Small Data Regimes, Sociedad Española para el Procesamiento del Lenguaje Natural (2023) 165–177. doi:10.26342/2023-71-13.

[43] R. García Cerdá, M. Miró Maestre, M. Canal, Building a Lightweight Classifier to Distinguish Closely

Related Language Varieties with Limited Supervision: The Case of Catalan vs Valencian, in: E. L. Estevanell-Valladares, A. Picazo-Izquierdo, T. Ranasinghe, B. Mikaberidze, S. Ostermann, D. Gurgurov, P. Mueller, C. Borg, M. Šimko (Eds.), Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages, Incoma Ltd., Shoumen, Bulgaria, Barna, Bulgaria, 2025, pp. 7–11. doi:10.26615/978-954-452-100-4-002.

[44] P. Chocron, A. Abella, G. de Maeztu, ContextMEL: Classifying contextual modifiers in clinical text, Sociedad Española para el Procesamiento del Lenguaje Natural (2020) 45–52. doi:10.26342/2020-65-5.

[45] C. España-Bonet, A. Barrón-Cedeño, L. Màrquez, Tailoring and evaluating the Wikipedia for in-domain comparable corpora extraction, Knowledge and Information Systems 65 (2023) 1365–1397. doi:10.1007/s10115-022-01767-5.

[46] R. Nazar, Inducción automática de una taxonomía multilingüe de marcadores discursivos: primeros resultados en castellano, inglés, francés, alemán y catalán, Sociedad Española para el Procesamiento del Lenguaje Natural (2021) 127–138. doi:10.26342/2021-67-11.

[47] Y. Chen, A. Toral, Z. Li, M. Farrús, Improving NMT from a Low-Resource Source Language: A Use Case from Catalan to Chinese via Spanish, in: C. Scarton, C. Prescott, C. Bayliss, C. Oakley, J. Wright, S. Wrigley, X. Song, E. Gow-Smith, R. Bawden, V. M. Sánchez-Cartagena, P. Cadwell, E. Lapshinova-Koltunski, V. Cabarrão, K. Chatzitheodorou, M. Nurminen, D. Kanojia, H. Moniz (Eds.), Proceedings of the 25th Annual Conference of the European Association for Machine Translation, volume 1, European Association for Machine Translation (EAMT), Sheffield, United Kingdom, 2024, pp. 229–245.

[48] S. Álvarez-Vidal, Resolució anafòrica en traducció automàtica: el cas de l'espanyol i el català, Linguamática 16 (2024) 3–13. doi:10.21814/lm.16.1.424.

[49] J. Rodríguez-Miret, E. Farré-Maduell, S. Lima-López, L. Vigil, V. Briva-Iglesias, M. Krallinger, Exploring the Potential of Neural Machine Translation for Cross-Language Clinical Natural Language Processing (NLP) Resource Generation through Annotation Projection, Information 15 (2024) 585. doi:10.3390/info15100585.

[50] S. Plaza, J. Vilaplana, J. Mateo, J. Rius, F. Solsona, CatDetect, a framework for detecting Catalan tweets, Multimedia Tools and Applications 80 (2021) 10657–10677. doi:10.1007/s11042-020-10182-3.

[51] J. Fernández de Landa, R. Agerri, Language Independent Stance Detection: Social Interaction-based Embeddings and Large Language Models, Sociedad Española para el Procesamiento del Lenguaje Natural (2025) 139–157. doi:10.26342/2025-74-10.

[52] O. Pražák, M. Konopík, J. Sido, Multilingual Coreference Resolution with Harmonized Annotations, in: R. Mitkov, G. Angelova (Eds.), Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Incoma Ltd, online, 2021, pp. 1119–1123. doi:10.26615/978-954-452-072-4_125.

[53] G. Cadevall Ferreres, M. Serrano Sanz, M. Bardeli Gámez, P. Gerdt Basullas, F. Tarres Ruiz, R. Quijada Ferrero, NERCat: Fine-Tuning for Enhanced Named Entity Recognition in Catalan (2025). doi:10.48550/ARXIV.2503.14173.

[54] W. Dai, D. Han, T. Tohti, Y. Liang, Z. Zuo, Y. Liao, Q. Yang, TFT-TL: Token-Level Filter Training Transfer Learning for Low-Resource Neural Machine Translation, ACM Transactions on Asian and Low-Resource Language Information Processing 24 (2025) 1–19. doi:10.1145/3732779.

[55] M. Villegas Montserrat, El projecte AINA, la IA i les tecnologies del llenguatge, Terminàlia (2023) 80–84. doi:10.2436/20.2503.01.189.