Expanding the FLORES+ Multilingual Benchmark with Translations for Aragonese, Aranese, Asturian, and Valencian

Juan Antonio Pérez-Ortiz,[*†] **Felipe Sánchez-Martínez,**[†] **Víctor M. Sánchez-Cartagena,**[†] **Miquel Esplà-Gomis,**[†] **Aarón Galiano-Jiménez,**[†] **Antoni Oliver,**[‡] **Claudi Aventín-Boya,**[‡] **Alejandro Pardos,**[◇] **Cristina Valdés,**[•] **Jusèp Loís Sans Socasau,**[□] **Juan Pablo Martínez**[△]

[†]Universitat d'Alacant {japerez,fsanchez,vm.sanchez,miquel.espla,aaron.galiano}@ua.es
[*]Valencian Graduate School and Research Network of Artificial Intelligence, ValgrAI
[‡]Universitat Oberta de Catalunya {aoliverg,caventinb}@uoc.edu
[◇]Universidad de Zaragoza apardoscalvo@gmail.com
[•]Academia de la Llingua Asturiana / Universidad de Oviedo cris@uniovi.es
[□]Institut d'Estudis Aranesi – Acadèmia Aranesa dera Lengua Occitana sanssocasau@gmail.com
[△]Academia Aragonesa de la Lengua / Universidad de Zaragoza jpmart@unizar.es

Abstract

In this paper, we describe the process of creating the FLORES+ datasets for several Romance languages spoken in Spain, namely Aragonese, Aranese, Asturian, and Valencian. The Aragonese and Aranese datasets are entirely new additions to the FLORES+ multilingual benchmark. An initial version of the Asturian dataset was already available in FLORES+, and our work focused on a thorough revision. Similarly, FLORES+ included a Catalan dataset, which we adapted to the Valencian variety spoken in the Valencian Community. The development of the Aragonese, Aranese, and revised Asturian FLORES+ datasets was undertaken as part of a WMT24 shared task on translation into low-resource languages of Spain.

1 Introduction

Although notable advances have been reported in the realm of machine translation (MT) in recent years, performance for the so-called low-resource languages (Ranathunga et al., 2023) still lags behind that of languages with more extensive linguistic resources. Increasing the availability of such resources is a key factor for enabling MT to progressively and ultimately cover all languages in the world. Several relevant initiatives, such as FLORES-101 (Goyal et al., 2022), FLORES-200 (Costa-jussà et al., 2024), Seed (Maillard et al., 2023), and NTREX (Federmann et al., 2022), have recently addressed this challenge by providing multilingual datasets covering up to 200 languages. While these datasets were originally created as part of other projects, the Open Language Data Initiative[1] (OLDI) now leads a collective endeavor to expand the number of supported languages. In particular, a shared task was proposed to extend OLDI's open datasets to more languages for the Ninth Conference on Machine Translation (WMT24).[2] This paper presents our efforts to extend the OLDI's FLORES+ dataset to four low-resource Romance languages spoken in Spain as part of this task, namely Aragonese, Aranese, Asturian, and Valencian. The resulting datasets can be downloaded from the repository of the PILAR corpus (Galiano-Jiménez et al., 2024).[3]

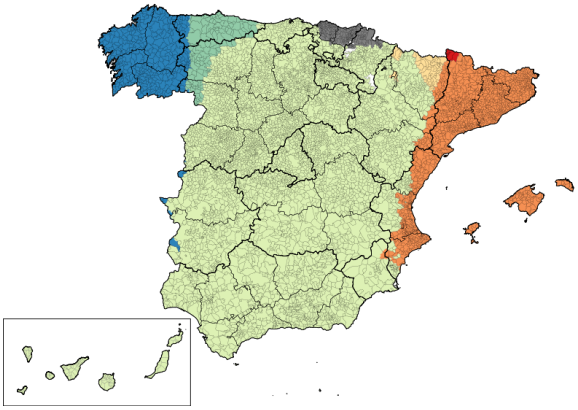

Figure 1: Some of the languages of Spain: Spanish, Catalan, Galician, Basque, Asturian, Aragonese and Aranese. The language names listed are shown in the corresponding colors of their regions on the map. Note that Spanish is spoken throughout the entire country, but the map highlights the regions where other languages are co-official or regionally predominant. [Source: Wikimedia Commons, https://commons.wikimedia.org/wiki/File:Spain_languages.svg]

Most languages spoken in Spain belong to the Romance language family, except for Basque, which is a *language isolate* with no known genetic

[1] https://oldi.org

[2] https://www2.statmt.org/wmt24/open-data.html

[3] https://github.com/transducens/PILAR

Proceedings of the Ninth Conference on Machine Translation, pages 547–555
November 15-16, 2024

relationship to any other language. Among the languages of Spain, Spanish is the only one with official status across the entire country. Other languages, including Catalan (primarily spoken in Catalonia, the Valencian Community, and the Balearic Islands, with a few thousand speakers in Aragon), Galician (mainly spoken in Galicia and some adjacent areas across the border), and Basque (spoken in the Basque Country and parts of Navarre), have official status in their respective autonomous regions (in certain areas for Navarre, and not in Aragon for Catalan). Additionally, Aranese is official in the Val d'Aran in accordance with the Statute of Autonomy of Catalonia. However, Asturian and Aragonese do not have official status in their respective regions, though they are recognized and protected as cultural heritage.[4] The Acadèmia Valenciana de la Llengua[5] (AVL) considers Valencian to be an alternative name for the Catalan language, as well as the term used to refer specifically to the variety of this language spoken in the Valencian Community.

Figure 1 illustrates the geographical distribution of these languages, while Table 1 provides estimates of their number of speakers in Spain. It is important to note that Spanish is spoken nationwide, while proficiency in regional languages varies among people in bilingual or diglossic areas.

We present our contribution to the FLORES+ benchmark,[6] focusing on four Romance languages spoken in Spain. Our target languages are Aragonese, Asturian, Aranese, and Valencian. These languages were chosen due to the evident social and governmental interest in preserving them. For Asturian, there are ongoing efforts to achieve official language status in Asturias. In contrast, Aranese is already an official language in the small central Pyrenees valley where it is moslty spoken, although it has very few speakers and limited textual resources. Aragonese is recognized and protected in Aragon as a heritage language, yet political support for granting it official status remains limited. Lastly, Valencian, a variant of Catalan with notable lexical differences, was included because having MT systems generate it directly rather than adapting from Catalan significantly reduces the need for additional post-editing.

2 Target Languages Overview

This section provides a brief overview of the languages covered in this paper.

2.1 The Aragonese Language

Aragonese is a Romance language spoken in the Pyrenees valleys of Aragon, primarily in the *comarcas* of Somontano de Barbastro, Jacetania, Alto Gállego, Sobrarbe, Ribagorza and Hoya de Huesca/Plana de Uesca.

Until recently, Aragonese has had several alternative orthographic norms, none of them official. In 1987, a number of associations organized a congress where a quasi-phonetic spelling system for Aragonese was approved, called *Normas Graficas de l'Aragonés*. It was commonly used in the subsequent two decades. The *Sociedat de Lingüistica Aragonesa* (SLA, Aragonese Linguistic Society), an association established in 2004, published a set of spelling rules (SLA rules) in 2006 based on the written tradition of Medieval Aragonese. In 2010, the *Estudio de Filología Aragonesa – Academia de l'Aragonés* (EFA-ACAR), a private entity created by the II Congress on Aragonese Language, approved and published the *Propuesta Ortografica de l'Academia de l'Aragonés*, more aligned with etymology. Finally, the *Academia Aragonesa de la Lengua*[7] (AAL, Aragonese Language Academy), a public body created by the Law of the Languages of Aragon, was established in 2021, and approved in April 2023 the standard orthographic norm for Aragonese. This standard orthography has been adopted by associations (including EFA-ACAR) as well as by the main pub-

[4] It is also worth noting that there are other languages in Spain that could be considered, but there is less consensus on whether they are part of an enclosing language or distinct languages on their own, debates that echo the famous saying attributed to sociolinguist Max Weinreich that "a language is a dialect with an army and a navy." For example, Eonavian or Galician-Asturian is a set of Romance dialects that has been classified either as northeastern varieties of Galician, as a linguistic group of its own, or as a transitional dialect between western Asturian and Galician. Another example is the Leonese language, which is currently spoken in the northern and western parts of the historical region of León in Spain (the modern provinces of León, Zamora, and Salamanca). Leonese, however, is considered part of the Asturleonese linguistic group, along with the dialects of Asturian. Similarly, Fala (spoken in the northwestern part of Extremadura) and Cantabrian also exhibit this ambiguity being classified, depending on the perspective, as distinct languages, hybrid dialects, or intermediate linguistic varieties.

[5] https://www.avl.gva.es

[6] FLORES+ is OLDI's evaluation benchmark for multilingual machine translation, building upon the FLORES-200 benchmark (Costa-jussà et al., 2024), which spans over 200 languages and comprises 2 009 sentences split into development (dev) and devtest sets. Texts come from English Wikimedia (more exactly, Wikinews, Wikijunior and Wikivoyage).

[7] https://www.academiaaragonesadelalengua.org

Language	Speakers	Bibliographic reference
Spanish	47 000 000	(Instituto Nacional de Estadística, 2022, pages 6–7)
Catalan (incl. Valencian)	9 000 000	(Generalitat de Catalunya, 2007, Table 2)
Valencian	3 500 000	(Generalitat Valenciana, 2021, page 6)
Galician	2 100 000	(Observatorio da Lingua Galega, 2007, page 11)
Basque	1 200 000	(Depto. de Cultura y Política Lingüística, 2023, page 2)
Asturian	250 000	(Llera Ramo, 2018, Figure 6)
Aragonese	25 000	(Reyes et al., 2017, Table 5)
Aranese	4 500	(Generalitat de Catalunya, 2019, page 4)

Table 1: Approximate number of speakers *in Spain* for some of the languages spoken in the country. When the reference includes different figures for the number of speakers depending on their ability to speak, understand, read, or write the language, we provide the data for the number of people who can speak it, including those with even a basic level of proficiency.

lishers in Aragonese. In August 2024, the Government of Aragon established this standard spelling for administrative and educational uses.[8]

A recent study reveals that Aragonese remains alive across the entire area where it is traditionally spoken (Eito et al., 2024). As of today, it is estimated that around 8 000 people use it on a daily basis, according to the most optimistic estimates, within a broader group of approximately 25 000 people who claim to have knowledge of the language (Reyes et al., 2017). Its use is much more prevalent in family and neighbourly interactions, especially among older individuals and in rural communities. However, the use of the language declines significantly in more formal contexts and outside the immediate environment of the speakers. Consequently, intergenerational transmission is severely at risk; some initial measures, such as teaching in schools or the language limited presence in the media, may not be sufficient to ensure the language survival.

2.2 The Aranese Language

Aranese is a standardized form of the Pyrenean Gascon variety of the Occitan language spoken in the Val d'Aran in northwestern Catalonia. Aranese is one of the three official languages in Catalonia, alongside Catalan and Spanish.[9] In the case of Aranese, its official status is limited to the territory of the Val d'Aran.

The Aranese language is regulated by the *Institut d'Estudis Aranesi–Acadèmia Aranesa dera Lengua Occitana*.[10] The main goal of this institution is to establish and update the linguistic norms of the Aranese variety of Occitan and ensure that the standardization process is consistent throughout its linguistic area.

Regarding education, the Occitan language is included, along with Catalan, Spanish, English, and optionally French, in the school system of the valley at all compulsory levels. However, the presence of Aranese in higher education, such as in the baccalaureate program, is virtually non-existent. According to the Linguistic Census of Aranese 2001,[11] only 34.2% of the population in the Val d'Aran have Aranese as their mother tongue, and only 25.8% use Aranese exclusively at home. The estimated number of speakers is approximately 4 500 (Generalitat de Catalunya, 2019).

2.3 The Asturian Language

Asturian is one of the Iberian Romance languages, closely related to Galician-Portuguese and Castilian Spanish, and a language historically influencing the Silver Way[12] to the Portuguese border (in fact, Asturian-based Mirandese is official in Miranda do Douro, in Portugal).

[8] https://www.boa.aragon.es/cgi-bin/EBOA/BRSCGI?CMD=VEROBJ&MLKOB=1347070761212

[9] The Val d'Aran was relatively well connected to Occitania in the north, but remained isolated to the south until the construction of the Vielha Tunnel, first in 1948 and then with a new one in 2007. The changes experienced by the valley due to its shift toward the south following the opening of the Vielha Tunnel, and especially the development of tourism, have led to a massive influx of migrants initially from the rest of Spain and later from other countries. This migration has resulted in drastic changes in the linguistic practices of the Aranese people. Occitan has increasingly been confronted with Spanish and Catalan, losing the dominance it once had in past centuries.

[10] https://www.institutestudisaranesi.cat

[11] https://llengua.gencat.cat/web/.content/documents/altres/arxius/aran_cens.pdf

[12] The Silver Way is a route of the *Camino de Santiago* that runs from southern Spain to Santiago de Compostela.

The *Academia de la Llingua Asturiana*[13] (Academy of the Asturian Language) is an official institution of the Government of the Principality of Asturias that promotes and regulates the Asturian language. Its main objectives include researching and standardizing the Asturian language, developing language usage norms and dictionaries, promoting its use and education, compiling its lexicon, fostering research related to Asturian, awarding literary prizes, and protecting the language rights of Asturian users. As a result, the Asturian language has clear and widely accepted spelling and orthographic rules.

Although it does not have official language status, Asturian is protected under the Statute of Autonomy of Asturias. In many schools, children can take Asturian-language classes, and in some schools, it is offered as an elective language.

In a report for the BBC, the President of the Academy of the Asturian Language pointed out that "at present about 250 000 people are able to understand, speak, read and write in Asturian, that is, a 25% of the population of the region" (Hernández, 2022; Llera Ramo, 2018). As the latest 3rd Sociolinguistic Study on the Asturian Language (Llera Ramo, 2018) points out, 87% of the population identify with both Asturian and Spanish identities. Of these, 64% have a balanced sense of both identities, while 18% feels predominantly Asturian over Spanish. Literacy rates are high: 90% of the population report understanding Asturian, though 29% describe their knowledge as passive, while 38% are able to read and 25% to write.

2.4 The Catalan and Valencian Languages

Catalan is a Romance language with official status in three autonomous communities in Spain: Catalonia, the Balearic Islands and the Valencian Community. It is also the official language of Andorra, a small state in the Pyrenees, and it has semi-official status in the Italian *comune* of Alghero, in the island of Sardinia. Catalan is also spoken in the south of France and in some parts of the Spanish regions of Aragon and Murcia.

Catalan is usually named Valencian in the Valencian Community, but both terms academically refer to the same language. Catalan is divided into two major dialect groups: Eastern and Western (see Figure 2). The main difference is in the pronunciation of unstressed 'a' and 'e'. In Eastern dialects, these

Figure 2: Catalan's two main dialects (Eastern/Western) shown divided by the dashed line. [Source: modified from the original in Wikimedia Commons, https://commons.wikimedia.org/wiki/File:Catalan_dialects-en.png]

sounds have merged into /ə/, while in Western dialects, they remain distinct as /a/ and /e/. There are also some other differences in pronunciation, verb forms, and vocabulary. Western Catalan includes the Northwestern Catalan and Valencian dialects. The Eastern group includes four dialects: Central Catalan, Balearic, Rossellonese, and Algherese.

The *Institut d'Estudis Catalans*[14] in Catalonia and the *Acadèmia Valenciana de la Llengua*[15] are the two institutions that regulate the Catalan and Valencian varieties, respectively. The relationship between these two institutions has not always been easy, but in recent years, they have achieved some degree of coordination.

In 2018[16], 94.4% of the population aged 15 or older in Catalonia could understand Catalan, 81.2% could speak it, 85.5% could read it, and 65.3% could write it. In the Valencian Community, 79.4% of the population aged 15 or older could understand Valencian, 54.9% could speak it, 60.9% could read it, and 44.4% could write it (Generalitat Valenciana, 2021). Valencian speakers make up approximately 3.5 million of the total 9 million Catalan speakers in Spain (Generalitat Valenciana, 2021; Generalitat de Catalunya, 2007).

Although standard Catalan and Valencian are highly similar and largely mutually intelligible, the differences and the effort required to adapt Cata-

[13] https://alladixital.org

[14] https://www.iec.cat

[15] https://www.avl.gva.es

[16] https://www.idescat.cat/indicadors/?id=basics&n=10367&lang=en

lan texts into Valencian are significant to justify the creation of a dedicated FLORES+ dataset for Valencian. This will support the development and evaluation of MT systems specifically designed for this variant.

3 Development of the FLORES+ Datasets

This section describes the workflow and resources used to produce the FLORES+ datasets for Aragonese, Aranese, Asturian and Valencian.

3.1 The Aragonese FLORES+ Dataset

A first draft of the FLORES+ dev and devtest sentences for Aragonese was initially obtained from Spanish using the Spanish–Aragonese[17] rule-based machine translation system Apertium (Forcada et al., 2011). This translation was subsequently post-edited by specialists proficient in Aragonese. Finally, the post-edited translation was reviewed by a member of the Academia Aragonesa de la Lengua. In this last step, the reviewer had also in view the English version of the FLORES+ dataset; an important number of translations were modified for better agreement with the English original data.

The translators and reviewers relied on the following resources to inform their decisions:

- The orthography of Aragonese[18] published by the Academia Aragonesa de la Lengua in 2023.
- The grammar of Aragonese published by the EFA-ACAR.[19]
- The dictionary of Aragonese,[20] published by the EFA-ACAR.
- The Aragonese dictionary *Tresoro d'a Luenga Aragonesa*,[21] a lexicographical research project developed by the *Instituto de Estudios Altoaragoneses*, in collaboration with the Government of Aragon.
- The Spanish–Aragonese and Aragonese–Spanish dictionary *Aragonario*,[22] developed under the project Linguatec and published by the Government of Aragon.

Justification for the use of MT. While it might be argued that using machine translation (MT) in a dataset like FLORES+, conceived as a benchmark for MT systems, is not the ideal procedure, our decision is justified by three main reasons: first, the two-step workflow of MT followed by post-editing is common practice for this language, with many existing texts produced in this manner; second, the scarcity of qualified linguists and translators for Aragonese made it difficult to complete the task within the required timeframe without the support of an MT system; third, rule-based systems like Apertium typically exhibit strong performance for closely related Romance languages, resulting in minimal translationese that does not significantly impact the overall quality.

3.2 The Aranese FLORES+ Dataset

Similarly to Aragonese (see above), the data for Aranese was initially obtained by translating the sentences in the Catalan FLORES+ dev and devtest datasets using the Apertium rule-based MT system (Forcada et al., 2011) for Catalan–Aranese.[23] The revision process was then performed in two steps. Firstly, a professional reviewer with wide experience in translation and revision with proficiency in Aranese was presented with the French, Catalan and Occitan versions of the FLORES+, along with the machine translated version into Aranese. Finally, the post-edited translation was reviewed by different individuals, who are native speakers, from the Institut d'Estudis Aranesi (IEA). The use of MT is motivated by the same reasons as those for Aragonese (see the end of the previous section).

During the post-editing and the subsequent review process, the guidelines provided by the IEA were strictly followed to ensure that the translation into Aranese aligned with their recommendations. Specifically, the translators and reviewers based their decisions on the following resources published by the IEA:

[17] https://github.com/apertium/apertium-spa-arg, release 0.5.0, the latest release available at the time of translation.

[18] https://academiaaragonesadelalengua.org/sites/default/files/ficheros-pdf/ortografia-de-laragones_web_an.pdf

[19] http://www.academiadelaragones.org/biblio/Edacar10_GBAprovisional.pdf

[20] http://www.academiadelaragones.org/biblio/Edacar13.pdf

[21] http://diccionario.sipca.es/fabla/faces/index.xhtml

[22] https://aragonario.aragon.es/

[23] https://github.com/apertium/apertium-oci-cat, release 1.0.8.

- Dictionary for Aranese;[24] erratum 2021;[25] extensions 2021,[26] 2022[27] and 2023.[28]
- Grammar for Aranese.[29]
- Orthographic vocabulary of Aranese.[30]
- Winter sports vocabulary of Aranese.[31]
- Computer and informatics vocabulary of Aranese.[32]
- The Aranese verbs.[33]

3.3 The Asturian FLORES+ Dataset

The dev and devtest datasets we are contributing for Asturian are a corrected version of the original FLORES+ dataset, which were initially translated from English by Meta using professional translators as part of their *no language left behind* initiative (Costa-jussà et al., 2024). We had this translation into Asturian reviewed by native speakers, some of whom are members of the *Academia de la Llingua Asturiana*, philologists and a renowned writer, translator and activist for the Asturian language. The revision process was carried out twice by different people. In the first round, the reviewers were presented with the Spanish text and the existing version of the Asturian FLORES+, and in the second round, with the Spanish FLORES+ sentences and the first revised version.

During the review process, special attention was paid to adhering to the guidelines provided by the Academia to ensure that the translation aligns with their recommendations by relying on the following resources:

- Diccionariu de la Llingua Asturiana.[34]
- Gramática de la Llingua Asturiana.[35]
- Normes ortográfiques.[36]

3.4 The Valencian FLORES+ Dataset

The dataset for Valencian was created by adapting the existing Catalan version of the FLORES+ devtest set.[37] The work was carried out by a single native speaker of the Valencian variant, who holds a university degree and has experience translating into Valencian and reviewing texts in this language.

To expedite the process, LanguageTool[38] was employed to apply an initial set of changes to the original text. These changes were then manually reviewed, and additional necessary modifications were made, considering a list of lexical items that differ across the various Catalan dialects (see below). In cases of uncertainty, the dictionary of the Acadèmia Valenciana de la Llengua was consulted. The following resources were used to guide the modifications to the original Catalan text:

- List of lexical items that differ across the different dialects of the Catalan language.[39]
- Diccionari normatiu valencià.[40]
- Linguistic criteria for the institutional use of the Valencian dialect in the Valencian universities.[41]

3.5 Modifications Across Versions of the Datasets

As previously mentioned, the datasets for Aragonese, Aranese, and Asturian have undergone several iterations during their development. Table 2 shows a comparison of the translation error rate (TER) between these versions, labeled as v1,

[24] http://www.institutestudisaranesi.cat/wp-content/uploads/2021/04/DICCIONARI-DER-ARAN%C3%89S.pdf
[25] http://www.institutestudisaranesi.cat/wp-content/uploads/2020/12/ERRATA-WEB.pdf
[26] http://www.institutestudisaranesi.cat/wp-content/uploads/2020/12/500-1-g%C3%A8r-2021-WEB.pdf
[27] http://www.institutestudisaranesi.cat/wp-content/uploads/2022/01/WEB-AMPLIACION-01-01-2022.pdf
[28] http://www.institutestudisaranesi.cat/wp-content/uploads/2023/03/Ampliacion-2023.pdf
[29] http://www.institutestudisaranesi.cat/wp-content/uploads/2021/04/gramatica-aranes.pdf
[30] http://www.institutestudisaranesi.cat/wp-content/uploads/2019/11/33520-vocabulari-ortografic.pdf
[31] http://www.institutestudisaranesi.cat/wp-content/uploads/2019/11/33288-vocabulari-esports-iuern.pdf
[32] http://www.institutestudisaranesi.cat/wp-content/uploads/2018/11/TECNOLOGIA.pdf
[33] http://www.institutestudisaranesi.cat/wp-content/uploads/2019/11/33287-els-ve%CC%80rbs.pdf

[34] https://diccionariu.alladixital.org/
[35] https://alladixital.org/wp-content/uploads/2022/08/Gramatica-de-la-Llingua-Asturiana.pdf
[36] https://alladixital.org/wp-content/uploads/2024/01/Normes-Ortografiques-8a-edicion-FINAL-3.pdf
[37] It is important to note that Valencian is the only target language for which only one of the two FLORES+ datasets (the devtest set) has been translated, instead of both.
[38] https://languagetool.org
[39] https://ca.wikipedia.org/wiki/Llista_diat%C3%B2pica_del_l%C3%A8xic_catal%C3%A0
[40] https://www.avl.gva.es/lexicval
[41] https://sl.ua.es/en/assessorament/documentos/criteris-linguistics.pdf

Language	v1 → v2	v2 → v3	v1 → v3
Aragonese dev	4.8	22.2	24.3
Aragonese devtest	4.0	26.8	28.4
Aranese dev	1.8	54.2	54.7
Aranese devtest	0.1	70.3	70.2
Asturian dev	6.0	4.4	10.0
Asturian devtest	5.5	0.1	5.6

Table 2: TER scores comparing different versions (v1, v2, and v3) of translations for Aragonese, Aranese, and Asturian. v1 represents the original translations, v2 the post-edited or improved versions (depending on the language; see the main text), and v3 the standardized versions produced under the supervision of language academies.

v2, and v3.[42] The TER metric (Snover et al., 2006) quantifies the number of edits required to transform one sentence into another, referred to as the *reference*. It is calculated by dividing the total number of edits (insertions, deletions, substitutions, and shifts) by the total number of words in the reference. TER is a widely accepted metric in machine translation evaluation, providing insight into how close a system's output is to a human-generated reference. Lower TER scores correspond to higher translation quality. In this study, we use TER to compare different versions of the datasets for each language, offering an estimate of the extent of changes introduced by the reviewers during the revision process. In our case, higher TER scores correspond to a larger number of edits.

Version v1 refers to the original translations. As already mentioned, for Aragonese and Aranese, the translations in v1 were initially produced by Apertium, whereas v2 represents the post-edited translations of v1. For Asturian, the v1 corresponds to the version already included in the FLORES+ dataset, and v2 incorporates some normative and stylistic corrections to improve the quality of the translations. Finally, v3 refers in all cases to a version that was further adapted to conform to the linguistic standards promoted by the respective language academies. The revisions carried out under the direct supervision of these academies were designed to ensure the highest quality and compliance with current linguistic norms.

The data reveals that in most cases, the second round of quality checks leading to v3 introduced significant differences compared to v2. This is particularly notable for Aragonese and Aranese, where the adaptation to standard forms resulted in substantial changes, as reflected in the TER scores. Asturian is an exception, especially in the devtest set, where the differences between v2 and v3 were minimal, suggesting that the initial revisions in v2 already addressed most of the necessary improvements. This underscores the importance of multiple quality review stages, as the adaptation to standardized forms can introduce notable refinements to the translations.

4 Validation of the FLORES+ datasets

The dev sets of FLORES+ for Asturian, Aragonese, and Aranese were distributed and utilized in an WMT24 shared task[43] which focused on translating from Spanish into low-resource languages of Spain (Sánchez-Martínez et al., 2024). These resources proved valuable for the participants as validation sets for training neural MT systems. The corresponding devtest subsets were then used by the organizers of the shared task to evaluate the submitted systems. To ensure fairness in the evaluation, the devtest sets were withheld until the end of the submission period. Several submitted systems outperformed the respective Apertium-based baselines, achieving higher performance scores based on the BLEU, chrF++ and TER metrics.

The Valencian FLORES+ dataset was utilized in the development of a Spanish–Valencian neural MT system. One experiment involved training models with varying proportions of training data in the Valencian and Central Catalan dialects. As expected, results indicated that the higher the proportion of Valencian data in the training set, the higher the BLEU score achieved on the Valencian FLORES+ dataset and the greater the presence of Valencian dialectal forms in the output.

As a tentative indicator of the baseline performance expected on the newly created devtest sets, Table 3 presents standard automatic evaluation metrics for Apertium-based systems (Forcada et al., 2011) translating the Spanish side of the devtest. The results suggest that the use of MT (specifically the Apertium system) during the initial phase of corpus creation likely contributed to the significantly higher evaluation scores for the Aragonese and Aranese Apertium-based systems compared to Asturian. This points to a potential bias in the dataset

[42] SacreBLEU (Post, 2018) TER signature: nrefs:1 | case:lc | tok:tercom | norm:no | punct:yes | asian:no | version:2.0.0

[43] https://www2.statmt.org/wmt24/romance-task.html

Apertium system	BLEU	chrF++	TER
Spanish–Aragonese	61.1	79.3	27.2
Spanish–Aranese	28.8	49.4	72.3
Spanish–Asturian	17.0	50.8	80.4

Table 3: Automatic evaluation scores of the Apertium-based systems on the devtest FLORES+ data. The Apertium data correspond to the following releases: `spa-arg`, 0.6.0; `oc-es`, 1.0.8; `spa-ast`, 1.1.1.

favoring rule-based MT approaches over non-rule-based systems, and even human translations, when evaluating the devtest translations from Spanish into these languages. Nonetheless, results from the WMT24 shared task (Sánchez-Martínez et al., 2024) show that neural systems can still outperform Apertium, even for Aragonese and Aranese. This suggests that the use of MT in the early stages of creating the Aragonese and Aranese FLORES+ datasets did not significantly compromise the utility of the data.

5 Concluding remarks

In this paper, we have detailed the development of the FLORES+ datasets for four low-resource Romance languages spoken in Spain: Aragonese, Aranese, Asturian, and the Valencian variant of Catalan. Each dataset was meticulously curated through a two-step manual review process involving native speakers and professionals from the respective language academies. The creation of these datasets is particularly significant given the ongoing efforts to revitalize and promote these languages. Our work also highlights several key challenges in resource development for low-resource languages, such as the scarcity of expert translators, or the constraints of limited time and funding for dataset production.

Acknowledgements

This work was supported by the R+D+i projects PID2021-127999NB-I00 (*LiLowLa: Lightweight neural translation technologies for low-resource languages*) and PID2021-124663OB-I00 (*TANIBE: Neural Machine Translation for the Romance languages of the Iberian Peninsula*), founded by MCIN /AEI /10.13039/501100011033 / FEDER, UE. The development of the FLORES+ dataset for Valencian was supported by project CENID2023/10, funded by Diputació d'Alacant.

We thank Academia Aragonesa de la Lengua, Institut d'Estudis Aranesi, and Academia de la Llingua Asturiana for their support and assistance in the translation of the FLORES+ sentences.

References

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Jeff Wang, and NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Depto. de Cultura y Política Lingüística. 2023. *VII Encuesta Sociolingüística 2021. Comunidad Autónoma de Euskadi. Informe Resumen*. Departamento de Cultura y Política Lingüística, Gobierno Vasco, Viceconsejería de Política Lingüística, Donostia-San Sebastián.

Antonio Eito, José Ángel Iranzo, and Chaime Marcuello. 2024. Hablando aragonés: análisis de la encuesta de uso de la lengua aragonesa en la zona de uso predominante y su evolución. Unpublished.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Aarón Galiano-Jiménez, Felipe Sánchez-Martínez, and Juan Antonio Pérez-Ortiz. 2024. Pan-iberian language archival resource.

Generalitat de Catalunya. 2007. *El català, llengua d'Europa*. Generalitat de Catalunya, Departament de la Vicepresidència, Secretaria de Política Lingüística, Barcelona.

Generalitat de Catalunya. 2019. *Els usos lingüístics de la població de l'Aran: Principals resultats de l'Enquesta d'usos lingüístics de la població. 2018*. Generalitat de Catalunya, Barcelona.

Generalitat Valenciana. 2021. *Conocimiento y uso social del valenciano: síntesis de resultados*. Generalitat Valenciana, Conselleria d'Educació, Cultura i Esport, València.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Alicia Hernández. 2022. Dónde se habla el bable y por qué dicen que es un idioma en peligro. https://www.bbc.com/mundo/noticias-internacional-59547573.

Instituto Nacional de Estadística. 2022. Encuesta de características esenciales de la población y viviendas. Año 2021. Datos definitivos. Nota de prensa.

Francisco J. Llera Ramo. 2018. *III Estudio Sociolingüístico de Asturias 2017: Avance de resultados*. Academia de la Llingua Asturiana, Uvieu. Estaya Sociollingüística, colección 7.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. 2023. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

Observatorio da Lingua Galega. 2007. *Situación da lingua galega na sociedade: Observación no ámbito da cidadanía: resumo executivo*. Observatorio da Lingua Galega, Xunta de Galicia, Santiago de Compostela.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*, 55(11).

Anchel Reyes, Chabier Gimeno, Miguel Montañés, Natxo Sorolla, Pep Espluga, and Juan Pablo Martínez. 2017. *L'aragonés y lo catalán en l'actualidat: analisi d'o censo de población y viviendas de 2011*. Seminario Aragonés de Sociolingüística, Asociación Aragonesa de Sociolochía, Universidad de Zaragoza. Primera parte, febrero 2017.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Felipe Sánchez-Martínez, Juan Antonio Pérez-Ortiz, Aarón Galiano-Jiménez, and Antoni Oliver. 2024. Findings of the WMT 2024 Shared Task on Translating into Low-Resource Languages of Spain: Blending rule-based and neural systems. In *Proceedings of the Ninth Conference on Machine Translation (WMT24)*, Miami, Florida, USA. Association for Computational Linguistics.