Enhaced Apertium System: Translation into Low-Resource Languages of Spain
Spanish–Asturian

Sofía García González
imaxin|software
Rúa dos Salgueiriños de Abaixo, 11, L-6, 15703
Santiago de Compostela, A Coruña
sofia.garcia@imaxin.com

Abstract

We present the Spanish–Asturian Apertium translation system, which has been enhanced and refined by our team of linguists for the shared task: Low Resource Languages of Spain of this WMT24 under the closed submission. While our system did not rank among the top 10 in terms of results, we believe that Apertium's translations are of a commendable standard and demonstrate competitiveness with respect to the other systems.

1 Introduction

In this shared task: Translation into Low-Resource Languages of Spain, we present an enhancement of the machine translator *Eslema* system for Spanish–Asturian pair (Viejo et al., 2008). We present our system under the closed submission.

Asturian is not an officially recognised language in the Spanish state. The 1981 Statute of Autonomy of Asturias makes only passing reference to Asturian, citing the need to protect and disseminate it. However, it does not accord the language the same privileges as are enjoyed by other official languages of Spain, such as Galician, Catalan or Basque. Subsequently, on 23 March 1998, the Asturian Parliament enacted the Law on the Use and Promotion of Asturian. The aforementioned legislation stipulates that all citizens are entitled to utilise Asturian in verbal and written communication, and that such communication shall be deemed valid. Furthermore, it acknowledges the necessity for the dissemination of Asturian in educational and media contexts (Galán y González, 2015). Consequently, Asturian is categorised as a minority and Low-Resource Language (LRL), exhibiting a paucity of resources and a diminished presence in Natural Language Processing (NLP) relative to other co-official languages of Spain.

The objective of this shared task is to develop innovative systems and data resources for this low-resource language, the Aragonese and the Aranese. In light of the aforementioned considerations, we present an enhacement of Apertium (Forcada et al., 2011), the foundational system of the current Spanish–Asturian MT translator, *Eslema*. Based on this open-source system, a series of grammatical, syntactic and lexical improvements have been implemented in order to participate in this shared task. While the results obtained have not been sufficient to maintain a position within the top 10, they have been noteworthy.

2 Eslema

Eslema was initiated as a project of the University of Oviedo in 2004 with the objective of assembling corpora in Asturian language. The Asturian Philology Group (*Seminariu de Filoloxía Asturiana*) within the Spanish Philology Department was responsible for its research, with the aim of compiling texts of diverse typology, format and historical periods (Viejo et al., 2008).

Subsequently, at the conclusion of 2008, the Principality of Asturies (*Conseyería de Cultura del Principáu d'Asturies*) assumed the economic responsibility for the establishment of the regulatory framework for the development of a rule-based machine translator for the Spanish–Asturian language pair. This project was carried out in collaboration between the University of Oviedo and the Apertium community. The report published in early 2010 about this translator acknowledged that its functionality was satisfactory, especially in the Spanish–Asturian direction. However, it was also noted that the software still presented some residual problems that would be solved in subsequent updates (Universidad de Oviedo, 2010).

Nevertheless, it is important to recognise that, despite the best efforts of the developers to rectify all potential errors, no machine translator can be considered perfect. In particular, Rule Based Machine Translators require ongoing maintenance and revision to ensure optimal performance. Subsequently,

Proceedings of the Ninth Conference on Machine Translation, pages 878–884

Eslema has been provided to the Administration and citizens free of charge until nowadays. Right now there is a new version of the machine translator in the Linguistic Policy (*Política Llingüística*) webpage.[1]

3 Apertium

Apertium is a free/open-source platform for Rule-Based Machine Translation (RBMT) (Forcada et al., 2011). It is designed to provide high-quality translation tools for several LRL pairs. Apertium was initially developed as part of a project developed by the Alacant University and different public and private Spanish institutions and companies such as **imaxin**|software[2] and El Huyar.[3] But since then, it has evolved into a collaborative endeavour involving developers, linguists and researches (Khanna et al., 2021). The platform is constructed on a rule-based translation system, which relies on predefined linguistic rules and a modular architecture. In the most recent versions of Apertium, these modules are divided into two monolingual packages and one bilingual package for each language pair. The following subsections will provide an explanation of the monolingual packages 3.1 and the bilingual package 3.2.[4] These modules constitute the various components of the Apertium engine pipeline. Modifications have been made to them with the objective of enhancing the original *Eslema* translator.

3.1 Monolingual Packages

There is a monolingual package for each language in the language-pair. For example, in this case, an Asturian package and a Spanish package. Each of these packages is formed by a dictionary 3.1.1, a post-generator 3.1.2 and a Constraint Grammar 3.1.3.

3.1.1 Monolingual Dictionary

Monolingual dictionaries[5] serve the function of regulatory modules for the system's lexicon. The dictionary is comprised of two principal sections. The initial one is the paradigm section, wherein paradigms are defined as patterns or models that delineate the potential declensions of each term, contingent on its category or morphology. The subsequent section is where the lexicon is incorporated. Each novel word is introduced in the format of an entry, which encompasses the term in its fundamental form and the paradigm ascribed to it.

3.1.2 Post-generator

The post-generator[6] is a module that is employed to rectify minor spelling issues in each language. To illustrate, in languages where contractions or the use of apostrophes is prevalent, these orthographic phenomena are regulated by the post-generator. It is typically a module that remains consistent for each language and does not necessitate significant updates.

3.1.3 Constraint Grammar

As posited by Bick and Didriksen (2015), the Constraint Grammar (CG) may be conceived of as a declarative whole of contextual possibilities and impossibilities for a language or genre. However, in programming terms, it is implemented procedurally as a set of consecutively iterated rules that add, remove or select tagged-encoded information.

In Apertium, each language package is equipped with a CG tool, which serves to clarify the source text. One illustrative example of a CG rule is as follows[7]:

The rule "SELECT VERB IF (1 (det))" indicates that the verb category must be selected whenever the following word is a determiner.

This tool is of vital importance in an RBMT engine, as disambiguation errors can lead to significant translation errors. Therefore, the more developed the CG is, the more accurate the translation will be. For this particular pair, the Spanish CG[8] has been used, which was previously created by the Apertium community and, due to errors detected, has had to be modified on some occasions.

3.2 Bilingual Package

There is one bilingual package for each language pair. Each bilingual package is formed by a bilin-

[1] https://politicallinguistica.asturias.es/eslema

[2] https://imaxin.com/gl/

[3] https://www.elhuyar.eus/eu

[4] The majority of the information pertaining to these modules has been derived from the Apertium Wiki: https://wiki.apertium.org/wiki/Main_Page

[5] See:https://wiki.apertium.org/wiki/Monodix_basicshttps://wiki.apertium.org/wiki/Monodix_basics

[6] See:https://wiki.apertium.org/wiki/Post-generator

[7] See:https://wiki.apertium.org/wiki/Constraint_Grammar

[8] See: https://github.com/apertium/apertium-spa/blob/master/apertium-spa.spa.rlx

gual dictionary 3.2.1, the transfer rules 3.2.2 and the lexical selection rules 3.2.3.

3.2.1 Dictionary

In bilingual dictionaries,[9] the terms of both languages are aligned with one another. As a general rule, in this type of engine, each term in the source language can only have one correspondence in the target language. In other words, a term in the source language will be translated by the same term in all contexts, with the exception of specific cases which will be addressed in the subsequent modules.[10]

3.2.2 Transfer Rules

Transfer rules[11] are employed to oversee the most intricate structural divergences between two languages, whether pertaining to syntax, morphology, or grammar. To illustrate, transfer rules facilitate the rearrangement of a sentence in the target language, the alteration or insertion of tags by category, and other modifications that enhance the coherence of the target language. In essence, this module is responsible for managing the majority of complex changes that are contingent upon grammatical or lexical context.

3.2.3 Lexical Selection Rules

In instances where a term in the source language has two or more potential translations in the target language, the lexical transfer rules module[12] is responsible for selecting one or the other option, depending on the surrounding context. This context may be either grammatical or lexical in nature.

4 Dependencies

The dependencies of our translation system are presented in the following list. It is imperative that all modules are in place for the correct functioning of the pair: Apertium-3.8.3, lttoolbox-3.7.1, apertium-lex-tools-0.4.2 and cg3-3.9.

[9] See:https://wiki.apertium.org/wiki/Bilingual_dictionary

[10] The latest versions of Apertium include the Lexical Selection Rules module 3.2.3, which enables to assign a specific meaning and translation to the target language depending on the context. This module will be explained in subsection.

[11] See:https://wiki.apertium.org/wiki/A_long_introduction_to_transfer_rules

[12] See:https://wiki.apertium.org/wiki/Constraint-based_lexical_selection_module

5 Methodology

In order to participate in this shared task, the team at **imaxin**|software has utilized the open-source translator published in 2010 by *Eslema*[13] and made available on the Apertium project website,[14] to enhance it in the morphological 5.1 and lexical 5.2 linguistic areas. The implementation of these alterations and enhancements was overseen at all times by a team of linguists with expertise in Asturian, over a period of 18 months, during which not only the Spanish–Asturian direction was considered, but also the Asturian–Spanish.

5.1 Morphological Enhacement

With regard to morphological errors, three principal categories may be identified. Firstly, this pair presented a multitude of disambiguation issues. To illustrate, the preposition *para* (for) in Spanish was frequently analyzed as the third person singular of the verb *parar* (to stop). This resulted in errors such as: *Ir para casa* (go home) in Spanish was translated to Asturian as *ir **para** casa* innstead of *Ir **pa** casa*. These types of errors were corrected in a generic way by making use of the Constraint Grammar that had already been created by the Apertium community. However, it was also necessary to create new rules for specific cases.

Furthermore, the paradigms created for different grammatical categories contained various errors, either because the term had been assigned the wrong paradigm or because the assigned paradigm contained errors in its definition. To illustrate, Asturian verbs ending in *-ñir* or *-xir* (e.g. *teñir* (to dye) and *dirixir* (to address)) exhibited erroneous conjugation of the third person singular present indicative and subjunctive forms. This resulted in the generation of incorrect forms, such as *tiñió* (dyed) or *dirixió* (addressed), rather than the intended *tiñó* and *dirixó*. This was due to an erroneous assignment of the paradigm. It was thus necessary to create a specific paradigm for this type of verbs. Furthermore, a considerable number of Asturian verbs with enclitic pronouns were not correctly translated, resulting in the translation of their infinitive form instead of the expected conjugation. To address this issue, it was imperative to rectify the verb paradigms, which, due to inconsistencies with the paradigms of the Spanish dictionary, led to this

[13] https://eslema.it.uniovi.es/comun/traductor.php

[14] https://github.com/apertium/apertium-spa-ast

type of error.

Finally, the transfer rules for this pair were found to contain numerous errors, which resulted in a significant decline in the quality of the translation. These errors manifested in various ways, including inconsistencies in gender or number between related nouns and adjectives, the absence of verb conjugation or declension, incorrect disambiguation, and other issues that affected the whole translation.

The aforementioned examples illustrate the work that has been carried out on the monolingual and bilingual packages. As a result, 172 transfer selection rules were corrected for both translation directions and 11 paradigms for each monolingual dictionary.

5.2 Lexical Enhacement

The dictionaries have been expanded to include new terms drawn from a number of fields, including administration, toponymy from both Asturias and Spain, and anthroponymy. In total, 3100 new terms have been incorporated into the dictionaries, with the inclusion of each new term informed by the preferences of the Dictionary of the Asturian Academy (*Diccionariu de la Academia de la Llingua Asturiana*) (DALLA[15]). Indeed, one of the most significant alterations implemented at the generic level within the Apertium dictionaries has been the selection of the cultured form in lieu of the vocalised form of the term. To illustrate, the choice of *-pt-* instead of *-ut-* in terms such as *conceptu/conceutu*, the choice of *-ps-* instead of *-us-* in terms like *cápsula/cáusula*, the choice of *-cd-* instead of *-ud-* in terms such as *anécdota/anéuduta* and the preference for the intervocalic *-x-* rather than the *-s-* found in terms such as *exame/esame* are examples of the aforementioned changes. Similarly, the *-zar* ending is preferred for verbs such as *forzar/forciar*, in contrast to the *-ciar* ending. Otherwise, as mentioned above, the DALLA shape was always preferred in all cases where there were two possibilities.

6 Results

Table 1 presents the BLEU (Papineni et al., 2002) and chrF++ (Popović, 2015) scores received from the OCELoT system. The table includes the ten best systems presented to this shared task and our own system, identified by the ID 580.

[15]See:https://www.diccionariu.alladixital.org/

ID	BLEU	chrF++
576	**23.2**	**55.2**
606	19.8	52.2
574	19.7	52.2
528	18.4	52.1
609	19.8	52.1
551	18.2	51.6
557	17.9	51.6
629	18.0	51.6
568	18.0	51.6
564	18.0	51.6
580	17.6	51.2

Table 1: The best 10 scores obtained in the OCELoT system in the WMT24 Shared Task: Low Resource Languages in Spain (Spanish–Asturian) and the enhaced Apertium system, ID 580.

7 Analysis

In order to elucidate the outcomes yielded by our system and the top ten in this shared task, it is essential to examine the functioning of the BLEU and chrF++ metrics, on the one hand, and the FLORES test, on the other.

From one perspective, BLEU and chrF++ are lexical-based metrics that rely on a reference corpus to assess the quality of a translation. In essence, BLEU assesses the quality of the system by comparing the MT output with the reference test token by token at sentence or corpus level (Papineni et al., 2002). In contrast, the chrF++ metric functions in a comparable manner, albeit by comparing character by character rather than token by token (Popović, 2015). Both metrics have been the subject of criticism on the grounds of their reliance on a reference corpus, which presents a significant challenge for low-resource languages, such as Asturian. In the absence of the requisite test datasets, the evaluation with these metrics is often impractical. Furthermore, these metrics fail to account for the inherent variability and versatility of languages. In many cases, multiple translations may be equally valid for a given source sentence. However, these metrics treat any deviation from the reference corpus as an error, leading to artificially low metrics when the deviation is linguistically correct (Lee et al., 2023).

In light of the aforementioned considerations, it is pertinent to highlight that the FLORES+ corpus, which serves as the basis for the evaluation of the systems in this shared task, comprises 3001

English sentences extracted from Wikimedia and translated manually by linguists into 200 minority languages, including Asturian. Subsequently, these translations are subjected to automatic revision and post-editing as required (Costa-jussà et al., 2022). It should be noted that the parallel corpora generated by this project are not direct translations. However, the translated text maintains the meaning of the original sentence while deviating from the structure of the source language. This may be due to the fact that the corpus was generated from English. Furthermore, as stipulated in the terms of reference for this shared task, the Asturian corpus has been duly revised by the Asturian Academy for use as a reference corpus in this shared task.

For illustrative purposes, three examples can be found in Table 2. In this table it can be found the original sentence in Spanish from the FLORES+ devtest in the first column, the original sentence in Asturian from the FLORES+ devtest in the second column and the version of the same sentence in English in the third column. This version in English is also taken from FLORES+ devtest. It is evident from these sentences that the translation from Spanish to Asturian is not a literal one, but rather a free rendering. In some cases, the meaning may even change. For instance, in the first sentence, the English meaning is retained in the Spanish sentence, but is lost in the Asturian translation. The direct translation of the sentence from Asturian to English would be: "They all ran back **when the accident happened**". "From where the accident had happened" and "when the accident happened" is not meaning the same. Furthermore, information can also be lost, as evidenced by the second sentence. Once more, the Asturian translation does not convey the same information as the original English sentence and the Spanish translation. In this instance, information is lost. Rather than referencing the navigable canals, the translation merely states that they are located inland, and instead of indicating that they are an optimal destination for "holidays", the translation simply uses the word *viaxes* (travels) which is not an equivalent expression. Furthermore, as evidenced in the third sentence, this phenomenon also occurs in the context of English–Spanish sentence translation. In such instances, the order of the sentence may undergo a change from English to Spanish, even when there is no necessity to align with the grammatical conventions of the target language. Even minor alterations such as this one have a deleterious impact on evaluation using lexical-based metrics.

These discrepancies within the parallel test corpora give rise to suboptimal results in metrics such as BLEU or chrF++, particularly in instances where the translated text may not be technically incorrect. This is not only the case for our system, but for all of them. The highest metric for BLEU is 23.2, while for chrF++ it is 55.2. These results are considerably low.

In regard to the results obtained by our system, it can be stated that Apertium, as a RBMT, produces translations that are literal in nature. In other words, unless it is a syntactic or grammatical feature intrinsic to the target language, the structure of the source language will always be replicated. In the case of Spanish and Asturian, which are two closely related languages, the MT output produced by Apertium will invariably adhere to the structure of Spanish, rather than exhibiting a more Asturian-specific structure. For illustrative purposes, consider the sentences presented in Table 3. This table presents the same sentences as in Table 2, with the second column displaying the translations generated by our system instead of the FLORES+ devtest Asturian sentences. The examples illustrate that the translation produced by Apertium preserves the structure of the source sentence in Spanish, but the translations are all accurate. It should be noted, however, that the system itself is not without limitations. It should first be noted that a word in Spanish has only one possible translation into Asturian, irrespective of context. While this can be managed in some specific cases, it may result in the translation failing in other sentences. It is possible that the inflexibility of this system, which does not always permit adaptation to context, may have had an adverse effect on our results, extending beyond the aforementioned metrics and the test employed. Nevertheless, we consider the output of our system to be a satisfactory translation that could be competitive with other systems despite the results. However, it would be necessary to carry out more tests in order to go deeper and identify the aspects in which the quality of our translation system could be improved, since the RBMT systems, as already mentioned, require constant revision and improvement.

Finally, and this is a strong point of our proposal, this type of system does not have a high computational consumption like Statistical Machine

Original Sentences in Spanish	Original Sentences in Asturian	Original Sentences in English
Todos volvieron corriendo desde el lugar del accidente.	*Volvieron p'atrás corriendo cuando ocurrió l'accidente.*	They all ran back from where the accident had happened.
Los canales navegables internos pueden ser una buena temática para las vacaciones.	*Les canales d'interior son un bon tema de viaxe.*	Inland waterways can be a good theme to base a holiday around.
En Inglaterra, las vías de tren ya se habían instalado hacia el siglo XVI.	*Les primeres víes foron construyíes n'Inglaterra nel sieglu XVI.*	Wagonways were built in England as early as the 16th Century.

Table 2: Illustrative sentences of the FLORES test dataset taken from the FLORES+ devtest in Spanish, Asturian and English languages.

Original Sentences in Spanish	Apertium MT output	Original Sentences in English
Todos volvieron corriendo desde el lugar del accidente.	*Toos volvieron corriendo dende'l llugar del accidente.*	They all ran back from where the accident had happened.
Los canales navegables internos pueden ser una buena temática para las vacaciones.	*Les canales navegables internos puen ser una bona temática pa les vacaciones.*	Inland waterways can be a good theme to base a holiday around.
En Inglaterra, las vías de tren ya se habían instalado hacia el siglo XVI.	*N'Inglaterra, les víes de tren yá s'instalaren escontra'l sieglu XVI.*	Wagonways were built in England as early as the 16th Century.

Table 3: Apertium MT output from the Spanish–Asturian translation of three sentences taken from FLORES+ devtest in their Spanish and English version.

Translation (SMT) and Neural Machine Translation (NMT) in its training, as signalled by Shterionov and Vanmassenhove (2023). Comparing the quality produced by this system with its consumption, both for training/development and for use, Apertium is still a competitive system for low-resource languages such as Asturian. Furthermore, it is also more cost-effective to produce than other types of systems. Therefore, it is essential to consider the trade-off between quality, consumption and price in order to assess the performance of the different systems.

8 Conclusions

In conclusion, although our enhanced Apertium system has not yet achieved a position among the top ten systems in this shared task, the results obtained in terms of machine translation quality have been exemplary. As previously stated in the Section 7, the test employed and the metrics utilized do not permit an accurate assessment of the quality of a MT system. Additionally, it is noteworthy that a RBMT exhibits a markedly lower consumption rate in comparison to NMTs. Consequently, we regard our system as being competitive with those submitted to this shared task, although it still necessitates further enhancements and revisions.

References

Eckhard Bick and Tino Didriksen. 2015. Cg-3—beyond classical constraint grammar. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 31–39.

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 Mejía González, Prangthip Hansanti, John Hoffman, Semarley Jarret, 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, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

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*, 25:127–144.

Inaciu Galán y González. 2015. Asturianu sos: una güeyada sobre la situación de la llingua asturiana y les sos perspectives de futuru. *Luenga & fablas: publicazión añal de rechiras, treballos e decumentazión arredol de l'aragonés ea suya literatura*, (19):67–72.

Tanmai Khanna, Jonathan N Washington, Francis M Tyers, Sevilay Bayatlı, Daniel G Swanson, Tommi A

Pirinen, Irene Tang, and Hector Alos i Font. 2021. Recent advances in apertium, a free/open-source rule-based machine translation platform for low-resource languages. *Machine Translation*, 35(4):475–502.

Seungjun Lee, Jungseob Lee, Hyeonseok Moon, Chanjun Park, Jaehyung Seo, Sugyeong Eo, Seonmin Koo, and Heuiseok Lim. 2023. A survey on evaluation metrics for machine translation. *Mathematics*, 11(4):1006.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Dimitar Shterionov and Eva Vanmassenhove. 2023. *The Ecological Footprint of Neural Machine Translation Systems*, volume 4, pages 185–213. Springer Nature Switzerland AG, Switzerland. 25 pages, 3 figures, 10 tables Copyright © 2023, The Author(s), under exclusive license to Springer Nature Switzerland AG.

Universidad de Oviedo. 2010. Traductor automático castellano-asturiano-castellano algunos datos.

Xulio Viejo, Roser Saurı, and Angel Neira. 2008. Eslema. towards a corpus for asturian. In *Collaboration: Interoperability between people in the creation of language resources for less-resourced languages. A SALTMIL workshop*.