"We Are (Language) Family": Adapting Transformer models to related minority languages with linguistic data

Miguel López-Otal and Jorge Gracia
Aragon Institute of Engineering Research, Universidad de Zaragoza
mlopezotal@unizar.es, jogracia@unizar.es

Abstract

Transformer-based language models, despite their widespread use, remain mostly unavailable for low-resourced languages (LRLs), due to their lack of texts for pre-training. While solutions have emerged to remedy this, they still almost exclusively rely on raw text corpora, which may be almost non-existent for some languages. A recent line of work has attempted to circumvent this by replacing these with linguistics-based materials, such as grammars, to adapt LRLs to these models. However, many approaches tend to work with languages that are typologically very distant to each other.

In this work we investigate whether adapting closely related languages, belonging to the same family, with linguistics-based data can facilitate this process. For this, we look into the adaptation of two Spanish-based Transformer encoders –a monolingual and multilingual models– to Aragonese, a low-resourced Romance language spoken in Northern Spain, with whom it shares similar syntax but differing lexical and morphological phenomena. We rely on several knowledge injection methods, with which we report results, for a monolingual model, above some baselines in a set of Natural Language Understanding (NLU) benchmarks, proving the efficiency of relying on linguistics materials –or combined with a small amount of text– when languages belong to the same family.

1 Introduction

Transformer-based models (Vaswani et al., 2017) haven risen in popularity over the last years, replacing many dedicated solutions for downstream NLP tasks, such as rule-based systems or alternative neural architectures –e.g. LSTM (Yu et al., 2019). Most Transformer-based models, however, are only available for majority languages such as English or Spanish, with limited or no support for low-resourced languages (LRLs). This a consequence of the design of the Transformer architecture: for pre-training, these models rely on large amounts of text corpora, which is unaffordable for low-resourced languages (LRLs). To counteract this, many approaches have emerged to adapt existing Transformer-based models to LRLs –e.g. via continual pre-training (CPT), such as in Fujii et al. (2024)–, but most methods still almost exclusively rely on raw text corpora, which may be unavailable in large enough quantities for many languages. Meanwhile, they neglect the availability of other linguistic descriptions of these, such as grammars, dictionaries or lexicons –which include rich information about the language, complementary to raw text corpora. As a result, a recent line of work (Gurgurov et al., 2024, 2025; Tanzer et al., 2023; Zhang et al., 2024; Ramos et al., 2025) has explored potential ways to leverage this type of data for this, relying on methods such as knowledge injection (Hu et al., 2023; Yang et al., 2024a,b) or in-context prompting. This, however, still remains a relatively underexplored area.

In the context of this paper we lean into this line of research, by using linguistic resources, created by linguists for the study of LRLs, to add support for such languages to existing Transformer-based models. We distance ourselves, however, from existing works that tend to adapt models to typologically-distant languages[1] and instead look onto adaptation between closely-related languages, specifically belonging to the same family –e.g. Romance. A language family is a group of related tongues that, owing to their shared linguistic and/or geographical origin, present similar traits between them, such as conflated syntax or morphological phenomena. We hypothesize that the usage of linguistic resources, in contexts where i) there is very limited text corpora in an LRL, and ii) the lan-

[1]For example, English –Germanic– to Kalamang –Trans–New Guinea language (Tanzer et al., 2023)

guages are related to each other and belong to the same family, can prove an effective method for model adaptation between them. The influence of these shared linguistic traits, when adapting models via linguistic resources, has been seldomly analyzed in the literature except for a few cases, such as Senel et al. (2024); Bafna et al. (2025).

For this work, we attempt to adapt two encoder-based Transformer models trained in Spanish –`roberta-base-bne`, a monolingual model (Gutiérrez-Fandiño et al., 2021)[2], and `mRoBERTa`, a multilingual model[3]– to Aragonese, a low-resourced language from the same family –Romance– spoken by less than 50,000 speakers in Northern Spain. The standardized variant of Aragonese is selected for our research. This language pair is chosen given their shared similarities: while Aragonese presents noticeable lexical and morphological differences compared to Spanish, they also have an overall similar syntax. Moreover, while Aragonese has a growing corpus of texts –including a sizable monolingual Wikipedia–, its size can still prove insufficient for use in existing adaptation techniques. Additionally, as with many LRLs, Aragonese also has many curated traditional linguistic resources, such as dictionaries or grammars (de Filología Aragonesa, 2021).

We first present a custom linguistic resource, an inflected lexicon of Aragonese and Spanish, meant to help bridge the differences between the two languages. An inflected lexicon is a special type of lexicon that contains not only lemma-based forms of a language, but also all of its inflected forms, alongside their respective Part-of-Speech information.

We then investigate the use of two adapter-based knowledge injection methods, Lauscher et al. (2020) and K-Adapter (Wang et al., 2021), to integrate this lexical and morphological knowledge into existing Spanish models.[4] We present the results of injecting on these models both lemma-based and morphological knowledge separately, as well as combined, under different settings. We also replicate the methodology presented by Gurgurov et al. (2024), which combines injection of linguistic knowledge –in their case, of semantics– with post-training on raw text from an LRL, inspired by the MAD-X framework (Pfeiffer et al., 2020b).

We test our adapted models against a series of silver-standard Aragonese Natural Language Understanding (NLU) evaluation benchmarks[5], which have been machine-translated from a series of corresponding Spanish datasets (Baucells et al., 2025). Overall, we show performance gains over a series of baselines –including a re-training of the state-of-the-art MAD-X (Pfeiffer et al., 2020b) solution in Aragonese– for one of our tested models –the monolingual Spanish model–, demonstrating the effectiveness of using linguistically-based information for adapting language models to LRLs.[6]

In the remainder of this paper, we first present a series of prior research work similar to our proposal, followed by a description of the set of linguistic materials we have used for our experiments. Afterwards, we describe the knowledge injection solutions used for integrating linguistic information into our researched models, alongside information on how these are trained, and we finally report the obtained results with these methods as well as providing a series of concluding remarks.

2 Related work

In this section we present some work lines similar to our proposals, including the development of specific language technologies for Aragonese, common techniques for model adaptation to LRLs and the specific line of research that relies on linguistic resources for the latter objective.

2.1 NLP tools for Aragonese

As a minority language, Aragonese has not had a major support in the NLP community, unlike other low-resourced Iberian Romance languages such as Catalan or Galician. For instance, while the latter has had efforts to create decoder-based Transformer

[2]Originally located in `https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne`, since removed, we use an archived version instead.

[3]This model was pre-trained on a large corpus of English, Spanish, Catalan and Galician (Gonzalez-Agirre et al., 2025), among other languages, and is located at `https://huggingface.co/BSC-LT/mRoBERTa`

[4]This approach is most similar in nature to that of Gurgurov et al. (2024), which also relies on PEFT-based knowledge injection for LRL model adaptation. Contrary to their methods, we focus on a different type of level of linguistic knowledge –morphology and syntax, as opposed to semantics– and we investigate the effects of relying on closely-related languages for this process.

[5]These datasets are expected to undergo a manual revision effort by speakers of Aragonese, in order to ensure their correctness. The final versions of these datasets are to be released in an upcoming publication.

[6]We release all our resources –training, test data and evaluation scripts, as well as prototype inflected lexicon resource– on the following repository: `https://github.com/sid-unizar/lrl-transformer-adaptation`.

models (Gonzalez-Agirre et al., 2025; Gamallo et al., 2024), these do not exist for Aragonese. Furthermore, no dedicated NLP tools, such as PoS taggers, exist for Aragonese. Instead, most focus for this language has been centered in the development of Machine Translation tools, specially from Spanish to Aragonese. One example is the Aragonese-Spanish translation module in the Apertium framework (Forcada et al., 2011), created by native speakers of the language. Recently, efforts have also been devoted to the development of neural-based machine translation tools for Aragonese (Sánchez-Martínez et al., 2024; González, 2024), with the latter research also leading to the creation of a sole Aragonese-based language benchmark (Pérez-Ortiz et al., 2024), extended from the FLORES-200 (Costa-Jussà et al., 2022) MT benchmark.

2.2 Cross-lingual transfer and LRL model adaptation

Traditionally, there has been a steady line of work that has studied cross-lingual transfer of specific downstream tasks –e.g. named entity recognition (Johnson et al., 2019)– between majority and minority languages. These works have mainly been based on adapting existing task-oriented models, trained in one majority language, to perform that same task in a minority language, most times with little to no language-specific data. Some of these works rely on non-Transformer architectures, while others use Transformer models fine-tuned for it. The main limitation behind these works is that, while showing effective cross-lingual transfer results, they are mostly oriented towards a single specific task, not necessarily generalizable to other tasks. Meanwhile, a recent line of work has looked instead into directly adapting base, pre-trained Transformer models to LRLs, before downstream task fine-tuning. This is based on the capabilities of these models to adapt to many different types of tasks (either via fine-tuning or prompting). These works perform their adaptation on either pre-trained models, prior to any task modification (both encoders and decoder models), or on decoder models that have been post-trained for instruction following. One of the most common methods used for decoder-based models is continual pre-training (CPT), used to create many Transformer base models for other Iberian languages such as Catalan (Gonzalez-Agirre et al., 2025), Basque (Etxaniz et al., 2024; Corral et al., 2025) and Galician (Gamallo et al., 2024). This method consists of leveraging a pre-trained base model, trained in a majority language –e.g. English–, and continue training it with a corpus consisting of texts of an LRL combined with those of the language(s) the model was originally trained on. This allows the model to adapt to an LRL while keeping its original capabilities intact. This solution, however, still requires a moderately-sized amount of text in an LRL, which may not be available for many languages. Another common solution also includes the adaptation of the original tokenizer of a model with one trained specifically for an LRL (Artetxe et al., 2020b), as well as synthetically augmenting pre-training texts via machine translation –e.g. Boughorbel et al. (2024)–, or the use of models with scaled-down architectures –e.g. in the number of layers and attention heads, such as MicroBERT (Gessler and Zeldes, 2022)– to accommodate the low volume of available text. Another strategy for adaptation, in cases where there is less but still available corpora, is the use of Language Adapters, a special type of adapter (Houlsby et al., 2019; Pfeiffer et al., 2021) trained on text from an LRL, and which is accompanied by another adapter specifically trained for a downstream task. This solution was introduced by Pfeiffer et al. (2020b), and further refined by other frameworks (Parović et al., 2022; Chronopoulou et al., 2023). The Language Adapter solution offers a Parameter Efficient Fine Tuning (PEFT) method for language adaptation, which is advantageous as language models grow in size. Some solutions explore alternative PEFT methods to adapters, such as the use of LoRA or soft prompts (Vykopal et al., 2025).

2.3 Use of linguistic resources for model adaptation

In cases where there is little to no general-purpose text corpora available for a language, there is a recent line of work that has attempted to make use of linguistic materials available for the study of such language. Generally, these methods are mostly geared towards translation tasks (Tanzer et al., 2023; Merx et al., 2024; Zhang et al., 2025), although other works attempt to apply it to other downstream tasks (Zhang et al., 2024). Most of these methods rely on instruction-based models, as they require advanced instruction following capabilities to understand their provided sources. In most cases, these consist of raw texts explaining grammatical phenomena in an LRL –such as grammars (Tanzer et al., 2023; Zhang et al., 2025) or

language learning materials (Merx et al., 2024). These grammar texts are generally provided to a model within the contents of a prompt, which in many cases forbids the inclusion of an entire grammar book at once in a single prompt. This is usually solved by designing external solutions –akin to Retrieval Augmented Generation (RAG)– that strategically select only dedicated grammar book chapters to specific linguistic phenomena present in an input text. A dedicated RAG implementation, for instance, is in use by a system built for Quechua (Chen et al., 2024). Other studies provide other types of linguistic information alongside grammar book excerpts, such as PoS tags (Zhang et al., 2024), or even choose to provide parallel sentences but no grammar books at all –questioning the overall effectiveness of the latter for enhanced results (Pei et al., 2025). While most of the aforementioned studies adapt languages that are typologically distant to each other, there is one work (Bafna et al., 2025) that, similarly to ours, uses linguistics-based information to adapt languages belonging to the same family –including experiments in families such as Romance and Turkic. Their approach, however, does not use linguistic resources *per se* for training, but rather indirectly rely on them to create synthetic training texts imitating those of a target language. Other than instruction-based models, a recent line of work (Gurgurov et al., 2024, 2025) has also investigated the addition of semantics-based information from ConceptNet –a large-scale semantics network available for several LRLs– for both encoder and decoder-based models to perform LRL adaptation, alongside a Language Adapter module trained in a minority language. These solutions are based on the use of PEFT-based methods such as adapters or LoRA modules. Our work shares many similarities with the aforementioned works, but differs in the use of a different type of linguistic data –lexical and morphological information– as well as on focusing on the effects of studying the effectiveness of these solutions on related languages.

3 Linguistic Data

We present two sources of linguistic-specific data, used for training and evaluating our adapted models respectively. The first one is an inflected Aragonese-Spanish lexicon, serving as the main training dataset for LRL adaptation, and the second are a series of machine-translated NLU benchmarks in Aragonese, used for testing the effectiveness of the proposed methods.

3.1 Inflected lexicon

As our main source of data for knowledge injection, we set out to create a bilingual Spanish-Aragonese inflected lexicon. An inflected lexicon encompasses both lexical and morphological level information of a language, comprising all possible inflections of every word in a language alongside their Part-Of-Speech tags. Considering the linguistic features of Aragonese, in comparison to Spanish –i.e. similar syntax but different morphology and lexical content–, we consider this resource to represent a good approach for this specific case.

The resource used for the context of this publication is based on the language descriptions of Aragonese and Spanish provided by Apertium (Forcada et al., 2011), a rule-based translation system which includes direct translation between texts in Spanish and Aragonese. Our resource consists of a series of TSV files, each for each individual word category, which contain a series of Aragonese forms, alongside their PoS tags, and their corresponding forms in Spanish. These forms have been extracted from a series of source Apertium XML files, authored by linguists, used for guiding the steps in the custom Apertium translation pipeline. Our inflected lexicon contains a large number of surface forms, as listed in Table 1, which provide a rich source of language-based information for our injection experiments. However, it still lacks representation for inflected forms in some categories –specifically articles– due to technical difficulties in representing and extracting them, and as such is meant to be improved in future releases. The used representation system is also bound to change for final release, but serves our purposes for the context of this publication. This source of information is later converted to a series of tailor-suited training datasets designed for each of the knowledge injection methods described in Section 4.

3.2 Raw Aragonese text corpora

Some experiments in our research also compare existing methods that either only rely on LRL text alone (Pfeiffer et al., 2020b) or in a combination of both linguistic and raw text data (Gurgurov et al., 2024). As such, we also leverage an Aragonese corpus of text from a monolingual

Category	No. of lemmas	No. of inflected forms
Verbs	2,056	94,118
Nouns	6,494	13,834
Adjectives	4,093	27,576
Determiners	20	-
Indeterminate articles	25	-
Pronouns	178	
Adverbs	3,648	
Numerals	118	
Prepositions	232	
Interjections	40	
Conjunctions	124	

Table 1: Number of Aragonese word forms present in our inflected lexicon. Each word form has its corresponding Spanish translated form. Inflected forms for Articles (determiners and indeterminate articles) are planned to be added in future releases of the lexicon.

version of Wikipedia for that language.[7] Other Aragonese-based text materials could not be located in time due to their sparsity –e.g. being found only in printed materials–, use of outdated orthographic standards and/or an active copyright status.

3.3 Evaluation benchmarks

The non-existence of Aragonese NLU benchmarks –except for a recently released subset of the FLORES-200 translation benchmark (Pérez-Ortiz et al., 2024)– hinders the development of appropriate language tools for this language. As such, we set out to create a set of preliminary language benchmarks for Aragonese, in order to test the abilities of Transformer models in this language.

The chosen datasets were machine translated from existing, established Spanish benchmarks (Baucells et al., 2025), relying on the Apertium framework (Forcada et al., 2011). The choice of benchmarks was motivated by several factors, including the availability of corresponding human-curated datasets in Spanish –due to the existing tools for MT only being targeted for Spanish-Aragonese translation– and whether they tested linguistics-related phenomena. These datasets are meant to undergo a manual revision effort in a later stage by native Aragonese speakers, ensuring an upcoming release of a series of gold-standard Aragonese NLU benchmarks. For the purpose of our experiments, we settled on the following three silver-standard evaluation datasets:

- COPA (Corpus of Plausible Alternatives): a translation of the Spanish COPA dataset[8], itself adapted from the original COPA benchmark (Roemmele et al., 2011). Despite its small size (600 instances), we have divided it into a training (500 samples) and test sets (100 samples).
- XNLI (Cross-Lingual Natural Language Inference): a common task in NLP, we deviate from the use of the XNLI dataset (Conneau et al., 2018) –the standard multilingual NLI benchmark, used by research such as Gonzalez-Agirre et al. (2025)– due to, i) some inaccuracies in the Spanish version, that could lead to poor translation quality in Aragonese, and ii) its big dataset size –around 400,000 entries, which would make it unsuitable for a manual correction effort later on. We instead adopt an alternative human-curated Spanish NLI dataset, ES-XNLI (Artetxe et al., 2020a), created for testing biases in models trained with machine-translated versions of NLI datasets.[9]
- PAWS-X (Paraphrase Adversaries from Word Scrambling): Aragonese version of the Spanish subset of the PAWS-X (Yang et al., 2019) dataset, the multilingual version of the PAWS (Zhang et al., 2019) benchmark. We follow this distribution: 49,401 entries for training, 1,999 for development and 2,000 for test.

While none of the aforementioned datasets test linguistic phenomena per se, they still require deep NLU skills in a target language, thus serving for our experiments. Additionally, since these datasets are also available in Spanish, cross-lingual transfer experiments from models trained in Spanish to Aragonese can also be performed.

4 Methodology

We rely on two knowledge injection methods for model adaptation, Lauscher et al. (2020) and K-Adapter (Wang et al., 2021), modified or updated for our specific uses. In our configuration, we decide to separate lexical information (lemmas) from

[7] https://an.wikipedia.org/

[8] https://huggingface.co/datasets/BSC-LT/COPA-es

[9] While this specific dataset, with 2,490 entries, was originally designed as a development set replacement for XNLI (Conneau et al., 2018) in downstream tasks, we have instead repurposed it here as a full-fledged dataset divided in dedicated train (1,990 entries) and test sets (500 entries).

morphological data (inflected forms), by training dedicated adapters, each with their own custom training task. We performed these experiments to test the effect of only providing lexical information to a model compared to additional morphological data, as well as to test out the effectiveness of using different injection strategies for each type of information.

4.1 Lauscher et al. (2020)

Lauscher et al. (2020) is a knowledge injection method that trains adapter modules[10] on raw text corpora with a MLM (Masked Language Modeling) or CLM (Causal Language Modeling) objective. This corpus consists of a series of triples from a structured knowledge source that are verbalized using a series of predefined templates. We replicated their experiments here for linguistic information by performing a verbalization effort of our source inflectional lexicon, done using Spanish-based templates.

We trained adapters with lemma-only information exclusively, reserving morphological information for another type of knowledge injection (see section 4.2). For lemma-only approaches, we first followed a simple verbalization strategy, consisting of basic natural language templates, featured in Appendix A.1.

The former approach, however, was limited, due to it only encompassing one possible tokenization of each word pair; as a result, we also underwent an additional, more complex verbalization strategy for our lemma-only approach. We devised a total of 18 different verbalization templates –featured in Appendix A.2–, with each lemma pair appearing for each one of the provided templates. The latter ensured that each word pair appeared more than once in the training set, hence ensuring broad coverage.

All adapters were trained for 10 epochs with a learning rate of 1e-4.

4.2 K-Adapter

K-Adapter (Wang et al., 2021) is a knowledge injection framework that inserts two types of adapters, depending on the source of information –factual or linguistic–, inside Transformer-based models. Contrary to Lauscher et al. (2020), these adapters are not trained with a MLM or CLM objective on raw text, but follow custom training tasks based on the type of injected information:

- Factual information: relying on a text corpus of Wikipedia abstracts containing pairs of named entities, linked together via a shared relationship, the adapter is trained on predicting the relation between the two entities (relation classification task).
- Linguistic information: a series of sentences, alongside their parse trees, are fed into an adapter, which is trained to deduce the correct head of specific words.

Despite our focus on improving linguistic performance in models, we do not use the linguistic information approach, since it would require an expensive manual acquisition process of parse trees for a set of Aragonese texts. We instead rely on a variation of the factual information approach, in which we treat lexicon data as a type of factual information. We use this knowledge injection method exclusively for morphologically inflected words, specifically for three types of words: nouns, adjectives and simple verb forms. Nouns and adjectives are trained in one adapter, whereas verbs are trained in a separate adapter, due to inconsistent PoS tags between those word categories. Regarding verbs, only simple tenses are included, excluding composite forms –e.g. "heba cantau". This decision, while linguistically controversial, is made to facilitate cross-lingual transfer in our experiments, since i) composite tenses are the same across Spanish and Aragonese and, ii) they systematically consist of a combination of an inflected form, in a *simple* tense, of the auxiliary verb "haber" with a past participle of a main verb.

We report the results of combining our lemma-only adapters with each of the two morphological adapters separately, as well as combined. Among these configurations, we do not test the Morphological Adapters in isolation, due to their lower coverage of word forms (only nouns, adjectives and verbs) compared to the lemma-based adapter. Our datasets built for training this type of adapters follow a similar format to those of Lauscher et al. (2020): they contain multiple verbalized sentences from our source inflectional lexicon TSV files –relying on several templates for exhaustiveness, featured in Appendix B–, but within each sentence we assign a label, representing a PoS tag, to the inflected word forms mentioned in those sentences. As such, we repurpose the relation classification

[10] They originally used adapters proposed by Houlsby et al. (2019), but we experiment with those of Pfeiffer et al. (2021) instead

task, as proposed originally by Wang et al. (2021), as a labeling task, intended to guess the correct PoS tag for each form.

4.3 Training settings and baselines

Our linguistic adapters (lemma-based and morphological) are added to a model alongside an additional, untrained adapter module, called a 'Task Adapter', following the methodology introduced by Pfeiffer et al. (2020b). A custom task-specific layer is also added on top of the model. The task adapters, alongside the task-specific layers, are trained when adapting the models to a downstream task, while keeping the rest of the model and the other linguistic adapters frozen. We test the following combinations of adapter configurations:

- Lemma (Simple||Multiple) + *Task Adapter*
- Lemma (Multiple)+Verbs (K-Adapter)+*T.A.*
- Lemma (Multiple)+Nouns-Adjectives+*T.A.*
- Lemma (Multiple)+Nouns-Adjectives (K-Adapter)+Verbs (K-Adapter) + *T.A.*

All of our configurations are trained in each of the provided NLU datasets –COPA, XNLI and PAWS-X–, averaged across three independent runs, relying on the Adapterhub (Pfeiffer et al., 2020a) library. We test the combination of the different adapters (linguistic and task adapters) with an AdapterFusion setting (Pfeiffer et al., 2021), following a similar implementation by Gurgurov et al. (2024). We choose a learning rate for the task adapters of 5e-5 for 5 epochs. These configurations are tested on both the monolingual `roberta-base-bne` model as well as on the multilingual `mRoBERTa` model. As baselines, we also provide the results of testing the latter models, without any LRL adaptation, directly in our Aragonese benchmarks, using a Task Adapter setting. This experiment is performed by i) training the models in Spanish data and testing them in Aragonese test data, and ii) training and testing the models exclusively in Aragonese data. As an additional baseline, we also replicate the cross-lingual transfer strategy proposed by MAD-X (Pfeiffer et al., 2020b), with a Language Adapter trained in Aragonese Wikipedia data and tested with a similar configuration, albeit with some differences: whereas the original paper, during training, trained with data and a language adapter from a majority language (e.g. English) –later swapping them out at test time for those of an LRL–, we train and test with data and language adapters from Aragonese exclusively. Finally, we also replicate the experiments performed by Gurgurov et al. (2024), with our own source of information, and test our models combining our Aragonese text-based Language Adapter with our lemma-based adapters. For all these baselines, we follow the same hyperparameters as the ones used in our proposed configurations.

5 Results

The obtained results for our experiments, as well as a series of baselines, are reported in Table 2. We should note that the reliance on Apertium-based data for both training –inflectional lexicon– and test –MT-based benchmarks– may have biased the results. This is a limitation that is meant to be solved in future work via manual revision of the evaluation benchmarks.

We first address the results for `roberta-base-bne`. The metrics for the lexical adapters using the simple verbalization strategies are not very positive compared to the baselines, however the lexical adapters that rely on multiple verbalizations show, for `roberta-bne`, in many cases, results well above the reported baselines. For COPA and XNLI, the results are better than those reported for the baseline by Gurgurov et al. (2024) –whereas PAWS-X, oppositely, shows better results in that baseline. This is perhaps explained by the nature of the datasets: whereas COPA and XNLI are mostly lexical, PAWS-X –as a paraphrase detection dataset– is more syntactical in nature, something that is better addressed by a language adapter, trained in multiple real-life Aragonese sentences, as used by Gurgurov et al. (2024).

We should note that, while both COPA and PAWS-X benchmark provide good results for this model, the COPA results can be unreliable due to its small test size, hence limiting potential claims. In the case of XNLI, although the baseline trained in Spanish and tested in Aragonese still shows the highest results, our solutions still display metrics that are above another baseline in which the model was trained with Aragonese data exclusively –as well as the MAD-X (Pfeiffer et al., 2020b) and the Gurgurov et al. (2024) setting–, hence indicating that the injected lemma-based data –and/or LRL text– may still have a positive impact in the results for this specific benchmark. Additionally, the relatively high standard deviation rate of the XNLI

Model	Lexical Adapter (Lauscher)	Morph. Adapter (K-Adapter)	COPA	XNLI	PAWS-X
roberta-BNE	**Baseline:**train ES,test AN		0.526 ± 0.057	**0.413 ± 0.066**	0.7059 ± 0.015
mRoBERTa			0.48 ± 0.045	**0.3973 ± 0.023**	0.8075 ± 0.007
roberta-BNE	**Baseline:**train+test AN		0.516 ± 0.049	0.349 ± 0.004	0.6978 ± 0.007
mRoBERTa			0.51 ± 0.115	0.38 ± 0.004	**0.8196 ± 0.002**
roberta-BNE	**Baseline:** Pfeiffer et al., 2020b		0.49 ± 0.036	0.36 ± 0.01	0.682 ± 0.015
mRoBERTa			0.513 ± 0.08	0.348 ± 0.034	0.697 ± 0.006
roberta-BNE	**Baseline:** Gurgurov et al., 2024		0.56 ± 0.085	0.366 ± 0.018	**0.747 ± 0.012**
mRoBERTa	Multiple Lemma+Lang. Adapter		**0.54 ± 0.026**	0.375 ± 0.025	0.751 ± 0.029
roberta-BNE	Simple	-	0.483 ± 0.04	0.358 ± 0.019	0.67 ± 0.004
mRoBERTa	(AN-ES)		0.49 ± 0.043	0.3226 ± 0.013	0.6525 ± 0.025
roberta-bne	Simple	-	0.536 ± 0.04	0.358 ± 0.012	0.673 ± 0.012
mRoBERTa	(ES-AN)		0.49 ± 0.045	0.338 ± 0.003	0.6073 ± 0.041
roberta-BNE	Multiple	-	**0.59 ± 0.01**	0.3713 ± 0.01	0.728 ± 0.006
mRoBERTa			0.536 ± 0.055	0.339 ± 0.022	0.779 ± 0.008
roberta-BNE	Multiple	Verbs	0.54 ± 0.034	0.378 ± 0.004	0.743 ± 0.009
mRoBERTa			0.526 ± 0.06	0.377 ± 0.063	0.79 ± 0.004
roberta-BNE	Multiple	Nouns-Adjs	0.559 ± 0.045	0.362 ± 0.007	0.738 ± 0.005
mRoBERTa			0.516 ± 0.032	0.328 ± 0.003	0.79 ± 0.002
roberta-BNE	Multiple	Verbs+Nouns-Adjs	0.556 ± 0.005	0.366 ± 0.009	0.742 ± 0.003
mRoBERTa			0.5 ± 0.02	0.345 ± 0.009	0.791 ± 0.002

Table 2: Results of a series of baselines and our proposed adaptation methods on a series of NLU benchmarks. COPA and XNLI results are reported in Accuracy metric, while PAWS-X is reported in F1.

results in the Spanish-Aragonese baseline (0.066), compared to those of the other settings for the same model (which stand at between 0.004 and 0.019), may hint at a much more variable performance in that specific cross-lingual setting, hence making its results potentially more unstable.

For `roberta-base-bne`, the combination of lexical adapters with morphological adapters provides very little improvements in the obtained results; more specifically, the results are even worse in some cases compared to the use of a lexical adapter alone –e.g. in some of the settings of the XNLI benchmark–, albeit still show improvements compared to non-adapted baselines. The exception is PAWS-X, that manages to achieve in this setting results not only above the baselines but also comparable to Gurgurov et al. (2024) without the use of a language adapter, which is a remarkable achievement. Overall, the worse results of these morphological adapters could hint at limitations of the K-Adapter method for learning morphological information, or at the fact that COPA and XNLI benefit from lemma information exclusively, with no need for providing all inflected forms. Meanwhile, while not being an outright replacement for text corpora from a language, PAWS-X manages to achieve comparable results with linguistic data alone, which by itself is a notable finding.

`mRoBERTa`, on the other hand, shows worse results in our solutions compared to baselines, except for small gains in the COPA benchmark –which, nevertheless, may not be reliable, as stated prior, owing to its lower dataset size. In the best performing scenarios, PAWS-X manages to achieve a performance comparable to the unadapted baselines, but still falls below it. Overall, we attribute the different performance between the multilingual and the monolingual models to the fact that `mRoBERTa`

was pre-trained on a corpus containing not only Spanish but several others Romance languages, including Catalan, Galician and French. Incidentally, the latter languages present some linguistic features that Aragonese possesses –i.e. similar lexical content– but Spanish does not. Interpretability works on the linguistic abilities of encoder-based Transformer models (López-Otal et al., 2025) have demonstrated that multilingual models may come to encode cross-lingual phenomena in shared representations, which could potentially help generalization to unseen languages with similar traits –in this case, Aragonese– and thus lead to the observed results in these baseline models. In this sense, it remains to be investigated whether this implies that the proposed solutions can only show improvements for monolingual models trained in a related language, or whether this is the result of experimenting with this specific multilingual model that shows this distribution of languages. Regarding the latter point, it could be tested whether relying on a different multilingual model, including Spanish but not the other mentioned languages, could lead to a different set of results. Unfortunately, this could not be tested due to time constraints.

6 Conclusions

Through our provided solutions we aim to boost low-resourced languages NLU skills in Transformer-based encoders trained in languages from the same family, via the use of linguistic resources as either a replacement or additional support for raw text. Through a series of experiments in Aragonese, a language closely related to Spanish, we observe that simple knowledge injection of morphological and lexical information –the main differing traits between the two languages– can slightly improve performance in a series of Aragonese NLU benchmarks for a monolingual Spanish model, although results are not as positive for a tested multilingual model –which we hypothesize is due to its specific training on multiple Romance languages that bear some similarities to Aragonese that Spanish, in itself, does not have. For the monolingual model, while performance improves with the use of an additional adapter trained with text from the target language, results are still competitive with linguistic information alone.

While this could open the door in the future to potential corpora-less adaptation efforts of Transformer-based language models, specifically if the languages to be adapted are related, this line of research still needs to be further validated under different configurations and setups. In this sense, future work should focus on the use of alternative methodologies for LRL adaptation as well as reporting on a wider array of NLU benchmarks. Additionally, the use of machine-translated data for testing is not ideal, as it may have impacted the end results. This is meant to be solved in the future via a manual curation effort of these resources. The small size of one of the tested benchmarks, COPA, may have also led to unstable results.

Future work should include experiments in decoder-based language models –a target preliminary analyzed by Gurgurov et al. (2025)–, due to their increased popularity, with the usage of either prompt-based analyses –which is the majority direction for most studies relying on grammar resources for these models (e.g. Tanzer et al., 2023; Zhang et al., 2024)– as well as fine-tuning based strategies. Experiments on decoder-based models should also include evaluations on Natural Language Generation (NLG) tasks, in order to test the effectiveness of LRL adaptation in this dimension.

Limitations

This work relies on a shared pool of linguistic data for both training and evaluation: linguistic data originating from the Apertium (Forcada et al., 2011) machine translation framework. This could have had some biases in the provided results. This problem is meant to be solved in a later stage with a manual correction effort of the Aragonese MT-based evaluation benchmarks.

Additionally, the usage of the COPA dataset for training and test in our configuration is not ideal, as the dataset is generally used as a test set (Ponti et al., 2020) and training is usually performed with other existing, similar datasets, such as the English-based SIQA dataset (Sap et al., 2019).[11] We opted to rely on this uncommon training setting, nevertheless, due to the time constraints.[12] Nevertheless, we recognize that the limited size of the test set may have led to unreliable results for this specific benchmark.

[11] For instance, this is a choice made by Senel et al. (2024), which also report results on the COPA task.

[12] While there is a recent release of a Spanish-based version of the SIQA development set –located at https://huggingface.co/datasets/BSC-LT/SIQA_es–, it was released too late to incorporate it into our current analysis.

Acknowledgments

We would like to appreciate the help and feedback received from fellow Aragonese speakers, including Juan Pablo Martínez Cortés and José Ignacio López Susín. This work was supported by the research project PID2024-159530OB-I00, funded by MICIU /AEI /10.13039/501100011033 / FEDER, UE. It has also been partially funded by DGA Government predoctoral fellowship.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020a. Translation artifacts in cross-lingual transfer learning. *arXiv preprint arXiv:2004.04721*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020b. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4623–4637.

Niyati Bafna, Emily Chang, Nathaniel Romney Robinson, David R. Mortensen, Kenton Murray, David Yarowsky, and Hale Sirin. 2025. DialUp! modeling the language continuum by adapting models to dialects and dialects to models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20188–20233, Vienna, Austria. Association for Computational Linguistics.

Irene Baucells, Javier Aula-Blasco, Iria de Dios-Flores, Silvia Paniagua Suárez, Naiara Perez, Anna Salles, Susana Sotelo Docio, Júlia Falcão, José Javier Saiz, Robiert Sepúlveda-Torres, and 1 others. 2025. Iberobench: A benchmark for llm evaluation in iberian languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10491–10519.

Sabri Boughorbel, Md Rizwan Parvez, and Majd Hawasly. 2024. Improving language models trained on translated data with continual pre-training and dictionary learning analysis. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 73–88, Bangkok, Thailand. Association for Computational Linguistics.

Junhao Chen, Peng Shu, Yiwei Li, Huaqin Zhao, Hanqi Jiang, Yi Pan, Yifan Zhou, Zhengliang Liu, Lewis C Howe, and Tianming Liu. 2024. Queen: A large language model for quechua-english translation. *Preprint*, arXiv:2412.05184.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2023. Language-family adapters for low-resource multilingual neural machine translation. In *Proceedings of the Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 59–72, Dubrovnik, Croatia. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2475–2485.

Ander Corral, Ixak Sarasua Antero, and Xabier Saralegi. 2025. Pipeline analysis for developing instruct LLMs in low-resource languages: A case study on Basque. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12636–12655, Albuquerque, New Mexico. Association for Computational Linguistics.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Estudio de Filología Aragonesa. 2021. *Gramática básica de l'aragonés*. Prensas de la Universidad de Zaragoza.

Julen Etxaniz, Oscar Sainz, Naiara Miguel, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. Latxa: An open language model and evaluation suite for basque. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14952–14972.

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.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. *Preprint*, arXiv:2404.17790.

Pablo Gamallo, Pablo Rodríguez, Iria de Dios-Flores, Susana Sotelo, Silvia Paniagua, Daniel Bardanca, José Ramom Pichel, and Marcos Garcia. 2024. Open generative large language models for galician. *arXiv preprint arXiv:2406.13893*.

Luke Gessler and Amir Zeldes. 2022. Microbert: Effective training of low-resource monolingual berts through parameter reduction and multitask learning. *arXiv preprint arXiv:2212.12510*.

Antoni Oliver González. 2024. Sistemas de traducción automática para el aragonés. *Luenga & fablas: publicazión añal de rechiras, treballos e decumentazión arredol de l'aragonés ea suya literatura*, (28):91–100.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, and 1 others. 2025. Salamandra technical report. *arXiv preprint arXiv:2502.08489*.

Daniil Gurgurov, Mareike Hartmann, and Simon Ostermann. 2024. Adapting multilingual llms to low-resource languages with knowledge graphs via adapters. *arXiv preprint arXiv:2407.01406*.

Daniil Gurgurov, Ivan Vykopal, Josef van Genabith, and Simon Ostermann. 2025. Small models, big impact: Efficient corpus and graph-based adaptation of small multilingual language models for low-resource languages. *arXiv preprint arXiv:2502.10140*.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Aitor Gonzalez-Agirre, Carme Armentano-Oller, Carlos Rodriguez-Penagos, and Marta Villegas. 2021. Maria: Spanish language models. *arXiv preprint arXiv:2107.07253*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. 2023. A survey of knowledge enhanced pre-trained language models. *Preprint*, arXiv:2211.05994.

Andrew Johnson, Penny Karanasou, Judith Gaspers, and Dietrich Klakow. 2019. Cross-lingual transfer learning for Japanese named entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 182–189, Minneapolis, Minnesota. Association for Computational Linguistics.

Anne Lauscher, Olga Majewska, Leonardo FR Ribeiro, Iryna Gurevych, Nikolai Rozanov, and Goran Glavaš. 2020. Common sense or world knowledge? investigating adapter-based knowledge injection into pretrained transformers. *arXiv preprint arXiv:2005.11787*.

Miguel López-Otal, Jorge Gracia, Jordi Bernad, Carlos Bobed, Lucía Pitarch-Ballesteros, and Emma Anglés-Herrero. 2025. Linguistic interpretability of transformer-based language models: a systematic review. *Preprint*, arXiv:2504.08001.

Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo, and Ekaterina Vylomova. 2024. Low-resource machine translation through retrieval-augmented llm prompting: A study on the mambai language. *arXiv preprint arXiv:2404.04809*.

Marinela Parović, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2022. Bad-x: Bilingual adapters improve zero-shot cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1791–1799.

Renhao Pei, Yihong Liu, Peiqin Lin, François Yvon, and Hinrich Schuetze. 2025. Understanding in-context machine translation for low-resource languages: A case study on Manchu. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8767–8788, Vienna, Austria. Association for Computational Linguistics.

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, and 1 others. 2024. Expanding the flores+ multilingual benchmark with translations for aragonese, aranese, asturian, and valencian. In *Proceedings of the Ninth Conference on Machine Translation*, pages 547–555.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume*, pages 487–503.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. *arXiv preprint arXiv:2005.00052*.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Rita Ramos, Everlyn Asiko Chimoto, Maartje Ter Hoeve, and Natalie Schluter. 2025. GrammaMT: Improving machine translation with grammar-informed

in-context learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29920–29940, Vienna, Austria. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pages 90–95.

Felipe Sánchez-Martínez, Juan Antonio Perez-Ortiz, Aaron Galiano Jimenez, and Antoni Oliver. 2024. Findings of the WMT 2024 shared task translation into low-resource languages of Spain: Blending rule-based and neural systems. In *Proceedings of the Ninth Conference on Machine Translation*, pages 684–698, Miami, Florida, USA. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Lütfi Kerem Senel, Benedikt Ebing, Konul Baghirova, Hinrich Schuetze, and Goran Glavaš. 2024. Kardeş-NLU: Transfer to low-resource languages with the help of a high-resource cousin – a benchmark and evaluation for Turkic languages. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1672–1688, St. Julian's, Malta. Association for Computational Linguistics.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2023. A benchmark for learning to translate a new language from one grammar book. *arXiv preprint arXiv:2309.16575*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ivan Vykopal, Simon Ostermann, and Marián Šimko. 2025. Soft language prompts for language transfer. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10294–10313.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuan-Jing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021. K-adapter: Infusing knowledge into pre-trained models with adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418.

Jian Yang, Xinyu Hu, Gang Xiao, and Yulong Shen. 2024a. A survey of knowledge enhanced pre-trained language models. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

Linyao Yang, Hongyang Chen, Zhao Li, Xiao Ding, and Xindong Wu. 2024b. Give us the facts: Enhancing large language models with knowledge graphs for fact-aware language modeling. *IEEE Trans. on Knowl. and Data Eng.*, 36(7):3091–3110.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. *arXiv preprint arXiv:1908.11828*.

Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. 2019. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270.

Chen Zhang, Jiuheng Lin, Xiao Liu, Zekai Zhang, and Yansong Feng. 2025. Read it in two steps: Translating extremely low-resource languages with code-augmented grammar books. *Preprint*, arXiv:2506.01796.

Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. Hire a linguist!: Learning endangered languages in llms with in-context linguistic descriptions. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15654–15669.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. *arXiv preprint arXiv:1904.01130*.

A Lemma-based Verbalizations Templates

In this Appendix we present the series of natural language templates that were used for verbalizing a series of Aragonese lemmas alongside their Spanish corresponding forms, as exposed in section 4.1. There is a distinction between "simple verbalizations", in which a simple template format was used –Appendix A.1, and "multiple verbalizations", where 18 different templates were devised for comprehensiveness –Appendix A.2. Please note that, for all templates, the variation in the forms, marked by a slash, between "el/la" and "un/una" reflect the gender of the word category noun in Spanish (e.g. "el sustantivo" against "la preposición") and are not part of the actual verbalization.

A.1 Simple templates

- Aragonese-Spanish: "*El/La* <word-category> <form in Aragonese> *es lo mismo que el*/la <word-category> <form in Spanish>"

- Spanish-Aragonese: "*El/La* <word-category> <form in Spanish> *es lo mismo que el/la* <word-category> <form in Aragonese>".

A.2 Multiple templates

1. *El/La* <word category> <Aragonese lemma> *es* <Spanish lemma>.
2. *En aragonés,* <Aragonese lemma> *es un/una* <word category> *que en español significa* <Spanish lemma>
3. *En aragonés,* <Aragonese lemma> *es un/una* <word category> *que en castellano significa* <Aragonese lemma>
4. <Aragonese lemma> *significa lo mismo que el/la* <word category> <Spanish lemma>
5. *el/la* <word category> <Aragonese lemma> *significa lo mismo que* <Spanish lemma>
6. *el/la* <word category> <Aragonese lemma> *significa lo mismo que* <Spanish lemma> *en español*
7. *el/la* <word category> <Aragonese lemma> *significa lo mismo que el/la* <word category> <Spanish lemma>
8. *el/la* <word category> <Aragonese lemma> *significa lo mismo que el/la* <word category> <Spanish lemma> *en español*
9. <Aragonese lemma> *quiere decir lo mismo que el/la* <word category> <Spanish lemma>
10. *El/La* <word category> <Aragonese lemma> *quiere decir lo mismo que* <Spanish lemma>
11. *El/La* <word category> <Aragonese lemma> *quiere decir lo mismo que* <Spanish lemma> *en español*
12. *El/La* <word category> <Aragonese lemma> *quiere decir lo mismo que el/la* <word category> <Spanish lemma>
13. *El* <word category> <Aragonese lemma> *quiere decir lo mismo que el/la* <word category> <Spanish lemma> *en español*
14. <Aragonese lemma> *es un/una* <word category> *en aragonés que quiere decir* <Spanish lemma> *en español*
15. <Aragonese lemma> *es un/una* <word category> *en aragonés que quiere decir* <Spanish lemma> *en castellano*
16. <Spanish lemma> *es un/una* <word category> *en español que quiere decir* <Aragonese lemma> *en aragonés*
17. <Spanish lemma> *es un/una* <word category> *en castellano que quiere decir* <Aragonese lemma> *en aragonés*
18. <Aragonese lemma> *es lo mismo que* <Spanish lemma>

B Morphological Information Verbalizations Templates and dataset format

For building our training datasets for the knowledge injection strategy used in section 4.2, we first relied on the following set of 6 templates in Spanish for verbalizing Aragonese inflected nouns, adjectives and verbs alongside their lemmas:

1. *<inflected form> viene del <word category> <lemma>*
2. *<inflected form> es del <word category> <lemma>*
3. *La forma <inflected form> viene del <word category> <lemma>*
4. *El <word category> <lemma> tiene la forma <inflected form>*
5. *En el <word category> <lemma> encontramos la forma <inflected form>*
6. *En el <word category> <lemma> tenemos <inflected form>*

With these verbalized sentences, distributed randomly between training, development and test sets, we then labeled each inflected form with the following set of labels, which follow the naming convention for PoS-tags established by the Apertium framework (Forcada et al., 2011):

- Nouns and Adjectives: `n_m_sg` (noun, masculine, singular), `n_f_sg` (noun, feminine, singular), `n_m_pl` (noun, masculine, plural), `n_f_pl` (noun, feminine, plural), `adj_m_sg` (adjective, masculine, singular), `adj_f_sg` (adjective, feminine, singular), `adj_m_pl` (adjective, masculine, plural), `adj_f_pl` (adjective, feminine, plural).

- Verbs (only simple tenses –see section 4.2):
  - Present Tense, Indicative (`pri`): `pri_p1_sg` (1st person, singular), `pri_p2_sg` (2nd p., sg.), `pri_p3_sg` (3rd p., sg.), `pri_p1_pl` (1st person, plural), `pri_p2_pl` (2nd p., pl.), `pri_p3_pl` (3rd p., pl.)
  - Imperfect Past Tense, Indicative (`pii`): `pii_p1_sg` (1st p., sg.), `pii_p2_sg` (2nd p., sg.), `pii_p3_sg` (3rd p., sg.), `pii_p1_pl` (1st p., pl.), `pii_p2_pl` (2nd p., pl.), `pii_p3_pl` (3rd p., pl.)
  - Perfect Past Tense, Indicative (`ifi`): `ifi_p1_sg` (1st p., sg.), `ifi_p2_sg` (2nd p., sg.), `ifi_p3_sg` (3rd p., sg.), `ifi_p1_pl` (1st p., pl.), `ifi_p2_pl` (2nd p., pl.), `ifi_p3_pl` (3rd p., pl.)
  - Future Tense, Indicative (`fti`): `fti_p1_sg` (1st p., sg.), `fti_p2_sg` (2nd p., sg.), `fti_p3_sg` (3rd p., sg.), `fti_p1_pl` (1st p., pl.), `fti_p2_pl` (2nd p., pl.), `fti_p3_pl` (3rd p., pl.)
  - Present Tense, Subjunctive (`prs`): `prs_p1_sg` (1st p., sg.), `prs_p2_sg` (2nd p., sg.), `prs_p3_sg` (3rd p., sg.), `prs_p1_pl` (1st p., pl.), `prs_p2_pl` (2nd p., pl.), `prs_p3_pl` (3rd p., pl.)
  - Imperfect Past, Subjunctive (`pis`): `pis_p1_sg` (1st p., sg.), `pis_p2_sg` (2nd p., sg.), `pis_p3_sg` (3rd p., sg.), `pis_p1_pl` (1st p., pl.), `pis_p2_pl` (2nd p., pl.), `pis_p3_pl` (3rd p., pl.)
  - Conditional (`cni`): `cni_p1_sg` (1st p., sg.), `cni_p2_sg` (2nd p., sg.), `cni_p3_sg` (3rd p., sg.), `cni_p1_pl` (1st p., pl.), `cni_p2_pl` (2nd p., pl.), `cni_p3_pl` (3rd p., pl.)
  - Imperative (`imp`): `imp_p2_sg` (2nd p., sg.), `imp_p3_sg` (3rd p., sg.), `imp_p1_pl` (1st p., pl.), `imp_p2_pl` (2nd p., pl.), `imp_p3_pl` (3rd p., pl.)

Additionally, lemmas are also labeled with their corresponding word category: `verb`, `noun` or `adjective`.