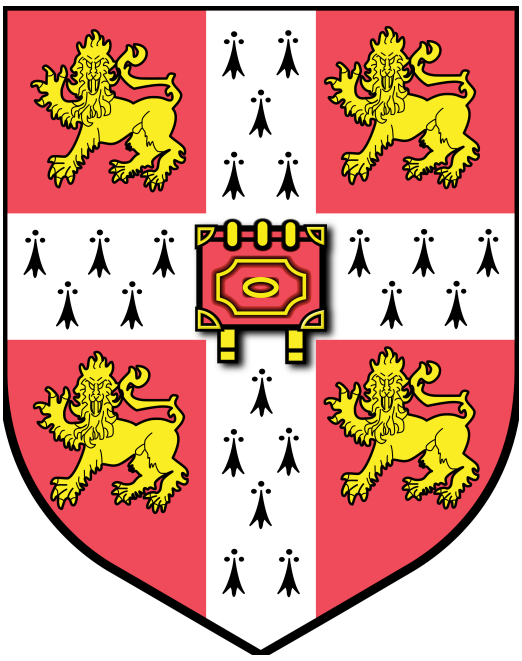
Exploring and Understanding Cross-Lingual Transfer to Catalan and Galician via High-Resource Typological Relatives

Oli Carlos Hepworth

Word Count: 10,034



University of Cambridge
Department of Theoretical and Applied Linguistics
Part IIB Dissertation

Contents

1 Introduction and Motivation

In recent years, machine learning has come to dominate the field of Natural Language Processing (NLP) – the application of computational techniques to language. The advent of powerful architectures like the transformer has seen language models rapidly increase in sophistication, and it seems likely that these technologies will become a major presence in our society [1]. There exists an increasingly relevant problem, however, when applying these techniques to the majority of the world's languages: a lack of resources [2].

1.1 The Problem With Low-Resource Languages

To be effective, language models require vast amounts of training data. For English and a handful of other high-resource languages, this is generally not a problem: high-quality labelled datasets exist for almost every relevant NLP task, and large unlabelled datasets can be scraped from the internet. However, this is certainly not the case for the majority of the world's languages, and it is very common for a language to lack task-specific datasets or even unlabelled corpora large enough to train a powerful model on [2]. Moreover, even when the relevant data does exist, it is often of a lower quality to that of high-resource languages [3].

Many solutions to this problem have been proposed, with one of the most prominent being cross-lingual transfer learning [4]. Under this paradigm, a language model is able to perform a task in a target language by leveraging knowledge learned in a different language, frequently one which is much richer in resources [5]. Many variations of cross-lingual transfer learning exist, spanning both single-source approaches (where knowledge from a single source language is transferred) and multi-source approaches (where knowledge from multiple source languages is transferred). With regards to the latter of these, there exists a debate within the field as to the best type of model to use: some argue that transfer from "language family" models – trained on languages typologically related to the target language – elicits the highest performance [6], whereas others have found that transfer from Massively Multilingual Transformers (MMTs) – frequently trained on over 100 languages – is preferable [7].

1.2 Catalan and Galician

In this dissertation, I have focused on two low-resource languages, Catalan and Galician.

Catalan is a Western Romance language spoken by 9.3 million people, primarily in and around Catalonia [8]. It is considered a moderately low-resource language, but has several high-resource relatives, including French, Spanish, Italian, and Portuguese. Interestingly, it is difficult to propose any one of these languages as an ideal candidate for cross-lingual transfer. Catalan's exact typological classification is a subject of significant debate, and it is often described as a bridging language between the Gallo-Romance and Ibero-Romance families [9]. This would place French, Spanish, and Portuguese as its closest high-resource relatives, and Catalan indeed bears a complex resemblance to each of these languages in phonology and structure [10]. However, genetic relationships alone do not tell the full story here: Ethnologue states that Catalan shares the greatest lexical similarity with Italian [8], and quantitative morphosyntactic analyses have suggested that these two languages cluster most closely to one another [11].

Galician is an Ibero-Romance language spoken by 3.1 million people, primarily in Galicia [12]. It is also considered a low-resource language and shares the same high-resource relatives as Catalan, though it differs in having a more obvious primary candidate for cross-lingual transfer: Portuguese. The two languages share a single common ancestor and may even be classified as dialects of one another, though this is a strongly contentious issue [9] [13].

Catalan and Galician are thus good languages with which to evaluate the effectiveness of cross-lingual transfer from language family models over MMTs or monolingual models, as they share several high-resource relatives. The languages also provide an interesting foil to one another: it can be assumed that a model performing downstream tasks in Galician would primarily benefit from Portuguese cross-lingual transfer, but the case with Catalan is not nearly as clear-cut. As such, it could be hypothesized that Galician models stand to benefit from single-source transfer to a greater extent than Catalan models, or that the reverse is true for multi-source transfer. Moreover, Catalan provides a good setting in which to assess the most important characteristics

of a high-resource relative for effective cross-lingual transfer, be they deep typological relationships, surface similarities, or something else entirely.

1.3 Research Questions, Methodology, and Main Findings

My work here asks two primary research questions:

1. Which cross-lingual transfer scenario is the most effective for low-resource target languages: transfer from MMTs, transfer from language family models, or transfer from monolingual models?
    - Is this impacted by the typology of the target language?
2. What types of typological relationship between source and low-resource target language facilitate the most effective single-source cross-lingual transfer?

To investigate these, I performed cross-lingual transfer to Catalan and Galician from monolingual models, language family models, and MMTs. I employed multiple transfer protocols across a wide variety of NLP tasks, utilising two transformer model architectures. My findings indicate that multi-source transfer is a superior option to single-source transfer, but that the success of language family models over MMTs depends on the target language's relationship with the source languages used. Additionally, I found that the degree of language contact between Catalan and each source language made good predictions of the success of single-source transfer, and that more direct measures of language similarity were largely inconsistent across all tasks.

2 Background and Related Work

2.1 Machine Learning and Language Modelling

Machine learning approaches to language modelling utilize various forms of Artificial Neural Network (ANN). ANNs are composed of many small computational units, called neurons [14]. Each neuron takes an input vector $\boldsymbol{x}$, weights it using a corresponding weight vector $\boldsymbol{w}$ and a bias term b, and applies a non-linear activation function g to produce an output y:

$$y = g(\boldsymbol{w} \cdot \boldsymbol{x} + b) \tag{1}$$

At its simplest, an ANN takes the form of a Feedforward Neural Network (FNN), which consists of an input layer, an output layer, and at least one hidden layer [14]. Each of these layers is composed of neural units as described above, and they are arranged such that the output of each layer becomes the input of the next layer. Typically, FNNs are fully-connected, with each node in a layer taking every node in the previous layer as an input. This can be visualized in Figure 1:

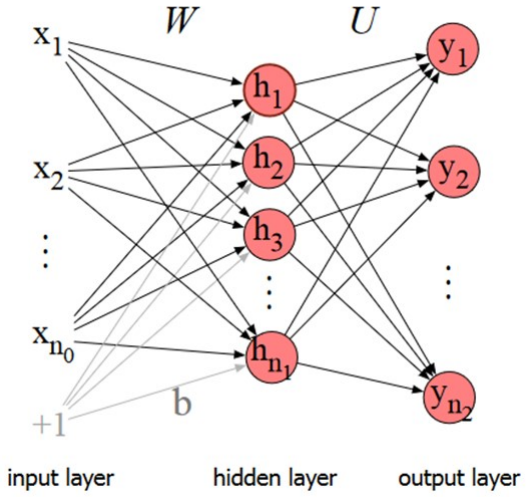


Figure 1: A fully-connected Feedforward Neural Network (FNN), taken from [14].

FNNs can be used as Language Models (LMs), outputting the most probable next word in a sentence given an input of some number of previous words [15]. To do this, a tokenizer is used to map words to numerical representations called tokens, which often involves splitting words into subwords. These are then further converted to embeddings – vectors which encode the token's meaning in some fashion [16]. In a forward pass of a model like Figure 1, the embedding vectors for a small number of input tokens are multiplied by the weight vectors of each node in the hidden layer, summed with the biases, and passed through the activation function to produce a new representation of the input; this is then used similarly by the output layer to generate the final output. The activation function of the output layer is generally softmax, which produces a probability distribution over all the tokens in the model's vocabulary. As such, the model's prediction for the next token is the node in the output layer with the highest probability.

The weights and biases for each unit in the model are randomly initialized, and it must first be trained on some training data before it can be used for effective inference. The goal of training is to learn the weight matrices and bias vectors for each layer such that the model's prediction of an output $\hat{y}$ is as close as possible to the correct output y, given an input x. This is done by computing the distance between $\hat{y}$ and y via a loss function, before finding the weights and biases that minimize this function using gradient descent and error backpropagation [14].

2.2 The Transformer Architecture and Key Innovations

In recent years, a single neural LM architecture has dominated the field: the transformer. The original transformer was an encoder-decoder model, with an encoder network that produces a contextualized representation of the input, and a separate decoder network that autoregressively generates an output using both its own output embeddings and the output of the encoder [17]. The key innovation underpinning the architecture's prominence is its use of an attention mechanism, which allows the model to efficiently incorporate information from an arbitrarily large context of tokens. This is very useful in language modelling, where the ability to capture long-distance relationships and dependencies between words is crucial. Moreover, unlike earlier models which encode these relationships through recurrent hidden layers, the attention mechanism does not rely on sequential computation, and can thus be computed much more efficiently. However, due to the model's lack of recurrence, embeddings must initially be modified with positional encodings, which embed some information about the word's position in the input.

Attention can be thought of as a method of weighting each token in an input by comparing it to every other token in the input. In practice, this is done by projecting each embedding vector into a query, key, and value vector using three respective weight matrices. Attention is then calculated for a given token by taking a sum of every value in the input, each weighted by the compatibility of the value's respective key and the query of the token being given attention to [17]. More specifically, the attention function used by transformers is Scaled Dot-Product Attention, which can be defined as follows:

$$Attention(Q, K, V) = softmax(\frac{QK.T}{\sqrt{d_k}})V \tag{2}$$

Here, the query, key, and value vectors are packed into matrices, thus allowing the attention scores for each token to be calculated simultaneously. The output is scaled by a factor of $\sqrt{d_k}$ (where d_k is the dimensionality of the query/key vectors) to avoid numerical issues during training [18]. Moreover, for autoregression purposes, the score may optionally be masked before the softmax is applied, so that only the current and previous tokens are considered. The mechanism discussed so far is self-attention; it is also possible to have cross-attention by using keys and values from the encoder and queries from the decoder.

Attention is implemented in the architecture as multi-head attention, using smaller, parallel attention layers rather than one large one. This allows the model to capture information from multiple information subspaces at once [17]. After each attention layer, these outputs are concatenated and projected to the original dimensions of the input vector, summed with the original vector (by means of a residual connection), normalized (by subtracting the mean of the vector and dividing by its standard deviation), and passed through a standard FNN. This together forms a "transformer block". The original transformer architecture thus resembles Figure 2:

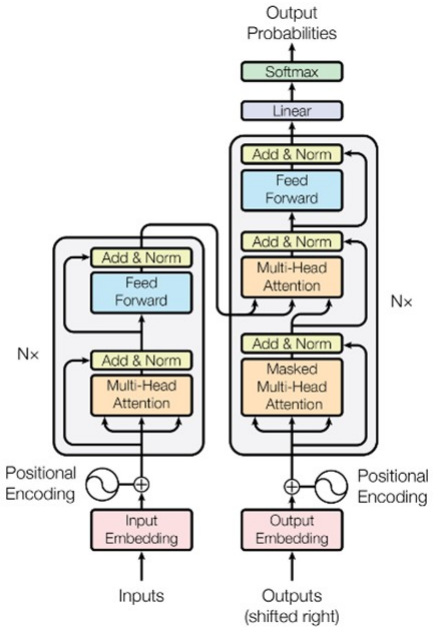


Figure 2: The original Transformer model, taken directly from Vaswani et al. [17].

2.3 Encoder-Decoder, "Encoder-Only", and "Decoder-Only" Models

The encoder-decoder structure of the original transformer architecture renders it especially well-suited to sequence-to-sequence (seq2seq) tasks (where the model's inputs and outputs are both strings of text [19]), but it can prove unnecessary or even detrimental for other NLP tasks [20]. As such, several alterations to the architecture have been developed.

Many models, for instance, are designed to capitalize on the bidirectional capabilities of "encoder" transformer blocks [20]. As shown in Figure 2, these blocks do not mask the tokens that follow the current token when calculating multi-head attention, and instead include the entire input; this could theoretically lead to a more sophisticated understanding of language. As such, several models utilize only "encoder" transformer blocks, and are known as "encoder-only" models. Likewise, several models seek to capitalize on the autoregressive potential of "decoder" transformer blocks, and eschew unmasked "encoder" blocks altogether [21]. These are known as "decoder-only" models.

2.4 Pretraining, Fine-Tuning, and Transfer Learning

Thus far, the only NLP task examined has been language modelling, but transformer models can be trained for many other related purposes; for instance, the original model was used for machine translation [17]. In recent years, however, it has become standard practice for transformers to utilize transfer learning when training on downstream tasks [22]. Transfer learning is the process of improving a model's performance on a task by leveraging knowledge learned from a different task [23]. Within the context of transformer models, this typically involves a model first being trained on a pretraining task that allows it to develop a general, multifaceted understanding of language, before being fine-tuned on a specific downstream NLP task by training the weights for some or all of its layers again, typically using data labelled to indicate the correct output for each input.

The specific pretraining task varies, but it often amounts to either Causal Language Modelling (CLM) or Masked Language Modelling (MLM) [24]. CLM is language modelling as described thus far, predicting the next token given some input of tokens. It is typically used to train encoder-decoder and decoder-only models. MLM, on the other hand, involves randomly masking a certain percentage of the input tokens, and predicting those. It is typically used to train encoder-only models. Both of these pretraining tasks are self-supervised – they do not require human-annotated labels – and as such are easy to compile extremely large corpora for. In this way, transfer learning provides an effective way for models to leverage very large amounts of textual data in downstream tasks where it is otherwise less available.

With regards to MLM, there are several variations of the task used in practice that deviate slightly from the description outlined above. In its simplest form, tokens for words and subwords are masked alike, but this can be adjusted so that all subword tokens in a word are masked [25]. Alternatively, entire spans of text may be replaced by a single mask token [22].

In a substantial way, transfer learning in and of itself alleviates some of the issues low-resource languages face: pretraining can improve a model's ability at a task when labelled task-specific data is scarce [26]. However, it frequently remains insufficient, and many techniques have subsequently been developed to improve low-resource NLP.

2.5 Cross-lingual Transfer Learning

Cross-lingual transfer learning is transfer learning where knowledge of one or more source languages is transferred to a separate target language [5]. Like standard transfer learning, it often revolves around the pretraining and fine-tuning processes within the context of transformers. Cross-lingual transfer is one of the primary methods used to increase performance over downstream tasks in low-resource languages, as it allows models to leverage data from high-resource languages that may otherwise be unavailable. For instance, if a language has a good task-specific dataset suitable for finetuning but no real resources for pretraining, pretraining can instead be done in a high-resource language, and the broad understanding of language learned from this can still be transferred to the downstream task. Frequently, transfer is from languages related to the target language in some way, though it has also been shown to be effective where this is not the case [27].

There are many cross-lingual transfer scenarios that can be devised. In its purest form, a model might be trained to perform a task in a low-resource language entirely through high-resource data, both in pretraining and fine-tuning. Alternatively, the low-resource language might be used for pretraining, but the task-specific data is entirely in a high-resource language. These are known as zero-shot cross-lingual transfer [27], closely related to few-shot transfer, where a small number of target language examples are seen in fine-tuning. As described above, a setup can also be envisioned where a model is pretrained in a high-resource language, but fine-tuned on data in the target low-resource language [6].

2.6 Multi-Source Transfer and MMTs

In addition to single-source scenarios, cross-lingual transfer can take place in a multi-source setup, where data from several languages is leveraged in the target language. This may involve both high and low-resource source languages [28], and the transfer may resemble any of the scenarios described above.

One of the most common types of multilingual model is the Massively Multilingual Transformer (MMT), which is pretrained on data from many high and low-resource languages – frequently over 100 [29]. MMTs are able to use their extremely varied input to develop a sophisticated understanding of language, and have frequently been used for cross-lingual transfer to low-resource languages, included in the pretraining data or otherwise. MMTs have seen significant success in the field, though they have been found to perform poorly on the lowest-resource languages [30]. Moreover, they face a key problem known as the "curse of multilinguality": the addition of more languages in the pretraining data leads to lower performance in low-resource languages after a certain point [28].

An alternative approach to multi-source transfer is to pretrain "language family" models consisting of multiple languages of the same typological family. These can then be used for cross-lingual transfer to other languages in or related to the family, and it has been suggested that this transfer protocol may be superior to MMTs or monolingual models if several high-resource relatives exist [6].

2.7 RoBERTa and T5

For a more complete picture of cross-lingual transfer, I used models employing two different architectures throughout my experiments. The first of these was RoBERTa, a versatile encoder-only transformer architecture that has been found to achieve strong performance across a wide range of NLP tasks [31] [32] [28]. The second was T5, an encoder-decoder transformer architecture designed for seq2seq tasks. Effective performance in

seq2seq tasks requires autoregressive capabilities and a fundamentally different understanding of language, so this transfer was worth investigating.

2.7.1 RoBERTa

RoBERTa is a bidirectional encoder-only transformer architecture, developed as the successor to the extremely influential BERT (Bidirectional Encoder Representations from Transformers) [31]. It improves on its predecessor by adjusting its pretraining process, which consists solely of a dynamic MLM task where the masking patterns are generated each time an input is passed to the model. It also makes use of an improved Byte Pair Encoding (BPE) tokenizer, which statistically forms a vocabulary of subwords by iteratively merging the most common subword pair in a text and appending it to the vocabulary.

A prominent RoBERTa-architecture MMT is XLM-RoBERTa (XLM-R), which is pretrained on 100 languages (including Catalan and Galician) [28]. XLM-R utilizes a SentencePiece tokenizer, which treats entire raw sentences as inputs (including spaces) before applying a subword segmentation algorithm [33]; standard BPE, by contrast, requires an inefficient and cross-linguistically dubious pre-tokenization stage. Here, a unigram algorithm is used to construct subword units, which initializes a large base vocabulary, and progressively removes subwords of low importance based on a heuristic of their probability of occurring [34].

2.7.2 T5

Text-to-Text Transfer Transformer (T5) is an encoder-decoder transformer architecture, developed explicitly for seq2seq tasks [22]. It provides a unifying framework for NLP problems by treating them all as a single sequence-to-sequence task, where the model produces an output text given some input text. Tasks are then differentiated through a task-specific prefix (such as "summarize: "), which must be prepended to every input. T5 is pretrained using a span-masking MLM task as outlined above, with the original model additionally trained on 8 supervised tasks such as question answering. A SentencePiece tokenizer is used, constructing subword units with the WordPiece algorithm – a modification of BPE.

A prominent T5-architecture MMT is mT5, which is pretrained on data from 101 languages (including Catalan and Galician) [29]. The model is faithful to the original architecture, but does not use any supervised, task-specific pretraining, instead relying entirely on MLM.

2.8 Related Work

Snæbjarnarson et al. [6] investigate cross-lingual transfer to Faroese, a low-resource language in the Western Scandinavian family. The study focuses on the MMT XLM-R, monolingual Danish and Icelandic models, and a language family model pretrained on higher-resource Scandinavian relatives, all of which are encoder-only transformers. Each of these models is fine-tuned on and evaluated over four Faroese downstream tasks, and it is found that the Scandinavian family model is largely the best, surprisingly outperforming an alternative Scandinavian model that included Faroese data in pretraining. With regards to monolingual models, the Icelandic model performed best, which could be expected as Icelandic is the closer relative. The authors conclude that language family models may well be a better option than MMTs for cross-lingual transfer to low-resource languages.

Similarly, de Vries et al. [35] investigate the performance of monolingual English, German, and Dutch encoder-only models for part-of-speech (POS) tagging in Gronings and West Frisian, both spoken in the north of the Netherlands. The transformer layers for each model are fine-tuned on labelled data in the source language, and their lexical embedding layers are separately trained using MLM in the target language; the two are then combined and evaluated. It is found that transfer from Dutch elicits the best results for both languages, which is in accordance with the authors' estimates of linguistic similarity between each source and target language, but stands against the fact that Gronings and West Frisian are more closely related to German and English, respectively. Moreover, the Dutch and German models are both found to outperform the MMT mBERT. The authors conclude that measures of linguistic distance can be used to select an ideal candidate for single-source cross-lingual transfer, and that transfer from MMTs may not always be the best option for low-resource languages.

With regards to Catalan and Galician, de Vries et al. [36] finetune XLM-R for POS tagging in 105 target languages in a zero-shot setting, using 65 source languages. The model is fine-tuned on labelled data in each source language, and each fine-tuned model is then evaluated on every target language. Catalan and Galician are used as both source and target languages, and the authors make note that Catalan and Spanish form one of several "optimal language pairs" that are each other's best source language. It is more broadly concluded that factors such as lexical-phonetic distance and word order can have a significant effect on the effectiveness of cross-lingual transfer. Moreover, Armengol-Estapé et al. [37] and Vilares et al. [38] each trained monolingual Catalan and Galician models, and found that they outperformed setups involving transfer from MMTs.

3 Methodology

3.1 Overview

To investigate the research questions outlined in my introduction, I performed cross-lingual transfer to Catalan and Galician by fine-tuning several pretrained transformer language models on a range of target language downstream tasks.

For each Catalan downstream task, I performed transfer from three monolingual models: Spanish, French, and Italian. I compared the performances of these models to that of an MMT, and two Italo-Western language family models – one which utilizes knowledge of Spanish, French, Italian, and Portuguese, and one which includes Catalan and Galician in addition to these. The effects of including target language knowledge on model performance are worth investigating, as low-resource languages often do have some pretraining data available that can be leveraged in a multi-source setup. For each Galician downstream task, I compared the performance of transfer from a monolingual Portuguese model, the two Italo-Western models, and an MMT.

I used RoBERTa models for most of my tasks, fine-tuning them for Named Entity Recognition (NER), POS tagging, Dependency Parsing (DP), and Machine Reading Comprehension (MRC). Together, these allow for a multifaceted and linguistically detailed analysis of each transfer setup's strengths and weaknesses. Unlike the studies outlined above, I also investigated transfer via a seq2seq protocol, fine-tuning T5 models for Summarization and Machine Translation (MT). Due to a lack of appropriate seq2seq resources in Galician, these experiments were only ran for Catalan.

In the following sections, I summarize the models, tasks, and datasets I experimented on.

3.2 Models

3.2.1 RoBERTa Models

As previously mentioned, I used monolingual Spanish, French, Italian, and Portuguese RoBERTa models, two Italo-Western (IW) family models, and an MMT for my experiments. Pretrained RoBERTa models for Spanish, French, and Italian already existed, and I selected BERTIN [32], the base-sized version of CamemBERT [25], and GilBERTo [39], respectively. The former of these is faithful to the original RoBERTa architecture, whereas the latter two introduce a SentencePiece tokenizer. CamemBERT additionally utilizes the whole-word masking variation of MLM, as previously described. For the MMT, I utilized the base-sized version of XLM-R, which was sufficiently powerful while being compact enough to use with the limited resources available to me.

I was not aware of any suitable pretrained Portuguese or IW RoBERTa models, and I did not have the resources to perform any pretraining myself. As such, I chose to create these models from XLM-R-base through a process of pruning unnecessary embeddings, first proposed in Abdaoui et al. [40] as a method of compressing the size of MMTs. Whereas the input and output embeddings in MMTs span many languages, those in monolingual and language family models span only the languages included, and it is thus possible to produce a model comparable to these by removing all but the embeddings relevant to the desired languages from an MMT. As such, I selected the most frequent 30,000 tokens for each language by tokenizing its relevant Leipzig corpus [41], before updating the model's tokenizer vocabulary and input and output embeddings to only include these[1].

[1] I would like to thank David Dale for providing me with the code to implement this process, which can be found at https://colab.research.google.com/drive/1f-n3zBQjmtMrp7oHzvunHPSC5aIMNe_N

Unsurprisingly, there was significant overlap of tokens between languages, and the IW family models have vocabularies comparable to that of the monolingual Portuguese model.

Table 1 summarizes some details about the RoBERTa models selected.

Model	Parameters	Vocabulary Size	Pretraining Corpus Size	Languages
BERTIN (es)	125M	50262	200GB	1
CamemBERT-base (fr)	111M	32005	138GB	1
GilBERTo (it)	111M	32005	71GB	1
XLM-R-pt	109M	30000	2.4TB	1
XLM-R-IW	121M	44923	2.4TB	4
XLM-R-IW-ca-gl	123M	48349	2.4TB	6
XLM-R-base	279M	250002	2.4TB	100

Table 1: Information about the RoBERTa models used in this dissertation. Within this context, "parameters" refers to the number of weights and biases internal to the model.

3.2.2 T5 Models

The T5 models were only used for my Catalan experiments, and as such I required pretrained monolingual models for Spanish, French, and Italian, as well as the two IW family models and an MMT. Of these, I was not aware of any appropriate French or IW models, so I decided to create every necessary model by applying the same "trimming" process as described above to the base-sized version of the MMT mT5[2].

Table 2 summarizes some details about the T5 models selected.

Model	Parameters	Vocabulary Size	Pretraining Corpus Size	Languages
mT5-es	221M	30000	6.3T	1
mT5-fr	221M	30000	6.3T	1
mT5-it	221M	30000	6.3T	1
mT5-IW	240M	54167	6.3T	4
mT5-IW-ca-gl	244M	59137	6.3T	6
mT5-base	580M	250112	6.3T	101

Table 2: Information about the T5 models used in this dissertation.

3.3 Tasks and Datasets

3.3.1 Named Entity Recognition

Named Entity Recognition is the task of identifying named entities – real-world objects with proper names – in some input of text [42]. An example can be seen in Figure 3:

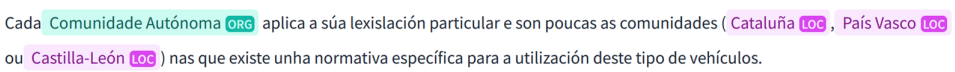


Figure 3: An example of NER, taken from the SLI Galician NERC Gold corpus. Generated using the HuggingFace API.

To perform NER with transformer models, a token classification head must be attached to the output of the final hidden layer, which produces a probability distribution over each possible class of named entity. As the model classifies tokens rather than words, I also had to preprocess my tokenized data so that only the first token in a word was classified.

[2]I would once again like to thank David Dale for his implementation of this task for mT5, which can be found at https://colab.research.google.com/gist/avidale/44cd35bfcdaf8bedf51d97c468cc8001/create_rut5-base.ipynb

The datasets I chose to use for fine-tuning were the AnCora Catalan NER dataset [37] and the SLI Galician NERC Gold Corpus [43]. Both datasets follow the standard CoNLL format [44], and utilize the same named entity classes: persons, organisations, locations, and miscellaneous. As with nearly all of the datasets I used, both of these were divided into train, validation, and test splits, which were used for fine-tuning, model development, and evaluation, respectively. I did this manually for the SLI corpus.

To evaluate NER performance, I primarily relied on the F1 metric, which can be defined as the harmonic mean of precision (the ratio of correctly predicted labels to all predicted labels) and recall (the ratio of correctly predicted labels to all labels that should be predicted) [42]:

$$F_1 = 2\frac{precision \times recall}{precision + recall} \tag{3}$$

I also chose to record each model's accuracy over the test splits, which can be defined simply as the number of correct predictions over the number of total predictions. I implemented both of these metrics using the Python library seqeval [45].

3.3.2 Part Of Speech Tagging

Part of Speech Tagging is the task of classifying each word in an input of text by its grammatical category [42]. An example can be seen in Figure 4:

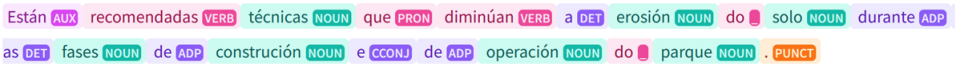


Figure 4: An example of POS, taken from the UD CTG Galician corpus. Generated using the HuggingFace API.

Like NER, POS tagging is a token classification task, and my models needed the same token classification head and data preprocessing algorithm for fine-tuning. I also used the same evaluation metrics: primarily F1, with accuracy for additional insight.

For both languages, I used corpora from the Universal Dependencies project, which provides a unified cross-linguistic framework for the grammatical annotation of text [46]. As such, both datasets utilized identical POS tags, known as UPOS tags. The two datasets selected were the AnCora Catalan and CTG Galician corpora.

3.3.3 Dependency Parsing

Dependency Parsing is the task of automatically generating a dependency tree structure from an input sentence [47]. Dependency structures consist of syntactic relations between heads and dependents, which are labelled for their grammatical function. An example can be seen in Figure 5:

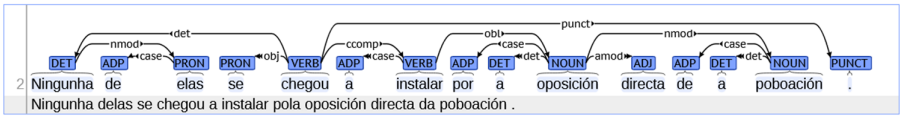


Figure 5: An example of DP, taken from the UD CTG Galician corpus. Generated at https://spyysalo.github.io/conllu.js

DP can thus be thought of as a process of annotating each dependent in a sentence with its associated head and dependency relation label. To perform this, I attached a deep biaffine parser to my models, a graph-based parser which utilizes two classifiers that predict heads and labels separately [48]. In each classifier, the output embeddings of the model are fed through two separate FNNs, which individually encode their roles as a head

and as a dependent. These embeddings are then passed through a biaffine transformation, the output of which is used to produce a probability distribution over heads/labels [47].

UD entries are treebanks, and each sentence is fully annotated with its dependency structure. As such, I was able to use the same datasets for DP that I used for POS tagging. For evaluation, I used Labelled Attachment Score (LAS), the percentage of words assigned both the correct head and dependency relation label. I also measured Unlabelled Attachment Score (UAS), the percentage of words assigned just the correct head.

3.3.4 Machine Reading Comprehension

Machine Reading Comprehension is the task of processing, understanding, and answering questions about an input text in a manner similar to that of humans [49]. It can be used as a method of evaluating a model's capability for natural language understanding (NLU), and is often cast as a multiple-choice question answering task. An example can be seen in Figure 6:

Passage: Assegureu-vos que teniu la mà dreta tan relaxada com sigui possible, sense deixar de tocar totes les notes correctament; a més, heu de mirar de no fer gaires moviments superflus amb els dits. D'aquesta manera, us cansareu el mínim possible. Recordeu que no cal polsar les tecles amb més força per aconseguir més volum, com passa amb el piano. Per tenir més volum en l'acordió s'utilitza la manxa amb més pressió o velocitat.

Question: Segons el text, quin dels següents no es considera un bon consell en tocar l'acordió?

Answer 1: Augmenteu la força amb la qual polseu les tecles per aconseguir més volum

Answer 2: No realitzeu moviments innecessaris, així mantindreu la vostra resistència

Answer 3: No toqueu les notes amb la mà excessivament relaxada

Answer 4: Augmenteu la velocitat amb la qual utilitzeu la manxa per tenir més volum

Correct Answer: 1

Figure 6: An example of MRC, taken from the Catalan portion of the Belebele dataset.

To perform multiple-choice question answering, I attached a multiple-choice classification head to my models, which produces a probability distribution over each possible answer. For preprocessing, I structured my inputs such that each answer is prepended by its associated passage and question.

The dataset I chose for this task was Belebele, a multilingual dataset for MRC evaluation spanning 122 languages [50]. For each language, it contains 900 parallel sets of questions and answers, each linked to a passage from the FLORES-200 dataset [51]. The developers propose several settings in which models can be evaluated on Belebele, and I elected to use the translate-train method, where models are first fine-tuned on training data that has been machine translated from English to the target language. This suited my needs particularly well, as it was difficult to find relevant training data in Catalan or Galician, and the developers had provided a script to assemble a set of appropriate English examples from six larger multiple-choice datasets[3]. I thus machine translated the passages, questions, and answers for each entry in the assembled training data ("Beletrain") sentence by sentence, using a CTranslate2[4] implementation of NLLB-200-distilled-1.3B [51]. I then divided Beletrain into train, validation, and test splits.

Belebele has no Galician entry, so I created one by machine translating each set of questions and answers, and aligning them with the relevant passage from the Galician portion of FLORES-200. Following the example of the authors, I chose accuracy as my evaluation metric, and measured it both over Belebele and the test split of Beletrain.

3.3.5 Summarization

Summarization is the task of shortening the length of some input text while still retaining all of its important information [52]. An example can be seen in Figure 7:

[3] https://github.com/facebookresearch/belebele/blob/main/assemble_training_set.py
[4] https://github.com/OpenNMT/CTranslate2

Text: El Grec Festival de Barcelona 2018 ofereix aquest any una programació especial extraordinària dedicada a celebrar les 50 edicions del Festival Internacional de Jazz de Barcelona. Snarky Puppy, Pat Metheny, Cory Henry & The Funk Apostles, R+R = Now i Cécile Mclorin Salvant són els noms que s'han triat per celebrar l'efemèride. Els cinc concerts es faran entre el 6 i el 25 de juliol, quatre dels quals a la sala Barts, mentre que el cinquè es farà a l'amfiteatre de Montjuïc. Les cinc cites estan coprogramades pels dos festivals barcelonins. Les entrades per tots els concerts, tret del de Pat Metheny (que sortirà a la venda pròximament), ja estan disponibles a la pàgina web www.jazz.barcelona.

Summary: El Grec Festival celebra el 50 aniversari del Festival Internacional de Jazz de Barcelona.

Figure 7: An example of Summarisation, taken from the CaSum dataset.

Summarization is a seq2seq task, and as such requires a language modelling head that produces a probability distribution over the tokenizer's entire vocabulary. Preprocessing was minimal, but as I used T5 models, I prepended the prefix "summarize: " to each input.

The dataset I used for summarization was CaSum, extracted from a Catalan newswire corpus [53]. I evaluated performance using ROUGE scores: ROUGE-1 and ROUGE-2, based on the overlap of unigrams and bigrams between predicted and reference summaries, and ROUGE-L, based on their longest common subsequences [54]. Of these, I considered ROUGE-2 the most important, following previous work [55]. I also utilized BERTScore, which allowed me to use mT5's pretrained contextual embeddings to measure the cosine similarity between words in predictions and references [56].

3.3.6 Machine Translation

Machine Translation is the task of translating sentences from one language to another [57]. An example can be seen in Figure 8:

Fr: Je ne peux pas rester assise là en attendant que tu reviennes.

Ca: No em seuré aquí tot esperant que tornis.

Figure 8: An example of MT, taken from the Catalan-French portion of the OpenSubtitles corpus.

As MT is an inherently bilingual task, I decided against fine-tuning my multilingual models for it. Instead, I evaluated translation to Catalan from the language of each monolingual model, which I felt would enrich the discussions surrounding my second research question. Like summarization, MT is a seq2seq task that requires a language modelling head. For preprocessing, I prepended the prefix "translate [Spanish/French/Italian] to Catalan: " to each input.

I thus required three bilingual datasets comparable to one another in size and theme. As such, I decided to use subsections of the OpenSubtitles corpus, consisting of movie and TV subtitles across many languages [58]. For each source language, I compiled a set of 300,000 parallel sentences in that language and Catalan, before dividing them into train, validation, and test splits. To evaluate MT performance, I primarily considered the BLEU metric (implemented using sacreBLEU [59]), based on the n-gram word overlap between predicted and reference translations. I also measured chrF++, which includes the n-gram character overlap in addition to this [60].

Table 3 summarizes some details about the datasets used, and Table 4 outlines the broad experimental setup.

Dataset	Task(s)	Language(s)	Training Split Size (rows)	Validation Split Size (rows)	Test Split Size (rows)
AnCora-ca-NER	NER	ca	10.6k	1.43k	1.53k
SLI-NERC-gl-Gold	NER	gl	5.19k	648	649
UD-AnCora	POS & DP	ca	13.1k	1.71k	1.85k
UD-CTG	POS & DP	gl	2.27k	860	861
Beletrain-ca	NLU	ca	57.1k	7.13k	7.13k
Beletrain-gl	NLU	gl	57.1k	7.13k	7.13k
Belebele	NLU	ca & gl	NA	NA	900
CaSum	Summarization	ca	198k	10k	10k
OpenSubtitles-es-ca	Translation	ca & es	240k	30k	30k
OpenSubtitles-fr-ca	Translation	ca & fr	240k	30k	30k
OpenSubtitles-it-ca	Translation	ca & it	240k	30k	30k

Table 3: Information about the datasets used in this dissertation.

Model/Task	NER	POS	DP	MRC	Summarization	MT
BERTIN (es)	ca	ca	ca	ca		
CamemBERT-base (fr)	ca	ca	ca	ca		
GilBERTo (it)	ca	ca	ca	ca		
XLM-R-pt	gl	gl	gl	gl		
XLM-R-IW	ca, gl	ca, gl	ca, gl	ca, gl		
XLM-R-IW-ca-gl	ca, gl	ca, gl	ca, gl	ca, gl		
XLM-R-base	ca, gl	ca, gl	ca, gl	ca, gl		
mT5-es					ca	ca
mT5-fr					ca	ca
mT5-it					ca	ca
mT5-IW					ca	
mT5-IW-ca-gl					ca	
mT5-base					ca	

Table 4: The models used in this dissertation, and the tasks they are fine-tuned for.

3.4 Experimental Setup

To fine-tune my models, I used the HuggingFace Transformers Python library[5], which provides a framework for model training with PyTorch[6]. To implement and fine-tune deep biaffine parsers for DP, I additionally used the SuPar library[7], which is built on top of Transformers.

In all tasks, each model was fine-tuned for 4 epochs. At the end of every epoch, each model was saved and evaluated on the validation set, with the best performing checkpoint (according to the task's primary evaluation metric) selected for further evaluation on the test set. These scores were then used as my final results. As the training process involves a random data sampling process, I trained most of my models multiple times, and calculated the mean and standard error of their final performances across all runs. Resource limitations prevented me from doing this for the seq2seq tasks.

To ensure that fine-tuning was effective, I tuned the hyperparameters of the training process. Transformers are only able to operate on sequences of tokens of a certain size, and it was thus necessary to truncate inputs to task-specific lengths. Model learning is affected by batch size – the number of examples seen before weights are updated – which subsequently also needed to be controlled for. In some instances, a shortage of memory necessitated the use of gradient accumulation steps – where weights are only updated after a certain number

[5]https://github.com/huggingface/transformers

[6]I would like to thank HuggingFace for their substantial collection of tutorial and example scripts, which I used and modified extensively throughout this project

[7]https://github.com/yzhangcs/parser

of batches – to achieve an effective batch size. Memory limitations also occasionally necessitated the use of mixed precision training, where the precision of floating point variables used in training is lowered. Further hyperparameters to control were learning rate – the degree to which a model's weights change after each batch – and weight decay – which penalizes complexity and helps to prevent the model becoming too specialized for the training data. For all tasks, I used the AdamW optimizer [61] to update model parameters during training. Any remaining hyperparameters were set to the default values of either the HuggingFace TrainingArguments class or the SuPar dep-biaffine-xlmr configuration.

The hyperparameters used for each task are outlined in Table 5.

Task	Selection Metric	Runs	Epochs Per Run	Learning Rate	Batch Size	Gradient Accumulation Steps	Max Sequence Length	Weight Decay	Mixed Precision
NER	F1	5	4	2e-5	16	1	512	0.01	No
POS	F1	5	4	2e-5	16	1	512	0.01	No
DP	LAS	5	4	2e-5	16	1	NA	NA	No
MRC	Accuracy	5	4	1e-5	16	1	512	0.01	Yes
Summarization	ROUGE-2	1	4	2e-5	16/8	1/2	1024 (256 for summaries)	0.01	Yes
MT	BLEU	1	4	2e-5	16	1	128	0.01	No

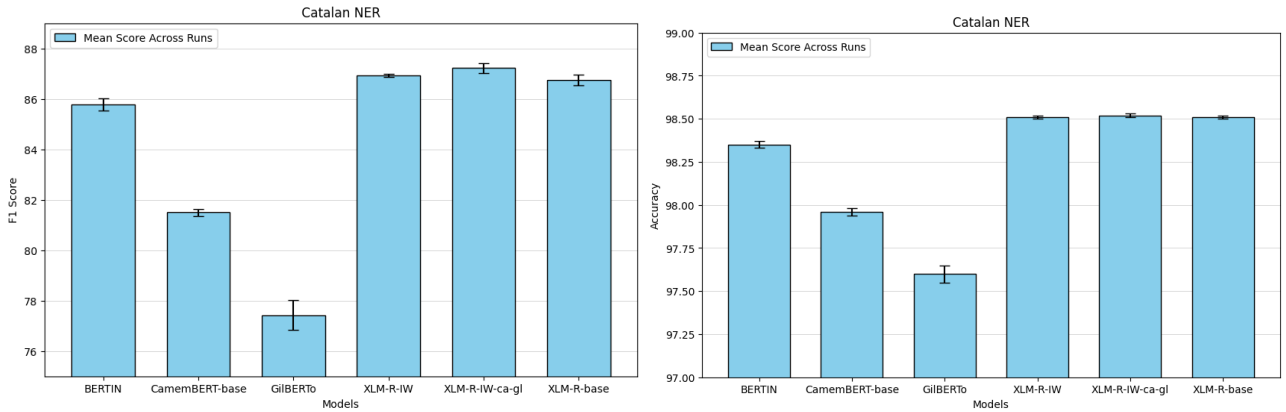
Table 5: Information about the datasets used in this dissertation.

4 Results

4.1 Named Entity Recognition

Model	F1	Accuracy
BERTIN (es)	85.78 ± 0.24	98.35 ± 0.02
CamemBERT-base (fr)	81.50 ± 0.14	97.96 ± 0.02
GilBERTo (it)	77.42 ± 0.59	97.60 ± 0.05
XLM-R-IW	86.93 ± 0.07	98.51 ± 0.01
XLM-R-IW-ca-gl	**87.23 ± 0.20**	**98.52 ± 0.01**
XLM-R-base	86.75 ± 0.22	98.51 ± 0.01

Table 6: Results for Catalan NER. Mean performance across runs and standard error intervals are reported.



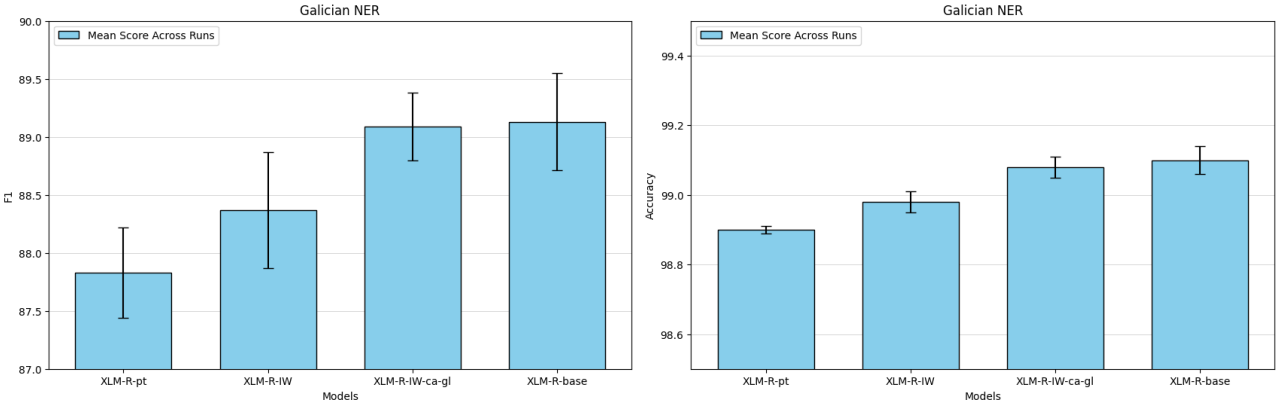
(a) F1 Score

(b) Accuracy

Figure 9: Results for Catalan NER.

Model	F1	Accuracy
XLM-R-pt	87.83 ± 0.39	98.90 ± 0.01
XLM-R-IW	88.37 ± 0.50	98.98 ± 0.03
XLM-R-IW-ca-gl	89.09 ± 0.29	99.08 ± 0.03
XLM-R-base	**89.13 ± 0.42**	**99.10 ± 0.04**

Table 7: Results for Galician NER. Mean performance across runs and standard error intervals are reported.



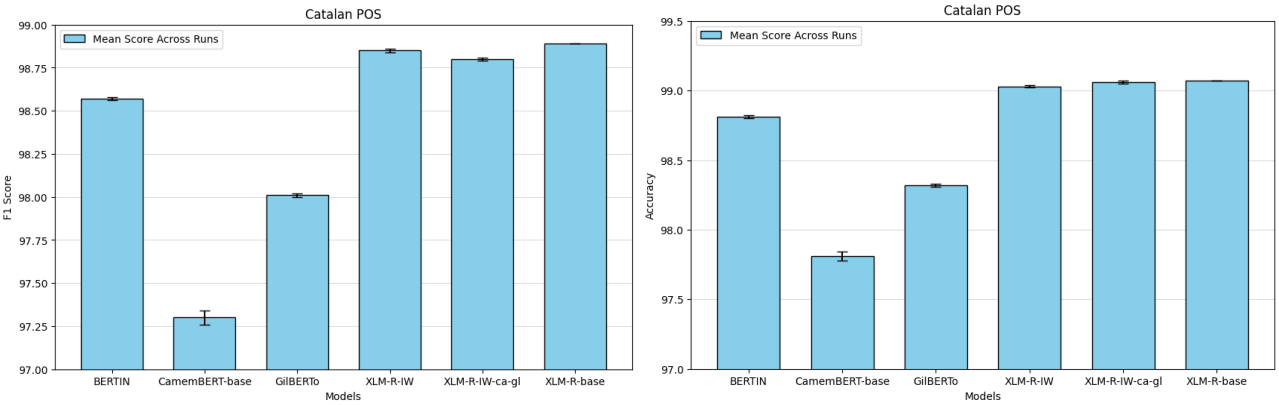
(a) F1 Score (b) Accuracy

Figure 10: Results for Galician NER.

4.2 Part Of Speech Tagging

Model	F1	Accuracy
BERTIN (es)	98.57 ± 0.01	98.81 ± 0.01
CamemBERT-base (fr)	97.36 ± 0.04	97.81 ± 0.03
GilBERTo (it)	98.01 ± 0.01	98.32 ± 0.01
XLM-R-IW	98.85 ± 0.01	99.03 ± 0.01
XLM-R-IW-ca-gl	98.87 ± 0.01	99.06 ± 0.01
XLM-R-base	**98.89 ± 0.00**	**99.07 ± 0.00**

Table 8: Results for Catalan POS Tagging. Mean performance across runs and standard error intervals are reported.

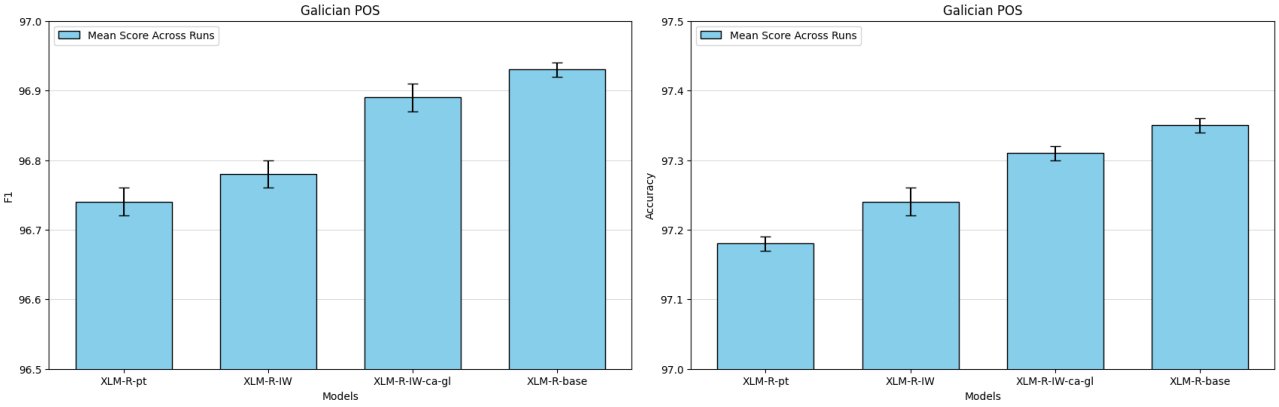


(a) F1 Score (b) Accuracy

Figure 11: Results for Catalan POS Tagging.

Model	F1	Accuracy
XLM-R-pt	96.74 ± 0.02	97.18 ± 0.01
XLM-R-IW	96.78 ± 0.02	97.24 ± 0.02
XLM-R-IW-ca-gl	96.89 ± 0.02	97.31 ± 0.01
XLM-R-base	**96.93 ± 0.01**	**97.35 ± 0.01**

Table 9: Results for Galician POS Tagging. Mean performance across runs and standard error intervals are reported.



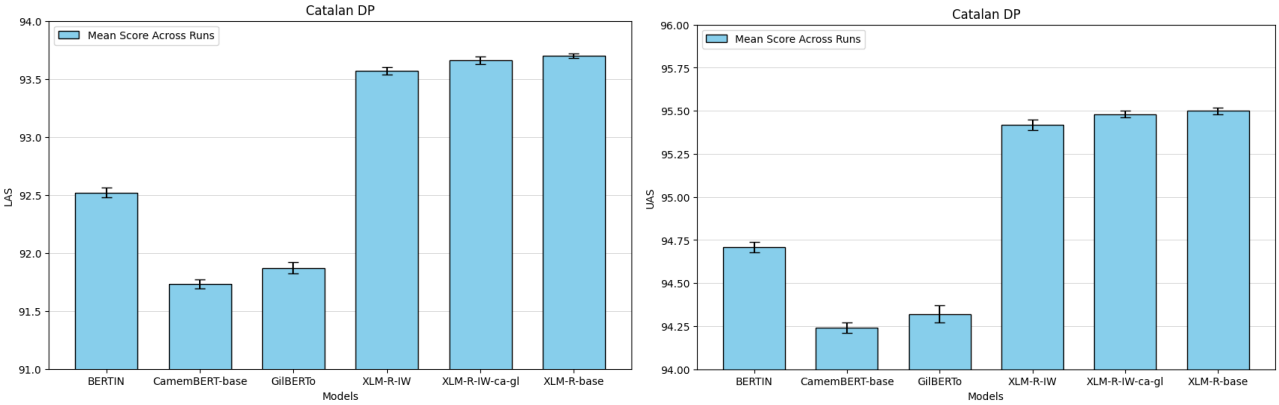
(a) F1 Score (b) Accuracy

Figure 12: Results for Galician POS Tagging.

4.3 Dependency Parsing

Model	LAS	UAS
BERTIN (es)	92.52 ± 0.04	94.71 ± 0.03
CamemBERT-base (fr)	91.73 ± 0.04	94.24 ± 0.03
GilBERTo (it)	91.87 ± 0.05	94.32 ± 0.05
XLM-R-IW	93.57 ± 0.03	95.42 ± 0.03
XLM-R-IW-ca-gl	93.66 ± 0.03	95.48 ± 0.02
XLM-R-base	**93.70 ± 0.02**	**95.50 ± 0.02**

Table 10: Results for Catalan DP. Mean performance across runs and standard error intervals are reported.

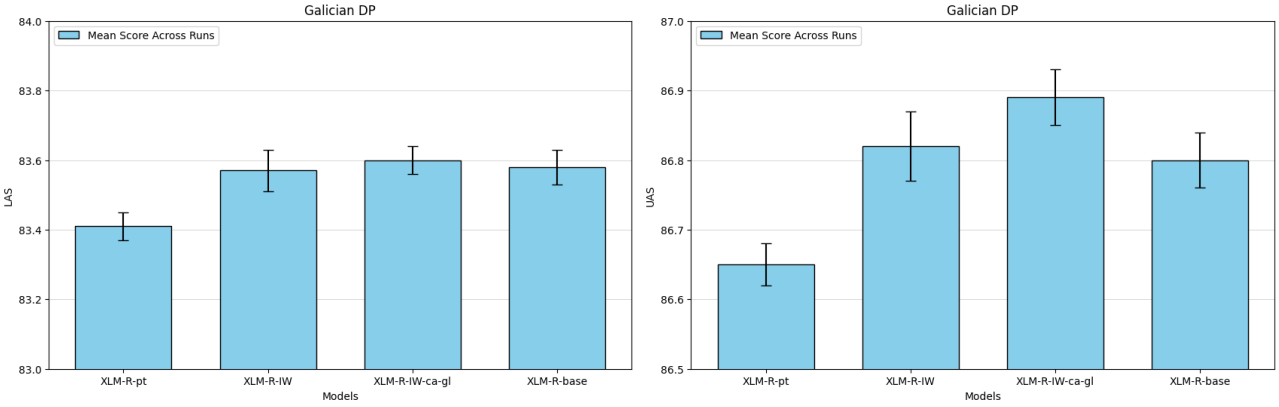


(a) LAS (b) UAS

Figure 13: Results for Catalan DP.

Model	LAS	UAS
XLM-R-pt	83.41 ± 0.04	86.65 ± 0.03
XLM-R-IW	83.57 ± 0.06	86.82 ± 0.05
XLM-R-IW-ca-gl	**83.60 ± 0.04**	**86.89 ± 0.04**
XLM-R-base	83.58 ± 0.05	86.80 ± 0.04

Table 11: Results for Galician DP. Mean performance across runs and standard error intervals are reported.



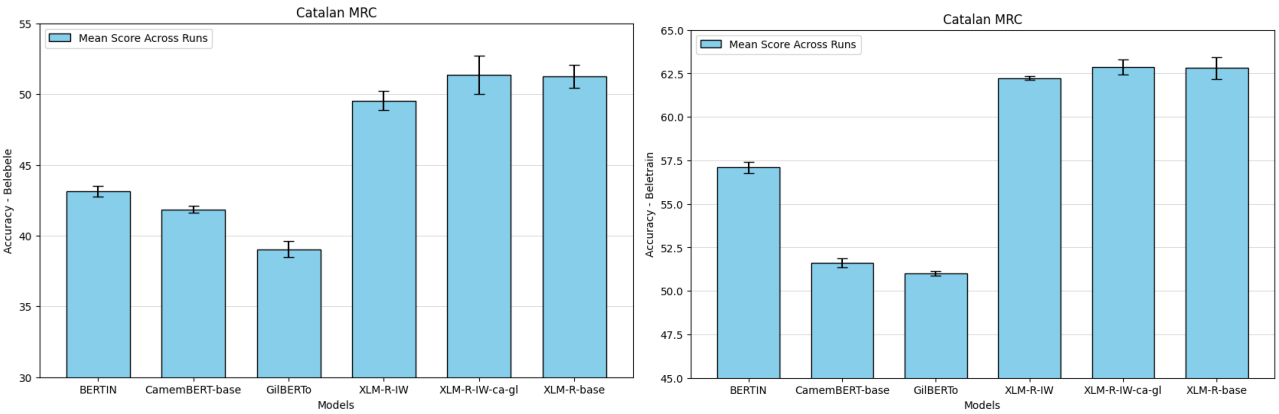
(a) LAS (b) UAS

Figure 14: Results for Galician DP.

4.4 Machine Reading Comprehension

Model	Accuracy - Belebele	Accuracy - Beletrain
BERTIN (es)	43.13 ± 0.39	57.09 ± 0.34
CamemBERT-base (fr)	41.86 ± 0.22	51.63 ± 0.26
GilBERTo (it)	39.04 ± 0.58	51.00 ± 0.14
XLM-R-IW	49.53 ± 0.68	62.22 ± 0.11
XLM-R-IW-ca-gl	**51.36 ± 1.36**	**62.88 ± 0.44**
XLM-R-base	51.25 ± 0.83	62.82 ± 0.63

Table 12: Results for Catalan MRC. Mean performance across runs and standard error intervals are reported.

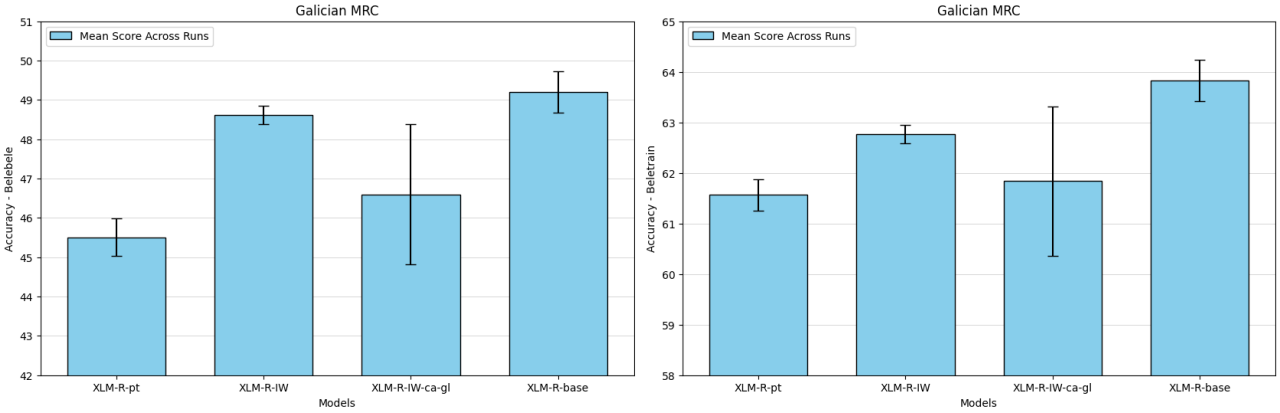


(a) Accuracy - Belebele (b) Accuracy - Beletrain

Figure 15: Results for Catalan MRC.

Model	Accuracy - Belebele	Accuracy - Beletrain
XLM-R-pt	45.51 ± 0.47	61.57 ± 0.31
XLM-R-IW	48.62 ± 0.23	62.77 ± 0.18
XLM-R-IW-ca-gl	46.60 ± 1.78	61.84 ± 1.48
XLM-R-base	**49.20 ± 0.52**	**63.83 ± 0.41**

Table 13: Results for Galician MRC. Mean performance across runs and standard error intervals are reported.



(a) Accuracy - Belebele

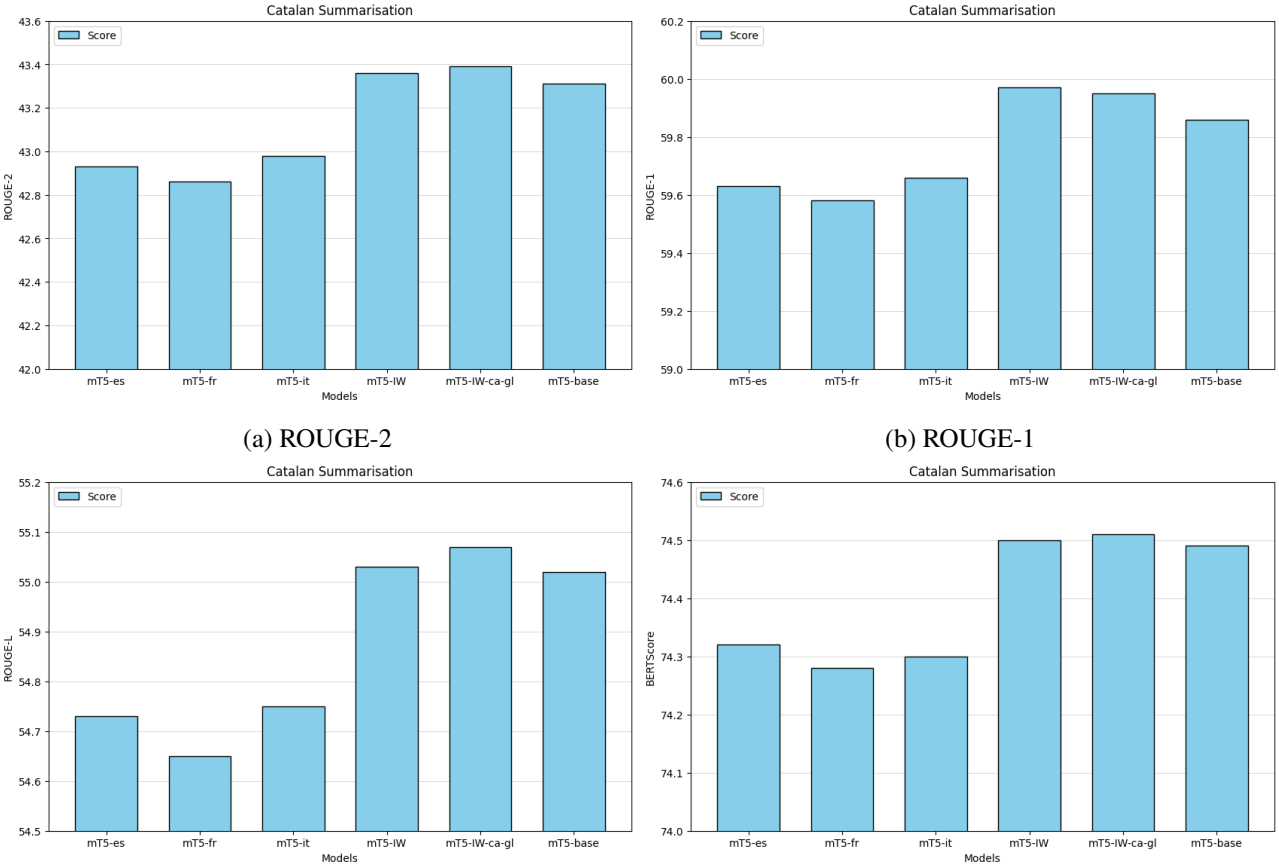
(b) Accuracy - Beletrain

Figure 16: Results for Galician MRC.

4.5 Summarization

Model	ROUGE-2	ROUGE-1	ROUGE-L	BERTScore
mT5-es	42.93	59.63	54.73	74.32
mT5-fr	42.86	59.58	54.65	74.28
mT5-it	42.98	59.66	54.75	74.30
mT5-IW	43.36	**59.97**	55.03	74.50
mT5-IW-ca-gl	**43.39**	59.95	**55.07**	**74.51**
mT5-base	43.31	59.86	55.02	74.49

Table 14: Results for Catalan Summarization.



(a) ROUGE-2

(b) ROUGE-1

(c) ROUGE-L

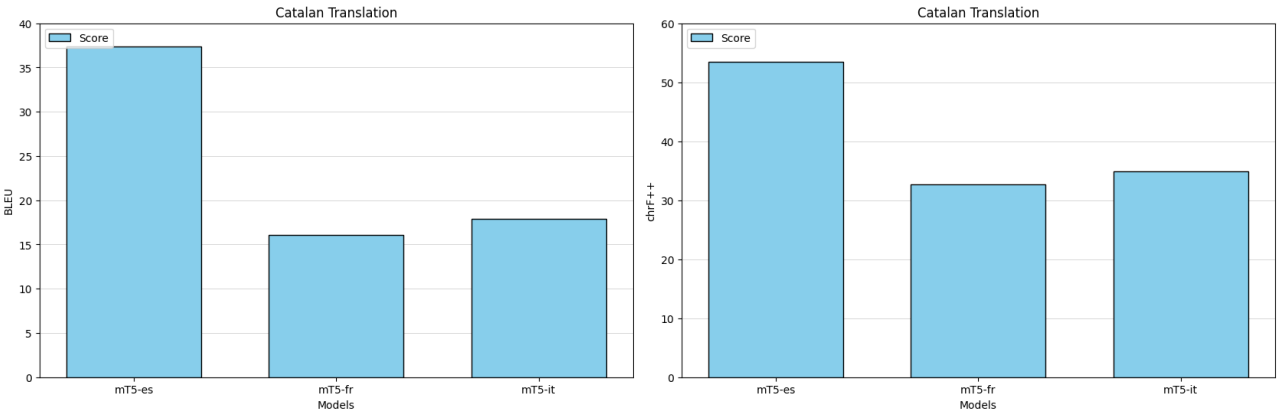
(d) BERTScore

Figure 17: Results for Catalan Summarization.

4.6 Machine Translation

Model	BLEU	chrF++
mT5-es	**37.38**	**53.49**
mT5-fr	16.10	32.63
mT5-it	17.88	34.93

Table 15: Results for Catalan MT.



(a) BLEU

(b) chrF++

Figure 18: Results for Catalan MT.

4.7 Summary and Key Findings

4.7.1 Research Question 1

For both languages, monolingual models elicited the worst transfer, with performance consistently lower than multilingual models across all tasks and metrics. In Galician, however, a standard error overlap between the monolingual model and the lowest-performing multilingual model can be seen on the graphs for NER, POS tagging, and MRC; no such pattern can be observed in the Catalan data. It thus seems that monolingual models are a better choice for Galician than for Catalan, though this may be a consequence of the only monolingual model used for the Galician experiments – XLM-R-pt – being a subsection of an MMT.

With regards to the multilingual models, the results are less clear-cut. For Catalan, the language family model XLM-R-IW-ca-gl performs best over 3 of the 5 relevant tasks, with the MMT XLM-R-base exhibiting higher scores in POS tagging and DP. For Galician, XLM-R-base is the best model for 3 out of 4 tasks, with XLM-R-IW-ca-gl performing better in DP. As such, it seems that language family models are generally a better choice than MMTs for Catalan, with the reverse being true for Galician.

Additionally, XLM-R-IW-ca-gl generally outperforms XLM-R-IW across both languages, with the exceptions being Galician MRC (in which the performance of XLM-R-IW-ca-gl fluctuates wildly across runs, shown by the standard error bars in Figure 16), and the ROUGE-1 score for Catalan summarization (demonstrated in Figure 17). This indicates that including some target language vocabulary in a language family model may lead to better transfer.

4.7.2 Research Question 2

Spanish models outperformed both of the other monolingual models over 5 of the 6 tasks. The only exception to this was found in summarization, where mT5-it scored higher than mT5-es over all evaluation metrics except BERTScore (as shown in Figure 17). It thus seems sensible to posit Spanish as the best language for cross-lingual transfer to Catalan overall. For the other languages, Italian models outperformed French models in 4 of the 6 Catalan tasks: POS tagging, DP, summarization, and MT. This suggests that Italian is broadly the better source language for cross-lingual transfer to Catalan, though there are still 2 tasks (NER and MRC) where French is the better choice.

4.8 Experimental Issues

I encountered some technical difficulties when performing these experiments that may have affected my results. When fine-tuning for MRC, the Spanish models consistently exhibited overfitting to the training data, and performance on the validation set tended to decline after about 2 epochs. As such, earlier checkpoints were selected than with the other models. It should also be noted that several of the multilingual models occasionally failed to converge on the MRC training data at all, with validation accuracy never growing greater than chance – where this was the case, the run was repeated.

A further potential issue concerns summarization, where the mT5-es checkpoint selected (on the basis of ROUGE-2 score) was that of the third epoch, rather than the fourth. This was not the case for any other model in the task, and it is likely responsible for the Spanish model's poor performance in comparison to the Italian model.

5 Discussion

5.1 Research Question 1

5.1.1 Monolingual vs Multilingual Models

The fact that cross-lingual transfer from monolingual models was worse than from multilingual models makes intuitive sense: a model with knowledge of several languages will likely have a more abstract and sophisticated understanding of language than a model with knowledge of just one. Given that no target language will align completely with a separate source language, this is especially beneficial for cross-lingual transfer learning, and

indeed many studies on the subject have reported similar findings to mine [6] [7] [30]. The aforementioned methodological issues make it difficult to conclusively state that monolingual models are more effective for transfer to Galician than to Catalan, as my results suggest; however, if this finding were to be treated as valid, it could be hypothesized that it results from a higher degree of linguistic similarity between Galician and Portuguese than between Catalan and any of its source languages, thus enabling more effective transfer.

5.1.2 Language Family Models vs MMTs

Linguistic similarity may also help to explain the discrepancies observed between the most effective multi-source model types across the two languages. The fact that XLM-R-IW-ca-gl outperforms XLM-R-base for Catalan but not Galician seems counterintuitive – if Galician is more similar to source languages like Portuguese and Spanish, it would make sense if language family transfer to it was more effective than to Catalan. However, it may well be the case that Galician is *too* similar to source languages like Portuguese and Spanish for XLM-R-IW/XLM-R-IW-ca-gl to be maximally effective, as other source languages like French – which is comparatively much more dissimilar to Galician – stand to contribute significantly less to the transfer, and may even serve as "distractions". Catalan, by comparison, may lie in a "sweet spot" of similarity with these languages, being similar enough to benefit from the transfer in the first place, but dissimilar enough for each language to productively contribute in some fashion.

To investigate this hypothesis, it is beneficial to consider both quantitative and qualitative measures of linguistic similarity. With regards to the former, a simple heuristic of lexico-phonetic similarity can be obtained by calculating the Normalized Divided Levenshtein Distance (LDND) [62] between 40-item wordlists from the ASJP database [63], which has been found to be a good predictor of cross-lingual transfer efficiency [35] [36]. Table 16 lists the LDND measures between the languages relevant to this study[8], which can be visualized through multidimensional scaling (MDS) in Figure 19:

	Catalan	French	Galician	Italian	Portuguese	Spanish
Catalan	0	75.45	63.51	65.41	72.74	72.12
French	75.45	0	79.86	78.25	78.74	84.03
Galician	63.51	79.86	0	49.90	55.01	54.82
Italian	65.41	78.25	49.90	0	65.86	57.26
Portuguese	72.74	78.74	55.01	65.86	0	67.96
Spanish	72.12	84.03	54.82	57.26	67.96	0

Table 16: Distance matrix for the LDND measures between each language's ASJP list.

[8] Obtained using the asjp62 program [64]

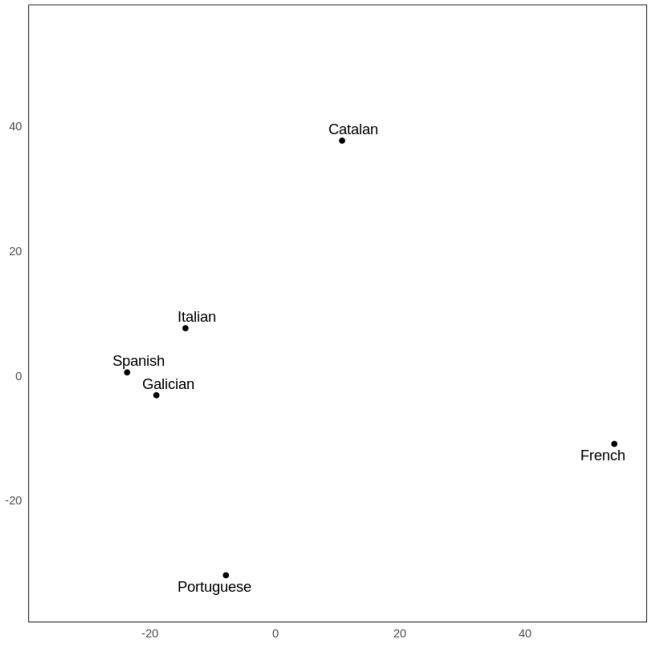


Figure 19: MDS plot for the LDND measures between each language's ASJP list.

It is worth noting that this metric is quite simplistic and unnuanced, and as such the data presented is somewhat dubious. However, it can be clearly seen that Catalan's position is almost equidistant between each of the high-resource languages, whereas Galician is much closer to Portuguese, Spanish, and Italian, and much further from French. This seems convincing evidence in favour of the sweet spot hypothesis.

For a more sophisticated account of linguistic similarity, it is perhaps best to examine the nature of the tasks investigated, and the linguistic knowledge that models require for optimal performance in them. Interestingly, tasks requiring a strong degree of syntactic knowledge seem to be exceptional across both languages: the best results for Catalan DP and POS tagging were produced by XLM-R-base, and the best results for Galician DP were produced by XLM-R-IW-ca-gl. These findings can be elegantly accounted for if it is posited that the sweet spot patterns seen thus far are inverted for syntactic tasks, with Galician's syntax placing it at the sweet spot of the source languages, and Catalan's rendering it either "too sweet" (too similar to a subset of the source languages) or "too sour" (too dissimilar from the source languages). Indeed, such a hypothesis may well be warranted by the linguistic evidence. Whereas Catalan morphology bears many resemblances to French (for instance, the lack of masculine gender markers [65]), its syntax is perhaps more similar to Spanish, Italian, and Portuguese, with whom it shares features such as the permission of null subjects [65] – this may render it "too sweet". This notion is further supported by the fact that the French model exhibited poorer performance than the Spanish and Italian models over both syntactic tasks. Galician syntax, on the other hand, while arguably more dissimilar to French syntax than Catalan, is also arguably less similar to the syntax of Spanish or Italian, with distinctive features such as enclitic object pronouns [13]. It thus seems reasonable to suggest that it might lie in more of a syntactic sweet spot.

A simple, quantitative heuristic for estimating syntactic similarity can be devised by measuring the distance between lang2vec vectors from the URIEL database, which encode properties of languages in vector form [66]. URIEL's syntax lang2vec vectors are constructed based on sets of each language's syntactic features, sourced from databases such as WALS [67] and SSWL [68]. For my purposes, the syntax_knn vectors – which include statistical predictions for any values missing from these sources – were the most suitable to use, as the database entries for several languages (especially Galician) were quite sparse. Table 17 lists the Euclidian distances between the syntax_knn vectors for the languages relevant to this study, and Figure 20 provides a visualisation of these through MDS:

	Catalan	French	Galician	Italian	Portuguese	Spanish
Catalan	0	3.61	2.45	1	2	2
French	3.61	0	3.87	3.74	3.32	3.61
Galician	2.45	3.87	0	2.24	2.45	2.45
Italian	1	3.74	2.24	0	1.73	1.75
Portuguese	2	3.32	2.45	1.73	0	2
Spanish	2	3.61	2.45	1.73	2	0

Table 17: Distance matrix for the Euclidian distances between each language's URIEL syntax_knn vector.

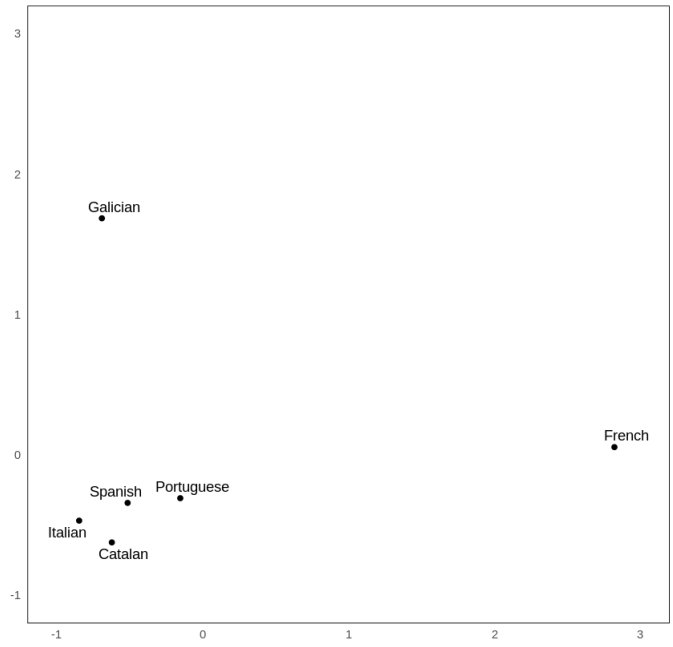


Figure 20: MDS plot for the Euclidian distances between each language's URIEL syntax_knn vectors.

As with the LDND measures, this data must be treated with caution, and it is worth noting that the Galician syntax_knn vector featured a much higher number of predicted values (70 out of 103 total) than any of the other languages. Nonetheless, there is a clear pattern present in Figure 20, with Catalan and Galician almost swapping places when compared to Figure 19. This again suggests that Catalan syntax may be "too sweet" for optimal transfer from the source languages, and that Galician syntax resides in more of a sweet spot, which is consistent with my findings.

An alternative analysis might focus on subword overlap, the degree to which the subwords in a model's vocabulary are shared across languages [69]. It has been found that a higher degree of overlap facilitates better cross-lingual transfer [70], but this is not necessarily universal across all tasks – Limisiewicz et al. [69] find that a higher subword overlap is beneficial for tasks like NER and NLU, but detrimental to others like POS tagging and dependency labelling. Given that XLM-R-base almost certainly has a lower overlap than XLM-R-IW-ca-gl, these findings mirror those that I have obtained for Catalan, which could suggest that subword overlap is responsible for the patterns observed. However, this hypothesis does not provide a good account of my Galician results, or the poor performance of XLM-R-IW (which likely has a higher subword overlap than XLM-R-IW-ca-gl), and I do not believe it should be favoured over the sweet spot hypothesis.

5.1.3 Including Target-Language Examples in Pretrained Model

The fact that XLM-R-IW-ca-gl consistently outperformed XLM-R-IW across both languages is another finding that seems intuitively sensible, but it in fact diverges somewhat from the existing literature – Snæbjarnarson et al. [6] report the opposite effect for language family transfer to Faroese. However, these patterns were primarily

found with language family models that had been further adapted to Faroese via an additional pretraining stage; this was theorized to lower the subword overlap with the other source languages, thus reducing the potential for transfer. Where this adaptation did not take place, the models including Faroese vocabulary largely performed better, and I thus believe that these findings can be reconciled with my own.

5.1.4 Summary

In all, then, I believe that my results have led me to answer my first research question as follows: a language family model may be a better choice than an MMT for cross-lingual transfer to a low-resource language, but only if the target language lies in a "sweet spot" of typological similarity to the source languages included. Regardless of this, some form of multi-source cross-lingual transfer is likely a better choice than single-source cross-lingual transfer.

5.2 Research Question 2

5.2.1 Genealogy

Firstly, my results seem to indicate that a close genealogical relationship between a source and target language is not necessarily a good predictor of strong cross-lingual transfer. Although the question as to whether Catalan is more closely related to Spanish or French is fiercely debated, it could be expected on the basis of genealogy alone that cross-lingual transfer from these languages to Catalan would be somewhat similar – this is not the case with my results, where the gap between Spanish and French is consistently large. Moreover, the fact that the French models were broadly outperformed by the Italian models is completely unpredicted by genealogy, with French being a closer relative to Catalan than Italian. Ultimately, these findings are not particularly surprising, as a language's ancestry alone does not necessarily reveal much about the social or environmental factors that have come to define it.

5.2.2 Surface-Level Similarity

It is somewhat simpler to reconcile more surface-level relationships with my results, though not without exception. As stated previously, the ASJP distances presented in Table 16 posit Italian as the most similar language to Catalan, followed by Spanish, and then French. My small sample size makes it difficult to calculate statistical measures of correlation (such as Pearson's correlation coefficient) with any confidence, but this broad pattern is largely inconsistent with my findings: while it can account for Italian's success over French, it is difficult to reconcile with Spanish's success over Italian. In this respect, my results stand against existing studies that have found strong correlations between ASJP distances and cross-lingual transfer performance [35] [36], though I do not consider this to be altogether surprising given the simplicity of the metric. It seems reasonable for a high surface similarity to facilitate efficient cross-lingual transfer, given the increases in subword overlap and shared embeddings that come with it. However, it does not appear to be sufficient for optimal transfer in and of itself, and there must be other typological relationships at play.

Interestingly, if the methodological concerns surrounding the task are to be overlooked, ASJP distances align closely with my summarization results. The Italian model's comparatively poor performance in MT makes it difficult to suggest that this finding has any real implications for seq2seq tasks in general, but it could perhaps be hypothesized that close surface-level relationships are crucial for successful cross-lingual transfer in seq2seq tasks where both encoding and decoding are performed in the target language.

5.2.3 Morphosyntactic Similarity

Unfortunately, it is difficult to achieve a more fruitful account of single-source transfer by investigating the morphosyntactic nature of the tasks performed. The URIEL syntactic distances presented in Table 17 align well with the fact that Italian models outperform French models over the primary syntactic tasks, but not with the superiority of the Spanish models over both of these. Indeed, it is also difficult to account for this pattern through more qualitative syntactic means, which render the situation much more complex: while Catalan syntax does resemble that of Spanish and Italian with its loose word order and permission of null subjects, key structures like negation are much more reminiscent of their French counterparts. Catalan's morphology provides no

clearer an account: its lack of masculine gender marking may go some way to explain the French model's success in NER over the Italian model, but its clear lack of success in comparison to the Spanish model remains somewhat surprising. Moreover, it is difficult to identify any particular aspects of linguistic knowledge that might underly the Spanish model's dominance in more general, high-level tasks such as MRC and MT, with the variable success of the Italian and French models making the situation even less clear. As such, while it seems appropriate to consider the linguistic nature of the task when deciding on an ideal source language, this alone seems not to be fully adequate.

5.2.4 Language Contact

A final typological relationship to consider in this case is language contact, wherein a language is influenced through its interactions with other languages [71]. As a regional language, the vast majority of Catalan speakers are bilingual, primarily with Spanish, but also to a lesser extent with French in the smaller region of Northern Catalonia, and Italian in the Sardinian city of Alghero [72]. Contact effects can thus be expected for each of these languages, but it is almost certainly fair to say that Catalan's contact with Spanish has historically been the most intense. Castilian Spanish gradually became the dominant language of Catalonia from as early as 1497, leading to the persecution of Catalan and strong influences in syntax, phonology, and vocabulary [72]. A similar fate belied the region of Northern Catalonia with French, but the effects of this contact are largely restricted to the smaller North Catalan dialects [72]. Contact with Italian has been much more limited, being primarily restricted to Alghero, and it is difficult to account for the success of the Italian models on this basis alone. However, the relationship provides a better explanation for the success of Spanish models in comparison to the other languages than any other discussed thus far, and I thus believe it is reasonable to posit it as a particularly good indicator of successful transfer.

It should be noted, though, that this will likely not be the case for many low-resource languages. In the case of Catalan, contact is relatively simple in that it has been predominantly bilingual, involving only two languages. This is certainly not the case universally, and highly multilingual contact environments are common [73], potentially making it difficult to identify any language as being of primary influence. Moreover, there are many examples of language contact wherein cross-lingual transfer would likely prove ineffective due to the stark differences of the languages involved, for example between English and Hindi. Alternatively, a language may have had little to no contact with other languages at all, rendering the relationship irrelevant.

5.2.5 Summary

In all, I do not believe that my results have identified a single typological relationship between source and low-resource target language that can be consistently relied upon to facilitate optimal cross-lingual transfer. While measures such as morphosyntactic and surface lexical similarity seem to make more reliable predictions than deeper genealogical ties, none are accurate over all tasks, and are as such rather limited in scope. In the case of Catalan, it may be most appropriate to consider the impacts of language contact, which seem to pattern closely with my results; however, there are many low-resource languages for which this would likely prove either unfruitful or impractical. As such, it is difficult to posit any general rule for the selection of an optimal source language for single-source cross-lingual transfer – instead I must conclude that it is more appropriate to tailor each decision to the relevant target language and task.

5.3 Limitations

Perhaps the most significant limitation of my study is the fact that I did not have the resources to pretrain my language family models and many of my monolingual models from scratch. While the process of creating models by pruning redundant embeddings from MMTs served as a good compromise, there are substantial differences between a model formed in this way and a model pretrained on just the source languages. For instance, the weights of the embeddings in a "true" monolingual/language family model will be trained on only the relevant source languages; the embeddings in my models, on the other hand, retain the same weights as are present in XLM-R/mT5, and as such are additionally informed by the many other languages present in the training data. As is evident from the success of models like XLM-R-IW-ca-gl, this was not necessarily a barrier to increases in performance, and it can be argued that its implications are mitigated somewhat for the Galician

and seq2seq experiments (where all models used are subsections of MMTs). However, it is certainly still an issue, and it renders my results less reliable.

There are also smaller methodological issues that may have limited the reliability of my results. Firstly, I did not have the resources to perform multiple runs for any of my seq2seq experiments, which prevented me from taking an average of scores and calculating standard error intervals. Given the strong variance in performance found in some tasks, this could be problematic. The number of runs I performed for the rest of my experiments (five) is also somewhat low, and increasing this number would afford a greater degree of confidence to my conclusions. Moreover, the lack of Galician language NLP resources constrained my discussions on the seq2seq experiments to Catalan, thus limiting them in scope.

A further limitation concerns the Italian RoBERTa architecture model, GilBERTo. Unlike every other RoBERTa model used in this study, GilBERTo is uncased, meaning it lowercases and strips accent markers from all text during processing. This is potentially problematic, as case is of huge potential significance in tasks like NER, and given that my experiments found that the Italian model performed poorly here, I felt that it was worth investigating further. As such, I performed a single run of each of my RoBERTa experiments again using a cased Italian RoBERTa model comparable in size to GilBERTo – UmBERTo [74]. Table 18 outlines some details about UmBERTo, and Tables 19 (a)-(d) present the results of these supplementary experiments.

Model	Parameters	Vocabulary Size	Pretraining Corpus Size	Languages
UmBERTo (it)	111M	32005	69GB	1
GilBERTo (it)	111M	32005	71GB	1

Table 18: Information about UmBERTo, with information about GilBERTo for reference.

Model	F1	Accuracy
UmBERTo (it)	80.83	97.85
BERTIN (es)	85.78 ± 0.24	98.35 ± 0.02
CamemBERT-base (fr)	81.50 ± 0.14	97.96 ± 0.02
GilBERTo (it)	77.42 ± 0.59	97.60 ± 0.05

(a) NER

Model	F1	Accuracy
UmBERTo (it)	97.92	98.25
BERTIN (es)	98.57 ± 0.01	98.81 ± 0.01
CamemBERT-base (fr)	97.36 ± 0.04	97.81 ± 0.03
GilBERTo (it)	98.01 ± 0.01	98.32 ± 0.01

(b) POS

Model	LAS	UAS
UmBERTo (it)	92.00	94.44
BERTIN (es)	92.52 ± 0.04	94.71 ± 0.03
CamemBERT-base (fr)	91.73 ± 0.04	94.24 ± 0.03
GilBERTo (it)	91.87 ± 0.05	94.32 ± 0.05

(c) DP

Model	Accuracy - Belebele	Accuracy - Beletrain
UmBERTo (it)	38.11	50.45
BERTIN (es)	43.13 ± 0.39	57.09 ± 0.34
CamemBERT-base (fr)	41.86 ± 0.22	51.63 ± 0.26
GilBERTo (it)	39.04 ± 0.58	51.00 ± 0.14

(d) MRC

Table 19: UmBERTo results for Catalan NER, POS, DP, and MRC, with the results from the other monolingual models for reference.

UmBERTo outperforms GilBERTo in NER and DP, but performs worse in POS tagging and MRC. More importantly, none of the UmBERTo scores would change the "ranking" of an Italian model in comparison to Spanish and French: no score is better than the average of the next highest performing model, or lower than the average of the next lowest performing model. I thus do not believe that the use of GilBERTo has substantially impacted the results of my study.

5.4 Further Work

In addition to rectifying the limitations outlined above, there are several directions that further research could take to expand on these results. Firstly, it would be of significant benefit to apply my methodology to other low-resource languages, and investigate whether the patterns I have identified for Catalan and Galician can be further generalized. Results from a more diverse pool of target languages might help to identify a better indicator of successful single-source transfer than can be gleaned from Catalan alone – as previously discussed, there is significant reason to doubt that language contact will be universally informative. Moreover, it would be beneficial to evaluate the validity of the "sweet spot" hypothesis for language family transfer in less trivial environments: for instance, where the most appropriate typological relatives for transfer are substantially less high-resource than the Italo-Western languages, or are much more distant from the target language, if genealogically related at all.

A second area of expansion would be to perform cross-lingual transfer via a more diverse set of transfer protocols. The present study has investigated several protocols, but they largely follow a consistent structure: pretrain on the source languages, and fine-tune on the target language. There are many other techniques available, however, and it is worth investigating whether changes made to the nature of the transfer would impact the conclusions I have drawn. Indeed, I have already discussed a potential example of this: the model adaptation protocol adopted in Snæbjarnarson et al. [6] elicited results that contradicted my findings of XLM-R-IW-ca-gl outperforming XLM-R-IW.

It may also prove enlightening to investigate transfer to a greater number of NLP tasks. While the tasks I have performed in this study are varied in nature and have facilitated a nuanced linguistic analysis of the transfer effects observed, there remain several types of application that I did not consider. For instance, I did not include any sequence classification tasks, which involve assigning labels to whole strings of text. An expansion in this respect could provide greater insight into my second research question, for example, potentially demonstrating a better typological predictor of successful transfer.

6 Conclusion

In this dissertation, I have investigated cross-lingual transfer to two low-resource languages with several high-resource relatives. I performed transfer from both monolingual and multilingual transformer language models, utilising multiple transfer protocols across a wide variety of NLP tasks.

I have analysed my results with respect to two primary research questions. The first of these aimed to identify the most effective type of model for transfer to a low-resource target language, and the degree to which this was affected by the language's typology. I found that multi-source transfer setups near-universally outperformed single-source transfer setups, and discussed the possibility of this pattern being stronger for target languages typologically more distant from the source languages. My results also indicated that MMTs trained on over 100 languages could be outperformed by much smaller language family models, though only where the target language lay in a "sweet spot" of typological similarity (with respect to the nature of the task being performed) to the source languages included in them. Finally, I found that the incorporation of target language examples into a model's vocabulary led to an increase in performance.

My second research question aimed to identify the typological relationships between source and target languages that facilitated the best single-source transfer. I found that no typological relationship was able to account for the patterns observed in my results across all tasks, but suggested that the degree of language contact between a source and target language acts as a good heuristic for successful transfer. However, I also discussed the ways in which extrapolating this finding to other languages might prove to be dubious, and I emphasized

the necessity of examining each target language and task on an individual basis. For this research question in particular, I feel that further work examining a wider range of low-resource languages would be beneficial.

Acknowledgements

I would like to extend my warmest thanks to my supervisors Dr Ivan Vulić and Professor Anna Korhonen for their guidance and support throughout this project, without which this dissertation would not have been possible. I would also like to thank David Dale for his assistance with implementing the process of trimming redundant embeddings from MMTs, as well as the HuggingFace team for their truly excellent resources for transformer language models and NLP in general.

Gràcies! Grazas!

References

[1] "Prepare for truly useful large language models," *Nature Biomedical Engineering*, vol. 7, no. 2, pp. 85–86, 2023.

[2] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The state and fate of linguistic diversity and inclusion in the NLP world," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020, pp. 6282–6293.

[3] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow, "A survey on recent approaches for natural language processing in low-resource scenarios," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2021, pp. 2545–2568.

[4] A. Lauscher, V. Ravishankar, I. Vulić, and G. Glavaš, "From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020, pp. 4483–4499.

[5] C.-H. Lee and H.-Y. Lee, "Cross-lingual transfer learning for question answering," *arXiv preprint arXiv:1907.06042*, 2019.

[6] V. Snæbjarnarson, A. Simonsen, G. Glavaš, and I. Vulić, "Transfer to a low-resource language via close relatives: The case study on Faroese," in *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. Tórshavn, Faroe Islands: Association for Computational Linguistics, 2023, pp. 728–737.

[7] Y. Fujinuma, J. Boyd-Graber, and K. Kann, "Match the script, adapt if multilingual: Analyzing the effect of multilingual pretraining on cross-lingual transferability," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 1500–1512.

[8] "Catalan," in *Ethnologue: Languages of the World*, 27th ed., D. M. Eberhard, G. F. Simons, and C. D. Fennig, Eds. Dallas, Texas: SIL International, 2024.

[9] G. Bossong, "Classifications," in *The Oxford Guide to the Romance Languages*, 1st ed., A. Ledgeway and M. Maiden, Eds. Oxford University Press, 2016, pp. 63–72.

[10] A. Alsina, "Catalan," in *The Oxford Guide to the Romance Languages*, A. Ledgeway and M. Maiden, Eds. Oxford University Press, 2016, pp. 363–381.

[11] H. Liu and C. Xu, "Quantitative typological analysis of Romance languages," *Poznań Studies in Contemporary Linguistics*, vol. 48, no. 4, pp. 597–625, 2012.

[12] "Galician," in *Ethnologue: Languages of the World*, 27th ed., D. M. Eberhard, G. F. Simons, and C. D. Fennig, Eds. Dallas, Texas: SIL International, 2024.

[13] F. Dubert and C. Galves, "Galician and Portuguese," in *The Oxford Guide to the Romance Languages*, A. Ledgeway and M. Maiden, Eds. Oxford University Press, 2016, pp. 411–446.

[14] D. Jurafsky and J. H. Martin, "7: Neural networks and neural language models," in *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd ed. Unpublished, 2024. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[15] Y. Goldberg, "Language modeling," in *Neural Network Methods for Natural Language Processing*, 1st ed., ser. Synthesis Lectures on Human Language Technologies, G. Hirst, Ed. Springer Nature, 2017, no. 37, pp. 105–113.

[16] D. Jurafsky and J. H. Martin, "6: Vector semantics and embeddings," in *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd ed. Unpublished, 2024. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Long Beach, CA: Curran Associates, Inc., 2017, pp. 6000–6010.

[18] D. Jurafsky and J. H. Martin, "10: Transformers and large language models," in *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd ed. Unpublished, 2024. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[19] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. Montreal, Canada: MIT Press, 2014, pp. 3104–3112.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186.

[21] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI, Technical Report, 2018.

[22] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[23] S. Niu, Y. Liu, J. Wang, and H. Song, "A decade survey of transfer learning (2010–2020)," *IEEE Transactions on Artificial Intelligence*, vol. 1, no. 2, pp. 151–166, 2020.

[24] D. Jurafsky and J. H. Martin, "11: Fine-tuning and masked language models," in *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd ed. Unpublished, 2024. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[25] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de La Clergerie, D. Seddah, and B. Sagot, "CamemBERT: A tasty French language model," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020.

[26] J. C. B. Cruz and C. Cheng, "Evaluating language model finetuning techniques for low-resource languages," *arXiv preprint arXiv:1907.00409*, 2019.

[27] E. Gogoulou, A. Ekgren, T. Isbister, and M. Sahlgren, "Cross-lingual transfer of monolingual models," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2022, pp. 948–955.

[28] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of*

the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020, pp. 8440–8451.

[29] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mT5: A massively multilingual pre-trained text-to-text transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2021, pp. 483–498.

[30] S. Wu and M. Dredze, "Are all languages created equal in multilingual BERT?" in *Proceedings of the 5th Workshop on Representation Learning for NLP*. Online: Association for Computational Linguistics, 2020, pp. 120–130.

[31] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[32] J. De la Rosa, E. G. Ponferrada, P. Villegas, P. G. d. P. Salas, M. Romero, and M. Grandury, "BERTIN: efficient pre-training of a Spanish language model using perplexity sampling," *Procesamiento del Lenguaje Natural*, vol. 68, pp. 13–23, 2022.

[33] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 66–71.

[34] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 66–75.

[35] W. de Vries, M. Bartelds, M. Nissim, and M. Wieling, "Adapting monolingual models: Data can be scarce when language similarity is high," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, 2021, pp. 4901–4907.

[36] W. de Vries, M. Wieling, and M. Nissim, "Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 7676–7685.

[37] J. Armengol-Estapé, C. P. Carrino, C. Rodriguez-Penagos, O. d. G. Bonet, C. Armentano-Oller, A. Gonzalez-Agirre, M. Melero, and M. Villegas, "Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, 2021, pp. 4933–4946.

[38] D. Vilares, M. Garcia, and C. Gómez-Rodríguez, "Bertinho: Galician BERT representations," *Procesamiento del Lenguaje Natural*, vol. 66, pp. 13–26, 2021.

[39] G. Ravasio and L. Di Perna, "GilBERTo: An Italian pretrained language model based on RoBERTa," 2020. [Online]. Available: www.github.com/idb-ita/GilBERTo

[40] A. Abdaoui, C. Pradel, and G. Sigel, "Load what you need: Smaller versions of multilingual BERT," in *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*. Online: Association for Computational Linguistics, 2020, pp. 119–123.

[41] D. Goldhahn, T. Eckart, and U. Quasthoff, "Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages." in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey: European Language Resources Association (ELRA), 2012, pp. 759–765.

[42] D. Jurafsky and J. H. Martin, "8: Sequence labeling for parts of speech and named entities," in *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd ed. Unpublished, 2024. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[43] R. Agerri, X. G. Guinovart, G. Rigau, and M. A. S. Portela, "Developing new linguistic resources and tools for the Galician language," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), 2018.

[44] E. F. Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. Edmonton, Canada: Association for Computational Linguistics, 2003, pp. 142–147.

[45] H. Nakayama, "seqeval: A python framework for sequence labeling evaluation," 2018. [Online]. Available: www.github.com/chakki-works/seqeval

[46] M.-C. de Marneffe, C. D. Manning, J. Nivre, and D. Zeman, "Universal dependencies," *Computational Linguistics*, vol. 47, no. 2, pp. 255–308, 2021.

[47] D. Jurafsky and J. H. Martin, "18: Dependency parsing," in *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd ed. Unpublished, 2024. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[48] T. Dozat and C. D. Manning, "Deep biaffine attention for neural dependency parsing," in *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*, Toulon, France, 2017.

[49] R. Baradaran, R. Ghiasi, and H. Amirkhani, "A survey on machine reading comprehension systems," *Natural Language Engineering*, vol. 28, no. 6, pp. 683–732, 2022.

[50] L. Bandarkar, D. Liang, B. Muller, M. Artetxe, S. N. Shukla, D. Husa, N. Goyal, A. Krishnan, L. Zettlemoyer, and M. Khabsa, "The Belebele benchmark: A parallel reading comprehension dataset in 122 language variants," *arXiv preprint arXiv:2308.16884*, 2023.

[51] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang, "No language left behind: Scaling human-centered machine translation," *arXiv preprint arXiv:2207.04672*, 2022.

[52] J.-M. Torres-Moreno, "Why summarize texts?" in *Automatic Text Summarization*, 1st ed. John Wiley & Sons, 2014, pp. 3–21.

[53] O. de Gibert, K. Kharitonova, B. C. Figueras, J. Armengol-Estapé, and M. Melero, "Sequence-to-sequence resources for Catalan," *arXiv preprint arXiv:2202.06871*, 2022.

[54] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81.

[55] M. La Quatra and L. Cagliero, "BART-IT: An efficient sequence-to-sequence model for Italian text summarization," *Future Internet*, vol. 15, no. 1, p. 15, 2022.

[56] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *Proceedings of the The 8th International Conference on Learning Representations (ICLR 2020)*, Online, 2020.

[57] D. Jurafsky and J. H. Martin, "13: Machine translation," in *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd ed. Unpublished, 2024. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[58] P. Lison, J. Tiedemann, and M. Kouylekov, "OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora," in *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), 2018.

[59] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 186–191.

[60] M. Popović, "chrF++: Words helping character n-grams," in *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 612–618.

[61] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*, New Orleans, Louisiana, 2019.

[62] S. Wichmann, E. W. Holman, D. Bakker, and C. H. Brown, "Evaluating linguistic distance measures," *Physica A: Statistical Mechanics and its Applications*, vol. 389, no. 17, pp. 3632–3639, 2010.

[63] S. Wichmann, E. W. Holman, and C. H. Brown, "The ASJP database (version 20)," 2022. [Online]. Available: asjp.clld.org

[64] E. W. Holman, "Programs for calculating ASJP distance matrices (version 2.1)," 2011. [Online]. Available: asjp.clld.org/software

[65] J. I. Hualde, *Catalan*, 1st ed. Routledge, 1992.

[66] P. Littell, D. R. Mortensen, K. Lin, K. Kairis, C. Turner, and L. Levin, "URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, 2017, pp. 8–14.

[67] M. S. Dryer and M. Haspelmath, "The world atlas of language structures online," 2013. [Online]. Available: www.wals.info

[68] C. Collins and R. Kayne, "Syntactic structures of the world's languages," 2009. [Online]. Available: www.terraling.com/groups/7

[69] T. Limisiewicz, J. Balhar, and D. Mareček, "Tokenization impacts multilingual language modeling: Assessing vocabulary allocation and overlap across languages," in *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, 2023, pp. 5661–5681.

[70] S. Wu and M. Dredze, "Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT," 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)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 833–844.

[71] A. P. Grant, "Contact-induced linguistic change: An introduction," in *The Oxford Handbook of Language Contact*, A. P. Grant, Ed. Oxford University Press, 2020, pp. 1–48.

[72] M. W. Wheeler, "Catalan," in *Concise Encyclopedia of Languages of the World*, 1st ed., K. Brown and S. Ogilvie, Eds. Elsevier, 2009, pp. 188–192.

[73] S. G. Thomason, "Multilingualism in nations and individuals," in *Language Contact: An Introduction*, 1st ed. Edinburgh University Press, 2001, pp. 27–58.

[74] L. Parisi, S. Francia, and P. Magnani, "UmBERTo: An Italian language model trained with whole word masking," 2020. [Online]. Available: www.github.com/musixmatchresearch/umberto