Empowering Low-Resource Regional Languages with Lexicons : A Comparative Study of NLP Tools for Morphosyntactic Analysis

Cristina Garcia Holgado, Marianne Vergez-Couret

Empowering Low-Resource Regional Languages with Lexicons : A Comparative Study of NLP Tools for Morphosyntactic Analysis

Cristina Garcia Holgado[1,2], **Marianne Vergez-Couret**[1]
[1]FoReLLIS, UR 15076, University of Poitiers
[2]UMR CNRS 6240 LISA, University of Corsica - Pasquale Paoli
5 Rue Théodore Lefebvre, 86000 Poitiers, France
Avenue Jean Nicoli, 20250 Corte, France
cristina.garcia.holgado@univ-poitiers.fr, marianne.vergez.couret@univ-poitiers.fr

Abstract

We investigate the effect of integrating lexicon information to an extremely low-resource language when annotated data is scarce for morpho-syntactic analysis. Obtaining such data and linguistic resources for these languages are usually constrained by a lack of human and financial resources making this task particularly challenging. In this paper, we describe the collection and leverage of a bilingual lexicon for Poitevin-Saintongeais, a regional language of France, to create augmented data through a neighbor-based distributional method. We assess this lexicon-driven approach in improving POS tagging while using different lexicon and augmented data sizes. To evaluate this strategy, we compare two distinct paradigms: neural networks, which typically require extensive data, and a conventional probabilistic approach, in which a lexicon is instrumental in its performance. Our findings reveal that the lexicon is a valuable asset for all models, but in particular for neural, demonstrating an enhanced generalization across diverse classes without requiring an extensive lexicon size.

Keywords: low-resource languages, lexical resources, POS tagging, Poitevin-Saintongeais, regional languages, evaluation, distributional neighbors

1. Introduction

This article investigates the effect of integrating lexicon information to a very low-resource language, Poitevin-Saintongeais, a regional language of France, when annotated data is scarce for morphosyntactic analysis. Current techniques rely on large annotated data, but for low-resource languages obtaining those annotations is often constrained by a lack of human and financial resources making this task particularly challenging.

We present here the first steps in providing lexical resources for this language while comparing the effectiveness and limitations of transferring lexical information for POS tagging. As a secondary task, we evaluate the effect for lemmatization. Since this language is in its initial stages of linguistic endowment, the objectives of this work are twofold: Firstly, evaluate two different POS tagging methods as a starting point for an automatic annotation method when only very few linguistic resources are available to complement human annotation. Secondly, to establish a groundwork for addressing dialectal variation in future work.

The work presented here is part of the DIVITAL project[1], a project aiming to provide linguistic resources and aligned corpora for various regional languages of France to increase their digital visibility. It constitutes the first efforts on automatic morphosyntactic analysis for Poitevin-Saintongeais.

The article is structured as follows: Section 2 provides a description of the language and introduces the motivation behind this work. Section 3 presents relevant prior approaches in morphosyntactic analysis, as well as recent corpus annotation methods commonly used to handle LR languages. A description of the corpus and the linguistic resources used in the experiments are provided in section 4. The methodology for model training, corpus augmentation and the utilized text representations are detailed in section 5. In section 6, we outline the experimental setup and the training parameters. The results and discussion of each method, including an evaluation of using varying lexicon sizes are presented in section 7. We conduct an error analysis in section 8, and discuss the strengths and limits of our lexicon-based strategy. Lastly, we summarize our findings and outline future directions in section 9.

2. Context

2.1. Poitevin-Saintongeais

Poitevin-Saintongeais is a Romance language spoken between the Loire and Garonne rivers, with a strong Occitan substratum. It is morphologically close to French, but differs from other Oïl idioms by

[1]Project funded by the French National Research Agency, ANR-21-CE27-0004

a few salient features, for instance the 1st and 4th person subject pronoun *i*, derived from the Latin "ego" in the areas of "grammatical words" and verbal inflection and the palatalization of groups from Latin, such as [p+l] noted pll- in standardized spelling: *pllanjhe* (calm) in the graphic variations that transpose phonetic evolutions and variations.

Another crucial property of Poitevin-Saintongeais from the NLP point of view is that it is not a standardized language. It has two varieties called Poitevin and Saintongeais. While there is a recent spelling standard called *graphie normalisée* (standardized spelling), other spellings have been or are still used. This dialectal and spelling diversity manifests itself on the lexical and morphological levels making Poitevin-Saintongeais particularly challenging for NLP tasks, since it aggravates the data sparsity issue.

Regarding the status of this language, it suffers from a breakdown in transmission, and as of today, there is no estimate available for the number of speakers in the region.

Nevertheless, we provide the first effort towards endowing Poitevin-Saintongeais with essential NLP resources and discuss next our strategy doing so.

2.2. Motivation

From the NLP perspective, Poitevin-Saintongeais stands as one of the many languages in France falling within the category of low-resource languages. In this sense, the availability of annotated data and other other linguistic assets the language are limited. Although there are a few linguistic descriptions and a relatively large number of digital texts available, the former are incomplete and the latter are characterized by orthographic and dialectal variations, which complicates the task of developing language processing applications.

On the other side, obtaining annotated texts in morphosyntax is one of the main objectives of the project on which this article is based. While there is ongoing manual annotation work, the scarcity of linguistic experts hampers the creation of high-quality annotated corpora. In fact, the difficulty in finding annotators limits the data and complicates the linguistic endowment from a methodological standpoint. Considering this, manual approaches have been carried alongside this work but annotators are unequivocal in this case: it is faster to annotate pre-annotated data than raw data. Therefore, we opted for a computational approach after acquiring an initial set of manually annotated texts seeking to expedite the annotation process by concurrently utilizing automatic methods alongside human efforts, especially in a time when the language still requires considerable attention and various external factors hinder the development of manually annotated corpora.

However, any supervised machine learning method requires a minimum size of training data. While cross-lingual transfer learning has proven effective for low-resource languages with insufficient or no annotated training data, it does not ensure consistent annotation quality across all languages, especially without a small amount of training data to fine-tune the model. Therefore, in this work we aim to leverage an available lexical resource to create augmented training data through distributional neighbors, utilizing French as the intermediary language to transfer potential replacements for tokens within the corpus.

Through this approach, our goal is to augment the training corpus seeking to enhance the morphosyntactic analysis, especially for DL methods as they normally require a significant amount of annotated data. Also, we decided to compare this approach with a widely implemented POS tagging method based on a probabilistic model, as they are suitable for low-resource settings.

Thus, we investigate the effectiveness of these methods utilizing varying lexicon and augmented corpus sizes, in increasing the annotation performance when very small training data is available.

3. Related work

3.1. Morphosyntactic analysis for low-resource languages

In recent years, there has been a growing emphasis on developing digital resources for regional languages in France, aiming to preserve and transmit them as integral part of the country's heritage (Bernhard et al., 2018a). After digitizing these resources, an annotation task is initiated. Since this is a very slow and costly process, it is often accompanied by different automated strategies. However, this task remains complex due to the limited resources typically available for these languages, and unlike French, they often lack of a standardized spelling. To address the issue of data scarcity in these languages, prior efforts have explored various approaches, such as adapting tools like Talismane for Occitan (Vergez-Couret and Urieli, 2015), crowdsourcing for Alsatian (Millour and Fort, 2018) or pre-tagging Alsatian texts with a close language like German with the support of linguists for manual correction (Bernhard et al., 2018b).
In other languages, probabilistic taggers have been

adapted to a close related language (Scherrer, 2014), addressed word alignment of indigenous languages to a high resource one (Ebrahimi et al., 2023) or even explored neural glossing (Cross et al., 2023) to improve the learning of morphological patterns and handle unknown words, an important feature in languages presenting variation.

3.2. Embeddings

Text representation algorithms yield impressive results, but they require significantly larger volumes of training data that are typically not available for low-resource languages. Pretrained word embeddings in a related high resource language have emerged as one efficient approach to address this problem due to the lack of sufficient data to train a model in the low-resource language. While (Jiang et al., 2018) and (Dunn et al., 2022) have attempted to address this issue by training embeddings directly for low-resource languages by reducing the corpus size to simulate a low-resource scenario, the size of the data remains large when compared to what is typically available for truly low-resource languages, given the limited digitalized content that is available. As a consequence, it is neither possible to evaluate their reliability.

4. Resources for Poitevin-Saintongeais

4.1. Lexicon

The first effort to increase the number of linguistic resources for Poitevin-Saintongeais began with the creation of of a compact lexicon containing inflected grammatical forms and conjugated auxiliary verbs, all standardized in spelling. To increase its lexical coverage, the lexicon was expanded through the extraction and transformation of an online bilingual dictionary (Pivetea, 2019). This dictionary covers both French and Poitevin Saintongeais, featuring around 23,000 entries. It provides some morphological information, which was formatted according to the Universal Dependencies guidelines, and includes information such as Gender, Number, Tense, VerbForm and PronType. While the dictionary naturally lacks of inflected forms, it provides their possible realizations (mainly in number and gender) facilitating the addition of inflected forms with a few preprocessing operations. After this step, we counted 41,047 forms in the lexicon (see Figure 1) with a predominance of nouns, adjectives and verbs (mostly infinitives).

4.2. Unannotated corpus

Unannotated corpora have been already gathered in a previous project. The text base for Poitevin-Saintongeais (Dourdet et al., 2019) contains more than 125 bibliographic references for literary texts.

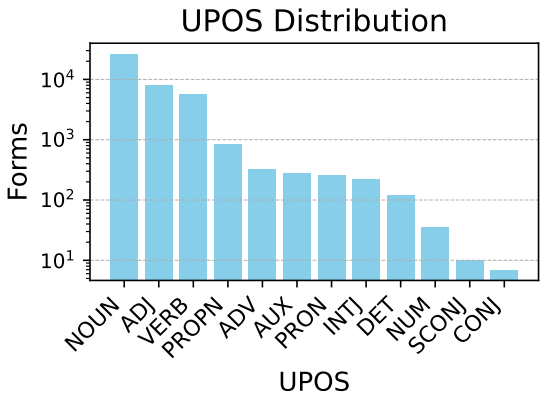

Figure 1: UPOS distribution of the lexicon (log scale).

UPOS	Count	Percentage (%)
NOUN	327	13.6
PRON	314	13.1
VERB	265	11.0
DET	223	9.3
ADP	212	8.8
ADV	151	6.3
ADJ	96	4.0
CCONJ	80	3.3
SCONJ	63	2.6
AUX	62	2.6
NUM	46	1.9
ADP+DET	46	1.9
PROPN	29	1.2
X	4	0.2
INTJ	5	0.2
	Total tokens: 2,399	
	Total Types: 806	

Table 1: Corpus size and UPOS distribution

These textual resources are characterized by different spellings, where only a few use the standard spelling.

4.3. Annotated corpus

The baseline dataset comes from a corpus that contains 2.4k tokens (see Table 1), each of which has been manually annotated with part-of-speech information (UD), lemma and French glosses. As displayed in Table 2, it is a very small corpus and it only consists of narrative texts. Since it is a very small corpus, we considered relevant to provide detailed information such as the proportion of the different UPOS, in order to understand the coverage of each in the data when evaluating the results.

5. Approach

Although recent neural methods for morphosyntactic annotation have mostly marginalized the role

	Sentences	Tokens
Train	100	1,700
Dev	21	360
Test	22	390

Table 2: Data partition size

of lexicons, LR languages do sometimes dispose of these resources in different proportions, which constitute a valuable asset for them as it can supplement the learning process with additional knowledge as noted in (Haddow et al., 2022). Here, we describe how we make use of the lexicon to perform a data augmentation technique to handle the data sparsity and evaluate its impact in the annotation accuracy. Our main objective is to evaluate to which extent they are beneficial. Thus, we evaluate how expanding the corpus size affects the model's performance when incorporating varying proportions of lexicon information.

Initially, we trained the models with the base corpus. Subsequently, we systematically increased the size of the annotated corpus to evaluate model performance across different lexicon proportions. For TreeTagger, we employed a step-wise approach, beginning with the base lexicon (training corpus) and progressively increasing its internal tagging lexicon. For Flair and Pie, we implemented a distributional neighbor replacement (DNR) method as described in section 5.2.

5.1. Dataset

We used the annotated corpus as a baseline for the experiments by dividing it into 5 distinct folds. This segmentation allows us to make the most of the little available data ensuring robustness and reliability in our evaluation. The data was segmented at the sentence level, distributing random sentences across the respective training (70-80%), development (15-10%), and test (15-10%) sets. The sentence boundaries were determined by periods, exclamation marks, and colons. The selected texts in this corpus are written in standardized spelling. Table 2 shows the number of sentences for each set and the corresponding average token count.

5.2. Data Augmentation

We performed a distributional neighbour replacement to augment the training and development sets.

The task of distributional neighbor replacement involves the substitution of tokens in a text with others that share a similar meaning based on their contextual patterns observed in a large corpus. This unsupervised approach can help to reduce the sparsity of the training corpus and improve the generalization ability of a model when the data is very scarce, especially for under-represented classes. However, it can introduce noise into the corpus if the replaced words are not semantically close or adapted to the semantic and syntactic context. To handle that, we verify whether each candidate aligns with an existing entry in the lexicon and matches the same UPOS tag. A sample is provided in Table 3.

Input Sentence (pos)
Lés movement de quéle armàie étiant réglläe coume qués d'in balét d'opéra.
Translation (fr)
Les mouvements de cette armée étaient réglés comme ceux d'un ballet d'opéra.
Proposed Sentences
Lés movement de quéle énfanterie étiant réglläe coume qués d'in capucine d'opéra.
Lés trvirajhes de quéle armàie furant réglläe coume qués d'in courente d'opéra.
Lés vrlitour de quéle armàie sirant réglläe coume qués d'in dance d'opéra.

Table 3: Sample of generated sentences in Poitevin-Saintongeais (pos) and French translation (fr)

To accomplish this, we have first leveraged the French glosses available in the gold annotations and identified, for each, the top-20 most similar words using the French FastText embeddings (Bojanowski et al., 2016) with a similarity threshold set to $\geq$ 0.60. Secondly, for each proposed neighbour, we searched for corresponding French glosses in the lexicon. Thirdly, for every instance, if a word matched and the UPOS corresponded to the same category, the group *token-upos-lemma* in the gold data was replaced with the new neighbour candidate, therefore generating new sentences. To prevent inconsistencies, certain categories were excluded since they were already well-represented in the corpus. These categories include grammatical classes such as SCONJ, CCONJ, DET, ADP+DET, ADP and PRON.

It is important to highlight that while glosses in the gold corpus are naturally inflected, this is not always the case in the lexicon, as it mainly originates from a dictionary. This disparity may reduce the ability to find a match for inflected forms. Consequently, verbs are the forms that have found the fewest equivalents, except for those manually conjugated in the source lexicon, which primarily correspond to auxiliary verbs.

Table 4 illustrates the gradual augmentation of the dataset. In TreeTagger, it presents the various proportions of lexicons integrated into the tagger lexicon. For TreeTagger, we follow this approach as we aim to evaluate the extent to which the size of a

	TreeTagger	Flair/Pie	
	Lexicon	Train	Dev
10%	4,519	∼200	∼57
25%	10,242	∼430	∼117
50%	19,726	∼730	∼210
75%	29,083	∼990	∼290
100%	38,376	∼1,200	∼360

Table 4: Number of added unique forms (Pie/Flair) and TreeTagger lexicon size

lexicon benefits us when dealing with a very small corpus. This allows us to make a comparison with the distributional neighbour replacement approach implemented in Flair and Pie, which involves the insertion of new *token-upos-lemma* groups. For Flair and Pie, we deducted the corresponding percentage of replacements for each UPOS on each iteration. For example, we keep 50% of the the proposed neighbouring words for NOUN, and so for each category. The same rule is applied on the TreeTagger lexicon, keeping only a fixed proportion of lexical entries for each UPOS.

6. Experimental setup

6.1. Methods

As suggested by (Wiechetek et al., 2021), it is advisable to first investigate traditional machine learning methods when dealing with small datasets before turning to deep learning. Traditional machine learning models are lightweight in terms of computational resources compared to modern deep learning-based models and relatively more efficient in such scenarios. Hence, we used the following tools in our experiments:

TreeTagger (Schmid, 1994) TreeTagger is a Markov Model based tagger constituting one of the most common probabilistic models for POS tagging. Operating on the principle that current tags hinge on recent words and tags in the sequence, it captures the transition probabilities between POS tags, indicating the likelihood of transitioning from one tag to another in the sequence. Also, it integrates a lexicon that maps the words to their corresponding POS tags. When encountering a word with multiple potential POS tags, it employs contextual cues from neighboring words and tags, in conjunction with the computed probabilities, to resolve ambiguity and determine the most probable POS tag for the word. TreeTagger achieves satisfactory results without needing extensive annotated data.

Flair (Akbik et al., 2019) It represents one of the current state-of-the-art frameworks for sequence tagging. A notable strength of Flair lies in its extensive collection of pretrained embeddings, which can be used for cross-lingual transfer learning. This means that even if a particular language has limited labeled data available, Flair can leverage the knowledge gained from well-resourced languages during training, thereby benefiting resource-constrained settings for etymologically related languages.

Pie (Manjavacas et al., 2019) Pie is a specialized deep learning framework designed specifically to handle languages with spelling variation. We incorporated it in our analysis to additionally include an evaluation of its lemmatization capability when feeding different proportions of lexicon information.

6.2. Embeddings

In Flair, we combined two types of embeddings:

(a) **Contextual string embeddings** (`FlairEmbeddings`). We integrated `fr-backward` and `fr-forward` embeddings into the Flair embedding constructor, both of which were trained on the French Wikipedia as it is an etymologically related language, which belongs to the same family of Language d'oïl.

(b) **Character embeddings**. We observed that including character-level embeddings improved substantially the model accuracy during parameter optimization.

We additionally trained a custom embedding which we integrated to the French embeddings. Initial experiments demonstrated that relying solely on these embeddings resulted in lower accuracy. However, when used in conjunction with the French embeddings, they yielded a modest improvement in the macro-level performance.

(c) **Custom Flair Embeddings**, trained from scratch using solely texts in Poitevin-Saintongeais. We integrated this embedding to the French ones.

For Pie, we trained a word2vec model, besides the limited number of tokens as it only allows Word2Vec as input embeddings. Both word2vec model in Pie and the custom embeddings on Flair, were trained over 32k sentences (∼700k tokens) using a number of available texts from the Telpos database that were already clean and in an machine-accesible format. However, it must be pointed out that they include texts from various genres, dialects and spellings.

6.3. Training parameters

TreeTagger was trained with the following parameters: Tagging context length `-cl`=2, decision threshold `-clg`=60. For **Flair**, we

	TT	Flair	Pie
baseline	0.773	0.813	0.743
10%	0.777	0.833	0.774
25%	0.793	**0.842**	0.762
50%	0.805	0.837	0.767
75%	0.824	0.838	0.700
100%	**0.846**	0.839	**0.785**

Table 5: UPOS F1-score (micro)

	TT	Flair	Pie
baseline	0.636	0.754	0.572
10%	0.641	0.753	0.665
25%	0.665	**0.776**	**0.669**
50%	0.667	0.774	0.666
75%	0.703	0.773	0.652
100%	**0.729**	0.773	0.666

Table 6: UPOS F1-score (macro)

used the best parameter combination: `lr`=0.15, `hidden_size`=64, `mini_batch_size`=16, `rnn_layers`=2, `dropout`=0.0263. **Pie** was trained using custom pretrained embeddings, a CRF decoder and `lr`=0.001.

6.4. Evaluation metrics

We have used F1-score (micro and macro) as metrics to assess the predictions over the average on the 5 folds. While POS tagging tasks usually contain imbalanced class distributions, as certain parts of speech are naturally more prevalent than others, the situation becomes more pronounced in very small datasets. Thus, macro is an important metric for us as it considers the performance on each independent class. Micro-F1 score calculates the overall performance by considering the total number of true positives, false positives, and false negatives across all classes, providing equal weight to each instance. On the other hand, macro-F1 score calculates the average F1 score for each class separately and then averages these scores, giving equal weight to each class, regardless of its prevalence in the dataset. Punctuation was omitted from evaluation to avoid a virtual increase of the scores.

7. Experimental results

Tables 5, 6 and 7 show the results for the POS tagging and lemmatization tasks. It presents the baseline results (only corpus) and the gradual increase in corpus size for Flair and Pie through DNR, along with the different lexicon sizes passed to TreeTagger. Both the results for PIE and FLAIR involve the utilization of custom embeddings. However, for FLAIR, their integration did not yield any improvement. In fact, the predictions were poorer when relying solely on these embeddings. In contrast, PIE exhibited a slight macro-level improvement, and thus, we decided to keep them.

In order to better understand how well the models generalised in every augmented dataset, we display in Figure 2 the percentage of unknown words from test in the training data.

	TT		Pie	
	F1-micro	F1-macro	F1-micro	F1-macro
baseline	0.829	0.514	0.881	0.361
10%	0.833	0.525	0.916	0.463
25%	0.830	0.519	0.910	0.426
50%	0.841	0.548	0.925	0.533
75%	0.848	0.568	0.917	0.488
100%	**0.852**	**0.579**	**0.926**	**0.549**

Table 7: Lemma F1-score (PUNCT excluded)

7.1. Task 1: POS

In this task, the obtained results show that all the tools benefited from leveraging the lexicon, but to a different extent. Overall, the best result combining F1-micro and F1-macro scores are achieved by Flair. While micro score is slightly increased in TreeTagger (0,846), we found that we require a much larger lexicon size to approach the performance of Flair (0,842) when using only a 23% of the word neighbours. This represents a significant disparity between the number of added forms observing table 4: TreeTagger at 100% contains 38k lexical forms, while Flair was fed only with only 430 new words for training and 117 for validation. Although results are lower for Pie, we consider it reasonably satisfactory considering the extremely small corpus size and, in contrast to Flair, it does not benefit from a large language model transfer learning. Also, Pie particularly benefits from the augmented corpus, as both metrics increase 0.094 (macro) and 0.042 (micro) points respectively from the baseline.

Figure 2 illustrates the proportion of out-of-vocabulary words (OOV) in the training and development sets when new lexical information has been incorporated. In the case of TreeTagger, this pertains to the annotation lexicon, while for Pie and Flair, it corresponds to the number of new lexical items integrated by DNR. When it comes to neural models, they are not sensitive to OOV. In fact, the introduction of new words enhances their ability to generalize and improve precision across various categories, as demonstrated in Table 6.

In section 9, we provide a finer analysis of the the errors for each task with a focus on POS tagging.

7.2. Task 2: Lemma

This task is more complex given the size of our corpus, as lemmatization involves mapping each word to its base or dictionary, which translate into a high proportion of OOV. However, we wanted to evaluate lemma as secondary task to investigate as well the effect of the transferring different lexicon sizes.

TreeTagger As it is well known for TreeTagger, its lemmatization ability relies on the POS tagging task. Therefore, when encountering an OOV word, it either provides no lemma or returns the token. This implies that the accuracy of this task depends on whether the training data or lexicon includes the lemma associated with a known token. As shows in Table 7 and demonstrated by Figure 2, the lower is the proportion of OOV words, the higher the f-scores. The low macro is explained by the predominance of inflected lexical forms, where the tagger is more likely to find unknown words.

Pie Pie exhibits high micro but a low macro. This discrepancy is explained by several factors: Pie's framework is based in a label encoder-decoder architecture. In this context, as we examine the errors in lemmatization, we observe that these different inflectional paradigm of some classes are not properly modeled due to the rich inflectional paradigm of the language, which may not be well represented when the corpus is very small. The number of lexical categories where inflection is frequent and diverse, as is the case of nouns (*abolle*, *pilai*, *vilajh**ous***, *aprenti**ve***, *aprenti**i***), adjectives (*abena**i***, *aribaud**ae*** *acti**ve***, *afaer**ous***, *ajhoufr**i***) and verbs (*abatardi*, *boulit**ae***), are more prone to errors. In fact most of the errors manifested in these categories. Notably, inflected verbs were not sufficiently represented in the lexicon and in the training data.

When it comes to the replacement of distributional neighbors, it would be necessary to investigate how many of these newly inserted words contribute to providing a more comprehensive representation of these various paradigms. Also, while Pie requires a larger dataset for optimal results, we hypothesize that it is essential for the inflectional paradigms of these categories to be well-represented in the augmented corpus to improve the models capability for word reconstruction. Another problematic category is PROPN that Pie fails to model accurately. In fact, proper nouns obtain the lowest macro score in 4 out of the 5 sets, as the model reconstructs the lemma while it is invariable. Moreover, Pie enriches lemmatization using sentence context information, but in our study by performing DNR over the same sentences, we maintain the same contextual information which could contribute to a reduced macro.

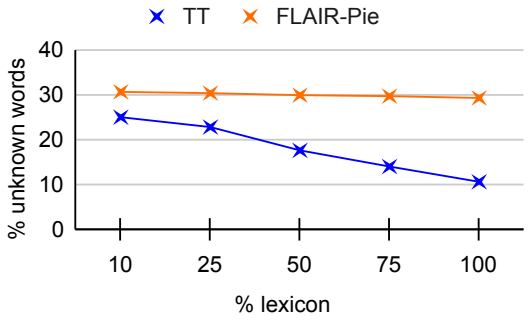

Figure 2: Coverage of OOV per lexicon size

8. Error analysis

8.1. TreeTagger

As shown in Tables 5 and 6, there is a progressive increase in the F1-macro and F1-micro scores for TreeTagger. This observed trend is expected as TreeTagger's performance is partly reliant on its lexicon to handle unknown words. The gradual improvements in both F1-macro and F1-micro underscore the significance of lexicon enrichment in enhancing the tagging scores. However, the lower macro indicates that the model's performance varies significantly across different classes. While the improvement is consistently seen in all classes, there are certain categories where the rate of errors, particularly VERB, is notably higher. We attribute this phenomenon to the absence of inflected forms in the lexicon (as it isn't the case for AUX), highlighting the inherent limitation of TreeTagger at dealing with OOV words. Even if it is supported by the lexicon, we hypothesize that the insufficient training data does not allow the model to properly capture the relationships between certain tags.

Also, we noted that its performance is very sensitive to tagging inconsistencies and ambiguity. In this sense, it has difficulty at distinguishing between VERB and AUX, as auxiliary verbs also appear as verbs in the corpus. For example, the form *oghisse* from the verb *to have* which occurs as both VERB and AUX. Addressing this issue would necessitate an better representation of specific syntactical structures, such as AUX+VERB in this context. The same tagging decision problem is persists among ADP/ADV, PRON/DET and ADJ/NOUN due to a significant overlap of tokens appearing in both classes.

8.2. Flair

With Flair, we observe that both macro and micro benefit from the addition of new lexical items.

But a larger lexicon does not lead to a greater improvement. In fact, improvement stagnates at 25%. On one hand, this indicates that reasonably good scores can be achieved with a limited lexicon. On the other hand, it raises questions about whether using the same sentences even if with different neighbors could pose a challenge, as the model is learning from the same syntactic structure repeatedly. The same observation was made for Pie.

Very rare categories such as X and INTJ were poorly predicted. The DNR strategy could not match any lexical entry for these forms, so these classes remained under-represented, leading to a lower macro. While to a lesser extent than TreeTagger, Flair was also affected in distinguishing between auxiliaries and non-auxiliary verbs. Moreover, tokens belonging to different categories account for the majority of errors made by Flair, notably between DET and PRON, where forms like *le* can function as both definite articles and personal pronouns, and *que* (PRON and SCONJ).

8.3. Pie

Most of the errors in Pie predictions correspond to the classes PRO/DET and VERB. This correspond to the same type of errors found in TreeTagger and Flair, where we encounter several tokens appearing in different classes. However, we should note that the model relies on sentence context to make predictions. If the context that is provided in the training data is insufficient, the model may not have the information it needs to distinguish between the different classes. Thus, for the DNR strategy, we estimate that replacing tokens with neighbors consistently and without considering to feed some variations in sentence context, the model is receiving identical contexts for different words. This homogeneity in context may limit the model's ability to differentiate between words with different meanings or parts of speech when tagging unseen texts that reflect different contexts.

Also, while the model has the ability to handle OOV words, the replacement strategy feeds it with a reduced number of new lexical forms. In the case of PRO and DET, we avoided the replacement operation as they are not very diverse (they constitute a fixed number of forms) and were well represented in the corpus. Also, we wanted to avoid introducing syntactic inconsistencies. PIE tries to maximize the probability of the target character sequence (Manjavacas et al., 2019), it is important to note that Pie does not benefit from external knowledge via large pre-trained language models, as it is the case of Flair. This causes the model to struggle with less frequent UPOS.

9. Conclusions and Future work

This work shows the feasibility of leveraging a lexicon to the advantage of a corpus with a very scarce number of annotations. Moreover, using transfer learning, the significance of the required lexicon is minimized, which offers an advantage for low-resource languages as such resources are either not available or very scarce. Nevertheless, the effectiveness of our strategy depends on two factors: first, having glosses in a major language already available, and second, the number of possible matches between the similar distributional neighbors found in the lexicon.

Traditional tools like TreeTagger continue to demonstrate their utility for low-resource languages, as indicated by the results. Nevertheless, it requires a lexicon to be available, particularly for a task like lemmatization, in order to handle out-of-vocabulary words which are frequent in smaller corpora.

We also point the necessity of enhancing the representation of certain syntactic structures in the training data to improve the quality of the pos-tagging. This will require a more in-depth analysis in future work. Also, improving the performance for these tools at this stage, involves expanding the lexicon, particularly with conjugated verbs. This also includes augmenting the lexicon for word forms that correspond to multiple lexical or grammatical units (for example, auxiliaries that can also function as verbs, or the adverb *si* which can also be a conjunction). Utilizing larger annotated corpora with more context can help mitigate ambiguities with polysemous words, such as *la*, which can be both a determiner (DET) or a noun (NOUN).

On the other hand, improving the performance of neural models at this stage involves increasing the annotated corpora and training the embeddings on a larger corpus of texts in normalized spelling. While it seems that the lexicon has exhausted its potential in terms of quantity, there is room for quantitative improvement, particularly with conjugated verbs. Expanding the experiment with distributional neighbors while respecting morphosyntactic features is another avenue. However, we do not have access to this information in the annotated corpora, which would require finding a way to provide finer morphological information to perform more accurate replacements.

Finally, we hypothesize that enhancing the representation of diverse inflectional paradigms could facilitate the model in capturing the language's internal morphology more effectively. In the case of nouns and adjectives, there exists a multitude

of shared inflection types, and we believe that an over-representation of a particular inflection within one category could potentially disadvantage another category where the same inflection appears in a word. A collection of inflectional paradigms is currently underway to address the shortage of conjugated verbs in the lexicon, which should help the model to better capture the inner morphology of the language.

The objective of this study has been to assess the extent to which lexicon can be a valuable resource when other assets are unavailable for a less-resourced language. In doing so, we aimed to evaluate how more traditional methods can be helpful without the need for extensive data, which is typically required by neural models. However, the latter benefit from this approach without requiring a substantial lexicon, making it a viable strategy for improving accuracy across different classes. This is one feasible step to increase the number of annotated texts, and eventually, we believe that considering an ensemble method to validate such annotations could be useful to facilitate the augmentation of the annotated corpora for these languages.

10. Acknowledgements

This work has been carried out within the framework of the DIVITAL project (ANR-21-CE27-0004) supported by the Agence National de la Recherche. We would like to give special thanks to Vianney Piveteau for granting us permission to utilize the dictionary. This valuable resource has not only facilitated the development of an exploitable lexical resource, but has also enabled us to conduct our initial experiments for this language.

11. Bibliographical References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Delphine Bernhard, Anne-Laure Ligozat, Fanny Martin, Myriam Bras, Pierre Magistry, Marianne Vergez-Couret, Lucie Steiblé, Pascale Erhart, Nabil Hathout, Dominique Huck, Christophe Rey, Philippe Reynés, Sophie Rosset, Jean Sibille, and Thomas Lavergne. 2018a. Corpora with part-of-speech annotations for three regional languages of France: Alsatian, Occitan and Picard. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Delphine Bernhard, Anne-Laure Ligozat, Fanny Martin, Myriam Bras, Pierre Magistry, Marianne Vergez-Couret, Lucie Steiblé, Pascale Erhart, Nabil Hathout, Dominique Huck, Christophe Rey, Philippe Reynés, Sophie Rosset, Jean Sibille, and Thomas Lavergne. 2018b. Corpora with part-of-speech annotations for three regional languages of France: Alsatian, Occitan and Picard. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Ziggy Cross, Michelle Yun, Ananya Apparaju, Jata MacCabe, Garrett Nicolai, and Miikka Silfverberg. 2023. Glossy bytes: Neural glossing using subword encoding. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 222–229, Toronto, Canada. Association for Computational Linguistics.

Jean-Christophe Dourdet, Marianne Vergez-Couret, and Marie-Helene Lay. 2019. Telpos - texte electronique en poitevin-saintongeais, enjeux et difficultes. In *Colloque "Langues minoritaires" : quels acteurs pour quel avenir ?*, Strasbourg, France.

Jonathan Dunn, Haipeng Li, and Damian Sastre. 2022. Predicting embedding reliability in low-resource settings using corpus similarity measures. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6461–6470, Marseille, France. European Language Resources Association.

Abteen Ebrahimi, Arya D. McCarthy, Arturo Oncevay, John E. Ortega, Luis Chiruzzo, Gustavo Giménez-Lugo, Rolando Coto-Solano, and Katharina Kann. 2023. Meeting the needs of low-resource languages: The value of automatic alignments via pretrained models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3912–3926, Dubrovnik, Croatia. Association for Computational Linguistics.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of Low-Resource Ma-

chine Translation. *Computational Linguistics*, 48(3):673–732.

Chao Jiang, Hsiang-Fu Yu, Cho-Jui Hsieh, and Kai-Wei Chang. 2018. Learning word embeddings for low-resource languages by PU learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1024–1034, New Orleans, Louisiana. Association for Computational Linguistics.

Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. 2019. Improving lemmatization of non-standard languages with joint learning. 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)*, pages 1493–1503, Minneapolis, Minnesota. Association for Computational Linguistics.

Alice Millour and Karën Fort. 2018. Toward a lightweight solution for less-resourced languages: Creating a POS tagger for Alsatian using voluntary crowdsourcing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Vianney Pivetea. 2019. *Dictionnaire français poitevin-saintongeais*. Geste.

Yves Scherrer. 2014. Unsupervised adaptation of supervised part-of-speech taggers for closely related languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 30–38, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

Marianne Vergez-Couret and Assaf Urieli. 2015. Analyse morphosyntaxique de l'occitan languedocien : l'amitié entre un petit languedocien et un gros catalan. In *TALARE 2015*, Caen, France.

Linda Wiechetek, Flammie Pirinen, Mika Hämäläinen, and Chiara Argese. 2021. Rules ruling neural networks - neural vs. rule-based grammar checking for a low resource language. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1526–1535, Held Online. INCOMA Ltd.