Agettivu, aggitivu o aghjettivu? POS Tagging Corsican Dialects

Alice Millour, Laurent Kevers, Lorenza Brasile, Alberto Ghia

Agettivu, aggitivu o aghjettivu? POS Tagging Corsican Dialects

Alice Millour[1], Laurent Kevers[2], Lorenza Brasile[2,3], Alberto Ghia[2,3]
[1] LIASD - Pastis, Université Paris 8 Vincennes Saint-Denis (France)
[2] LISA UMR 6240 - Université de Corse (France)
[3] DIRAAS - Università di Genova (Italy)
[4] StudiUm - Università di Torino (Italy)
am@up8.edu, laurent@kevers.org,
brasile_l@univ-corse.fr / lorenza.brasile@edu.unige.it, albertoghia.mg@gmail.com

Abstract

In this paper we present a series of experiments towards POS tagging Corsican, a less-resourced language spoken in Corsica and linguistically related to Italian. The first contribution is CORSICAN-POS, the first gold standard POS-tagged corpus for Corsica, composed of 500 sentences manually annotated with the Universal POS tagset. Our second contribution is a set of experiments and evaluation of POS tagging models which starts with a baseline model for Italian and is aimed at finding the best training configuration, namely in terms of the size and combination strategy of the existing raw and annotated resources.These experiments result in (i) the first POS tagger for Corsican, reaching an accuracy of 93.38 %, (ii) a quantification of the gain provided by the use of each available resource. We find that the optimal configuration uses Italian word embeddings further specialized with Corsican embeddings and trained on the largest gold corpus for Corsican available so far.

Keywords: POS tagging, less-resourced languages, Corsican

1. Introduction and Objectives

This work[1] is part of the initiative to provide a Basic Language Resource Kit (BLARK, Krauwer (2003)) for Corsican (Kevers and Retali-Medori, 2020), a less-resourced language spoken in Corsica. Following previous work introducing the existing resources and tools existing for Corsican (Kevers et al., 2021) and CoSwID, a language identification tool (Kevers, 2022), we present the recent advances regarding POS tagging Corsican.

The purpose of this article is not only to present a new state of the art for this task, but also to explore the benefits provided by the use of the existing resources. In fact, in a scarce context both in term of linguistic datasets and human resources to build them, putting the effort in the right place is critical. We thus lead a comparative study to answer questions such as: is it more beneficial to go with unsupervised training or is it worth it to annotate a small training corpus? Or else, how much training data do you need in a context where you can rely on resources and models for a closely related language?

The paper is organized as follows: first, we present the existing works and resources for Corsican dialects, as well as the recent methods to POS tag varied languages in a scarce resource context. Then, we present the process of annotation that lead to the creation of the first gold POS tagged dataset for Corsican dialects. In a fourth section, we describe the series of experiments we lead, which involve both unsupervised and supervised training, and in which we vary: the size and language of the datasets used to train embeddings, and the size and language of the training corpus.

2. Previous work

2.1. Existing Language Resources for Corsican

Corsican is a continuum of four to five dialectal variants spoken mainly on the island of Corsica, and fits naturally into the Italo-Romance family (Ledgeway, 2016; Dalbera-Stefanaggi, 2002). Its spelling is not standardized by a consensual norm. Kevers et al. (2021) and Kevers and Millour (2022) provide a review of the existing resources and initiative to progress towards the digital inclusion of Corsican. Among these resources is the first open source and accessible digital corpus for Corsican[2], published in 2019. The sources used for this corpus are: (i) a dump of the Corsican Wikipedia extracted in 2019 (919K words), (ii) a translation of the Bible by Christian Dubois (500K words), (iii) and a series of articles published on a blog between 2010 and 2019[3] (770K words).

[1]The paper was conceived and written collectively by all the authors. Notwithstanding, for academic purposes the paragraphs §§1., 2., 3.2., 4., 5. are to be attributed to A. Millour and L. Kevers, §3.1. to A. Ghia, and §3.3. to L. Brasile.

[2]See: https://bdlc.univ-corse.fr/tal/index.php?page=res.

[3]See: https://www.apiazzetta.com/.

Each of these corpora is highly peculiar in terms of content, authorship or both. Although being diverse, this corpus lacks representativeness especially in terms of writing practices. This is why it was decided in 2022 to conclude a convention with the Corsican branch of the Canopé network, which is a public instance of the National Education Ministry. The instance provides documents including teaching material, pieces of literature (for both youth and adults), and documentation about Corsican history and patrimony. This curated corpus, the *Corpus Canopé de Corse* (CCdC, Kevers and MacLean (2023)), complements the existing resource with about 500K word. In the following, we will refer to the complete available corpus as C_{Cos}.

2.2. Previous work on POS tagging less-resourced non standardized languages

Recent works on POS tagging in a scarce resource context are mainly based, depending on the resources at disposal, on cross-lingual learning or cross-lingual transfer. Kann et al. (2020) have shown that, although weak supervision appears as a convenient perspective when resources are scarce, systems perform poorly in truly low-resourced languages. Similarly, the experiments of Laméris and Stymne (2021) led on POS tagging Scots show that having at disposal a small annotated corpus was very beneficial with respect to zero-shot annotation from English. We reproduce the methodology they proposed, using Italian as a transfer language for Corsican.
Tagging non-standardized languages presents an additional challenge: the multiplicity of dialectal forms is not standardized by a consensual spelling, hence numerous forms coexist, causing the number of out-of-vocabulary words to explode. In such contexts, various strategies can be tested. One of them is the use of a character-level analysis combined with morphosyntactic properties of a target word and its context as experimented by (Magistry et al., 2019) showing good results on three regional languages. Yet, this method has shown limitations to encompass dialectal variation. Other options include using normalization (experimented successfully for instance for Finnish dialects (Partanen et al., 2019)), or variation patterns integration (see for instance the work of Millour and Fort (2019) based on crowdsourced spelling variants), yet these methods require knowledge of the variation mechanisms or having at disposal parallel data, two resources that are not yet available for Corsican.

Works on less resourced non standardized languages globally insist on the necessity of building curated resources, for both unsupervised and supervised training, and evaluation. This last point is particularly important, since in a context of dialectal and spelling variation, limiting the evaluation to a partial sample of the language may lead to unreliable results, undermined as soon as the systems are used out of their training/evaluation initial context.

3. Corsican-POS: a gold POS-tagged corpus for Corsican

3.1. Methodology

We describe in this section the methodology used to create Corsican-POS, from corpus collection to curation of the annotations.

3.1.1. Corpus

The annotated corpus comprises 220 sentences from the Corpus Canopé de Corse; 20 sentences from the Banque de Données Langue Corse (BDLC, Corsican Language Database) corpus ; 60 sentences from three corpora in XML TEI format (Kevers and Retali-Medori, 2020): *A Sacra Bibbia* (20 sentences), *Wikipedia, enciclopedia libara in lingua corsa* (20 sentences), and *A piazzetta, giurnale in lingua corsa* (20 sentences); and finally 200 sentences taken from Gino Bottiglioni's *Atlante Linguistico ed Etnografico Italiano della Corsica*, the ALEIC (Bottiglioni, 1933 - 1942). The resulting corpus of 500 sentences is thus quite varied in terms of the type of source (oral, written), textual genres (literary, religious, historical, encyclopedic, educational, informational contents), length and complexity of the sentences, and geographical origin (dialectal varieties are represented).

This last point is particularly important for Corsican, since in its standardisation process it has embraced 'polynomy', *i.e.* the unity of the Corsican language "is an abtract concept which is the result of a dialectal movement rather than of the ossification of a single norm, and which existence relies on the massive decision of whom speak it to give it a particular name and declare it autonomous from the other recognized languages.[4]" (Marcellesi, 1984). This implies the presence in the texts of phonetic and morphological variants for the same lemma, reflected in the spelling (*ziteddu/zitellu* 'child'; *croci/cruci* 'crosses'; *faci/face* '(he) does'),

[4] "*[...] dont l'unité est abstraite et résulte d'un mouvement dialectique et non de la simple ossification d'une norme unique, et dont l'existence est fondée sur la décision massive de ceux qui la parlent de lui donner un nom particulier et de déclarer autonome des autres langues reconnues.*" (Marcellesi, 1984), personal translation.

	Iter01	Iter02	Iter03	Iter04	Iter05
#tokens	1 792	1 824	943	1 325	900
Tagger	ITA	ITA	ITA	COS-0.1	COS-0.2
Accuracy	0.71	0.71	0.65	0.90	0.92

Table 1: Pre-annotation tools and their accuracies.

as well as geosynonyms (such as *cascia/fronda* 'leaf' and *cane/ghjacaru* 'dog').

3.1.2. Corpus preparation

Tokenization: The tokenization step was curated manually. In fact, according to the UD guidelines, the sentences should be split into "syntactic" words, including the separation of clitics and decomposition of contractions. No satisfying tokenizer exists so far for Corsican. This is mainly due to concurrent spelling practices that we observe for instance in the case of pronominal verbs that exist in their agglomerated form (eg. *spassassi* 'to have fun') and separated by a space (*spassà si*), or the irregular use of punctuation signs within tokens (*eg. cum'è*, which is the single token 'like'). The manually annotated corpus adds up to 6 784 tokens.

Pre-annotation: All the texts were pre-annotated before correction by human annotators. For the three first batches of 100 sentences, the pre-annotation was carried out by Flair (Akbik et al., 2019) trained on the largest Italian corpus available (tagger ITA, see section 4.1), Italian being the closest well-resourced language to Corsican[5]. As the campaign progressed, a Corsican tagger was trained on the gold corpus (see 4.3) to improve the quality of the pre-annotation in an iterative manner: COS-0.1 was trained on Iter01, COS-0.2 was trained on Iter01+Iter02. Table 1 summarizes the pre-annotation tools used as well as their accuracy (calculated afterwards).

3.1.3. Annotation campaign

The sentences to be annotated were divided into five series of one hundred sentences each. Specifically, series 1, 2 and 4 consisted of sentences extracted from the previously processed corpus of digital Corsican language tests (CCdC, BDLC, Wikipedia, etc.); series 3 and 5, on the other hand, consisted of ALEIC data only. The annotation campaign thus comprised five iterations, in three stages each: an automatic pre-annotation, a manual annotation and a meeting to review and comment on the previous work. The annotation process was entrusted to five annotators with different experiences and language skills, all affiliated to the NALC-BDLC project: a *Studi corsi* Master student, a research engineer, two PhD students in Linguistics and a Post-doc researcher in Linguistics. Two of them are native Corsican speakers while the other three are native Italian speakers. The process was supervised by a NLP research engineer.

The first four sets of sentences were annotated by all of the annotators, while the last set by three of them only. Each was given a batch of 100 pre-annotated sentences each time. Each sentence being processed by three to four different annotators, the choices were further compared and errors and inconsistencies limited. The pre-annotation appeared on one line, while the annotator had to insert his own annotation proposal on another line. Only for the fifth and final iteration did the annotator have to directly correct the pre-annotation, acting on the same line. Cases of different annotation of the same word by two or more annotators were then discussed in collective work sessions, in order to elaborate on the annotation choices and standardise them across the corpus. These reflections, as well as further and subsequent revisions and curation, finally led to annotation guidelines (currently kept for internal use only).

The annotation was carried out on the INCEpTION interface[6] as exemplified in figure 1 using the CONLL-U format. The POS tagset is from the Universal Dependencies framework[7]. We chose this tagset because it allows both a comparability of the data with those of other corpora and the possibility of benefiting from the prior experience of other languages, as well as the international visibility of the project and the potential perpetuity of the data due to the well-documented standard formats. Since the annotation manual for Corsican was only compiled after the annotation itself, the annotators began their task resorting to their personal metalinguistic expertise and, above all, using the following resources as a point of reference: the UD guidelines for each POS tag (with a special attention to specific indications for other Romance languages, especially French and Italian); the annotations made for corpora in other languages available on Grew-Match[8]; dictionaries (Marchetti, 2001), the INFCOR[9], and grammars (Durand, 2003; Comiti, 2011; Romani, 2005) of the Corsican language; the same tools made for Italian (Treccani and Nuovo De Mauro online dictionaries[10]; (Serianni, 2000, 2005)) and for French

[5] See: (Ledgeway, 2016).

[6] See: inception-project.github.io.

[7] See https://universaldependencies.org/u/pos/index.html.

[8] See: http://universal.grew.fr/.

[9] See: http://infcor.adecec.net.

[10] https://www.treccani.it/vocabolario/, https://dizionario.internazionale.it/.

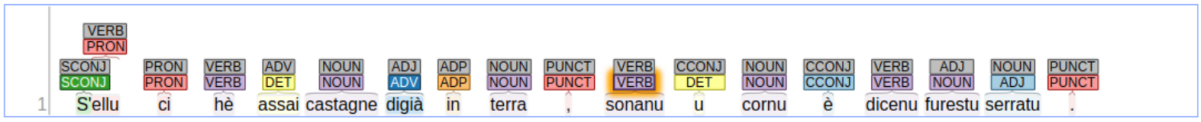

Figure 1: An annotated sentence: "If there are already chestnuts on the ground, they play the corn and say "free grazing forbidden" "; the gray layer corresponds to pre-annotation; the colored layer to manual annotation.

(Larousse online dictionary[11]; Trésor de la Langue Française informatisé[12]).

3.2. Results

Table 2 shows the statistics of the corpus in terms of tags representation. The PART tag was not used, as in other Romance languages[13].

We used Cohen's Kappa (Cohen, 1960) to evaluate the understanding of the guidelines among annotators (see Table 3). We observe that the agreement is good and that, as the campaign progresses, the coefficient of agreement globally increases.

Note that a potential problem in automatic POS tagging Corsican is its diatopic variation (see section 3.1.1). However, the dialectal variants did not pose any particular problems of interpretation, since their matching patterns are fairly regular (*eg.* <ll>/<dd>, or <nghj>/<gn>) and were well represented in the corpus. We observed that dialectal variation was fairly embraced by the pre-annotation tools which could properly handle spelling variants as well as some mistakes.

3.3. Discussion

Doubts and problems were encountered during the annotation of the gold corpus, both of a general nature and specific to not entirely standardised languages. With regard to the former, although we wanted to stick as closely as possible to the UD guidelines, in some cases we deviated from them. This is the case of complex proper names, i.e. made up of more than one element, such as *U Vescuvatu* ('the Bishopric', name of a municipality) and *Valli di i setti mulini* ('Valley of the seven mills', name of a place). Within UD, each of their components should be annotated according to its prototypical POS[14], disregarding it being a unique proper name in this context. This is inconsistent with the treatment of those proper names which are not as transparent, such as *Olmeta di Tuda*, *Penta di Casinca* or *Munacìa d'Auddè*: *Tuda*, *Casinca*, *Munacìa* e *Auddè* must be tagged as PROPN since they are exactly "name[s] (or part[s] of the name) of a unique entity, be it an individual, a place, or an object". Moreover, the researcher who will use the annotated corpus to search for proper names (PROPN) will have an obvious difficulty in finding the fully transparent compound ones (no part of which is tagged PROPN). Thus, the usefulness of the PROPN label, which itself represents a sub-category of the set of names (i.e., it concerns a group of tokens that would otherwise be labelled NOUN) escapes notice. It is worth remembering that annotation is not the *purpose*, but rather a *tool* at the service of researchers. Sometimes UD's annotation approach seems to lose sight of this goal, ending up facilitating the learning of the automatic annotator, instead of the research for potential users of the annotated corpus.

Participles pose some problems too, for their ambivalent nature in Romance languages, straddling verb and adjective. The word *aperta*, for instance, is a verbal form in a sentence such as *A latrina hè aperta da Ghjuvanni* 'the toilet is opened by Ghjuvanni' (passive diathesis, *hè* being AUX)[15], while it clearly is an adjective in *A latrina hè aperta da tempu* 'the toilet is open for a long time' (nominal predicate, to be compared with *a latrina hè grande* 'the toilet is big'), *hè* being a copula). The first case is labelled VERB, the second ADJ, indeed. However, in between there are numerous controversial cases: in the annotated corpus we find *mughji spavintosi diffusi in tutta a vaddi* 'frightful screams spread through the valley', as well as *i paisani, emuziunati è rispittosi* 'villagers, excited and respectful' and *una cudetta turchina bella strinta à a vita* 'a belt very tight to the waist (literally)'. There are indicators that can allow disambiguation: the presence of an agent (as if we had *diffusi da...* 'spread by...') or of an adverb (*diffusi bè...* 'well spread...') would enhance the verbal value. None of these indicators occur in

[11] https://www.larousse.fr/dictionnaires/francais/.

[12] https://www.cnrtl.fr/definition/.

[13] The usage of this tag is very limited in English, too (see languanges subsections in https://universaldependencies.org/u/pos/PART.html)

[14] https://universaldependencies.org/u/overview/morphology.html.

[15] As to copulas, and modal verbs as well, our choice was to tag them as VERBs, while UD suggests AUX.

Tag	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM
Proportion	3.53%	11.43%	5.60%	2.03%	2.50%	13.59%	0.29%	15.15%	1.26%
	PART	PROPN	PRON	PUNCT	SCONJ	SYM	VERB	X	
	0%	3.44%	8.35%	14.16%	2.80%	0.01%	14.75%	0.31%	

Table 2: Tags repartition in the CORSICAN-POS corpus.

	Iter01	Iter02	Iter03	Iter04	Iter05
Cohen's Kappa	0.88	0.89	0.93	0.91	0.95

Table 3: Inter-annotator agreement.

the controversial sentences above. Therefore, a second point worth remembering is that labels are - and cannot but be - rigidly flat entities that we try to apply onto a curved linguistic reality, rich in nuances and gradualness: it can be easy to make them adhere to one point, the prototypical one, but what strays from it remains uncovered.

The last case of ambiguous status is that of substantivated POS, as in *a u so piacè e u so volè* 'to its liking or willing', or *porta ancu u male* 'it brings the bad', or *facenu l'evvive à i sposi* 'they say the "horray" for the spouses', or *u perché* 'the why'. No indication about how to deal with these cases was found in UD guidelines. For our Corsican corpus, we have chosen to label them NOUN, according to the new POS assigned by the substantivation process. The same choice has been made for other corpora available on Grew-Match: in sentences like *degli indigeni e dei "bianchi"* (corpus ISDT@2.12), *il domani alle porte* (corpus VIT@2.12), *la mise à disposition* (corpus ParTUT@2.12), *un devoir* (corpus GSD@2.12), the words *indigeni*, *bianchi*, *domani*, *mise* and *devoir* have been tagged NOUN. Here, again, one may notice how simplistic it is to reduce these cases under a single tag, and therefore how useful it would be having the chance of a "double" tag.

4. Combining Resources to Find the Right Balance

The newly created resource enabled us to perform a series of experiments by using the gold corpus as an evaluation resource and a training resource for supervised systems and finetuning Italian taggers. We used the Flair (Akbik et al., 2019) tagger as our baseline and further model to finetune. The performance measure is the annotation accuracy, calculated using a 10-fold cross-validation. The optimal number of epochs and size of the embeddings were systematically explored. In the following we restitute the results of our experiments as answers to the questions that working on a less-resourced language raises.

4.1. Using a Closely-Related Language as a Baseline

Question 1: *When no resource is available, is it possible to take advantage of resources and models available for a closely related languages?*

Config.	Iter01-05
C_{Ita} – E_{Ita} / 10	**62.84**
C_{Ita} – E_{Ita} / 100	59.32

Table 4: Baseline tagger evaluation.

Italian being the closest language to Corsican, we use as a baseline the Flair Italian model which we retrained according to the defined experimental set-up. The corpus used for training and evaluation is the Italian Stanford Dependency Treebank, composed of 14 167 sentences adding up to 278 429 tokens. We used the 'it-forward' embeddings[16] generated from a corpus of 1.5G tokens extracted from the Italian Wikipedia. This model reaches a F-score of 0.98 on the Italian corpus[17], and the accuracy calculated on the CORSICAN-POS corpus once annotated reaches 62.84 %. Table 4 shows that in this configuration, short training performed best. Because of its relatively low performance, such a system is unusable in a real applicative context, yet proved to be a viable tool to perform pre-annotation and serves as a reasonable baseline for further experiments.

In the following, we perform various experiments to measure the gain brought by the resources available: (i) the CCdC corpus of 500K tokens, the C_{Cos} corpus of 2,7M tokens, and CORSICAN-POS as a training corpus on the Corsican side, and (ii) the Italian existing models and embeddings on the Italian side.

4.2. Using the Raw Curated Corpus to Train Embeddings in the target language

Question 2a: "*For a less-resourced language with a minimal raw dataset available, can we improve a tagset trained for a closely related language thanks to embeddings adapted to our target language?*"

[16]See: https://github.com/flairNLP/flair-lms.

[17]See: https://github.com/stefan-it/flair-pos-tagging#citing.

Config.	Iter01-05
Q1: Baseline	62.84
C_{Ita} – E_{CCdC}-1/10	50.07
C_{Ita} – E_{CCdC}-1/100	48.05
C_{Ita} – E_{CCdC}-2/10	67.83
C_{Ita} – E_{CCdC}-2/100	**68.80**
C_{Ita} – E_{CCdC}-3/10	63.81
C_{Ita} – E_{CCdC}-3/100	63.82

Table 5: Results for question 2a (Italian training Corpus and Corsican limited embeddings).

We set the first experiment in a very low resource context where only a corpus of around 500K tokens is available. We use the CCdC to train embeddings further combined with the Flair Italian POS tagging model. To enable further comparison, we do not use in this experiment the Italian embeddings. We compare configurations using the three options available in terms of embedding sizes: E_{CCdC}-1 (minimal), E_{CCdC}-2 (conventional), and E_{CCdC}-3 (extended), and comparing training duration: short (10 epochs) and long (100 epochs). Training is stopped after more than three consecutive epochs without improvement, meaning that the learning rate is too small. Results are reported in table 5. The best configuration is obtained with the conventional size and a long training which provides a gain of 18.73 points with respect to short training. With an accuracy of 68.80, the baseline is outperformed by 5.96 points. Using embeddings, *even trained on a limited dataset*, thus already brings a benefit.

Question 2b : *When no training corpus is available, what is the gain provided by adding data to the embeddings training corpus?*
In this second experiment, we use the C_{Cos} corpus to train the Corsican embeddings E_{Cos}.

Config.	Iter01-05	*vs* Q2a
Q1: Baseline	62.84	-
Q2: C_{Ita} – E_{CCdC}	68.80	-
C_{Ita} - E_{Cos}-1/10	32.98	−17.09
C_{Ita} - E_{Cos}-1/100	31.32	−16.73
C_{Ita} - E_{Cos}-2/10	**80.12**	+12.29
C_{Ita} - E_{Cos}-2/100	79.69	+10.89
C_{Ita} - E_{Cos}-3/10	**80.21**	+16.40
C_{Ita} - E_{Cos}-3/100	79.95	+16.13

Table 6: Results for question 2b (Italian training corpus. Corsican embeddings trained on C_{Cos}).

Results are reported in table 6 in which we provide the comparison with respect to the baseline and best result obtained earlier (C_{Ita} – E_{CCdC}). In this experiment the length of training has a smaller impact than in the limited training corpus setup. We also observe a significant drop when using the minimal embeddings configuration while the conventional and extended embeddings show a significant improvement to reach similar accuracy. As a conclusion to this experiment, an increase of the size and diversity oh the training corpus allows to reach an accuracy of 80 %.

Question 3: *Rather than using embeddings trained solely on the target language, can we use embeddings from a closely related language and specialize them with the available data for the target language?*

As a third experiment to take advantage of our raw corpus only, we specialize the Italian embeddings with C_{CCdC} resulting in the $E_{Ita-CCdC}$ embeddings, and on C_{Cos}, resulting in the $E_{Ita-Cos}$ embeddings. Only two configurations (minimal and conventional) were available in this setup. The results are reported in table 7.

Config.	Iter01-05
Q1 : Baseline	62.84
Q2b : C_{Ita} - E_{Cos}-3/10	80.21
C_{Ita} - $E_{Ita-CCdC}$-1/10	73.96
C_{Ita} - $E_{Ita-CCdC}$-1/100	73.73
C_{Ita} - $E_{Ita-CCdC}$-2/10	83.68
C_{Ita} - $E_{Ita-CCdC}$-2/100	**83.83**
C_{Ita} - $E_{Ita-Cos}$-1/10	74.34
C_{Ita} - $E_{Ita-Cos}$-1/100	74.21
C_{Ita} - $E_{Ita-Cos}$-2/10	84.27
C_{Ita} - $E_{Ita-Cos}$-2/100	**85.12**

Table 7: Results for question 3 (Italian training corpus, Italian embeddings specialized on the Corsican datasets).

With no manually annotated data, the best result is obtained using the largest corpus available (+1.19 point with respect to the $E_{Ita-CCdC}$ configuration) as an accuracy of 85.12 % is reached. As for question 2, the length of training is not decisive, and results with minimal embeddings are significantly lower. Note that this configuration is also interesting because it does not require heavy calculation, the closely related language embeddings being already trained.

As an intermediary conclusion, we have observed that using all the available raw data is beneficial. The gain brought by the use of the complete raw corpus at disposal (although being a limited increase in the size of the dataset) might be due to the increase in coverage on dialectal and spelling variants that it carries.

Config.	Iter01	Iter01-02	Iter01-03	Iter01-04	Iter01-05	*vs* Q2a	*vs* Q2b
size (sentences)	100	200	300	400	500		
Q2a : C_{Ita} – E_{CCdC}-2/100					68.80	0	-
Q2b : C_{Ita} - E_{Cos}-3/10					80.21	-	0
C_{Cos} - E_{CCdC}-1/10	26.84	38.61	46.68	49.82	52.37	−16.43	−27.84
C_{Cos} - E_{CCdC}-1/100	50.33	54.39	57.87	58.62	59.48	− 9.32	−20.73
C_{Cos} - E_{CCdC}-2/10	69.69	75.82	78.98	79.93	81.84	+13.04	+ 1.63
C_{Cos} - E_{CCdC}-2/100	**81.50**	83.59	83.90	85.83	**87.14**	+18.34	+ 6.93
C_{Cos} - E_{CCdC}-3/10	69.51	74.71	78.23	80.87	82.61	+13.81	+ 2.40
C_{Cos} - E_{CCdC}-3/100	**81.55**	83.59	84.61	86.29	**87.56**	+18.76	+ 7.35

Table 8: Results for question 4a. Models trained on the CORSICAN-POS corpus with embeddings trained on CCdC.

4.3. Using a Small Gold Corpus to Train a POS-Tagger

In this section, we explore the opportunities offered by having at our disposal a gold corpus in our target language.

Question 4a: *When annotated data is available for the target language, how much data is needed to dispense with the gold corpus of a closely related language and match the baseline?*
To answer this question, we replace the Italian training corpus (C_{Ita}) by CORSICAN-POS (C_{Cos}), and train embeddings on the CCdC corpus to match the conditions of question 2a.

We present in table 8 the evolution of the performances as the Corsican training corpus grows. In this configuration, it is always more beneficial to use a long training of 100 epochs. Similarly to the previous experiment, small embeddings show the poorest results while conventional and extended ones are close. As for the required size of training dataset, we observe that with embeddings trained on 500K tokens, the performance of the Italian tagger are outperformed as soon as 100 sentences are available, no matter the size of the corpus used to train embeddings. Adding training data brings a gain in performance that leads in our case to an accuracy of 87.56.

Question 4b: *When annotated data is available for the target language, how beneficial is it to train embeddings on a larger corpus?*
We repeated the previous experiment using the whole dataset available to train the Corsican embeddings. The conclusions are stable (see table 9): a training corpus (even of reduced size) is highly beneficial, since with only 100 sentences, an accuracy of 88.79 is reached. The best performance is achieved when the whole training corpus is used (92.41). This last configuration also outperforms the use of the Italian embeddings fine-tuned with the Corsican dataset.

Question 5: *When annotated data is available for the target language, is it still beneficial to fine-tune embeddings from a closely related language?*
To confirm the benefits provided by each resource, we use in this experiment CORSICAN-POS as a training corpus and the fine-tuned embeddings. With long training and conventional embeddings, we observe that the size of the training corpus for the embeddings is less relevant, but that the use of a target language training corpus with fine-tuned embeddings is the best configuration, reaching an accuracy of 93.38 with 400 sentences in the training corpus (see table 10).

Question 6: *Is using all the data available always beneficial?*
As a final set of experiments we tested configurations in which the training corpus was formed by the concatenation of the Italian training corpus and CORSICAN-POS. Whether when using embeddings trained on the Corsican corpus or fine-tuned from Italian, good results were obtained yet did not outperform the configurations using only CORSICAN-POS for training combined with the fine tuned embeddings.

5. Conclusion

In this paper, we have presented the development of CORSICAN-POS, the first gold standard corpus manually annotated for Corsican, covering northern and southern variants. We also provide the first POS tagging model for Corsican, evaluated on a varied corpus and reaching 93.38 accuracy with 400 training sentences. Section 3.3 discusses the annotation process to unveil the difficulties that manual annotation brings, and the limitations of the UD POS tagset. The limited number of tags available induced some arbitrary choices in annotation, yet we hypothesise that the good results obtained when using supervised training on this corpus is a consequence of the high consistency achieved by the manual annotation.

In fact, the series of experiments that we led show that a gold training corpus, even if small, leads to an important increase in performances

Config.	Iter01	Iter01-02	Iter01-03	Iter01-04	Iter01-05	*vs* Q2b	*vs* Q3a
Q2b: C_{Ita} - E_{Cos}-3/10					80.21	-	0
Q3a: C_{Cos} - E_{CCdC}-3/10					87.56	-	0
C_{Cos} - E_{Cos}-1/10	22.77	27.82	33.87	38.54	41.49	−38.21	−46.07
C_{Cos} - E_{Cos}-1/100	27.42	41.42	43.05	43.42	44.49	−35.72	−43.07
C_{Cos} - E_{Cos}-2/10	74.45	83.47	85.84	87.59	88.16	+ 7.95	0.60
C_{Cos} - E_{Cos}-2/100	**86.54**	89.46	90.28	91.01	**91.67**	+11.46	+ 4.11
C_{Cos} - E_{Cos}-3/10	78.39	84.58	86.63	89.03	89.57	+ 9.36	+ 2.01
C_{Cos} - E_{Cos}-3/100	**88.79**	89.95	91.17	92.04	**92.41**	+12.20	+ 4.85

Table 9: Results for question 4b. Models trained on the CORSICAN-POS corpus with embeddings trained on C_{Cos}.

Config.	Iter01	Iter01-02	Iter01-03	Iter01-04	Iter01-05	*vs* Q3b	*vs* Q4b
Q3b: C_{Ita}-$E_{Ita-Cos}$-2/100					85.12	0	-
Q4b: C_{Cos}-E_{Cos}3/100					92.41	-	0
C_{Cos}-$E_{Ita-CCdC}$-1/10	49.43	62.43	67.32	71.62	75.97	−16.44	−9.15
C_{Cos}-$E_{Ita-CCdC}$-1/100	74.01	82.06	84.32	85.41	86.04	− 6.37	+0.92
C_{Cos}-$E_{Ita-CCdC}$-2/10	72.43	79.10	85.16	87.85	88.41	− 4.00	+3.29
C_{Cos}-$E_{Ita-CCdC}$-2/100	87.80	91.65	91.96	93.18	**93.35**	+ 0.94	+8.23
C_{Cos}-$E_{Ita-Cos}$-1/10	53.19	63.42	68.28	71.74	75.28	−17.13	−9.84
C_{Cos}-$E_{Ita-Cos}$-1/100	76.04	81.98	82.57	83.33	84.10	− 8.31	−1.02
C_{Cos}-$E_{Ita-Cos}$-2/10	72.88	80.01	84.75	87.61	89.53	− 2.88	+4.41
C_{Cos}-$E_{Ita-Cos}$-2/100	88.07	91.83	92.25	**93.38**	**93.38**	+ 0.97	+8.26

Table 10: Results for question 5. Fine-tuned embeddings are used in combination with the CORSICAN-POS training corpus.

with respect to using training corpora in a closely related language and to cross lingual training, and that fine-tuned embeddings from a closely related language are highly beneficial. We have shown that having at disposal a curated corpus for training that presents a diversity in genres, authorship and writing practices, leads to good results in a very low resource context.

To pursue this work, we will explore the following perspectives: (i) increasing the embeddings training dataset in Corsican to find how much data is required to match the fine-tuned embeddings, and whether this would enable to outperform our best configuration that still takes advantage of existing resources for a closely related language, (ii) complementing our quantitative study with a qualitative analysis of the automatically annotated corpora, in order to get a better understanding of the improvement perspectives, (iii) exploring in depth the parameters of the neural networks available and (iv) reproducing this methodology to evaluate its robustness across less resourced non standardized contexts.

The reference corpus and the best model trained are freely available.[18]

[18]The URL to the repository will be added in the final version.

6. Acknowledgements

This work has been carried out within the framework of the ANR-21-CE27-0004 DIVITAL project supported by the French National Research Agency. We thank all the annotators for their work.

Bibliography

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.

Gino Bottiglioni. 1933 - 1942. *Atlante linguistico etnografico italiano della Corsica).*

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Jean-Marie Comiti. 2011. *A Pratica è a grammatica.* Albiana/Università di Corsica, Ajaccio/Corte.

Marie-Josée Dalbera-Stefanaggi. 2002. *La langue corse.*

Olivier Durand. 2003. *La lingua còrsa*. Studi grammaticali e linguistici. Paideia, Brescia.

Katharina Kann, Ophélie Lacroix, and Anders Søgaard. 2020. Weakly supervised pos taggers perform poorly on truly low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8066–8073.

Laurent Kevers. 2022. CoSwID, a code switching identification method suitable for under-resourced languages. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 112–121, Marseille, France. European Language Resources Association.

Laurent Kevers and Connor MacLean. 2023. Corpus canopé de corse (ccdc). ORTOLANG (Open Resources and TOols for LANGuage) – www.ortolang.fr.

Laurent Kevers and Alice Millour. 2022. Réalisations, obstacles et perspectives pour l'outillage du corse. In *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)*, pages 154–161, Marseille, France. CNRS.

Laurent Kevers and Stella Retali-Medori. 2020. Towards a corsican basic language resource kit. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2726–2735, Marseille, France. European Language Resources Association.

Laurent Kevers, Stella Retali Medori, and A. Ghjacumina Tognotti. 2021. A Survey of Language Technologies Resources and Tools for Corsican. Research report, UMR 6240 CNRS LISA - Université de Corse.

Steven Krauwer. 2003. The basic language resource kit (blark) as the first milestone for the language resources roadmap. In *Proceedings of SPECOM*, volume 2003, page 15.

H. J. Laméris and Sara Stymne. 2021. Whit's the richt pairt o speech: Pos tagging for scots. In *Workshop on NLP for Similar Languages, Varieties and Dialects*.

Adam Ledgeway. 2016. Italian, tuscan, and corsican. In *The Oxford Guide to Romance Languages*, pages 206–227, Oxford. Oxford University Press.

Pierre Magistry, Anne-Laure Ligozat, and Sophie Rosset. 2019. Exploiting languages proximity for part-of-speech tagging of three French regional languages. *Language Resources and Evaluation*, pages 1–26.

Jean-Baptiste Marcellesi. 1984. La définition des langues en domaine roman : les enseignements à tirer de la situation corse. In *Actes du Congrès de Linguistique et de Philologie Romanes 5*, pages 307–314, Aix-en-Provence.

Pascal Marchetti. 2001. *L'usu córsu*. Sammarcelli, Biguglia.

Alice Millour and Karën Fort. 2019. Unsupervised data augmentation for less-resourced languages with no standardized spelling. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 776–784.

Niko Partanen, Mika Hämäläinen, and Khalid Alnajjar. 2019. Dialect text normalization to normative standard finnish. In *The Fifth Workshop on Noisy User-generated Text (W-NUT 2019)*. The Association for Computational Linguistics.

Gilbert Romani. 2005. *Grammaire corse pour le collège et l'école*. DCL éditions, Bastia.

Luca Serianni. 2000. *Italiano: grammatica, sintassi, dubbi*. Garzanti, Milano.

Luca Serianni. 2005. *Grammatica italiana: italiano comune e lingua letteraria*. UTET, Torino.