Part-of-Speech Tagging on an Endangered Language: a Parallel Griko-Italian Resource

Antonis Anastasopoulos[♠] **Marika Lekakou**[♢] **Josep Quer**[♣] **Eleni Zimianiti**[♢]
Justin DeBenedetto[♠] **David Chiang**[♠]
[♠]Department of Computer Science and Engineering, University of Notre Dame
[♢]Department of Philology, University of Ioannina
[♣]ICREA & Department of Translation and Language Sciences, Universitat Pompeu Fabra
aanastas@nd.edu, mlekakou@cc.uoi.gr, josep.quer@upf.edu

Abstract

Most work on part-of-speech (POS) tagging is focused on high resource languages, or examines low-resource and active learning settings through simulated studies. We evaluate POS tagging techniques on an actual endangered language, Griko. We present a resource that contains 114 narratives in Griko, along with sentence-level translations in Italian, and provides gold annotations for the test set. Based on a previously collected small corpus, we investigate several traditional methods, as well as methods that take advantage of monolingual data or project cross-lingual POS tags. We show that the combination of a semi-supervised method with cross-lingual transfer is more appropriate for this extremely challenging setting, with the best tagger achieving an accuracy of 72.9%. With an applied active learning scheme, which we use to collect sentence-level annotations over the test set, we achieve improvements of more than 21 percentage points.

1 Introduction

Most natural language processing (NLP) applications have been developed for and tested on only a handful of languages. The majority of the world's languages are under-represented in the field, mostly due to the lack of proper resources. Endangered languages pose additional problems, as the lack of resources is further exacerbated by the lack of standard orthography. Therefore, there is an obvious need for both resources and technologies adapted to languages of under-represented communities. Especially in the case of endangered languages, computational approaches can be used to scale and accelerate documentation and revitalization efforts.

For the purposes of this study, we focus on Griko, a Greek dialect spoken in southern Italy, in the Grecìa Salentina area southeast of Lecce.[1] There is another endangered Italo-Greek variety in southern Italy spoken in the region of Calabria, known as Grecanico or Greco. Both languages, jointly referred to as *Italiot Greek*, were included as seriously endangered in the UNESCO *Red Book of Endangered Languages* in 1999. Griko is only partially intelligible with modern Greek, and unlike other Greek dialects, it uses the Latin alphabet. Less than 20,000 people (mostly people over 60 years old) are believed to be native speakers (Horrocks, 2009; Douri and De Santis, 2015), a number which is quite likely an overestimation (Chatzikyriakidis, 2010).

In general, the lack of annotated resources can be addressed through several directions. The obvious first one is the collection of human annotations, an expensive and time-consuming process. Another option is to collect additional monolingual data and use them in weakly-supervised approaches. Finally, one could use methods that transfer knowledge across languages, in which case they could focus on collecting translations or bilingual data. With this paper we attempt to quantify how effective each of these directions can be at the beginning of building applications for an endangered language.

We present a corpus of Griko narratives, first collected in the beginning of the 20th century, along with their Italian translations, suitable for computational research on the language.[2] We focus on the task

[1] A discussion on the possible origins of Griko can be found in the paper by Manolessou (2005).

[2] Our corpus is available for download here: https://bitbucket.org/antonis/grikoresource

Proceedings of the 27th International Conference on Computational Linguistics, pages 2529–2539
Santa Fe, New Mexico, USA, August 20-26, 2018.

	Stories	Sentences	Griko		Italian	
			Types	Tokens	Types	Tokens
train	104	9.2k	13.5k	197.6k	10.6k	169.7k
test	10	885	2.4k	14.0k	2.3k	13.1k
all	114	10.1k	14.1k	211.6k	11.0k	182.7k

Table 1: Statistics on our collected Griko-Italian resource.

of part-of-speech (POS) tagging, an important subcomponent of many downstream NLP applications, although our long-term goal is to annotate the whole corpus with morphosyntactic tags. Three linguists were tasked with creating gold-standard annotations for a portion of the resource, which we use as a test set in our experiments. Subsequently, our corpus includes human annotations as well as monolingual data and translations, allowing us to explore all approaches.

Based on another small Griko corpus with POS annotations, we explore computational methods for annotating our collected resource with POS tags. We evaluate several commonly used models for POS-tagging. We start with a traditional feature-based cross-entropy tagger, a Conditional Random Field (CRF) tagger, and a neural bi-LSTM tagger. We further experiment with methods that take advantage of additional monolingual data, and with methods that project cross-lingual tags through alignments. In addition, we are the first to combine the two latter approaches for POS-tagging in a low-resource setting, and we show that their combination achieves higher accuracy. Finally, we study how additional human annotations can be incorporated through an applied active learning scenario. In line with previous work, we show that this greatly improves the tagging accuracy.

Our contributions are three-fold: first, we aim to provide coarse insights into what type of annotations and resources are more effective at the early stages of developing a tagging tool for an endangered language. Second, we hope that the release of our resource will spark further interest in the computational community, so that previous, and future, methods are tested in actual endangered language settings. Finally, we benchmark the accuracy of several models on our dataset, and test them in a traditional setting and in a *transduction* setting where we also use the translations of the test set. We also evaluate all methods in an active learning scenario, showing that different approaches are suitable for different amounts of annotated data.

When only 360 annotated sentences are available, the best method is the one that combines them with both cross-lingual projections and monolingual data, achieving an accuracy of 72.9%. As the active learning scheme unfolds, however, we show that there is no need for either semi-supervised nor transfer learning approaches: a simpler feature-based CRF model achieves the highest scores, with more than 94% accuracy.

2 Resource

Resources in Griko are very scarce. The German scholar Gerhard Rohlfs pioneered research on Griko and composed the first grammar of the language (Rohlfs, 1977), also heavily influencing the subsequent grammar created by Karanastasis (1997). Although the language has been further studied, almost no corpora are available for linguistics research.

The only Griko corpus available online[3] (Lekakou et al., 2013) consists of about 20 minutes of speech in Griko, along with text translations into Italian. The corpus (henceforth `UoI` corpus, as it is hosted at the University of Ioannina, Greece) consists of 330 mostly elicited utterances by nine native speakers, annotated with transcriptions, morphosyntactic tags, and glossing in Italian.

The most noted Griko scholar is Vito Domenico Palumbo (1854–1928) who made the first serious attempts to create a literary Griko for the dialect of Calimera (the most populous of the nine remaining communities where Griko is still spoken), based on modified Italian orthography. Salvatore Tommasi and

[3] `http://griko.project.uoi.gr`

tag	frequency	tag	frequency	tag	frequency
V *(verb)*	24.4	Prt *(particle)*	2.2	Adv+Adv	0.4
PUNCT *(punctuation)*	18.3	P+D	1.8	X *(other)*	0.3
Pr *(pronoun)*	12.5	P *(adposition)*	1.6	V+Pr	0.3
N *(noun)*	11.6	Adj *(adjective)*	1.2	P+P	0.1
C *(complementizer)*	11.4	Num *(numeral)*	0.7	Pr+Pr	0.1
D *(determiner)*	7.2	N+Pr	0.6	C+Pr	0.1
Adv *(adverb)*	5.0	V+C	0.4		
Adv+P, Adv+Pr, Adv+Prt, Adj+Pr, Prt+N, Prt+Pt, D+N, Adv+Adv+Prt, Adv+Adv+Pr					< 0.1

Table 2: List of tags and their frequency in the annotated test part of the corpus.

Salvatore Sicuro then edited and published Palumbo's manuscripts (Palumbo, 1998; Palumbo, 1999), a part of which we now make available for computational and linguistic research.

After scraping from their website[4] 114 narratives that Palumbo had collected, along with their Italian translations, we removed all HTML markup and normalized the orthography: we substituted all curly quotes and apostrophes with simple ones, and substituted the vowels with circumflex (â, ô, û) that were used in a few contractions with the more common accented vowel–apostrophe combination (à', ò' ù'). Using the Moses tools (Koehn et al., 2007) with the Italian settings, we lowercased and tokenized our parallel dataset. For completeness purposes, we also make available the untokenized and proper-case versions of the corpus. The statistics of the resource are shown in Table 1.

We chose the first 10 narratives to be our test set, as they correspond to about 10% of all sentences. The rest of the narratives are treated as a monolingual or parallel resource to be leveraged. The test set was, in addition, hand-annotated by linguists: they corrected any tokenization errors that were introduced by the automatic process (for example, regarding the use of the apostrophe) and produced POS tags for every test sentence.

For every narrative, one of the linguists was presented with the produced output of the tagger, and proceeded to correct it. In order to ensure the quality of the annotations, a second linguist was then presented with the result of the work of the first linguist and tasked with correcting it, until all disagreements were resolved. Although it significantly slowed down the annotation process, we hope that this scheme ensured the quality of our annotations.

2.1 Differences from previous Griko resources

Orthography Griko has never had a consistent orthography. The transcriptions in the `UoI` corpus are based on orthographic conventions found in the few textual resources such as the local magazine *Spìtta*, that closely follow conventions adopted in Italian, aiming to be familiar to the speakers of the language. This non-standardization of the orthography leads to variations in the transcription of the same words.

In addition, we find that the word segmentation in our collected narratives follows more the concept of a phonological word. As a result, words that are segmented in the `UoI` corpus, in our narratives are often fused in a single token. The most common case that also appears in both Italian and Greek, is the contraction of prepositions and subsequent articles, such as the Italian *alla* or the Greek *στη (sti)* 'to the.Fem'. Other examples of word fusion that is not permitted in either Italian or Greek but appear in our narratives are nouns and possessive pronouns, or adverbs with other adverbs or prepositions. A direct result of this phenomenon is that annotating such tokens with single POS tags does not capture all of the necessary information.

Therefore, we chose to annotate such words with multiple POS-tags, effectively making our tag dictionary the superset of the universal tagset. The final tags that appear in practice in our corpus, and their respective frequencies, are listed in Table 2. Examples of fused words and their glosses and associated tags are shown in Table 3.

[4] `https://www.ciuricepedi.it`

word:	*stì*	*mànassu*	*cikau*	*ènna*	*vàleti*	*pànuti*
morphemes:	*s[e]-tì*	*màna-su*	*ci-kau*	*è-na*	*vàle-ti*	*pànu-ti*
POS tag:	P+D	N+Pr	Adv+Adv	V+C	V+Pr	Adv+Pr
gloss:	to-the.Fem.SG	mother-your.SG	there-down	have-COMP	put-her	on-her
translation:	'to the'	'your mother'	'down there'	'will'	'put her'	'on her'

Table 3: Examples of fused types that receive multiple tags in our annotation. The first example is a common preposition-determiner contraction, while the second and last example denote the common fusion of pronouns that follow nouns or adverbs. Notice the doubled consonants in the second and fourth instance, due to *raddoppiamento fonosintattico*.

Phonosyntactic Gemination One important difference is that the UoI corpus explicitly annotates the phenomenon of *raddoppiamento fonosintattico* (phonosyntactic gemination, or doubling of the initial consonant of the word in certain contexts) with a hyphen that separates the two words. The transcriptions that we collected do not mark for this phenomenon. The two words are often fused into a single token, and the doubling is not always present. For example, both following types appear in our corpus: *aderfòmmu* and *aderfòmu* 'my brother'.

Furthermore, the UoI corpus also uses apostrophes to mark word boundaries within which the *raddoppiamento fonosintattico* takes place. The use of apostrophes in our collected narratives is more loose. They are used both to mark elision/apocope, stress, as well as what it seems to be instances of raddoppiamento fonosintattico. This poses further issues that are discussed in the next paragraph.

Code Switching There are three languages present in the region of Salento: the regional variety of Italian, the Italo-Romance dialect of Salentino, and Griko.[5] In modern day all members of the Griko community are bilingual or trilingual. The generations before the Second World War are considered to have been predominantly monolingual, and our narratives were collected at that time, around the beginning of the 20th century. However, elements of Salentino do appear in the narratives, either as passing words, or as full sentences, mostly in dialogue turns. Note that resources on Salentino are also extremely scarce if not non-existent.

In order to deal with such examples, we decided to distinguish two scenarios. Tag switching or intra-sentential switching instances were fully annotated. So, any Salentino words or phrases that appear *within* a Griko sentence, are used for training and evaluation. However, in the few cases where we encounter full sentences in Salentino, we opt to not use them for training or evaluation. Such sentences are marked with distinctive tags in the released corpus. Note that the UoI corpus does not include any non-Griko words or phrases. An extensive study of the code-switching phenomena that occur in our corpus is left for future work.

The following is an example of usage of a Salentino phrase (italicized) within a Griko sentence, taken from story 4. Note that there exists a Griko word for 'olive oil', namely *alài* or *alàdi*, as well words for 'good', namely *kalòn* or *brao*. However, the Salentino phrase *oju finu* 'fine oil' is chosen:

leo	ti	vastò	*oju*	*finu*
say-1SG	COMP	hold-1SG	oil	fine
V	C	V	N	Adj

'I say that I have good olive oil'

Tokenization The UoI corpus has been carefully crafted to make sure that word boundaries are clearly denoted by spaces or hyphens. This unfortunately is not the case in our collected narratives. The "loose" use of apostrophes complicates the work of the tokenizer. We chose to tokenize all apostrophes as a single token, except for the cases of known elisions that were present in previous corpora, such as the case of the conjunction *c' (ce)* 'and'. In addition, in the manually annotated test set, the linguists corrected any clear tokenization issues regarding the apostrophe.

[5] See (Golovko and Panov, 2013) for a broader overview of the linguistic diversity in the Salento area.

Stress Marking In the UoI corpus, all words with two or more syllables have a diacritic mark to indicate the location of stress. However, the resources that we collected are not consistent in the use of such a diacritic. Its use is, besides, not standardized and not well studied. Although in most cases such a diacritic is used, there are several instances of polysyllabic words that have no stress marks.

2.2 Metadata

We further provide as much information as possible for each narrative, in the form of metadata. This includes the original url of the narrative, the title of the narrative in Griko and its translation in Italian. Whenever they were reported (more than 95% of the narratives) we include the location where the narrative was collected, and we anticipate that further analysis could possibly reveal any regional variations. The vast majority of the stories were naturally collected in Calimera, the largest village and the center of the Griko community, but the resource also includes 10 stories collected in Martano, as well as stories collected in Corigliano and Martignano, two smaller villages. We also include information about the date that a story was collected, as well as the narrator of the story. There are a total of 37 different narrators, while the 10 stories from Martano were retrieved from anonymous manuscripts. There are also 11 stories where the narrator is not known. Two thirds of the stories were narrated by women, while 15% of the narrators were male. The oldest manuscript dates back to 1883, while the most recent story was collected in 1998. We hope that this additional information will further allow us to investigate morphosyntactic phenomena in relation to their temporal or location context, but this is left as future work.

3 Related Work

POS tagging is a very well studied problem; probabilistic models like Hidden Markov Models and Conditional Random Fields (CRF) were initially proposed (Lafferty et al., 2001; Toutanova et al., 2003), with neural network approaches taking over in recent years (Mikolov et al., 2010; Huang et al., 2015).

The use of parallel data for projecting POS tag information across languages was introduced by Yarowsky and Ngai (2001), and further improved at a large scale by Das and Petrov (2011) who used graph-based label propagation to expand the coverage of labeled tokens. Täckström et al. (2013) used high-quality alignments to construct type and token level dictionaries. In the neural realm, Zhang et al. (2016) used only a few word translations in order to train cross-lingual word embeddings, using them in an unsupervised setting. Fang and Cohn (2017), on the other hand, used parallel dictionaries of 20k entries along with 20 annotated sentences.

Most of the previous approaches are rarely tested on under-represented languages, with research on POS tagging for endangered languages being sporadic. Ptaszynski and Momouchi (2012), for example, presented an HMM-based POS tagger for the extremely endangered Ainu language, based on dictionaries, 12 narratives (yukar), using one annotated story (200 words) for evaluation. To our knowledge, no other previous work has extensively tested several approaches on an actual endangered language.

The lack of high quality annotated data led to approaches that attempt to use monolingual resources in a semi-supervised setting. Notably, Garrette and Baldridge (2013) used about 200 annotated sentences along with monolingual corpora improving the accuracy of an HMM-based model. They tested their model on two low-resource African languages, Kinyarwanda and Malagasy and they found that in this time-constrained scenario type-level annotation leads to slightly higher improvements than token-level annotation, increasing the accuracy of their taggers to slightly less than 80%. Similar conclusions were reached in Garrette et al. (2013): 4 hours of annotation are more wisely spent if annotating at the type-level, provided there exist additional raw monolingual data. This line of work adequately addressed the question of what labeled data are preferable when there is (exceptionally) restricted access to annotators.

However, language documentation neither is nor needs to be restricted to such minimal amounts of annotation work. In addition, recently proposed endangered language documentation frameworks (Bird et al., 2014b) advocate the collection of translations (Bird et al., 2014a) which render the resource interpretable. In the case of our resource, we argue that the translations are enough for providing type-level supervision. Possibly, this is only feasible because the two languages belong in the same family *and* have been in contact for centuries, so care needs to be taken with the application of this claim.

Model	Data		
	no transduction		*transduction*
	UoI	+clp	+clp-all
stanford	62.90	67.10	67.11
crf	57.79	59.12	59.26
crf-mod	67.52	62.89	66.50
neural	45.27	53.24	58.50
	UoI+mono	+clp	+clp-all
G&B	71.67	**72.92**	72.07

Table 4: The best performing model is the one that combines semi-supervised learning with cross-lingual projected tags (G&B+clp). All models except for crf-mod benefit from transfer learning through alignments (+clp). Transduction does not significantly affect performance, except for the neural model.

4 Part-of-Speech Tagging

First, we construct a mapping of the tags of the UoI corpus to the Universal Part-of-Speech tagset (Petrov et al., 2012). This mapping is available as part of the complementary material of our resource.

Starting with the tagged UoI corpus, we can use several methods to train a tagger, which we use as baselines. We use the Stanford Log-linear POS-tagger (Toutanova et al., 2003) (henceforth stanford), trained and tested with the default settings. We also test a simple feature-based CRF tagger (henceforth crf), using the implementation of the nltk toolkit (Bird and Loper, 2004). We extended the implementation to also use prefix and suffix features of up to 4 characters, along with bigram and trigram features.[6] We will refer to this method as crf-mod.

Finally, we also investigate the use of a simple neural model. It uses a single bi-LSTM layer to encode the input sentence, and it outputs tags after a fully connected layer applied on the output of the recurrent encoder, as was described in Lample et al. (2016). The model is implemented in DyNet[7] (Neubig et al., 2017), with input embedding and hidden sizes of 128, and output (tag) embedding size of 32. It is trained with the Adam optimizer with an initial learning rate of 0.0002 and for a maximum of 50 epochs. We select the best model based on the performance on a small dev set of 40 sentences that we sampled randomly from the training set.

The tagging performance of all methods is shown in the first column of Table 4. We find that the crf-mod model is the best baseline model. With such few data to train on, both the crf and the neural model do not perform well. The bi- and tri-gram features that the crf-mod model uses are very sparse, while the neural model has to deal with a very large number of unknown words, as discussed below in the Analysis subsection.

In line with previous work, we find that semi-supervised training achieves better results in such low-resource settings. We exploit all the narratives that we collected by treating them as an additional monolingual corpus, used in the framework proposed by Garrette and Baldridge (2013). This approach (henceforth G&B) significantly improves upon all baselines, achieving an accuracy of 71.67% in the test set, an improvement of more than 4 percentage points.

Cross-lingual projected tags So far, our results have not used the Italian translations of our resource. We can follow a procedure similar to the one of Täckström et al. (2013), and extract word alignments from the Griko-Italian parallel data of the training set. We use a pre-trained Italian tagger[8] in order to tag the Italian side, and we map those tags to the universal tagset. We can then project the tags of the Italian tokens to the aligned Griko ones.[9] For the cross-lingual projected tags, we found that in practice

[6] Our extensions will be submitted to the nltk codebase.

[7] We will make the code available online.

[8] http://elearning.unistrapg.it/TreeTaggerWeb/TreeTagger.html

[9] The type-level projections are also provided with the Supplementary Material.

type-level predictions work better, and thus we only report results with such models. The tags of the Italian side of our resource, the Griko-Italian alignments, and the cross-lingual POS projections on Griko types are available through the complementary material of our resource.

Augmenting the training set with the type-level projected tags (`clp` in Table 4), we achieve improvements for all models, except `crf-mod`. The `crf-mod` method uses sparser features and is more prone to errors due to the noise of the projections. The best performance is achieved when we combine the projected tags, as type-level supervision, with the `G&B` method that leverages monolingual data. Their combination achieves the best overall performance, with an accuracy of 72.9%, a significant improvement over all other methods. As far as we know, this is the first time that cross-lingual projected tags are combined with the method of Garrette and Baldridge (2013).

Transduction An additional approach that needs to be studied is the transductive approach. Since we have translations both for the training and the test sets, we can extract word alignments and project POS tags also for the test set. The results of the transductive approach using cross-lingual projected tags from all the data that we have are shown in the third column of Table 4 (under `clp-all`).

We find that most methods benefit from the transductive approach, with the `stanford` and `crf` methods exhibiting minimal improvements, while the `neural` method improves significantly by about 5 percentage points as now there are even less out-of-vocabulary words in the input. The `crf-mod` method improves over the `UoI+clp` version, but still does not surpass the `UoI` only version. The only method that does not benefit from the transduction setting is the `G&B` method, where the performance drops.

An additional transductive step that can be taken with the `G&B` method is to also add the test set as part of the monolingual data that it uses. However, including the test set in the monolingual data also resulted in a drop in performance. Using all monolingual data along with the train-only cross-lingual types (`clp`) leads to accuracy around 69.9% (a drop of 3 points from the best model), while using all monolingual data with `clp-all` leads to a drop of another 1.4 points, to an accuracy of only 68.5%, which however is still better than all other taggers. These accuracy drops are probably justifiable, since `G&B` was not developed under a transductive assumption.

Analysis It is worth noting that our choice of using combined tags for fused/contracted words means that our training sets, under all settings, do not contain all tags that we encounter in the test set. The tagset of the `UoI` corpus only had 14 tags (the 12 universal ones plus `P+D` and `C+Pr`), indicative of its small size. As more narratives were annotated, the size of the necessary tagset increased to the final 29. However, the additional tags that we had to use are rather rare and do not severely affect the performance of our models. The tags that are present in the `UoI` corpus in fact account for 96.7% of all target tags in the test set, a value that could be considered as a skyline for all methods.

The explanation of our models' performance lies in vocabulary coverage. The `UoI` corpus only includes 46.6% of the test set tokens (8.9% of the test set types). The augmented training set with type-level projections increases those numbers to 48.7% of test tokens and 14.8% of test types. Even though we restrict ourselves to high quality alignments,[10] we are able to project tags to 3870 types (3911 in the transductive scenario), an amount higher than the amount of tags that a trained linguist can produce within four hours of annotation (Garrette et al., 2013).

The `G&B` method deals with the vocabulary coverage issue by introducing a tag dictionary expansion as a first step. They use a label propagation algorithm —Modified Adsorption (Talukdar and Crammer, 2009)— in order to spread labels between related items. In our framework, the cross-lingual projected tags provide labels for a subset of the types, in a way similar that an annotator would, partially alleviating the difficulty of the method's first step. This leads to less noise in the created tag dictionary, leading to increased accuracy. Note that, out of the cross-lingual projected tags that correspond to types that appear in the test set (about 10% of the test set types, in the *no transduction* setting), more than 65% were correctly projected.

[10] An alignment is used if either its probability is 1, or its probability is higher than 0.9 *and* the frequency of both tokens is higher than 5. Relaxing those conditions leads to worse performance due to noise.

Iteration	Narrative	Best accuracy without AL	Best accuracy with AL	Δ	Accuracy on story 9	Best method
1	story-1	77.89	—	0.0	78.13	G&B+clp
2	story-8	72.76	78.48	5.72	82.12	
3	story-7	75.07	85.17	10.10	83.57	
4	story-10	70.88	79.98	9.10	85.08	
5	story-5	72.26	82.34	10.08	88.21	crf-mod+clp
6	story-4	74.03	86.30	12.27	90.32	
7	story-3	72.48	89.67	17.19	92.13	crf-mod
8	story-6	74.67	91.80	17.13	93.64	
9	story-2	70.78	92.67	21.89	94.17	
10	story-9	72.97	94.17	21.20	—	

Table 5: Tagging accuracy for each test narrative with and without active learning. We obtain significant improvements (shown in the Δ column) by adding each annotated narrative to the training set before retraining and tagging the next narrative. The last two columns further outline the improvements from active learning, showing performance on the last and longest narrative of the test set (story-9) in each iteration, also showing the best method in each iteration. The impact of using monolingual data and type-level cross-lingual projections disappears when more training data are available.

5 Active learning

We further explored the use of active learning while tagging our test set. Our active learning scheme is as follows: We first sorted the test set narratives according to length, and starting only with the UoI corpus, we trained all taggers, producing annotations for the first story of the test set. After the corrections on the annotation of each narrative were completed, it was added as gold training data and the taggers were re-trained. For each subsequent story, the linguists were provided with the output of the tagger that achieved the highest accuracy in the previous iteration.

The main reason why we decided to follow this narrative-level active learning scheme instead of collecting type-level annotations is that a noisy corpus is not very helpful for linguistics research; at least some part of the resource should have to be checked for quality and accuracy by hand. In addition, the translations of the narratives can provide such information, as we already showed in the previous section. As we expand the coverage of our POS annotations over the whole corpus, we will explore other methods for selecting the types or sentences to be annotated through an active learning scheme.

The results, per narrative, with and without active learning, in the order that they were annotated by our linguists (from the shortest narrative to the longest) are outlined in Table 5. The results for each narrative in the active learning scenario ("with AL" column) report the best performing model that is trained on the concatenation of the UoI corpus and all the stories that were annotated in previous iterations. It is clear that the performance of the taggers improved continuously, as we added more training data. This is further outlined by each iteration's tagging accuracy on story 9, the last and longest narrative of the test set. Of course, when a narrative is added in the training set, it is then excluded from the test set, and the performance is reported on the rest of the narratives.

All methods display notable improvement as we added the annotated narratives to the training set. The performance trends are outlined in Figure 1. Firstly, it is notable that as the training set increases, the advantage of the model of Garrette and Baldridge (2013) that leverages monolingual data diminishes, compared to our simpler crf-mod tagger, both with or without cross-lingual projected tags. Before the first iteration, the accuracy gap is 6.4 percentage points in favor of G&B. However, after adding around 4-5 narratives so that there are around 500 training sentences, our crf-mod+clp method surpasses the G&B method and keeps improving. This is also outlined by the dashed line in Table 5. As we add more training instances, the accuracy of the G&B method plateaus around 85% and does not improve further.

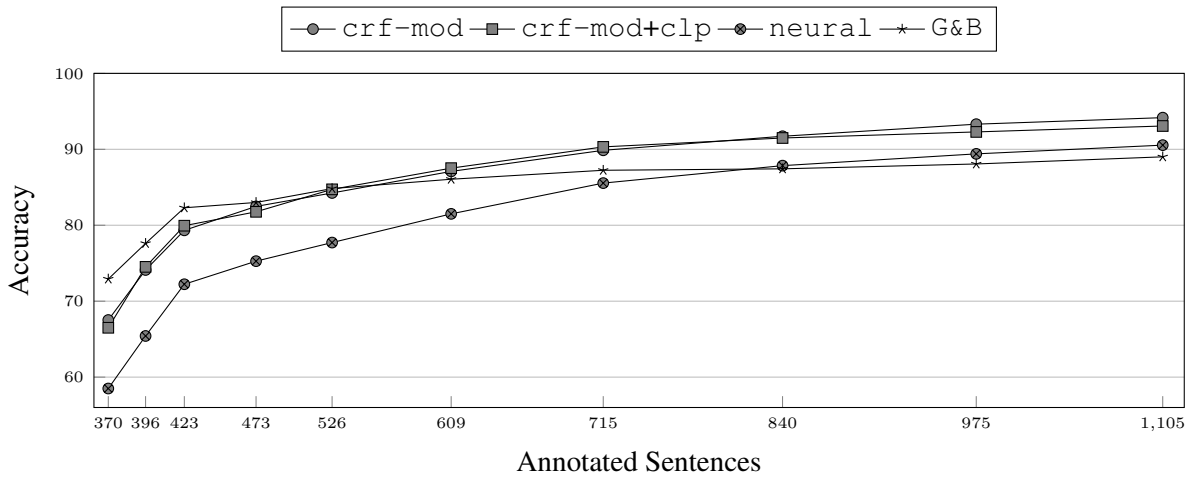

Figure 1: Accuracy on the (remaining) test set as we add annotated narratives to the training set. All methods benefit from the active learning approach, with G&B displaying better performance due to its use of monolingual data in the first iterations, but the crf-mod approach achieving the best results in the last iterations, eventually not even needing the cross-lingual type-level projections (+clp).

Furthermore, after a couple more iterations, when more than 800 annotated sentences are available for training, the crf-mod method without cross-lingual projected tags achieves higher accuracy than all others. We identify this point as the one where simple token-level supervision is efficient enough to outperform semi-supervised or transfer-learning approaches.

Finally, we observe that the accuracy of the neural bi-LSTM approach that only uses the tagged corpus without further use of monolingual data, improves significantly as the training set increases. With only 370 training sentences, the gap between the neural and the best method is more than 14 percentage points. With 1,100 training sentences, the accuracy gap diminishes to only 2 percentage points.

6 Cross-Validation

Our end goal is to annotate the whole corpus with POS tags, as well as richer annotations. Towards that direction, our gold annotated test data could be used to train a higher quality POS tagger, which we will use to annotate the rest of the corpus. In Section §5, we found that including all but one annotated narratives for training, and testing on the last one (story-9) we were able to obtain an accuracy of more than 94.17%. In order to get a better estimation of how well a tagger trained on our gold data would work, we perform a cross-validation experiment, using crf-mod, our best performing model.

For each cross-validation instance, one of the annotated narratives becomes the test set, and the rest will be included in the training set. This allows us to obtain an average performance over 10 instances. The average accuracy of the crf-mod model is about 91.9%, with a standard deviation of about 2 percentage points (minimum is 88.5% on story 5, and maximum is 94.9% on story 2).

The main obstacle to annotating the rest of the corpus with higher quality is out-of-vocabulary words. The combined vocabulary of the UoI corpus and our 10 annotated narratives covers 16% of the vocabulary of the 104 unannotated sentences (but 85% of the total tokens). As part of our future work, we plan to incorporate word-level active learning in our annotation/correction scheme, similar to the approaches proposed by Fang and Cohn (2017).

7 Conclusion

We presented a parallel corpus of 114 narratives on an endangered language, Griko, with translations in Italian. For now, a test set of 10 narratives is hand-annotated with Part-of-Speech tags, but in the future we will enrich the resource with annotations on the rest of the corpus, as well as with richer syntactic and morphological annotations. We also plan on contributing our corpus to the Universal Dependencies treebanks (Nivre et al., 2016) as Griko is absent from the supported languages.

We extensively evaluated several POS tagging approaches, and found that the method of Garrette and Baldridge (2013) can be combined with cross-lingual type-level projected tags, outperforming all other methods, with an accuracy of 72.9%, when less than 500 sentences are available. As data was added in the training set in an active learning scenario, a simple feature-based CRF approach outperforms all other models, with accuracy improvements of over 21 percentage points and over 94% accuracy on the last narrative. In fact, when more than 800 sentences are available for training, cross-lingual tag projections hurt performance.

The collected annotations from our test set could form the basis for training a high-accuracy POS tagger for Griko, so that we can expand the POS annotations to the rest of our corpus with only a small amount of noise. We aim to explore this direction in our future work, along with other active learning methods that require less human intervention. In addition, we plan to further enrich the annotations of our corpus with morphological tags similar to the `UoI` corpus, that will provide even more insight in Griko and its usage. When the full annotations of the corpus are completed, we plan to use statistical methods to study specific phenomena regarding the grammar and syntax of Griko.

Finally, and most importantly, we hope that the release of this corpus will spark further interest for computational approaches applied on endangered languages documentation and on under-represented languages in general.

Acknowledgements This work was generously partially supported by NSF Award 1464553. We are also grateful to the anonymous reviewers for their thougthful reviews and useful comments.

References

Steven Bird and Edward Loper. 2004. Nltk: the natural language toolkit. In *Proc. ACL*, page 31. Association for Computational Linguistics.

Steven Bird, Lauren Gawne, Katie Gelbart, and Isaac McAlister. 2014a. Collecting bilingual audio in remote indigenous communities. In *Proc. COLING*.

Steven Bird, Florian R. Hanke, Oliver Adams, and Haejoong Lee. 2014b. Aikuma: A mobile app for collaborative language documentation. In *Proc. of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*.

Stergios Chatzikyriakidis. 2010. *Clitics in four dialects of Modern Greek: A dynamic account*. Ph.D. thesis, University of London.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proc. ACL*, pages 600–609. Association for Computational Linguistics.

Angeliki Douri and Dario De Santis. 2015. Griko and modern Greek in Grecìa Salentina: an overview. *L'Idomeneo*, 2015(19):187–198.

Meng Fang and Trevor Cohn. 2017. Model transfer for tagging low-resource languages using a bilingual dictionary. In *Proc. ACL*, pages 587–593. Association for Computational Linguistics.

Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proc. NAACL-HLT*, pages 138–147. Association for Computational Linguistics.

Dan Garrette, Jason Mielens, and Jason Baldridge. 2013. Real-world semi-supervised learning of pos-taggers for low-resource languages. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 583–592.

Ekaterina Golovko and Vladimir Panov. 2013. Salentino dialect, Griko and regional Italian: Linguistic diversity of Salento. *Working Papers of the Linguistics Circle of the University of Victoria*, 23(1):51.

Geoffrey Horrocks. 2009. *Greek: A History of the Language and its Speakers*. Wiley-Blackwell.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. arXiv:1508.01991.

Anastasios Karanastasis. 1997. *Grammatiki ton ellinikon idiomaton tis Kato Italias [Grammar of the Greek dialects of south Italy]*. Akadimia Athinon.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. ACL*, pages 177–180. Association for Computational Linguistics.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proc. NAACL-HLT*, pages 260–270.

Marika Lekakou, Valeria Baldiserra, and Antonis Anastasopoulos. 2013. Documentation and analysis of an endangered language: aspects of the grammar of Griko. `http://griko.project.uoi.gr`.

Ioanna Manolessou. 2005. The greek dialects of southern Italy: an overview. *KAMPOS: Cambridge Papers in Modern Greek*, 13:103–125.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proc. Interspeech*.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, et al. 2017. DyNet: The dynamic neural network toolkit. arXiv:1701.03980.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proc. LREC*.

Vito Domenico Palumbo. 1998. *Io' mia fora' - Fiabe e Racconti della Grecìa Salentina [Once upon a time - Fairy Tales and Stories from Grecìa Salentina]*. Calimera (LE): Ghetonìa. a cura di S. Tommasi.

Vito Domenico Palumbo. 1999. *'Itela na su pò - Canti popolari della Grecìa Salentina [I wanted to tell you - Folk songs of Grecìa Salentina]*. Calimera (LE): Ghetonìa. a cura di S. Sicuro.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proc. LREC*.

Michal Ptaszynski and Yoshio Momouchi. 2012. Part-of-speech tagger for Ainu language based on higher order Hidden Markov Model. *Expert Systems with Applications*, 39(14):11576–11582.

Gerhard Rohlfs. 1977. *Grammatica storica dei dialetti italogreci (Calabria, Salento) dt. Original [1949–1954] [Historical Grammar of the Italiot Greek dialects (Calabria, Salento)]*. CH Beck.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Partha Pratim Talukdar and Koby Crammer. 2009. New regularized algorithms for transductive learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 442–457. Springer.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. NAACL-HLT*, pages 173–180. Association for Computational Linguistics.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual pos taggers and np brackets via robust projection across aligned corpora. In *Proc. NAACL*.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag – multilingual pos tagging via coarse mapping between embeddings. In *Proc. NAACL-HLT*, pages 1307–1317. Association for Computational Linguistics.