Lemmatization Experiments on Two Low-Resourced Languages: Low Saxon and Occitan

Aleksandra Miletić
University of Helsinki
Department of Digital Humanities
aleksandra.miletic@helsinki.fi

Janine Siewert
University of Helsinki
Department of Digital Humanities
janine.siewert@helsinki.fi

Abstract

We present lemmatization experiments on the unstandardized low-resourced languages Low Saxon and Occitan using two machine-learning-based approaches represented by MaChAmp and Stanza. We show different ways to increase training data by leveraging historical corpora, small amounts of gold data and dictionary information, and discuss the usefulness of this additional data. In the results, we find some differences in the performance of the models depending on the language. This variation is likely to be partly due to differences in the corpora we used, such as the amount of internal variation. However, we also observe common tendencies, for instance that sequential models trained only on gold-annotated data often yield the best overall performance and generalize better to unknown tokens.

1 Introduction

Lemmatization consists in finding the base form of a given inflected form. The definition of the base-form for a grammatical category can vary across languages. It can include, e.g., finding the masculine singular for an adjective (*bèlas* 'beautiful.F.PL' > *bèu* 'beautiful.M.SG), or finding the infinitive for a verb (*atten* 'eat.3PL.IND.PRES' > *eaten* 'eat.INF'). The main benefit of lemmatization lies in reducing data sparsity by grouping together all surface forms stemming from the same lemma. It is especially useful for morphologically rich languages, for which the high number of surface forms leads to lower token – type ratios. For such languages, lemmatization is systematically used as a preprocessing step for downstream tasks such as parsing, and it is essential for building efficient corpus querying systems.

We approach this task from the perspective of two low-resourced, non standardized minority languages: Occitan and Low Saxon. In the case of non standardized varieties, acquiring even minimal amounts of manually lemmatized data can be difficult. One of the reasons is the definition of lemmatization itself: in the absence of a common standard, which approach to lemmatization should be adopted? Should lemmatization respect different levels of variation (lexical, morphological, orthographic) which are present in multi-dialect datasets? Or should one variety be chosen for lemmatization purposes and used across all dialects? The former solution allows for the preservation of dialectal differences, but limits the positive impact of lemmatization on data sparsity. The latter is more effective in this respect but it compounds lemmatization and normalization, arguably making the task more difficult. Furthermore, it can be deemed problematic by the speakers of the language in question.

In this paper, we explore both of these approaches: our Low Saxon dataset adopts an interdialectal lemmatization approach, whereas the Occitan dataset's lemmas are dialect-specific. We evaluate the effects of using small, manually annotated datasets for training lemmatization models vs relying on a larger, automatically preannotated corpus. We investigate the utility of developing one general model for all dialects vs training dialect-specific models. Since lemmatization typically relies on PoS information to aid the processing of ambiguous tokens, we look into different ways of using this annotation layer in our corpora by evaluating two learning paradigms: joint learning and classical, sequential learning for PoS-tagging and lemmatization.

2 Related Work

Lemmatization methods based on machine learning can be divided into edit tree-based approaches and string transduction methods. The edit tree-based algorithms (ges; Grzegorz Chrupala and van Genabith, 2008; Müller et al., 2015) derive the sequence of edit operations needed to transform the inflected form into the lemma. The edit tree is

Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), pages 163–173
May 5, 2023 ©2023 Association for Computational Linguistics

used as a label for each wordform – lemma pair. The model learns to predict the edit tree and not the lemma itself, thus treating lemmatization as a classification task.

With the advent of neural methods, lemmatization has been recast as a string-transduction task (e.g. Bergmanis and Goldwater, 2018; Manjavacas et al., 2019). Currently, the main contribution of these approaches to the state of the art seems to be better generalization capacities, measured as the model's ability to correctly lemmatize unseen wordforms. Bergmanis and Goldwater (2018) report an important improvement on unknown tokens over non-neural approaches, and similar observations are made by Manjavacas et al. (2019). However, both works remark that the neural networks do not seem to outperform edit tree-based approaches on ambiguous tokens. In general, the capacity to deal with ambiguous tokens is believed to depend on the availability of contextual information, which is supposed to facilitate disambiguation. Bergmanis and Goldwater (2018) use a sliding character window as context, and Manjavacas et al. (2019) condition the decoder on sentence-level embeddings. These efforts to include contextual information do not seem sufficient to beat the edit-based methods on this type of tokens.

The well-suitedness of one type of lemmatization algorithm over the other may also depend on the linguistic properties of a given language. Manjavacas et al. (2019) note that, when evaluating on modern languages, the edit tree-based method outperforms the neural model on both West European and Uralic languages, whereas for Slavic languages the neural model yields better results. These results would, however, need to be confirmed, since Ljubešić and Dobrovoljc (2019) find that the edit tree-based approach beats the neural model on South Slavic languages they investigate.

Lemmatization is often paired with PoS (Part of Speech) tagging. Since inflected forms can be ambiguous as to their lemma, relying on PoS-tags can help the disambiguation process. This information can be exploited as part of joint multi-task learning (Kondratyuk et al., 2018; Manjavacas et al., 2019; van der Goot et al., 2021) or, more traditionally, in a sequential approach, in which the models for two tasks are learned separately, but the lemmatizer relies on the morphological information during training and prediction (e.g. Qi et al., 2020). Vatri and McGillivray (2020) compare lemmatizers for Ancient Greek based on dictionary lookup that exploit PoS information to distinguish ambiguous tokens. Alternatively, some approaches do not rely on this type of information at all (e.g. Bergmanis and Goldwater, 2018), which may simplify lemmatization for low-resource languages.

More generally, lemmatization in the low-resource setting has also received attention in recent work. Bergmanis and Goldwater (2018) evaluate their models both on the full amount of available data and on 10k samples. Saunack et al. (2021) explore the lower bound for training data size on Indian languages: they compare a standard setting with low-resource settings with only 500 and 100 training instances, in which they rely on data augmentation techniques. Saurav et al. (2020) investigate cross-lingual approaches for lemmatizing low-resourced Indian languages.

In this work, we are particularly interested in the low-resource setting, since the gold standard datasets available for Low Saxon and Occitan are limited in size. We also experiment with the cross-lingual and cross-lectal approach by using historical data and related languages. We opt for neural models since we expect a high proportion of unknown tokens in our datasets due to the fact that we are dealing with non standardized languages. We examine both the joint and sequential learning in an attempt to identify the optimal approach to exploit the PoS tagging information present in our datasets.

3 Languages

3.1 Low Saxon

Low Saxon is a West Germanic language spoken by approximately 4.8 million people primarily in the north-eastern Netherlands and northern Germany (Moseley, 2010). Despite official recognition in both countries, no interdialectal standard variety has been established so far.

Dialect classification of Low Saxon is normally more finegrained than the three-fold subdivision we use here. Dutch Low Saxon is traditionally divided into Gronings, Stellingwerfs, Drents, Sallands, Twents, Veluws, Achterhoeks and Urkers (Bloemhoff et al., 2019, 20), but due to scarcity of data we treat it as one group. The traditional classification of German Low Saxon (see for instance Schröder, 2004 and Stellmacher, 1983) assumes an East-West division based on, among others, the history of settlement and the plural suffix of verbs in the present tense. However, we have not found this traditional division to correspond to overall dialect similarity in our previous dialectometric experiments. Therefore, we instead adopt a north-south division following Lameli (2016) and our own observations. The northern group consists of North Saxon and Mecklenburgish - West Pomeranian, and the southern group of Westphalian and Eastphalian. We excluded Brandenburgish, East Pomeranian and Low Prussian due to data scarcity.

Compared with Middle Low Saxon, the number of inflectional categories has decreased, and there is dialectal variation in the number of categories preserved. For instance, while nouns in Middle Low Saxon inflected for four cases, nominative, genitive, dative and accusative (Lasch, 1974), only a few of the southern varieties in Westphalia and Eastphalia still distinguish the dative and accusative (Lindow et al., 1998). Most Low Saxon varieties in Germany distinguish the nominative and the accusative, whereas Dutch Low Saxon typically does not. Usage of the genitive is very restricted in all Low Saxon varieties.

At the phonological level, we find noticeable variation in the number of distinct vowel phonemes preserved and in the ways vowel phonemes have merged. A typical example is the merger of Proto-Germanic *â and lengthened *a[1] that has occurred outside of Westphalia (Niebaum, 2008; Bloemhoff et al., 2019). As a result, we find the same phoneme in *Spraak*[2] 'language' and *Water* 'water' in the north-western dialects, while Westphalian, here Münsterlandic, shows distinct phonemes in *Spraoke* and *Water*.

In addition to the dialectal variation, there is considerable orthographic variation as most Low Saxon writers follow regional writing traditions to different degrees or might devise their own spelling systems. These regional or personal spellings often draw some inspiration from the majority language orthographies. This can be seen, e.g., in the frequent capitalization of nouns by German Low Saxon writers and in the representation of the voiced sibilant /z/ with the grapheme <z> by Dutch Low Saxon writers, while German Low Saxon writers commonly use <s> for the same phoneme.

Our corpus reflects this orthographic and dialectal variation that poses significant challenges to NLP.

3.2 Occitan

Occitan is Romance language which belongs to the Gallo-Romance group. It is closest to Catalan, with which it forms a subgroup called occitano-roman (Bec, 1970). It is spoken in southern France (except in the Basque and Catalan areas), in several valleys of the Italian Piedmont and in the Val d'Aran in Spain. When it comes to its linguistic properties, Occitan is a null subject language with tense, person and number inflection marks on finite verbs. Number and gender are marked on all components of the noun phrase in many dialects.

The most widely accepted classification proposed by Bec (1995) includes 6 major dialectal groups: Auvernhat, Gascon, Lengadocian, Lemosin, Provençau and Vivaroaupenc[3], each of them with areas of greater or lesser variation. Geographic variation affects all levels of linguistic structure. In this paper we focus on Lengadocian, Gascon, Provençau and Lemosin, due to the availability of annotated material for these dialect groups. Geographical variation affects all levels of linguistic structure. Different phonological processes have resulted in series of wordforms specific to each dialect group, e.g. the word *son* translates to *hilh* in Gascon, *filh* in Lengaodcian and Lemosin, and *fiu* in Provençau. On the lexical level, the word *potato* corresponds to *mandòrra*, whereas it is *trufa/trufet* or *patana/patanon* in Lengadocian. On the morpho-syntactic level, verb inflection varies from one dialect to another, and there is also an important degree of intra-dialectal variation. To illustrate, *we are* corresponds to *èm* in Gascon, *sem* in Lemosin, *sèm* in Lengadocian and *siam* in Provençau based on the most frequent paradigm for each dialect group.

This situation is further complicated by the existence of several orthographic norms, out of which two seem to dominate today: the so-called *Mistralian* orthography, inspired by French writing conventions, and the *classical* orthography, closer to the medieval troubadours' spelling (Sibille, 2002). The data used in our experiments is limited to the classical orthography.

[1] This lengthening happened relatively regularly in open syllables.

[2] Notice the apocope of final *-e* that has occurred in most northern dialects. Vowel length is often marked by doubling the letter in closed syllables.

		Dataset	Sent.	Tok.	Types	Sent. len.
Low Saxon	SMALL	All dialects	904	19258	6000	21.30
		Dutch LS	310	6716	2297	21.66
		North Ger. LS	265	5415	1961	20.43
		South Ger. LS	326	7127	2635	21.86
	LARGE	All dialects	126359	2431944	166625	19.25
Occitan	SMALL	All dialects	1522	26122	6196	17.16
		Gascon	255	4170	1429	16.35
		Lengadocian	1113	19315	4499	17.35
		Lemosin	77	1344	596	17.45
		Provençau	77	1293	583	16.79
	LARGE	-	100000	2037723	147070	20.38

Table 1: SMALL and LARGE dataset information for Low Saxon and Occitan

4 Datasets

For both languages, we use two basic datasets: the SMALL dataset is manually annotated and it was available for both languages at the beginning of the experiments reported here. The LARGE datasets are an order of magnitude greater than their SMALL counterparts, but contain only automatic preannotation with PoS-tags and lemmas. In the experiments presented here, the SMALL datasets were used for initial training of our models, and we make use of their dev and test splits for training and for evaluation. The LARGE datasets were used as additional training material in various setups in an attempt to improve model accuracies. In the remainder of this section, we provide some quantitative details and descriptions of each dataset. Note that the LARGE datasets were not annotated at the beginning of our work. The strategies used to palliate this are described in Section 6.

In the case of Low Saxon, both the SMALL and the LARGE dataset stem from the same corpus, described in Siewert et al. (2020), and contain several genres, for instance fiction texts such as fairytales or novels, and non-fiction texts such as letters, announcements or political speeches. The Low Saxon dataset is roughly split into two time periods: 1800–1939 and 1980–2022. The distribution within the dialect groups is as follows: Dutch Low Saxon 20% and 80%, North German Low Saxon 87% and 13%, South German Low Saxon 44% and 56%.

For Occitan, the SMALL dataset is based on the treebank presented in Miletic et al. (2020) and contains predominantly literary texts. The LARGE Occitan corpus contains Occitan Wikipedia articles from 2021, taken from the Leipzig Corpora Collection[4].

For both languages, the gold dataset has also been stratified into dialect groups in order to examine the usefulness of dialect-specific training data and evaluate model performance for different dialects. The gold sets are split into train, test and development sets (except in the case of Occitan, for which two dialect groups do not have a dev set, the total amount of annotated data being too small)[5]. A quantitative overview of gold splits is given in Table 2. The unknown and ambiguous tokens are defined in relation to the gold annotated train set including all dialects.

5 Tools

We make use of two training paradigms: multi-task learning applied to PoS-tagging and lemmatization, in which both tasks are learned as part of the same model, and traditional sequential learning, in which a separate model is trained for each task. We explore the former with MaChAmp (van der Goot et al., 2021) and use the Stanza NLP pipeline (Qi et al., 2020) for the latter.

5.1 MaChAmp

MaChAmp is a toolkit that allows for easy finetuning and joint learning of a wide range of NLP tasks, including PoS-tagging, lemmatization, parsing, masked language modelling and text generation. MaChAmp takes a pretrained contextualized model as the initial encoder and fine-tunes it according to a given set of downstream tasks. Each task has its own decoder for task-specific predictions. The tool also allows an initial round of training on a specific task, and then fine-tune it in a second round of training. We put this functionalty to test in our lemmatization experiments. As the default embeddings, MaChAmp uses mBERT (Devlin et al., 2019). For a detailed description of the tool and the model it is based on, the reader is referred to van der Goot et al. (2021)

5.2 Stanza

Stanza is a Python NLP pipeline currently supporting 66 languages (which do not include Occitan and Low Saxon). The tool supports tokenization,

[3] Names of dialects are given in Occitan (each one in its dialect) as there is no standardized orthographic form for those names in English.

[4] https://corpora.uni-leipzig.de/en?corpusId=oci_wikipedia_2021

[5] Since the original corpus did not have *dev* splits, the corpus was re-split into *train*, *dev* and *test* for the needs of the experiments we describe.

	Dataset	train			test					dev				
		Sent.	Tok.	Types	Sent.	Tok.	Types	Unk. (%)	Amb. (%)	Sent.	Tok.	Types	Unk. (%)	Amb. (%)
Low Saxon	All dialects	723	15346	5083	91	1972	1020	26.88	36.76	90	1940	930	26.29	32.84
	Dutch LS	249	5072	1878	31	925	469	24.54	37.73	30	719	410	23.5	32.55
	North German LS	213	4447	1683	26	391	241	24.3	37.6	26	577	352	25.12	33.62
	South German LS	262	5827	2220	32	656	407	31.71	34.91	32	644	406	30.43	32.45
Occitan	All dialects	1196	20551	5292	202	3179	1054	22.11	28.18	124	2392	1009	16.39	31.77
	Gascon	195	3258	1173	35	421	230	26.37	23.28	25	491	267	19.35	33.60
	Lengadocian	884	15494	3937	130	1920	577	19.64	27.50	99	1901	814	15.62	31.30
	Lemosin	56	919	434	16	413	211	27.76	31.76	-	-	-	-	-
	Provençau	61	880	424	16	413	211	23.49	32.69	-	-	-	-	-

Table 2: SMALL dataset split into train, dev and test

multi-word token expansion, lemmatization, PoS and morphological feature tagging, dependency parsing, and named entity recognition. In this work, we utilize its PoS-tagger, based on a biLSTM model, and its lemmatizer, a neural seq2seq model. For more details, please see Qi et al. (2020).

6 Strategies for Creating Large(r) Amounts of Annotated Data

One of the dimensions of lemmatization we explored in this work relates to the size and the nature of the training material. Specifically, we compared the performance of tools trained on small amounts of gold-annotated data with using larger corpora that were automatically preannotated. As mentioned in Section 1, the corpora we used as our LARGE datasets were not annotated at the outset of the experiments presented here. There were, to the best of our knowledge, no freely available models based on neural approaches for the PoS-taggging and the lemmatization of Low Saxon and Occitan. The first round of our experiments was therefore dedicated to creating initial models for both tasks which would allow us to produce reliable automatic preannotation. For Low Saxon, we leveraged an existing historical corpus of Middle Low Saxon to train models that were then transferred to Modern Low Saxon. This had the advantage of using a corpus that was larger than the available gold standard in Modern Low Saxon. For Occitan, no comparable historical corpus was available. We therefore relied on bootstrapping using the SMALL dataset.

6.1 Leveraging a Historical Corpus for the Preannotation of Modern Text

The initial preannotation of the Low Saxon lemmas was done with MaChAmp and the reference corpus Middle Low German[6] / Low Rhenish (Peters, 2017). The reference corpus uses a tagset specifically designed for the needs of Middle Low Saxon. Therefore, we instead made use of the automatic PoS annotation provided by Siewert et al. (2022).

The reference corpus consists of two parts: an annotated part that comes with supradialectal lemmatization following primarily the *Mittelniederdeutsches Handwörterbuch* by Lasch et al. and the one by Lübben (1995 - 1888) and a transcribed part without annotation. We converted the annotated part to the ConLLU-format required by the tools we used. The MaChAmp lemmatization model achieved an accuracy of 89.9% on this data. This model was subsequently finetuned on a small set of manually annotated modern data in order to annotate the rest of the corpus.

Our modern Low Saxon gold annotated dataset does not employ the Middle Low Saxon dictionary spelling, but the *Nysassiske Skryvwyse*[7] 'New Saxon spelling', an interregional spelling based on historical sound correspondences and used by, for instance, the Dutch Low Saxon Wikipedia. As this spelling does not reduce all dialectal variation, the lemma form is, as far as possible, chosen based on the Middle Low Saxon dictionary form, attested Old Saxon forms or Proto-Germanic reconstructions. For future comparisons with the historical corpus, it would be desirable to add a Middle Low Saxon lemmatization layer to the modern data.

The final pretrained MaChAmp model for modern Low Saxon achieved a lemma accuracy of 87%, and a PoS accuracy of 94% on the manually annotated development set. These relatively good results (compared with our later experiments) might be explained by some overfitting as we used the same development set in two consecutive training steps: Original lemmatization finetuning and later joint training of lemmatization and PoS tagging.

[6] Called "Middle Low German" in the official English name of the reference corpus; We otherwise refer to this language as "Middle Low Saxon".

[7] https://skryvwyse.eu

6.2 Bootstrapping Using a Small Gold Standard Corpus and a Lexicon

For Occitan, we used MaChAmp in order to train a PoS-tagger and a lemmatizer which would allow us to preprocess the LARGE dataset. Since this was a preliminary experiment with the tool on this language, we opted for training independent models for each of the tasks in order to evaluate the baseline performance for each task on gold data. In this scenario, we did a single training run on the full SMALL dataset, using the default embeddings.

The PoS-tagger achieved global accuracy of 92.26% on the test set comprised of all dialects, the highest being 92.97% on Lengadocian and the lowest 89.10% on Provençau.

The lemmatizer's global accuracy reached 89.30%, ranging from 88.6% on Gascon to 93.33 on Lengadocian and Lemosin (detailed results for the global evaluation are available in Table 6, and for the dialect-specific evaluation in Table 7).

These models were ensembled with the morphological lexicon Loflòc (Vergez-Couret, 2016; Bras et al., 2020). If a wordform was present in the lexicon and only had one entry, it was annotated with information found in the lexicon. Otherwise, the models' predictions were used. Around a third of the wordforms received lexicon-based annotations.

The preannotated corpus was used as the LARGE dataset in the experiments described in the sections below.

7 Large Preannotated Corpora and Small Gold Datasets

Following the creation of additional annotated material, we trained new models with both MaChAmp and Stanza.

With MaChAmp, we used the LARGE datasets for the initial round of training, and the SMALL dataset for the fine-tuning of the model. The fine-tuning was done both with the full SMALL dataset and using dialect-specific subsets of it. We trained for lemmatization and PoS-tagging jointly, resulting in one model capable of performing both tasks.

With Stanza, we trained lemmatizers both on the SMALL dataset on its own and on a combined dataset, concatenating SMALL and LARGE datasets. This approach was chosen because the current version of Stanza does not support retraining. We leveraged the available morphological information in the training process. We also trained the corresponding PoS-tagging models: they are used to approximate a pipeline setup and evaluate the Stanza lemmatizers on predicted PoS-tags.

Additionally, we trained a lemmatizer that does not rely on morphological annotation with both tools. These models were intended as a baseline, but they also correspond to a real-life usecase in which a lemmatized corpus for a given language is available, but contains no PoS tags.

The global lemmatization results are given in Table 6, whereas the dialect-specific results are available in Tables 7 and 8. We report mean accuracy and standard deviation over three training runs on the test set[8]. In addition to results on the full evaluation set, we also report performance on unknown and ambiguous tokens. We consider as unknown tokens those that do not appear in any of the training material. We define as ambiguous all tokens having more than one possible lemma in the training material. In the case of dialect-specific evaluations, we evaluate the dialect-specific model trained using MaChAmp along with the general models trained with both tools. Our goal is to assess if dialect-specific training is useful even if it entails using less training data than for the general model.

7.1 General results

As an overall tendency, Occitan seems to be easier to lemmatize than Low Saxon, with the former's accuracy ranging often around 10% higher than the latter's. In case of the unknown tokens, the difference is even bigger. Given the greater orthographic variation in our Low Saxon dataset, this does not come as a surprise.

The sequential approach of the Stanza pipeline most of the time yields the best results for both Low Saxon and Occitan. Surprisingly, we found the MaChAmp base model[9] to perform best for Low Saxon, with an almost 5% advantage over the finetuned model. On Occitan, finetuning the MaChAmp model does bring an improvement, albeit a small one (around 1.5%)

Large automatically annotated corpora seem to bring some benefit for the overall accuracy but they do not generally outperform the smaller Stanza models which have access to the PoS information. In the case of unknown tokens in particular, we see that the Stanza model trained only on gold data with gold PoS performs best.

[8] The results on the *dev* set are available in Appendix A.

[9] Only trained on a large corpus of automatically annotated data, no finetuning on gold data.

	Tool	Training set	Task	Train cond.	Test cond.	ALL	UNK	AMB
Occitan	MaChAmp	SMALL	LEM	no POS, gold LEM	no POS	$91.28^{\pm 0.42}$	$72.22^{\pm 1.55}$	$96.23^{\pm 0.37}$
		LARGE	POS+LEM	pred. POS+LEM	no POS	$91.77^{\pm 0.23}$	$68.54^{\pm 1.86}$	$92.19^{\pm 0.14}$
		L+S	POS+LEM	pred. POS+LEM	no POS	$92.16^{\pm 0.25}$	$67.2^{\pm 0.33}$	$93.05^{\pm 0.45}$
	Stanza	SMALL	LEM	no POS	no POS	$90.35^{\pm 0.42}$	$66.86^{\pm 1.85}$	$95.78^{\pm 0.0}$
		SMALL	LEM	gold POS+LEM	pred. POS	**93.21** $^{\pm 0.09}$	**78.43** $^{\pm 0.41}$	**96.69** $^{\pm 0.0}$
		COMB	LEM	pred. POS+LEM	pred. POS	$92.49^{\pm 0.08}$	$68.4^{\pm 0.98}$	$92.63^{\pm 0.0}$
Low Saxon	MaChAmp	SMALL	LEM	no POS, gold LEM	no POS	$70.74^{\pm 0.09}$	$17.47^{\pm 0.48}$	$88.63^{\pm 0.26}$
		LARGE	POS+LEM	pred. POS+LEM	no POS	**83.42** $^{\pm 0.21}$	$30.19^{\pm 1.33}$	$85.19^{\pm 0.47}$
		L+S	POS+LEM	pred. POS+LEM	no POS	$78.14^{\pm 0.31}$	$20.44^{\pm 1.18}$	$81.2^{\pm 0.22}$
	Stanza	SMALL	LEM	no POS	no POS	$75.33^{\pm 0.11}$	$36.41^{\pm 0.42}$	$82.03^{\pm 0.0}$
		SMALL	LEM	gold POS+LEM	pred. POS	$80.52^{\pm 0.43}$	**45.66** $^{\pm 1.59}$	**89.42** $^{\pm 0.0}$
		COMB	LEM	pred. POS+LEM	pred. POS	$81.31^{\pm 0.05}$	$20.12^{\pm 0.89}$	$82.16^{\pm 0.0}$

Table 3: Global Lemmatization Accuracy for Occitan and Low Saxon

		Gascon					Lemosin		
Tool	Train	ALL	UNK	AMB	Tool	Train	ALL	UNK	AMB
MaChAmp	L+S	$89.66^{\pm 0.52}$	$57.01^{\pm 1.24}$	$90.28^{\pm 0.57}$	MaChAmp	L+S	**90.91** $^{\pm 0.2}$	**74.42** $^{\pm 1.9}$	$94.35^{\pm 0.46}$
MaChAmp	L+GAS	$88.86^{\pm 0.41}$	$54.38^{\pm 1.24}$	$89.58^{\pm 0.98}$	MaChAmp	L+LEM	$87.64^{\pm 0.57}$	$64.34^{\pm 1.1}$	$92.66^{\pm 0.8}$
Stanza	SMALL	**90.71** $^{\pm 0.75}$	**77.78** $^{\pm 2.79}$	**91.49** $^{\pm 0.0}$	Stanza	SMALL	$90.59^{\pm 0.41}$	$72.6^{\pm 0.8}$	**99.22** $^{\pm 0.0}$
Stanza	COMB	$90.06^{\pm 0.11}$	$67.54^{\pm 1.24}$	$89.58^{\pm 0.0}$	Stanza	COMB	$89.79^{\pm 0.23}$	$66.67^{\pm 1.09}$	$92.66^{\pm 0.0}$

		Lengadocian					Provençau		
Tool	Train	ALL	UNK	AMB	Tool	Train	ALL	UNK	AMB
MaChAmp	L+S	$93.08^{\pm 0.48}$	$69.91^{\pm 0.33}$	$92.76^{\pm 0.69}$	MaChAmp	L+S	$91.67^{\pm 0.0}$	$54.67^{\pm 1.89}$	$95.14^{\pm 0.44}$
MaChAmp	L+LEN	$92.56^{\pm 0.6}$	$68.29^{\pm 0.33}$	$92.29^{\pm 0.8}$	MaChAmp	L+PRO	$86.6^{\pm 0.11}$	$52.0^{\pm 0.0}$	$89.55^{\pm 0.25}$
Stanza	SMALL	**94.42** $^{\pm 0.13}$	**81.35** $^{\pm 0.9}$	**96.54** $^{\pm 0.0}$	Stanza	SMALL	**92.81** $^{\pm 0.31}$	**74.92** $^{\pm 1.28}$	**98.51** $^{\pm 0.0}$
Stanza	COMB	$93.72^{\pm 0.11}$	$71.53^{\pm 1.5}$	$92.98^{\pm 0.0}$	Stanza	COMB	$92.08^{\pm 0.12}$	$54.67^{\pm 1.89}$	$93.51^{\pm 0.0}$

Table 4: Dialect-Specific Lemmatization Accuracy on Occitan

7.2 Dialect-Specific Results

When testing on individual dialects, too, the sequential approach of the Stanza model most often yields a higher accuracy for both Low Saxon and Occitan. As in case of the general tests, we do not find the automatically annotated data to benefit the model performance on Occitan. However, for both North and South German Low Saxon, we observed an improvement of the overall accuracy. Furthermore, we find MaChAmp to generalise particularly well to Lemosin.

When comparing the performance of the general and the dialect-specific MaChAmp models, the finetuning on a small dialect-specific dataset does not bring any improvement except for unknown tokens in German North Low Saxon. The MaChAmp models in fact consistently show a better overall accuracy when finetuned on the general gold train data. Since the general gold train data combines all the dialect-specific train sets, it is reasonable to suppose that these results are driven by the size difference between the finetuning datasets.

For Stanza, a more focused approach – here exclusive training on gold data without adding automatically annotated data – leads to a higher accuracy for lemmatizing unknown tokens. This holds true for both Low Saxon and Occitan, with the exception of Lemosin.

8 Discussion and Conclusion

The overall accuracy results for Low Saxon are noticeably lower than for Occitan, around 10% on average. One possible explanation could be the greater orthographic variation that is likely the reason behind the higher percentages of unknown and ambiguous tokens in Low Saxon seen in Table 2. While our Occitan corpus makes use of the same spelling convention throughout, the Low Saxon corpus contains various writing systems even within the same dialect group. Furthermore, we trained the models for Occitan on major dialects, whereas we used groups of major dialects for Low Saxon. Another reason might be found in the different diachronic structure of the datasets: Whereas the Occitan data mostly comes from the 20th and 21st century, the Low Saxon dataset covers the period

Dutch Low Saxon				
Tool	Train	All	Unk	Amb
MaChAmp	L+S	$77.46^{\pm0.24}$	$11.11^{\pm1.13}$	$82.39^{\pm0.22}$
MaChAmp	L+DLS	$76.31^{\pm0.23}$	$10.65^{\pm0.66}$	$81.16^{\pm0.08}$
Stanza	SMALL	**80.41** $^{\pm0.81}$	**21.3** $^{\pm4.72}$	**84.45** $^{\pm0.66}$
Stanza	COMB	$78.93^{\pm0.11}$	$14.35^{\pm1.31}$	$81.98^{\pm0.0}$

(German) North Low Saxon				
Tool	Train	All	Unk	Amb
MaChAmp	L+S	$86.77^{\pm0.8}$	$30.55^{\pm3.93}$	**90.35** $^{\pm0.62}$
MaChAmp	L+NLS	$82.65^{\pm0.32}$	**33.33** $^{\pm6.81}$	$85.35^{\pm0.79}$
Stanza	SMALL	$84.79^{\pm0.92}$	**33.33** $^{\pm6.81}$	$89.01^{\pm1.08}$
Stanza	COMB	**89.6** $^{\pm0.12}$	$30.55^{\pm3.93}$	$89.01^{\pm0.0}$

(German) South Low Saxon				
Tool	Train	All	Unk	Amb
MaChAmp	L+S	$73.97^{\pm0.22}$	$45.45^{\pm0.00}$	$74.49^{\pm0.26}$
MaChAmp	L+SLS	$72.74^{\pm0.54}$	$42.42^{\pm4.29}$	$73.57^{\pm0.62}$
Stanza	SMALL	$78.15^{\pm0.56}$	**46.97** $^{\pm2.14}$	**79.42** $^{\pm1.22}$
Stanza	COMB	**79.68** $^{\pm0.08}$	$33.33^{\pm2.14}$	$78.44^{\pm0.0}$

Table 5: Dialect-Specific Lemmatization Accuracy for Low Saxon

from the 19th century to the 21st.

This greater variation might be the reason why a sequential approach proves particularly useful for Low Saxon. As a result of the dialectal and orthographic variation, there are many ambiguous tokens that need to be lemmatized differently depending on the writing system and dialect. For instance, the character string *doe* typically refers to the feminine or masculine definite article in eastern Westphalian, where it should be lemmatized as *de*, whereas it should be lemmatized as *du* in Gronings, where it represents the pronoun of the second person singular. In addition, this string can stand for the 1st person singular in the present tense of the verb *doon* ‘to do’ in many dialects throughout the language area. PoS-tagging effectively disambiguates these three usages.

When it comes to Occitan, we noted that the Stanza model trained only on the gold data performs better than its counterpart trained on both preannotated and gold data. This may be due to the genre mismatch between the gold corpus (which is predominantly literary) and the automatically annotated corpus (which is extracted from Wikipedia). MaChAmp’s finetuning approach seems to be more robust to this, since the model trained on both preannotated and gold data achieves better general results than the the one limited to the gold dataset.

The different model behaviour we have observed in our two low-resourced languages also warrants a more general question: How faithfully can low-resource scenarios be simulated by using small amounts of data from standardized high-resource languages? As this seems to be a relatively common practice, it would be worth investigating how this approach actually compares to the task it is supposed to simulate.

In conclusion, we found that the sequential approach implemented by Stanza was a good fit for both languages. The amount of training data also seemed to have more of an impact than dialect-level specificity, given that the MaChAmp models finetuned on the full gold dataset systematically outperformed the dialect-specific models.Another common tendency for both languages is the positive effect of using only gold data for training on the performance of the Stanza model over unknown tokens. This is a particularly interesting finding because it could be expected that a larger amount of training would make the model generalize better. It seems that in our case the reliability of the training data was more important.

Data Access

The new annotated corpora created as part of this work are distributed on Zenodo.

The datasets for Low Saxon are available here: https://doi.org/10.5281/zenodo.7777282.

The large dataset for Occitan is available here: https://doi.org/10.5281/zenodo.7777340.

Limitations

The MaChAmp and Stanza results are not fully comparable as we did not present the performance of dialect-specific Stanza models here. Since Stanza does not allow finetuning, we do not expect the small individual dialect-specific train sets to have a strong effect compared with the much larger amount of automatically annotated data. We defer testing this hypothesis to future work.

Acknowledgements

This work has been supported by the Academy of Finland through project No. 342859 “CorCoDial – Corpus-based computational dialectology”.

References

Pierre Bec. 1970. *Manuel pratique de philologie romane*, volume Vol. 1. Picard.

Pierre Bec. 1995. *La langue occitane*, 6th edition. PUF.

Toms Bergmanis and Sharon Goldwater. 2018. Context sensitive neural lemmatization with Lematus. 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 1391–1400, New Orleans, Louisiana. Association for Computational Linguistics.

Henk Bloemhoff, Philomèle Bloemhoff de Bruijn, Jan Nijen Twilhaar, Henk Nijkeuter, and Harrie Scholtmeijer. 2019. *Nedersaksisch in een notendop – Inleiding in de Nedersaksische taal en literatuur.* Koninklijke Van Gorcum, Assen.

Myriam Bras, Marianne Vergez-Couret, Nabil Hathout, Jean Sibille, Aure Séguier, and Benazet Dazéas. 2020. Loflòc : Lexic obèrt flechit occitan. In *Fidélités et dissidences (Actes du* XIIe *congrès de l'Association Internationale d'Études Occitanes)*, Albi. Centre d'Etude de la Littérature Occitane.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Georgiana Dinu Grzegorz Chrupala and Josef van Genabith. 2008. Learning morphology with morfette. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). Http://www.lrec-conf.org/proceedings/lrec2008/.

Daniel Kondratyuk, Tomáš Gavenčiak, Milan Straka, and Jan Hajič. 2018. LemmaTag: Jointly tagging and lemmatizing for morphologically rich languages with BRNNs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4921–4928, Brussels, Belgium. Association for Computational Linguistics.

Alfred Lameli. 2016. Raumstrukturen im Niederdeutschen. Eine Re-Analyse der Wenkerdaten. *Niederdeutsches Jahrbuch. Jahrbuch des Vereins für niederdeutsche Sprachforschung*, 139:131–152.

Agathe. Lasch. 1974. *Mittelniederdeutsche Grammatik.* Sammlung kurzer Grammatiken germanischer Dialekte ; 9. Niemeyer, Halle a. S.

Agathe Lasch, Conrad Borchling, Gerhard Cordes, Dieter Möhn, Ingrid Schröder, Jürgen Meier, and Sabina Tsapaeva. *Mittelniederdeutsches Handwörterbuch.*

Wolfgang Lindow, Dieter Möhn, Hermann Niebaum, Dieter Stellmacher, Hans Taubken, and Jan Wirrer. 1998. *Niederdeutsche Grammatik.* Schuster, Leer.

Nikola Ljubešić and Kaja Dobrovoljc. 2019. What does neural bring? analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 29–34, Florence, Italy. Association for Computational Linguistics.

August. Lübben. 1995 - 1888. *Mittelniederdeutsches Handwörterbuch.* Wissenschaftliche Buchgesellschaft, Darmstadt.

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.

Aleksandra Miletic, Myriam Bras, Marianne Vergez-Couret, Louise Esher, Clamença Poujade, and Jean Sibille. 2020. A four-dialect treebank for Occitan: Building process and parsing experiments. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 140–149, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Christopher Moseley, editor. 2010. *Atlas of the World's Languages in Danger*, 3 edition. UNESCO Publishing, Paris. Online version: `http://www.unesco.org/culture/en/endangeredlanguages/atlas`.

Thomas Müller, Ryan Cotterell, Alexander Fraser, and Hinrich Schütze. 2015. Joint lemmatization and morphological tagging with lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2268–2274, Lisbon, Portugal. Association for Computational Linguistics.

Hermann Niebaum. 2008. Het Nederduits. In Henk Bloemhoff, Jurjen van der Kooi, Hermann Niebaum, and Siemon Reker, editors, *Handboek Nedersaksische Taal- en Letterkunde*, pages 430–447. Koninklijke Van Gorcum, Assen.

Robert Peters. 2017. Das referenzkorpus mittelniederdeutsch/ niederrheinisch (1200–1650). *Niederdeutsches Jahrbuch. Jahrbuch des Vereins für niederdeutsche Sprachforschung*, 140:35–42.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Kumar Saunack, Kumar Saurav, and Pushpak Bhattacharyya. 2021. How low is too low? a monolingual take on lemmatisation in Indian languages. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4088–4094, Online. Association for Computational Linguistics.

Kumar Saurav, Kumar Saunack, and Pushpak Bhattacharyya. 2020. Analysing cross-lingual transfer in lemmatisation for Indian languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6070–6076, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ingrid Schröder. 2004. Niederdeutsch in der Gegenwart: Sprachgebiet – Grammatisches – Binnendifferenzierung. In Dieter Stellmacher, editor, *Niederdeutsche Sprache und Literatur der Gegenwart*, pages 35–97. Georg Olms Verlag, Hildesheim, Zürich and New York.

Jean Sibille. 2002. Ecrire l'occitan : essai de présentation et de synthèse. In *Les langues de France et leur codification. Ecrits divers – Ecrits ouverts*, Paris, France. Inalco / Association Universitaire des Langues de France, L'Harmattan.

Janine Siewert, Yves Scherrer, and Martijn Wieling. 2022. Low Saxon dialect distances at the orthographic and syntactic level. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 119–124, Dublin, Ireland. Association for Computational Linguistics.

Janine Siewert, Yves Scherrer, Martijn Wieling, and Jörg Tiedemann. 2020. LSDC - a comprehensive dataset for low Saxon dialect classification. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 25–35, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Dieter Stellmacher. 1983. Neuniederdeutsche Grammatik – Phonologie und Morphologie. In Gerhard Cordes and Dieter Möhn, editors, *Handbuch zur niederdeutschen Sprach- und Literaturwissenschaft*, pages 238–278. Erich Schmidt Verlag, Berlin, Germany.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Alessandro Vatri and Barbara McGillivray. 2020. Lemmatization for ancient greek: An experimental assessment of the state of the art. *Journal of Greek linguistics*, 20(2):179–196.

Marianne Vergez-Couret. 2016. Description du lexique Loflòc. Research report, CLLE-ERSS.

A Complementary Evaluation Results

	Tool	Training set	Task	Train cond.	Test cond.	ALL	UNK	AMB
Occitan	MaChAmp	SMALL	LEM	no POS, gold LEM	no POS	$93.57^{\pm0.06}$	$78.74^{\pm1.14}$	$95.08^{\pm0.28}$
		LARGE	POS+LEM	pred. POS+LEM	no POS	$93.32^{\pm0.09}$	$76.07^{\pm0.5}$	$91.94^{\pm0.15}$
		L+S	POS+LEM	pred. POS+LEM	no POS	$94.24^{\pm0.17}$	$73.49^{\pm0.74}$	$93.47^{\pm0.29}$
	Stanza	SMALL	LEM	no POS	no POS	$92.84^{\pm0.14}$	$75.43^{\pm0.84}$	$93.1^{\pm0.0}$
		SMALL	LEM	gold POS+LEM	pred. POS	$94.68^{\pm0.03}$	$83.16^{\pm0.21}$	$94.86^{\pm0.0}$
		COMB	LEM	pred. POS+LEM	pred. POS	$93.53^{\pm0.06}$	$74.01^{\pm1.11}$	$91.19^{\pm0.0}$
Low Saxon	MaChAmp	SMALL	LEM	no POS, gold LEM	no POS	$74.72^{\pm0.62}$	$25.48^{\pm1.67}$	$91.84^{\pm0.44}$
		LARGE	POS+LEM	pred. POS+LEM	no POS	$86.69^{\pm0.31}$	$52.01^{\pm0.89}$	$89.14^{\pm0.46}$
		L+S	POS+LEM	pred. POS+LEM	no POS	$81.64^{\pm0.43}$	$56.74^{\pm0.58}$	$83.41^{\pm0.38}$
	Stanza	SMALL	LEM	no POS	no POS	$78.7^{\pm0.23}$	$42.38^{\pm0.9}$	$85.09^{\pm0.0}$
		SMALL	LEM	gold POS+LEM	pred. POS	$82.4^{\pm0.16}$	$47.64^{\pm0.61}$	$92.15^{\pm0.0}$
		COMB	LEM	pred. POS+LEM	pred. POS	$83.54^{\pm0.17}$	$55.79^{\pm2.34}$	$83.95^{\pm0.0}$

Table 6: Global Lemmatization Accuracy for Occitan and Low Saxon. *Dev* set.

		Gascon					**Lengadocian**		
Tool	Train	ALL	UNK	AMB	Tool	Train	ALL	UNK	AMB
MaChAmp	L+S	$93.6^{\pm0.36}$	$69.1^{\pm1.15}$	$93.55^{\pm1.11}$	MaChAmp	L+S	$94.4^{\pm0.14}$	$75.58^{\pm0.95}$	$93.45^{\pm0.1}$
MaChAmp	L+GAS	$92.83^{\pm0.1}$	$69.1^{\pm2.3}$	$92.14^{\pm0.22}$	MaChAmp	L+LEN	$94.12^{\pm0.13}$	$74.03^{\pm1.09}$	$93.25^{\pm0.11}$
Stanza	SMALL	$94.29^{\pm0.2}$	$79.65^{\pm0.99}$	$96.15^{\pm0.0}$	Stanza	SMALL	$94.78^{\pm0.02}$	$84.29^{\pm0.16}$	$94.51^{\pm0.0}$
Stanza	COMB	$90.4^{\pm0.0}$	$68.29^{\pm0.0}$	$87.26^{\pm0.0}$	Stanza	COMB	$94.33^{\pm0.08}$	$76.75^{\pm1.65}$	$92.16^{\pm0.0}$

Table 7: Dialect-Specific Lemmatization Accuracy on Occitan. *Dev* set (there are no dialect-specific *dev* sets for Lemosin and Provençau.)

		Dutch Low Saxon					**(German) North Low Saxon**		
Tool	Train	All	Unk	Amb	Tool	Train	All	Unk	Amb
MaChAmp	L+S	$83.64^{\pm0.75}$	$60.0^{\pm1.26}$	$86.57^{\pm0.61}$	MaChAmp	L+S	$84.49^{\pm0.46}$	$60.49^{\pm1.74}$	$86.7^{\pm0.49}$
MaChAmp	L+DLS	$81.35^{\pm0.13}$	$52.31^{\pm0.00}$	$84.77^{\pm0.35}$	MaChAmp	L+NLS	$80.96^{\pm0.22}$	$60.49^{\pm3.49}$	$83.42^{\pm0.37}$
Stanza	SMALL	$84.2^{\pm0.5}$	$50.26^{\pm2.62}$	$87.55^{\pm0.11}$	Stanza	SMALL	$83.97^{\pm0.46}$	$45.68^{\pm1.75}$	$87.39^{\pm0.95}$
Stanza	COMB	$84.11^{\pm0.35}$	$55.9^{\pm3.84}$	$84.52^{\pm0.0}$	Stanza	COMB	$87.96^{\pm0.08}$	$65.43^{\pm1.75}$	$88.08^{\pm0.0}$

		(German) South Low Saxon		
Tool	Train	All	Unk	Amb
MaChAmp	L+S	$76.82^{\pm0.07}$	$50.34^{\pm0.96}$	$77.63^{\pm0.11}$
MaChAmp	L+SLS	$74.25^{\pm0.71}$	$48.3^{\pm1.92}$	$74.68^{\pm0.88}$
Stanza	SMALL	$78.96^{\pm0.34}$	$42.86^{\pm0.0}$	$81.41^{\pm0.49}$
Stanza	COMB	$78.91^{\pm0.15}$	$50.34^{\pm1.92}$	$79.82^{\pm0.0}$

Table 8: Dialect-Specific Lemmatization Accuracy for Low Saxon. *Dev* set.