<https://helda.helsinki.fi>

Adaptation of morphosyntactic taggers : Cross-lectal and multilectal approaches

Scherrer, Yves

Cambridge University Press
2021-08

Scherrer , Y 2021 , Adaptation of morphosyntactic taggers : Cross-lectal and multilectal approaches . in M Zampieri & P Nakov (eds) , Similar languages, varieties, and dialects : A computational perspective . Studies in Natural Language Processing , Cambridge University Press , pp. 138-166 . <https://doi.org/10.1017/9781108565080.011>

<http://hdl.handle.net/10138/334252>

<https://doi.org/10.1017/9781108565080.011>

cc_by_sa

acceptedVersion

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.

7

Adaptation of Morphosyntactic Taggers: Cross-lectal and Multilectal Approaches

Yves Scherrer

7.1 Introduction

The topic of this book, “similar languages, varieties, and dialects,” encompasses several linguistic and sociolinguistic configurations that call for adapted solutions. Most research in this area approaches these configurations from the angle of data availability and assumes that they show an imbalance of high-resource and low-resource varieties. A popular approach is thus to create models for the high-resource varieties and adapt them in some way to the low-resource varieties, which is facilitated by the posited similarity between the varieties. While such cross-lingual (or more generally, cross-lectal) approaches have been proposed for a range of natural language processing (NLP) tasks such as dependency parsing (see Chapter 8), named-entity recognition, or semantic role labeling, this chapter focuses on the task of morphosyntactic tagging. Morphosyntactic tagging is considered one of the core language technologies (Rehm and Uszkoreit, 2013) and often serves as a pre-annotation layer for higher-level tasks such as parsing. Furthermore, morphosyntactic annotation is useful for corpus-linguistic research as it enables search queries that are independent of the lexical forms, which is especially practical for non-standardized varieties such as dialects.

In this introduction, we present the task of morphosyntactic tagging and the most popular model architectures currently in use (Section 7.1.1), we describe different linguistic and sociolinguistic scenarios and their impact on data availability (Section 7.1.2) and give a general overview of cross-lingual and cross-lectal transfer learning (Section 7.1.3). In the remaining sections, we describe the different families of cross-lingual learning approaches in detail, providing examples of recent research on tagging for similar language varieties.

7.1.1 Morphosyntactic Tagging

Tagging in general refers to any kind of sequence labeling task. In natural language processing, sequences are typically sentences consisting of words, and every word of this sequence is assigned a label out of a predefined set of labels. Morphosyntactic tagging is one of the most prominent sequence labeling tasks, and two flavors – *part-of-speech tagging* and (for lack of a better term) *full morphological tagging* – can be distinguished.

Part-of-speech (POS) tagging refers to the process of assigning a part-of-speech label to each word of a sequence. Part-of-speech labels correspond to the main word classes or syntactic categories, such as noun, verb, adjective, pronoun, adverb, and punctuation. The number of part-of-speech labels varies across different annotation schemes and traditions,¹ but recent work has striven to define a core label set of ten to twenty labels that is as language-independent as possible (Petrov et al., 2012). For part-of-speech tagging, it is common to achieve more than 95 percent of accuracy with the adequate training data, to the point that it has been called a “solved task” (see, e.g., Manning, 2011, for a critical discussion).

In contrast to part-of-speech tagging, full morphological tagging relies on richer tag sets. The tags include, in addition to the main word class, a set of values for morphological features such as gender, number, case, or person, depending on the morphological complexity of the language (Hajič, 2000). As a rule of thumb, the full morphological label should, together with the lemma, enable the unambiguous reconstruction of the inflected word form. Whereas the set of part-of-speech tags is fairly language-independent, the inventory of morphosyntactic tags depends to a large extent on the particular linguistic systems. For instance, not all languages morphologically encode gender, and not all of those that do share the same values of the gender feature. Moreover, the combinatorics of different features results in an explosion of the number of distinct tags seen in a training corpus. For morphologically rich languages, it is common to have tag inventories of more than 1000 distinct elements (Erjavec, 2010). This makes full morphological tagging considerably more difficult than part-of-speech tagging.

Historically, there has not been a clear distinction between part-of-speech tagging and full morphological tagging. One of the most influ-

¹ For example, there is some debate about whether proper nouns should be labeled differently from common nouns or not, or whether auxiliary verbs should be distinguished from lexical verbs or not (Nivre et al., 2016).

ential tag sets is the Penn Treebank tag set for English (Marcus et al., 1993), which defines forty-five labels. Due to the low complexity of English inflectional morphology, this tag set is sufficient for full morphological tagging of English, as per the definition given. Simultaneous annotation efforts for other languages provided tag sets of similar sizes (an overview is given in Petrov et al., 2012), but these could most often not provide full morphological disambiguation due to the higher morphological complexity of the corresponding languages. It is only with recent work on “universal” annotation that this distinction has been clarified, first through the development of a language-independent universal part-of-speech tag set (Petrov et al., 2012) and of a unified feature-value inventory for morphological features (Zeman, 2008), and later through the harmonization of the two within the universal dependencies project (Nivre et al., 2016).

Morphosyntactic tagging is a disambiguation task. For those words that occur only with a single label in the training corpus, the prediction is trivial. Depending on the task (part-of-speech tagging or morphological tagging) and the language, the proportion of such unambiguous words is considerable and may exceed 50 percent (Jurafsky and Martin, 2009). Hence, the two main difficulties that a tagger needs to solve are ambiguous words – word forms associated with more than one label in the training corpus –, and unknown words – word forms not encountered at all in the training corpus.

In contrast to other NLP tasks such as parsing, morphological analysis and machine translation, rule-based taggers relying on manually coded (morpho-)syntactic knowledge have never been popular. The few references in that domain use a full-fledged parser internally and restrict its output to the labels assigned to the terminal nodes (e.g., Tapanainen and Voutilainen, 1994; Scherrer, 2008). Tagging as such has typically been viewed as a disambiguation task that is most effectively handled using data-driven methods trained on hand-annotated corpora.

One of the simplest model architectures for tagging is the Hidden Markov Model (HMM), which models sequence labeling as a combination of emission probabilities (the association of tags with words) and transition probabilities (the probabilities of tag sequences). The former generally include some back-off mechanism based on suffix analysis for handling unknown words. Two of the historically most popular taggers, TreeTagger (Schmid, 1994) and TnT (Brants, 2000), are based on a HMM architecture and have been found hard to beat, especially in low-resource scenarios (Plank et al., 2016). Other, more practical reasons for

the popularity of these systems include the availability of pre-trained TreeTagger models for a large number of languages, and the fact that TnT – unlike its open-source reimplementation HunPos (Halácsy et al., 2007) – stores the model parameters in plain text files that are easy to modify after training.

Most subsequent work has focused on alternative machine learning algorithms that support more sophisticated feature representations than HMMs. For example, Ratnaparkhi (1996) and Toutanova et al. (2003) use Maximum-Entropy Markov Models instead of HMMs, Giménez and Màrquez (2004) use support vector machine (SVM) classifiers, and Müller et al. (2013) rely on a conditional random fields (CRF) architecture. The latter proved to be especially well suited for large tag sets such as those used in full morphological tagging.

In recent years, tagger architectures based on neural networks have been introduced successfully, following Collobert et al. (2011). The advantage of neural networks lies in the fact that the feature representations (e.g., whether the tagger should focus on suffixes of a certain length, on capitalization, or on particular combinations of words) do not have to be specified explicitly in advance, but are learned directly within the model. Neural network taggers generally use a combination of character-level representations that inform about internal word structure (*character embeddings* or, more precisely, *character-level word embeddings*) and word-level representations that inform about the distributional similarities of words in their sentential context (*word embeddings*). Most neural network taggers use recurrent neural networks, in which the internal states of the network at each word (and at each decision) are connected. Different types of connections, such as long short-term memory (LSTM) or gated recurrent unit (GRU) cells, are available, and connections can be unidirectional (from left to right) or bidirectional. A representative example of such a tagger architecture is described in Plank et al. (2016). Other architectures use convolutional neural networks (CNNs) to represent the input data (e.g., Yu et al., 2017), or add a CRF on top of the output layer to capture the contextual dependencies between adjacent tags (e.g., Lample et al., 2016; Yang and Zhang, 2018).

For full morphological tagging, the question about the representation of complex tags arises. A tag is typically encoded as a sequence of feature-value pairs, and two tags may thus be distinct to several degrees, as in examples (a)–(c):

- (a) POS=VERB|Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin
- (b) POS=VERB|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin

(c) POS=NOUN|Animacy=Anim|Case=Nom|Gender=Masc|Number=Sing

The most straightforward option is to consider each tag string as a distinct entity. Tagging a word thus amounts to picking one tag string among all distinct tag strings seen in the training data. With this option, the same model architectures as for POS-tagging can be used, but it is impossible for the model to learn generalizations about particular features, e.g., that (a) and (b) both represent finite verb forms.

Another option is to model each feature separately. Tagging a word then amounts to picking one value for each feature (absent features are typically given a null value that is removed later) and concatenating them into a tag string. This approach requires changes to the model architecture, but allows the model to generalize better. While most work is based on the first option, Müller et al. (2013) and Pinter et al. (2017) are notable representatives of the second option, the former relying on a CRF model, the latter on a recurrent neural network model. Our own recent experiments (Scherrer and Rabus, 2017; Scherrer et al., 2018; Scherrer and Rabus, 2019) use these two tagging architectures. A related approach has been proposed by Silfverberg and Drobac (2018), who view the tag string as a sequence that is generated one feature at a time.

A recent addition to neural NLP are multi-task models (Ruder, 2017). Such models are trained not only on the main task (part-of-speech or morphological tagging in the present case) but also on some auxiliary task(s). A popular auxiliary task is the prediction of word frequency, following the intuition that low-frequency words have different tag distributions than high-frequency words (Plank et al., 2016). In the case of similar language varieties, it may be useful to train a single tagger model on multiple varieties and include variety identification as an auxiliary task. This allows the tagger to take advantage of the increased training corpus size, but also to learn differences between the varieties. Such an approach has been proposed by Cotterell and Heigold (2017) and has been applied successfully in Scherrer and Rabus (2019).

Different tagging models have been compared in Plank et al. (2016) for part-of-speech tagging, and in Horsmann and Zesch (2017) for finer-grained tag sets that lie somewhere between part-of-speech tagging and full morphological tagging. These comparisons show (1) that neural-network-based taggers generally surpass traditional methods, but only if the training corpora are large enough, and (2) that neural-network-based taggers are especially well suited to tagging problems with large tag sets.

However, results have been found to depend a lot on the considered languages and tag sets.

7.1.2 Scenarios

This book intends to offer an overview of NLP methods for similar languages, similar varieties, and dialects. While it is well known that the distinction between a language, a language variety, and a dialect is determined to a large extent by political and sociolinguistic criteria rather than purely linguistic ones (Trudgill, 2000), these criteria often have a direct impact on the usage of the varieties, and therefore on the type of data that are available for computational research (see also Chapter 3).

Similar languages are generally characterized by the fact that each of them has some degree of official recognition (nationally or regionally), is standardized, and can be qualified as *autonomous* or *Ausbau* language (Trudgill, 2004) as a result of its independent cultural status. These attributes are typically lacking for *dialects*: most dialects are not autonomous language varieties, but are dependent on one variety of particular status, which is referred to as their *roofing language*. For example, Swiss German dialects depend on Standard German in the sense that formal written communication in Switzerland is carried out in Standard German, just like Arabic dialects are roofed by standardized pan-Arabic varieties such as Modern Standard Arabic (cf. Chapter 13). In the context of cross-lectal learning, the roofing language is often a natural choice as the high-resource variety for its associated (low-resource) dialects. Moreover, the absence of standardization for dialects also means that pronunciation differences are reflected more easily in writing than in standardized varieties; this greater variability in the textual data poses particular challenges for NLP applications such as tagging.

The relation between dialects and their roofing variety mirrors the relation between different diachronic states of a language and its present-day variety: in most linguistic areas of Europe, orthographic standardization has set in only a few centuries ago (Baddeley and Voeste, 2012), such that earlier stages of written language exhibit the same type of spelling variation that is observed today in dialectal variation. In consequence, methods that have been developed in the context of historical language varieties may also prove useful for the processing of synchronic dialectal varieties.

Similar varieties typically refer to different national or regional varieties of a pluricentric language (Clyne, 1992). Such varieties follow to

a large extent the same standardization guidelines. Surprisingly, discussion of such phenomena is almost completely lacking in NLP research, and many data sets do not even specify the variety they represent.² The underlying assumptions seem to be that (1) the inter-variety differences are too infrequent to require specific modeling, and that (2) some of the data sets include data from several varieties, such that models trained on these data sets end up being relatively variety-independent. Indeed, in one of the rare studies involving tagging of different national varieties of Spanish (Francom et al., 2014), the authors find that data annotation inconsistencies and register mismatch affect tagging accuracy more than regional variation.

The status of similar languages and varieties as autonomous languages facilitates the collection of textual material for corpus construction, as popular domains and genres such as newspaper text, legal documents, film subtitles, and literary texts are in principle available. This contrasts with dialects that lack official status and therefore see limited written production. This especially holds in situations of *diglossia* (Ferguson, 1959), where the different varieties are used in different functional contexts. As a result, dialect corpora most often contain transcriptions of speech or text from computer-mediated communication, which are fundamentally different from the domains and genres typically found in NLP corpora of standardized varieties (see Chapter 6).

In all scenarios, the overarching question is about how to deal with the variation that occurs in the considered varieties: should it be neutralized so that simpler, variation-independent models can be built, or should variety-specific models be created? In NLP for historical language varieties, this question has largely been answered in favor of the former approach, using normalization techniques to reduce the massive amounts of variation (see Section 7.3). For related varieties and dialects, both approaches are technically viable, and the answer to this question ultimately lies in the aims of the computational modeling rather than in the technical possibilities.

² For example, none of the six English treebanks available in the Universal Dependencies repository (as of June 29, 2018) specifies explicitly which national or regional variety of English it covers, although the data sources and the places of data collection may hint at the regional variety of the data. The same holds for the six French treebanks and the three Spanish treebanks. Only one of the Portuguese treebanks explicitly states that it contains material from both European and Brazilian variants.

7.1.3 Cross-lingual and Cross-lectal Learning

In the previous section, we sketched different scenarios, all of which concern sets of closely related language varieties. It is typical of these scenarios that the availability of linguistic resources (in the form of annotated corpora for training) varies widely across the members of such a set. Therefore, the wish may occur to leverage the better-resourced varieties of the set in order to annotate the less-resourced ones. This is the core idea of cross-lectal learning.

It is not always obvious to identify the high-resource and low-resource varieties in a set of closely related varieties. Whereas dialects generally sport fewer resources than their corresponding roofing languages, the resource distributions of pluricentric language varieties often depend on coincidental factors and may even be task dependent. For example, Latin American varieties of Spanish tend to be less rich in NLP resources than European varieties because research has historically been more active on the Iberic peninsula than in Latin America. Danish may be more resourced than Norwegian when it comes to parallel data thanks to Denmark’s membership in the European Union and the related translation activities, but Norwegian is currently more resourced in terms of syntactically annotated data than Danish.³

We argue here that the similar languages scenarios can often be cast as a low-resource languages problem, for which a number of cross-lingual transfer learning methods have been proposed (Tiedemann and Agić, 2016). The general goal of cross-lingual transfer learning is to create tools for a *low-resource language* (LRL) when training data are only available for a (not necessarily related) *high-resource language* (HRL); in this context, the high-resource language is also referred to as *donor language* or *source language*, and the low-resource language as the *recipient language* or *target language*. Although there seems to be a general consensus that transfer learning methods work better when the involved language varieties are closely related, the majority of the proposed methods aims to be as language-independent as possible and forgoes the explicit modeling of features that hold only for related varieties (e.g., surface similarities like large number of cognates or structural similarities like word order). Here, we would like to take a different stance and present models that sacri-

³ As of June 2018, the <http://universaldependencies.org> repository contains one Danish corpus with 100,000 words, whereas both varieties of Norwegian, Bokmål and Nynorsk, are represented by corpora with more than 300,000 words.

face language independence for better adaptation to related language varieties.

In the following sections, we present the three main approaches to cross-lingual transfer learning for similar language varieties. The first approach, subsumed under the name *model transfer*, aims at training tagging models directly on annotated HRL data in such a way that they end up being generic enough to be applied without change to a LRL (Section 7.2). The second approach operates on the data level and has consequently been called *data transfer*. Its goal is to make the data of the HRL and LRL look more similar than they actually are, such that a tagger trained on one variety can be applied to the modified version of the data of the other variety (Section 7.3). The most popular flavor of this approach consists in *normalizing* the LRL towards the HRL. The third approach acknowledges the fact that there are different varieties with different properties and thus explicitly aims at building multilingual, or multilectal, models. The underlying assumption is that small amounts of data can be made available even for the low-resource varieties and that these data sets can be combined intelligently with the larger data sets of the high-resource ones (Section 7.4).

7.2 Model Transfer Methods

7.2.1 Plain Model Transfer

One of the most straightforward techniques to tag a LRL when only training data from a related HRL are available is to assume that the two varieties are in fact the same, or at least so similar that they do not require different models. In this case, a tagger is trained exclusively on HRL data and applied directly and without any modification to data from the LRL. While this assumption is too naive in most cases, this approach is often presented as a simple baseline with which more sophisticated approaches are compared. However, most work in the area of similar varieties of pluricentric languages implicitly follows this strategy by assuming that the overlap in the lexicon and in morphosyntactic structures is so high that it is not worth while training distinct taggers for each variety.

In Scherrer (2014), we report plain model transfer baseline results for various West Germanic, West Slavic, and Ibero-Romance languages: a TnT tagger is trained on a HRL and applied without modification to

HRL → LRL	TnT	BTagger	MarMoT	CLSTM
Czech → Polish	66.4%	69.3%	67.2%	66.9%
Czech → Slovak	80.6%	81.1%	80.1%	82.0%
German → Dutch	24.5%	32.7%	25.3%	32.0%
Spanish → Catalan	58.4%	67.2%	62.8%	59.1%
Spanish → Aragonese	71.6%	72.7%	72.0%	73.2%

Table 7.1 *Tagging Accuracies for Systems Trained on the HRL and Tested on the LRL. The data sets are described in Scherrer (2014).*

a related LRL. Table 7.1 extends these experiments to a larger set of tagger architectures:

- **TnT** (Brants, 2000): a non-neural, HMM-based tagger that relies on suffix analysis for unknown words;
- **BTagger** (Gesmundo, 2011): a non-neural perceptron-based tagger based on bidirectional guided learning and semi-automatic feature template extraction;
- **MarMoT** (Müller et al., 2013): a non-neural CRF-based tagger;
- **CLSTM** (Scherrer and Rabus, 2019): a neural tagger using a character-level LSTM to represent words.

The results show that this naive approach can work surprisingly well even in configurations where the HRL and LRL are considered distinct languages (e.g., Czech → Slovak with over 80 percent accuracy, Spanish → Aragonese with over 70 percent accuracy). The success of plain model transfer depends a lot on the type of feature representations used when training the HRL tagger: modern discriminative or neural tagging models such as BTagger or CLSTM provide fair tagging performance even with high out-of-vocabulary rates, thanks to their fine-grained feature representations. The differences are especially striking for language pairs with different spelling conventions such as German → Dutch.

7.2.2 Multi-source Transfer

There are linguistic configurations with a clearly defined low-resource variety, but with several related high-resource varieties, all of which are expected to contribute to various extents to the tagging of the low-resource variety. One such case is Rusyn, a Slavic variety spoken predominantly in Transcarpathian Ukraine, eastern Slovakia, and southeastern Poland,

and related most closely to Ukrainian but also to Polish and Slovak. Scherrer and Rabus (2017) train a single tagger on the concatenation of Slovak, Ukrainian, Polish, and Russian data and apply this multi-lingual tagger directly to Rusyn, achieving 79.2 percent part-of-speech accuracy and 72.0 percent full morphological tagging accuracy, with an out-of-vocabulary (OOV) rate of 26.4 percent. For comparison, the best single-source model is the one trained on Ukrainian, which yields 69.7 percent POS accuracy, 63.4 percent morphology accuracy, and 37.1 percent OOV rate. An ablation study furthermore showed that in agreement with linguistic knowledge, removing Ukrainian from the multi-source tagger hurts its performance most, whereas the contributions of Polish and Slovak are minor and the Russian corpora do not facilitate Rusyn tagging at all. Follow-up experiments with neural taggers on the same multi-source setup showed improvements of up to 5 percent absolute in terms of full morphological tagging accuracy (Scherrer and Rabus, 2019).

Multi-source transfer approaches must meet two important requirements. First, all source languages need to share the same tag set and annotation conventions. This requirement has become much easier to meet in recent years thanks to the Universal Dependencies project and the related standardized annotation schemes. Second, all languages of the experiment need to share the same script, such that the taggers are able to learn generalizations about subword features (e.g., suffixes) that cover multiple languages. This is not naturally the case in the study cited earlier: Polish and Slovak are written in Latin script, but Rusyn, Ukrainian, and Russian use Cyrillic script. We decided to transliterate the Polish and Slovak data to Cyrillic script, so that no modification would be required for the target language Rusyn.

In simple model transfer approaches, the LRL and HRL words are assumed to be close enough that a sufficient number of HRL features learned by the tagger would fire for the relevant LRL words. By choosing character-level features, such as word prefixes or suffixes, this assumption can be satisfied in many cases. However, it is possible to define the HRL features in such a way that they become more language-independent (this case is called *delexicalization*), or to explicitly convert them to LRL features (*relexicalization*). These two approaches will be presented in the following sections.

7.2.3 Delexicalization and Cross-lingual Word Embeddings

Delexicalization was first proposed for dependency parsing (Zeman and Resnik, 2008; McDonald et al., 2011). Common dependency parsers use features based on word forms and part-of-speech tags to predict the dependency relations, and delexicalization simply removes the word form features in order to obtain a language-independent parser, the underlying idea being that the part-of-speech tags – which still need to be predicted in a language-dependent way – provide sufficient information for parsing.

For tagging itself, only the word forms (and parts thereof) are traditionally used as features, and there is no obvious choice of alternative, language-independent features. Yu et al. (2016) have attempted to use several statistics derived from the word forms as language-independent features, such as word length, word frequency, or entropy of the previous and following words, but without much success.

Another delexicalization approach for tagging has been pioneered by Täckström et al. (2012). They cluster all words of the HRL according to their distributional properties and replace the word forms by their cluster numbers in the training data. For example, words like *very*, *mildly*, *terribly* would be clustered together in cluster 101, and the tagger would then learn that cluster 101 words should be tagged as adverbs. Using a word-aligned parallel corpus, the cluster numbers are transferred to the LRL corpus, which can then be annotated using the same tagger. For similar language varieties, the association of HRL and LRL clusters could be inferred in a different way than by using a parallel corpus. Word clusters have been used in a similar way by Owoputi et al. (2013) to reduce the out-of-vocabulary rates in informal online conversations, but not in a completely delexicalized manner.

In recent years, vector-based representations of the distributional properties of words (so-called word embeddings) have been used with success in various NLP tasks. The approach of Täckström et al. (2012) can be adapted easily by replacing the cluster numbers with word embedding vectors. In order for this approach to work in a cross-lingual way, the word embeddings of the HRL and LRL need to be in the same multidimensional space. This is generally not the case, as word embeddings are typically inferred separately for each language, using large unannotated corpora such as Wikipedia dumps or web crawls (see Chapter 6). There is a large (and still growing) amount of research on how to project word embeddings of several languages into a common space (for overviews, see

Upadhyay et al., 2016; Ruder et al., 2019). A typical approach is to use a seed dictionary to find a linear transformation that projects the vectors of one language into the space of the other language such that the words end up as close as possible to their translational equivalents defined in the dictionary. An application to POS-tagging has been presented in Zhang et al. (2016).

A bottleneck of this approach is that the embeddings have to be trained on large amounts of raw data, typically containing millions of words. For low-resource varieties such as dialects or similar varieties, this requirement is unrealistic. For example, Magistry et al. (2019) train cross-lingual word embeddings for three French regional languages – Alsatian, Occitan, and Picard – in view of using them for part-of-speech tagging. They show that embeddings trained on corpora containing one to two million words are not able to yield taggers that are competitive with much simpler adaptation strategies.

The current trend in cross-lingual embeddings research seems to be to relax the seed dictionary requirement as much as possible, to the point that no attention is paid to the existence of cognates and otherwise identical words (e.g., Artetxe et al., 2017), which however occur massively in similar languages. One notable exception is Sharoff (2018), who infers Levenshtein distance weights from the seed dictionary of related languages and uses this weighted Levenshtein distance to optimize the mapping between monolingual embeddings.

Another issue that is often overlooked concerns the fact that the projection of multilingual word embeddings into a single space works reliably only if the corpora used to train them are comparable in terms of domains and genres (Søgaard et al., 2018). This requirement can be difficult to satisfy in the configurations of interest here, in particular with dialects, whose domains of usage are inherently different from the ones of the respective roofing languages.

7.2.4 Relexicalization

A more traditional approach to model transfer assumes that the features of the HRL tagger are actual word forms or subword features (as opposed to numerical representations like embeddings) and can be modified after training. For example, the HRL word forms in a trained tagging model can be replaced by LRL equivalents extracted from a bilingual dictionary, yielding a LRL tagger. This approach works well with traditional tagger architectures such as HMMs, where emission and transition

probabilities can be modified separately, according to the linguistic requirements.

The pioneering work in this domain is Feldman et al. (2006), who replace the emission probabilities of a HRL tagger by a lexicon generated independently from a morphological generator of the LRL. In related work (Scherrer, 2014; Scherrer and Sagot, 2014), we adapted the relexicalization approach to similar language varieties by building bilingual dictionaries in an unsupervised way, taking advantage of the prevalence of cognates, similar word forms and phrases. We then “translated” the HRL emission parameters using this dictionary. With this approach, we obtained error reduction rates between 36.2 and 73.7 percent for seven West Germanic, West Slavic, and Ibero-Romance languages. Whether it is easier to induce a bilingual dictionary or to create or adapt a morphological generator for the target language depends on the actual linguistic configuration.

The transition probabilities are often considered to be stable across related languages, but Babych and Sharoff (2016) propose a method to rewrite these parameters as well, which can be useful if the tag sets are not identical across languages. While the delexicalization method with cross-lingual word embeddings tends to work better under laboratory conditions, the relexicalization method can still be competitive thanks to its lower and easier-to-satisfy data requirements.

7.3 Normalization and Other Data Transfer Methods

Tagging text of a low-resource variety does not necessarily require building a tagger for that variety. Instead, one could translate the LRL text to the HRL and tag it with an off-the-shelf HRL tagger. This strategy has encountered considerable success in the area of historical language processing. In these cases, taggers are readily available for the standardized modern variety (the HRL), but not for the historical varieties for which documents need to be annotated (the LRLs). Thus, the chosen solution has been the following:

1. Train a tagger for the modern variety.
2. Create a *normalization* (or *modernization*) system that converts historical text into modern text.

3. Convert the historical data set to the modern variety and annotate it with the modern tagger.
4. Revert the modernized word forms to the original ones, keeping the annotations.

In this scenario, normalization has two crucial benefits: (1) it eliminates the requirement for annotated training data of the low-resource variety by enabling the use of a high-resource variety tagger, and (2) it reduces the amount of spelling variation in the low-resource variety to a level that is manageable by standard taggers, diminishing the out-of-vocabulary rates.

Historical varieties, just like dialectal varieties, exhibit large amounts of spelling variation. In both situations, it makes sense to forgo creating taggers specifically designed to deal with such variation and rather reduce the variation by means of some independent normalization method.

7.3.1 Normalization Methods

Text normalization – whether applied to historical varieties, dialectal varieties or user-generated content on social media – can be seen as a typical instance of the noisy channel model (see Koehn, 2010): the observed spellings are “corrupted” variants of the standardized spellings, which have to be reconstructed. Normalization methods therefore typically rely on a channel model and a language model. The channel model defines what types of spelling changes are allowed, and the language model specifies what strings are correct standardized spellings.

One instance of such a model has been presented by Pettersson et al. (2013a). Their aim is to approximately match historical words in a modern-language lexicon, measuring similarity with Levenshtein distance. In this case, Levenshtein distance figures as the channel model, and the lexicon is used as the language model. This method is completely unsupervised in the sense that it does not require any parallel data (word pairs) and uses a language-independent similarity measure. If parallel data are available, these can be used to adapt the weights of a weighted edit distance measure, as in Mann and Yarowsky (2001) for closely related languages, Scherrer (2007) for dialects, or Pettersson et al. (2013a) for historical varieties. Reffle (2011) and Etxeberria et al. (2019) propose to learn weighted finite-state transducers instead of simpler edit-distance-based models.

Bollmann (2012) presents a method for automatically extracting context-

sensitive rewrite rules from a parallel corpus. These rules are then used as the channel model, either on their own or in conjunction with a language model consisting of a reference lexicon of the standardized variety. A combination of approaches is used in the VARD 2 normalization system (Rayson et al., 2007; Baron and Rayson, 2008), which views text normalization as a particular case of spellchecking. It relies on a lexicon of known variants, a set of letter replacement rules, phonetic matching, and edit distance.

Statistical machine translation systems also follow the noisy channel metaphor (Koehn, 2010). In particular, Vilar et al. (2007) and Tiedemann (2009) proposed character-level statistical machine translation (CSMT) as an alternative approach to translating words between closely related languages. Character-level SMT is different from standard (i.e., word-level) SMT in that, instead of translating words occurring in sentence pairs, it translates characters occurring in word pairs. CSMT has been successfully applied to translation between closely related languages (Tiedemann, 2009), standardization of user-generated content (De Clercq et al., 2013; Ljubešić et al., 2014; Ljubešić et al., 2016) and normalization of historical words (Pettersson et al., 2013b; Sánchez-Martínez et al., 2013; Pettersson et al., 2014; Scherrer and Erjavec, 2016).

Scherrer and Erjavec (2016) present a bootstrapping approach that allows CSMT to be used without parallel training data: they generate noisy parallel data with the unsupervised Levenshtein model of Pettersson et al. (2013a) and train the CSMT model on these noisy data. CSMT successfully regularizes the model, leading to better normalization than with the Levenshtein-based model alone.

All of these methods normalize each word independently without looking at the sentential context. This setup has been viewed as sufficient, on one hand, because the varieties in question generally do not exhibit much syntactic variation and, on the other hand, because the main goal of normalization is not to produce fluent modern-language text, but rather to provide glossing intended to improve corpus search, linguistic analysis, and, last but not least, automatic annotation such as morphosyntactic tagging (Gotscharek et al., 2009; Samardzić et al., 2016).

Nevertheless, context can be useful for disambiguating ambiguous spellings and to account for phonological effects at word boundaries that affect spelling. Ljubešić et al. (2016) and Scherrer and Ljubešić (2016) extend the scope of the CSMT model to entire segments (utterances or

sentences) and obtain improvements for historical, dialectal, and social media data, depending on the level of ambiguity present in the data.⁴

As the machine translation community moves from statistical to neural models, neural-network based normalization models have also started to appear. Bollmann and Søgaard (2016) are the first to our knowledge to use neural networks for historical spelling normalization. They train a bidirectional LSTM on character-aligned word pairs; alignment is performed using iterated Levenshtein alignment, but their method outperforms previous work (Bollmann, 2012) only in a multi-task setting where training data with different spelling characteristics are included. Korchagina (2017) uses a character-level neural machine translation framework that obviates the need for preliminary character alignment. Tang et al. (2018) compare various neural machine translation architectures and basic translation units (single characters or pre-segmented character sequences) on the task of word-by-word spelling normalization in five languages; neural models outperform CSMT models in two out of five languages, but no clear winning parameter setting was found. Lusetti et al. (2018) present a neural normalization method that includes external language models and find improvements over comparable CSMT models.

At the moment – and unlike in general machine translation – neural normalization models do not clearly outperform statistical ones (Bollmann, 2019). I see two main reasons for this. First, normalization models are generally characterized by rather simple channel models and heavy reliance on language models, but neural translation models typically provide more powerful channel models at the expense of an explicit language model. Second, the generally small training data sizes do not favor the more data-hungry neural models either.

It should be noted that normalization as a task may encounter resistance for political or sociolinguistic reasons, in particular in the context of minority languages. Typically, some language varieties have been involved in long identitary struggles to emphasize their (cultural and linguistic) difference from a majority language, and the last thing that the main actors of these struggles want to see is to have it reduced to that

⁴ Sentence-level normalization systems have an important drawback, namely that it is difficult to retrieve the word boundaries, especially if the inputs and outputs have different numbers of tokens. To this end, the work cited here uses a specifically designed re-alignment script that is based on the phrase alignment information output by the CSMT system and a set of heuristics. This script is included in a larger set of scripts designed to simplify the use of CSMT systems: <https://github.com/clarinsi/csmtiser>.

majority language by some normalization procedure, even though that might be the most adapted approach from a technological point of view. The relation between contemporary language varieties is often more ideologically charged than the relation between different diachronic states of a language.

7.3.2 Normalization as a Preprocessing Step for Tagging

The usefulness of normalization for tagging has been demonstrated in various case studies. For example, Scherrer and Erjavec (2016) applied a modern Slovene tagger to unnormalized or normalized historical Slovene and showed that normalization indeed improved tagging accuracy by up to 40 percent absolute, depending on the normalization setting and the time period. However, a supervised tagger directly trained on manually annotated historical data always outperformed the normalization-based taggers. Obviously, in many situations, training a supervised tagger is not a viable option due to the lack of annotated training data.

The CLIN27 shared task on historical text modernization (Tjong Kim Sang et al., 2017) also included an evaluation of part-of-speech tagging: the same modern Dutch tagger was applied to the unnormalized text, the texts normalized with one of the submitted systems, and the manually normalized text. While the unnormalized data yielded a tagging accuracy of 70.9 percent and an OOV rate of 40.7 percent, the best submitted normalization system obtained an accuracy of 85.6 percent, reducing the OOV rate to 14.8 percent. For comparison, the manually normalized data resulted in an accuracy of 87.5 percent and an OOV rate of 9.3 percent. The detailed results also show that normalization performance and tagging performance are not necessarily correlated: for example, the Bochum system was ranked first on the normalization task (on par with two other systems), but performed much worse on the tagging task. The OOV rates suggest that the Bochum system did a good job on normalizing rare words, but was less successful with frequent words.

If the goal of normalization lies merely in tagging, it can be sufficient to focus on normalizing frequent words and let the unknown word module of the HRL tagger deal with the rare words. In their case study, Bernhard and Ligozat (2013) focus on tagging Alsatian, a German variety spoken in the Alsace region of France, using a tagger trained on Standard German. They manually create a normalization dictionary that relates the 107 most frequent Alsatian word types to their Standard

German equivalents, leaving all other Alsatian words untouched. This minimalistic normalization improves tagging accuracy by 11 to 33 percent absolute, depending on the text genre and tagging algorithm used.

A study by Bollmann (2013) further illustrates the impact of normalization quality on tagging. He trained four normalization models using 100, 250, 500, and 1000 word pairs respectively, and created four normalized variants of the test corpus, which were then annotated with a tagger of the normalized variety. While each step yielded additional improvements in tagging accuracy, the biggest increases were seen with the first 100 or 250 word pairs.

In some cases, a normalization approach may be in competition with a supervised tagging approach. Samardžić et al. (2016) discuss two scenarios for tagging transcriptions of Swiss German dialects (cf. also Chapter 6). The supervised option relies on a collection of part-of-speech annotated Swiss German texts from various written sources (Hollenstein and Aepli, 2014). This corpus is rather small (73,000 tokens) and differs in genre (written news data and literature versus transcribed speech), but contains the same dialectal varieties as the target texts. The normalization approach uses a Standard German tagger trained on TüBa-D/S, an annotated corpus of spontaneous dialogues conducted in Standard German (360,000 tokens). In this case, the text genre matches, but the normalization of the dialectal data to Standard German may introduce additional errors. The first option slightly outperformed the second one, indicating that using less training data from a non-matching genre in the same variety can give better results than using more data from a matching genre in a different variety together with normalization. However, the test data normalization approach is sufficiently competitive to provide a viable approach if training data of the same variety are too scarce.

Recent experiments on the same data set (Scherrer et al., 2019) show that a combination of both approaches may be most promising: adding the normalized forms as additional features to the supervised tagger (in the training set as well as in the test set) increased tagging accuracy by up to 10 percent absolute. In turn, the gains obtained with sentence-level normalization suggests that better knowledge of the syntactic structure, e.g., in the form of POS-tags, may be helpful for normalization. An interesting strand for future research will thus be to jointly learn normalization and tagging. This would, however, require at least minimal amounts of training data for both tasks.

7.3.3 Annotation Projection

A different variant of data transfer techniques is called *annotation projection*, and has been introduced by Yarowsky and Ngai (2001). It requires two resources: a tagger of the source language (or an annotated corpus from which it can be trained) and a parallel corpus relating the source with the target language. The source side of the parallel corpus is annotated with the existing tagger, and the labels are projected to the target side using word alignment links produced by an alignment tool. The resulting annotated target corpus can then serve as training data for a target tagger.

On the surface, this approach looks rather simple, but complications arise through word alignment errors, as well as many-to-one and one-to-many alignments. Yarowsky and Ngai (2001) introduce simple heuristics to reduce such noise, and follow-up work by Das and Petrov (2011) and Duong et al. (2013) provides interesting extensions. However, none of these works specifically address configurations with similar languages.

A multilingual extension has been introduced by Agić et al. (2015). They argue that the reliance on a single source language, in most cases English, is not optimal, and that multiple source languages can be used for projection by means of massively multilingual parallel corpora, such as the Bible. For a given target language, they project annotations from all available source corpora and use simple majority voting to resolve ambiguities. They also provide an optional second retraining step, in which low-resource taggers (such as those obtained at the end of the majority voting step) are included to guide the annotations of another low-resource language. This second step could be helpful in scenarios where several low-resource languages are related among each other, but no related high-resource language is available.

Annotation projection is not a particularly popular choice for configurations involving closely related varieties, mainly because large parallel corpora can be hard to find: for instance, parallel corpora involving a dialect and its corresponding standard variety are not naturally available, as dialects do not have official status and speakers are generally fluent in both varieties, obviating the need for translations.

One exception is the work of Aepli et al. (2014), who use a similar approach as Agić et al. (2015) to create a full morphological tagger for Macedonian using majority voting from the related languages Bulgarian, Czech, Slovene, and Serbian as well as from English. Their full tagging setup is somewhat more complicated and allows different morphological

features to be inferred from different sets of source languages. The voting process is also guided by partial tagging information for Macedonian. The authors obtain an overall tagging accuracy of 88 percent, which is close to state-of-the-art performance for full tagging of morphologically rich languages (for contemporary results, see, e.g., Müller et al., 2013).

7.3.4 Training Data Translation

Tiedemann et al. (2014) introduce yet another variant of data transfer that relies on the same resources as annotation projection, namely an annotated corpus of the source language and a parallel corpus relating the source with the target language. Instead of annotating the parallel corpus, they use the parallel corpus to train a machine translation system (from source to target language) and use that system to translate the annotated source corpus to the target language, keeping the annotations intact.

Tiedemann et al. (2014) mention two inherent advantages of training data translation compared to annotation projection. First, instead of using an external word alignment tool that can add errors, the translation process and word alignment are intrinsically linked, at least in statistical machine translation models. As a result, training data translation contains one step less, and thus fewer chances of accumulating errors. Second, (statistical) machine translation tends to produce more literal translations than the human translations in parallel texts usually are, which reduces the amount of misalignments. The choice between annotation projection and training data translation may also be guided by the target domain: if the annotated source corpus matches the target domain better, then training data translation should be preferred, and if the parallel corpus is closer to the target domain, annotation projection may be the better choice.

In any case, this approach suffers from the same data availability constraints as annotation projection and is thus of limited use for the closely related language scenarios. Moreover, it requires translation from the HRL to the LRL. This translation direction is unintuitive for a scenario where the former is standardized and the latter is not: it corresponds to “un-normalization” and hence to the introduction of noise during the translation process.

Nevertheless, Boujelbane et al. (2014) provide an example of this approach: they translate a Modern Standard Arabic (MSA) part-of-speech annotated corpus to Tunisian dialect (TD) and train a tagger on the

resulting TD data. The translation model is not trained on a large parallel corpus, but mostly relies on a dictionary and a set of hand-written morphosyntactic translation rules. They improve tagging accuracy by 15 percent absolute compared to applying an MSA tagger directly to TD.

7.4 Tagger Adaptation and Multilingual Models

The approaches presented in this section make the assumption that no training data annotated with part-of-speech (or morphological) tags is available for the target variety, but that other types of data (raw text for training word embeddings, parallel corpora for training normalization systems, bilingual dictionaries, etc.) can be used instead. However, in many cases, at least small amounts of annotated data of the target variety can be made available. In this section, we review some approaches that take advantage of these types of data.

7.4.1 From Domain Adaptation to Language Adaptation

Mismatches in text genre or domain have caused problems in NLP research for a long time: the situation is common where the annotated training data stems from a particular domain (most often newswire) but the trained model is supposed to be applied with high accuracy to data of a different domain (e.g., tweets, biomedical documents, or speech transcripts). Algorithms for resolving such mismatches have been researched in the area called *domain adaptation*. The typical domain adaptation scenario involves a large annotated “out-of-domain” training corpus and a small annotated “in-domain” corpus, and the goal of the learning algorithm is to learn general knowledge about the language from the out-of-domain data and complement it with domain-specific knowledge from the in-domain data (e.g., Daumé III, 2007). Domain adaptation can also be done in a semi-supervised fashion, where the in-domain corpus is not annotated (e.g., Kübler and Baucom, 2011). The domain adaptation problem can be reformulated as a language adaptation problem if the involved languages are very similar.

For example, Jørgensen et al. (2016) create a tagger for African American Vernacular English (AAVE) tweets using domain adaptation strategies. They start by training a tagger on a large, non-AAVE-specific Twitter corpus and add domain-specific information like word embeddings, unlabeled in-domain data, and tag dictionaries extracted from the web.

They also make use of a strategy called *self-training*: self-training consists in annotating unlabeled data using an imperfect tagging method, correcting the annotations in some way, and adding the corrected data to the training set of a second tagger, which will hopefully perform better than the initial imperfect one. A variety of methods can be used for the correction step, e.g., majority voting from different taggers (Kübler and Baucom, 2011; Aepli et al., 2014) or annotation projection (Duong et al., 2013).

Another option that is frequently employed in the context of neural network models is *curriculum training* or *fine-tuning*. Here, a model is trained in two phases, first on out-of-domain data and second on in-domain data. The motivation of this approach is that the small in-domain data set should not be “wasted” for learning the generic properties of the language in question. This approach could well be applied to tagging of low-resource varieties, but we are not aware of any such work.

7.4.2 Multilingual and Multilectal Models

We have already alluded to the possibility of using multiple source languages for tagging a different target language (Section 7.2.2). This idea can be combined with domain adaptation if small amounts of annotated target language data are added into the multilingual model, as suggested by Cotterell and Heigold (2017).

Multilingual – or multilectal – models can be trained in a completely language-agnostic way, i.e., by making the model believe that all training instances stem from the same language. This may work reasonably well if the involved languages are indeed very similar, but multilectal models generally work better if they are given information about the language variety of each training instance. When using neural network architectures, there are essentially two options for providing language information. The first option is to add a *language label* at the beginning (and/or at the end) of each sentence, as if it were an additional word (a). This information will be passed along successive states of the neural network and can influence the tagging decisions. The second option is a multi-task learning scenario, in which the model is trained to predict the language labels added on the output side (b). In this scenario, one may think of the language label as an additional morphological feature *Lang* that is appended to the tag of each word of the training data. By forcing the model to predict the language variety of each word, it can learn

some generalizations about the interactions of morphosyntactic tags and language varieties.

(a)	$\langle \text{EN} \rangle$	I	am	a	student	$\langle \text{EN} \rangle$
	NA	PRON	VERB	DET	NOUN	NA
(b)	I	am	a	student		
	PRON Lang=EN	VERB Lang=EN	DET Lang=EN	NOUN Lang=EN		

The crucial difference between these two options is their behavior at test time. With option (a), the language labels need to be specified in the input. With option (b), the input does not need to be modified, as the language labels are generated by the model. This option can therefore be used when the test data cannot be unambiguously assigned a language label.

With both options, the neural network tagger will build internal, vector-based representations of the language labels. By mapping these representations to a low-dimensional space, one can visualize the “landscape” of language varieties as it has been inferred by the tagger. Östling and Tiedemann (2017) present an example of such work, using language modelling instead of tagging as the underlying task.

Multilectal models are particularly well suited for situations where several target language varieties are to be supported. It would be interesting to apply this approach to complex dialect landscapes whose varieties are known to be distinct, but not distinct enough to warrant completely independent models.

7.5 Conclusions

The goal of this chapter was to give a high-level overview of different approaches to tagging similar language varieties. If sufficient manually annotated training data are available, one may obviously attempt to create distinct tagging models for each language variety. On the opposite, one may also attempt to create a single tagging model for all varieties, using all available training data without distinction. However, most linguistic configurations call for some intermediate approach, whenever the variety for which training data are available is not the same as the variety for which a tagger is needed. For such scenarios, a range of cross-lingual learning techniques have been developed, but most of them adopt a broader scope in the sense that they do not assume any etymological relation between the involved language varieties. Also, some of these techniques have been developed with other tasks such as parsing

in mind. We have striven to present the different approaches, giving examples from morphosyntactic tagging tasks and closely related language settings.

Two opposing trends can be seen in recent work. One aims at reducing inter-linguistic variation, e.g., by using normalization systems, such that traditional tagging algorithms can continue to be applied. The other one aims at embracing variation to create taggers that intrinsically handle data with internal variation, as witnessed by the increasing use of character-level features and vector-based word representations. While the most popular approach has without doubt been the one based on normalization, a slow shift towards the second type of models can be perceived, although examples of their application to similar language settings are still scarce. In both approaches however, the internal states of the models (of the normalization model in the former, or of the tagging model in the latter) can give rise to linguistic interpretations about the relationship between the modeled language varieties. Tagging of similar languages can thus be much more than just a particular application of a decade-old sequence labeling task.

References

- Aepli, Noëmi, von Waldenfels, Ruprecht, and Samardžić, Tanja. 2014. Part-of-speech tag disambiguation by cross-linguistic majority vote. Pages 76–84 of: *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*.
- Agić, Željko, Hovy, Dirk, and Søgaard, Anders. 2015. If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. Pages 268–272 of: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*.
- Artetxe, Mikel, Labaka, Gorka, and Agirre, Eneko. 2017. Learning bilingual word embeddings with (almost) no bilingual data. Pages 451–462 of: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Babych, Bogdan, and Sharoff, Serge. 2016. Ukrainian part-of-speech tagger for hybrid MT: Rapid induction of morphological disambiguation resources from a closely related language. Pages 1–12 of: *Proceedings of the Fifth Workshop on Hybrid Approaches to Translation (HyTra), European Association for Machine Translation (EAMT)*.
- Baddeley, Susan, and Voeste, Anja (eds). 2012. *Orthographies in Early Modern Europe*. Berlin, Boston: De Gruyter Mouton.
- Baron, Alistair, and Rayson, Paul. 2008. VARD 2: A tool for dealing with

- spelling variation in historical corpora. Pages 1–15 of: *Proceedings of the Postgraduate Conference in Corpus Linguistics*. Birmingham, UK: Aston University.
- Bernhard, Delphine, and Ligozat, Anne-Laure. 2013. Hassle-free POS-tagging for the Alsatian dialects. Pages 85–92 of: Zampieri, Marcos, and Diwersy, Sascha (eds), *Non-Standard Data Sources in Corpus Based-Research*. ZSM Studien. Shaker.
- Bollmann, Marcel. 2012. (Semi-)automatic normalization of historical texts using distance measures and the Norma tool. Pages 3–14 of: *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*.
- Bollmann, Marcel. 2013. POS tagging for historical texts with sparse training data. Pages 11–18 of: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*.
- Bollmann, Marcel. 2019. A large-scale comparison of historical text normalization systems. Pages 3885–3898 of: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Bollmann, Marcel, and Søgaard, Anders. 2016. Improving historical spelling normalization with bi-directional LSTMs and multi-task learning. Pages 131–139 of: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*.
- Boujelbane, Rahma, Mallek, Mariem, Ellouze, Mariem, and Belguith, Lamia Hadrich. 2014. Fine-grained POS tagging of spoken Tunisian dialect corpora. Pages 59–62 of: Métais, Elisabeth, Roche, Mathieu, and Teisseire, Maguelonne (eds), *Natural Language Processing and Information Systems*. Basel, Switzerland: Springer.
- Brants, Thorsten. 2000. TnT – A statistical part-of-speech tagger. Pages 224–231 of: *Proceedings of the Sixth Conference on Applied Natural Language Processing*.
- Clyne, Michael G. (ed). 1992. *Pluricentric Languages: Differing Norms in Different Nations*. Contributions to the Sociology of Language, no. 62. Berlin, New York: Mouton de Gruyter.
- Collobert, Ronan, Weston, Jason, Bottou, Léon, Karlen, Michael, Kavukcuoglu, Koray, and Kuksa, Pavel. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, **12**, 2493–2537.
- Cotterell, Ryan, and Heigold, Georg. 2017. Cross-lingual character-level neural morphological tagging. Pages 748–759 of: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.
- Das, Dipanjan, and Petrov, Slav. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. Pages 600–609 of: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.
- Daumé III, Hal. 2007. Frustratingly easy domain adaptation. Pages 256–263 of: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.

- De Clercq, Orphée, Desmet, Bart, Schulz, Sarah, Lefever, Els, and Hoste, Véronique. 2013. Normalization of Dutch user-generated content. Pages 179–188 of: *Proceedings of RANLP 2013*.
- Duong, Long, Cook, Paul, Bird, Steven, and Pecina, Pavel. 2013. Simpler unsupervised POS tagging with bilingual projections. Pages 634–639 of: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.
- Erjavec, Tomaž. 2010. MULTEXT-East version 4: Multilingual morphosyntactic specifications, lexicons and corpora. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA).
- Etxeberria, Izaskun, Alegria, Iñaki, and Uria, Larraitz. 2019. Weighted finite-state transducers for normalization of historical texts. *Natural Language Engineering*, **25**(2), 307–321.
- Feldman, Anna, Hana, Jirka, and Brew, Chris. 2006. A cross-language approach to rapid creation of new morphosyntactically annotated resources. Pages 549–554 of: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*. European Language Resources Association (ELRA).
- Ferguson, Charles A. 1959. Diglossia. *Word*, **15**(2), 325–340.
- Francom, Jerid, Hulten, Mans, and Ussishkin, Adam. 2014. ACTIV-ES: A comparable, cross-dialect corpus of “everyday” Spanish from Argentina, Mexico, and Spain. Pages 1733–1737 of: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA).
- Gesmundo, Andrea. 2011. Bidirectional sequence classification for tagging tasks with guided learning. In: *Actes de la 18e confrence sur le Traitement Automatique des Langues Naturelles*.
- Giménez, Jesús, and Màrquez, Lluís. 2004. SVMTool: A general POS tagger generator based on support vector machines. Pages 43–46 of: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*. Lisbon, Portugal: European Language Resources Association (ELRA).
- Gotscharek, Annette, Neumann, Andreas, Reffle, Ulrich, Ringlstetter, Christoph, and Schulz, Klaus U. 2009. Enabling information retrieval on historical document collections: The role of matching procedures and special lexica. Pages 69–76 of: *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data (AND'09)*.
- Hajič, Jan. 2000. Morphological tagging: Data vs. dictionaries. In: *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference (NAACL 2000)*.
- Halácsy, Péter, Kornai, András, and Oravecz, Csaba. 2007. HunPos – An open source trigram tagger. Pages 209–212 of: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Companion Volume: Demo and Poster Sessions*.

- Hollenstein, Nora, and Aepli, Noëmi. 2014. Compilation of a Swiss German dialect corpus and its application to PoS tagging. In: *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*.
- Horsmann, Tobias, and Zesch, Torsten. 2017. Do LSTMs really work so well for PoS tagging? – A replication study. Pages 727–736 of: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.
- Jørgensen, Anna, Hovy, Dirk, and Søgaard, Anders. 2016. Learning a POS tagger for AAVE-like language. Pages 1115–1120 of: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Jurafsky, Daniel, and Martin, James H. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. 2nd edn. Upper Saddle River, NJ, USA: Prentice-Hall.
- Koehn, Philipp. 2010. *Statistical Machine Translation*. Cambridge, UK: Cambridge University Press.
- Korchagina, Natalia. 2017. Normalizing medieval German texts: From rules to deep learning. Pages 12–17 of: *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*.
- Kübler, Sandra, and Baucom, Eric. 2011. Fast domain adaptation for part of speech tagging for dialogues. Pages 41–48 of: *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*.
- Lample, Guillaume, Ballesteros, Miguel, Subramanian, Sandeep, Kawakami, Kazuya, and Dyer, Chris. 2016. Neural architectures for named entity recognition. Pages 260–270 of: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, CA: Association for Computational Linguistics.
- Ljubešić, Nikola, Erjavec, Tomaž, and Fišer, Darja. 2014. Standardizing tweets with character-level machine translation. Pages 164–175 of: *Proceedings of CICLing 2014*. Lecture Notes in Computer Science. Kathmandu, Nepal: Springer.
- Ljubešić, Nikola, Zupan, Katja, Fišer, Darja, and Erjavec, Tomaž. 2016. Normalising Slovene data: Historical texts vs. user-generated content. Pages 146–155 of: *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*.
- Lusetti, Massimo, Ruzsics, Tatyana, Göhring, Anne, Samardžić, Tanja, and Stark, Elisabeth. 2018. Encoder-decoder methods for text normalization. Pages 18–28 of: *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*. Association for Computational Linguistics.
- Magistry, Pierre, Ligozat, Anne-Laure, and Rosset, Sophie. 2019. Exploiting languages proximity for part-of-speech tagging of three French regional languages. *Language Resources and Evaluation*, **53**, 865–888.

- Mann, Gideon S., and Yarowsky, David. 2001. Multipath translation lexicon induction via bridge languages. Pages 151–158 of: *Proceedings of NAACL 2001*.
- Manning, Christopher D. 2011. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? Pages 171–189 of: Gelbukh, Alexander (ed), *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2011)*. Lecture Notes in Computer Science, no. 6608. Berlin/Heidelberg, Germany: Springer.
- Marcus, Mitchell P., Santorini, Beatrice, and Marcinkiewicz, Mary Ann. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 2(19), 313–330.
- McDonald, Ryan, Petrov, Slav, and Hall, Keith. 2011. Multi-source transfer of delexicalized dependency parsers. Pages 62–72 of: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.
- Müller, Thomas, Schmid, Helmut, and Schütze, Hinrich. 2013. Efficient higher-order CRFs for morphological tagging. Pages 322–332 of: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.
- Nivre, Joakim, de Marneffe, Marie-Catherine, Ginter, Filip, Goldberg, Yoav, Hajič, Jan, Manning, Christopher D., McDonald, Ryan, Petrov, Slav, Pyysalo, Sampo, Silveira, Natalia, Tsarfaty, Reut, and Zeman, Daniel. 2016. Universal Dependencies v1: A multilingual treebank collection. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia: European Language Resources Association (ELRA).
- Östling, Robert, and Tiedemann, Jörg. 2017. Continuous multilinguality with language vectors. Pages 644–649 of: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*.
- Owoputi, Olutobi, O'Connor, Brendan, Dyer, Chris, Gimpel, Kevin, Schneider, Nathan, and Smith, Noah A. 2013. Improved part-of-speech tagging for online conversational text with word clusters. Pages 380–390 of: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Petrov, Slav, Das, Dipanjan, and McDonald, Ryan. 2012. A universal part-of-speech tagset. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).
- Pettersson, Eva, Megyesi, Beáta B., and Nivre, Joakim. 2013a. Normalisation of historical text using context-sensitive weighted Levenshtein distance and compound splitting. Pages 163–179 of: *Proceedings of the 19th Nordic Conference of Computational Linguistics (Nodalida 2013)*.
- Pettersson, Eva, Megyesi, Beáta B., and Tiedemann, Jörg. 2013b. An SMT approach to automatic annotation of historical text. Pages 54–69 of: *Proceedings of the Nodalida Workshop on Computational Historical Linguistics*.

- Pettersson, Eva, Megyesi, Beáta B., and Nivre, Joakim. 2014. A multilingual evaluation of three spelling normalisation methods for historical text. Pages 32–41 of: *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*.
- Pinter, Yuval, Guthrie, Robert, and Eisenstein, Jacob. 2017. Mimicking word embeddings using subword RNNs. Pages 102–112 of: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.
- Plank, Barbara, Søgaard, Anders, and Goldberg, Yoav. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. Pages 412–418 of: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.
- Ratnaparkhi, Adwait. 1996. A maximum entropy model for part-of-speech tagging. Pages 133–142 of: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- Rayson, Paul, Archer, Dawn, Baron, Alistair, and Smith, Nick. 2007. Tagging historical corpora – The problem of spelling variation. In: *Proceedings of Digital Historical Corpora, Dagstuhl-Seminar 06491*. Wadern, Germany: International Conference and Research Center for Computer Science, Schloss Dagstuhl.
- Reffle, Ulrich. 2011. Efficiently generating correction suggestions for garbled tokens of historical language. *Natural Language Engineering*, **17**, 265–282.
- Rehm, Georg, and Uszkoreit, Hans (eds). 2013. *Strategic Research Agenda for Multilingual Europe 2020*. Berlin/Heidelberg, Germany: Springer.
- Ruder, Sebastian. 2017. An overview of multi-task learning in deep neural networks. *arXiv e-prints*, arXiv:1706.05098.
- Ruder, Sebastian, Vulić, Ivan, and Søgaard, Anders. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, **65**, 569–631.
- Samardžić, Tanja, Scherrer, Yves, and Glaser, Elvira. 2016. ArchiMob – A corpus of spoken Swiss German. Pages 4061–4066 of: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).
- Sánchez-Martínez, Felipe, Martínez-Sempere, Isabel, Ivars-Ribes, Xavier, and Carrasco, Rafael C. 2013. *An open diachronic corpus of historical Spanish: annotation criteria and automatic modernisation of spelling*. Research Report. Departament de Llenguatges i Sistemes Informàtics, Universitat d’Alacant, Alicante.
- Scherrer, Yves. 2007. Adaptive string distance measures for bilingual dialect lexicon induction. Pages 55–60 of: *Proceedings of the ACL 2007 Student Research Workshop*.
- Scherrer, Yves. 2008. Part-of-speech tagging with a symbolic full parser: Using the TIGER treebank to evaluate Fips. Pages 16–23 of: *Proceedings of the Workshop on Parsing German*.
- Scherrer, Yves. 2014. Unsupervised adaptation of supervised part-of-speech taggers for closely related languages. Pages 30–38 of: *Proceedings of the*

First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects.

- Scherrer, Yves, and Erjavec, Tomaž. 2016. Modernising historical Slovene words. *Natural Language Engineering*, **6**(22), 881–905.
- Scherrer, Yves, and Ljubešić, Nikola. 2016. Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation. Pages 248–255 of: *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*.
- Scherrer, Yves, and Rabus, Achim. 2017. Multi-source morphosyntactic tagging for spoken Rusyn. Pages 84–92 of: *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.
- Scherrer, Yves, and Rabus, Achim. 2019. Neural morphosyntactic tagging for Rusyn. *Natural Language Engineering*, **25**(5), 633–650.
- Scherrer, Yves, and Sagot, Benoit. 2014. A language-independent and fully unsupervised approach to lexicon induction and part-of-speech tagging for closely related languages. Pages 502–508 of: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA).
- Scherrer, Yves, Mocken, Susanne, and Rabus, Achim. 2018. New developments in tagging pre-modern orthodox Slavic texts. *Scripta & e-Scripta*, **18**, 9–34.
- Scherrer, Yves, Samardžić, Tanja, and Glaser, Elvira. 2019. Digitising Swiss German: How to process and study a polycentric spoken language. *Language Resources and Evaluation*, **53**, 735–769.
- Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In: *Proceedings of International Conference on New Methods in Language Processing*.
- Sharoff, Serge. 2018. Language adaptation experiments via cross-lingual embeddings for related languages. Pages 844–849 of: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).
- Silfverberg, Miikka, and Drobac, Senka. 2018. Sub-label dependencies for neural morphological tagging – The joint submission of University of Colorado and University of Helsinki for VarDial 2018. Pages 37–45 of: *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*.
- Søgaard, Anders, Ruder, Sebastian, and Vulić, Ivan. 2018. On the limitations of unsupervised bilingual dictionary induction. Pages 778–788 of: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Täckström, Oscar, McDonald, Ryan, and Uszkoreit, Jakob. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. Pages 477–487 of: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

- Tang, Gongbo, Cap, Fabienne, Pettersson, Eva, and Nivre, Joakim. 2018. An evaluation of neural machine translation models on historical spelling normalization. Pages 1320–1331 of: *Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics*.
- Tapanainen, Pasi, and Voutilainen, Atro. 1994. Tagging accurately – Don’t guess if you know. Pages 47–52 of: *Proceedings of the Fourth Conference on Applied Natural Language Processing*.
- Tiedemann, Jörg. 2009. Character-based PSMT for closely related languages. Pages 12–19 of: *Proceedings of EAMT 2009*.
- Tiedemann, Jörg, and Agić, Željko. 2016. Synthetic treebanking for cross-lingual dependency parsing. *Journal of Artificial Intelligence Research*, **55**, 209–248.
- Tiedemann, Jörg, Agić, Željko, and Nivre, Joakim. 2014. Treebank translation for cross-lingual parser induction. Pages 130–140 of: *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*.
- Tjong Kim Sang, Erik, Bollmann, Marcel, Boschker, Remko, Casacuberta, Francisco, Dietz, Feike, Dipper, Stefanie, Domingo, Miguel, van der Goot, Rob, van Koppen, Marjo, Ljubešić, Nikola, Östling, Robert, Petran, Florian, Pettersson, Eva, Scherrer, Yves, Schraagen, Marijn, Sevens, Leen, Tiedemann, Jörg, Vanallemeersch, Tom, and Zervanou, Kalliopi. 2017. The CLIN27 shared task: Translating historical text to contemporary language for improving automatic linguistic annotation. *Computational Linguistics in the Netherlands Journal*, **7**, 53–64.
- Toutanova, Kristina, Klein, Dan, Manning, Christopher D., and Singer, Yoram. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. Pages 173–180 of: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL 2003)*.
- Trudgill, Peter. 2000. *Sociolinguistics: An Introduction to Language and Society*. London: Penguin Books.
- Trudgill, Peter. 2004. Globalisation and the Ausbau sociolinguistics of modern Europe. Pages 35–49 of: Duszak, A., and Okulska, U. (eds), *Speaking from the Margin: Global English from a European Perspective*. Frankfurt: Peter Lang.
- Upadhyay, Shyam, Faruqui, Manaal, Dyer, Chris, and Roth, Dan. 2016. Cross-lingual models of word embeddings: An empirical comparison. Pages 1661–1670 of: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Vilar, David, Peter, Jan-Thorsten, and Ney, Hermann. 2007. Can we translate letters? Pages 33–39 of: *Proceedings of the Second Workshop on Statistical Machine Translation*.
- Yang, Jie, and Zhang, Yue. 2018. NCRF++: An open-source neural sequence labeling toolkit. Pages 74–79 of: *Proceedings of ACL 2018, System Demonstrations*.
- Yarowsky, David, and Ngai, Grace. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. Pages 200–207 of: *Proceedings of NAACL 2001*.