Natural Language Engineering (2020), **26**, pp. 595–612
doi:10.1017/S1351324920000492

CAMBRIDGE UNIVERSITY PRESS

SURVEY PAPER

Natural language processing for similar languages, varieties, and dialects: A survey

Marcos Zampieri [1,*] Preslav Nakov [2] and Yves Scherrer [3]

[1]Rochester Institute of Technology, USA, [2]Qatar Computing Research Institute, HBKU, Qatar and
[3]University of Helsinki, Finland
*Corresponding author. E-mail: marcos.zampieri@rit.edu

Abstract

There has been a lot of recent interest in the natural language processing (NLP) community in the computational processing of language varieties and dialects, with the aim to improve the performance of applications such as machine translation, speech recognition, and dialogue systems. Here, we attempt to survey this growing field of research, with focus on computational methods for processing similar languages, varieties, and dialects. In particular, we discuss the most important challenges when dealing with diatopic language variation, and we present some of the available datasets, the process of data collection, and the most common data collection strategies used to compile datasets for similar languages, varieties, and dialects. We further present a number of studies on computational methods developed and/or adapted for preprocessing, normalization, part-of-speech tagging, and parsing similar languages, language varieties, and dialects. Finally, we discuss relevant applications such as language and dialect identification and machine translation for closely related languages, language varieties, and dialects.

Keywords: Dialects; similar languages; language varieties; language identification machine; translation parsing

1. Introduction

Variation is intrinsic to human language and it is manifested in different ways. There are four generally accepted dimensions of language variation, namely *diaphasic*, *diastratic*, *diachronic*, and *diatopic*. *Diaphasic* variation is related to the setting or the medium of communication, for example, different levels of style and register, oral versus written language. *Diastratic* variation is related to language variation in different social groups (e.g., age, gender), whereas *diachronic* is language variation across time. Finally, *diatopic* variation is language variation in space such as different dialects or national varieties of the same languages (e.g., British and American English). All these dimensions of language variation pose challenges for Natural Language Processing (NLP) applications developed to process text and speech. As a result, there has been a growing interest in language variation in the NLP community, as evidenced by a large number of publications and events, for example, conferences, shared tasks, tutorials, and workshops on the topic, some of which we cover in this survey.

One of these initiatives is the series of workshops on NLP for Similar Languages, Language Varieties, and Dialects (VarDial), a workshop series with a special focus on diatopic language variation. VarDial started in 2014, and since then it has become an important venue for work on the study of language variation from a computational perspective, co-located with international NLP conferences such as COLING, EACL, and NAACL. Past editions of the workshop included papers on machine translation (MT) (Shapiro and Duh 2019; Myint Oo, Kyaw Thu, and Mar Soe 2019;

Popović *et al.* 2020), part-of-speech tagging (Huck, Dutka, and Fraser 2019; AlGhamdi and Diab 2019), text normalization (Lusetti *et al.* 2018), and many other relevant topics applied to the computational processing of similar languages, varieties, and dialects. The workshop also featured evaluation campaigns with multiple shared tasks on a number of topics such as cross-lingual morphological analysis, cross-lingual parsing, language and dialect identification, and morphosyntactic tagging (Zampieri *et al.* 2018, 2019; Găman *et al.* 2020). These shared tasks have provided important datasets for the community, and we will present some of them in Section 2.

Below, we will use the terms *varieties* and *language varieties* interchangeably as synonyms for (standard) *national language varieties* of pluricentric languages, that is, languages with multiple interacting standard forms in different countries. Examples of pluricentric languages include English, French, Portuguese, and Spanish; see Clyne (1992) for a discussion on the status of national language varieties and pluricentric languages. Furthermore, in this survey, we do not draw a clear separating line between languages, language varieties, and dialects. This is on purpose, as in many cases this is a political rather than a linguistic distinction. From a computational perspective, problems faced by systems processing, for example, Croatian and Serbian are very similar to those that occur when dealing with Dutch and Flemish, with Brazilian and European Portuguese, or with the various dialects of Arabic.

Our focus below is on the computational processing of diatopic language variation. Section 2 describes the process of data collection and presents some available corpora. Section 3 discusses some part-of-speech tagging and parsing methods that have been successful for processing similar languages, varieties, and dialects. Section 4 focuses on relevant applications such as language and dialect identification, and MT. Finally, Section 5 concludes this survey, and Section 6 presents some avenues for future work.

2. Available corpora and data collection

It is well-known that the performance of NLP systems degrades when faced with language variation and ideally, applications should be trained on data that enables it to model the different dimensions of variation discussed in the introduction of this article: spoken vs. written language, different registers and genres, and regional variation, etc. Thus, it is somewhat simplistic to assume that corpora could fully represent a language without considering variation. In corpus linguistics, researchers have tried to address variation and represent it in corpora. One such example is the Brown corpus for English (Francis and Kucera 1979).

A well-known early attempt to represent diatopic language variation in corpora is the International Corpus of English (Greenbaum 1991), which follows sampling methods similar to those used for the Brown corpus, and includes multiple varieties of English with texts from thirteen countries including Canada, Great Britain, Ireland, Jamaica, and the USA.

One of the main challenges when dealing with diatopic variation for low-resource languages is finding suitable resources and tools. Acquiring text corpora for dialects is particularly challenging as dialects are typically vastly underrepresented in written text. Thus, the typical solution is to produce text corpora by transcribing speech, as it is much easier to obtain spoken dialectal data. The transcription can be done automatically, for example, using Automatic Speech Recognition, which was used to produce Arabic dialectal text corpora (Ali *et al.* 2016), or manually, which was used to build the ArchiMob corpus for (Swiss) German dialects (Scherrer, Samardžić, and Glaser 2019) used in past editions of the German Dialect Identification shared tasks at VarDial. Alternative approaches to dialectal data collection include social media, for example, Twitter, (Cotterell and Callison-Burch 2014) and translations (Bouamor *et al.* 2018) as in the case of the MADAR corpus used in the MADAR shared task (Bouamor, Hassan, and Habash 2019) on fine-grained identification of Arabic dialects.

The case of national language varieties is generally less challenging. Each such variety (e.g., British vs. American English) has its own written standard, which often differs from other varieties

https://doi.org/10.1017/S1351324920000492 Published online by Cambridge University Press

of the same language in several aspects. This includes spelling, for example, *favour* in the UK versus *favor* in the USA and lexical preferences for example, *rubbish* in the UK versus *garbage* in the USA. Most books, magazines, and newspapers reflect these differences, which makes them suitable training resources for most language varieties. One example of a corpus collected for language varieties is the Discriminating between Similar Language (DSL) Corpus Collection (DSLCC)[a] (Tan *et al.* 2014), which contains short excerpts of texts (from 20 to 100 tokens each), collected from multiple newspapers per country.

The DSLCC was created to serve as a dataset for discriminating between similar languages and language varieties for the DSL shared tasks, organized annually within the scope of the VarDial workshop (Zampieri *et al.* 2014, 2015, 2017; Malmasi *et al.* 2016). The texts in DSLCC were compiled from existing corpora such as HC Corpora (Christensen 2014), the SETimes Corpus (Tyers and Alperen 2010), and the Leipzig Corpora Collection (Biemann *et al.* 2007). Five versions including 20,000–22,000 texts per language or language variety have been released with data from several pairs or groups of similar languages such as Bulgarian and Macedonian, Czech and Slovak, Bosnian, Croatian, and Serbian, Malay and Indonesian, and pairs or groups of language varieties such as Brazilian and European Portuguese, British and American English, and several varieties of Spanish for example, Argentinian and Peninsular Spanish. The languages and the language varieties included in all versions of the corpus collection are presented in Table 1.

The DSLCC features journalistic texts collected from multiple newspapers in each target country in order to alleviate potential topical and stylistic biases, which are intrinsic to any newspaper, that is, in order to prevent systems from learning a specific newspaper's writing style as opposed to learning the language variety it represents. Newspaper texts were chosen with the assumption that they are the most accurate representation of the contemporary written standard of a language in a given country and therefore could be used to represent national language varieties.

We should note that other popular data sources, which have been used in a variety of NLP tasks, for example, Wikipedia, are not suited to serve as training data for modeling diatopic variation in language as they disregard language varieties. Wikipedia is a collaborative resource, which allows speakers of multiple language varieties and non-native speakers to contribute to the same article(s) available in a single English, Portuguese, or Spanish Wikipedia. Notable exceptions are the *Simple English* Wikipedia, which, as the name suggests, contains simplified English, and a few (small) dialect Wikipedias.

Finally, movie and TV subtitles have also been used as data sources for Dutch and Flemish (van der Lee and van den Bosch 2017), as well as in related NLP applications such as MT between Brazilian and European Portuguese (Costa-jussà, Zampieri, and Pal 2018).

3. POS tagging and parsing

Part-of-speech tagging and parsing are two core morphosyntactic annotation tasks. Their output often serves as a pre-annotation for downstream applications such as information retrieval or natural language understanding, but morphosyntactic annotation is also useful for corpus linguistics research as it enables search queries that are independent of the lexical forms. This is especially practical for nonstandardized language varieties such as dialects.

Recent research on tagging and parsing similar languages, varieties, and dialects typically assumes that the availability of linguistic resources is asymmetric, in the sense that some varieties have more resources than other ones, and that low-resource language varieties can benefit from resources for high-resource ones. In such a scenario, a tagger or a parser for a new language variety can be produced using cross-lingual transfer learning (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

[a] http://ttg.uni-saarland.de/resources/DSLCC/.

https://doi.org/10.1017/S1351324920000492 Published online by Cambridge University Press

Table 1. Languages and language varieties used in the five versions of the DSLCC, grouped by language similarity. The checkboxes show which language variety was present in a particular version of the corpus

Language/Variety	v1.0	v2.0	v2.1	v3.0	v4.0
Bosnian	X	X	X	X	X
Croatian	X	X	X	X	X
Serbian	X	X	X	X	X
Czech	X	X	X		
Slovak	X	X	X		
Indonesian	X	X	X	X	X
Malay	X	X	X	X	X
Brazilian Portuguese	X	X	X	X	X
European Portuguese	X	X	X	X	X
Macanese Portuguese			X		
Argentine Spanish	X	X	X	X	X
Mexican Spanish			X	X	
Peninsular Spanish	X	X	X	X	X
Peruvian Spanish					X
Bulgarian		X	X		
Macedonian		X	X		
Canadian French				X	X
Hexagonal French				X	X
American English	X				
British English	X				
Persian					X
Dari					X

context, the high-resource language is also referred to as a *donor language* or a *source language* and the low-resource language as the *recipient language* or the *target language*. While transfer learning as such is not restricted to similar language varieties, the common assumption is that the closer the HRL and the LRL are, the simpler the transfer would be.

A straightforward cross-lingual transfer learning technique is plain model transfer, or using a more recent name, zero-shot learning. It assumes that the HRL and the LRL are the same, and that a model trained on HRL data can be applied directly and without any modification to LRL data. While this assumption is too naïve in most cases, plain model transfer results are often presented as simple baselines to which more sophisticated approaches are compared. For instance, Scherrer (2014) reported baseline tagging results for various related languages, and Huck *et al.* (2019) used zero-shot learning as a baseline for more sophisticated experiments of Russian → Ukrainian tagging transfer. Zampieri *et al.* (2017) trained dependency cross-lingual dependency parsing baselines for pairs like Slovenian → Croatian, Danish/Swedish → Norwegian, and Czech → Slovak.

https://doi.org/10.1017/S1351324920000492 Published online by Cambridge University Press

Zero-shot learning can be extended to multi-source scenarios, where there is a clearly defined low-resource variety, but several related high-resource varieties, all of which are expected to contribute to various extent to the analysis of the low-resource variety. Scherrer and Rabus (2017, 2019) trained a tagger on the concatenation of Slovak, Ukrainian, Polish, and Russian data and applied it directly to Rusyn (a Slavic variety spoken predominantly in Transcarpathian Ukraine, Eastern Slovakia, and Southeastern Poland). The only preprocessing consists in transliterating all data into a common script, in the present case Cyrillic. Likewise, Zampieri *et al.* (2017) showed that dependency parsing for Norwegian works best with a model trained on both Danish and Swedish data, rather than on just one of the source languages.

Another popular cross-lingual transfer learning technique is annotation projection (Yarowsky and Ngai 2001), which crucially requires a parallel corpus relating the HRL and the LRL. The HRL side of the parallel corpus is annotated using an existing model, and the labels are projected to the LRL side along the word alignment links. The annotated LRL side can then serve as a training corpus. A multilingual variant of annotation projection was introduced in Agić, Hovy, and Søgaard (2015). They projected annotations from all available source languages and used simple majority voting to resolve the ambiguities. A similar approach was proposed in Aepli, von Waldenfels, and Samardžić (2014). They created a tagger for Macedonian using majority voting from related languages such as Bulgarian, Czech, Slovene, and Serbian, as well as from English. Their full setup is somewhat more complicated and allows different morphological features to be inferred from different sets of source languages. In the cross-lingual dependency parsing task at VarDial 2017 (Zampieri *et al.* 2017), all participants opted for some variant of annotation projection. In particular, two teams relied on word-by-word translation models inferred from the provided parallel data (Rosa *et al.* 2017; Çöltekin and Rama 2017).

Despite these examples, annotation projection is often not a particularly popular choice for configurations involving closely related varieties because large parallel corpora can be hard to find. For example, parallel corpora involving a dialect and its corresponding standard variety are not naturally available, as dialects do not have an official status and speakers are generally fluent in both varieties, which obviates the need for translations.

Delexicalization is an alternative transfer learning technique that attempts to create models that do not rely on language-specific information, but rather on language-independent features and representations. It was first proposed for dependency parsing (Zeman and Resnik 2008; McDonald, Petrov, and Hall 2011), where the (language-dependent) word forms were replaced as input features by (language-independent) part-of-speech tags. For tagging, Täckström, McDonald, and Uszkoreit (2012) proposed to replace the (language-dependent) word forms by (language-independent) cluster ids obtained by clustering together all words with the same distributional properties. More recently, clusters were superseded by cross-lingual word embeddings as language-independent input features. However, a bottleneck of this approach is that the embeddings have to be trained on large amounts of raw data, typically in the order of millions of words. For low-resource varieties such as dialects or similar varieties, this requirement is unrealistic. For example, Magistry, Ligozat, and Rosset (2019) trained cross-lingual word embeddings for three French regional languages – Alsatian, Occitan, and Picard – in view of using them for part-of-speech tagging. They showed that embeddings trained on corpora containing one to two million words were not sufficient to train taggers that would be competitive to much simpler adaptation strategies.

Relexicalization is a variant of delexicalization in which the HRL input features are replaced by LRL features within the model. For example, the HRL word forms of the original models can be replaced by LRL word forms extracted from a bilingual dictionary (Feldman, Hana, and Brew 2006). This approach was adapted to closely related language varieties by Scherrer (2014), who built the bilingual dictionary in an unsupervised way, taking advantage of the prevalence of cognates, similar word forms and phrases. Following the success of cross-lingual word embeddings, relexicalization became less popular in recent years.

https://doi.org/10.1017/S1351324920000492 Published online by Cambridge University Press

The cross-lingual transfer techniques presented above assume that no annotated training data are available for the target variety, but that other types of data can be obtained, for example, raw text for training word embeddings, parallel corpora, bilingual dictionaries, etc. However, in many cases, at least small amounts of annotated data of the target variety can be made available. In this situation, the domain adaptation problem – where the annotated training data have different properties than the data the trained model is supposed to be applied to – can be reformulated as a language adaptation problem. The resulting approaches are typically subsumed as multilingual or multi-lectal models.

For example, Jørgensen, Hovy, and Søgaard (2016) created a tagger for African American Vernacular English (AAVE) tweets by first training a model on a large, non-AAVE-specific Twitter corpus and then added variety-specific information into the model. Similarly, in a range of experiments on tagging Ukrainian, Huck *et al.* (2019) obtained the best results with a multilingual model trained on a large corpus of Russian and a small corpus of Ukrainian. In earlier work, Cotterell and Heigold (2017) used several related resource-rich source languages to improve the performance of their taggers.

Multilingual models can be trained in a completely language-agnostic way, that is, by making the model believe that all training instances stem from the same language, but they generally work better if they are given information about the language variety of each training instance. One way of doing so is to use multi-task learning (Ruder 2017), where the task of detecting the language variety is introduced as an auxiliary task on top of the main task, for example, tagging or parsing (Cotterell and Heigold 2017; Scherrer and Rabus 2019).

All approaches discussed above rely on two crucial assumptions: (1) that there is significant overlap of the input features (which, in most cases, are word forms) across the languages or that the model is able to share information across similar but non-identical input features and (2) that the output labels (part-of-speech tags, dependency labels, constituent names, etc.) are unified across the languages or the varieties.

Let us first discuss assumption 2, that is, that the output labels (part-of-speech tags, dependency labels, constituent names, etc.) are unified across the languages or the varieties. For decades, the exact task definitions and annotation conventions have been determined more or less independently for each language, which made the generalization across languages difficult, both when creating models and when evaluating them (Scherrer 2014). It is only in recent years that this situation has improved through the development of a language-independent universal part-of-speech tag set (Petrov, Das, and McDonald 2012), a language-independent universal dependency annotation scheme (McDonald *et al.* 2013), a unified feature-value inventory for morphological features (Zeman 2008), and the subsequent merging of the three schemes within the Universal Dependencies project (Nivre *et al.* 2016). However, despite these huge harmonization efforts, different annotation traditions still shine in the currently available corpora (Zupan, Ljubešić, and Erjavec 2019), and as a result even recent research sometimes resorts to some kind of ad hoc label normalization (Rosa *et al.* 2017).

Regarding assumption 1, that is, that there is significant overlap of the input features (which, in most cases, are word forms) across the languages or that the model is able to share information across similar but non-identical input features, there have been essentially two extreme strategies: either one aggressively normalizes and standardizes the data, so that related words or word forms from different varieties are made to look the same, or one avoids all types of normalization and makes sure to choose a model architecture that is still capable of generalizing from variation within the data.

In NLP for historical language varieties, this question has largely been answered in favor of the first approach: namely, the massive amounts of variation occurring in the original data are normalized to some canonical spelling, which usually coincides with the modern-day standardized spelling (Tjong Kim Sang *et al.* 2017). While the normalization approach has been applied to dialectal varieties (Samardžić, Scherrer, and Glaser 2016), recent work (Scherrer, Rabus, and

https://doi.org/10.1017/S1351324920000492 Published online by Cambridge University Press

Mocken 2018) suggested that neural networks can actually extract sufficient information from raw (non-normalized) data, provided that the words are represented as character sequences rather than as atomic units. In this case, information can be shared across similarly spelled words both within the same language variety (e.g., if there is no orthographic standard) as well as across language varieties (e.g, in a multilingual model). For example, Scherrer and Rabus (2019) represented the input words using a bidirectional character-level long short-term memory (LSTM) recurrent neural network and obtained up to 13% absolute boost in terms of F1-score compared to using atomic word-level representations. Zupan *et al.* (2019) also stressed the importance of character-level input representations. In high-resource settings, character-level input representations, which are computationally costly, have lately been replaced by fixed-size vocabularies obtained by unsupervised subword segmentation methods such as byte-pair encodings and word-pieces (Sennrich, Haddow, and Birch 2016; Kudo and Richardson 2018); however, for the moment this change appears to be less relevant in low-resource tagging and parsing settings.

4. Applications

In this section, we describe studies addressing diatopic language variation on two relevant NLP applications: language and dialect identification and MT. Language and dialect identification is an important part of many NLP pipelines. For example, it can be used to help collecting suitable language-specific training data (Bergsma *et al.* 2012), or to aid geolocation prediction systems (Han, Cook, and Baldwin 2012). MT can address diatopic variation when translating between pairs of closely related languages (Nakov and Tiedemann 2012; Lakew, Cettolo, and Federico 2018), language varieties (Costa-jussà *et al.* 2018), and dialects (Zbib *et al.* 2012), or when using similar language data to augment the training data for low-resource languages, that is, by merging datasets from closely related languages to form larger parallel corpora (Nakov and Ng 2009, 2012).

4.1. Language and dialect identification

Language identification is a well-known research topic in NLP, and it represents an important part of many NLP applications. Language identification has been applied to speech (Zissman and Berkling 2001), to sign language (Gebre, Wittenburg, and Heskes 2013), and to written texts (Jauhiainen *et al.* 2019b), as we will see below. When applied to texts, language identification systems are trained to identify the language a document or a part of a document is written in.

Language identification is most commonly modeled as a supervised multi-class single-label document classification task, where each document is assigned one class from a small inventory of classes (languages, dialects, varieties). Given a set of n documents, a language identification system will typically implement the following four major steps (Lui 2014):

1. Represent the texts using characters, words, linguistically motivated features such as POS tags, or a combination of multiple features;
2. Train a model or build a language profile from documents known to be written in each of the target languages;
3. Define a classification function that best represents the similarity between a document and each language model or language profile;
4. Compute the probability of each class using the models to determine the most likely language for a given test document.

While most work has addressed language identification in sentences, paragraphs, or full texts, for example, newspaper articles, including the work we discuss in this section, some papers have focused at the word level (Nguyen and Dogruoz 2014).

https://doi.org/10.1017/S1351324920000492 Published online by Cambridge University Press

As discussed in a recent survey (Jauhiainen *et al.* 2019b), in the early 2000s, language identification was widely considered to be a solved task as n-gram methods performed very well at discriminating between unrelated languages in standard contemporary texts (McNamee 2005). There are, however, several challenging scenarios that have been explored in recent years, where the performance of language identification systems is far from perfect.

This is the case of multilingual documents (Lui, Lau, and Baldwin 2014), of short and noisy texts (Vogel and Tresner-Kirsch 2012) (such as user-generated content, e.g., microblogs and social media posts), of data containing code-switching, or code-mixing (Solorio *et al.* 2014), and finally, of similar languages, language varieties, and dialects, which we address in this section.

There have been a number of studies discussing methods to discriminate between similar languages. The study by Suzuki *et al.* (2002), for example, addressed the identification of language, script, and text encoding and showed that closely related language pairs that also share a common script and text encoding (e.g., Hindi and Marathi) are most difficult to discriminate between. Other studies have investigated the limitations of the use of general-purpose n-gram language identification methods, which are commonly trained using character trigrams. Some of these studies have proposed solutions tailored to particular pairs or groups of languages. Ranaivo-Malançon (2006) added a list of words exclusive to one of the languages to help a language identification system discriminate between Malay and Indonesian. In the same vein, Tiedemann and Ljubešić (2012) improved the performance of a baseline language identification system and reported 98% accuracy for discriminating between Bosnian, Croatian, and Serbian by applying a blacklist, that is, a list of words that do not appear in one of the three languages.

Previous research on the topic, including the aforementioned VarDial shared tasks, has shown that high-performing approaches to the task of discriminating between related languages, language varieties, and dialects tend to use word-based representations or character n-gram models of higher order (4-, 5-, or 6-grams), which can also cover entire words (Goutte *et al.* 2016). There have been studies that went beyond lexical features in an attempt to capture some of the abstract systemic differences between similar languages using linguistically motivated features. This includes the use of semi-delexicalized text representations in which named entities or content words are replaced by placeholders, or fully de-lexicalized representations using POS tags and other morphosyntactic information (Zampieri, Gebre, and Diwersy 2013; Diwersy, Evert, and Neumann 2014; Lui *et al.* 2014; Bestgen 2017).

In terms of computational methods, the bulk of research on this topic and the systems submitted to the DSL shared tasks at VarDial have shown that traditional machine learning classifiers such as support vector machines (Cortes and Vapnik 1995) tend to outperform dense neural network approaches for similar languages and language varieties (Bestgen 2017; Medvedeva, Kroon, and Plank 2017). The best system (Bernier-Colborne, Goutte, and Léger 2019) submitted to the VarDial 2019 Cuneiform Language Identification (CLI) shared task (Zampieri *et al.* 2019), however, has outperformed traditional machine learning methods using a BERT-based (Devlin *et al.* 2019) system to discriminating between Sumerian and Akkadian historical dialects in Cuneiform script (Jauhiainen *et al.* 2019a). BERT and other Transformer-based contextual representations have been recently applied to various NLP tasks achieving state-of-the-art results. The results by Bernier-Colborne *et al.* (2019) in the CLI shared task seem to indicate that recent developments in contextual embedding representations may also yield performance improvement in language identification applied to similar languages, varieties, and dialects.

Language identification was studied for closely related languages such as Malay–Indonesian (Ranaivo-Malançon 2006), South Slavic languages (Ljubešić, Mikelić, and Boras 2007; Tiedemann and Ljubešić 2012), and languages of the Iberian Peninsula (Zubiaga *et al.* 2014). It was also applied to national varieties of English (Lui and Cook 2013; Simaki *et al.* 2017), French (Mokhov 2010; Diwersy *et al.* 2014), Chinese (Huang and Lee 2008), and Portuguese (Zampieri and Gebre 2012; Zampieri *et al.* 2016), as well as to dialects of Romanian (Ciobanu and Dinu 2016), Arabic (Elfardy and Diab 2013; Zaidan and Callison-Burch 2014; Tillmann, Al-Onaizan, and

https://doi.org/10.1017/S1351324920000492 Published online by Cambridge University Press

Mansour 2014; Sadat, Kazemi, and Farzindar 2014; Wray 2018), and German (Hollenstein and Aepli 2015). The VarDial shared tasks included the languages in the DSLCC, as well as Chinese varieties, Dutch and Flemish, dialects of Arabic, Romanian, and German, and many others.

Arabic is particularly interesting as its standard form coexists with several regional dialects in a large dialect continuum. This has motivated the bulk of recent work on processing Arabic dialects including a number of studies on the identification of Arabic dialects (Elfardy and Diab 2013; Zaidan and Callison-Burch 2014). Tillmann *et al.* (2014) used a linear-kernel SVM to distinguish between MSA and Egyptian Arabic in the Arabic online commentary dataset (Zaidan and Callison-Burch 2011). Salloum *et al.* (2014) carried out an extrinsic evaluation of an Arabic dialect identification used as part of the preprocessing steps of a MT system. Salloum *et al.* (2014) reported improvements in terms of BLEU score compared to a baseline that did not differentiate between dialects. Moreover, shared tasks on Arabic dialect identification were organized in recent years providing participants with annotated dialectal data from various genres and domains. This includes the Arabic Dialect Identification (ADI) task at the VarDial workshop (Malmasi et al. 2016) and the Multi-Genre Broadcast (MGB) challenge (Ali, Vogel, and Renals 2017), which included broadcast speech, and the MADAR shared task (Bouamor *et al.* 2019), which included translations of tourism-related texts. There have been also a number of other multi-dialectal corpora compiled for Arabic including a parallel corpus of 2000 sentences in English, MSA, and multiple Arabic dialects (Bouamor, Habash, and Oflazer 2014); a corpus from web forums with data from eighteen Arabic-speaking countries (Sadat *et al.* 2014); as well as some multi-dialect corpora consisting of Twitter posts (Elgabou and Kazakov 2017; Alshutayri and Atwell 2017).

4.2. Machine translation

Early work on machine translation between closely related languages and dialects used word-for-word translation and manual language-specific rules to handle morphological and syntactic transformations. This was tried for a number of language pairs such as Czech–Slovak (Hajič, Hric, and Kuboň 2000), Czech–Russian (Bemova, Oliva, and Panevova 1988), Turkish–Crimean Tatar (Altintas and Cicekli 2002), Irish–Scottish Gaelic (Scannell 2006), Punjabi–Hindi (Josan and Lehal 2008), Levantine/Egyptian/Iraqi/Gulf–Standard Arabic (Salloum and Habash 2012), and Cantonese–Mandarin (Zhang 1998).

The Apertium MT platform (Corbí-Bellot *et al.* 2005) used bilingual dictionaries and manual rules to translate between a number of related Romance languages such as Spanish–Catalan, Spanish–Galician, Occitan–Catalan, and Portuguese–Spanish (Armentano-Oller et al. 2006), and it also supports some very small languages such as the Aranese variety of Occitan (Forcada 2006). There has also been work on rule-based MT using related language *pairs*, for example, improving Norwegian–English using Danish–English (Bick and Nygaard 2007). Finally, there have been rule-based MT systems for translating from English to American sign language (Zhao *et al.* 2000).

A special case is the translation between different dialects of the same language, for example, between Cantonese and Mandarin Chinese (Zhang 1998), or between a dialect of a language and a standard version of that language, for example, between Arabic dialects (Bakr, Shaalan, and Ziedan 2008; Sawaf 2010; Salloum and Habash 2011; Sajjad, Darwish, and Belinkov 2013). Here again, manual rules and/or language-specific tools and resources are typically used. In the case of Arabic dialects, a further complication arises due to their informal status, which means that they are primarily used in spoken interactions, or in informal text, for example, in social media, chats, forums, and SMS messages; this also causes mismatches in domain and genre. Thus, translating from Arabic dialects to Modern Standard Arabic requires, among other things, normalizing informal text to a formal form. For example, Sajjad *et al.* (2013) first normalized a dialectal Egyptian Arabic to look like MSA, and then translated the transformed text to English. In fact, this kind of

https://doi.org/10.1017/S1351324920000492 Published online by Cambridge University Press

adaptation is a more general problem, which arises with informal sources such as SMS messages and tweets for just any language (Aw et al. 2006; Han and Baldwin 2011; Wang and Ng 2013; Bojja, Nedunchezhian, and Wang 2015). Here the main focus is on coping with spelling errors, abbreviations, and slang, which are typically addressed using string edit distance, while also taking pronunciation into account.

Another line of research is on language adaptation and normalization, when done specifically for improving MT into another language. For example, Marujo *et al.* (2011) built a rule-based system for adapting Brazilian Portuguese (BP) to European Portuguese (EP), which they used to adapt BP–English bitexts to EP–English. They reported small improvements in BLEU for EP–English translation when training on the adapted "EP"–English bitext compared to using the unadapted BP–English, or when an EP–English bitext is used in addition to the adapted/unadapted one.

For closely related languages and dialects, especially when they use the same writing script, many differences might occur at the spelling/morphological level. Thus, there have been many successful attempts at performing translation using character-level representations, especially with phrase-based statistical machine translation (PBSMT) (Vilar, Peter, and Ney 2007; Tiedemann 2012; Nakov and Tiedemann 2012). As switching to characters as the basic unit of representation yields a severely reduced vocabulary, this causes a problem for word alignment, which is an essential step for PBSMT; one solution is to align character n-grams instead of single characters (Nakov and Tiedemann 2012; Tiedemann 2012; Tiedemann and Nakov 2013).

A third line of research is on reusing bitexts between related languages without or with very little adaptation. For example, Nakov and Ng (2009, 2012) experimented with various techniques to combine a small bitext for a resource-poor language, for example, Indonesian–English, with a much larger bitext for a related resource-rich language, for example, Malay–English. It has been further shown that it makes sense to combine the two ideas, that is, to adapt the resource-rich training bitext to look more similar to the resource-poor one, while also applying certain smart text combination techniques (Wang, Nakov, and Ng 2012, 2016).

A further idea is to use cascaded translation using a resource-rich language as a pivot, for example, translating from Macedonian to English by pivoting over Bulgarian (Tiedemann and Nakov 2013). The closer the pivot and the source, the better the results, that is, for Macedonian to English translation it is better to pivot over Bulgarian than over Slovenian or Czech, which are less related Slavic languages.

Most of the above work relates to rule-based or statistical MT, which has now become somewhat obsolete, as a result of the ongoing neural networks revolution in the field. The rise of word embeddings in 2013 had an immediate impact on MT, as word embeddings proved helpful for translating between related languages (Mikolov, Le, and Sutskever 2013). The subsequent Neural MT revolution of 2014 (Cho *et al.* 2014; Sutskever, Vinyals, and Le 2014; Bahdanau, Cho, and Bengio 2015) yielded LSTM-based recurrent neural language models with attention, known as seq2seq, which enabled easy translation between multiple languages, including many-to-many and zero-shot translation (Johnson *et al.* 2017; Aharoni, Johnson, and Firat 2019), and proved to be especially effective for related languages. It has been further shown that neural MT outperforms phrase-based statistical MT for closely related language varieties such as European–Brazilian Portuguese (Costa-jussà, Zampieri, and Pal 2018).

Neural MT also allows for an easy transfer from a resource-rich "parent"–target language pair such as Uzbek–English to a related resource-poor "child"–target language pair such as Uyghur–English (Nguyen and Chiang 2017) by simply pre-training on the "parent"–target language pair and then training on the "child"–target pair, using a form of transfer learning, originally proposed in Zoph *et al.* (2016). In 2017, along came the Transformer variant (Vaswani *et al.* 2017) of neural MT, which yielded improvements over RNN-based seq2seq, for example, for Romanian–Italian and Dutch–German in bilingual, multilingual and zero-shot setups (Lakew *et al.* 2018).

https://doi.org/10.1017/S1351324920000492 Published online by Cambridge University Press

Neural MT also makes it very easy to train multilingual models with multiple languages on the source side, on the target side, or on both sides (Johnson *et al*. 2017; Lakew *et al*. 2018; Aharoni *et al*. 2019). This is especially useful for closely related languages and language variants, as the model can learn from many languages simultaneously, but it has been shown that even distant languages can help in this setup. Part of the benefit comes from sharing subword-level vocabularies with tied embeddings, which allows models to learn some spelling and morphological variations between related languages.

Last but not least, there has been a lot of recent research interest in building cross-language word embeddings (Lample *et al*. 2018) or sentence representations without the need for parallel training bitexts or using both parallel and non-parallel data (Joty *et al.* 2017; Artetxe and Schwenk 2019; Conneau and Lample 2019; Søgaard *et al.* 2019; Guzmán *et al.* 2019; Cao, Kitaev, and Klein 2020), which can be quite helpful in a low-resource setting.

5. Conclusion

We have presented a survey of the growing field of research that focuses on computational methods for processing similar languages, language varieties, and dialects, with focus on diatopic language variation and integration in NLP applications. We have described some of the available datasets and the most common strategies used to create datasets for similar languages, language varieties, and dialects. We further noted that popular data sources used in NLP such as *Wikipedia* are not suited for language varieties and dialects, which motivated researchers to look for alternative data sources such as social media posts and speech transcripts. We further presented a number of studies describing methods for preprocessing, normalization, part-of-speech tagging, and parsing applied to similar languages, language varieties, and dialects. Finally, we discussed how closely related languages, language varieties, and dialects are handled in two prominent NLP applications: language and dialect identification, and MT.

6. Future perspectives

Given our discussion above, it is clear that more and more NLP applications developed in industry and in academia are addressing issues related to closely related languages, language variants, and dialects, with specific emphasis on diatopic variation. In recent years, we have seen a substantial increase in the number of resources such as corpora and tools created for similar languages, language varieties, and dialects, and we have described some of them in this survey. We have also seen an increase in the number of publications on these topics in scientific journals as well as in the main international conferences in Computational Linguistics such as ACL, EMNLP, NAACL, EACL, COLING, and LREC. The success of recent initiatives such as the aforementioned annual VarDial workshop series and the associated VarDial evaluation campaigns, which keep attracting a large number of participants, are another indication of the importance of these topics.

Finally, another recent evidence of the interest of the community in the computational processing of diatopic variation is the special issue of the Journal of Natural Language Engineering (NLE) on NLP for Similar Languages, Varieties and, Dialects, where this survey appears.[b]

The special issue received a record number of submissions for an NLE special issue and it was eventually split into two parts. The articles published in Part 1 (NLE 25:5) have demonstrated the vibrancy of research on the topic, covering a number of applications areas such as morphosyntactic tagging (Scherrer and Rabus 2019), text normalization (Martinc and Pollak 2019), and language identification (Jauhiainen, Lindén, and Jauhiainen 2019).

[b] http://sites.google.com/view/nledialects.

https://doi.org/10.1017/S1351324920000492 Published online by Cambridge University Press

References

Aepli N., von Waldenfels R. and Samardžić T. (2014). Part-of-speech tag disambiguation by cross-linguistic majority vote. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, Dublin, Ireland, pp. 76–84.

Agić ž., Hovy D. and Søgaard A. (2015). If all you have is a bit of the bible: Learning pos taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Beijing, China, pp. 268–272.

Aharoni R., Johnson M. and Firat O. (2019). Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL-HLT'19, Minneapolis, Minnesota, pp. 3874–3884.

AlGhamdi F. and Diab M. (2019). Leveraging pretrained word embeddings for part-of-speech tagging of code switching data. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, Ann Arbor, Michigan, June. Association for Computational Linguistics, pp. 99–109.

Ali A., Dehak N., Cardinal P., Khurana S., Yella S.H., Glass J., Bell P. and Renals S. (2016). Automatic dialect detection in Arabic broadcast speech. In *Proceedings of INTERSPEECH*, San Francisco, USA, pp. 2934–2938.

Ali A., Vogel S. and Renals S. (2017). Speech recognition challenge in the wild: Arabic MGB-3. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Okinawa, Japan, pp. 316–322.

Alshutayri, A. and Atwell E. (2017). Exploring twitter as a source of an Arabic dialect corpus. *International Journal of Computational Linguistics (IJCL)* **8**(2), 37–44.

Altintas K. and Cicekli I. (2002). A machine translation system between a pair of closely related languages. In *Proceedings of the 17th International Symposium on Computer and Information Sciences*, ISCIS'02, Orlando, Florida, USA, pp. 192–196.

Armentano-Oller C., Carrasco R.C., Corbí-Bellot A.M., Forcada M.L., Ginestí-Rosell M., Ortiz-Rojas S., Pérez-Ortiz J.A., Ramírez-Sánchez G., Sánchez-Martínez F. and Scalco M.A. (2006). Open-source Portuguese-Spanish machine translation. In *Proceedings of the 7th International Workshop on Computational Processing of the Portuguese Language*, PROPOR '06, Itatiaia, Brazil, pp. 50–59.

Artetxe M. and Schwenk H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics (TACL)* **7**, 597–610.

Aw A., Zhang M., Xiao J. and Su J. (2006). A phrase-based statistical model for SMS text normalization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, ACL-COLING'06, Sydney, Australia, pp. 33–40.

Bahdanau D., Cho K. and Bengio Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, ICLR'15, San Diego, California, USA.

Bakr H.A., Shaalan K. and Ziedan I. (2008). A hybrid approach for converting written Egyptian colloquial dialect into diacritized Arabic. In *Proceedings of the 6th International Conference on Informatics and Systems*, INFOS'08, Egypt, pp. 27–33.

Bemova A., Oliva K. and Panevova J. (1988). Some problems of machine translation between closely related languages. In *Proceedings of the International Conference on Computational Linguistics*, COLING'88, Budapest, Hungary.

Bergsma S., McNamee P., Bagdouri M., Fink C. and Wilson T. (2012). Language identification for creating language-specific Twitter collections. In *Proceedings of the Second Workshop on Language in Social Media*, pp. 65–74.

Bernier-Colborne G., Goutte C. and Léger S. (2019). Improving cuneiform language identification with BERT. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Minneapolis, USA, pp. 17–25.

Bestgen Y. (2017). Improving the character ngram model for the DSL task with BM25 weighting and less frequently used feature sets. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Valencia, Spain, pp. 115–123.

Bick E. and Nygaard L. (2007). Using Danish as a CG interlingua: A wide-coverage Norwegian-English machine translation system. In *Proceedings of the 16th Nordic Conference of Computational Linguistics*, NODALIDA'07, Tartu, Estonia, pp. 21–28.

Biemann C., Heyer G., Quasthoff U. and Richter M. (2007). The Leipzig corpora collection-monolingual corpora of standard size. In *Proceedings of Corpus Linguistics*.

Bojja N., Nedunchezhian A. and Wang P. (2015). Machine translation in mobile games: Augmenting social media text normalization with incentivized feedback. In *Proceedings of the 15th Machine Translation Summit (MT Users' Track)*, vol. 2, Miami, Florida, USA, pp. 11–16.

Bouamor H., Habash N. and Oflazer K. (2014). A multidialectal parallel corpus of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, pp. 1240–1245.

Bouamor H., Habash N., Salameh M., Zaghouani W., Rambow O., Abdulrahim D., Obeid O., Khalifa S., Eryani F., Erdmann A., Oflazer K. (2018). The MADAR arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, pp. 3387–3396.

https://doi.org/10.1017/S1351324920000492 Published online by Cambridge University Press

Bouamor H., Hassan S. and Habash N. (2019). The MADAR shared task on arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pp. 199–207.

Cao S., Kitaev N. and Klein D. (2020). Multilingual alignment of contextual word representations. In *Proceedings of the 8th International Conference on Learning Representations*, ICLR'20, Addis Ababa, Ethiopia.

Çöltekin Ç. and Rama T. (2017). Tübingen system in VarDial 2017 shared task: Experiments with language identification and cross-lingual parsing. In *Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Cho K., van Merriënboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H. and Bengio Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP'14, Doha, Qatar, pp. 1724–1734.

Christensen H. (2014). Hc corpora. http://www.corpora.heliohost.org/.

Ciobanu A.M. and Dinu L.P. (2016). A computational perspective on the Romanian dialects. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, May, pp. 3281–3285.

Clyne M. (1992). *Pluricentric Languages: Different Norms in Different Nations*, Amsterdam: De Gruyter Mouton.

Conneau A. and Lample G. (2019). Cross-lingual language model pretraining. In Wallach H., Larochelle H., Beygelzimer A., dAlché-Buc F., Fox E. and Garnett R. (eds), Advances in Neural Information Processing Systems 32, Vancouver, Canada, pp. 7059–7069.

Corbí-Bellot A.M., Forcada M.L., Ortiz-Rojas S., Pérez-Ortiz J.A., Ramírez-Sánchez G., Sánchez-Martínez F., Alegria I., Mayor A. and Sarasola K. (2005). An open-source shallow-transfer machine translation engine for the romance languages of Spain. In *Proceedings of the Tenth Conference of the European Association for Machine Translation*, EAMT'05, Budapest, Hungary, pp. 79–86.

Cortes C. and Vapnik V. (1995). Support-vector networks. *Machine Learning* **20**(3), 273–297.

Costa-jussà M.R., Zampieri M. and Pal S. (2018). A neural approach to language variety translation. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects*, VarDial'18, Santa Fe, New Mexico, USA, pp. 275–282.

Cotterell R. and Callison-Burch C. (2014). A multi-dialect, multi-genre corpus of informal written Arabic. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, pp. 241–245.

Cotterell R. and Heigold G. (2017). Cross-lingual character-level neural morphological tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp. 748–759.

Devlin J., Chang M.-W., Lee K. and Toutanova K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4171–4186.

Diwersy S., Evert S. and Neumann S. (2014). A weakly supervised multivariate approach to the study of language variation. In Szmrecsanyi B. and Wälchli B. (eds), *Aggregating Dialectology, Typology, and Register Analysis. Linguistic Variation in Text and Speech*. Berlin: De Gruyter.

Elfardy H. and Diab M. (2013). Sentence level dialect identification in Arabic. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Sofia, Bulgaria, pp. 456–461.

Elgabou H.A. and Kazakov D. (2017). Building dialectal Arabic corpora. In *The Proceedings of the First Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT)*, Varna, Bulgaria, pp. 52–57.

Feldman A., Hana J. and Brew C. (2006). A cross-language approach to rapid creation of new morphosyntactically annotated resources. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*. European Language Resources Association (ELRA), pp. 549–554.

Forcada M.L. (2006). Open-source machine translation: An opportunity for minor languages. In *Proceedings of the LREC'06 Workshop on Strategies for Developing Machine Translation for Minority Languages*, Genoa, Italy.

Francis W.N. and Kucera H. (1979). *Brown Corpus Manual*.

Găman M., Hovy D., Ionescu R.T., Jauhiainen H., Jauhiainen T., Lindén K., Ljubešić N., Partanen N., Purschke C., Scherrer Y. and Zampieri M. (2020). A report on the VarDial evaluation campaign 2020. In *Proceedings of the Seventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Gebre B.G., Wittenburg P. and Heskes T. (2013). Automatic sign language identification. In *Proceedings of the IEEE International Conference on Image Processing*. IEEE, pp. 2626–2630.

Goutte C., Léger S., Malmasi S. and Zampieri M. (2016). Discriminating similar languages: Evaluations and explorations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016*, pp. 1800–1807.

Greenbaum S. (1991). Ice: The international corpus of english. *English Today* **7**(4), 3–7.

Guzmán F., Chen P.-J., Ott M., Pino J., Lample G., Koehn P., Chaudhary V. and Ranzato M. (2019). The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, EMNLP-IJCNLP'19, Hong Kong, China, pp. 6098–6111.

Hajič J., Hric J. and Kuboň V. (2000). Machine translation of very close languages. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLP'00, Seattle, Washington, USA, pp. 7–12.

Han B. and Baldwin T. (2011). Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, ACL-HLT'11, Portland, Oregon, USA, pp. 368–378.

https://doi.org/10.1017/S1351324920000492 Published online by Cambridge University Press

Han B., Cook P. and Baldwin T. (2012). Geolocation prediction in social media data by finding location indicative words. In *Proceedings of the International Conference in Computational Linguistics (COLING)*, pp. 1045–1062.

Hollenstein N. and Aepli N. (2015). A resource for natural language processing of Swiss German dialects. In *Proceedings of GSCL*, pp. 108–109.

Huang C.-R. and Lee L.-H. (2008). Contrastive approach towards text source classification based on top-bag-of-word similarity. In *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation*, Cebu City, Philippines, November, pp. 404–410.

Huck M., Dutka D. and Fraser A. (2019). Cross-lingual annotation projection is effective for neural part-of-speech tagging. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, Ann Arbor, Michigan, June. Association for Computational Linguistics, pp. 223–233.

Jauhiainen T., Lindén K. and Jauhiainen H. (2019). Language model adaptation for language and dialect identification of text. *Natural Language Engineering* **25**(5), 561–583.

Jauhiainen T., Jauhiainen H., Alstola T. and Lindén K. (2019a). Language and dialect identification of Cuneiform texts. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Ann Arbor, Michigan. Association for Computational Linguistics, pp. 89–98.

Jauhiainen T., Lui M., Zampieri M., Baldwin T. and Lindén K. (2019b). Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, pp. 675–782.

Johnson M., Schuster M., Le Q.V., Krikun M., Wu Y., Chen Z., Thorat N., Viégas F., Wattenberg M., Corrado G., Hughes M. and Dean J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics* **5**, 339–351.

Jørgensen A., Hovy D. and Søgaard A. (2016). Learning a POS tagger for AAVE-like language. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California, pp. 1115–1120.

Josan G.S. and Lehal G.S. (2008). A Punjabi to Hindi machine translation system. In *Proceedings of the 22nd International Conference on on Computational Linguistics*, COLING'08, Manchester, UK, pp. 157–160.

Joty S., Nakov P., Màrquez L. and Jaradat I. (2017). Cross-language learning with adversarial neural networks. In *Proceedings of the 21st Conference on Computational Natural Language Learning*, CoNLL'17, Vancouver, Canada, pp. 226–237.

Kudo T. and Richardson J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Brussels, Belgium, November. Association for Computational Linguistics, pp. 66–71.

Lakew S.M., Cettolo M. and Federico M. (2018). A comparison of transformer and recurrent neural networks on multilingual neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, COLING'18, Santa Fe, New Mexico, USA, pp. 641–652.

Lample G., Conneau A., Ranzato M., Denoyer L. and Jégou H. (2018). Word translation without parallel data. In *Proceedings of the 6th International Conference on Learning Representations*, ICLR'18, Vancouver, BC, Canada.

Ljubešić N., Mikelić N. and Boras D. (2007). Language identification: How to distinguish similar languages? In *Proceedings of the 29th International Conference on Information Technology Interfaces (ITI 2007)*, Cavtat/Dubrovnik, Croatia, pp. 541–546.

Lui M. (2014). *Generalized Language Identification*. PhD Thesis, University of Melbourne.

Lui M. and Cook P. (2013). Classifying English documents by national dialect. In *Proceedings of Australasian Language Technology Association Workshop 2013 (ALTA 2013)*, Brisbane, Australia, December, pp. 5–15.

Lui M., Lau J.H. and Baldwin T. (2014). Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics* **2**, 27–40.

Lui M., Letcher N., Adams O., Duong L., Cook P. and Baldwin T. (2014). Exploring methods and resources for discriminating similar languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, Dublin, Ireland, August, pp. 129–138.

Lusetti M., Ruzsics T., Göhring A., Samardžić T. and Stark E. 2018. Encoder-decoder methods for text normalization. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics, pp. 18–28.

Magistry P., Ligozat A.-L. and Rosset S. (2019). Exploiting languages proximity for part-of-speech tagging of three French regional languages. *Language Resources and Evaluation* **53**, 865–888.

Malmasi S., Zampieri M., Ljubešić N., Nakov P., Ali A. and Tiedemann J. (2016). Discriminating between similar languages and Arabic dialect identification: A report on the third DSL shared task. In *Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Osaka, Japan, pp. 1–14.

Martinc M. and Pollak S. (2019). Combining N-grams and deep convolutional features for language variety classification. *Natural Language Engineering* **25**(5), 607–632.

Marujo L., Grazina N., Luís T., Ling W., Coheur L. and Trancoso I. (2011). BP2EP - adaptation of Brazilian Portuguese texts to European Portuguese. In *Proceedings of the 15th Conference of the European Association for Machine Translation*, EAMT'11, Leuven, Belgium, pp. 129–136.

McDonald R., Nivre J., Quirmbach-Brundage Y., Goldberg Y., Das D., Ganchev K., Hall K., Petrov S., Zhang H., Täckström O., Bedini C., Castelló N.B. and Lee J. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of ACL*.

McDonald R., Petrov S. and Hall K. (2011). Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK, pp. 62–72.

McNamee P. (2005). Language identification: A solved problem suitable for undergraduate instruction. *Journal of Computing Sciences in Colleges* **20**(3), 94–101.

Medvedeva M., Kroon M. and Plank B. (2017) When sparse traditional models outperform dense neural networks: The curious case of discriminating between similar languages. In *Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pp. 156–163.

Mikolov T., Le Q.V. and Sutskever I. (2013). Exploiting similarities among languages for machine translation. CoRR, abs/1309.4168.

Mokhov S.A. (2010). A MARF approach to DEFT 2010. In *Proceedings of the 6th DEFT Workshop (DEFT'10)*, pp. 35–49.

Myint Oo T., Kyaw Thu Y. and Mar Soe K. (2019) Neural machine translation between Myanmar (Burmese) and rakhine (arakanese). In *Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Minneapolis, USA, pp. 80—88.

Nakov P. and Ng H.T. (2009). Improved statistical machine translation for resource-poor languages using related resource-rich languages. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, EMNLP'09, Singapore, pp. 1358–1367.

Nakov P. and Ng H.T. (2012). Improving statistical machine translation for a resource-poor language using related resource-rich languages. *Journal of Artificial Intelligence Research*, **44**, 179–222.

Nakov P. and Tiedemann J. (2012). Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ACL'12, Jeju Island, Korea, pp. 301–305.

Nguyen D. and Dogruoz A.S. (2014). Word level language identification in online multilingual communication. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pp. 18–21.

Nguyen T.Q. and Chiang D. (2017). Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, IJCNLP'17, Taipei, Taiwan, pp. 296–301.

Nivre J., de Marneffe M.-C., Ginter F., Goldberg Y., Hajič J., Manning C.D., McDonald R., Petrov S., Pyysalo S., Silveira N., Silveira R., Zeman D. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Portoroz, Slovenia, pp. 1659–1666.

Petrov S., Das D. and McDonald R. (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).

Popović M., Poncelas A., Brkic M. and Way A. (2020). Neural machine translation for translating into Croatian and Serbian. In *Proceedings of the Seventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Ranaivo-Malançon B. (2006). Automatic identification of close languages – case study: Malay and Indonesian. *ECTI Transactions on Computer and Information Technology* **2**(2), 126–134.

Rosa R., Zeman D., Mareček D. and Žabokrtský Z. (2017). Slavic forest, Norwegian wood. In *Proceedings of the VarDial Workshop (VarDial)*.

Ruder S. (2017). An overview of multi-task learning in deep neural networks. arXiv e-prints, page *arXiv:1706.05098*.

Sadat F., Kazemi F. and Farzindar A. (2014). Automatic identification of Arabic dialects in social media. In *Proceedings of the First International Workshop on Social Media Retrieval and Analysis (SoMeRA 2014)*, Gold Coast, Australia. ACM, pp. 35–40.

Sajjad H., Darwish K. and Belinkov Y. (2013). Translating dialectal Arabic to English. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ACL'13, Sofia, Bulgaria, pp. 1–6.

Salloum W., Elfardy H., Alamir-Salloum L., Habash N. and Diab M. (2014). Sentence level dialect identification for machine translation system selection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Baltimore, USA, pp. 772–778.

Salloum W. and Habash N. (2011). Dialectal to Standard Arabic paraphrasing to improve Arabic-English statistical machine translation. In *Proceedings of the Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, Stroudsburg, Pennsylvania, USA, pp. 10–21.

Salloum, W. and Habash N. (2012). Elissa: A dialectal to standard Arabic machine translation system. In *Proceedings of COLING 2012: Demonstration Papers*, COLING'12, Mumbai, India, pp. 385–392.

Samardžić T., Scherrer Y. and Glaser E. (2016). ArchiMob – a corpus of spoken Swiss German. In *Proceedings of LREC*.

Sawaf H. (2010). Arabic dialect handling in hybrid machine translation. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas*, AMTA'10, Denver, Colorado, USA.

https://doi.org/10.1017/S1351324920000492 Published online by Cambridge University Press

Scannell K.P. (2006). Machine translation for closely related language pairs. In *Proceedings of the LREC 2006 Workshop on Strategies for Developing Machine Translation for Minority Languages*, Genoa, Italy, pp. 103–109.

Scherrer Y. (2014). Unsupervised adaptation of supervised part-of-speech taggers for closely related languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, Dublin, Ireland, pp. 30–38.

Scherrer Y. and Rabus A. (2017). Multi-source morphosyntactic tagging for spoken rusyn. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Valencia, Spain, pp. 84–92.

Scherrer Y. and Rabus A. (2019). Neural morphosyntactic tagging for Rusyn. *Natural Language Engineering* **25**(5), 633–650.

Scherrer Y., Rabus A. and Mocken S. (2018). New developments in tagging pre-modern orthodox slavic texts. *Scripta & e-Scripta* **18**, 9–33.

Scherrer Y., Samardžić T. and Glaser E. (2019). Digitising Swiss German – how to process and study a polycentric spoken language. *Language Resources and Evaluation* **53**(4), 735–769.

Sennrich R., Haddow B. and Birch A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, August. Association for Computational Linguistics, pp. 1715–1725.

Shapiro P. and Duh K. (2019). Comparing pipelined and integrated approaches to dialectal Arabic neural machine translation. In *Proceedings of the Workshop on NLP for Similar Languages Varieties and Dialects (VarDial)*, Minneapolis, USA, pp. 214—222.

Simaki V., Simakis P., Paradis C. and Kerren A. (2017). Identifying the authors' national variety of English in social media texts. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2017)*, Varna, Bulgaria, September. INCOMA Ltd., pp. 671–678.

Søgaard A., Vulic I., Ruder S. and Faruqui M. (2019). *Cross-Lingual Word Embeddings*. *Synthesis Lectures on Human Language Technologies*. San Rafael: Morgan & Claypool Publishers.

Solorio T., Blair E., Maharjan S., Bethard S., Diab M., Ghoneim M., Hawwari A., AlGhamdi F., Hirschberg J., Chang A. and Fung P. (2014). Overview for the first shared task on language identification in code-switched data. In *Proceedings of the Workshop on Computational Approaches to Code Switching*, Doha, Qatar, pp. 62–72.

Sutskever I., Vinyals O. and Le Q.V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing System*, NIPS'14, Montreal, Canada, pp. 3104–3112.

Suzuki I., Mikami Y., Ohsato A. and Chubachi Y. (2002). A language and character set determination method based on N-gram statistics. *ACM Transactions on Asian Language Information Processing (TALIP)* **1**(3), 269–278.

Täckström O., McDonald R. and Uszkoreit J. (2012). Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montréal, Canada, pp. 477–487.

Tan L., Zampieri M., Ljubešić N. and Tiedemann J. (2014). Merging comparable data sources for the discrimination of similar languages: The DSL corpus collection. In *Proceedings of the Workshop on Building and Using Comparable Corpora (BUCC)*, Reykjavik, Iceland, pp. 6–10.

Tiedemann J. (2012). Character-based pivot translation for under-resourced languages and domains. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL'12, Avignon, France, pp. 141–151.

Tiedemann J. and Agić Ž. (2016). Synthetic treebanking for cross-lingual dependency parsing. *Journal of Artificial Intelligence Research* **55**, 209–248.

Tiedemann J. and Ljubešić N. (2012). Efficient discrimination between closely related languages. In *Proceedings of the International Conference in Computational Linguistics (COLING)*, Mumbai, India, pp. 2619–2634.

Tiedemann J. and Nakov P. (2013). Analyzing the use of character-level translation with sparse and noisy datasets. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, RANLP'13, Hissar, Bulgaria, pp. 676–684.

Tillmann C., Al-Onaizan Y. and Mansour S. (2014). Improved sentence-level Arabic dialect classification. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, Dublin, Ireland, pp. 110–119.

Tjong Kim Sang E., Bollmann M., Boschker R., Casacuberta F., Dietz F., Dipper S., Domingo M., van der Goot R., van Koppen M., Ljubešić N., Östling R., Petran F., Pettersson E., Scherrer Y., Schraagen M., Sevens L., Tiedemann J., Vanallemeersch T. and Zervanou K. (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.

Tyers F. and Alperen M.S. (2010). South-East European Times: A parallel corpus of Balkan languages. In *Proceedings of the LREC workshop on Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages*.

van der Lee C. and van den Bosch A. (2017). Exploring lexical and syntactic features for language variety identification. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Valencia, Spain, pp. 190–199.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L. and Polosukhin I. (2017). Attention is all you need. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, NIPS'17, Long Beach, California, USA, pp. 5998–6008.

Vilar D., Peter J.-T. and Ney H. (2007). Can we translate letters? In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT'07, Prague, Czech Republic, pp. 33–39.

Vogel J. and Tresner-Kirsch D. (2012). Robust language identification in short, noisy texts: Improvements to LIGA. In *Third International Workshop on Mining Ubiquitous and Social Environments (MUSE 2012)*.

Wang P., Nakov P. and Ng H.T. (2012). Source language adaptation for resource-poor machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL'12, Jeju Island, Korea, pp. 286–296.

Wang P., Nakov P. and Ng H.T. (2016). Source language adaptation approaches for resource-poor machine translation. *Computational Linguistics* **42**(2), 277–306.

Wang P. and Ng H.T. (2013). A beam-search decoder for normalization of social media text with application to machine translation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT'13, Atlanta, Georgia, USA, pp. 471–481.

Wray S. (2018). Classification of closely related sub-dialects of Arabic using support-vector machines. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC), Miyazaki, Japan*, pp. 3671–3674.

Yarowsky D. and Ngai G. (2001). Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Pittsburgh, USA*, pp. 200–207.

Zaidan O.F. and Callison-Burch C. (2011). The Arabic online commentary dataset: An annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*, Portland, Oregon, USA, June, pp. 37–41.

Zaidan O.F. and Callison-Burch C. (2014). Arabic dialect identification. *Computational Linguistics* **40**(1), 171–202.

Zampieri M. and Gebre B.G. (2012). Automatic identification of language varieties: The case of Portuguese. In *Proceedings of The 11th Conference on Natural Language Processing (KONVENS 2012)*, Vienna, Austria, pp. 233–237.

Zampieri M., Gebre B.G. and Diwersy S. (2013). N-gram language models and POS distribution for the identification of Spanish varieties. In *Proceedings of la 20ème conférence du Traitement Automatique du Langage Naturel (TALN)*, Sables d'Olonne, France, pp. 580–587.

Zampieri M., Malmasi S., Ljubešić N., Nakov P., Ali A., Tiedemann J., Scherrer Y. and Aepli N. (2017). Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Valencia, Spain. Association for Computational Linguistics, pp. 1–15.

Zampieri M., Malmasi S., Nakov P., Ali A., Shon S., Glass J., Scherrer Y., Samardžić T., Ljubešić N., Tiedemann J., van der Lee C., Grondelaers S., Oostdijk N., Speelman D., van den Bosch A., Kumar R., Lahiri B. and Jain M. (2018). Language identification and morphosyntactic tagging: The second VarDial evaluation campaign. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*. Association for Computational Linguistics, pp. 1–17.

Zampieri M., Malmasi S., Scherrer Y., Samardžić T., Tyers F., Silfverberg M., Klyueva N., Pan T.-L., Huang C.-R., Ionescu R.T., Butnaru A. and Jauhiainen T. (2019). A report on the third VarDial evaluation campaign. In *Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Association for Computational Linguistics, pp. 1–16.

Zampieri M., Malmasi S., Sulea O.-M. and Dinu L.P. (2016). A computational approach to the study of Portuguese newspapers published in Macau. In *Proceedings of the Workshop on Natural Language Processing meets Journalism (NLPMJ 2016)*, New York City, NY, USA, pp. 47–51.

Zampieri M., Tan L., Ljubešić N. and Tiedemann J. (2014). A report on the DSL shared task 2014. In *Proceedings of the Workshop on NLP for Similar Languages Varieites and Dialects (VarDial)*, Dublin, Ireland, pp. 58–67.

Zampieri M., Tan L., Ljubešić N., Tiedemann J. and Nakov P. (2015). Overview of the DSL shared task 2015. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LTVarDial)*, Hissar, Bulgaria, pp. 1–9.

Zbib R., Malchiodi E., Devlin J., Stallard D., Matsoukas S., Schwartz R., Makhoul J., Zaidan O.F. and Callison-Burch C. (2012). Machine translation of Arabic dialects. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT)*, Montreal, Canada, pp. 49–59.

Zeman D. (2008). Reusable tagset conversion using tagset drivers. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, pp. 213–218.

Zeman D. and Resnik P. (2008). Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, Hyderabad, India, pp. 35–42.

Zhang X. (1998). Dialect MT: A case study between Cantonese and Mandarin. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, ACL-COLING'98, Quebec, Canada, pp. 1460–1464.

Zhao L., Kipper K., Schuler W., Vogler C., Badler N.I. and Palmer M. (2000). A machine translation system from English to American Sign Language. In *Proceedings of the 4th Conference of the Association for Machine Translation in the Americas on Envisioning Machine Translation in the Information Future*, AMTA'00, London, UK, pp. 54–67.

Zissman M.A. and Berkling K.M. (2001). Automatic language identification. *Speech Communication* **35**(1–2), 115–124.

Zoph B., **Yuret D.**, **May J. and Knight K.** (2016). Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, EMNLP'16, Austin, Texas, USA, pp. 1568–1575.

Zubiaga A., **Vicente I.S.**, **Gamallo P.**, **Pichel J.R.**, **Alegria I.**, **Aranberri N.**, **Ezeiza A. and Fresno V.** (2014). Overview of TweetLID: Tweet language identification at SEPLN 2014. In *Proceedings of the Tweet Language Identification Workshop 2014 co-located with 30th Conference of the Spanish Society for Natural Language Processing (SEPLN 2014)*, Girona, Spain, pp. 1–11.

Zupan K., **Ljubešić N. and Erjavec T.** (2019). How to tag non-standard language: Normalisation versus domain adaptation for slovene historical and user-generated texts. *Natural Language Engineering* **25**(5), 651–674.

Cite this article: Zampieri M, Nakov P and Scherrer Y (2020). Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering* **26**, 595–612. https://doi.org/10.1017/S1351324920000492