Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing

Edoardo Maria Ponti
LTL
University of Cambridge
ep490@cam.ac.uk

Helen O'Horan
LTL
University of Cambridge
helen.ohoran@gmail.com

Yevgeni Berzak
Department of Brain and Cognitive
Sciences MIT
berzak@mit.edu

Ivan Vulić
LTL
University of Cambridge
iv250@cam.ac.uk

Roi Reichart
Faculty of Industrial Engineering
and Management
Technion – IIT
roiri@ie.technion.ac.il

Thierry Poibeau
LATTICE Lab
CNRS, ENS/PSL, Universite Sorbonne
nouvelle/USPC
thierry.poibeau@ens.fr

Ekaterina Shutova
ILLC
University of Amsterdam
e.shutova@uva.nl

Anna Korhonen
LTL
University of Cambridge
alk23@cam.ac.uk

Linguistic typology aims to capture structural and semantic variation across the world's languages. A large-scale typology could provide excellent guidance for multilingual Natural Language Processing (NLP), particularly for languages that suffer from the lack of human labeled resources. We present an extensive literature survey on the use of typological information in the development of NLP techniques. Our survey demonstrates that to date, the use of information in existing typological databases has resulted in consistent but modest improvements in system performance. We show that this is due to both intrinsic limitations of databases (in terms of coverage and feature granularity) and under-utilization of the typological features included in them. We advocate for a new approach that adapts the broad and discrete nature of typological categories to the contextual and continuous nature of machine learning algorithms used in contemporary NLP. In particular, we suggest that such an approach could be facilitated by recent developments in data-driven induction of typological knowledge.

1. Introduction

The world's languages may share universal features at a deep, abstract level, but the structures found in real-world, surface-level texts can vary significantly. This cross-lingual variation has challenged the development of robust, multilingually applicable Natural Language Processing (NLP) technology, and as a consequence, existing NLP is still largely limited to a handful of resource-rich languages. The architecture design, training, and hyper-parameter tuning of most current algorithms are far from being language-agnostic, and often inadvertently incorporate language-specific biases (Bender 2009, 2011). In addition, most state-of-the-art machine learning models rely on supervision from (large amounts of) labeled data—a requirement that cannot be met for the majority of the world's languages (Snyder 2010).

Over time, approaches have been developed to address the data bottleneck in multilingual NLP. These include unsupervised models that do not rely on the availability of manually annotated resources (Snyder and Barzilay 2008; Vulić, De Smet, and Moens 2011, inter alia) and techniques that transfer data or models from resource-rich to resource-poor languages (Padó and Lapata 2005; Das and Petrov 2011; Täckström, McDonald, and Uszkoreit 2012, inter alia). Some multilingual applications, such as Neural Machine Translation and Information Retrieval, have been facilitated by learning joint models that learn from several languages (Ammar et al. 2016; Johnson et al. 2017, inter alia) or via multilingual distributed representations of words and sentences (Mikolov, Le, and Sutskever 2013, inter alia). Such techniques can lead to significant improvements in performance and parameter efficiency over monolingual baselines (Pappas and Popescu-Belis 2017).

Another, highly promising source of information for modeling cross-lingual variation can be found in the field of Linguistic Typology. This discipline aims to systematically compare and document the world's languages based on the empirical observation of their variation with respect to cross-lingual benchmarks (Comrie 1989; Croft 2003). Research efforts in this field have resulted in large typological databases—for example, most prominently the World Atlas of Language Structures (WALS) (Dryer

and Haspelmath 2013). Such databases can serve as a source of guidance for feature choice, algorithm design, and data selection or generation in multilingual NLP.

Previous surveys on this topic have covered earlier research integrating typological knowledge into NLP (O'Horan et al. 2016; Bender 2016). However, there is still no consensus on the general effectiveness of this approach. For instance, Sproat (2016) has argued that data-driven machine learning should not need to commit to any assumptions about categorical and manually defined language types as defined in typological databases.

In this article, we provide an extensive survey of typologically informed NLP methods to date, including the more recent neural approaches not previously surveyed in this area. We consider the impact of typological (including both structural and semantic) information on system performance and discuss the optimal sources for such information. Traditionally, typological information has been obtained from hand-crafted databases and, therefore, it tends to be coarse-grained and incomplete. Recent research has focused on inferring typological information automatically from multilingual data (Asgari and Schütze 2017, inter alia), with the specific purpose of obtaining a more complete and finer-grained set of feature values. We survey these techniques and discuss ways to integrate their predictions into the current NLP algorithms. To the best of our knowledge, this has not yet been covered in the existing literature.

In short, the key questions our paper addresses can be summarized as follows: (i) Which NLP tasks and applications can benefit from typology? (ii) What are the advantages and limitations of currently available typological databases? Can data-driven inference of typological features offer an alternative source of information? (iii) Which methods have been proposed to incorporate typological information in NLP systems, and how should such information be encoded? (iv) To what extent does the performance of typology-savvy methods surpass typology-agnostic baselines? How does typology compare with other criteria of language classification, such as genealogy? (v) How can typology be harnessed for data selection, rule-based systems, and model interpretation?

We start this survey with a brief overview of Linguistic Typology (§ 2) and multilingual NLP (§ 3). After these introductory sections we proceed to examine the development of typological information for NLP, including that in hand-crafted typological databases and that derived through automatic inference from linguistic data (§ 4). In the same section, we also describe typological features commonly selected for application in NLP. In § 5 we discuss ways in which typological information has been integrated into NLP algorithms, identifying the main trends and comparing the performance of a range of methods. Finally, in § 6 we discuss the current limitations in the use of typology in NLP and propose novel research directions inspired by our findings.

2. Overview of Linguistic Typology

There is no consensus on the precise number of languages in the world. For example, Glottolog provides the estimate of 7,748 (Hammarström et al. 2016), whereas Ethnologue (Lewis, Simons, and Fennig 2016) refers to 7,097.[1] This is because defining what constitutes a 'language' is in part arbitrary. Mutual intelligibility, which is used as the main criterion for including different language variants under the same label, is

---

1 These counts include only languages traditionally spoken by a community as their principal means of communication, and exclude unattested, pidgin, whistled, and sign languages.

gradient in nature. Moreover, social and political factors play a role in the definition of language.

Linguistic Typology is the discipline that studies the variation among the world's languages through their systematic comparison (Comrie 1989; Croft 2003). The comparison is challenging because linguistic categories cannot be predefined (Haspelmath 2007). Rather, cross-linguistically significant categories emerge inductively from the comparison of known languages, and are progressively refined with the discovery of new languages. Crucially, the comparison needs to be based on functional criteria, rather than formal criteria. Typologists distinguish between **constructions**, abstract and universal functions, and **strategies**, the type of expressions adopted by each language to codify a specific construction (Croft et al. 2017). For instance, the passive voice is considered a strategy that emphasizes the semantic role of patient: some languages lack this strategy and use other strategies to express the construction. For instance, Awtuw (Sepik family) simply allows for the subject to be omitted.

The classification of the strategies in each language is grounded in typological **documentation** (Bickel 2007, page 248). Documentation is empirical in nature and involves collecting texts or speech excerpts, and assessing the features of a language based on their analysis. The resulting information is stored in large databases (see § 4.1) of attribute–values (this pair is henceforth referred to as *typological feature*), where usually each attribute corresponds to a construction and each value to the most widespread strategy in a specific language.

Analysis of cross-lingual patterns reveals that cross-lingual variation is bounded and far from random (Greenberg 1966b). Indeed, typological features can be interdependent: The presence of one feature may implicate another (in one direction or both). This interdependence is called **restricted universal**, as opposed to unrestricted universals, which specify properties shared unconditionally by all languages. Such typological universals (restricted or not) are rarely absolute (i.e., exceptionless); rather, they are tendencies (Corbett 2010), hence they are called "statistical." For example, consider this restricted universal: If a language (such as Hmong Njua, Hmong–Mien family) has prepositions, then genitive-like modifiers follow their head. If, instead, a language (such as Slavey, Na–Dené family) has postpositions, the order of heads and genitive-like modifiers is swapped. However, there are known exceptions: Norwegian (Indo–European) has prepositions but genitives precede their syntactic heads.[2] Moreover, some typological features are rare whereas others are highly frequent. Interestingly, this also means that some languages are intuitively more plausible than others. Implications and frequencies of features are important, as they unravel the deeper explanatory factors underlying the patterns of cross-linguistic variation (Dryer 1998).

Cross-lingual variation can be found at all levels of linguistic structure. The seminal works on Linguistic Typology were concerned with morphosyntax, mainly morphological systems (Sapir 2014 [1921], page 128) and word order (Greenberg 1966b). This level of analysis deals with the form of meaningful elements (morphemes and words) and their combination, hence it is called *structural* typology. As an example, consider the alignment of the nominal case system (Dixon 1994): Some languages like Nenets (Uralic) use the same case for subjects of both transitive and intransitive verbs, and a different one for objects (nominative–accusative alignment). Other languages like

---

2 Exception-less generalizations are known as **absolute universals**. However, properties that have been proposed as such are often controversial, because they are too vacuous or have been eventually falsified (Evans and Levinson 2009).

Lezgian (Northeast Caucasian) group together intransitive subjects and objects, and treat transitive subjects differently (ergative–absolutive alignment).

On the other hand, *semantic* typology studies languages at the semantic and pragmatic levels. This area was pioneered by anthropologists interested in kinship (d'Andrade 1995) and colors (Berlin and Kay 1969), and was expanded by studies on lexical classes (Dixon 1977). The main focus of semantic typology has been to categorize languages in terms of concepts (Evans 2011) in the lexicon, in particular with respect to the 1) granularity, 2) division (boundary location), and 3) membership criteria (grouping and dissection). For instance, consider the event expressed by *to open* (something). It lacks a precise equivalent in languages such as Korean, where similar verbs overlap in meaning only in part (Bowerman and Choi 2001). For instance, *ppaeda* means 'to remove an object from tight fit' (used, e.g., for drawers) and *pyeolchida* means 'to spread out a flat thing' (used, e.g., for hands). Moreover, in most expressions, the English verb encodes the resulting state of the event, whereas an equivalent verb in another language such as Spanish (*abrir*) rather expresses the manner of the event (Talmy 1991). Although variation in the categories is pervasive due to their partly arbitrary nature, it is constrained cross-lingually via shared cognitive constraints (Majid et al. 2007).

Similarities between languages do not always arise from language-internal dynamics but also from external factors. In particular, similarities can be inherited from a common ancestor (genealogical bias) or borrowed by contact with a neighbor (areal bias) (Bakker 2010). Owing to genealogical inheritance, there are features that are widespread within a family but extremely rare elsewhere (e.g., the presence of click phonemes in the Khoisan languages). As an example of geographic percolation, most languages in the Balkan area (Albanian, Bulgarian, Macedonian, Romanian, Torlakian) have developed, even without a common ancestor, a definite article that is put after its noun simply because of their close proximity.

Research in linguistic typology has sought to disentangle such factors and to integrate them into a single framework aimed at answering the question "what's where why?" (Nichols 1992). Language can be viewed as a hybrid biological and cultural system. The two components co-evolved in a twin track, developing partly independently and partly via mutual interaction (Durham 1991). The causes of cross-lingual variation can therefore be studied from two complementary perspectives—from the perspective of functional theories or event-based theories (Bickel 2015). The former theories involve cognitive and communicative principles (internal factors) and account for the origin of variation, whereas the latter ones emphasize the imitation of patterns found in other languages (external factors) and account for the propagation (or extinction) of typological features (Croft 1995, 2000).

Examples of functional principles include factors associated with language use, such as the frequency or processing complexity of a pattern (Cristofaro and Ramat 1999). Patterns that are easy or widespread become integrated into the grammar (Haspelmath 1999, inter alia). On the other hand, functional principles allow the speakers to draw similar inferences from similar contexts, leading to locally motivated pathways of diachronic change through the process known as **grammaticalization** (Greenberg 1966a, 1978; Bybee 1988). For instance, in the world's languages (including English) the future tense marker almost always originates from verbs expressing direction, duty, will, or attempt because they imply a future situation.

The diachronic and gradual origin of the changes in language patterns and the statistical nature of the universals explain why languages do not behave monolithically. Each language can adopt several strategies for a given construction and partly inconsistent semantic categories. In other words, typological patterns tend to be gradient.

For instance, the semantics of grammatical and lexical categories can be represented on continuous multi-dimensional maps (Croft and Poole 2008). Bybee and McClelland (2005) have noted how this gradience resembles the patterns learned by connectionist networks (and statistical machine learning algorithms in general). In particular, they argue that such architectures are sensitive to both local (contextual) information and general patterns, as well as to their frequency of use, similarly to natural languages.

Typological documentation is limited by the fact that the evidence available for each language is highly unbalanced and many languages are not even recorded in a written form.[3] However, large typological databases such as WALS (Dryer and Haspelmath 2013) nevertheless have an impressive coverage (syntactic features for up to 1,519 languages). Where such information can be usefully integrated in machine learning, it can provide an alternative form of guidance to manual construction of resources that are now largely lacking for low resource languages. We discuss the existing typological databases and the integration of their features into NLP models in sections 4 and 5.

3. Overview of Multilingual NLP

The scarcity of data and resources in many languages represents a major challenge for multilingual NLP. Many state-of-the-art methods rely on supervised learning, hence their performance depends on the availability of manually crafted data sets annotated with linguistic information (e.g., treebanks, parallel corpora) and/or lexical databases (e.g., terminology databases, dictionaries). Although similar resources are available for key tasks in a few well-researched languages, the majority of the world's languages lack them almost entirely. This gap cannot be easily bridged: The creation of linguistic resources is a time-consuming process and requires skilled labor. Furthermore, the immense range of possible tasks and languages makes the aim of a complete coverage unrealistic.

One solution to this problem explored by the research community abandons the use of annotated resources altogether and instead focuses on unsupervised learning. This class of methods infers probabilistic models of the observations given some latent variables. In other words, it unravels the hidden structures within unlabeled text data. Although these methods have been used extensively for multilingual applications (Snyder and Barzilay 2008; Vulić, De Smet, and Moens 2011; Titov and Klementiev 2012, inter alia), their performance tends to lag behind the more linguistically informed supervised learning approaches (Täckström, McDonald, and Nivre 2013). Moreover, they have been rarely combined with typological knowledge. For these reasons, we do not review them in this section.

Other promising ways to overcome data scarcity include transferring models or data from resource-rich to resource-poor languages (§ 3.1) or learning joint models from annotated examples in multiple languages (§ 3.2) in order to leverage language interdependencies. Early approaches of this kind have relied on universal, high-level delexicalized features, such as part of speech (PoS) tags and dependency relations. More recently, however, the incompatibility of (language-specific) lexica has been countered by mapping equivalent words into the same multilingual semantic space through representation learning (§ 3.3). This has enriched language transfer and multilingual joint

---

3 According to Lewis, Simons, and Fennig (2016), 34.4% of the world's languages are threatened, not transmitted to younger generations, moribund, nearly extinct, or dormant. Moreover, 34% of the world's languages are vigorous but have not yet developed a system of writing.
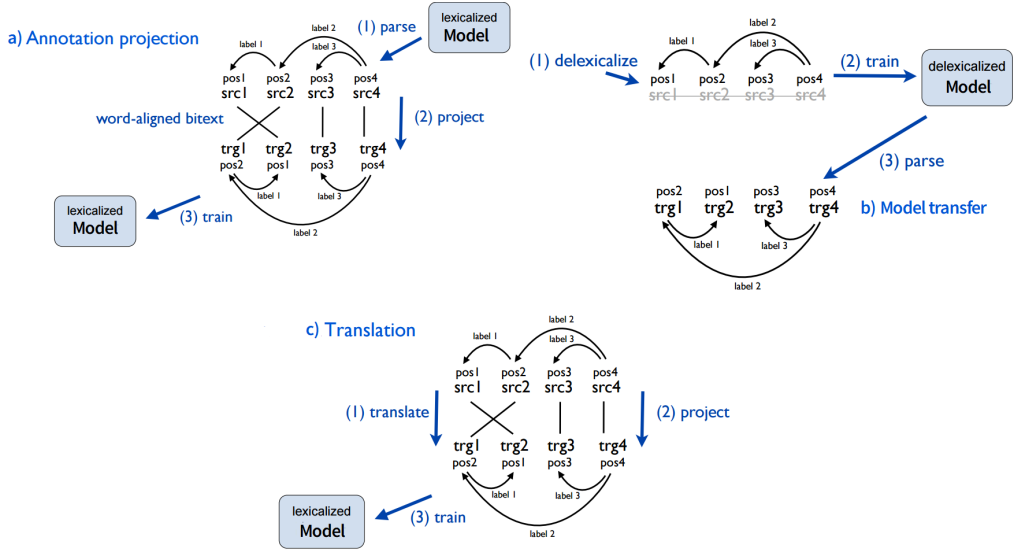
Figure 1
Three methods for language transfer: a) annotation projection, b) model transfer, and c) translation. The image has been adapted from Tiedemann (2015).

modeling with lexicalized features. In this section, we provide an overview of these methods, as they constitute the backbone of the typology-savvy algorithms surveyed in § 5.

3.1 Language Transfer

Linguistic information can be transferred from resource-rich languages to resource-poor languages; these are commonly referred to as source languages and target languages, respectively. Language transfer is challenging, as it requires us to match word sequences with different lexica and word orders, or syntactic trees with different (anisomorphic) structures (Ponti et al. 2018a). As a consequence, the information obtained from the source languages typically needs to be adapted, by tailoring it to the properties of the target languages. The methods developed for language transfer include annotation projection, (de)lexicalized model transfer, and translation (Agić et al. 2014). We illustrate them here using dependency parsing as an example.

Annotation projection was introduced in the seminal work of Yarowsky, Ngai, and Wicentowski (2001) and Hwa et al. (2005). In its original formulation, as illustrated in Figure 1(a), a source text is parsed and word-aligned with a target parallel raw text. Its annotation (e.g., PoS tags and dependency trees) is then projected directly between corresponding words and used to train a supervised model on the target language. Later refinements to this process are known as *soft* projection, where constraints can be used to complement alignment, based on distributional similarity (Das and Petrov 2011) or constituent membership (Padó and Lapata 2009). Moreover, source model expectations on labels (Wang and Manning 2014; Agić et al. 2016) or sets of most likely labels (Khapra et al. 2011; Wisniewski et al. 2014) can be projected instead of single categorical labels.

These can constrain unsupervised models by reducing the divergence between the expectations on target labels and on source labels or supporting "ambiguous learning" on the target language, respectively.

Model transfer instead involves training a model (e.g., a parser) on a source language and applying it on a target language (Zeman and Resnik 2008), as shown in Figure 1(b). Due to their incompatible vocabularies, models are typically delexicalized prior to transfer and take language-independent (Nivre et al. 2016) or harmonized (Zhang et al. 2012) features as input. In order to bridge the vocabulary gap, model transfer was later augmented with multilingual Brown word clusters (Täckström, McDonald, and Uszkoreit 2012) or multilingual distributed word representations (see § 3.3).

Machine translation offers an alternative to lexicalization in absence of annotated parallel data. As shown in Figure 1(c), a source sentence is machine translated into a target language (Banea et al. 2008), or through a bilingual lexicon (Durrett, Pauls, and Klein 2012). Its annotation is then projected and used to train a target-side supervised model. Translated documents can also be used to generate multilingual sentence representations, which facilitate language transfer (Zhou, Wan, and Xiao 2016).

Some of these methods are hampered by their resource requirements. In fact, annotation projection and translation need parallel texts to align words and train translation systems, respectively (Agić, Hovy, and Søgaard 2015). Moreover, comparisons of state-of-the-art algorithms revealed that model transfer is competitive with machine translation in terms of performance (Conneau et al. 2018). Partly owing to these reasons, typological knowledge has been mostly harnessed in connection with model transfer, as we discuss in § 5.2. Moreover, typological features can guide the selection of the best source language to match to a target language for language transfer (Agić et al. 2016, inter alia), which benefits all the above-mentioned methods (see § 5.3).

3.2 Multilingual Joint Supervised Learning

NLP models can be learned jointly from the data in multiple languages. In addition to facilitating intrinsically multilingual applications, such as Neural Machine Translation and Information Extraction, this approach often surpasses language-specific monolingual models, as it can leverage more (although noisier) data (Ammar et al. 2016, inter alia). This is particularly true in scenarios where either a target or all languages are resource-lean (Khapra et al. 2011) or in code-switching scenarios (Adel, Vu, and Schultz 2013). In fact, multilingual joint learning improves over pure model transfer also in scenarios with limited amounts of labeled data in target language(s) (Fang and Cohn 2017).[4]

A key strategy for multilingual joint learning is **parameter sharing** (Johnson et al. 2017). More specifically, in state-of-the-art neural architectures, input and hidden representations can be either private (language-specific) or shared across languages. Shared representations are the result of tying the parameters of a network component across languages, such as word embeddings (Guo et al. 2016), character embeddings (Yang, Salakhutdinov, and Cohen 2016), hidden layers (Duong et al. 2015b), or the attention mechanism (Pappas and Popescu-Belis 2017). Figure 2 shows an example where all the components of a PoS tagger are shared between two languages (Bambara on the left and Warlpiri on the right). Parameter sharing, however, does not necessarily imply parameter identity: It can be enforced by minimizing the distance between parameters

---

4 This approach is also more cost-effective in terms of parameters (Pappas and Popescu-Belis 2017).
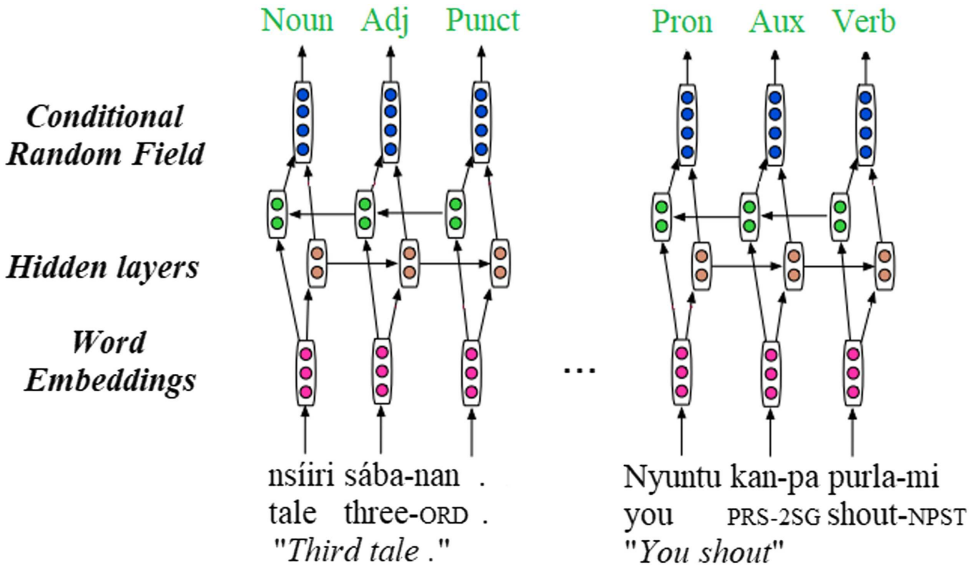
Figure 2
In multilingual joint learning, representations can be private or shared across languages. Tied parameters are shown as neurons with identical color. Image adapted from Fang and Cohn (2017), representing multilingual PoS tagging for Bambara (left) and Warlpiri (right).

(Duong et al. 2015a) or between latent representations of parallel sentences (Niehues et al. 2011; Zhou et al. 2015) in separate language-specific models.

Another common strategy in multilingual joint modeling is providing information about the properties of the language of the current text in the form of input **language vectors** (Guo et al. 2016). The intuition is that this helps tailoring the joint model toward specific languages. These vectors can be learned end-to-end in neural language modeling tasks (Tsvetkov et al. 2016; Östling and Tiedemann 2017) or neural machine translation tasks (Ha, Niehues, and Waibel 2016; Johnson et al. 2017). Ammar et al. (2016) instead used language vectors as a prior for language identity or typological features.

In § 5.2, we discuss ways in which typological knowledge is used to balance private and shared neural network components and provide informative input language vectors. In § 6.3, we argue that language vectors do not need to be limited to features extracted from typological databases, but should also include automatically induced typological information (Malaviya, Neubig, and Littell 2017, see § 4.3).

3.3 Multilingual Representation Learning

The multilingual algorithms reviewed in § 3.1 and § 3.2 are facilitated by dense real-valued vector representations of words, known as multilingual word embeddings. These can be learned from corpora and provide pivotal lexical features to several downstream NLP applications. In multilingual word embeddings, similar words (regardless of the actual language) obtain similar representations. Various methods to generate multilingual word embeddings have been developed. We follow the classification proposed by Ruder (2018), and we refer the reader to Upadhyay et al. (2016) for an empirical comparison.

Monolingual mapping generates independent monolingual representations and subsequently learns a linear map between a source language and a target language based on a bilingual lexicon (Mikolov, Le, and Sutskever 2013) or in an unsupervised fashion through adversarial networks (Conneau et al. 2017). Alternatively, both spaces can be cast into a new, lower-dimensional space through canonical correlation analysis based on dictionaries (Ammar et al. 2016) or word alignments (Guo et al. 2015).

Pseudo-cross-lingual approaches merge words with contexts of other languages and generate representations based on this mixed corpus. Substitutions are based on Wiktionary (Xiao and Guo 2014) or machine translation (Gouws and Søgaard 2015; Duong et al. 2016). Moreover, the mixed corpus can be produced by randomly shuffling words between aligned documents in two languages (Vulić and Moens 2015).

Cross-lingual training approaches jointly learn embeddings from parallel corpora and enforce cross-lingual constraints. This involves minimizing the distance of the hidden sentence representations of the two languages (Hermann and Blunsom 2014) or decoding one from the other (Lauly, Boulanger, and Larochelle 2013), possibly adding a correlation term to the loss (Chandar et al. 2014).

Joint optimization typically involves learning distinct monolingual embeddings, while enforcing cross-lingual constraints. These can be based on alignment-based translations (Klementiev, Titov, and Bhattarai 2012), cross-lingual word contexts (Luong, Pham, and Manning 2015), the average representations of parallel sentences (Gouws, Bengio, and Corrado 2015), or images (Rotman, Vulić, and Reichart 2018).

In this section, we have briefly outlined the most widely used methods in multilingual NLP. Although they offer a solution to data scarcity, cross-lingual variation remains a challenge for transferring knowledge across languages or learning from several languages simultaneously. Typological information offers promising ways to address this problem. In particular, we have noted that it can support model transfer, parameter sharing, and input biasing through language vectors. In the next two sections, we elaborate on these solutions. In particular, we review the development of typological information and the specific features that are selected for various NLP tasks (§ 4). Afterward, we discuss ways in which these features are integrated in NLP algorithms, for which applications they have been harnessed, and whether they truly benefit system performance (§ 5).

4. Selection and Development of Typological Information

In this section we first present major publicly available typological databases and then discuss how typological information relevant to NLP models is selected, pre-processed, and encoded. Finally, we highlight some limitations of database documentation with respect to coverage and feature granularity, and discuss how missing and finer-grained features can be obtained automatically.

4.1 Hand-Crafted Documentation in Typological Databases

Typological databases are created manually by linguists. They contain taxonomies of typological features, their possible values, as well as the documentation of feature values for the world's languages. Major typological databases, listed in Table 1, typically organize linguistic information in terms of universal features and language-specific values. For example, Figure 3 presents language-specific values for the feature *number of grammatical genders for nouns* on a world map. Note that each language is color-coded

Table 1
An overview of major publicly accessible databases of typological information. The databases are ordered by description level (and secondly by date of creation), along with their coverage. The table also provides feature examples: for each feature (in small capitals) we present two example languages with distinct feature values, and the total number of languages with each value in parenthesis (where applicable).

Name	Levels	Coverage	Feature Example
World Atlas of Language Structures (WALS)	Phonology, Morphosyntax, Lexical semantics	2,676 languages; 192 attributes; 17% values covered	ORDER OF OBJECT AND VERB Amele: OV (713) Gbaya Kara: VO (705)
Atlas of Pidgin and Creole Language Structures (APiCS)	Phonology, Morphosyntax	76 languages; 335 attributes	TENSE–ASPECT SYSTEMS Ternate Chabacano: purely aspectual (10) Afrikaans: purely temporal (1)
URIEL Typological Compendium	Phonology, Morphosyntax, Lexical semantics	8,070 languages; 284 attributes; ~439,000 values	CASE IS PREFIX Berber (Middle Atlas): yes (38) Hawaaian: no (993)
Syntactic Structures of the World's Languages (SSWL)	Morphosyntax	262 languages; 148 attributes; 45% values covered	STANDARD NEGATION IS SUFFIX Amharic: yes (21) Laal: no (170)
AUTOTYP	Morphosyntax	825 languages; ~1,000 attributes	PRESENCE OF CLUSIVITY !Kung (Ju): false Ik (Kuliak): true
Valency Patterns Leipzig (ValPaL)	Predicate–argument structures	36 languages; 80 attributes; 1,156 values	TO LAUGH Mandinka: 1 > V Sliammon: V.sbj[1] 1
Lyon–Albuquerque Phonological Systems Database (LAPSyD)	Phonology	422 languages; ~70 attributes	ɗ AND ʈ Sindhi: yes (1) Chuvash: no (421)
PHOIBLE Online	Phonology	2,155 languages; 2,160 attributes	m Vietnamese: yes (2053) Pirahã: no (102)
StressTyp2	Phonology	699 languages; 927 attributes	STRESS ON FIRST SYLLABLE Koromfé: yes (183) Cubeo: no (516)
World Loanword Database (WOLD)	Lexical semantics	41 languages; 24 attributes; ~2,000 values	HORSE Quechua: *kaballu* borrowed (24) Sakha: *sïlgï* no evidence (18)
Intercontinental Dictionary Series (IDS)	Lexical semantics	329 languages; 1,310 attributes	WORLD Russian: *mir* Tocharian A: *ārkiśoṣi*
Automated Similarity Judgment Program (ASJP)	Lexical semantics	7,221 languages; 40 attributes	I Ainu Maoka: *co7okay* Japanese: *watashi*
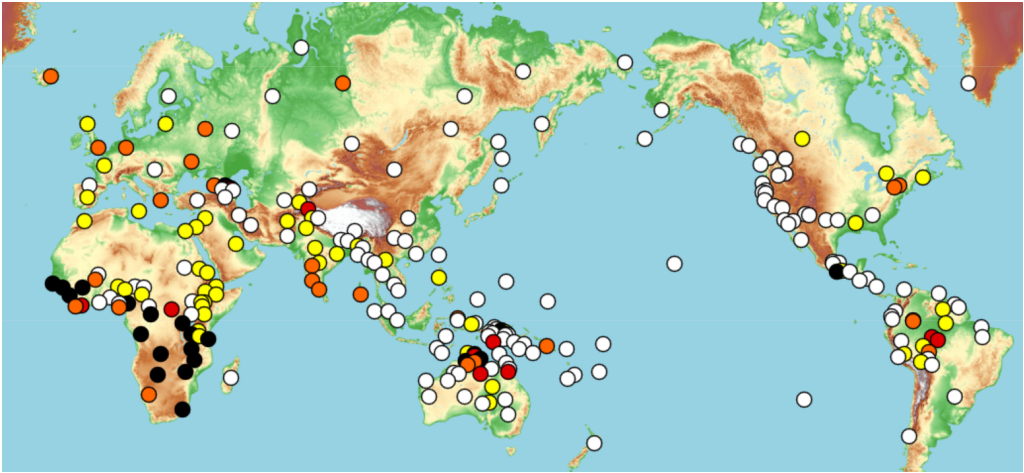
Figure 3
Number of grammatical genders for nouns in the world's languages according to WALS (Dryer and Haspelmath 2013): none (white), two (yellow), three (orange), four (red), five or more (black).

according to its value. Further examples for each database can be found in the rightmost column of Table 1.

Some databases store information pertaining to multiple levels of linguistic description. These include WALS (Dryer and Haspelmath 2013) and the Atlas of Pidgin and Creole Language Structures (APiCS) (Michaelis et al. 2013). Among all presently available databases, WALS has been the most widely used in NLP. In this resource, which has 142 typological features in total, features 1–19 deal with phonology, 20–29 with morphology, 30–57 with nominal categories, 58–64 with nominal syntax, 65–80 with verbal categories, 81–97 and 143–144 with word order, 98–121 with simple clauses, 122–128 with complex sentences, 129–138 with the lexicon, and 139–142 with other properties.

Other databases only cover features related to a specific level of linguistic description. For example, both Syntactic Structures of the World's Languages (SSWL) (Collins and Kayne 2009) and AUTOTYP (Bickel et al. 2017) focus on syntax. SSWL features are manually crafted, whereas AUTOTYP features are derived automatically from primary liguistic data using scripts. The Valency Patterns Leipzig (ValPaL) (Hartmann, Haspelmath, and Taylor 2013) provides verbs as attributes and predicate–argument structures as their values (including both valency and morphosyntactic constraints). For example, in both Mandinka and Sliammon, the verb *to laugh* has a valency of 1; in other words, it requires only one mandatory argument, the subject. In Mandinka the subject precedes the verb, but there is no agreement requirement; in Sliammon, on the other hand, the word order does not matter, but the verb is required to morphologically agree with the subject.

For phonology, the Phonetics Information Base and Lexicon (PHOIBLE) (Moran, McCloy, and Wright 2014) collates information on segments (binary phonetic features). In the Lyon–Albuquerque Phonological Systems Database (LAPSyD) (Maddieson et al. 2013), attributes are articulatory traits, syllabic structures, or tonal systems. Finally, StressTyp2 (Goedemans, Heinz, and der Hulst 2014) deals with stress and accent patterns. For instance, in Koromfé each word's first syllable has to be stressed, but not in Cubeo.

Other databases document various aspects of semantics. The World Loanword Database (WOLD) (Haspelmath and Tadmor 2009) documents loanwords by identifying the donor languages and the source words. The Automated Similarity Judgment Program (ASJP) (Wichmann, Holman, and Brown 2016) and the Intercontinental Dictionary Series (IDS) (Key and Comrie 2015) indicate how a meaning is lexicalized across languages: For example, the concept of WORLD is expressed as *mir* in Russian, and as *ārkiśoṣi* in Tocharian A.

Although typological databases store abundant information on many languages, they suffer from shortcomings that limit their usefulness. Perhaps the most significant shortcoming of such resources is their limited coverage. In fact, feature values are missing for most languages in most databases. Other shortcomings are related to feature granularity. In particular, most databases fail to account for feature value variation within each language: They report only majority value rather than the full range of possible values and their corresponding frequencies. For example, the dominant adjective–noun word order in Italian is adjective before noun; however, the opposite order is also attested. The latter information is often missing from typological databases.

Further challenges are posed by restricted feature applicability and feature hierarchies. Firstly, some features apply, by definition, only to subsets of languages that share another feature value. For instance, WALS feature 113A documents "Symmetric and Asymmetric Standard Negation," whereas WALS feature 114A "Subtypes of Asymmetric Standard Negation." Although a special NA value is assigned for symmetric-negation languages in the latter, there are cases where languages without the prerequisite feature are simply omitted from the sample. Secondly, features can be partially redundant, and subsume other features. For instance, WALS feature 81A "Order of Subject, Object and Verb" encodes the same information as WALS feature 82A "Order of Subject and Verb" and 83A "Order of Object and Verb," with the addition of the order of subject and object.

4.2 Feature Selection from Typological Databases

The databases presented above can serve as a rich source of typological information for NLP. In this section, we survey the feature sets that have been extracted from these databases in typologically informed NLP studies. In § 5.4, we review in which ways and to what degree of success these features have been integrated in machine learning algorithms.

Most NLP studies informed by typology only incorporated a subset of word order features from WALS (Dryer and Haspelmath 2013). Most of these studies focused on the task of syntactic dependency parsing, where word order provides crucial guidance (Naseem, Barzilay, and Globerson 2012), using the feature subsets shown in Figure 4. As depicted in the figure, these studies utilized quite similar word order features. The feature set first established by Naseem, Barzilay, and Globerson (2012) served as inspiration for subsequent works. The main differences in these sets results from the practice of discarding features that are not discriminative, when they are identical for all the languages in the sample.

Another group of studies used more comprehensive feature sets. The feature set of Daiber, Stanojević, and Sima'an (2016) included not only WALS word order features but also nominal categories (e.g., "Conjunctions and Universal Quantifiers") and nominal syntax (e.g., "Possessive Classification"). Berzak, Reichart, and Katz (2015) considered all features from WALS associated with morphosyntax and pruned out the redundant ones, resulting in a total of 119 features. Søgaard and Wulff (2012) utilized all the
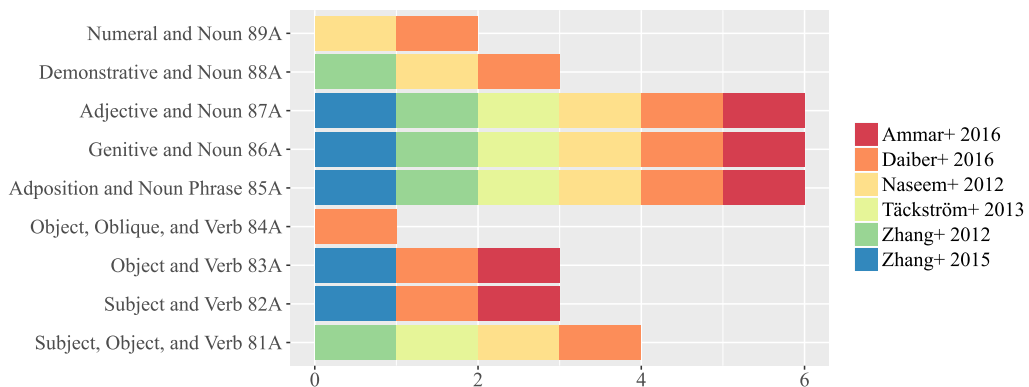
Figure 4
Feature sets used in a sample of typologically informed experiments for dependency parsing.
The numbers refer to WALS ordering (Dryer and Haspelmath 2013).

features in WALS with the exception of phonological features. Tsvetkov et al. (2016)
selected 190 binarized phonological features from URIEL (Littel, Mortensen, and Levin
2016). These features encoded the presence of single segments, classes of segments,
minimal contrasts in a language inventory, and the number of segments in a class. For
instance, they record whether a language allows two sounds to differ only in voicing,
such as /t/ and /d/.

Finally, a small number of experiments adopted the entire feature inventory of
typological databases, without any sort of pre-selection. In particular, Agić (2017) and
Ammar et al. (2016) extracted all the features in WALS, whereas Deri and Knight
(2016) extracted all the features in URIEL. Schone and Jurafsky (2001) did not resort
to basic typological features, but rather to "several hundred [implicational universals]
applicable to syntax" drawn from the Universal Archive (Plank and Filiminova 1996).

Typological attributes that are extracted from typological databases are typically
represented as feature vectors in which each dimension encodes a feature value. This
feature representation is often binarized (Georgi, Xia, and Lewis 2010): For each pos-
sible value v of each database attribute a, a new feature is created with value 1 if it
corresponds to the actual value for a specific language and 0 otherwise. Note that this
increases the number of features by a factor of $\frac{1}{||a||} \sum_{i=1}^{||a||} ||v_{a_i}||$. Although binarization
helps harmonizing different features and different databases, it muddies the different
types of typological variables.

To what extent do the limitations of typological databases mentioned in § 4.1
affect the feature sets surveyed in this section? The coverage is generally broad for the
languages used in these experiments, as they tend to be well-documented. For instance,
on average, 79.8% of the feature values are populated for the 14 languages appearing in
Berzak, Reichart, and Katz (2015), as opposed to 17% for all the languages in WALS.

It is hard to assess at a large scale how informative a set of typological features is.
However, these can be meaningfully compared with genealogical information. Ideally,
these two properties should not be completely equivalent (otherwise they would be
redundant),[5] but at the same time they should partly overlap (language cognates inherit

---

5 This does not apply to isolates, however: by definition, no genealogical information is available for these
  languages. Hence, typology is the only source of information about their properties.

(a) Word-order features

(b) All WALS features with average genus values
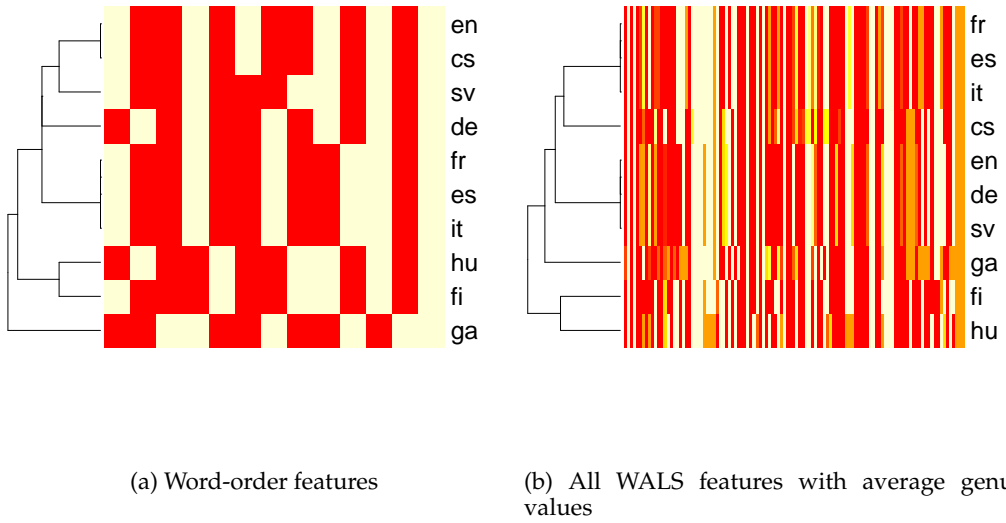
Figure 5
Heat maps of encodings for different subsets of typological WALS features taken from Ammar et al. (2016): rows stand for languages, dimensions for attributes, and color intensities for feature values. Encodings are clustered hierarchically by similarity. The meaning of language codes is: DE German, CS Czech, EN English, ES Spanish, FR French, FI Finnish, GA Irish Gaelic, HU Hungarian, IT Italian, SV Swedish.

typological properties from the same ancestors). In Figure 5, we show two feature sets appearing in Ammar et al. (2016), each depicted as a heatmap. Each row represents a language in the data; each cell is colored according to the feature value, ranging from 0 to 1. In particular, the feature set of Figure 5(a) is the subset of word order features listed in Figure 4; and Figure 5(b) is a large set of WALS features, where values are averaged by language genus to fill in missing values.

In order to compare the similarities of the typological feature vectors among languages, we clustered languages hierarchically based on such vectors.[6] Intuitively, the more this hierarchy resembles their actual family tree, the more redundant the typological information is. This is the case for Figure 5(b), where the lowest-lever clusters correspond exactly to a genus or family (top–down: Romance, Slavic, Germanic, Celtic, Uralic). Still, the language vectors belonging to the same cluster display some microvariations in individual features. On the other hand, 5(a) shows clusters differing from language genealogy: for instance, English and Czech are merged, although they belong to different genera (Germanic and Slavic). However, this feature set fails to account for fine-grained differences among related languages: For instance, French, Spanish, and Italian receive the same encoding.[7]

To sum up, this section's survey on typological feature sets reveals that most experiments have taken into account a small number of databases and features therein. However, several studies did utilize a larger set of coherent features or full databases.

---

6 Clustering was performed through the complete linkage method.
7 Notwithstanding they have different *preferences* over word orders (Liu 2010).

Although well-documented languages do not suffer much from coverage issues, we showed how difficult it is to select typological features that are non-redundant with genealogy, fully discriminative, and informative. The next section addresses these problems, proposing automatic prediction as a solution.

4.3 Automatic Prediction of Typological Features

The partial coverage and coarse granularity of existing typological resources sparked a line of research on automatic acquisition of typological information. Missing feature values can be predicted based on: i) heuristics from morphosyntactic annotation that pre-exists, such as in treebanks, or is transferred from aligned texts (§ 4.3.1); ii) unsupervised propagation from other values in a database based on clustering or language similarity metrics (§ 4.3.2); iii) supervised learning with Bayesian models or artificial neural networks (§ 4.3.3); or iv) heuristics based on co-occurrence metrics, typically applied to multi-parallel texts (§ 4.3.4). These strategies are summarized in Table 2.

With the exception of Naseem, Barzilay, and Globerson (2012), who treated typological information as a latent variable, automatically acquired typological features have not been integrated into algorithms for NLP applications to date. However, they have several advantages over manually crafted features. Unsupervised propagation and supervised learning fill in missing values in databases, thereby extending their coverage. Moreover, heuristics based on morphosyntactic annotation and co-occurrence metrics extract additional information that is not recorded in typological databases. Further, they can account for the distribution of feature values within single languages, rather than just the majority value. Finally, they do not make use of discrete cross-lingual categories to compare languages; rather, language properties are reflected in continuous representations, which is in line with their gradient nature (see § 2).

4.3.1 Heuristics Based on Morphosyntactic Annotation. Morphosyntactic feature values can be extracted via heuristics from morphologically and syntactically annotated texts. For example, word order features can be calculated by counting the average direction of dependency relations or constituency hierarchies (Liu 2010). Consider the tree of a sentence in Welsh from Bender et al. (2013) in Figure 6. The relative order of verb–subject, and verb–object can be deduced from the position of the relevant nodes *VBD*, NN_S, and NN_O (highlighted).

Morphosyntactic annotation is often unavailable for resource-lean languages. In such cases, it can be projected from a source language to a target language through language transfer. For instance, Östling (2015) projects source morphosyntactic annotation directly to several languages through a multilingual word alignment. After the alignment and projection, word order features are calculated by the average direction of dependency relations. Similarly, Zhang et al. (2016) transfer PoS annotation with a model transfer technique relying on multilingual embeddings, created through monolingual mapping (see § 3.3). After the projection, they predict feature values with a multiclass support vector machine using PoS tag *n*-gram features.

Finally, typological information can be extracted from Interlinear Glossed Texts (IGT). Such collections of example sentences are collated by linguists and contain grammatical glosses with morphological information. These can guide alignment between the example sentence and its English translation. Lewis and Xia (2008) and Bender et al. (2013) project chunking information from English and train context free grammars on target languages. After collapsing identical rules, they arrange them by frequency and infer word order features.

Table 2
An overview of the strategies for prediction of typological features.

|	Author	Details	Requirements	Languages	Features
Morphosyntactic Annotation	Liu (2010)	Treebank count	Treebank	20	word order
	Lewis and Xia (2008)	IGT projection	IGT, source chunker	97	word and morpheme order, determiners
	Bender et al. (2013)	IGT projection	IGT, source chunker	31	word order and case alignment
	Östling (2015)	Treebank projection	Parallel text, source tagger and parser	986	word order
	Zhang et al. (2016)	PoS projection	source tagger, seed dictionary	6	word order
Unsupervised Propagation	Teh, Daumé III, and Roy (2007)	Hierarchical typological cluster	WALS	2,150	whole
	Georgi, Xia, and Lewis (2010)	Majority value from k-means typological cluster	WALS	whole	whole
	Coke, King, and Radev (2016)	Majority value from genus	Genealogy and WALS	325	word order and passive
	Littel, Mortensen, and Levin (2016)	Family, area, and typology-based Nearest Neighbors	Genealogy and WALS	whole	whole
	Berzak, Reichart, and Katz (2014)	English as a Second Language–based Nearest Neighbors	ESL texts	14	whole
	Malaviya, Neubig, and Littell (2017)	Task-based language vector	NMT data set	1,017	whole
	Bjerva and Augenstein (2018)	Task-based language vector	PoS tag data set	27,824	phonology, morphology, syntax
Supervised Learning	Takamura, Nagata, and Kawasaki (2016)	Logistic regression	WALS	whole	whole
	Murawaki (2017)	Bayesian + feature and language interactions	Genealogy and WALS	2,607	whole
	Wang and Eisner (2017)	Feed-forward Neural Network	WALS, tagger, synthetic treebanks	37	word order
	Cotterell and Eisner (2017)	Determinant Point Process with neural features	WALS	200	vowel inventory
	Daumé III and Campbell (2007)	Implication universals	Genealogy and WALS	whole	whole
	Lu (2013)	Automatic discovery	Genealogy and WALS	1,646	word order
Cross-lingual distribution	Wälchli and Cysouw (2012)	Sentence edit distance	Multi-parallel texts, pivot	100	motion verbs
	Asgari and Schütze (2017)	Pivot alignment	Multi-parallel texts, pivot	1,163	tense markers
	Roy et al. (2014)	Correlations in counts and entropy	None	23	adposition word order
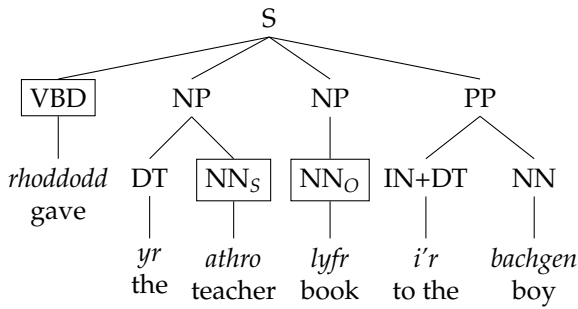
Figure 6
Constituency tree of a Welsh sentence.

4.3.2 Unsupervised Propagation. Another line of research seeks to increase the coverage of typological databases borrowing missing values from the known values in other languages. One approach is clustering languages according to some criterion and propagating the majority value within each cluster. Hierarchical clusters can be created either according to typological features (e.g., Teh, Daumé III, and Roy 2007) or based on language genus (Coke, King, and Radev 2016). Through extensive evaluation, Georgi, Xia, and Lewis (2010) demonstrate that typology based clustering outperforms genealogical clustering for unsupervised propagation of typological features. Among the clustering techniques examined, *k*-means appears to be the most reliable as compared to *k*-medoids, the Unweighted Pair Group Method with Arithmetic mean, repeated bisection, and hierarchical methods with partitional clusters.

Language similarity measures can also rely on a distributed representation of each language. These language vectors are trained end-to-end as part of neural models for downstream tasks such as many-to-one Neural Machine Translation (NMT). In particular, language vectors can be obtained from artificial trainable tokens concatenated to every input sentence, similar to Johnson et al. (2017), or from the aggregated values of the hidden states of a neural encoder. Using these language representations, typological feature values are propagated using *k* nearest neighbors (Bjerva and Augenstein 2018) or predicted with logistic regression (Malaviya, Neubig, and Littell 2017).

Language vectors can be conceived as data-driven, continuous typological representations of a language, and as such provide an alternative to manually crafted typological representations. Similar to the analysis carried out in § 4.2, we can investigate how much language vectors align with genealogical information. Figure 7 compares continuous representations based on artificial tokens (Figure 7(a)) and encoder hidden states (Figure 7(b)) with vectors of discrete WALS features from URIEL (Figure 7(c)). All the representations are reduced to two dimensions with *t*-Distributed Stochastic Neighbor Embedding (*t*-SNE), and color-coded based on their language family.

As the plots demonstrate, the information encoded in WALS vectors is akin to genealogical information, partly because of biases introduced by family-based propagation of missing values (Littel, Mortensen, and Levin 2016) (see § 4.3.2). On the other hand, artificial tokens and encoder hidden states cannot be reduced to genealogical clusters. Yet, their ability to predict missing values is not inferior to WALS features (as detailed in § 4.3.5). This implies that discrete and continuous representations appear to capture different aspects of the cross-lingual variation, while both being informative. For this reason, they are possibly complementary and could be combined in the future.

(a) Input vectors   (b) Cell states   (c) Predicted WALS
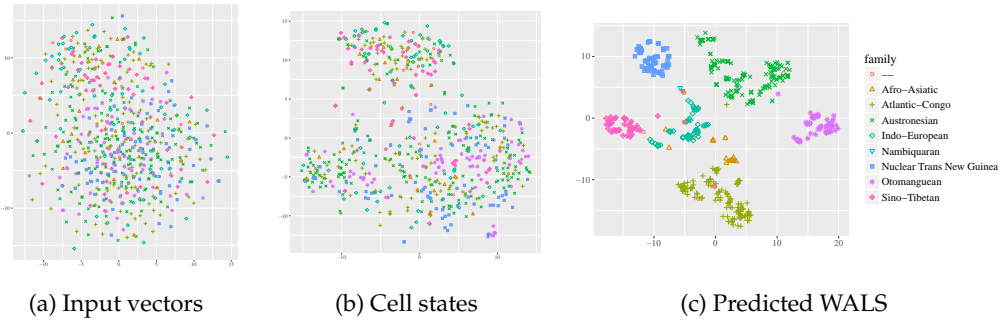
Figure 7
Language representations dimensionality-reduced with *t*-SNE.

4.3.3 Supervised Learning. As an alternative to unsupervised propagation, one can learn an explicit model for predicting feature values through supervised classification. For instance, Takamura, Nagata, and Kawasaki (2016) use logistic regression with WALS features and evaluate this model in a cross-validation setting where one language is held out in each fold. Wang and Eisner (2017) provide supervision to a feed-forward neural network with windows of PoS tags from natural and synthetic corpora.

Supervised learning of typology can also be guided by non-typological information (see § 2). Within the Bayesian framework, Murawaki (2017) exploits not only typological but also genealogical and areal dependencies among languages to represent each language as a binary latent parameter vector through a series of autologistic models. Cotterell and Eisner (2017, 2018) develop a point-process generative model of vowel inventories (represented as either IPA symbols or acoustic formants) based on some universal cognitive principles: dispersion (phonemes are as spread out as possible in the acoustic space) and focalization (some positions in the acoustic space are preferred due to the similarity of the main formants).

An alternative approach to supervised prediction of typology is based on learning implicational universals of the kind pioneered by Greenberg (1963), with probabilistic models from existing typological databases. Using such universals, features can be deduced by *modus ponens*. For instance, once it has been established that the presence of "High consonant/vowel ratio" and "No front-rounded vowels" implies "No tones," the latter feature value can be deduced from the premises if those are known. Daumé III and Campbell (2007) propose a Bayesian model for learning typological universals that predicts implications between features based on the intuition that their likelihood does not equal their prior probability, but rather is constrained by other features. Lu (2013) casts this problem as knowledge discovery, where language features are encoded in a directed acyclic graph. The strength of implication universals is represented as weights associated with the edges of this graph.

4.3.4 Heuristics Based on Cross-Lingual Distributional Information. Typological features can also emerge in a data-driven fashion, based on distributional information from multi-parallel texts. Wälchli and Cysouw (2012) create a matrix where each row is a parallel sentence, each column is a language, and cell values are lemmas of motion verbs occurring in those sentences. This matrix can be transformed to a (Hamming) distance matrix
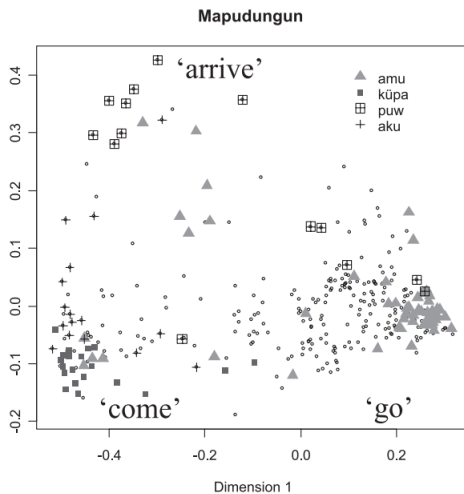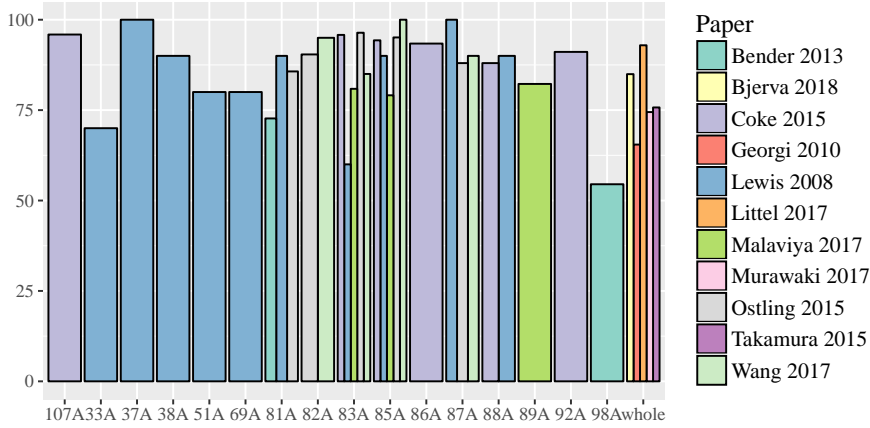
Figure 8
Wälchli and Cysouw's (2012) cross-lingual sentence visualization for Mapungundun. In the top-right corner is the legend of the motion verbs taken into consideration. Each data point is an instance of a verb in a sentence, positioned according to its contextualized sense. English glosses are the authors' interpretations of the main clusters.

between sentence pairs, and reduced to lower dimensionality via multidimensional scaling. This provides a continuous map of lexical semantics that is language-specific, but motivated by categories that emerge across languages. For instance, Figure 8 shows the first two dimensions of the multidimensional scaled similarity matrix in Mapudungun, where the first dimension can be interpreted as reflecting the direction of motion.

Asgari and Schütze (2017) devised a procedure to obtain markers of grammatical features across languages. Initially, they manually select a language containing an unambiguous and overt marker for a specific typological feature (called head pivot) based on linguistic expertise. For instance, *ti* in Seychellois Creole (French Creole) is a head pivot for past-tense marking. Then, this marker is connected to equivalent markers in other languages through alignment-based χ^2 tests in a multi-parallel corpus and n-gram counts.

Finally, typological features can be derived from raw texts in a completely unsupervised fashion, without multi-parallel texts. Roy et al. (2014) use heuristics to predict the order of adpositions and nouns. Adpositions are identified as the most frequent words. Afterward, the position of the noun is established based on whether selectional restrictions appear on the right context or the left context of the adposition, according to count-based and entropy-based metrics.

4.3.5 Comparison of the Strategies. Establishing which of the above-mentioned strategies is optimal in terms of prediction accuracy is not straightforward. In Figure 9, we collect the scores reported by several of the surveyed papers, provided that they concern specific features or the whole WALS data set (as opposed to subsets) and are numerical (as opposed to graphical plots). However, these results are not strictly comparable, because language samples and/or the split of data partitions may differ. The lack of standardization in this respect allows us to draw conclusions only about the difficulty

Figure 9
Accuracy of different approaches (see legend on the right) in predicting missing values of WALS typological features (specified on the vertical axis).

of predicting each feature relative to a specific strategy: For instance, the correct value of passive voice is harder to predict than word order, as claimed by Bender et al. (2013) and seen in Figure 9.

However, some papers carry out comparisons of the different strategies within the same experimental setting. According to Coke, King, and Radev (2016), propagation from the genus majority value outperforms logistic regression among word-order typological features. On the other hand, Georgi, Xia, and Lewis (2010) argue that typology-based clusters are to be preferred in general. This apparent contradiction stems from the nature of the target features: Genealogy excels in word order features because of their diachronic stability. As they tend to be preserved over time, they are often shared by all members of a family. In turn, majority value propagation is surpassed by supervised classification when evaluated on the entire WALS feature set (Takamura, Nagata, and Kawasaki 2016).

In general, there appears to be no "one-size-fits-all" algorithm. For instance, Coke, King, and Radev (2016) outperform Wang and Eisner (2017) for object–verb order (83A) but are inferior to it for adposition–noun (85A). In fact, each strategy is suited for different features, and requires different resources. Based on Figure 9, the extraction of information from morphosyntactic annotation is well suited for word order features, whereas distributional heuristics from multi-parallel texts are more informative about lexicalization patterns. On the other hand, unsupervised propagation and supervised learning are general-purpose strategies. Moreover, the first two presuppose some annotated and/or parallel texts, whereas the second two need pre-existing database documentation. Strategies may be preferred, according to which resources are available for a specific language.

Many strategies have a common weakness, however, as they postulate incorrectly that language samples are independent and identically distributed (Lu 2013; Cotterell and Eisner 2017). This is not the case, due to the interactions of family, area, and implicational universals. The solutions adopted to mitigate this weakness vary: Wang and Eisner (2017) balance the data distribution with synthetic examples, whereas Takamura,

Nagata, and Kawasaki (2016) model family and area interactions explicitly. However, according to Murawaki (2017), these interactions have different degrees of impact on typological features. In particular, inter-feature dependencies are more influential than inter-language dependencies, and horizontal diffusibility (borrowing from neighbors) is more prominent than vertical stability (inheriting from ancestors).

Finally, a potential direction for future investigation emerges from this section's survey. In addition to missing value completion, automatic prediction often also accounts for the variation internal to each language. However, some strategies go even further, and "open the way for a typology where generalizations can be made without there being any need to reduce the attested diversity of categorization patterns to discrete types" (Wälchli and Cysouw 2012). In fact, language vectors (Malaviya, Neubig, and Littell 2017; Bjerva and Augenstein 2018) and distributional information from multi-parallel texts (Asgari and Schütze 2017) are promising insofar they capture latent properties of languages in a bottom–up fashion, preserving their gradient nature. This offers an alternative to hand-crafted database features: In § 6.3 we make a case for integrating continuous, data-driven typological representations into NLP algorithms.

5. Uses of Typological Information in NLP Models

The typological features developed as discussed in § 4 are of significant importance for NLP algorithms. Particularly, they are used in three main ways. First, they can be manually converted into rules for expert systems (§5.1); second, they can be integrated into algorithms as constraints that inject prior knowledge or tie together specific parameters across languages (§ 5.2); and, finally, they can guide data selection and synthesis (§ 5.3). All of these approaches are summarized in Table 3 and described in detail in the following sections, with a particular focus on the second approach.

5.1 Rule-Based Systems

An interesting example of a rule-based system in our context is the Grammar Matrix kit, presented by Bender (2016), where rule-based grammars can be generated from typological features. These grammars are designed within the framework of Minimal Recursion Semantics (Copestake et al. 2005) and can parse a natural language input string into a semantic logical form.

The Grammar Matrix consists of a universal core grammar and language-specific libraries for phenomena where typological variation is attested. For instance, the module for coordination typology expects the specification of the kind, pattern, and position of a grammatical marking, as well as the phrase types it covers. For instance, the Ono language (Trans–New Guinea) expresses it with a lexical, monosyndetic, pre-nominal marker *so* in noun phrases. A collection of pre-defined grammars is available through the Language CoLLAGE initiative (Bender 2014).

5.2 Feature Engineering and Constraints

The most common usage of typological features in NLP is in feature engineering and constraint design for machine learning algorithms. Two popular approaches we consider here are *language transfer with selective sharing*, where the parameters of languages with similar typological features are tied together (§ 5.2.1), and *joint multilingual learning*, where typological information is used in order to bias models to reflect the properties of specific languages (see § 5.2.2).

Table 3
An overview of the approaches to use typological features in NLP models.

	Author	Details	Number of Languages / Families	Task
Rules	Bender (2016)	Grammar generation	12 / 8	semantic parsing
Feature engineering	Naseem, Barzilay, and Globerson (2012)	Generative	17 / 10	syntactic parsing
	Täckström, McDonald, and Nivre (2013)	Discriminative graph-based	16 / 7	syntactic parsing
	Zhang and Barzilay (2015)	Discriminative tensor-based	10 / 4	syntactic parsing
	Daiber, Stanojević, and Sima'an (2016)	One-to-many MLP	22 / 5	reordering for machine translation
	Ammar et al. (2016)	Multi-lingual transition-based	7 / 1	syntactic parsing
	Tsvetkov et al. (2016)	Phone-based polyglot language model	9 / 4	identification of lexical borrowings and speech synthesis
	Schone and Jurafsky (2001)	Design of Bayesian network	1 / 1	word cluster labeling
Data Manipulation	Deri and Knight (2016)	Typology-based selection	227	grapheme to phoneme
	Agić (2017)	PoS divergence metric	26 / 5	syntactic parsing
	Søgaard and Wulff (2012)	Typology-based weighing	12 / 1	syntactic parsing
	Wang and Eisner (2017)	Word-order-based tree synthesis	17 / 7	syntactic parsing
	Ponti et al. (2018a)	Construction-based tree preprocessing	6 / 3	machine translation, sentence similarity

5.2.1 Selective sharing. This framework was introduced by Naseem, Barzilay, and Globerson (2012) and was subsequently adopted by Täckström, McDonald, and Nivre (2013) and Zhang and Barzilay (2015). It aims at parsing sentences in a language transfer setting (see § 3.1) where there are multiple source languages and a single unobserved target language. It assumes that head–modifier relations between PoS pairs are universal, but the order of parts of speech within a sentence is language-specific. For instance, adjectives always modify nouns, but in Igbo (Niger–Congo) they linearly precede nouns, and in Nihali (isolate) they follow nouns. Leveraging this intuition, selective sharing models learn dependency relations from all source languages, while ordering is learned from typologically related languages only.

    Selective sharing was originally implemented in a generative framework, factorizing the recursive generation of dependency tree fragments into two steps (Naseem, Barzilay, and Globerson 2012). The first one is universal: The algorithm selects an

unordered (possibly empty) set of modifiers $\{M\}$ given a head h with probability $P(\{M\}|h)$, where both the head and the modifiers are characterized by their PoS tags. The second step is language-specific: Each dependent m is assigned a direction d (left or right) with respect to h based on the language l, with probability $P(d|m,h,l)$. Dependents in the same direction are eventually ordered with a probability drawn from a uniform distribution over their possible unique permutations. The total probability is then defined as follows:

$$P(n|h,\theta_1) \cdot \sigma_n \left(\sum_{m_i \in M} P(m_i|h,\theta_2) \right) \cdot \prod_{m_i \in M} \sigma\left(\mathbf{w} \cdot \mathbf{g}(m,h,l,\mathbf{f}_l)\right) \cdot \frac{1}{||M_R||\,||M_L||} \qquad (1)$$

In Equation (1), the first step is expressed as two factors: the estimation of the number n of modifiers, parametrized by θ_1, and the actual selection of modifiers, parametrized by θ_2, with the softmax function σ converting the n values into probabilities. The second step, overseeing the assignment of a direction to the dependencies, is parametrized by $\mathbf{w}$, which multiplies a feature function $\mathbf{g}()$, whose arguments include a typology feature vector $\mathbf{f}_l$. The values of all the parameters are estimated by maximizing the likelihood of the observations.

Täckström, McDonald, and Nivre (2013) proposed a discriminative version of the model, in order to amend the alleged limitations of the original generative variant. In particular, they dispose of the strong independence assumptions (e.g., between choice and ordering of modifiers) and invalid feature combinations. For instance, the WALS feature "Order of Subject, Verb, and Object" (81A) should be taken into account only when the head under consideration is a verb and the dependent is a noun, but in the generative model this feature was fed to $\mathbf{g}()$ regardless of the head–dependency pair. The method of Täckström, McDonald, and Nivre is a delexicalized, first-order, graph-based parser, based on a carefully selected feature set. From the set proposed by McDonald, Crammer, and Pereira (2005), they keep only (universal) features describing selectional preferences and dependency length. Moreover, they introduce (language-specific) features for the directionality of dependents, based on combinations of the PoS tags of the head and modifiers with corresponding WALS values.

This approach was further extended to tensor-based models by Zhang and Barzilay (2015), in order to avoid the shortcomings of manual feature selection. They induce a compact hidden representation of features and languages by factorizing a tensor constructed from their combination. The prior knowledge from the typological database enables the model to forbid the invalid interactions, by generating intermediate feature embeddings in a hierarchical structure. In particular, given n words and l dependency relations, each arc $h \rightarrow m$ is encoded as the tensor product of three feature vectors for heads $\Phi_h \in \mathbb{R}^n$, modifiers $\Phi_m \in \mathbb{R}^n$, and the arcs $\Phi_{h \rightarrow m} \in \mathbb{R}^l$. A score is obtained through the inner product of these and the corresponding r rank-1 dense parameter matrices for heads $H \in \mathbb{R}^{n \times r}$, dependents $M \in \mathbb{R}^{n \times r}$, and arcs $M \in \mathbb{R}^{l \times r}$. The resulting embedding is subsequently constrained through a summation with the typological features $T_u \phi_{t_u}$:

$$S(h \xrightarrow{l} m) = \sum_{i=1}^{r} [H_c \phi_{h_c}]_i [M_c \phi_{m_c}]_i \odot$$
$$\{[T_l \phi_{t_l}]_i + [L\phi_l]_i \odot \qquad\qquad (2)$$
$$\left([T_u \phi_{t_u}]_i + [H\phi_h]_i [M\phi_m]_i [D\phi_d]_i\right)\}$$

Equation (2) shows how the overall score of a labeled dependency is enriched (by element-wise product) with (1) the features and parameters for arc labels $L\phi_l$ constrained by the typological vector $T_l\phi_{t_l}$; and (2) features and parameters for head contexts $H_c\phi_{h_c}$ and dependent contexts $M_c\phi_{m_c}$. This loss is optimized within a maximum soft-margin objective through online passive–aggressive updates.

The different approaches to selective sharing presented here explicitly deal with cases where the typological features do not match any of the source languages, which may lead learning astray. Naseem, Barzilay, and Globerson (2012) propose a variant of their algorithm where the typological features are not observed (in WALS), treating them as latent variables, and where model parameters are learned in an unsupervised fashion with the Expectation Maximization algorithm (Dempster, Laird, and Rubin 1977). Täckström, McDonald, and Nivre (2013) tackle the same problem from the side of ambiguous learning. The discriminative model on the target language is trained on sets of automatically predicted ambiguous labels $\hat{\mathbf{y}}$. Finally, Zhang and Barzilay (2015) utilize semi-supervised techniques, where only a handful of annotated examples from the target language is available.

5.2.2 Multi-lingual Biasing. Some papers leverage typological features to gear the shared parameters of a joint multilingual model toward the properties of a specific language. Daiber, Stanojević, and Sima'an (2016) develop a reordering algorithm that estimates the permutation probabilities of aligned word pairs in multi-lingual parallel texts. The best sequence of permutations is inferred via *k*-best graph search in a finite state automaton, producing a lattice. This algorithm, which receives lexical, morphological, and syntactic features of the source word pairs and typological features of the target language as input, was shown to benefit a downstream machine translation task.

The joint multilingual parser of Ammar et al. (2016) shares hidden-layer parameters across languages, and combines both language-invariant and language-specific features in its copious lexicalized input feature set. This transition-based parser selects the next action z (e.g., SHIFT) from a pool of possible actions given its current state $\mathbf{p}_t$, as defined in Equation (3):

$$P(z|\mathbf{p}_t) = \sigma(\mathbf{g_z}^\top \max(\mathbf{0}, \mathbf{W}\,\mathbf{s}_t \oplus \mathbf{b}_t \oplus \mathbf{a}_t \oplus \mathbf{l}_{it} + \mathbf{b}) + q_z) \tag{3}$$

$P(z|\mathbf{p}_t)$ is defined in terms of a set of iteratively manipulated, densely represented data structures: a buffer $\mathbf{b}_t$, a stack $\mathbf{s}_t$, and an action history $\mathbf{a}_t$. The hidden representation of these modules are the output of stack-LSTMs, which are in turn fed with input word feature representations (stack and buffer) and action representations (history). The shared parameters are biased toward a particular language through language embeddings $\mathbf{l}_{it}$. The language embeddings consist of (a non-linear transformation of) either a mere one-hot identity vector or a vector of typological properties taken from WALS. In particular, they are added to both input feature and action vectors, to affect the three above-mentioned modules individually, and concatenated to the hidden module representations, to affect the entire parser state. The resulting state representation is propagated through an action-specific layer parametrized by $\mathbf{g}_t$ and q_t, and activated by a softmax function σ over actions.

Similarly, typological features have been used to bias input and hidden states of language models. For example, Tsvetkov et al. (2016) proposed a multilingual phoneme-level language model where an input phoneme $\mathbf{x}$ and a language vector ℓ at time t are linearly mapped to a local context representation and then passed to a global LSTM.

This hidden representation $\mathbf{G}_t^\ell$ is factored by a non-linear transformation of typological features t_ℓ, as shown in Equation (4):

$$\mathbf{G}_t^\ell = \text{LSTM}(W_{c_x}x_t + W_{c_\ell}x_\ell + b, \mathbf{g}_{t-1}) \otimes \tanh(W_\ell\, t_\ell + b_\ell)^\top \tag{4}$$

$$P(\phi_t|\phi_{<t}, \ell) = \sigma(\mathbf{W}\,\text{vec}(\mathbf{G}_t^\ell) + \mathbf{b}) \tag{5}$$

As described in Equation (5), $\mathbf{G}_t^\ell$ is then vectorized and mapped to a probability distribution of possible next phonemes ϕ_t. The phoneme vectors, learned by the language model in an end-to-end manner, were demonstrated to benefit two downstream applications: lexical borrowing identification and speech synthesis.

Moreover, typological features (in the form of implicational universals) can guide the design of Bayesian networks. Schone and Jurafsky (2001) assign part-of-speech labels to word clusters acquired in an unsupervised fashion. The underlying network is acyclic and directed, and is converted to a join-tree network to handle multiple parents (Jensen 1996). For instance, the sub-graph for the ordering of numerals and nouns is intertwined also with properties of adjectives and adpositions. The final objective maximizes the probability of a tag T_i and a feature set Φ_i, given the implicational universals U as $\text{argmax}_T P(\{\Phi_i, T_i\}_{i=1}^n|U)$.

5.3 Data Selection, Synthesis, and Preprocessing

Another way in which typological features are used in NLP is to guide data selection. This procedure is crucial for (1) language transfer methods, as it guides the choice of the most suitable source languages and examples; and (2) multilingual joint models, in order to weigh the contribution of each language and example. The selection is typically carried out through general language similarity metrics. For instance, Deri and Knight (2016) base their selection on the URIEL language typology database, considering information about genealogical, geographic, syntactic, and phonetic properties. This facilitates language transfer of grapheme-to-phoneme models, by guiding the choice of source languages and aligning phoneme inventories.

Metrics for source selection can also be extracted in a data-driven fashion, without explicit reference to structured taxonomies. For instance, Rosa and Zabokrtsky (2015) estimate the Kullback–Leibler divergence between PoS trigram distributions for delexicalized parser transfer. In order to approximate the divergence in syntactic structures between languages, Ponti et al. (2018a) utilize the Jaccard distance between morphological feature sets and the tree edit distance of delexicalized dependency parses of translationally equivalent sentences.

A priori and bottom–up approaches can also be combined. For delexicalized parser transfer, Agić (2017) relies on a weighted sum of distances based on (1) the PoS divergence defined by Rosa and Zabokrtsky (2015); (2) the character-based identity prediction of the target language; and (3) the Hamming distance from the target language typological vector. In fact, they have different weaknesses: Language identity (and consequently typology) fails to abstract away from language scripts. On the other hand, the accuracy of PoS-based metrics deteriorates easily in scenarios with scarce amounts of data.

Source language selection is a special case of source language weighting where weights are one-hot vectors. However, weights can also be gradient and consist of real numbers. Søgaard and Wulff (2012) adapt delexicalized parsers by weighting every

training instance based on the inverse of the Hamming distance between typological (or genealogical) features in source and target languages. An equivalent bottom–up approach is developed by Søgaard (2011), who weighs source language sentences based on the perplexity between their coarse PoS tags and the predictions of a sequential model trained on the target language.

Alternatively, the lack of target annotated data can be alleviated by synthesizing new examples, thus boosting the variety and amount of the source data. For instance, the Galactic Dependency Treebanks stem from real trees whose nodes have been permuted probabilistically, according to the word orders of nouns and verbs in other languages (Wang and Eisner 2016). Synthetic trees improve the performance of model transfer for parsing when the source is chosen in a supervised way (performance on target development data) and in an unsupervised way (coverage of target PoS sequences).

Rather than generating new synthetic data, Ponti et al. (2018a) leverage typological features to pre-process treebanks in order to reduce their variation in language transfer tasks. In particular, they adapt source trees to the typology of a target language with respect to several constructions in a rule-based fashion. For instance, relative clauses in Arabic (Afro–Asiatic) with an indefinite antecedent drop the relative pronoun, which is mandatory in Portuguese (Indo–European). Hence, the pronoun has to be added, or deleted in the other direction. Feeding pre-processed syntactic trees to lexicalized syntax-based neural models, such as feature-based recurrent encoders (Sennrich and Haddow 2016) or TreeLSTMs (Tai, Socher, and Manning 2015), achieves state-of-the-art results in Neural Machine Translation and cross-lingual sentence similarity classification.

5.4 Comparison

In light of the performance of the described methods, to what extent can typological features benefit downstream NLP tasks and applications? To answer this key question, consider the performance scores of each model reported in Figure 10. Each model has been evaluated in the original paper in one (or more) of the three main settings, with otherwise identical architecture and hyper-parameters: with gold database features (Typology), with latently inferred typological features (Data-driven), or without both (Baseline).

It is evident that typology-enriched models consistently outperform baselines across several NLP tasks. Indeed, the scores are higher for metrics that increase (Unlabeled Attachment Score, F1 Score, and BLEU) and lower for metrics that decrease (Word Error Rate, Mean Average Error, and Perplexity) with better predictions. Nevertheless, improvements tend to be moderate, and only a small number of experiments support them with statistical significance tests. In general, it appears that they fall short of the potential usefulness of typology: in § 6 we analyze the possible reasons for this.

Some of the experiments we have surveyed investigate the effect of substituting typological features with features related to Genealogy and Language Identity (e.g., one-hot encoding of languages). Based on the results in Figure 10, it is unclear whether typology should be preferred, as it is sometimes rivaled by other types of features. In particular, it is typology that excels according to Tsvetkov et al. (2016), genealogy according to Søgaard and Wulff (2012) and Täckström, McDonald, and Nivre (2013), and language identity according to Ammar et al. (2016). However, drawing conclusions from the last experiment seems incautious: In § 4.2, we argued that their selection of features (presented in Figure 5) is debatable because of low diversification or noise. Moreover, it should be emphasized that one-hot language encoding is limited to the
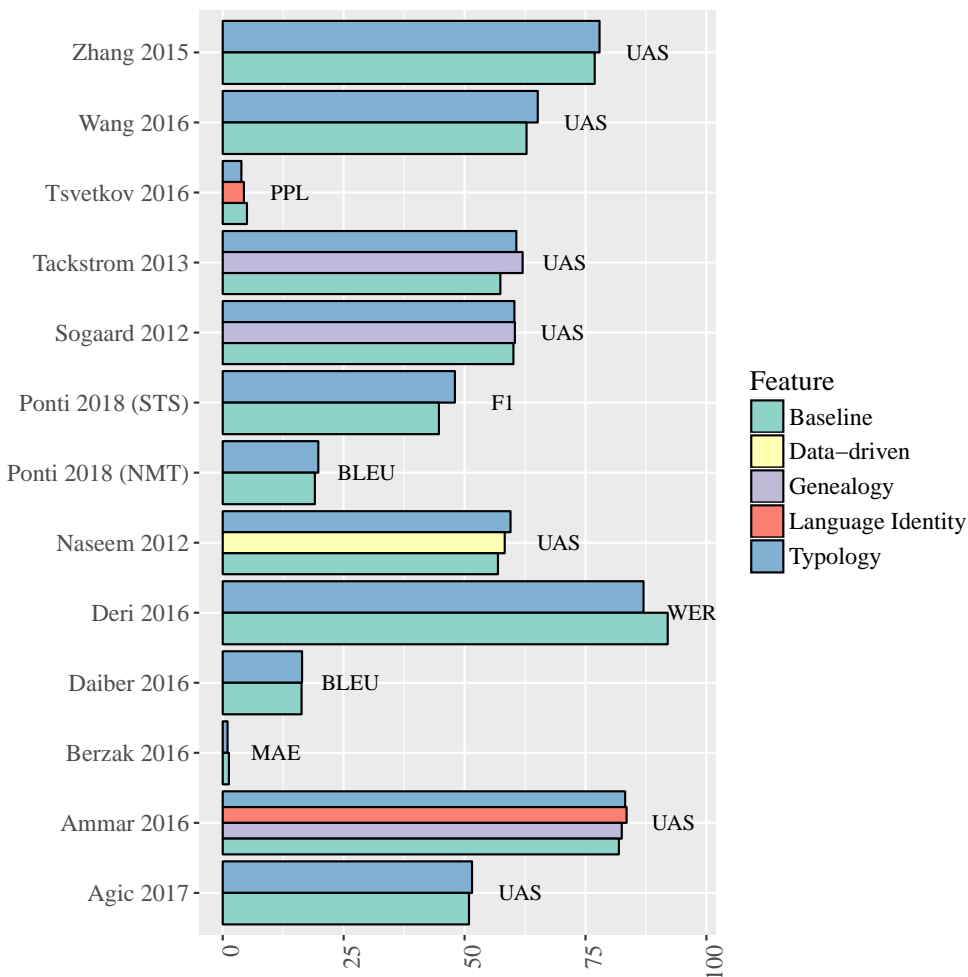
Figure 10
Performance of the surveyed algorithms for the tasks detailed in Table 3. The algorithms are
evaluated with different feature sets: no typological features (Baseline), latently inferred
typology (Data-driven), Genealogy, Language Identity, and gold database features (Typology).
Evaluation metrics are reported right of the bars: Unlabeled Attachment Score (UAS), Perplexity
(PPL), F1 Score, BiLingual Evaluation Understudy (BLEU), Word Error Rate (WER), and Mean
Absolute Error (MAE).

joint multilingual learning setting: Because it does not convey any information, it is of
no avail in language transfer.

Finally, let us consider the effectiveness of the methods described in §5.2 with
respect to incorporating typological features in NLP models. In case of selective shar-
ing, the tensor-based discriminative model (Zhang and Barzilay 2015) outperforms the
graph-based discriminative model (Täckström, McDonald, and Nivre 2013), which in
turn surpasses the generative model (Naseem, Barzilay, and Globerson 2012). With re-
gard to biasing multilingual models, there is a clear tendency toward letting typological
features interact not merely with the input representation, but also with deeper levels
of abstraction such as hidden layers.

Overall, this comparison supports the claim that typology can potentially aid in designing the architecture of algorithms, engineering their features, and selecting and pre-processing their data. Nonetheless, this discussion also revealed that many challenges lie ahead for each of these goals to be accomplished fully. We discuss them in the next section.

6. Future Research Avenues

In § 5 we surveyed the current uses of typological information in NLP. In this section we discuss potential future research avenues that may result in a closer and more effective integration of linguistic typology and multilingual NLP. In particular, we discuss: (1) the extension of existing methods to new tasks, possibly exploiting typological resources that have been neglected thus far (§ 6.1); (2) new methods for injecting typological information into NLP models as soft constraints or auxiliary objectives (§ 6.2); and (3) new ways to acquire and represent typological information that reflect the gradient and contextual nature of cross-lingual variation (§ 6.3).

6.1 Extending the Usage to New Tasks and Features

The trends observed in § 5 reveal that typology is integrated into NLP models mostly in the context of morphosyntactic tasks, and particularly syntactic parsing. Some exceptions include other levels of linguistic structure, such as phonology (Tsvetkov et al. 2016; Deri and Knight 2016) and semantics (Bender 2016; Ponti et al. 2018a). As a consequence, the set of selected typological features is mostly limited to a handful of word-order features from a single database, WALS. Nonetheless, the array of tasks that pertain to polyglot NLP is broad, and other typological data sets that have thus far been neglected may be relevant for them.

For example, typological frame semantics might benefit semantic role labeling, as it specifies the valency patterns of predicates across languages, including the number of arguments, their morphological markers, and their order. This information can be cast in the form of priors for unsupervised syntax-based Bayesian models (Titov and Klementiev 2012), guidance for alignments in annotation projection (Padó and Lapata 2009; Van der Plas, Merlo, and Henderson 2011), or regularizers for model transfer in order to tailor the source model to the grammar of the target language (Kozhevnikov and Titov 2013). Cross-lingual information about frame semantics can be extracted, for example, from the Valency Patterns Leipzig database (ValPaL).

Typological information regarding lexical semantics patterns can further assist various NLP tasks by providing information about translationally equivalent words across languages. Such information is provided in databases such as the World Loanword Database (WOLD), the Intercontinental Dictionary Series (IDS), and the Automated Similarity Judgment Program (ASJP). One example task is word sense disambiguation, as senses can be propagated from multilingual word graphs (Silberer and Ponzetto 2010) by bootstrapping from a few pivot pairs (Khapra et al. 2011), by imposing constraints in sentence alignments and harvesting bag-of-words features from these (Lefever, Hoste, and De Cock 2011), or by providing seeds for multilingual Word-Embedding-based lexicalized model transfer (Zennaki, Semmar, and Besacier 2016).

Another task where lexical semantics is crucial is sentiment analysis, for similar reasons: Bilingual lexicons constrain word alignments for annotation projection (Almeida et al. 2015) and provide pivots for shared multilingual representations in model transfer (Fernández, Esuli, and Sebastiani 2015; Ziser and Reichart 2018). Moreover, sentiment

analysis can leverage morphosyntactic typological information about constructions that alter polarity, such as negation (Ponti, Vulić, and Korhonen 2017).

Finally, morphological information was shown to aid interpreting the intrinsic difficulty of texts for language modeling and neural machine translation, both in supervised (Johnson et al. 2017) and in unsupervised (Artetxe et al. 2018) set-ups. In fact, the degree of fusion between roots and inflectional/derivative morphemes impacts the type/token ratio of texts, and consequently their rate of infrequent words. Moreover, the ambiguity of mapping between form and meaning of morphemes determines the usefulness of injecting character-level information (Gerz et al. 2018a, 2018b). This variation has to be taken into account in both language transfer and multilingual joint learning.

As a final note, we stress that the addition of new features does not concern just future work, but also the existing typology-savvy methods, which can widen their scope. For instance, the parsing experiments grounded on selective sharing (§ 5.2) could also take into consideration WALS features about Nominal Categories, Nominal Syntax, Verbal Categories, Simple Clauses, and Complex Sentences, as well as features from other databases such as SSWL, APiCS, and AUTOTYP. Likewise, models for phonological tasks (Tsvetkov et al. 2016; Deri and Knight 2016) could also extract features from typological databases such as LAPSyD and StressTyp2.

6.2 Injecting Typological Information into Machine Learning Algorithms

In § 5, we discussed the potential of typological information to provide guidance to NLP methods, and surveyed approaches such as network design in Bayesian models (Schone and Jurafsky 2001), selective sharing (Naseem, Barzilay, and Globerson 2012, inter alia), and biasing of multilingual joint models (Ammar et al. 2016, inter alia). However, many other frameworks (including those already mentioned in § 3) have been developed independently in order to allow the integration of expert and domain knowledge into traditional feature-based machine learning algorithms and neural networks. In this section we survey these frameworks and discuss their applicability to the integration of typological information into NLP models.

Encoding cross-language variation and preferences into a machine learning model requires a mechanism that can bias the *learning* (i.e., training and parameter estimation) and *inference* (prediction) of the model toward some pre-defined knowledge. In practice, learning algorithms, both linear (e.g., structured perceptron [Collins 2002], MIRA [Crammer and Singer 2003] and structured support vector machine [Taskar, Guestrin, and Koller 2004]) and non-linear (deep neural models) iterate between an inference step and a step of parameter update with respect to a gold standard. The inference step is the natural place where external knowledge could be encoded through constraints. This step biases the prediction of the model to agree with the external knowledge which, in turn, affects both the training process and the final prediction of the model at test time.

Information about cross-lingual variation, particularly when extracted empirically (see § 4), reflects tendencies rather than strict rules. As a consequence, *soft*, rather than *hard constraints* are a natural vehicle for their encoding. The goal of an inference algorithm is to predict the best output label according to the current state of the model parameters.[8] For this purpose, the algorithm searches the space of possible output labels

---

8 Generally speaking, an inference algorithm can make other predictions, such as computing expectations and marginal probabilities. Since in the context of this article we are mostly focused on the prediction of the best output label, we refer only to this type of inference problems.

in order to find the best one. Efficiency hence plays a key role in these algorithms. Introducing soft constraints into an inference algorithm, therefore, posits an algorithmic challenge: How can the output of the model be biased to agree with the constraints while the efficiency of the search procedure is kept? In this article we do not answer this question directly but rather survey a number of approaches that succeed in dealing with it.

Because linear models have been prominent in NLP research for a much longer time, it is not surprising that frameworks for the integration of soft constraints into these models are much more developed. The approaches proposed for this purpose include posterior regularization (PR) (Ganchev et al. 2010), generalized expectation (GE) (Mann and McCallum 2008), constraint-driven learning (CODL) (Chang, Ratinov, and Roth 2007), dual decomposition (DD) (Globerson and Jaakkola 2007; Komodakis, Paragios, and Tziritas 2011), and Bayesian modeling (Cohen 2016). These techniques use different types of knowledge encoding—for example, PR uses expectation constraints on the posterior parameter distribution, GE prefers parameter settings where the model's distribution on unsupervised data matches a predefined target distribution, CODL enriches existing statistical models with Integer Linear Programming constraints, and in Bayesian modeling a prior distribution is defined on the model parameters.

PR has already been used for incorporating universal linguistic knowledge into an unsupervised parsing model (Naseem et al. 2010). In the future, it could be extended to typological knowledge, which is a good fit for soft constraints. As another option, Bayesian modeling sets prior probability distributions according to the relationships encoded in typological features (Schone and Jurafsky 2001). Finally, DD has been applied to multi-task learning, which paves the way for typological knowledge encoding through a multi-task architecture in which one of the tasks is the actual NLP application and the other is the data-driven prediction of typological features. In fact, a modification of this architecture has already been applied to minimally supervised learning and domain adaptation with soft (non-typological) constraints (Reichart and Barzilay 2012; Rush et al. 2012).

The same ideas could be exploited in deep learning algorithms. We have seen in § 3.2 that multilingual joint models combine both shared and language-dependent parameters in order to capture the universal properties and cross-lingual differences, respectively. In order to enforce this division of roles more efficiently, these models could be augmented with the auxiliary task of predicting typological features automatically. This auxiliary objective could update parameters of the language-specific component, or those of the shared component, in an adversarial fashion, similar to what Chen et al. (2018) implemented by predicting language identity.

Recently, Hu et al. (2016a, 2016b) and Wang and Poon (2018) proposed frameworks that integrate deep neural models with manually specified or automatically induced constraints. Similar to CODL, the focus in Hu et al. (2016a) and Wang and Poon (2018) is on logical rules, while the ideas in Hu et al. (2016b) are related to PR. These frameworks provide a promising avenue for the integration of typological information and deep models.

A particular non-linear deep learning domain where knowledge integration is already prominent is multilingual representation learning (§ 3.3). In this domain, a number of works (Faruqui et al. 2015; Rothe and Schütze 2015; Mrkšić et al. 2016; Osborne, Narayan, and Cohen 2016) have proposed means through which external knowledge sourced from linguistic resources (such as WordNet, BabelNet, or lists of morphemes) can be encoded in word embeddings. Among the state-of-the-art specialization methods ATTRACT-REPEL (Mrkšić et al. 2017; Vulić et al. 2017) pushes together or pulls apart

vector pairs according to relational constraints, while preserving the relationship between words in the original space and possibly propagating the specialization knowledge to unseen words or transferring it to other languages (Ponti et al. 2018b). The success of these works suggests that a more extensive integration of external linguistic knowledge in general, and typological knowledge in particular, is likely to play a key role in the future development of word representations.

6.3 A New Typology: Gradience and Context-Sensitivity

As shown in § 4.2, most of the typology-savvy algorithms thus far exploited features extracted from manually crafted databases. However, this approach is riddled with several shortcomings, which are reflected in the small performance improvements observed in § 5.4. Luckily, these shortcomings may potentially be averted through the use of methods that allow typological information to emerge from the data in a bottom-up fashion, rather than being predetermined. In what follows we advocate for such a data-driven approach, based on several considerations.

First, typological databases provide *incomplete* documentation of the cross-lingual variation, in terms of features and languages. Raw textual data, which is easily accessible for many languages and is cost-effective, may provide a valid alternative that can facilitate automatic learning of more complete knowledge. Second, database information is *approximate*, as it is restricted to the majority strategy within a language. However, in theory each language allows for multiple strategies in different *contexts* and with different *frequencies*, hence databases risk hindering models from learning less-likely but plausible patterns (Sproat 2016). Inferring typological information from text would enable a system to discover patterns within individual examples, including both the frequent and the infrequent ones. Thirdly, typological features in databases are *discrete*, utilizing predefined categories devised to make high-level generalizations across languages. However, several categories in natural language are *gradient* (see for instance the discussion on semantic categorization in § 2), hence they are better captured by continuous features. In addition to being psychologically motivated, this sort of gradient representation is also more compatible with machine learning algorithms and particularly with deep neural models that naturally operate with real-valued multi-dimensional word embeddings and hidden states.

To sum up, the automatic development of typological information and its possible integration into machine learning algorithms have the potential to solve an important bottleneck in polyglot NLP. Current manually curated databases consist of incomplete, approximate, and discrete features that are intended to reflect contextual and gradient information implicitly present in text. These features are fed to continuous, probabilistic, and contextual machine learning models—which do not form a natural fit for the typological features. Instead, we believe that modeling cross-lingual variation directly from textual data can yield typological information that is more suitable for machine learning.

Several techniques surveyed in § 4.3 are suited to serve this purpose. In particular, the extraction from morphosyntactic annotation (Liu 2010, inter alia) and alignments from multi-parallel texts (Asgari and Schütze 2017, inter alia) provide information about typological constructions at the level of individual examples. Moreover, language vectors (Malaviya, Neubig, and Littell 2017; Bjerva and Augenstein 2018) and alignments from multi-parallel texts preserve the gradient nature of typology through continuous representations.

The successful integration of these components would affect the way multilingual feature engineering is performed. As opposed to using binary vectors of typological features, the information about language-internal variation could be encoded as real-valued vectors, where each dimension is a possible strategy for a given construction and its relative frequency within a language.

As an alternative, selective sharing and multilingual biasing could be performed at the level of individual examples rather than languages as a whole. In particular, model parameters could be transferred among similar examples; and input/hidden representations could be conditioned on contextual typological patterns. Finally, focusing on the various instantiations of a particular type rather than considering languages as indissoluble blocks would enhance data selection, similar to what Søgaard (2011) achieved using PoS *n*-grams for similarity measurement. The selection of similar sentences rather than similar languages as source data in language transfer is likely to yield large improvements, as demonstrated by Agić (2017) for parsing in an oracle setting.

Finally, the bottom–up development of typological features may also address radically resource-less languages that lack even raw textual data in a digital format. For this group, which still constitutes a large portion of the world's languages, there are often available reference grammars written by field linguists, which are the ultimate source for typological databases. These grammars could be queried automatically, and fine-grained typological information could be harvested through information extraction techniques.

7. Conclusions

In this article, we surveyed a wide range of approaches integrating typological information, derived from the empirical and systematic comparison of the world's languages, and NLP algorithms. The most fundamental problem for the advancement of this line of research is bridging the gap between the interpretable, language-wide, and discrete features of linguistic typology found in database documentation, and the opaque, contextual, and probabilistic models of NLP. We addressed this problem by exploring a series of questions: (i) for which tasks and applications is typology useful? (ii) What are the advantages and limitations of currently available typological databases? Can data-driven inference of typological features offer an alternative source of information? (iii) Which methods allow us to inject typological information from external resources, and how should such information be encoded? (iv) By which margin do typology-savvy methods surpass typology-agnostic baselines? How does typology compare to other criteria of language classification, such as genealogy? (v) In addition to augmenting machine learning algorithms, which other purposes do typology serve for NLP? We summarize our key findings here:

1. Typological information is currently used predominantly for morphosyntactic tasks, in particular dependency parsing. As a consequence, these approaches typically select a limited subset of features from a single data set (WALS) and focus on a single aspect of variation (typically word order). However, typological databases also cover other important features, related to predicate–argument structure (ValPaL), phonology (LAPSyD, PHOIBLE, StressTyp2), and lexical semantics (IDS, ASJP), which are currently largely neglected by the multilingual NLP community. In fact, these features have the potential to benefit many tasks addressed by language transfer or joint multilingual learning techniques,

such as semantic role labeling, word sense disambiguation, or sentiment analysis.

2. Typological databases tend to be incomplete, containing missing values for individual languages or features. This hinders the integration of the information in such databases into NLP models; and therefore, several techniques have been developed to predict missing values automatically. They include heuristics derived from morphosyntactic annotation; propagation from other languages based on hierarchical clusters or similarity metrics; supervised models; and distributional methods applied to multi-parallel texts. However, none of these techniques surpasses the others across the board in prediction accuracy, as each excels in different feature types. A challenge left for future work is creating ensembles of techniques to offset their individual disadvantages.

3. The most widespread approach to exploit typological features in NLP algorithms is "selective sharing" for language transfer. Its intuition is that a model should learn universal properties from all examples, but language-specific information only from examples with similar typological properties. Another successful approach is gearing multilingual joint models toward specific languages by concatenating typological features in input, or conditioning hidden layers and global sequence representations on them. New approaches could be inspired by traditional techniques for encoding external knowledge into machine learning algorithms through soft constraints on the inference step, semi-supervised prototype-driven methods, specialization of semantic spaces, or auxiliary objectives in a multi-task learning setting.

4. The integration of typological features into NLP models yields consistent (even if often moderate) improvements over baselines lacking such features. Moreover, guidance from typology should be preferred to features related to genealogy or other language properties. Models enriched with the latter features occasionally perform equally well due to their correlation with typological features, but fall short when it comes to modeling diversified language samples or fine-grained differences among languages.

5. In addition to feature engineering, typological information has served several other purposes. Firstly, it allows experts to define rule-based models, or to assign priors and independence assumptions in Bayesian graphical models. Secondly, it facilitates data selection and weighting, at the level of both languages and individual examples. Annotated data can also be synthesized or preprocessed according to typological criteria, in order to increase their coverage of phenomena or availability for further languages. Thirdly, typology enables researchers to interpret and reasonably foresee the difference in performance of algorithms across the sampled languages.

Finally, we advocated for a new approach to linguistic typology inspired by the most recent trends in the discipline and aimed at averting some fundamental limitations of the current approach. In fact, typological database documentation is incomplete, approximate, and discrete. As a consequence, it does not fit well with the gradient

and contextual models of machine learning. However, typological databases are originally created from raw linguistic data. An alternative approach could involve learning typology from such data automatically (i.e., from scratch). This would capture the variation within languages at the level of individual examples, and to naturally encode typological information into continuous representations. These goals have already been partly achieved by methods involving language vectors, heuristics derived from morphosyntactic annotation, or distributional information from multi-parallel texts. The main future challenge is the integration of these methods into machine learning models, as opposed to sourcing typological features from databases.

In general, we demonstrated that typology is relevant to a wide range of NLP tasks and provides a quite effective and principled way to carry out language transfer and multilingual joint learning. We hope that the research described in this survey will provide a platform for deeper integration of typological information and NLP techniques, thus furthering the advancement of multilingual NLP.

References

Adel, Heike, Ngoc Thang Vu, and Tanja Schultz. 2013. Combination of recurrent neural networks and factored language models for code-switching language modeling. In *Proceedings of ACL*, pages 206–211.

Agić, Željko. 2017. Cross-lingual parser selection for low-resource languages. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 1–10.

Agić, Željko, Dirk Hovy, and Anders Søgaard. 2015. If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. In *The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing*, pages 268–272.

Agić, Željko, Anders Johannsen, Barbara Plank, Héctor Alonso Martínez, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*.

Agić, Željko, Jörg Tiedemann, Kaja Dobrovoljc, Simon Krek, Danijela Merkler, and Sara Može. 2014. Cross-lingual dependency parsing of related languages with rich morphosyntactic tagsets. In *Proceedings of the EMNLP 2014 Workshop on Language Technology for Closely Related Languages and Language Variants*, pages 13–24.

Almeida, Mariana SC, Cláudia Pinto, Helena Figueira, Pedro Mendes, and André FT Martins. 2015. Aligning opinions: Cross-lingual opinion mining with dependencies. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 408–418.

Ammar, Waleed, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Many languages, one parser. *TACL*, 4:431–444.

Artetxe, Mikel, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*, pages 1–12.

Asgari, Ehsaneddin and Hinrich Schütze. 2017. Past, present, future: A computational investigation of the typology of tense in 1,000 languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 113–124.

Bakker, Dik. 2010. Language sampling. In J. J. Song, editor, *The Oxford Handbook of Linguistic Typology*, Oxford University Press, pages 100–127.

Banea, Carmen, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 127–135.

Bender, Emily M. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the*

EACL 2009 Workshop on the Interaction Between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?, pages 26–32.

Bender, Emily M. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 3(6):1–26.

Bender, Emily M. 2014. Language collage: Grammatical description with the lingo grammar matrix. In *Proceedings of LREC*, pages 2447–2451.

Bender, Emily M. 2016. Linguistic typology in natural language processing. *Linguistic Typology*, 20(3):645–660.

Bender, Emily M., Michael Wayne Goodman, Joshua Crowgey, and Fei Xia. 2013. Towards creating precision grammars from interlinear glossed text: Inferring large-scale typological properties. In *Proceedings of LaTeCH 2013*, pages 74–83, Sofia.

Berlin, Brent and Paul Kay. 1969. *Basic Color Terms: Their Universality and Evolution*. California University Press.

Berzak, Yevgeni, Roi Reichart, and Boris Katz. 2014. Reconstructing native language typology from foreign language usage. In *Proceedings of CoNLL*, pages 21–29, Baltimore, MD.

Berzak, Yevgeni, Roi Reichart, and Boris Katz. 2015. Contrastive analysis with predictive power: Typology driven estimation of grammatical error distributions in ESL. In *Proceedings of CoNLL*, pages 94–102, Beijing.

Bickel, Balthasar. 2007. Typology in the 21st century: Major current developments. *Linguistic Typology*, 11(1):239–251.

Bickel, Balthasar. 2015. Distributional typology: Statistical inquiries into the dynamics of linguistic diversity. In Bernd Heine and Heiko Narrog, editors, *Oxford Handbook of Linguistic Analysis*. 901–923.

Bickel, Balthasar, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Kristine Hildebrandt, Michael Rießler, Lennart Bierkandt, Fernando Zúñiga, and John Lowe. 2017. The AUTOTYP typological databases. version 0.1.0. Technical report, University of Zurich.

Bjerva, Johannes and Isabelle Augenstein. 2018. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 907–916.

Bowerman, Melissa and Soonja Choi. 2001. Shaping meanings for language: Universal and language-specific in the acquisition of semantic categories, In Melissa Bowerman and Stephen Levinson, editors, *Language Acquisition and Conceptual Development*, Cambridge University Press, pages 475–511.

Bybee, Joan and James L. McClelland. 2005. Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The Linguistic Review*, 22(2-4):381–410.

Bybee, Joan L. 1988. The diachronic dimension in explanation. In J. A. Hawkins, editor, *Explaining Language Universals*, Basil Blackwell, pages 350–379.

Chandar, Sarath, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C. Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1853–1861.

Chang, Ming Wei, Lev Ratinov, and Dan Roth. 2007. Guiding semi-supervision with constraint-driven learning. In *Proceedings of ACL*, pages 280–287, Prague.

Chen, Xilun, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

Cohen, Shay B. 2016. *Bayesian Analysis in Natural Language Processing*, Synthesis Lectures on Human Language Technologies. Morgan and Claypool.

Coke, Reed, Ben King, and Dragomir R. Radev. 2016. Classifying syntactic regularities for hundreds of languages. *CoRR*, abs/1603.08016.

Collins, Chris and Richard Kayne. 2009. Syntactic structures of the world's languages. `http://sswl.railsplayground.net/`.

Collins, Michael. 2002. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of EMNLP*, pages 1–8, Philadelphia, PA.

Comrie, Bernard. 1989. *Language Universals and Linguistic Typology: Syntax and Morphology*. University of Chicago Press.

Conneau, Alexis, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Conneau, Alexis, Ruty Rinott, Guillaume
    Lample, Adina Williams, Samuel Bowman,
    Holger Schwenk, and Veselin Stoyanov.
    2018. XNLI: Evaluating cross-lingual
    sentence representations. In *Proceedings of
    the 2018 Conference on Empirical Methods in
    Natural Language Processing*,
    pages 2475–2485.

Copestake, Ann, Dan Flickinger, Carl
    Pollard, and Ivan A. Sag. 2005. Minimal
    recursion semantics: An introduction.
    Research on Language and Computation,
    3(2-3):281–332.

Corbett, Greville G. 2010. Implicational
    hierarchies. In J. J. Song, editor, *The Oxford
    Handbook of Linguistic Typology*, Oxford
    University Press, pages 190–205.

Cotterell, Ryan and Jason Eisner. 2017.
    Probabilistic typology: Deep generative
    models of vowel inventories. In *Proceedings
    of the 55th Annual Meeting of the Association
    for Computational Linguistics (Volume 1:
    Long Papers)*, volume 1, pages 1182–1192.

Cotterell, Ryan and Jason Eisner. 2018. A
    deep generative model of vowel formant
    typology. In *Proceedings of the 2018
    Conference of the North American Chapter of
    the Association for Computational Linguistics:
    Human Language Technologies, Volume 1
    (Long Papers)*, volume 1, pages 37–46.

Crammer, Koby and Yoram Singer. 2003.
    Ultraconservative online algorithms for
    multiclass problems. *Journal of Machine
    Learning Research*, 3:951–991.

Cristofaro, S. and P. Ramat. 1999. *Introduzione
    alla tipologia linguistica*. Carocci.

Croft, William. 1995. Autonomy and
    functionalist linguistics. *Language*,
    71(3):490–532.

Croft, William. 2000. *Explaining Language
    Change: An Evolutionary Approach*. Pearson
    Education.

Croft, William. 2003. *Typology and Universals*.
    Cambridge University Press.

Croft, William, Dawn Nordquist, Katherine
    Looney, and Michael Regan. 2017.
    Linguistic typology meets universal
    dependencies. In *Proceedings of the
    15th International Workshop on Treebanks
    and Linguistic Theories (TLT15)*,
    pages 63–75.

Croft, William and Keith T. Poole. 2008.
    Inferring universals from grammatical
    variation: Multidimensional scaling for
    typological analysis. *Theoretical Linguistics*,
    34(1):1–37.

Daiber, Joachim, Miloš Stanojević, and Khalil
    Sima'an. 2016. Universal reordering via
    linguistic typology. In *Proceedings of
    COLING 2016, the 26th International
    Conference on Computational Linguistics:
    Technical Papers*, pages 3167–3176.

d'Andrade, Roy G. 1995. *The Development of
    Cognitive Anthropology*. Cambridge
    University Press.

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

Daumé III, Hal and Lyle Campbell. 2007. A
    Bayesian model for discovering
    typological implications. In *Proceedings of
    ACL*, pages 65–72, Prague.

Dempster, Arthur P., Nan M. Laird, and
    Donald B. Rubin. 1977. Maximum
    likelihood from incomplete data via
    the em algorithm. *Journal of the Royal
    Statistical Society. Series B (Methodological)*.
    39(1):1–38.

Deri, Aliya and Kevin Knight. 2016.
    Grapheme-to-phoneme models for
    (almost) any language. In *Proceedings of
    ACL*, pages 399–408, Berlin.

Dixon, Robert M. W. 1977. Where have all the
    adjectives gone? *Studies in Language*,
    1(1):19–80.

Dixon, Robert M. W. 1994. *Ergativity*.
    Cambridge University Press.

Dryer, Matthew S. 1998. Why statistical
    universals are better than absolute
    universals. In *Papers from the 33rd Regional
    Meeting of the Chicago Linguistic Society*,
    pages 1–23.

Dryer, Matthew S. and Martin Haspelmath,
    editors. 2013. *WALS Online*. Max Planck
    Institute for Evolutionary Anthropology,
    Leipzig.

Duong, Long, Trevor Cohn, Steven Bird, and
    Paul Cook. 2015a. Low resource
    dependency parsing: Cross-lingual
    parameter sharing in a neural network
    parser. In *Proceedings of the 53rd Annual
    Meeting of the Association for Computational
    Linguistics and the 7th International Joint
    Conference on Natural Language Processing*,
    volume 2, pages 845–850.

Duong, Long, Trevor Cohn, Steven Bird, and
    Paul Cook. 2015b. A neural network
    model for low-resource universal
    dependency parsing. In *Proceedings
    of the 2015 Conference on Empirical
    Methods in Natural Language Processing*,
    pages 339–348.

Duong, Long, Hiroshi Kanayama, Tengfei
    Ma, Steven Bird, and Trevor Cohn. 2016.
    Learning crosslingual word embeddings
    without bilingual corpora. In *Proceedings of
    the 2016 Conference on Empirical Methods in
    Natural Language Processing*,
    pages 1285–1295.

Durham, William H. 1991. *Coevolution: Genes, Culture, and Human Diversity*. Stanford University Press.

Durrett, Greg, Adam Pauls, and Dan Klein. 2012. Syntactic transfer using a bilingual lexicon. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1–11.

Evans, Nicholas. 2011. In Jae Jung Song editor, Semantic typology, *The Oxford Handbook of Linguistic Typology*, Oxford University Press, pages 504–533.

Evans, Nicholas and Stephen C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain sciences*, 32(5):429–448.

Fang, Meng and Trevor Cohn. 2017. Model transfer for tagging low-resource languages using a bilingual dictionary. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 587–593.

Faruqui, Manaal, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL-HLT*, pages 1606–1615, Denver, CO.

Fernández, Alejandro Moreo, Andrea Esuli, and Fabrizio Sebastiani. 2015. Distributional correspondence indexing for cross-lingual and cross-domain sentiment classification. *Journal of Artificial Intelligence Research*, 55:131–163.

Ganchev, Kuzman, Jennifer Gillenwater, Ben Taskar et al. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049.

Georgi, Ryan, Fei Xia, and William Lewis. 2010. Comparing language similarity across genetic and typologically based groupings. In *Proceedings of COLING*, pages 385–393, Beijing.

Gerz, Daniela, Edoardo Maria Ponti, Jason Naradowsky, Roi Reichart, Anna Korhonen, and Ivan Vulić. 2018a. Language modeling for morphologically rich languages: Character-aware modeling for word-level prediction. *Transactions of the Association for Computational Linguistics*, 6:451–466.

Gerz, Daniela, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018b. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 316–327.

Globerson, Amir and Tommi S. Jaakkola. 2007. Fixing max-product: Convergent message passing algorithms for MAP LP-relaxations. In *Proceedings of NIPS*, pages 553–560, Vancouver.

Goedemans, Rob, Jeffrey Heinz, and Harry Van der Hulst, editors . 2014. *Stresstyp2*. University of Connecticut, University of Delaware, Leiden University, and the U.S. National Science Foundation.

Gouws, Stephan, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *International Conference on Machine Learning*, pages 748–756.

Gouws, Stephan and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of NAACL-HLT*, pages 1386–1390, Denver, CO.

Greenberg. 1978. Diachrony, synchrony and language universals. In Joseph H. Greenberg, Charles A. Ferguson, and Edith A. Moravcsik, editors, *Universals of Human Language, Vol. 1: Method and Theory*, Stanford University Press, pages 61–92.

Greenberg, Joseph H. 1963. Some universals of grammar with particular reference to the order of meaningful elements. *Universals of Language*, 2:73–113.

Greenberg, Joseph H. 1966a. Synchronic and diachronic universals in phonology. *Language*, 42(2):508–517.

Greenberg, Joseph H. 1966b. *Universals of language*. MIT Press.

Guo, Jiang, Wanxiang Che, Haifeng Wang, and Ting Liu. 2016. A universal framework for inductive transfer parsing across multi-typed treebanks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 12–22.

Guo, Jiang, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 1, pages 1234–1244.

Ha, Thanh Le, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the 2016 International Workshop on Spoken Language Translation (IWSLT)*.

Hammarström, Harald, Robert Forkel, Martin Haspelmath, and Sebastian Bank, editors. 2016. *Glottolog 2.7*. Max Planck Institute for the Science of Human History, Jena.

Hartmann, Iren, Martin Haspelmath, and Bradley Taylor, editors . 2013. *Valency Patterns Leipzig*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Haspelmath, Martin. 1999. Optimality and diachronic adaptation. *Zeitschrift für Sprachwissenschaft*, 18(2):180–205.

Haspelmath, Martin. 2007. Pre-established categories don't exist: Consequences for language description and typology. *Linguistic Typology*, 11(1):119–132.

Haspelmath, Martin and Uri Tadmor, editors. 2009. *WOLD*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Hermann, Karl Moritz and Phil Blunsom. 2014. Multilingual distributed representations without word alignment. In *Proceedings of ICLR*.

Hu, Zhiting, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. 2016a. Harnessing deep neural networks with logic rules. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2410–2420.

Hu, Zhiting, Zichao Yang, Ruslan Salakhutdinov, and Eric Xing. 2016b. Deep neural networks with massive learned knowledge. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1670–1679.

Hwa, Rebecca, Philip Resnik, Amy Weinberg, Clara I. Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325.

Jensen, Finn V. 1996. *An Introduction to Bayesian Networks*, volume 210. University College London Press, London.

Johnson, Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5(1):339–351.

Key, Mary Ritchie and Bernard Comrie, editors. 2015. *IDS*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Khapra, Mitesh M., Salil Joshi, Arindam Chatterjee, and Pushpak Bhattacharyya. 2011. Together we can: Bilingual bootstrapping for WSD. In *Proceedings of ACL*, pages 561–569, Portland, OR.

Klementiev, Alexandre, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING*, pages 1459–1474, Mumbai.

Komodakis, Nikos, Nikos Paragios, and Georgios Tziritas. 2011. MRF energy minimization and beyond via dual decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):531–552.

Kozhevnikov, Mikhail and Ivan Titov. 2013. Cross-lingual transfer of semantic role labeling models. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1190–1200.

Lauly, Stanislas, Alex Boulanger, and Hugo Larochelle. 2013. Learning multilingual word representations using a bag-of-words autoencoder. In *Deep Learning Workshop at NIPS*.

Lefever, Els, Véronique Hoste, and Martine De Cock. 2011. Parasense or how to use parallel corpora for word sense disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 317–322.

Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig. 2016. *Ethnologue: Languages of the World*, 19th ed., SIL International.

Lewis, William D. and Fei Xia. 2008. Automatically identifying computationally relevant typological features. In *Proceedings of IJCNLP*, pages 685–690, Hyderabad.

Littel, Patrick, David R. Mortensen, and Lori Levin. 2016. URIEL Typological database. Carnegie Mellon University, Pittsburgh: PA.

Liu, Haitao. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 120(6):1567–1578.

Lu, Xia. 2013. Exploring word order universals: A probabilistic graphical model approach. In *Proceedings of ACL (Student Research Workshop)*, pages 150–157.

Luong, Thang, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159.

Maddieson, Ian, Sébastien Flavier, Egidio Marsico, Christophe Coupé, and François Pellegrino. 2013. LAPSyd: Lyon-Albuquerque phonological systems

database. In *Proceedings of INTERSPEECH*, pages 3022–3026, Lyon.

Majid, Asifa, Melissa Bowerman, Miriam van Staden, and James S. Boster. 2007. The semantic categories of cutting and breaking events: A crosslinguistic perspective. *Cognitive Linguistics*, 18(2):133–152.

Malaviya, Chaitanya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535.

Mann, Gideon S. and Andrew McCallum. 2008. Generalized expectation criteria for semi-supervised learning of conditional random fields. In *Proceedings of ACL*, pages 870–878, Columbus, OH.

McDonald, Ryan, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 91–98.

Michaelis, Susanne Maria, Philippe Maurer, Martin Haspelmath, and Magnus Huber, editors. 2013. *Atlas of Pidgin and Creole Language Structures Online*. Max Planck Institute for Evolutionary Anthropology.

Mikolov, Tomas, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Moran, Steven, Daniel McCloy, and Richard Wright, editors. 2014. *PHOIBLE Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Mrkšić, Nikola, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5(1):309–324.

Mrkšić, Nikola, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of NAACL-HLT*, pages 142–148, San Diego, CA.

Murawaki, Yugo. 2017. Diachrony-aware induction of binary latent representations from typological features. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 451–461.

Naseem, Tahira, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of ACL*, pages 629–637, Jeju Island.

Naseem, Tahira, Harr Chen, Regina Barzilay, and Mark Johnson. 2010. Using universal linguistic knowledge to guide grammar induction. In *Proceedings of EMNLP 2010*, pages 1234–1244.

Nichols, Johanna. 1992. *Language Diversity in Space and Time*. University of Chicago Press.

Niehues, Jan, Teresa Herrmann, Stephan Vogel, and Alex Waibel. 2011. Wider context by using bilingual language models in machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 198–206.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of LREC*, pages 1659–1666, Portorož.

O'Horan, Helen, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1297–1308.

Osborne, D., S. Narayan, and S. B. Cohen. 2016. Encoding prior knowledge with eigenword embeddings. *Transactions of the Association for Computational Linguistics*, 4:417–430.

Östling, Robert. 2015. Word order typology through multilingual word alignment. In *Proceedings of ACL*, pages 205–211, Beijing.

Östling, Robert and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 644–649.

Padó, Sebastian and Mirella Lapata. 2005. Cross-linguistic projection of role-semantic information. In *Proceedings of EMNLP*, pages 859–866, Vancouver.

Padó, Sebastian and Mirella Lapata. 2009. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36(1):307–340.

Pappas, Nikolaos and Andrei Popescu-Belis. 2017. Multilingual hierarchical attention networks for document classification. In *8th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 1015–1025.

Plank, Frans and Elena Filiminova. 1996. Universals archive. http://www.ling.unikonstanz.de/pages/proj/sprachbau.htm. Universität Konstanz.

Ponti, Edoardo Maria, Roi Reichart, Anna Korhonen, and Ivan Vulić. 2018a. Isomorphic transfer of syntactic structures in cross-lingual NLP. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1531–1542.

Ponti, Edoardo Maria, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018b. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 282–293.

Ponti, Edoardo Maria, Ivan Vulić, and Anna Korhonen. 2017. Decoding sentiment from distributed representations of sentences. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 22–32, Vancouver.

Reichart, Roi and Regina Barzilay. 2012. Multi event extraction guided by global constraints. In *Proceedings of NAACL*, pages 70–79, Montreal.

Rosa, Rudolf and Zdenek Zabokrtsky. 2015. KLcpos3—a language similarity measure for delexicalized parser transfer. In *Proceedings of ACL*, pages 243–249, Beijing.

Rothe, Sascha and Hinrich Schütze. 2015. AutoExtend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of ACL*, pages 1793–1803, Beijing.

Rotman, Guy, Ivan Vulić, and Roi Reichart. 2018. Bridging languages through images with deep partial canonical correlation analysis. In *Proceedings of ACL 2018*, pages 910–921.

Roy, Rishiraj Saha, Rahul Katare, Niloy Ganguly, and Monojit Choudhury. 2014. Automatic discovery of adposition typology. In *Proceedings of COLING*, pages 1037–1046.

Ruder, Sebastian. 2018. A survey of cross-lingual embedding models. *Journal of Artificial Intelligence Research*. To appear.

Rush, Alexander M., Roi Reichart, Michael Collins, and Amir Globerson. 2012.

Improved parsing and POS tagging using inter-sentence consistency constraints. In *Proceedings of EMNLP-CoNLL*, pages 1434–1444, Jeju Island.

Sapir, Edward. 2014 [1921]. *Language.* Cambridge University Press.

Schone, Patrick and Daniel Jurafsky. 2001. Language-independent induction of part of speech class labels using only language universals. In *IJCAI-2001 Workshop "Text Learning: Beyond Supervision."*

Sennrich, Rico and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation*, volume 1, pages 83–91.

Silberer, Carina and Simone Paolo Ponzetto. 2010. UHD: Cross-lingual word sense disambiguation using multilingual co-occurrence graphs. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 134–137.

Snyder, Ben. 2010. *Unsupervised Multilingual Learning*, PhD thesis. Massachussetts Institute of Technology.

Snyder, Benjamin and Regina Barzilay. 2008. Unsupervised multilingual learning for morphological segmentation. In *Proceedings of ACL-08: HLT*, pages 737–745.

Søgaard, Anders. 2011. Data point selection for cross-language adaptation of dependency parsers. In *Proceedings of ACL*, pages 682–686, Portland, OR.

Søgaard, Anders and Julie Wulff. 2012. An empirical study of non-lexical extensions to delexicalized transfer. *Proceedings of COLING 2012: Posters*, pages 1181–1190.

Sproat, Richard. 2016. Language typology in speech and language technology. *Linguistic Typology*, 20(3):635–644.

Täckström, Oscar, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers. In *Proceedings of NAACL-HLT*, pages 1061–1071, Atlanta, GA.

Täckström, Oscar, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of NAACL-HLT*, pages 477–487, Montreal.

Tai, Kai Sheng, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1556–1566.

Takamura, Hiroya, Ryo Nagata, and Yoshifumi Kawasaki. 2016. Discriminative analysis of linguistic features for typological study. In *Proceedings of LREC*, pages 69–76, Portorož.

Talmy, Leonard. 1991. Path to realization: A typology of event conflation. In *Proceedings of the Seventeenth Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on the Grammar of Event Structure*, pages 480–519.

Taskar, Ben, Carlos Guestrin, and Daphne Koller. 2004. Max-margin Markov networks. In *Proceedings of NIPS*, pages 25–32, Vancouver.

Teh, Yee Whye, Hal Daumé III, and Daniel M. Roy. 2007. Bayesian agglomerative clustering with coalescents. In *Proceedings of NIPS*, pages 1473–1480.

Tiedemann, Jörg. 2015. Cross-lingual dependency parsing with universal dependencies and predicted POS labels. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 340–349.

Titov, Ivan and Alexandre Klementiev. 2012. Crosslingual induction of semantic roles. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 647–656.

Tsvetkov, Yulia, Sunayana Sitaram, Manaal Faruqui, Guillaume Lample, Patrick Littell, David Mortensen, Alan W. Black, Lori Levin, and Chris Dyer. 2016. Polyglot neural language models: A case study in cross-lingual phonetic representation learning. In *Proceedings of NAACL*, pages 1357–1366, San Diego, CA.

Upadhyay, Shyam, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1661–1670.

Vulić, Ivan, Wim De Smet, and Marie-Francine Moens. 2011. Identifying word translations from comparable corpora using latent topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*, pages 479–484.

Vulić, Ivan and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and*

the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), volume 2, pages 719–725.

Vulić, Ivan, Nikola Mrkšić, Roi Reichart, Diarmuid Ó Séaghdha, Steve Young, and Anna Korhonen. 2017. Morph-fitting: Fine-tuning word vector spaces with simple language-specific rules. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 56–68.

Wälchli, Bernhard and Michael Cysouw. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics*, 50(3):671–710.

Wang, Dingquan and Jason Eisner. 2016. The galactic dependencies treebanks: Getting more data by synthesizing new languages. *Transactions of the Association for Computational Linguistics*, 4:491–505.

Wang, Dingquan and Jason Eisner. 2017. Fine-grained prediction of syntactic typology: Discovering latent structure with *supervised* learning. *Transactions of the Association for Computational Linguistics*, 5:147–162.

Wang, Hai and Hoifung Poon. 2018. Deep probabilistic logic: A unifying framework for indirect supervision. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1891–1902.

Wang, Mengqiu and Christopher D. Manning. 2014. Cross-lingual pseudo-projected expectation regularization for weakly supervised learning. *Transactions of the Association for Computational Linguistics*, 2:55–66.

Wichmann, Søren, Eric W. Holman, and Cecil H. Brown, editors . 2016. *The ASJP Database (version 17)*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Wisniewski, Guillaume, Nicolas Pécheux, Souhir Gahbiche-Braham, and François Yvon. 2014. Cross-lingual part-of-speech tagging through ambiguous learning. In *Proceedings of EMNLP*, pages 1779–1785, Doha.

Xiao, Min and Yuhong Guo. 2014. Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of CoNLL*, pages 119–129.

Yang, Zhilin, Ruslan Salakhutdinov, and William Cohen. 2016. Multi-task cross-lingual sequence tagging from scratch. *arXiv preprint arXiv:1603.06270*.

Yarowsky, David, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection

across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, pages 1–8.

Zeman, Daniel and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of IJCNLP*, pages 35–42.

Zennaki, Othman, Nasredine Semmar, and Laurent Besacier. 2016. Inducing multilingual text analysis tools using bidirectional recurrent neural networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 450–460.

Zhang, Yuan and Regina Barzilay. 2015. Hierarchical low-rank tensors for multilingual transfer parsing. In *Proceedings of EMNLP*, pages 1857–1867, Lisbon.

Zhang, Yuan, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag—multilingual POS tagging via coarse mapping between embeddings. In *Proceedings of NAACL*, pages 1307–1317, San Diego, CA.

Zhang, Yuan, Roi Reichart, Regina Barzilay, and Amir Globerson. 2012. Learning to map into a universal POS tagset. In *Proceedings of EMNLP*, pages 1368–1378, Jeju Island.

Zhou, Guangyou, Tingting He, Jun Zhao, and Wensheng Wu. 2015. A subspace learning framework for cross-lingual sentiment classification with partial parallel data. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pages 1426–1432.

Zhou, Xinjie, Xianjun Wan, and Jianguo Xiao. 2016. Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1403–1412.

Ziser, Yftah and Roi Reichart. 2018. Deep pivot-based modeling for cross-language cross-domain transfer with minimal guidance. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 238–249.