



**HAL**  
open science

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

Edoardo Maria Ponti, Helen O 'Horan, Yevgeni Berzak, Ivan Vulic, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, Anna Korhonen

► **To cite this version:**

Edoardo Maria Ponti, Helen O 'Horan, Yevgeni Berzak, Ivan Vulic, Roi Reichart, et al.. Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing. 2018. hal-02425462v1

**HAL Id: hal-02425462**

**<https://hal.science/hal-02425462v1>**

Preprint submitted on 9 Aug 2018 (v1), last revised 24 Oct 2020 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Edoardo Maria Ponti\*  
LTL, University of Cambridge

Helen O’Horan\*\*  
LTL, University of Cambridge

Yevgeni Berzak†  
Department of Brain and Cognitive  
Sciences, MIT

Ivan Vulić‡  
LTL, University of Cambridge

Roi Reichart§  
Faculty of Industrial Engineering and  
Management, Technion - IIT

Thierry Poibeau#  
LATTICE Lab, CNRS and ENS/PSL and  
Univ. Sorbonne nouvelle/USPC

Ekaterina Shutova\*\*  
ILLC, University of Amsterdam

Anna Korhonen††  
LTL, University of Cambridge

*Understanding cross-lingual variation is essential for the development of effective multilingual natural language processing (NLP) applications. The field of linguistic typology studies and classifies the world’s languages according to their structural and semantic features, with the aim of explaining both the common properties and the systematic diversity of languages. Typological information sourced from databases has been integrated into NLP algorithms, providing valuable guidance for several tasks. In turn, NLP techniques can be used to inform research on linguistic typology itself, facilitating data-driven induction of typological knowledge. This paper provides a comprehensive review of the research at the intersection between multilingual NLP and linguistic typology and outlines future research avenues, with the aim of encouraging further advances in the two fields and building a bridge between them. In particular, we advocate for a new typology that adapts the deterministic and discrete nature of typological categories to the contextual and continuous nature of machine learning algorithms used in NLP.*

## 1. Introduction

Languages may share universal features at a deep, abstract level, but the structures found in real-world, surface-level natural language vary significantly. This variation makes it

\* English Faculty Building, 9 West Road Cambridge CB3 9DA, United Kingdom. E-mail: ep490@cam.ac.uk

\*\* English Faculty Building. E-mail: helen.ohoran@gmail.com

† 77 Massachusetts Avenue, Cambridge, MA 02139, USA. E-mail: berzak@mit.edu

‡ English Faculty Building. E-mail: iv250@cam.ac.uk

§ Technion City, Haifa 3200003, Israel. E-mail: roiri@ie.technion.ac.il

# 1 Rue Maurice Arnoux, 92120 Montrouge, France. E-mail: thierry.poibeau@ens.fr

\*\* Science Park 107, 1098 XG Amsterdam, Netherlands. E-mail: shutova.e@uva.nl

†† English Faculty Building. E-mail: alk23@cam.ac.uk

challenging to transfer NLP models across languages or to develop systems that apply to a wide range of languages. As a consequence, the availability of NLP technology is limited to a handful of resource-rich languages, leaving many other languages behind. Understanding linguistic variation in a systematic way is crucial for the development of effective multilingual NLP applications, thus making NLP technology more accessible globally.

Cross-lingual variation has several undesired consequences for NLP. Firstly, the architecture and hyper-parameters of most algorithms, allegedly language-independent, are fine-tuned and tested on a small set of languages. However, when applied to new languages, their performance decreases substantially as they inadvertently incorporate language-specific biases (Bender 2009, 2011). Moreover, state-of-the-art machine learning models typically rely on supervision from labeled data. Since the vast majority of the world's languages lack hand-annotated resources, such models cannot be easily trained for them (Snyder 2010). Finally, independent models for individual languages are often preferred to a single joint model of multiple languages and, as a result, useful cross-lingual information is neglected (Pappas and Popescu-Belis 2017). Yet, several experiments have demonstrated that such information leads to performance improvements over monolingual baselines (Ammar et al. 2016; Tsvetkov et al. 2016) by utilizing larger (although noisier) datasets and capitalizing on the fact that languages disambiguate each other.

Over time, many approaches have been put forth in multilingual NLP to mitigate some of these problems. They include unsupervised models that do not assume the availability of manually-annotated resources (Snyder and Barzilay 2008; Cohen and Smith 2009; Vulić, De Smet, and Moens 2011, *inter alia*); the transfer of models or data from resource-rich languages to resource-poor languages (Padó and Lapata 2005; Khapra et al. 2011; Das and Petrov 2011; Täckström, McDonald, and Uszkoreit 2012, *inter alia*); joint learning from multiple languages (Ammar et al. 2016; Johnson et al. 2016, *inter alia*); and the creation of multilingual distributed word representations (Mikolov, Le, and Sutskever 2013, *inter alia*).

These approaches are grounded in a principled theoretical framework provided by Linguistic Typology. This discipline aims at comparing the world's languages systematically, based on the empirical observation of their variation with respect to cross-lingual benchmarks (Comrie 1989; Croft 2002). Its documentation effort has resulted in off-the-shelf typological databases, e.g. the World Atlas of Language Structures (WALS) (Dryer and Haspelmath 2013), which can serve as a source of features, guidance in algorithm design, and criteria for data selection in multilingual NLP. However, as such databases tend to be incomplete, automatic techniques have been proposed to infer typological information from linguistic data, with the aim of obtaining missing or finer-grained values.

The development and usage of typological information for NLP, as well as the field of multilingual NLP in general, has achieved extraordinary progress recently. Hence the need for a survey that both covers up-to-date experiments and aims for comprehensiveness. Via an extensive review and analysis of existing typological resources and relevant NLP models, and discussion of their unexplored potential, this article aims to provide a platform for novel and integrative research on linguistic variation and multilingual NLP, thus facilitating progress in this important area of language technology.

In particular, this article covers tasks pertaining to all levels of linguistic structure. In addition to the morphosyntactic level and already widespread typology-savvy techniques such as 'selective sharing' (Täckström, McDonald, and Uszkoreit 2012), already surveyed by O'Horan et al. (2016), it also includes contributions about the

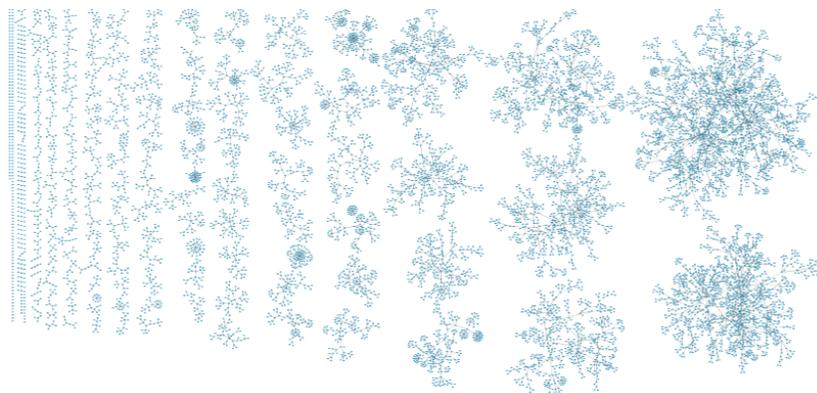


Figure 1: Networks of language families according to Glottolog data.

use of typological information in phonology and semantics. As research in these areas is still preliminary, we will pinpoint straightforward extensions to new tasks and the corresponding relevant typological features, based on the already established techniques.

Expanding on the previous surveys (O’Horan et al. 2016; Bender 2016), this article also provides an analysis of how typological constraints can be integrated in NLP methods from a machine learning perspective. In particular, we discuss such questions as: (1) how can typological information bias a model towards specific language properties?; (2) how can it be used to tie the parameters of different languages together?; (3) how can it aid in separating private (language-specific) from shared (language-invariant) parameters in multilingual models? Finally, we discuss some of the problems already pointed out in previous surveys (Sprout 2016) concerning the integration of categorical and manually-defined database features into probabilistic machine learning models. This survey explores the idea of a new approach to typology, that guarantees gradience in its categories through continuous representations, and bottom-up development of typological information from raw data through automatic inference.

The survey starts with an overview of Linguistic Typology, providing a detailed discussion of the available typological databases (§ 2). We then review the multilingual NLP tasks and techniques — a domain where typological information is potentially or has already been borne out to be useful (§ 3). Afterwards, we examine the selection and development of typological information for NLP (§ 4), and how it is implemented in a range of architectures, spanning from Bayesian models to artificial neural networks (§ 5). Finally, § 6 aims at identifying trends and future developments in this emerging field, whereas § 7 summarizes our key findings and conclusions.

## 2. Overview of Linguistic Typology

In this section, we focus on Linguistic Typology. Firstly, we illustrate the cross-lingual variation of the world’s languages, and show their genealogical and geographical relationships (§ 2.1). Afterwards, we describe Linguistic Typology as an empirical and systematic discipline (§ 2.2). We provide a summary of publicly accessible databases that store typological information in § 2.3, and discuss interaction with other fields in § 2.4.

## 2.1 Languages of the World

Providing the number of languages currently spoken around the world is challenging because of the difficulty of defining what constitutes a ‘language’. Languages are abstractions: every speaker has a lexicon and grammar slightly different to that of other members of the same community. A language is an ensemble of micro-variations (i.e. idiolects), rather than a monolithic system. The differences among idiolects are gradient and range from slight variations to the absence of mutual intelligibility. The latter is an important criterion for a definition of a language. Yet also cultural and political factors play a role, which is what led to the Weinrich’s famous saying that “a language is a dialect with an army and navy.” Bearing these caveats in mind, typological databases do attempt to define the number of languages in the world, as traditionally spoken by a community as their principal mean of communication. These are 7748 according to Glottolog (Hammarström et al. 2016), and 7097 according to Ethnologue (Lewis, Simons, and Fennig 2016).<sup>1</sup>

According to Ethnologue, the world’s languages can be organized into 141 families. The members of each family descend from the same ancestor language. Their differentiation is the result of the cumulative effect of mutations every language has undergone over time. Moreover, there are 88 isolates, i.e. languages without any relative. Each family is structured as a genealogical tree whose leaves are current languages and whose intermediate nodes are subgroups sharing common innovations (Ross 1997), possibly corresponding to an extinct language from a previous generation. For instance, both Sikkimese and Ladakhi originate from Old Tibetan and are part of the Tibetic branch within the Sino-Tibetan family.

To facilitate the visualization of language families, we show a plot of their genealogical trees in Figure 1. As it emerges clearly, the distribution of languages over families is fairly unbalanced. In fact, 6 families (Afro-Asiatic, Austronesian, Indo-European, Niger-Congo, Sino-Tibetan, and Trans-New Guinea) alone account for 63.13% of the world’s languages.

From the geographical point of view, languages are not equally distributed across macro-areas (Africa, Americas, Asia, Europe, and Pacific). Also the density of languages in an area is not proportional to their population. For instance, the 287 languages of Europe (4.0% of the world total) have around 1,673 M speakers (25.7%), whereas the 1,313 languages of the Pacific (18.5%) have around 7 M speakers (0.1%). The number of speakers for a language is expected to lie between 1 K and 10 K (which covers 1,979 languages, the 27.9 %), and follows a normal distribution: languages with more than 100 K (8, 0.1%) or less than 10 speakers (132, 1.9%) are highly infrequent.

Finally, the amount of documentation available for each language varies considerably. In fact, most languages are not recorded in written form. The 34.4% of the world’s languages are threatened, not transmitted to younger generations, moribund, nearly extinct or dormant, whereas the 34% is vigorous but has not developed yet a system of writing. Often, the communities of their speakers are too fragmented or remote to collect speech transcriptions. Even when we have some text available, it may be extremely scarce or unreliable. Annotated data or meta-linguistic information is even rarer because its creation requires additional resources. Considering currently available NLP resources, the largest existing collection of syntactic treebanks, Universal Dependencies (Nivre et al.

---

<sup>1</sup> These counts do not include unattested, pidgin, whistled, and sign languages.

2016), covers only 47 languages and only two languages have more than 1 M words. Overall, only a handful of the world's languages have large-scale, diverse annotations.

## 2.2 Empirical and Systematic Comparison

Linguistic Typology is a discipline that studies the variation across languages through their systematic comparison (Comrie 1989; Croft 2002). Comparison is challenging because linguistic categories cannot be predefined (Haspelmath 2007). There is a lot of cross-linguistic variation in lexicons and grammars and newly discovered languages often display unexpected properties. For instance, if we want to consider how languages express the passive voice, we soon discover this is a meaningless question for several languages like Qawasqar (isolate), which lack it altogether and behave totally differently from e.g. English. Rather, a *tertium comparationis* external to the cross-lingual data is required. In particular, the comparison should be based on *functional*, rather than *formal* criteria. According to functional criteria, the relevant question for the above example would be: how do languages emphasize the referent with semantic role of patient? We can distinguish between *constructions*, abstract and universal functions, and *strategies*, the type of expressions adopted by each language for a specific construction (Croft et al. 2017).

The *definition* of benchmarks for cross-lingual comparison is carried out jointly with *documentation* (Bickel 2007a, p. 248). Documentation is empirical in nature and involves both collection and observation of linguistic data. The resulting information is stored in large databases (see § 2.3) of attribute-values (this pair is henceforth referred to as *typological feature*), where each attribute corresponds to a construction and each value to the most widespread strategy in a specific language.

Given this evidence, it is possible to perform an *analysis* of cross-lingual patterns that emerge beyond chance, both synchronically and diachronically. One common observation, arising from such analysis, is that cross-lingual variation is bounded and not random (Greenberg 1966). Typological features tend to be interdependent: the presence of one feature may condition the likelihood of another (in one direction or both). Another observation is that some languages seem intuitively more plausible than others, since some typological features are rare while others are frequent. At the extremes of this spectrum lie 'impossible' and universally true features. The so-called *absolute universals* (e.g. all spoken languages have vowels (Croft 2002, ch. 3)) are understood to be shared by only the attested languages, rather than by all of them. *Implicational universals* are tendencies rather than actual rules (Corbett 2010): for example, if a language (such as Hmong Njua, Hmong-Mien) has prepositions, then the genitive-like modifier follows its head; instead if a language (such as Slavey, Na-Dené) has postpositions, their order is swapped. But exceptions are known: Norwegian has prepositions but genitives precede nouns.

Cross-lingual variation applies to all the levels of linguistic analysis. The seminal works on Linguistic Typology were concerned with morphosyntax, mainly morphological systems (Sapir 2014 [1921], p. 128) and word order (Greenberg 1966). This level of analysis deals with the form of meaningful elements (morphemes and words) and their combination. There is also work on phonology, i.e. comparison of elements that are distinctive but meaningless in isolation (phonemes). The systematic comparison of features at these levels 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 Lezgian (Northeast

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

On the other hand, *semantic* typology studies the semantic and pragmatic levels. This area was pioneered by anthropologists with works on kinship (d'Andrade 1995) and colors (Berlin and Kay 1969). The main focus of semantic typology is how languages categorize concepts (Evans 2011) in the lexicon, in particular with respect to the 1) number (granularity), 2) division (boundary location), and 3) membership criteria (grouping and dissection). For example, consider the event of *opening*. 1) It lexicalized as a single verb in English, but it can be split into a series of sub-events in a serial verb construction in Kalam (Highland Papuan) (Pawley 1993). 2) Moreover, it lacks a perfect equivalent in other languages, like Korean, where similar verbs overlap in meaning only in part (Bowerman and Choi 2001). 3) Finally, the English verb encodes the resulting state of the event, whereas an equivalent verb in another language like Spanish could rather express the manner of the event (Talmy 1991). Although variation in the categories is pervasive due to their partly arbitrary construal, languages share cognitive prototypes that constrain it (Majid et al. 2007).

Any cross-lingual generalization must be demonstrated through a representative sample of languages, but the selection of such a sample is highly problematic (Dryer 1989, *inter alia*). Firstly, the sample should be large enough to include even rarer features. Secondly, since many languages lack features due to sufficient documentation, the evidence is necessarily skewed by a bibliographical bias. Thirdly, similarities between languages do not always arise from language-internal dynamics but from external factors. In particular, they can be inherited from a common ancestor (genealogical bias) or borrowed by contact with a neighbor (areal bias) (Bakker 2010). As a consequence, since each language constitutes a single data point, data are radically sparse and not independent-and-identically-distributed (Cotterell and Eisner 2017).

Owing to genealogical inheritance, there are features that are widespread within a family but extremely rare elsewhere (e.g. such as 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 postponed to its noun simply because of their close proximity.

### 2.3 Documentation in Databases

Based on their observations in the variation among the world's languages, typologists have created open-source databases that collect typological features for a large number of languages. This information is invaluable for multilingual Natural Language Processing, as we will demonstrate in the rest of this survey. A list of current major typological databases is presented in Table 1. All of these databases organize the information in terms of universal attributes and language-specific values. Examples of such features can be found in the rightmost column of the Table. Moreover, some databases describe the features with examples (e.g. sentences for SSWL and ValPaL, or records for LAPSyD) and provide some theoretical discussion.

Some databases are general-purpose and store information pertaining to several levels of linguistic description. These include the World Atlas of Language Structures (WALS) (Dryer and Haspelmath 2013) and Atlas of Pidgin and Creole Language Structures (APiCS) (Michaelis et al. 2013). Moreover, there exist meta-repositories that wrap several databases together. An example is the URIEL Typological Compendium

Name	Levels	Coverage	Examples of feature
World Loanword Database (WOLD)	Loanwords (lexicon)	41 languages; ~2000 values; 24 attributes	HORSE Quechua : <i>kaballu</i> borrowed (24) Sakha : <i>silgi</i> no evidence (18)
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)
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)
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	ɸ AND ʈ Sindhi : yes (1) Chuvash : no (421)
PHOIBLE Online	Phonology	2155 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)
Intercontinental Dictionary Series (IDS)	Lexical semantics	329 languages; 1310 attributes	WORLD Russian : <i>mir</i> Tocharian A : <i>ārkišoṣi</i>
URIEL Typological Compendium	Phonology, Morphosyntax, Lexical semantics	8,070 languages; 284 attributes; ~439,000 values	CASE IS PREFIX Berber (Middle Atlas) : yes (38) Hawaian : no (993)
Automated Similarity Judgment Program (ASJP)	Lexical Semantics	7,221 languages; 40 attributes	I Ainu Maoka : <i>co7okay</i> Japanese : <i>watashi</i>
AUTOTYP	Morphosyntax	825 languages, ~1000 attributes	PRESENCE OF CLUSIVITY 'Kung (Ju) : false Ik (Kuliak) : true

Table 1: An overview of major publicly accessible databases of typological information. The table lists them in the order of creation, specifies the linguistic level they deal with and their coverage, and gives example features. In the examples, for each attribute (top, in small capital) we report two possible values (RHS) with a language name (LHS) and the total count of languages belonging to such type.

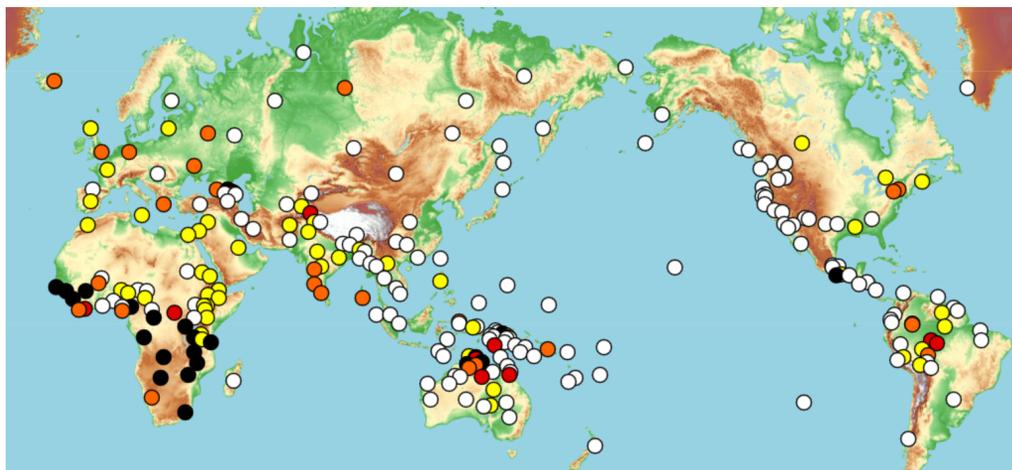


Figure 2: Number of genders in the world’s languages according to WALS (Dryer and Haspelmath 2013): none (white), two (yellow), three (orange), four (red), five or more (black).

(Littel, Mortensen, and Levin 2016), which standardizes features through binarization, fills in their missing values, and associates each language to a family and an area.

Among all databases, WALS has been the most widely used in NLP. In this resource, 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, and 129-138 with the lexicon, and 139-142 with other. WALS features can be plotted on a map to visualize their distribution. As an example, see Figure 2: the number of genders in nouns is the attribute, and each language is color-coded according to its value (an integer).

Some other databases only cover features at a specific level of linguistic description. For example, Syntactic Structures of the World’s Languages (SSWL) (Collins and Kayne 2009) and AUTOTYP (Bickel et al. 2017) focus on syntax only. In the former, attributes are constructions and values are strategies. In the latter, each language has multiple entries consisting of individual markers, and language-wide generalizations can be derived automatically. The Valency Patterns Leipzig (ValPaL) (Hartmann, Haspelmath, and Taylor 2013) provides verbs as attributes and predicate-argument structures as their values. As 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. Other databases concern various aspects of semantics. The World Loanword Database (WOLD) (Haspelmath and Tadmor 2009) documents loanwords, source words and donor languages. 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.

These databases have shortcomings that limit their usefulness. In particular, they suffer from inconsistencies due to the diversity of contributors, who might have followed different criteria and sources which might be outdated (Daumé III and Campbell 2007,

p. 2). For the same reasons, there are many gaps in attributes and missing values. Most databases also fail to account for the variation within a single language: for each language, only its majority value is reported rather than the full range of possible values and their corresponding frequencies. For example, the preponderant order in Italian is Adjective before Noun, but the opposite order is sometimes attested. Databases tend to neglect this information. Fortunately, many solutions have been devised to standardize features and extend the coverage of databases to unseen values and intra-language minority values. We will survey these solutions in § 4.2.

Typological databases are further limited by hierarchies and applicability of features. Firstly, some features are relevant, 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”. This creates a dependency between the two. 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. Features can also be redundant, and subsume partially or totally other features. For instance, WALs feature 81A “Order of Subject, Object and Verb” is the sum of WALs 82A “Order of Subject and Verb” and 83A “Order of Object and Verb”.

## 2.4 Recent Advancements

Traditional typology arose from early attempts to classify languages into ideal types (von Schlegel 1808, *inter alia*) that were later systematized by Greenberg (1963) in the probabilistic and empirical method sketched in § 2.2. This method was aimed at unveiling universal tendencies, and ultimately determining the limits of possible languages (Bickel 2007b). Nevertheless, typology has progressively emancipated from this original goal, and new perspectives have emerged advocating from more integrated and interdisciplinary methods. In particular, Nichols (1992) proposed a science of population typology answering “what’s where why?” and taking into account not only the variation, but also its diachronic evolution, its geographical distribution, and its cognitive and cultural underpinnings.

Thanks to this integrated approach, the temporal stability and horizontal diffusibility of typological features have been borne out to vary (Dediu and Cysouw 2013), partly in universal ways partly in lineage-specific ways (Dunn et al. 2011; Dediu and Levinson 2012). For instance, most features related to word order tend to be conservative (Daumé III 2009). On the other hand, some features are instable and provoke ripple effects. To sum up, typological generalizations can be considered as recurrent solutions in time and space, and outliers as rare happenstances triggered by unlikely preconditions (Evans and Levinson 2009).

Language is a hybrid biological and cultural system. These two components co-evolved in a twin track, with independent development but also mutual interaction (Durham 1991). Scholars have alternated in stressing these components to motivate the forces affecting language mutation and typological variation. Bickel (2015) differentiates between functional and event-based theories. Functional theories involve cognitive and communicative principles, while event-based theories emphasize the imitation of patterns found in other languages, which depends on its prestige and the contingencies of the contact.

The functional principles can be related to the content of an expression, such as the iconicity between form and meaning, or to (possibly conflicting) factors associated to its usage, such as its frequency or processing complexity (Cristofaro and Ramat

1999). As a consequence, patterns that are easy or widespread sediment in the grammar (Haspelmath 1999, *inter alia*). Similarly, these principles allow the speakers to draw similar inferences from similar contexts, leading to locally motivated pathways of change through the so-called *grammaticalization* process (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.

A *desideratum* for any artificial model that aims to mimic the faculty of language is to act in accordance of these above-mentioned cognitive principles. It has been noted by Bybee and McClelland (2005) that these principles can be adequately captured by the currently popular approach within NLP, namely neural networks. In particular, they argue that such architectures are sensitive to both local (contextual) information and general patterns, as well as to the frequency of use. They “address the issue of gradience and specificity found in postulated units, categories, and dichotomies such as regular and irregular.” This happens in so far as linguistic knowledge is represented by the strength of connections among processing units rather than rules as combinatorial systems would postulate.

This chapter has discussed cross-lingual variation and how the field of Linguistic Typology documents this variation in databases and explains it through historical, geographical, and cognitive causes. Understanding of cross-lingual variation can support the development multilingual NLP models. For instance, Bender (2009) shows how even sequential language models presuppose a rigid syntax that is absent in free-word-order, morphologically rich languages. Hence the recent trend to integrate also character-level information into language models (Gerz et al. 2018, *inter alia*). In the next sections, we focus on multilingual NLP and describe how it can be grounded on typological assumptions and enriched with typological features.

### 3. Overview of Multilingual NLP

The scarcity of resources is a major hurdle to any endeavor in multilingual NLP. State-of-the-art algorithms are based on supervised learning, hence their performance depends on the availability of manually crafted datasets 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. The immense range of possible tasks and languages makes the aim of a complete coverage unrealistic.

One solution to this problem, that has been explored by the research community, is disposing with annotated resources altogether through unsupervised learning, as discussed in § 3.1. However, the performance of this approach lags behind state-of-the-art algorithms. A more effective way to overcome the hurdle is either transferring models/data from resource-rich to resource-poor languages (§ 3.2) or learning joint models from annotated examples in multiple languages (§ 3.3) in order to verge on language independence. These approaches tend to leverage universal, high-level delexicalized features (e.g., PoS tags, dependency relations). However, the incompatibility of the (language-specific) lexica can be countered, too, by mapping equivalent words into the same multilingual semantic space through representation learning (§ 3.4).

This section provides a general background concerning all these strategies for multilingual NLP, whereas the subsequent sections focus on how they can be guided by typological information in particular, with respect to its selection (§ 4) and integration

(§ 5). In fact, these approaches often require some guidance with respect to several aspects of machine learning: data selection, feature engineering, and algorithm design. However, the entirety of this section may be also considered as a chart of the domains where typology may be potentially beneficial. This encompasses ways to inject external linguistic knowledge into a model, tasks where multilingual information is crucial, and the effects of sharing parameters in joint models. Hence, we speculate on these future perspectives in § 6. For this reason, we stress here that this section’s overview is not exhaustive, but limited to examples that corroborate our discussion.

### 3.1 Multilingual Unsupervised Learning

In this article, we refer to unsupervised learning as the set of methods that infer probabilistic models of the observations given some latent variables. In other words, it unravels the hidden structures from unlabeled data. Due to this, it is in principle suitable for any language in isolation, since it does not presuppose the availability of any resource. Meanwhile, it has been widely employed for multilingual applications, substantiating the notion that languages disambiguate each other (Snyder and Barzilay 2008, *inter alia*). In particular, multilingual applications include morphological segmentation (Snyder and Barzilay 2008), part-of-speech tagging (Snyder et al. 2009), semantic role labeling (Titov and Klementiev 2012), grammar induction (Cohen and Smith 2009), word sense discrimination (Navigli 2009), topic modeling (Vulić, De Smet, and Moens 2011), and neural machine translation (Artetxe et al. 2017, *inter alia*).

A major tendency grounded in the Bayesian framework is prearranging latent variables that record the similarities across languages (Snyder and Barzilay 2008) while training monolingual models on parallel data. For instance, an overlap in characters between morphemes may be a clue of their equivalence. The enhancement from a monolingual baseline grows with the number of languages involved (Snyder et al. 2009) and holds true for most of their combinations (although it varies markedly) (Naseem et al. 2009). In a similar spirit, monolingual objectives can be regularized with a constraint that penalizes disagreement in predictions across languages at inference time (Titov and Klementiev 2012). Nevertheless, Garg and Henderson (2016) raise a doubt on the usefulness of parameter sharing compared to simply procuring more monolingual examples.

However, multilingual corpora are often non-parallel, but comparable at most. In this scenario, parameters can be softly tied across languages through a shared logistic normal prior distribution (Cohen and Smith 2009) or estimated through multilingual Latent Dirichlet Allocation topic modeling (Vulić, De Smet, and Moens 2011). This technique maintains that words stem from language-specific vocabularies, but the latent variable generating topics is universal. It allows to tease out dictionaries of word translations, measure semantic similarity, and extract knowledge cross-lingually.

Another approach is based on disambiguating words by incorporating their translations into their representation. For instance, the senses of a polysemous word may be revealed as separate lexicalizations in other languages. Ide, Erjavec, and Tufis (2002) encode word occurrences as vectors of multilingual contexts, and subsequently cluster them. From these clusters, it is possible to further bootstrap a supervised model (Diab 2003). Although this method outperforms its monolingual counterpart, it suffers from a bottleneck in resource availability (Navigli 2009).

Some algorithms can be weakly supervised by providing prior knowledge declaratively. Bilingual lexica can equate the probabilities associated with translationally equivalent words for latent topic modeling (Zhang, Mei, and Zhai 2010). Naseem et al.

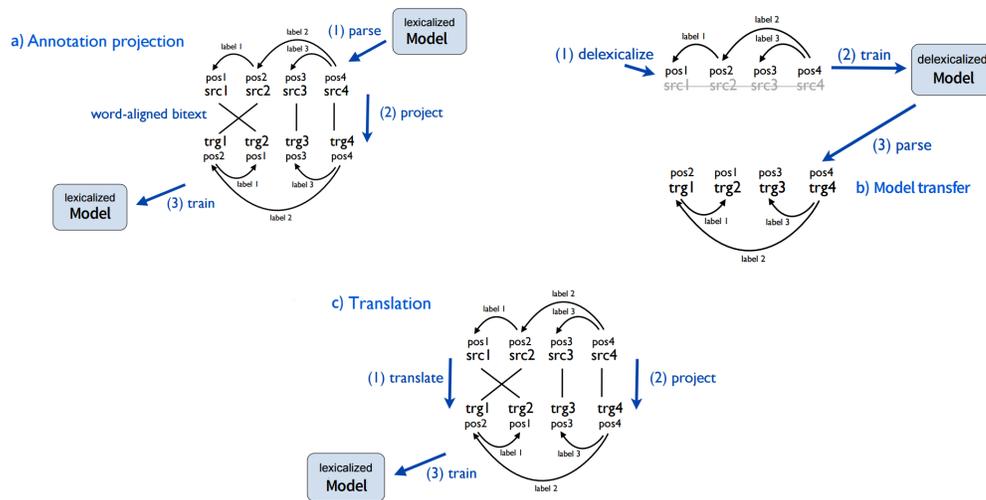


Figure 3: Three methods for language transfer: a) annotation projection, b) model transfer, and c) translation. The image is a revised version of the one appearing in [Tiedemann \(2015\)](#). As the example deals with parsing, projection regards dependencies and the delexicalized model is trained on PoS tags.

(2010) and [Grave and Elhadad \(2015\)](#) enforce some constraints on the structures the model is permitted to induce: they manually declare which PoS tags can be linked by a dependency. For instance, adjectives are always governed by nouns, and never the opposite. This sort of prior knowledge is ostensibly driven by typology, as it consists of absolute universals.

Unfortunately, unsupervised learning in general is not competitive in terms of performance with supervised learning ([Täckström, McDonald, and Nivre 2013](#)), and is plagued by attraction to local optima and slow convergence during optimization ([Spitkovsky, Alshawi, and Jurafsky 2011](#)). For these reasons, language transfer (§ 3.2) and multilingual joint learning (§ 3.3) have been recently preferred to perform multilingual NLP tasks. However, the intuition behind sharing parameters, disambiguating with multi-lingual contexts, and providing prior universal knowledge lends itself to these approaches as well.

### 3.2 Language Transfer

Linguistic information can be transferred to resource-poor languages, provided that it exists for some resource-rich languages: these are commonly referred to as target languages and source languages, respectively. Several methods have been developed for language transfer ([Agić et al. 2014](#)): they include annotation projection, (de)lexicalized model transfer, and translation. In this section, we review several representative methods for the three strategies and highlight their advantages and drawbacks. In general, language transfer is challenging because its requirement is to match sequences with different lexica and word orders ([Greenberg 1963](#)) or syntactic trees of translationally equivalent sentences with different (anisomorphic) structures ([Ponti et al. 2018](#)). Hence,

a focus of this section is on adaptation: how can linguistic information be tailored to a target language?

**Annotation projection** (schematized in Figure 3a) was inaugurated by the seminal work of Yarowsky, Ngai, and Wicentowski (2001) and Hwa et al. (2005). They word-align a source labeled text with a target raw text and project the annotation: this is defined as direct projection. Since this process is relatively noisy even when alignments are perfect, it was later refined by propagating labels over multiple steps based on bilingual graphs constructed with distributional similarity functions (Das and Petrov 2011) or constituents (Padó and Lapata 2009). In a similar vein, model expectations on labels (Wang and Manning 2014) or sets of most likely labels (Khapra et al. 2011; Wisniewski et al. 2014), rather than single categorical labels, can be projected in order to preserve the uncertainty of the source model. This is defined as soft projection.

The projection of structures (as opposed to sequences) such as syntactic trees is often partial and approximate, as it involves sets of vertices (e.g., words) and edges (e.g., dependencies) simultaneously. It is based on constraints that range from general structure preservation to language-specific rules (Ganchev, Gillenwater, and Taskar 2009) and can be relaxed over time (Rasooli and Collins 2015). In this case, too, the projection may regard model expectations on edges (weighted by the confidence in vertex alignments) (Agić et al. 2016).

If the projection is direct, the annotation can support the training of a target-side supervised model (Yarowsky, Ngai, and Wicentowski 2001). If it is soft, it can constrain target-side unsupervised models by reducing the divergence between their expectations and the source ones (Wang and Manning 2014; Ma and Xia 2014). If it involves multiple labels, the target model is supervised through ‘ambiguous learning’ (Khapra et al. 2011; Wisniewski et al. 2014).

This method can be enriched with auxiliary linguistic resources. Token-level constraints on labels (Li, Graça, and Taskar 2012; Täckström et al. 2013) or expectations (Ganchev and Das 2013) imposed by the alignment are combined with type-level constraints extracted from dictionaries during the projection. This can also be refined at a later time through manually written correction rules (Hwa et al. 2005) and through cross-lingual Wikipedia links (Kim, Toutanova, and Yu 2012).

**Model transfer** (illustrated in Figure 3b) instead is based on training a model on a source language and testing it on a target language (Zeman and Resnik 2008). Because of their incompatible vocabularies, the models are usually delexicalized prior to transfer and fed with language-independent (Nivre et al. 2016) or harmonized (Zhang et al. 2012) features. Alternatively, source delexicalized models can seed a target lexicalized model (McDonald, Petrov, and Hall 2011).

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 § 3.4). Zhang et al. (2016) use a source lexicalized model to regularize a target unsupervised Hidden Markov Model. The lexicalized model can be augmented with an adversarial objective that attempts to discriminate between source and target unlabeled texts, in order to enforce language independence on hidden representations (Chen et al. 2017).

As an alternative approach to lexicalization in absence of parallel data, **machine translation** (laid out in Figure 3c) of a source sentence is performed automatically (Banea et al. 2008) or through a bilingual lexicon (Durrett, Pauls, and Klein 2012) and its annotation is projected into a target language. If the pivot pairs in the dictionary are scarce, their annotation can be propagated to semantically similar words. This is defined as ‘prototype-driven learning’ (Prettenhofer and Stein 2010; Fernández, Esuli,

and Sebastiani 2015). The original and translated labeled data are sometimes combined in an ensemble Wan (2009) and used to bootstrap new examples from raw data. Translated documents are also employed to generate multilingual representations sharing semantic and sentiment content (Zhou, Wan, and Xiao 2016).

In all the aforementioned approaches, softening source constraints and enforcing linguistically motivated constraints aim at mitigating the cross-lingual differences in linear order and structures. Similarly, lexicalization and translation partially relate non-overlapping vocabularies (Agić et al. 2014). Nonetheless, these approaches remain impaired by their intrinsic limitations. The quality of alignments and translations deteriorates quickly in distant languages (Agić et al. 2016). Moreover, annotation projection presupposes the existence of parallel data (Agić, Hovy, and Søgaard 2015), and adds up noise incrementally in the pipeline unless trained jointly with the target model (Smith and Eisner 2009). Model transfer, on the other hand, overfits to the source language: although its features are often universal, it fails to conform to their language-specific interaction.

Typological information can be used to mitigate these limitations. In particular, it can simplify annotation projection, by carving source data (see § 5.3), and model transfer, by tying universal features together (see § 5.2) according to the properties of target languages. The employment of typology is even more momentous in multi-source transfer, where the methods of both Agić et al. (2016) and McDonald, Petrov, and Hall (2011) proved to generalize well.

### 3.3 Multilingual Joint Supervised Learning

Probabilistic models can be learnt jointly from multiple languages. This approach is endowed with multiple advantages compared to monolingual models: firstly, its performance often surpasses them as it can leverage more (although noisier) data (Ammar et al. 2016, *inter alia*). This is true especially 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, it improves over pure model transfer even with scant target labeled data, possibly resorting to active learning to discover the most relevant examples to annotate (Fang and Cohn 2017).

Secondly, multilingual joint learning is more cost-effective, because it allows to reduce the number of parameters, which scale up linearly, or even quadratically, with the number of languages (Pappas and Popescu-Belis 2017). Thirdly, it allows for zero-shot learning by triangulating or pivoting among languages. In other words, language transfer becomes possible even in absence of common resources for source and target, if a third (pivot) language can bridge between them (Johnson et al. 2016).

Among the niche NLP tasks, multilingual joint learning was borne out to achieve outstanding results in PoS tagging and NER (Yang, Salakhutdinov, and Cohen 2016), parsing (Ammar et al. 2016), discourse segmentation (Braud, Lacroix, and Søgaard 2017), language modeling (Tsvetkov et al. 2016, *inter alia*), neural machine translation (Johnson et al. 2016, *inter alia*), document classification (Pappas and Popescu-Belis 2017), and sentiment analysis (Niehues et al. 2011).

The key strategy for multilingual joint learning is parameter sharing (Johnson et al. 2016), as shown in Figure 4. More specifically, in state-of-art neural architectures input features and latent representations can be either private (language-specific) or shared across languages for distinct network components. For instance, these architectures incorporate both monolingual word embeddings and shared hidden layer parameters (Duong et al. 2015b), or shared parameters augmented with a language-dependent

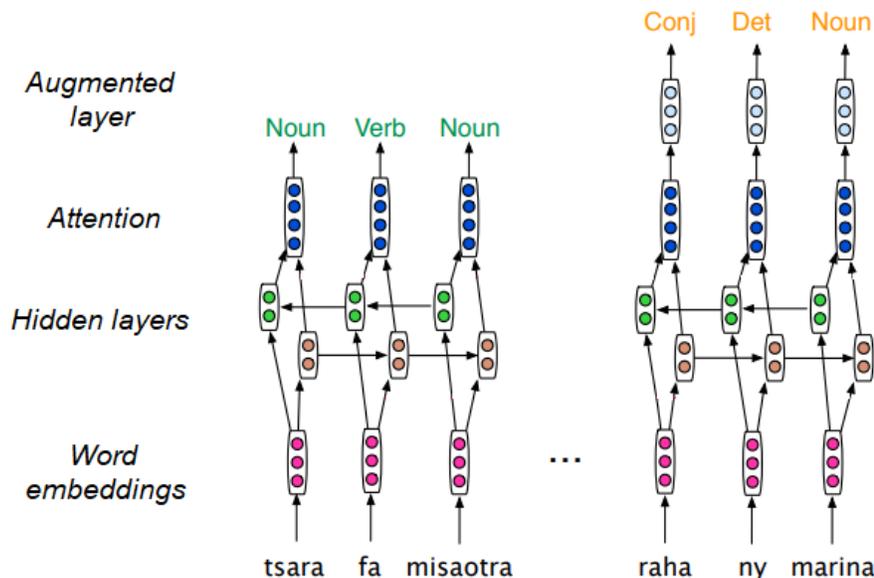


Figure 4: In multilingual joint learning, representations can be private or shared (among neurons with identical color) at each layer. Image adapted from Fang and Cohn (2017).

hidden layer (Fang and Cohn 2017). The choice of shared parameters also depends on the balance among the data for each language: the network proposed by Pappas and Popescu-Belis (2017) has a hierarchical structure and its attention mechanisms are always shared. However, word encoders and sentence encoders are shared effectively only in resource-lean scenarios.

Other discrepancies may arise from character encoding: Guo et al. (2016a) pool the representations of the inputs, but keep separate parameters for characters and action sequences of a transition-based parser, based on the assumption that languages differ in morphology and word order. On the contrary, Yang, Salakhutdinov, and Cohen (2016) share character encoders but not word encoders in order to leverage sub-string similarities of related languages. The problem of word and character compatibility is avoided altogether by Gillick et al. (2016), who train multilingual models directly on Unicode bytes.

The training procedure of these models has to be customized, since the data of each language are drawn from different distributions. An independent output layer may be specifically dedicated to individual languages (Braud, Lacroix, and Søgaard 2017): examples for each of them are fed together to the algorithm and their corresponding losses are weighted by the dataset size and summed (Duong et al. 2015b). Otherwise, examples from individual languages can be sampled individually with different probabilities (Guo et al. 2016a) or uniformly after oversampling (Johnson et al. 2016). Finally, multiple models can be trained separately: to bridge the language gap, the model can minimize the distance between their parameters (Duong et al. 2015a) or between latent representations of sentences in aligned data (Niehues et al. 2011; Zhou et al. 2015).

Information about the identity of the current language is sometimes provided in the form of input vectors (Guo et al. 2016a). These vectors can be learnt automatically in neural language modeling tasks (Tsvetkov et al. 2016; Östling and Tiedemann 2016) or neural machine translation tasks (Johnson et al. 2016; Ha, Niehues, and Waibel 2016). Ammar et al. (2016) instead considered the vector as a prior where they specified the language identity or some typological features. In language modeling, language identity can also be predicted in output at the token level to condition the prediction of the next token (Adel, Vu, and Schultz 2013).

As it emerged from this survey, crucial challenges involve tailoring the joint model toward a current language and striking a balance between private and shared neural network components. With respect to the former, multilingual learning is facilitated by explicit typological information in the form of input vectors (Ammar et al. 2016; Tsvetkov et al. 2016) (see § 5.2). Moreover, language-specific typological properties were successfully decoded from private representations (Malaviya, Neubig, and Littell 2017, see § 4.2) or visually identified in shared representations (Johnson et al. 2016). Therefore, unraveling the interaction between typological properties and neural network components seems to be a promising direction for future research (§ 6.3).

### 3.4 Multilingual Representation Learning

A large body of recent research in NLP is focused on learning dense real-valued vector representations or word embeddings (WEs), which serve as pivotal features in a range of downstream NLP tasks. They facilitate both language transfer and multilingual joint learning, as illustrated in § 3.2 and § 3.3, respectively. The extensions of WE models in multilingual settings abstract over language-specific features and attempt to represent words from both languages in a language-agnostic manner such that similar words (regardless of the actual language) obtain similar representations.

As opposed to static lexical repositories such as WordNet, the popularity of WE learning models stems from their adaptability and versatility. WEs can be automatically constructed from large corpora with little to no guidance, and they can be steered to capture multi-faceted linguistic similarity at the semantic (Mikolov et al. 2013; Rothe and Schütze 2015), syntactic (Levy and Goldberg 2014; Gouws and Søgaard 2015), or morphology levels (Botha and Blunsom 2014; Cotterell and Schütze 2015). Further, it is straightforward to inject external knowledge from structured knowledge bases and specialized linguistic resources – if these are available at all – to further influence the properties of such automatically induced WEs (Faruqui et al. 2015; Mrkšić et al. 2017; Vulić et al. 2017, inter alia).

The research on multilingual NLP reviewed in § 3.2 and § 3.3 makes use of various methods to generate multilingual WEs. We follow the classification proposed by Ruder (2018), whereas we refer the reader to Upadhyay et al. (2016) for an empirical comparison.

**Monolingual mapping** generates independent monolingual representations and subsequently learns a 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). Artetxe, Labaka, and Agirre (2017) explored a bootstrapping approach to acquire bilingual seeds. Alternatively, both spaces can be cast into a new, lower-dimensional one through canonical correlation analysis (CCA) based on dictionaries (Ammar et al. 2016) or word alignments (Guo et al. 2015). Figure 5 shows how equivalent words in the separate semantic spaces of different languages  $X$  and  $Y$  can be re-orientated through a learnt transformation  $W$ .

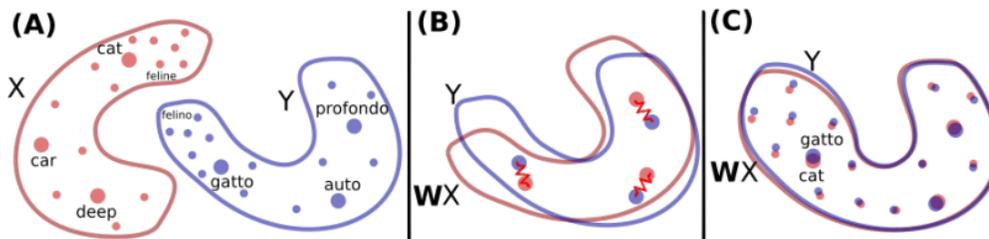


Figure 5: Bilingual mapping from [Conneau et al. \(2017\)](#): a linear mapping weight  $W$  is learnt to make monolingual semantic spaces of languages  $X$  and  $Y$  overlap.

**Pseudo-cross-lingual** approaches merge words with contexts of other languages and generate representations based on this corpus. [Xiao and Guo \(2014\)](#) substitute words with their translations in Wiktionary and force both of them to share the same vector. Other works resort to random machine translations, either before training ([Gouws and Søgaard 2015](#)) or on the fly to handle polysemy ([Duong et al. 2016](#)). Finally, the merge can result from random shuffling words in two languages from aligned documents ([Vulić and Moens 2015](#)).

**Cross-lingual training** jointly learns embeddings from parallel corpora and enforces cross-lingual constraints. After devising a composition function for words, some works minimize the distance of the resulting sentence representations in the two languages ([Hermann and Blunsom 2013](#)) or decode one from the other ([Lauzy, Boulanger, and Larochelle 2014](#)), possibly adding a correlation term to the loss ([Chandar et al. 2014](#)). With a distinctly different approach, [Søgaard et al. \(2015\)](#) leverage on the structured knowledge of Wikipedia by representing words by their occurrences in articles linked to the same concept and then learn dense vectors through dimensionality reduction.

**Joint optimization** takes into account both monolingual and multilingual constraints during training on parallel texts. [Klementiev, Titov, and Bhattarai \(2012\)](#) jointly learn distinct embedding models for each language and an alignment-based translation that regularizes them. [Zou et al. \(2013\)](#) create target embeddings as the product of the corresponding source embeddings in alignment-based matrices. A skipgram objective can be trained over both monolingual contexts and cross-lingual contexts ([Luong, Pham, and Manning 2015](#)) or can be regularized by the distance of the means of parallel sentences ([Gouws, Bengio, and Corrado 2015](#)). Finally, [Rotman, Vulić, and Reichart \(2018\)](#) treat images as language-independent constraints in addition to parallel texts.

We detect two main axes along which WE learning and linguistic typologies should cooperate. Firstly, monolingual WE methods are originally developed to fit English-language data; improvements to state-of-the-art models, such as the adoption of syntactic contexts in WE learning ([Levy and Goldberg 2014](#)), do not produce stable and predictable results when transferred cross-lingually ([Vulić and Korhonen 2016](#)). Data size issues aside, this varied cross-lingual performance suggests that different methods are more suited to modeling particular features, which may prove crucial for processing one language but less so for others (e.g., typically discarding morphology in WE modeling for English). It is thus not the case that “one model fits all” for monolingual WE generation in different languages. Future advances in WE learning for other languages should seek advice from typological knowledge to discern which features are more prominent in which

language, in order to achieve optimal performance. One step into this direction has been made recently by resorting to word embedding learning enriched by subword-level information (Bojanowski et al. 2017; Peters et al. 2018)

Secondly, multilingual WEs rely on the idea of a shared semantic (vector) space for data in two or more languages, induced in a scalable, data-driven manner. Yet it is not the case that all languages with arbitrary lexical profiles can simply be added into the same vector space to produce results usable for cross-lingual knowledge transfer. Typological factors should be of importance when: (1) making assumptions about how compatible the semantic spaces of multiple languages may be in the first place, (2) guiding development of more informed models by perturbing the spaces and make them akin topologically. Both these directions of research will be conjectured in more detail in § 6.

#### 4. Selection and Development of Typological Information

After having outlined the landscape of multilingual NLP in § 3, in the rest of this survey we demonstrate how typological information enhances work in this area, and provides a principled framework to exploit differences and similarities across languages. We start from illustrating in this section how typological information used in NLP models is selected from the databases listed in § 2.3, how it is preprocessed and encoded in a form that is compatible with NLP algorithms, and how missing and/or finer-grained features can be developed automatically. Subsequently, in § 5 we will explore how this information can support various NLP tasks.

The extraction of typological information from databases (§ 4.1) has focused on different feature subsets, mostly either a few word-order properties or the full range of properties available in a database. However, the documentation in such databases is often incomplete and heterogeneous, which limits their usability for NLP tasks.

In order to address these limitations, many methods have been developed to predict the missing values and acquire typological information automatically (§ 4.2). This results in a series of advantages over the database documentation. Firstly, these methods can account for the variation within single languages. Secondly, traditional types are discrete and partly arbitrary, as discussed in § 2.3. These methods can replace them with continuous representations of cross-lingual variation. A third advantage of automatic acquisition is allowing to explicitly model the correlations among features and with area and family.

Crucially, there is no single method that clearly outperforms the others because of these correlations. For instance, propagating the majority value from the other family members is effective for vertically stable features, but not for those unstable over time.

##### 4.1 Extraction from Manually Crafted Resources

Thus far, typological features in NLP models have been often sourced directly from manually crafted databases. However, the number of existing databases and features therein taken into account to date is small. In fact, the choice is strongly skewed toward a subset of word order features from WALS (Dryer and Haspelmath 2013, see § 2.3). Experiments that harness typology deal predominantly with morphosyntactic tasks, and this subset has so far been the most pertinent for this purpose (see § 5.4).

In Figure 6, we show the feature sets of this group of experiments. As it emerges, the features encode mostly the word order of nouns, verbs, and their modifiers. However, they do not overlap completely, as the subset firstly established by Naseem, Barzilay, and Globerson (2012) was adjusted minimally afterwards, for instance by discarding features

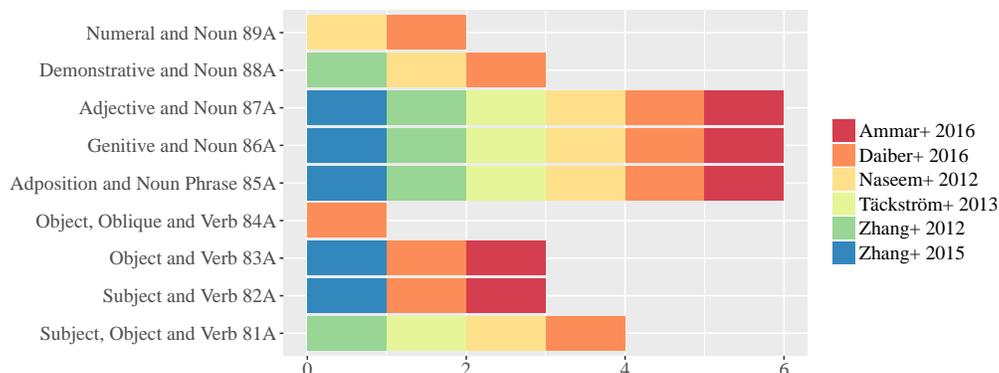


Figure 6: Different subsets of word-order features selected in different papers. The numbers refer to WALs ordering (Dryer and Haspelmath 2013).

with the same value for all the languages in the data. Moreover, because of hierarchies among features, selecting 81A often implies omitting 82A and 83A, and vice versa.

Although most of the papers reported here are limited to similar subsets, others are more comprehensive: the sample of Daiber, Stanojević, and Sima'an (2016) encompasses all the WALs features related to word order but also some that capture information about nominal categories (e.g. 'Conjunctions and Universal Quantifiers') and nominal syntax (e.g. 'Position of Tense-Aspect Affixes'). Berzak, Reichart, and Katz (2015) prune out all features from WALs not associated with morphosyntax or redundant, resulting in a total of 119 features. This feature set is augmented with a 104-dimensional binary vector, encoding whether each feature value in a given language agrees with the corresponding one in English. Tsvetkov et al. (2016) select 190 binarized phonological features from URIEL (Littel, Mortensen, and Levin 2016). These features encode the presence of single segments, classes of segments, minimal contrasts in a language inventory, and the number of segments in a class.

A small number of papers broadens the range of utilized typological features to the entire feature inventory of a given database. In particular Agić (2017) and Ammar et al. (2016) harvest all the features in WALs, while (Deri and Knight 2016) use all the features in URIEL. Similarly, Søgaard and Wulff (2012) utilize all the WALs features with the exception of phonological features. Finally, Schone and Jurafsky (2001) do not resort to basic features, but rather to "several hundred [implicational universals] applicable to syntax." These are drawn from the Universal Archive (Plank and Filiminova 1996).

The most crucial challenge to the creation of all-embracing and cross-lingually consistent feature sets remains the partial nature of the documentation available in manually crafted resources. For example, only about 17 percent of cells in the WALs language-feature matrix are currently populated. Nevertheless, the coverage for the languages involved in the above-mentioned papers is broader since they tend to be well-researched. For instance, 79.8 percent of the feature values are available on average for the 14 languages considered by Berzak, Reichart, and Katz (2015).

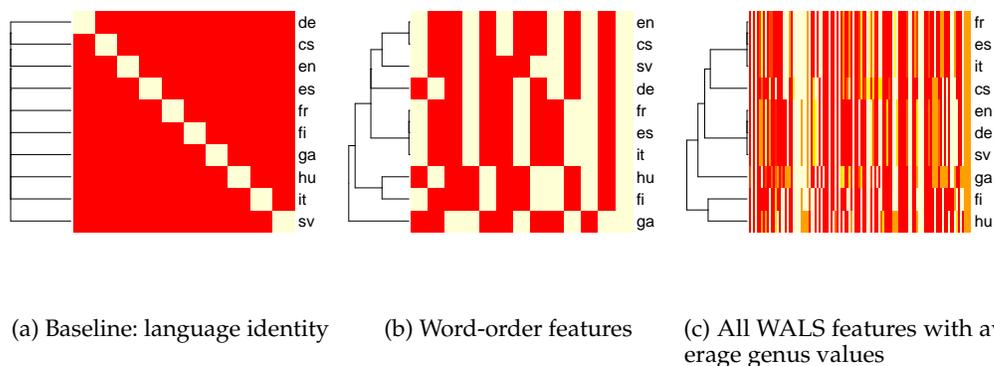


Figure 7: Heat maps of encodings for different subsets of typological WALS features taken from Ammar et al. (2016): rows stand for languages, dimensions for attributes, and color intensities for feature values. Encodings are clustered hierarchically by similarity.

Still, it is necessary to fill in the blank cells in order to restore the missing information and make the features comparable. In fact, features are usually encoded as vectors where each dimension is an attribute and each number represents the feature value. Moreover, these vectors are sometimes binarized (Georgi, Xia, and Lewis 2010): for each possible 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, 0 otherwise: note that this increases the number of features by a factor of  $\frac{1}{|a|} \sum_{i=1}^{|a|} \|v_{a_i}\|$ .

Unfortunately, this encoding strategy has many unwanted consequences. Firstly, the variables underlying database features are incompatible by nature: they include nominal, ordinal, and interval variables, as well as a blend of these. The binarization operation overshadows these differences. Moreover, because of the different applicability of features, filling in missing values may be meaningless, as some languages simply have no value for that attribute.

Despite the caveats of incomplete documentation and inconsistent feature nature presented above, typological resources do offer an abundance of underutilized valuable information. In fact, it is unclear whether a limited set of coherent features or the full, integrated database should be preferred. For a discussion based on their performance in comparable tasks, we refer to § 5.4.

In order to compare these types of features sets with respect to their content, consider the heat maps in Figure 7:<sup>2</sup> we take into account three feature sets appearing in Ammar et al. (2016). Rows represent feature encodings for single language, and the colors the feature values. In particular, Figure 7a is a single baseline with one-hot encoded language identities; Figure 7b is a subset of word order features; and Figure 7c shows a large set of WALS features where values are averaged by language genus. Finally, languages are hierarchically clustered through the complete linkage method according to the similarity of their typological feature encodings.

<sup>2</sup> 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.

Figure 7 reveals the impact of genealogical biases on predicting missing values: in Figure 7c the clusters are perfectly equivalent to known families and genera. However, also ‘golden’ word-order features fail to account for fine-grained differences between related languages: for instance, French, Spanish, and Italian receive the same encoding also in Figure 7b. Finally, the heat map in Figure 7a is completely uninformative. These examples show how difficult it is to choose and/or predict features that are non-redundant with other classifications (such as genealogy), fully discriminative, and informative.

## 4.2 Automatic Prediction of Typological Features

The partial coverage of existing resources sparked a line of research on automatic acquisition of typological information. Missing feature values can be predicted based on: i) heuristics from pre-existing or transferred morphosyntactic annotation, such as treebanks (§ 4.2.1); ii) propagation from other values in a database based on clustering or language similarity metrics (§ 4.2.2); iii) supervised learning with Bayesian models or artificial neural networks (§ 4.2.3); or heuristics based on co-occurrence metrics, typically applied to multi-parallel texts (§ 4.2.4). These strategies are summarized in Table 2.

Evaluation of methods for automatic prediction of typology is typically carried out using the existing feature documentation in typological databases, predominantly using WALS. However, the evaluation scores are hardly comparable, as they often result from different ways to partition WALS into training and test sets, considering different languages and features. Moreover, each strategy serves partially different purposes, and each is more suited to predict specific kinds of features. As a consequence, there has not been a clear preference in the literature for one evaluation strategy over the others.

Nevertheless, there is a general trend in opting for the automatic acquisition of typological features. Apart from filling in missing values, this allows to obtain information that is not recorded inside typological databases. Firstly, it accounts for the distribution of feature values within single languages, rather than just the majority value, possibly characterizing instantiations of a type in single examples rather than in languages as a whole. Secondly, this allows to avoid the problem of the partly arbitrary nature of cross-lingual categories (Haspelmath 2007, see 2.2 ) by representing them along a continuum.

**4.2.1 Morphosyntactic annotation.** Morphosyntactic feature values can be extracted through heuristics from morphologically and syntactically annotated text. In particular, word order features can be calculated by the token-based count of the directionality of equivalent dependency or constituency relations (Liu 2010). For instance, consider the tree of a sentence in Welsh from Bender et al. (2013) in Figure 8. The relative order of verb-subject, and verb-object can be deduced from the position of the relevant nodes *VBD* and *NN* (highlighted).

Typological information can also be harvested from Interlinear Glossed Texts (IGT). Such collections of example sentences are collated by linguists and contain grammatical glosses with morphological information. These can guide the 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.

Morphosyntactic annotation for typological prediction does not need to be pre-specified. It can also be projected from a source directly to several target languages

	Author	Details	Requirements	Langs	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
Propagation from database	Teh, Daumé III, and Roy (2007)	Hierarchical typological cluster	WALS	2150	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 dataset	1017	whole
Supervised inference	Bjerva and Augenstein (2018)	Task-based language vector	PoS tag dataset	27-824	phonology, morphology, syntax
	Takamura, Nagata, and Kawasaki (2016)	Logistic regression	WALS	whole	whole
	Murawaki (2017)	Bayesian + feature and language interactions	Genealogy and WALS	2607	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
Multi-alignments	Daumé III and Campbell (2007)	Implication universals	Genealogy and WALS	whole	whole
	Lu (2013)	Automatic discovery	Genealogy and WALS	1646	word order
	Wälchli and Cysouw (2012)	Sentence edit distance	Multi-parallel texts, pivot	100	motion verbs
Multi-alignments	Asgari and Schütze (2017)	Pivot alignment	Multi-parallel texts, pivot	1163	tense markers
	Roy et al. (2014)	Correlations in counts and entropy	None	23	adposition word order

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

through a multi-lingual alignment. For example, [Zhang et al. \(2016\)](#) transfer PoS annotation with a model transfer technique relying on multilingual embeddings created through monolingual mapping. After the projection, they predict feature values with a multi-class Support vector machine (SVM) trained on PoS tag n-gram features to predict typological features in WALS.

**4.2.2 Propagation within the database.** Another line of work seeks to increase the coverage of typological databases borrowing missing values from the known values other languages. The donor languages are chosen by clustering languages according to some criterion (e.g. genealogy) and propagating the majority value within the cluster, or by measuring language similarity according to some metric and propagating from nearest neighbors.

[Teh, Daumé III, and Roy \(2007\)](#) develop a Bayesian model for hierarchical clustering of languages according to typological features. After performing this operation, missing typological feature values can be inferred from the other languages in the same cluster. In this model, the prior is an exchangeable distribution over trees called Kingman’s coalescent. A full tree  $\pi$  is constructed bottom-up and greedily from observations  $\mathbf{x}$ , as a series of  $i = 1 \dots n$  coalescent events that merge two subtrees with leaves  $\rho l_i$  and  $\rho r_i$  and occurs at a time  $\delta_i$ . The choice of their combination is based on the product of a local prior  $e^{-\delta_i}$  and a local likelihood  $Z_{\rho_i}(x, \theta_i)$ . The probability of a hierarchical clustering is given by the product of each event with the tree root likelihood  $Z_{\rho_{-\infty}}(x, \theta_i)$ , as shown in Equation 1.

$$P(\mathbf{x}, \pi) = Z_{-\infty}(x, \theta_{n-1}) \prod_{i=1}^{n-1} e^{-\delta_i} Z_{\rho_i}(x, \theta_i) \quad (1)$$

Similarly, [Coke, King, and Radev \(2016\)](#) propagate the majority label among all languages of the same genus according to gold external knowledge. The sources and techniques of clustering for majority value propagation have been evaluated extensively by [Georgi, Xia, and Lewis \(2010\)](#). They demonstrate that typology is better than genealogical families for deriving effective clusters. Among the clustering techniques, k-means appear to be the most reliable as compared to k-medoids, the Unweighted Pair Group Method with Arithmetic mean (UPGMA), repeated bisection, and hierarchical methods with partitional clusters.

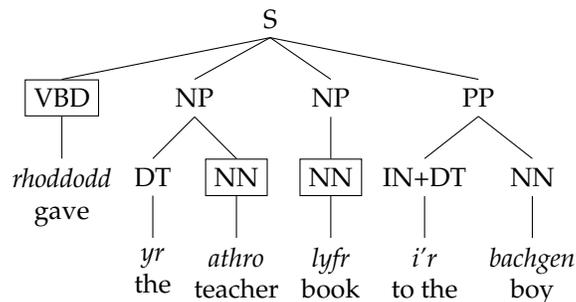


Figure 8: Constituency tree of a Welsh sentence.

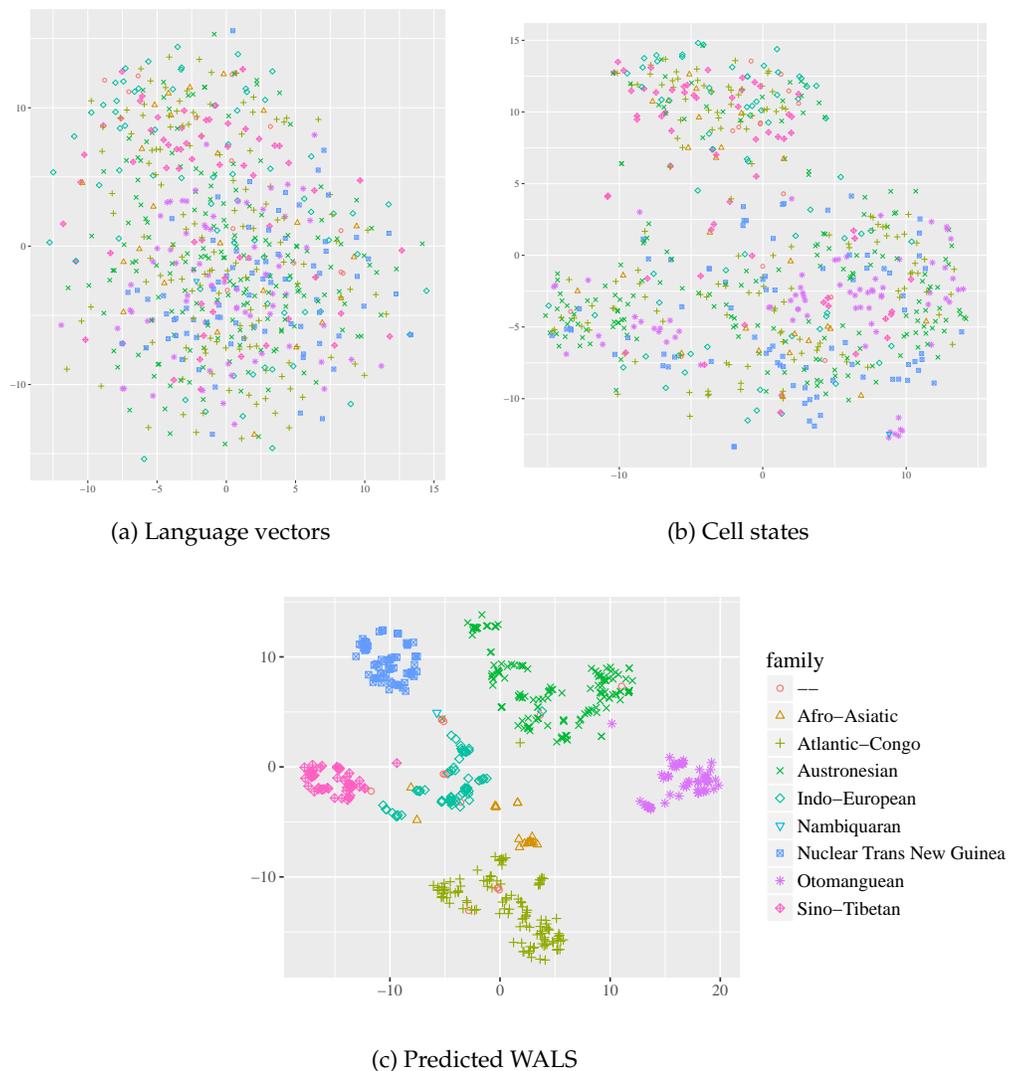


Figure 9: Dimensionality-reduced language representations after t-SNE.

Apart from automatic clustering, typological feature values can be propagated based on some measure of language similarity. For instance, [Littel, Mortensen, and Levin \(2016\)](#) take an average of genetic, geographical, and feature distances among languages from URIEL, and perform a weighted transfer from the  $k$  nearest-neighbor languages. [Berzak, Reichart, and Katz \(2014\)](#) derive a language distance measure based on delexicalized morphosyntactic features of English as a Second Language (ESL) texts. The intuition of this approach is that due to systematic first language influence on second language performance, similarity between languages can be approximated using similarity of morphosyntactic ESL usage by native speakers of those languages. After developing this metric, feature values are propagated by majority vote from the  $k$  nearest neighbor languages.

Language similarity measures can also rely on the distributed representation of each language. As opposed to completing current (discrete) typological databases which can then be used as feature pools for NLP models, the approaches which build such representations are based on a different set of assumptions: they recast the problem of learning language-relevant features into a continuous space. These representations are trained end-to-end as part of neural models.

For this purpose, [Malaviya, Neubig, and Littell \(2017\)](#) investigate two distinct approaches. Firstly, within a many-to-one multilingual Neural Machine Translation model, they concatenate an artificial token representing the identity of the current language to every input sentence, similarly to [Johnson et al. \(2016\)](#). As a result, the language identity gets encoded in the vector learned for this token. Alternatively, languages can be represented as the aggregated values of the hidden state of the encoder.

Distributed language representations implicitly embed features found in databases such as WALS, as borne out by [Malaviya, Neubig, and Littell \(2017\)](#): they concatenate the learned artificial tokens and the encoder hidden states and feed them to a logistic regression classifier. However discrete and continuous representations of typology appear to differ radically, as visualized in Figure 9. The Figure compares continuous representations based on artificial tokens (Figure 9a) and encoder hidden states (Figure 9b) with vectors of the WALS features available in URIEL (Figure 9c). All the representations were reduced to 2 dimensions using *t*-SNE, and color-coded according to the language family.

Not surprisingly, the information encoded in WALS vectors is akin to genealogical information, owing to intrinsic areal biases and because missing values are propagated within families. On the other hand, artificial tokens and encoder hidden states cannot be reduced to genealogical clusters. As shown in § 4.2.5, however, their ability to predict missing values is not degraded. This implies that closeness in the space of such representations is genuinely based on typological properties, rather than being biased by language-external factors. Overall, discrete and continuous representations appear to capture different aspects of the cross-lingual variation. For this reason, they are possibly complementary and could be leveraged together in the future.

[Bjerva and Augenstein \(2018\)](#) adopt the approach based on artificial tokens, but explore several more tasks across linguistic levels: phonology (grapheme-to-phoneme prediction and phoneme reconstruction), morphology (morphological inflection), and syntax (part-of-speech tagging). Typological feature values are inferred by k-NN classifiers according to the closeness of the token vectors. Whereas phonological tasks do not yield meaningful representations, morphosyntactic tasks are excellent proxies for the prediction, both of the feature subsets relevant to the corresponding linguistic level and the full set of database features.

**4.2.3 Supervised prediction.** Another line of research applied supervised classification approaches to predict feature values. For instance, [Takamura, Nagata, and Kawasaki \(2016\)](#) used logistic regression on WALS features leaving one language out as a development set in each iteration, and using a single feature as a gold label. [Wang and Eisner \(2017\)](#) provide supervision with both natural and synthetic languages. Given a PoS tagged corpus  $U$  the algorithm predicts the probability of a right directionality given a dependency  $r$  and a language  $l$  as  $P(d|r, l) \in [0, 1]$ . In particular, it minimizes an  $\epsilon$ -insensitive loss (so that it is not dominated by outliers) normalized by the probability of a relation  $r$  in language  $l$  (Equation 2).

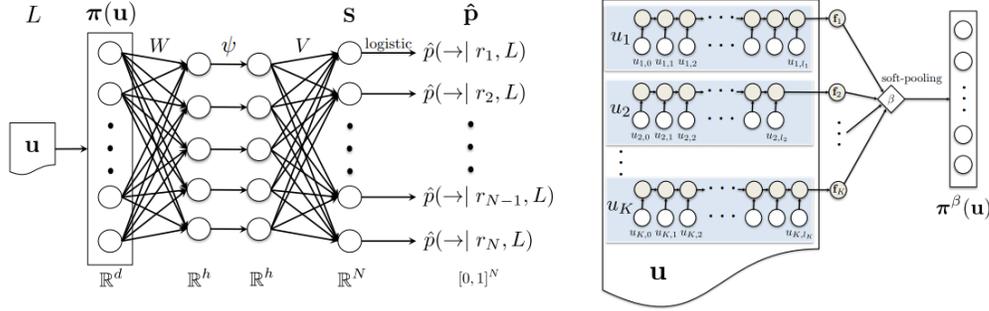


Figure 10: Architecture of Wang and Eisner (2017)'s supervised neural model: Feed-forward Neural Network (left) and feature extractor (right).

$$J = \sum_{r \in \mathfrak{R}} p(r|l) \max(|P(\hat{d}) - P(d)| - \epsilon, 0) \quad (2)$$

$$P(\hat{d}) = \sigma(V\psi(W\pi(u) + b_W) + b_V) \quad \pi^\beta(u) = \left( \frac{1}{n} \sum_{u_1}^{u_n} \text{GRU}(u_i)^\beta \right)^{\frac{1}{\beta}} \quad (3)$$

In turn, the prediction  $P(\hat{d})$  is the output of a feed-forward neural network with one hidden layer with non-linear activation  $\psi$  that sums over windows of tags  $u$  in the corpus  $U$ , as shown in Equation 3. In particular,  $\pi(u)$  extracts a feature vector of both hand-engineered co-occurrence features and neural features. The latter are extracted with a Gated Recurrent Unit (GRU) from each sentence and then soft-pooled with the inverse temperature  $\beta$ . The architecture for these equations is represented in Figure 10.

Within the Bayesian framework, Murawaki (2017) designs a model accounting for both implicational universals and genealogical/areal relationships. The traits of a language are represented as a series of binary latent parameters that can capture inter-feature dependencies. In turn, the series are linked to their phylogenetic and spatial neighbors. Each parameter  $k$  is associated with a vertical stability  $v_k > 0 \sim \Gamma(\kappa, \theta)$ , horizontal diffusibility  $h_k > 0 \sim \Gamma(\kappa, \theta)$ , and universality  $-\infty < u_k < \infty \sim \mathcal{N}(0, \sigma^2)$ . An auto-logistic model generates a binary parameter matrix  $Z$  based on these variables and functions that count features by family  $V_{z_{l,k}}$ , by area  $H_{z_{l,k}}$ , and in total  $U_{z_{l,k}}$  across each language  $l$ . This way, the three factors compete with each other for prevalence in the overall probability.

$$\theta_{l,m} = \sigma \left( \prod_{k=1}^K z_{l,k} w_{k,m} \right) \quad z_{l,k} = \sigma (v_k V_{z_{l,k}} h_k H_{z_{l,k}} u_k U_{z_{l,k}}) \quad (4)$$

In turn, as shown in Equation 4, the binary parameter matrix is combined with a weight  $W$  whose probability is drawn from Student's  $t$ -distribution with  $DF = 1$ , yielding a feature score matrix  $\theta_{l,m}$  for each language  $l$ 's  $m$ -th binarized typological feature. This way the latent feature matrix is able to generate surface typological features without missing values. This pipeline is schematized in Figure 11.

Cotterell and Eisner (2017, 2018) develop a generative model of vowel inventories (represented as either IPA symbols or acoustic formants) based on the cognitive principles of dispersion (phonemes are as spread out as possible in the acoustic space) and focalization (some positions in the acoustic space are preferred owing to the similarity of the main formants). Given a base set  $\mathcal{V}$  (all possible phonological inventories), the point process they develop returns a distribution over each subset  $V$ . In particular its probability is the determinant of a symmetric positive semidefinite matrix  $L$  which can be decomposed into the dot product of another matrix  $E$  and its transpose,  $P(V) \propto \det L_V$ . Each vector  $e_i$  is the embedding of a phoneme  $v_i$  and derives from a Multi-Layer Perceptron over the phoneme formants. The intuition is that the probability of phoneme pairs depends on the magnitude (its focalization) of both their embeddings and the sine of their angle (their dispersion), such that  $P(v_i, v_j) \propto (\|e_i\| \|e_j\| \sin\theta)^2$ .

Another, more indirect approach to supervised prediction is based on learning implicational universals of the kind pioneered by Greenberg (1963) with probabilistic models. For instance, once it has been established that the presence of ‘High consonant/vowel ratio’ and ‘No front-rounded vowels’ necessarily implies ‘No tones’, the missing consequence can be recovered from the premises if known. According to the model proposed by Daumé III and Campbell (2007), the likelihood of a single independent feature value  $f_1$  depends exclusively on its prior  $\pi_1$  (Equation 5). If an implication does not hold true (i.e. the special variable  $m = 0$ ) then also the likelihood of the dependent feature value  $f_2$  equals its prior  $\pi_1$  and  $\pi_2$ . However, if the implication is valid ( $m = 1$ ) the latter feature value is constrained, as shown in Equation 6.

$$p(f_1|\pi_1) = \pi_1^{f_1} (1 - \pi_1)^{1-f_1} \quad (5)$$

$$p(f_2|f_1, \pi_2, m) = \begin{cases} f_2 & f_1 = m = 1 \\ \pi_2^{f_2} (1 - \pi_2)^{1-f_2} & \text{otherwise} \end{cases} \quad (6)$$

This flat model is transformed into a hierarchical model in order to screen out the noise of similarities due to extra-linguistic factors. In particular, it is augmented with a hierarchy of  $m$  variables shaped by genealogical or areal classification. The prior of the root is a normal distribution  $\mathcal{N}(0, \sigma^2)$ , that of intermediate nodes  $\mathcal{N}(m_{par}, \sigma^2)$ , and that of the leaves the logistic binomial  $\mathcal{B}(\sigma(m_{par}))$ . Lu (2013) cast this problem as knowledge discovery, where language features are encoded in a Directed Acyclic Graph. The strength of implication universals can be learned as weights of its edges. However, the accuracy of this model lags behind that of Daumé III and Campbell (2007).

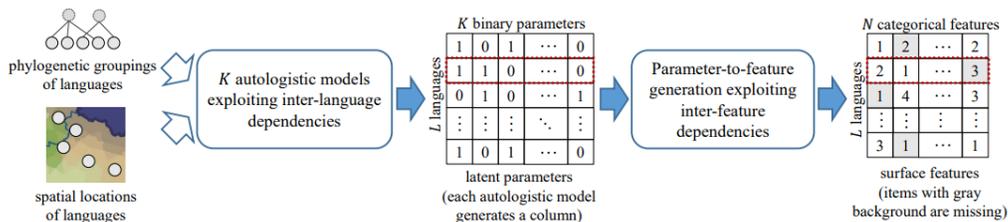


Figure 11: Pipeline of Murawaki (2017)’s Bayesian generative model.

**4.2.4 Multi-alignments.** Several approaches unravel typological features probing multi-parallel text. [Wälchli and Cysouw \(2012\)](#) represent each parallel sentence with a vector where each dimension is a language and its value is the lemma of the motion verb occurring therein. The similarity between each sentence encoding pair is estimated by Hamming distance, which corresponds to the agreement in lexicalization choices across languages. The dimensionality of the resulting similarity matrix can be reduced via Multi-Dimensional Scaling (MDS), which allows us to interpret the main dimensions of variation. In particular, each verb occurrence in a language is positioned in a continuum motivated by cross-lingually emergent categories. For instance, Figure 12 shows the first two dimensions of the MDS similarity matrix in Mapudungun. Note that dimensions are easily interpretable as e.g. the first accounts for deixis.

[Asgari and Schütze \(2017\)](#) initially search a language containing an unambiguous and overt marker for a specific typological feature (called head pivot) based on theoretical linguistic expertise. For instance, they opt for *ti* in Seychellois Creole (French Creole) as a head pivot for past tense marking. This is in turn projected to larger set of pivots through alignment-based  $\chi^2$  in a multi-parallel corpus. Finally, this set is aligned to n-grams in all the remaining languages. This allows to fetch markers of grammatical features across languages. Moreover, this procedure can reveal the similarity of languages in categorizing grammatical meanings. In particular, for each marker they calculate the (normalized) occurrence distribution over sentences. Then the similarity between language pairs is

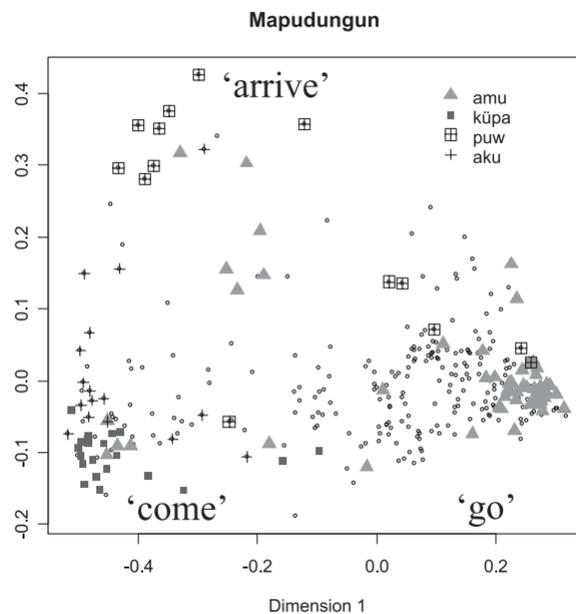


Figure 12: [Wälchli and Cysouw \(2012\)](#)'s cross-lingual sentence visualization for Mapudungun, where axes correspond to the first two dimensions of a MDS sentence similarity matrix. In the top-right corner is the legend of the motion verbs taken into consideration. Each data point is a sentence coded by the verb occurring therein, and positioned according to the cross-lingual matrix.



the Jensen-Shannon divergence between such distributions. This metric allows to cluster languages hierarchically, as shown in Figure 13.

Finally, typology can be derived from raw texts in an unsupervised fashion, without any need of multi-parallel texts. Roy et al. (2014) predict the position of adpositions based on heuristics. In particular, they assume that i) adpositions are frequent since they are function words; and ii) they have constrained selectional preferences for their complements. These assumptions can be quantified as follows: they draw up a list of the top words according to count- and entropy-based metrics of the left, right, and both contexts of frequent words. Then they estimate the rank correlation between lists derived from left, right and both contexts. If a language (e.g. Tamil, Dravidian) is post-positional, the right-both correlation is higher than left-both, and vice versa.

**4.2.5 Comparison of the strategies.** Establishing which of the above-mentioned strategies is optimal in terms of accuracy is not straightforward. In Figure 14, we collect the scores reported by several of the surveyed papers, provided that they concern specific features or the whole dataset (as opposed to subsets) and are numerical (as opposed to graphical plots). However, these results are not strictly comparable, since 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 relative difficulty of predicting each feature: for instance, the correct value of passive voice is less trivial to predict than word order according to Bender et al. (2013). Also, there appears to be no pre-eminent algorithm, as the properties of each are suited for some target features but detrimental for others. 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).

However, some papers carry out comparisons in the same experimental setting. According to Coke, King, and Radev (2016), the propagation from the genus majority value outperforms logistic regression both on linguistic features extracted from parallel texts and on other 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 target features: genealogy excels in word order features due to their diachronic stability. In turn majority propagation is surpassed by both supervised classification (Takamura, Nagata, and Kawasaki 2016) and ESL-based language similarity (Berzak, Reichart, and Katz 2014) based on evaluation on the entire WALS.

Another challenge in comparing different prediction mechanisms is that they target different features, and require different resources. The extraction of information from morphosyntactic annotation is more suited for word order features, whereas distributional metrics from multi-parallel texts are more informative of lexicalization patterns. On the other hand, propagation and supervised classification are general-purpose strategies. Moreover, the first two presuppose some annotated and/or parallel texts, whereas the second two need a pre-existing (although partial) database documentation. Different languages may lack one kind of resources or the other, limiting the mechanisms at their disposal.

In general, it should be stressed that many strategies have an evident weakness: they postulate incorrectly that language samples are independent and identically distributed (Lu 2013). This is not the case owing to family and area interactions. The solutions adopted to mitigate this bias vary: synthetic data can balance the distribution (Wang and Eisner 2017). Others include family and area features explicitly (Takamura, Nagata, and Kawasaki 2016; Malaviya, Neubig, and Littell 2017). Murawaki (2017) provide evidence with respect to the relative strength of interactions: inter-feature dependencies

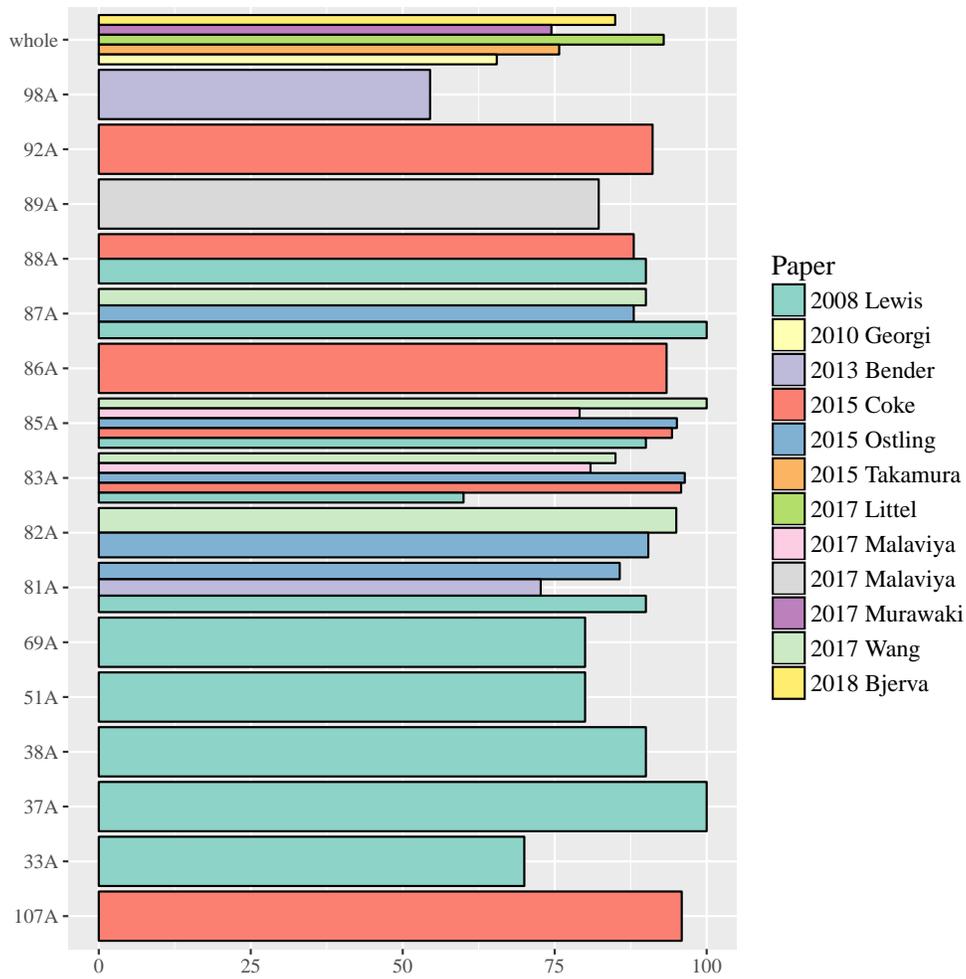


Figure 14: Accuracy of various algorithm on specific features or the whole WALS database.

are stronger typological indicators than inter-language dependencies (although they are complementary in nature), and horizontal diffusibility is more prominent than vertical stability.

Finally, a general trend emerges from the survey of automatic typology prediction. Apart from missing value completion, automatic prediction often accounts also for intra-language variation. 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 alignments from multi-parallel texts (Mayer and Cysouw 2012; Asgari and Schütze 2017) are promising insofar as they capture latent properties of languages in a bottom-up fashion, preserving their gradient nature.

	Author	Details	Number of Languages / Families	Task
Rules	Bender (2016)	Grammar generation	12 / 8	semantic parsing
	Schone and Jurafsky (2001)	Design of Bayesian network	1 / 1	word cluster labeling
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
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. (2018)	Construction-based tree preprocessing	6 / 3	machine translation, sentence similarity

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

## 5. Uses of Typological Information in NLP Models

The typological features developed in § 4 find many uses in NLP algorithms. Firstly, they can assist the design of algorithms, by being converted manually into rules, or priors / independence assumptions in Bayesian graphic model (§5.1). Secondly, they can be engineered to augment the input representations or tie together specific parameters across languages (§ 5.2). Finally, they can guide data selection and synthesis (§ 5.3). All these approaches are summarized in Table 3 and, as it will be demonstrated, consistently result in improvements in performance.

### 5.1 Rules and Priors

Typological features provide a watermark to the design of algorithms. In particular, they can be converted into instructions for rule-based algorithms (Bender 2016) or guide the choice of independence assumptions among the nodes in Bayesian networks (Schone and Jurafsky 2001).

Rule-based grammars can be generated from typological features through the Grammar Matrix kit, presented by Bender (2016). These grammars are couched within the framework of Minimal Recursion Semantics (Copestake et al. 2005) and can parse a string of a natural language into a semantic logical form, and vice versa. 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

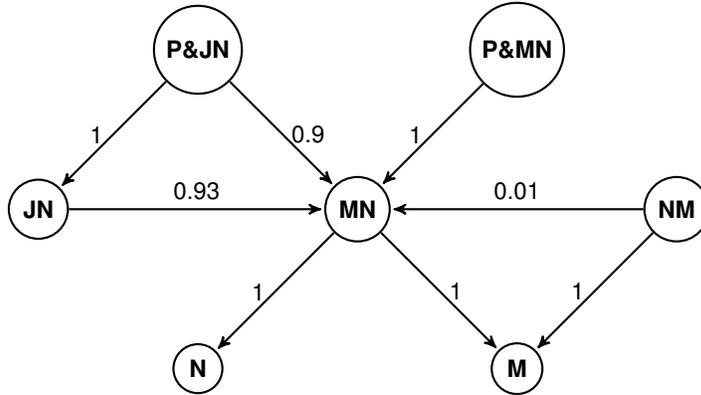


Figure 15: Subgraph of a Bayesian network for the ordering of numerals and nouns.

typology expects the specification of the kind, pattern, and position of grammatical marking, as well as the phrase types it covers: 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 made available through the Language CoLLAGE initiative (Bender 2014).

Moreover, typological features can guide the design of graphical models 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). The objective maximizes the probability of tag  $T_i$  and a feature set  $\Phi_i$  given the implicational universals  $U$  as  $\operatorname{argmax}_T P(\{\Phi_i, T_i\}_{i=1}^n | U)$ . By the chain rule it can be reformulated as Equation 7:

$$J = \operatorname{argmax}_T \prod_{i=1}^n P(T_i | \{\Phi_j, T_j\}_{j=1}^{i-1}, U) P(\Phi_i | T_i, \{\Phi_j, T_j\}_{j=1}^{i-1}, U) \quad (7)$$

Tags are processed by dependency order in the network, so  $\Phi_i$  can be removed from the first factor. In order to understand the effect of implicational universals, consider Figure 15: it shows the sub-graph for the ordering of numerals (M) and nouns (N), which is intertwined also with properties of adjectives (J) and adpositions (P).

## 5.2 Feature Engineering

The most widespread usage of typological features is tying specific parameters together and providing input representations of the current language properties in language transfer or multi-lingual joint learning (see § 3). The most prominent approach, introduced by Naseem, Barzilay, and Globerson (2012) and subsequently adopted by Täckström, McDonald, and Nivre (2013) and Zhang and Barzilay (2015), is called ‘selective sharing’. This approach aims at parsing sentences in a language transfer setting where there are multiple source languages and a single unobserved target language. It assumes that the parts of speech of head-dependent pairs are universal, but their ordering is language-specific. For instance, adjectives always depend on nouns syntactically, but with regard

to linear order in Igbo (Niger-Congo) they precede them, in Nihali (isolate) they follow them.

**5.2.1 Selective sharing.** This approach 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 dependents ( $\{D\}$ , characterized by their PoS tag) given a head  $h$ , with probability  $P(\{D\}|h)$ . The second step is language-specific: each dependent  $d$  is assigned a direction (left or right) with respect to  $h$  based on the language  $l$ , yielding  $\vec{d}$  with probability  $P(\vec{d}|d, h, l)$ . Dependents in the same direction are eventually ordered with a probability drawn from a uniform distribution of their possible unique permutations. The total probability is then defined as follows:

$$P(n|h, \theta_1) \cdot \sigma_n \left( \sum_{D_i \in D} P(D_i|h, \theta_2) \right) \cdot \prod_{d \in D} \sigma(\mathbf{w} \mathbf{g}(d, h, l, \mathbf{f}_l)) \cdot \frac{1}{\|D_R\| \|D_L\|} \quad (8)$$

In Equation 8, the first step is factorized in the estimation of the set size  $n$ , parametrized by  $\theta_1$ , and the actual selection of dependents, whose softmax function  $\sigma$  normalizes over different  $n$  values, is parametrized by  $\theta_2$ . The second step, overseeing the direction assignment, is parametrized by  $\mathbf{w}$  and hinges upon a function for feature extraction  $\mathbf{g}()$ , whose arguments include a typology feature vector  $\mathbf{f}_l$ . These features can be encoded explicitly, extracting them directly from WALS, or implicitly, treating them as latent features. The values of all the parameters are estimated by maximizing the likelihood of the observations.

Täckström, McDonald, and Nivre (2013) recast this algorithm into a discriminative model, in order to amend the alleged limitations of the generative one. In fact, this allows to dispose of strong independence assumptions (e.g. between choice and ordering of dependents) and invalid feature combinations. Their algorithm 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 about selectional preferences and dependency length. Moreover, they introduce (language-specific) features for the directionality of dependents. These are combinations of the PoS tags of the head and dependents with WALS values. For instance, ‘Order of subject, verb, and object’ (81A) is taken into account only when the head is a verb and the dependent is a noun.

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 atomic features and languages by factorizing a tensor constructed from their combination. The prior knowledge from the typological database enables the model to forbid 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 by being summed with the typological features  $T_u \phi_{t_u}$ . Moreover, the model 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 dependents contexts  $M_c\phi_{m_c}$ . The overall score for a labeled dependency is shown in Equation 9:

$$S(h \xrightarrow{l} m) = \sum_{i=1}^r [H_c\phi_{h_c}]_i [M_c\phi_{m_c}]_i \odot \{ [T_l\phi_{t_l}]_i + [L\phi_l]_i \odot ([T_u\phi_{t_u}]_i + [H\phi_h]_i [M\phi_m]_i [D\phi_d]_i) \} \quad (9)$$

The total loss function is the weighted sum of Equation 9 and the score of a flat model consisting in the tensor product of all the feature-parameter pairs introduced (excluding typology). The loss is optimized within a maximum soft-margin objective through on-line passive-aggressive updates.

All the presented approaches to selective sharing are robust to cases where the target typological features do not match any of the source language, which may lead learning astray. The strategies include unsupervised learning, ambiguous learning, and semi-supervised learning.

Naseem, Barzilay, and Globerson (2012) adapt the model in an unsupervised fashion through Expectation Maximization (Dempster, Laird, and Rubin 1977), marginalizing the likelihood over the latent parameters involved in the derivation of a target tree matching the observed PoS tag sequence. Täckström, McDonald, and Nivre (2013) tackle the same problem from the side of ambiguous learning, which consists in training a discriminative model on the target language from sets of automatically predicted ambiguous labels  $\hat{y}$ . This solution comes in two flavors: the ambiguous labels may derive from the (top-most likely) predictions of the source parser (self-learning) or their union with the predictions of other parsers (ensemble-learning). Finally, Zhang and Barzilay (2015) demonstrate the perks of semi-supervised learning. Even with a handful of annotated examples from the target language, it is possible to integrate the multi-lingual source model successfully.

To sum up, typological features are integrated differently within the framework of ‘selective sharing’. They can 1) be fed to a parametrized feature extractor in a sub-module for the ordering of dependents in a generative model; 2) be combined with PoS features for a discriminative graph-based model; 3) condition arcs and labels in a discriminative tensor-based model in order to avoid invalid parameter combinations.

**5.2.2 Multi-lingual Biasing.** Some papers leverage typology to gear a multilingual model toward the properties of a specific language. A basic approach is providing a vector of typological features for such language in input (Daiber, Stanojević, and Sima’an 2016; Berzak, Reichart, and Katz 2015). More sophisticated approaches also condition the hidden layers of a transition-based parser state (Ammar et al. 2016) or a global sequence representation in a language model (Tsvetkov et al. 2016).

Daiber, Stanojević, and Sima’an (2016) develop a one-to-many reordering algorithm which benefits downstream monotone translation (without reordering in the decoder). A feed-forward neural network is trained on a multi-parallel corpus to estimate the permutation probabilities of source word pairs. This network receives as input lexical and morphosyntactic features of the source word pairs and typological features of the target language. The best sequence of permutations is inferred via k-best graph search in a finite state automaton, producing a lattice.

The joint multilingual parser developed by Ammar et al. (2016) intertwines both language-specific and language-invariant features in its copious feature set. In particular, the former group includes universal coarse PoS tags, multi-lingual word embeddings (obtained through robust projection as detailed by Guo et al. (2016b)), and multilingual word clusters (Täckström, McDonald, and Uszkoreit 2012). The latter group consists of fine-grained PoS tags. Overall, the model maximizes the log-likelihood of:

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

This transition-based parser selects a next move from a pool of possible actions given its state  $\mathbf{p}_t$  at current time  $t$ . This in turn is defined as a set of iteratively manipulated, densely represented data structures, namely a buffer  $\mathbf{b}_t$ , a stack  $\mathbf{s}_t$ , and an action history  $\mathbf{a}_t$ . These modules are the output of stack-LSTMs represented with input feature representations (stack and buffer) and action representations (history). The entire parser is biased toward a particular language through language embeddings  $\mathbf{l}_{it}$ . These embeddings consist of (a non-linear transformation of) either a mere one-hot identity vector or a vector of typological properties. In particular, they are added to both input feature and action vectors (to affect the three above-mentioned modules) and concatenated to the modules themselves (to affect the entire parser state). Finally, the state is propagated through an action-specific layer parametrized by  $\mathbf{g}_t$  and  $q_t$ .

Similarly, also hidden states of language models can be conditioned on typological features. Character-level language models can be trained jointly for several languages, provided that these are encoded by universal symbols like IPA phonemes (Tsvetkov et al. 2016). An input  $\mathbf{x}$  and a language vector  $\ell$  at time  $t$  are initially mapped to a local context representation and then passed to a global LSTM. This hidden representation is factored by a non-linear transformation of typological features  $t_\ell$ , vectorized, and prompted to the output symbol  $\phi_t$ :

$$\mathbf{G}_t^\ell = \text{LSTM}(W_{c_x} x_t + W_{c_\ell} x_\ell + b, \mathbf{g}_{t-1}) \otimes \tanh(W_\ell t_\ell + b_\ell)^\top \quad (10)$$

$$P(\phi_t | \phi_{<t}, \ell) = \sigma(\mathbf{W} \text{vec}(\mathbf{G}_t^\ell) + \mathbf{b}) \quad (11)$$

Equation 10 is visualized in Figure 16: note how typological features interact with the global hidden representation of the sequence. The phoneme vectors learnt end-to-end by the language model are evaluated on two downstream applications. In particular, on the one hand the pair-wise cosine distance among such vectors weight the transitions between the corresponding phonemes in Finite State Transducer cascades for lexical borrowing identification. On the other hand, such vectors replace manual phonetic features in classification and regression trees for speech synthesis.

### 5.3 Data Selection and Synthesis

Another usage of typological features is data selection. This procedure is crucial for effective language transfer, as they superintend: i) to choosing the most suitable source language/examples; or ii) to weighting the contribution of each language/example in multilingual joint models. The selection is usually carried out through general language similarity metrics, or through measures of overlap in language-independent properties (such as PoS sequences).

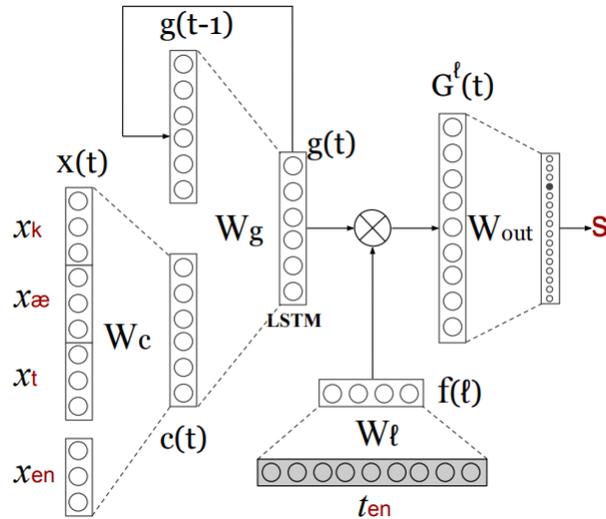


Figure 16: Architecture of Tsvetkov et al. (2016)'s phoneme-based polyglot language model.

Most of the language transfer experiments resort to typological (and/or genealogical) information behind the scenes. Indeed, they choose source and target languages based on their similarity: for instance, Czech and Russian (Hana, Feldman, and Brew 2004), Danish and Swedish, Hindi-Urdu (Zeman and Resnik 2008), Dutch and German (Spreyer and Kuhn 2009), Indo-European languages (McDonald, Petrov, and Hall 2011). However, they tend not to explain how languages are determined to be ‘closely related’.

Language distance metrics quantify these similarities explicitly. Deri and Knight (2016) extract them from URIEL, conflating information about genealogical, geographic, syntactic, and phonetic properties. Afterwards, they select the closest source with identical script for transferring grapheme-to-phoneme models. The source models are trained on Wikipedia data, either IPA help tables or Wiktionary. They are further adapted by mapping the phoneme inventory of the source language into that of the target language (before or after transfer), by establishing distances among phonemes based on Hamming distances among their feature vectors in Phoible.

Metrics for source selection can also be extracted in a data-driven fashion, without explicit reference to structured taxonomies. For example, Rosa and Zabokrtsky (2015) estimate the Kullback-Leibler divergence between part-of-speech trigram distributions. In order to approximate the divergence in syntactic structures, Ponti et al. (2018) employ the Jaccard distance of 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) character-based identity prediction of the target language (excluding its true identity); 3) Hamming distance from the target language typological vector. In fact, their perks are complementary: language identity (and consequently typology) are bound to character similarity, but generalize well. On the other hand, PoS-based metrics are universal, but deteriorate easily.

Source selection is a special case of source weighting where weights are integers. However, weights can be gradient and consist of real numbers, as proposed by [Søgaard and Wulff \(2012\)](#). In particular, they adapt delexicalized parsers by weighting every instance based on the inverse of the Hamming distance between typological (or genealogical) features in source and target languages. A bottom-up approach instead is developed by [Søgaard \(2011\)](#), who weights sentences in a source language based on the perplexity of their coarse PoS tags according to a sequential model trained on the target language.

Finally, the lack of target annotated data can be alleviated by boosting the variety and width of the source data by synthesizing new examples: for instance, the Galactic Dependencies Treebanks stem from real trees whose nodes have been permuted according to the word order rules for nouns and verbs in other languages ([Wang and Eisner 2016](#)). In particular, the probability of a permutation  $\pi$  for nodes  $i$  and  $j$  within a set of a head and its dependents is defined by a parametrized model:

$$P_{\theta}(\pi|x) = \sigma \sum_{1 \leq i < j \leq n} \theta \cdot \mathbf{f}(\pi, i, j)$$

The features  $\mathbf{f}$  taken into account include PoS tags and dependency relations of single nodes, siblings, and n-grams. Enlarging the pool of treebanks with synthetic data improves the performance of model transfer for parsing when the source is chosen in a supervised fashion (performance on target development data) and in an unsupervised fashion (coverage of target PoS sequences). However, adding new *real* languages to the pool is even more beneficial. This is possibly due to fact that the net contribution to diversity of these synthetic datasets is limited to word order.

Rather than generating new synthetic data, [Ponti et al. \(2018\)](#) 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 fact, syntactic structures of translationally equivalent sentences are not isomorphic. For instance, relative clauses in Arabic (Afro-Asiatic), with an indefinite antecedent, drop the relative pronoun, which is mandatory in Portuguese (Indo-European). The preprocessing method is rule-based: when it finds a source subtree matching a construction documented in a typological database, it converts it to the target strategy. The conversion hinges upon a sequence of node addition, node deletion, and label change.

Subsequently, [Ponti et al. \(2018\)](#) feed preprocessed syntactic representations to syntax-based neural models, achieving state-of-art results in several tasks. Firstly, they perform neural machine translation with an attentional encoder-decoder network that jointly learns to translate and align words ([Bahdanau, Cho, and Bengio 2015](#)) and filters linguistic features (including syntax) in input through a Convolutional Neural Network ([Sennrich and Haddow 2016](#)). Secondly, they encode dependency tree pairs with a TreeLSTM architecture ([Tai, Socher, and Manning 2015](#)) for cross-lingual sentence similarity classification. The model is lexicalized with multilingual word embeddings obtained with the iterative Procrustes method ([Artetxe, Labaka, and Agirre 2017](#)).

#### 5.4 Comparison

In order to compare the methods surveyed in this section, [Figure 17](#) provides the scores of each model in three main settings (each with identical architecture and hyper-parameters): with gold database features (Typology), latently inferred features (Data-

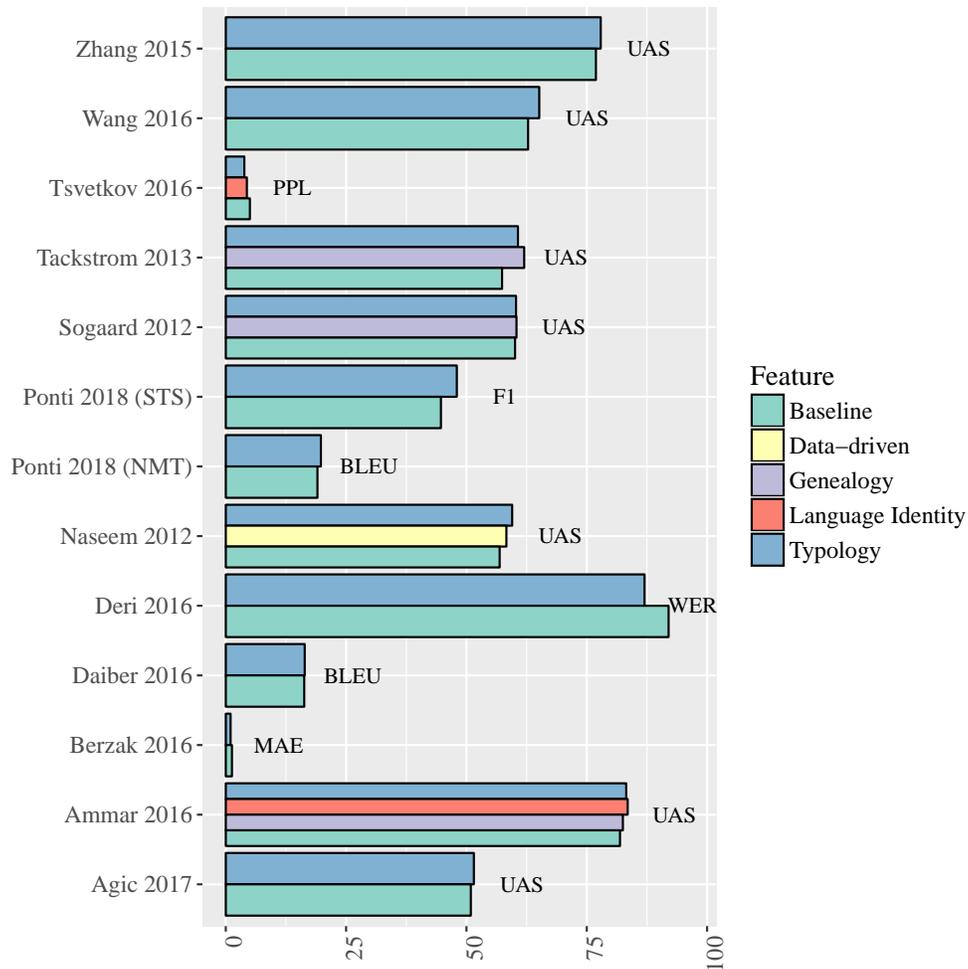


Figure 17: Performance of the surveyed algorithms for the tasks detailed in Table 3. They are evaluated with different feature sets: no features (Baseline), latently inferred typology (Data-driven), Genealogy, Language Identity, and gold database features (Typology). Evaluation metrics are reported right of the bars: Unlabeled Attachment Score (UAS), F1 Score, BiLingual Evaluation Understudy (BLEU), Word Error Rate (WER), and Mean Absolute Error (MAE).

driven), or without both (Baseline). Firstly, it is evident how typology consistently (although often moderately) ameliorates baseline performances across several NLP tasks. In particular, note that the scores are higher for for metrics that increase with better predictions (Unlabeled Attachment Score, F1 Score, BLEU) and lower for metrics that decrease (Word Error Rate, Mean Average Error, Perplexity). Secondly, gold database features appear to be more reliable than latently inferred features for typology (Naseem, Barzilay, and Globerson 2012).

In addition to performance improvements, typology-savvy methods tend to be more robust to a variety of scenarios of data paucity (Ammar et al. 2016), to better fit to languages outside the Indo-European family (Naseem, Barzilay, and Globerson 2012), and to be more tractable in terms of parameter numbers (Tsvetkov et al. 2016). It is then reasonable to conclude that typology is a valuable component for multilingual NLP algorithms.

Moreover, some of the experiments surveyed here compare typology with other language properties, by substituting typological features with features related to Genealogy and Language Identity. Based on Figure 17, it is unclear whether typology should be preferred, as it is sometimes rivaled in performance by the other properties. In particular, it is typology to excel according to Tsvetkov et al. (2016), but genealogy according to Søgaard and Wulff (2012); Täckström, McDonald, and Nivre (2013) and language identity according to Ammar et al. (2016). However, each of the experiments involves different settings: caution is required to justify generalizations. Features may fall short just because they have been selected poorly or are irrelevant for the task at hand, which makes the comparison invalid. In fact, through an in-depth discussion, we advocate a preference for typology.

The unexpected peak in performance of naive language identity features in Ammar et al. (2016) is probably due to a debatable selection of typological features for the two alternative experimental settings, given that the languages in their sample are mostly from the same family. In the first, the paucity and low diversification of features (5, all related to word order) fail to discriminate languages from one another. In the second, noise is pervasive because features are averaged by genus. Indeed, Tsvetkov et al. (2016) demonstrates that typology yields large gains with rich feature sets and representative language samples.

As for other experiments comparing typology with genealogy, the results are controversial because the performance of the compared models is almost equivalent. This is partly due to the high correlation between the two kinds of properties, and partly due to the design of the experiments. For instance, Täckström, McDonald, and Nivre (2013) define typological similarity as sharing *all* the features, whereas the definition of genealogy is broader: it suffices to belong to the same family. As a consequence, the latter receives supervision from more languages than the former.

A final question concerns which of the surveyed methods incorporates typological features in NLP algorithms more effectively. As for “selective sharing”, 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 regard to biasing multilingual models, there is a clear tendency toward letting typological features interact not just with the input representation, but also with deeper levels of abstraction such as hidden layers and global sequence representations.

Overall, the approaches surveyed in this section support the claim that typology can help design the architecture of algorithms, engineer their features, and select / pre-process their data. Nonetheless, many challenges lay ahead for each of these purposes. We discuss them in the next section.

## 6. Future Research Avenues

In § 5, we surveyed the current uses of typological information in NLP. Here we speculate about several future research avenues towards closer integration of linguistic typology and multilingual NLP. In particular, we discuss: i) the extension of existing

methods to new tasks, possibly exploiting typological resources that have been neglected thus far (§ 6.1); ii) new methods injecting typological information into NLP models as soft constraints or auxiliary objectives (§ 6.2); iii) new ways to acquire and represent typological information to reflect the gradient and contextual nature of cross-lingual variation and machine learning models (§ 6.3).

### 6.1 Extending the Usage to New Tasks and Features

The trends observed in § 5 reveal that experiments involving typology are mostly focused on morphosyntactic tasks, in particular 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. 2018). As a consequence, the set of typological features selected or acquired automatically is impoverished and is mostly limited to a handful of word-order features from a single database, *WALS* (see § 4). Nonetheless, the array of tasks that pertain to polyglot NLP is varied (see § 3), and other typological datasets thus far neglected (see § 2.3) 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 readily sourced from the Valency Patterns Leipzig database (*ValPaL*).

Lexical semantics could assist several tasks, by providing tables of translationally equivalent words across languages. These tables are provided by databases such as the *World Loanword Database (WOLD)*, the *Intercontinental Dictionary Series (IDS)*, or 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 the generation of multilingual word embeddings and enabling 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). 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 help interpreting the intrinsic difficulty of texts for language modeling or neural machine translation, both supervised (Johnson et al. 2016) and unsupervised (Artetxe et al. 2017). 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. 2018). This variation has to be taken into account in both language transfer and multilingual joint learning.

Nonetheless, 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, all 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, or other databases such as SSWL, APiCS, or AUTOTYP. On the other hand, experiments on phonological tasks (Tsvetkov et al. 2016; Deri and Knight 2016) could extract features also from LAPSyD and StressTyp2.

## 6.2 Injecting Typology in Machine Learning Algorithms

In § 5, we surveyed how typological information can provide guidance to NLP algorithms, including 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 typological resources and NLP algorithms.

Encoding cross-language variations 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 towards the pre-defined knowledge. In practice, learning algorithms (e.g. structured perceptron (Collins 2002), MIRA (Crammer and Singer 2003) and structured SVM (Taskar, Guestrin, and Koller 2004)) iterate between an inference step and a step of parameter update with respect to a gold standard. The inference step is the natural place of encoding external knowledge 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-language variation, especially 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. We next survey a number of existing approaches that can efficiently encode such constraints.

The goal of an inference algorithm is to predict the best output label according to the current state of the model parameters.<sup>3</sup> For this purpose, the algorithm searches the space of possible output labels 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 paper we do not answer this question directly but rather survey a number of approaches that succeeded in dealing with it.

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 employ different types of knowledge encoding, e.g. PR uses expectation constraints on the posterior parameter distribution,

---

<sup>3</sup> Generally speaking, an inference algorithm can make other predictions such as computing expectations and marginal probabilities. As in the context of this paper we are mostly focused on the prediction of the best output label, we refer only to this type of inference problems.

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 (ILP) constraints while 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 tendential and hence a good fit for soft constraints. Moreover, Bayesian modeling allows to set prior probability distributions according to the tendential relationships of typological features (Schone and Jurafsky 2001). Finally, DD is potentially useful to learn to perform multiple tasks jointly, where one is the actual NLP application and another is the data-driven prediction of typological features.

This same ideas could be exploited in deep learning algorithms. We have seen in § 3.3 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, similarly to what Chen et al. (2017) implemented by predicting language identity.

In the domain of multilingual representation learning (§ 3.4) a number of works (Faruqui et al. 2015; Rothe and Schütze 2015; Osborne, Narayan, and Cohen 2016; Mrkšić et al. 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) allows to push together and pull apart vector pairs according to relational constraints, while preserving the relationship between words in the original space and possibly propagating the specialization to unseen words or transferring it to other languages (Vulić et al. 2018). 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: Gradiance and Context-Sensitivity

As shown in § 4.1, most of the typology-savvy algorithms thus far exploited features extracted from manually-crafted databases. However, this approach is riddled by several shortcomings, which are reflected in the small margins of improvement in performance observed in § 5.4. Luckily, the shortcomings can be averted through some methods outlined in § 4.2 that allow typological information to emerge from the data in a bottom-up fashion, rather than being predetermined.

Firstly, typological databases provide *incomplete* documentation of the cross-lingual variation, in terms of features and languages. As raw textual data are more easily accessible and cost-effective, they are a valid alternative. Secondly, the 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 frequency*, hence they risk to hinder models from learning less likely but plausible patterns (Sproat 2016). Inferring typological information from texts would enable the system to discover patterns within individual examples, including both the frequent and the infrequent ones.

Thirdly, typological features inside datasets are *discrete*, i.e. 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 § 2.2), hence they are better captured by continuous features. In addition to being psychologically motivated, this sort of representations is also more compatible with machine learning algorithms as these work on real-valued multi-dimensional word embeddings and hidden states.

To sum up, the automatic development of typological information and its possible integration into end-to-end representational algorithms has the potential to solve an important bottleneck in polyglot NLP. Rather than transforming the contextual and gradient typological information implicitly present in texts into incomplete, approximate, and discrete features stored into databases, and subsequently feeding such features to continuous, probabilistic, and contextual models; the algorithms can skip the intermediate step and model cross-lingual variation directly from textual data.

Several techniques surveyed in § 4.2 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*) inform 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 ways in which feature engineering has been carried out thus far (§ 5.2). 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 (real) value its relative frequency within a language.

In 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 allows to enhance data selection, similarly to what Søgaard (2011) achieved using PoS n-grams as similarity metric. The selection of similar sentences rather than similar languages as source in language transfer is likely to yield large margins of improvement, as demonstrated by Agić (2017) for parsing in an oracle setting.

## 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 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, and which of its features are relevant? ii) which methods allow us to inject typological information from external resources, and how should such information be encoded? iii) can we interpret the typological information implicitly captured by distributed word representations and neural hidden states and exploit it? We summarize our key findings below:

1. Typological information is currently used predominantly for morpho-syntactic tasks, in particular dependency parsing. As a consequence, these approaches typically select a limited subset of features from a single dataset (WALS) and focus on a single aspect of variation (word order). However, typological databases 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. However, 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 towards 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 sometimes perform equally well because of 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 be also 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 allow to 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, rather than sourcing typological features from databases.

In general, we demonstrated that typology is relevant to a wide range of NLP tasks and provides the most 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.

## Acknowledgments

This work is supported by ERC Consolidator Grant LEXICAL (no 648909).

## 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 (ACL - IJCNLP 2015)*, 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*, 4:301.
- Agić, Željko, Jörg Tiedemann, Kaja Dobrovoljc, Simon Krek, Danijela Merkle, 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*.
- Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 451–462.

- Artetxe, Mikel, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.
- Asgari, Ehsaneddin and Hinrich Schütze. 2017. Past, present, future: A computational investigation of the typology of tense in 1000 languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 113–124.
- Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Bakker, Dik. 2010. Language sampling. In JJ 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, Association for Computational Linguistics.
- 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 *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 *LaTeCH 2013*.
- 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 *CoNLL*, pages 21–29.
- Berzak, Yevgeni, Roi Reichart, and Boris Katz. 2015. Contrastive analysis with predictive power: Typology driven estimation of grammatical error distributions in ESL. In *CoNLL*, pages 94–102.
- Bickel, Balthasar. 2007a. Typology in the 21st century: Major current developments. *Linguistic Typology*, 11(1):239–251.
- Bickel, Balthasar. 2007b. 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. *Oxford handbook of linguistic analysis*, pages 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.
- Bjerva, Johannes and Isabelle Augenstein. 2018. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. *arXiv preprint arXiv:1802.09375*.
- Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the ACL*, 5:135–146.
- Botha, Jan A. and Phil Blunsom. 2014. Compositional morphology for word representations and language modelling. In *ICML*, pages 1899–1907.
- Bowerman, Melissa and Soonja Choi. 2001. Shaping meanings for language: universal and language-specific in the acquisition of semantic categories. In *Language acquisition and conceptual development*. Cambridge University Press, pages 475–511.
- Braud, Chloé, Ophélie Lacroix, and Anders Søgaard. 2017. Cross-lingual and cross-domain discourse segmentation of entire documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 237–243.
- 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 JA 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 *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 *ACL*, pages 280–287.
- Chen, Xilun, Ben Athiwaratkun, Yu Sun, Kilian Weinberger, and Claire Cardie. 2017. Adversarial deep averaging networks for cross-lingual sentiment classification. *arXiv preprint arXiv:1606.01614*.
- Cohen, Shay and Noah A. Smith. 2009. Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction. In *NAACL-HLT*, pages 74–82.
- 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 *EMNLP*, pages 1–8.
- 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*.
- 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 JJ 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.
- Cotterell, Ryan and Hinrich Schütze. 2015. Morphological word-embeddings. In *NAACL-HLT*, pages 1287–1292.
- 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. 2002. *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.
- 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. 2009. Non-parametric bayesian areal linguistics. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics*, pages 593–601, Association for Computational Linguistics.
- Daumé III, Hal and Lyle Campbell. 2007. A Bayesian model for discovering typological implications. In *ACL*, pages 65–72.
- Dediu, Dan and Michael Cysouw. 2013. Some structural aspects of language are more stable than others: A comparison of seven methods. *PloS one*, 8(1):e55009.
- Dediu, Dan and Stephen C Levinson. 2012. Abstract profiles of structural stability point to universal tendencies, family-specific factors, and ancient connections between languages. *PLoS one*, 7(9):e45198.
- 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*), pages 1–38.
- Deri, Aliya and Kevin Knight. 2016. Grapheme-to-phoneme models for (almost) any language. In *ACL*, pages 399–408.
- Diab, Mona Talat. 2003. *Word sense disambiguation within a multilingual framework*. Ph.D. thesis.
- Dixon, Robert MW. 1994. *Ergativity*. Cambridge University Press.
- Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 13(2):257–292.
- Dryer, Matthew S. and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology.
- Dunn, Michael, Simon J Greenhill, Stephen C Levinson, and Russell D Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345):79.
- 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. *arXiv preprint arXiv:1606.09403*.
- 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. Semantic typology. In *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 *NAACL-HLT*, pages 1606–1615.
- 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 and Dipanjan Das. 2013. Cross-lingual discriminative learning of sequence models with posterior regularization. pages 1996–2006.
- Ganchev, Kuzman, Jennifer Gillenwater, and Ben Taskar. 2009. Dependency grammar induction via bitext projection constraints. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, volume 1, pages 369–377.
- Ganchev, Kuzman, Jennifer Gillenwater, Ben Taskar, et al. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049.
- Garg, Nikhil and James Henderson. 2016. A bayesian model of multilingual unsupervised semantic role induction. *arXiv preprint arXiv:1603.01514*.
- Georgi, Ryan, Fei Xia, and William Lewis. 2010. Comparing language similarity across genetic and typologically-based groupings. In *COLING*, pages 385–393.
- Gerz, Daniela, Edoardo Maria Ponti, Jason Naradowsky, Roi Reichart, Anna Korhonen, and Ivan Vulić. 2018. Language modeling for morphologically rich languages: Character-aware modeling for word-level prediction. *Transactions of the Association of Computational Linguistics*.
- Gillick, Dan, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. 2016. Multilingual language processing from bytes. In *Proceedings of NAACL-HLT*, pages 1296–1306.
- Globerson, Amir and Tommi S Jaakkola. 2007. Fixing max-product: Convergent message passing algorithms for MAP LP-relaxations. In *NIPS*, pages 553–560.
- 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 *NAACL-HLT*, pages 1386–1390.
- Grave, Edouard and Noémie Elhadad. 2015. A convex and feature-rich discriminative approach to dependency grammar induction. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1375–1384.
- 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. 1966. *Universals of language*. MIT Press.
- Guo, Jiang, Wanxiang Che, Haifeng Wang, and Ting Liu. 2016a. Exploiting multi-typed treebanks for parsing with deep multi-task learning. *arXiv preprint arXiv:1606.01161*.
- 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*, volume 1, pages 1234–1244.
- Guo, Jiang, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2016b. A representation learning framework for multi-source transfer parsing. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2734–2740.
- Ha, Thanh-Le, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *arXiv preprint arXiv:1611.04798*.
- Hammarström, Harald, Robert Forkel, Martin Haspelmath, and Sebastian Bank, editors. 2016. *Glottolog 2.7*. Max Planck Institute for the Science of Human History, Jena.
- Hana, Jiri, Anna Feldman, and Chris Brew. 2004. A resource-light approach to Russian morphology: Tagging Russian using Czech resources. In *EMNLP*, pages 222–229.
- 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. 2013. Multilingual distributed representations without word alignment. *arXiv preprint arXiv:1312.6173*.
- 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.
- Ide, Nancy, Tomaz Erjavec, and Dan Tufis. 2002. Sense discrimination with parallel corpora. In *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions-Volume 8*, pages 61–66, Association for Computational Linguistics.
- Jensen, Finn V. 1996. *An introduction to Bayesian networks*, volume 210. UCL 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. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.
- 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 *ACL*, pages 561–569.
- Kim, Sungchul, Kristina Toutanova, and Hwanjo Yu. 2012. Multilingual named entity recognition using parallel data and metadata from Wikipedia. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 694–702.
- Klementiev, Alexandre, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *COLING*, pages 1459–1474.

- 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. 2014. Learning multilingual word representations using a bag-of-words autoencoder. *arXiv preprint arXiv:1401.1803*.
- 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.
- Levy, Omer and Yoav Goldberg. 2014. Dependency-based word embeddings. In *ACL*, pages 302–308.
- Lewis, M Paul, Gary F Simons, and Charles D Fennig. 2016. *Ethnologue: Languages of the world*, 19th edition. SIL international.
- Lewis, William D and Fei Xia. 2008. Automatically identifying computationally relevant typological features. In *IJCNLP*, pages 685–690.
- Li, Shen, Joao V Graça, and Ben Taskar. 2012. Wiki-ly supervised part-of-speech tagging. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1389–1398.
- Littel, Patrick, David R. Mortensen, and Lori Levin. 2016. URIEL Typological database. Pittsburgh: CMU.
- 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.
- Ma, Xuezhe and Fei Xia. 2014. Unsupervised dependency parsing with transferring distribution via parallel guidance and entropy regularization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1337–1348.
- Maddieson, Ian, Sébastien Flavier, Egidio Marsico, Christophe Coupé, and François Pellegrino. 2013. LAPSyd: Lyon-Albuquerque phonological systems database. In *INTERSPEECH*, pages 3022–3026.
- 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 *ACL*, pages 870–878.
- Mayer, Thomas and Michael Cysouw. 2012. Language comparison through sparse multilingual word alignment. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 54–62, Association for Computational Linguistics.
- McDonald, Ryan, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 91–98, Association for Computational Linguistics.
- McDonald, Ryan, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *EMNLP*, pages 62–72.
- 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*.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

- 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 of 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 *NAACL-HLT*, pages 142–148.
- 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 *ACL*, pages 629–637.
- Naseem, Tahira, Benjamin B. Bonet, Jacob Eisenstein, and Regina Barzilay. 2009. Multilingual part-of-speech tagging: Two unsupervised approaches. *Journal of Artificial Intelligence Research*, 36:341–385.
- Naseem, Tahira, Harr Chen, Regina Barzilay, and Mark Johnson. 2010. Using universal linguistic knowledge to guide grammar induction. In *Proc. of EMNLP 2010*.
- Navigli, Roberto. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10.
- Nichols, Johanna. 1992. *Language diversity in space and time*.
- Niehuus, 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 *LREC*, pages 1659–1666.
- O’Horan, Helen, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. *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 ACL (to appear)*.
- Östling, Robert. 2015. Word order typology through multilingual word alignment. In *ACL*, pages 205–211.
- Östling, Robert and Jörg Tiedemann. 2016. Continuous multilinguality with language vectors. *arXiv preprint arXiv:1612.07486*.
- Padó, Sebastian and Mirella Lapata. 2005. Cross-linguistic projection of role-semantic information. In *EMNLP*, pages 859–866.
- 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)*, EPFL-CONF-231134.
- Pawley, Andrew. 1993. A language which defies description by ordinary means. In Andrew Pawley, editor, *The role of theory in language description*. pages 87–130.
- Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL-HLT*, pages 2227–2237.
- Plank, Frans and Elena Filiminova. 1996. Universals archive. <http://www.ling.unikonstanz.de/pages/proj/sprachbau.htm>. Universität Konstanz.
- Van der Plas, Lonneke, Paola Merlo, and James Henderson. 2011. Scaling up automatic cross-lingual semantic role annotation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 299–304.
- Ponti, Edoardo Maria, Roi Reichart, Anna Korhonen, and Ivan Vulić. 2018. Isomorphic transfer of syntactic structures for cross-lingual nlp. *Draft*.

- 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.
- Prettenhofer, Peter and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1118–1127, Association for Computational Linguistics.
- Rasooli, Mohammad Sadegh and Michael Collins. 2015. Density-driven cross-lingual transfer of dependency parsers. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 328–338.
- Rosa, Rudolf and Zdenek Zabokrtsky. 2015. KLcpos3 - a language similarity measure for delexicalized parser transfer. In *ACL*, pages 243–249.
- Ross, Malcolm. 1997. Social networks and kinds of speech community event. In Roger M. Blench and Matthew Spriggs, editors, *Archaeology and Language, I*. Routledge, pages 209–261.
- Rothe, Sascha and Hinrich Schütze. 2015. AutoExtend: Extending word embeddings to embeddings for synsets and lexemes. In *ACL*, pages 1793–1803.
- Rotman, Guy, Ivan Vulić, and Roi Reichart. 2018. Bridging languages through images with deep partial canonical correlation analysis. In *Proceedings of ACL 2018*.
- 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*.
- Sapir, Edward. 2014 [1921]. *Language*. Cambridge University Press.
- von Schlegel, Friedrich. 1808. *Über die Sprache und Weisheit der Indier*. Mohr und Zimmer.
- 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.
- Smith, David A and Jason Eisner. 2009. Parser adaptation and projection with quasi-synchronous grammar features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, volume 2, pages 822–831.
- Snyder, Ben. 2010. *Unsupervised Multilingual Learning*. PhD thesis. MIT.
- Snyder, Benjamin and Regina Barzilay. 2008. Unsupervised multilingual learning for morphological segmentation. In *Proceedings of ACL-08: HLT*, pages 737–745.
- Snyder, Benjamin, Tahira Naseem, Jacob Eisenstein, and Regina Barzilay. 2009. Adding more languages improves unsupervised multilingual part-of-speech tagging: A Bayesian non-parametric approach. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 83–91.
- Søgaard, Anders. 2011. Data point selection for cross-language adaptation of dependency parsers. In *ACL*, pages 682–686.
- Søgaard, Anders, Željko Agić, Héctor Martínez Alonso, Barbara Plank, Bernd Bohnet, and Anders Johannsen. 2015. Inverted indexing for cross-lingual nlp. 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 (ACL-IJCNLP 2015)*, page 1713–1722.
- Søgaard, Anders and Julie Wulff. 2012. An empirical study of non-lexical extensions to delexicalized transfer. *Proceedings of COLING 2012: Posters*, pages 1181–1190.
- Spitkovsky, Valentin I., Hiyan Alshawi, and Daniel Jurafsky. 2011. Lateen EM: Unsupervised training with multiple objectives, applied to dependency grammar induction. In *EMNLP*, pages 1269–1280.
- Spreyer, Kathrin and Jonas Kuhn. 2009. Data-driven dependency parsing of new languages using incomplete and noisy training data. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 12–20.
- Sproat, Richard. 2016. Language typology in speech and language technology. *Linguistic Typology*, 20(3):635–644.

- Täckström, Oscar, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.
- Täckström, Oscar, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers. In *NAACL-HLT*, pages 1061–1071.
- Täckström, Oscar, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *NAACL-HLT*, pages 477–487.
- 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 *LREC*, pages 69–76.
- 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 *NIPS*, pages 25–32.
- 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. *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, Association for Computational Linguistics.
- 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 *NAACL*, pages 1357–1366.
- 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, Association for Computational Linguistics.
- Vulić, Ivan, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Post-specialisation: Retrofitting vectors of words unseen in lexical resources. 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 516–527.
- Vulić, Ivan and Anna Korhonen. 2016. Is “universal syntax” universally useful for learning distributed word representations? In *The 54th Annual Meeting of the Association for Computational Linguistics*, pages 518–524.
- 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)*, volume 1, pages 56–68.
- 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 (ACL)*, 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.
- Wan, Xiaojun. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on*

- Natural Language Processing of the AFNLP: Volume 1-volume 1*, pages 235–243.
- 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 (ACL)*, 5.
- 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 *EMNLP*, pages 1779–1785.
- 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. pages 450–460.
- Zhang, Duo, Qiaozhu Mei, and ChengXiang Zhai. 2010. Cross-lingual latent topic extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1128–1137.
- Zhang, Yuan and Regina Barzilay. 2015. Hierarchical low-rank tensors for multilingual transfer parsing. In *EMNLP*, pages 1857–1867.
- Zhang, Yuan, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag – multilingual POS tagging via coarse mapping between embeddings. In *NAACL*, pages 1307–1317.
- Zhang, Yuan, Roi Reichart, Regina Barzilay, and Amir Globerson. 2012. Learning to map into a universal POS tagset. In *EMNLP*, pages 1368–1378.
- 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. pages 1403–1412.
- Zou, Will Y, Richard Socher, Daniel Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398.