Emily M. Bender

Linguistic typology in natural language processing

DOI 10.1515/lingty-2016-0035
Received August 3, 2016; revised September 6, 2016

Abstract: This paper explores the ways in which the field of natural language processing (NLP) can and does benefit from work in linguistic typology. I describe the recent increase in interest in multilingual natural language processing and give a high-level overview of the field. I then turn to a discussion of how linguistic knowledge in general is incorporated in NLP technology before describing how typological results in particular are used. I consider both rule-based and machine learning approaches to NLP and review literature on predicting typological features as well as that which leverages such features.

1 Introduction

In this paper, I describe the relationship of the field of typology to COMPUTATIONAL LINGUISTICS and NATURAL LANGUAGE PROCESSING. Those latter two terms are sometimes used interchangeably to describe the field concerned with the processing of human language by computers. If a distinction is drawn between them, computational linguistics is used to describe research interested in answering linguistic questions using computational methodology, while natural language processing describes research on automatic processing of human language for practical applications, including such applications as information retrieval (e.g., web search), machine translation, the processing of clinical narratives in patient records to support patient care or clinical research, and computer-assisted language learning, among many others. The term natural language processing also sometimes contrasts with SPEECH PROCESSING, where the former concerns the processing and generation of text while the latter encompasses speech recognition and speech synthesis. (The term computational linguistics, however, encompasses research on both speech and text.) In this paper, I will mostly use the term natural language processing, abbreviated as NLP, as I am focusing more on how typological results can inform work on practical applications, with a greater focus on text than speech.

Emily M. Bender, Department of Linguistics, University of Washington, Guggenheim Hall, 4th Floor, Box 352425, Seattle, WA 98195, U.S.A., E-mail: ebender@uw.edu

There is a broad range of motivations for work in multilingual NLP. Researchers whose primary interest is in machine learning methodology often argue for the effectiveness of their machine learning algorithms on the basis of their applicability to multiple languages. Military and intelligence funding agencies are interested in supporting research on technology that allows NLP developers to quickly get systems running for languages that are not well-researched from a computational linguistics point of view but suddenly gain strategic relevance. Similarly, organizations that respond to natural disasters and other crises may benefit from the ability to quickly set up machine translation or automatic summarization in some language not previously widely studied from that angle. Businesses developing commercial products with NLP are interested in the markets represented by LOW RESOURCE LANGUAGES (LRLs; i.e., those languages for which there are not many digitized data sets or basic NLP systems such as part-of-speech taggers or morphological or syntactic parsers), some of which represent very large populations in emerging economies. Finally, researchers looking to apply NLP techniques to assist in endangered language documentation are naturally interested in developing NLP systems that work across very diverse languages.

These sources of interest in multilingual NLP have led to a number of recent workshops on the topic, including the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies (CLIAWS3) at the annual meeting of the North American Chapter of the Association for Computational Linguistics (NAACL:HLT) 2009 (Bandyopadhyay et al. (eds.) 2009), the First Workshop on Multilingual Modeling at the annual meeting of the Association for Coimputational Linguistics (ACL) 2012 (Jagarlamudi (eds.) 2012), the Workshop on Multilingual Multi-document Summarization at ACL 2013 (Giannakopoulos & Petasis (eds.) 2013), the Workshop on the Use of Computational Methods in the Study of Endangered Languages at ACL 2014 (Good et al. (eds.) 2014), and the Workshop on Multilingual and Cross-lingual Methods in NLP at NAACL 2016 (https://sites.google.com/site/multilingnlp/).[1]

In addition, there are popular multilingual data sets, including those from a series of shared tasks[2] in multilingual dependency parsing (Buchholz & Marsi 2006; Nivre, Hall, Kübler et al. 2007; Hajič et al. 2009) at the Conference on

[1] For work on multilingual speech processing, see, for example, Schultz & Kirchhoff (eds.) (2006).

[2] A SHARED TASK is a competition in which different research groups submit NLP systems for evaluation on some data set developed by the shared task organizers.

Computational Natural Language Learning (CoNLL), the Morpho-Challenge morphological segmentation and analysis data sets (Kurimo et al. 2010), and ODIN, the (massively multilingual) On-line Database of Interlinear Glossed Text (Lewis 2006; Lewis & Xia 2010; Xia et al. 2016), a collection of IGT harvested from linguistics papers in pdf on the web.

Finally, there are emerging standards for multilingual annotation, including Petrov et al.'s (2012) universal coarse-grained part-of-speech tagset and McDonald et al.'s (2013) universal syntactic dependency annotation scheme. From a typological point of view these crosslinguistic standards may be problematic, as Petrov et al. (2012) acknowledge, citing Evans & Levinson (2009). From a practical point of view, work in NLP cannot proceed without annotated data at the very least for evaluating systems and in many cases for training systems (see Section 2 below).

As observed in Bender (2009, 2011), some work in machine learning based NLP promotes the avoidance of using linguistic knowledge as a means of creating language-independent NLP systems. However, as I argue there, the best way to achieve crosslinguistically useful NLP is to build on the fruits of typological research, and in fact, such an approach is gaining currency in NLP. In the remainder of this paper, I will outline how linguistic knowledge is collected and used in NLP systems (Section 2) and describe possible ways in which typological results can be and are used in NLP (Section 3).

2 Linguistic knowledge and natural language processing

In this section, I provide a very high-level overview of NLP in order to discuss how it makes use of linguistic knowledge.[3] In very abstract terms, NLP systems involve encoding linguistic knowledge in some machine-readable form and then deploying that knowledge according to an algorithm in order to associate inputs with outputs. For example, in parsing, the input is text and the output is text associated with syntactic or semantic structure. In morphological analysis, the input is word forms (either in isolation or in running text) and the output is

[3] For more on this topic, see the papers in Volume 6 of *Linguistic Issues in Language Technology*, a special issue on the interaction of linguistics and computational linguistics (Baldwin & Kordoni (eds.) 2011).

word forms segmented into sequences of morphemes or roots and sets of morphological features. In machine translation, the input is text in one language and the output is text in another. In speech recognition, the input is an audio signal and the output is a text string. In summarization, the input is a document or set of documents and the output is a single short document, and so on. In all such cases, it is typical for the machine to find multiple possible outputs and these are usually ranked with respect to each other so that a user or further application can request either the top-ranked output or some list of *n* top outputs. That is, ambiguity is an inherent property of natural language, and NLP algorithms must both detect ambiguity and provide a means of navigating it.

Broadly speaking, approaches to NLP tasks can be divided into those that are RULE-BASED and those that involve MACHINE LEARNING, according to how the linguistic knowledge used by the algorithm is acquired.[4] A rule-based system includes linguistic knowledge in the form of rules directly encoded by a linguist in some formalism and typically on the basis of analysis of some particular DEVELOPMENT DATA. Machine learning systems include linguistic knowledge which is compiled, typically, by counting occurrences of properties of linguistic items in context. This counting process is referred to as TRAINING and involves observations over TRAINING DATA. These properties could be made explicit through (human) annotation of the data ahead of time (e.g., the parse trees in the English Penn Treebank (Marcus et al. 1993)) or could involve only information about the words and their contexts themselves (e.g., initial or final substrings, preceding or following words, or words within some context window). Both rule-based and machine learning systems can then be applied to additional, unseen data (inputs) to produce the appropriate type of outputs for that data. In evaluating such systems, the unseen data is called TEST DATA, and the system outputs are either evaluated by humans or compared against annotations produced by humans in order to measure system quality.[5]

[4] There are also hybrid systems which combine these approaches.

[5] Approaches using machine learning can be further distinguished by how the annotations available on the training data relate to the final output annotations the system is expected to make: Systems trained with data annotated by humans with the type of labels they are supposed to produce on the test data instantiate SUPERVISED learning. Systems trained on data that is not annotated for labels of the target output type represent UNSUPERVISED learning. Finally, systems that make use of some training data annotated with target output labels but supplement this with unannotated data are called SEMI-SUPERVISED. For a general overview of machine learning in NLP, see Manning & Schütze (1999).

In rule-based approaches the role of linguistic knowledge is directly apparent: the rules that the linguist encodes reflect linguistic knowledge (or linguistic hypotheses). In machine learning approaches, even the unsupervised ones, there are several ways in which linguistic knowledge can play a role: (i) in the design of annotation schemes (for supervised and semi-supervised machine learning), (ii) in the deployment of annotation schemes in actual annotation, (iii) in error analysis, when system developers look through system output for patterns of errors that can point to directions for improvement, and (iv) in the design of FEATURES. Features (in machine learning) are the templates that determine what about the data gets counted in the training of the system. For example, a simple type of feature for unsupervised systems is n-grams: the occurrence of single word forms (unigrams), sequences of two word forms (bigrams), etc. For a more complex example, consider a machine learning system trained to assign dependency trees in the style of the Prague Dependency Treebank (Böhmová et al. 2003) such as MaltParser (Nivre, Hall, Nilsson et al. 2007). MaltParser makes use of features that describe the partial structures being manipulated in the course of parsing, and include patterns such as the word currently being processed, its leftmost and rightmost dependents and the leftmost dependent of the next word to be processed.

This has been a very brief and high-level overview of the different approaches to NLP and how they make use of linguistic knowledge. My purpose here is not to give a comprehensive introduction to the field, but rather to provide enough context to support the discussion of the use of typological linguistic knowledge, given in Section 3.

3 Typology in CL/NLP

Just as linguistic knowledge in general can be incorporated into both rule-based and machine learning approaches to NLP, typological knowledge in particular can also inform both styles of research in this field. In Section 3.1, I describe the Grammar Matrix customization system, a long-term multilingual grammar engineering project that explicitly builds on the results of typological research, with specific attention to how that work is incorporated. Section 3.2 looks at a variety of machine learning style work in NLP that seeks to compensate for the lack of annotated resources in most of the world's languages by projecting information from resource-rich languages (like English) to others. Typological knowledge is playing an increasingly important role in such work. Finally, in Section 3.3, I look at the various ways that work in computational linguistics and NLP have targeted the *World atlas of language structures online* (Dryer & Haspelmath (eds.) 2013), as both a source of input data and a “gold standard” for evaluation in the extraction of typological features from various sources.

3.1 Grammar Matrix

The LinGO Grammar Matrix (Bender, Flickinger, & Oepen 2002; Bender, Drellishak et al. 2010) is a starter-kit for creating linguistically-motivated implemented (rule-based) grammars for natural languages. These grammars are couched in the framework of Head-Driven Phrase Structure Grammar (HPSG; Pollard & Sag 1994) and map sentence strings to semantic representations in the formalism of Minimal Recursion Semantics (MRS; Copestake et al. 2005). The grammars can be used, together with appropriate algorithms in both parsing (strings to semantics) or generation (semantics to strings), but are themselves simply declarative collections of linguistic knowledge.[6]

The Matrix represents a rule-based approach to computational linguistics/NLP and draws directly on typological results. It consists of a core grammar, containing constraints hypothesized to be crosslinguistically useful, and a series of “libraries” of analyses of crosslinguistically varying phenomena. The Matrix is accessed through a web-based questionnaire, where linguist-users iteratively describe both typological and lexical properties of the languages they are creating grammars for. Once a sufficient amount of consistent information has been entered, the system can generate (for user download) a starter grammar package which includes the Grammar Matrix core grammar as well as additional constraints which implement the linguistic description input through the questionnaire.

When developing libraries to add to the questionnaire, Matrix developers begin by consulting the typological literature in order to understand the range of variation in the phenomenon they are targeting. Table 1 lists libraries in the Grammar Matrix customization system, the works describing them, and their sources for typological information.

The Grammar Matrix libraries are not direct implementations of the analyses in the typological literature, however. The differences stem from the differing goals of Matrix developers and (other) typologists. The products of typological research tend to describe one or a small handful of phenomena at a time, and to

[6] The Grammar Matrix is developed in the context of the DELPH-IN consortium (http://www.delph-in.net). Implementations of the relevant parsing and generation algorithms are available as open-source software from DELPH-IN.

Table 1: Libraries in the Grammar Matrix and their typological sources.

Library	Citation	Typological sources
Coordination	Drellishak & Bender (2005)	Payne (1985); Stassen (2000); Drellishak (2004)
Person	Drellishak (2009)	Cysouw (2003); Siewierska (2004)
Number	Drellishak (2009)	Corbett (2000)
Gender	Drellishak (2009)	Corbett (1991)
Agreement	Drellishak (2009)	Corbett (2006)
Case	Drellishak (2009)	Comrie (1989); Dixon (1994)
Direct-inverse	Drellishak (2009)	Givón (1994)
Argument Optionality	Saleem (2010); Saleem & Bender (2010)	Ackema et al. (eds.) (2006); Dryer (2008)
Tense	Poulson (2011)	Comrie (1985); Dahl (1985); Bybee et al. (1994), *inter alia*
Aspect	Poulson (2011)	Comrie (1976); Dahl (1985); Bybee et al. (1994), *inter alia*
Sentential Negation	Crowgey (2012)	Dahl (1979); Dryer (2005)
Information Structure	Song (2014)	Féry & Krifka (2009); Buring (2010), *inter alia*
Adjectives	Trimble (2014)	Stassen (2003, 2013); Dixon (2004); Dryer (2013a), *inter alia*

involve analytical categories which can be used to describe languages, investigate the distribution and co-occurrence of said analytical categories across languages, and support hypotheses about the reasons for patterns in language change over time. The goals of the Grammar Matrix developers, on the other hand, are to create analyses of particular phenomena which support the mapping of surface strings to semantic representations as well as the delineation of grammatical from ungrammatical strings and which all work together in the parsing (and generation) of particular sentences (Bender 2008).

Accordingly, the Grammar Matrix project hopes to also inform typological research; it is a scientific goal of the Grammar Matrix project "to use computational methods to combine the results of typological research and formal syntactic analysis into a single resource that achieves both typological breadth (handling the known range of realizations of the phenomena analyzed) and analytical depth (producing analyses which work together to map surface strings to semantic representations)'" (Bender, Drellishak et al. 2010: 24). Furthermore, the Grammar Matrix project can contribute to typological research by producing language profiles in the form of implemented grammars based on the Matrix (and language-specific descriptive resources) for a wide variety of languages, as well as associated documentation. These are distributed through the Language CoLLAGE (Bender 2014).[7]

3.2 Projecting resources across languages

As noted above, an important driver for NLP research is the development of language resources. The most basic language resource is a corpus (collection of texts). Other types of resources include pronouncing dictionaries, BITEXTS (parallel corpora, where text from one language is translated to the other and typically annotated for alignments at the sentence level), corpora annotated for linguistic and extralinguistic information such as part-of-speech tags, morphological information, named-entity types (e.g., “person”, “organization”, etc), TREEBANKS (corpora annotated with constituent structures or syntactic dependencies), and SEMBANKS (corpora annotated with semantic representations). Such resources can in turn be used to create language processing tools, such as part-of-speech taggers, morphological analyzers, parsers, etc. A few of the world’s languages (notably English, Spanish, French, Chinese,[8] and Arabic) have a wide variety of such resources. Perhaps a couple dozen more languages have robust resources of at least some of the types.[9] The vast majority of the world’s languages, however, including some widely spoken ones, are low resource languages (LRLs).

In creating NLP resources (tagged corpora, as well as text processing components, such as part-of-speech taggers or parsers) for LRLs, one strategy is to use information available for a high resource language and adapt it in some way to the LRL. This can be aided by using bitexts between the two languages, using statistical methods to align words across the bitexts (notably those developed for statistical machine translation by Brown et al. (1990)), and then using those links to align annotations from the source language resource to words in the target language. This technique was pioneered by Yarowsky et al. (2001), who applied it to the tasks of part-of-speech tagging, noun phrase bracketing, named-entity tagging, and morphological analysis. Hwa et al. (2005) extend this idea to the more elaborate representations of syntactic dependency

[7] Available from http://depts.washington.edu/uwcl/matrix/language-collage/

[8] Most such resources focus on the written language and are described as representing “Chinese” rather than some specific Chinese language.

[9] For an emerging catalog of resources, see the LRE Map (Calzolari et al. 2012): http://www.resourcebook.eu/searchll.php

structures. Georgi et al. (2012) take advantage of the IGT in ODIN (Lewis & Xia 2010) to show how this strategy can be deployed for a very wide variety of LRLs.[10]

It is also possible to share information across languages without bitexts (parallel corpora). This is explored by Zeman & Resnik (2008), Søgaard (2011), and others for the purpose of creating dependency parsers for low resource languages, under the rubric of DELEXICALIZED TRANSFER. The key idea here is that the training for the parser does not include the words of the source language directly in the features being counted, but rather works instead in terms of properties which will also be applicable in the target language, such as part-of-speech tags drawn from a universal set (like those of McDonald et al. 2013; see Section 1). The performance of such parsers has been improved by working with typological knowledge: Naseem et al. (2012) (and later Täckström et al. (2013)) break the parsing problem down into two steps, first choosing heads and dependents then choosing their order. The statistical model for the first step is the same for all languages, while the model for the second step varies depending on the language being parsed. In particular, the model is trained on a large set of source languages, but then customized to the target language based on its typological properties as encoded in the *World atlas of language structures* (*WALS*; Haspelmath et al. (eds.) 2005) features 81A and 85A to 89A (Dryer 2013a, b, c, d, e, f). In particular, these systems take training information for any given dependency type from source languages which share the value for the relevant *WALS* feature(s) with the language being processed, and ignore information from source languages which differ (for the phenomenon in question). In this way, language-level typological information is directly leveraged to improve the automatic processing of specific sentences in LRLs.[11]

3.3 Predicting or extracting typological features

As noted above, *WALS* – especially in its online form (Dryer & Haspelmath (eds.) 2013) – has been used as a source of information in typologically-informed approaches to sharing resources across languages. Computational linguists

[10] One advantage of IGT is that it supports the alignment of words from source to target language even when the amount of text available doesn't meet the needs of statistical word alignment.

[11] For further work using *WALS* features to inform dependency parsing, see Zhang & Barzilay (2015) and Ammar et al. (2016).

have also been interested in “predicting” values for *WALS* features. In some cases (e.g., Rama & Kolachina 2012) this is explicitly motivated in terms of determining which languages are good candidates as source/target pairs in sharing resources, even when they lack values for the specific features of interest in *WALS*. That is, if a statistical model can give a reasonable guess as to what the feature would be, that can be used as a basis for deciding how to share resources, even if the guess isn’t necessarily correct.

Models for predicting *WALS* feature values represent two main approaches. On the one hand, there is work (e.g., Teh et al. 2007; Rama & Kolachina 2012) that attempts to predict values for certain features based on other values in *WALS*, by making statistical models that can capture dependencies between the features. These are tested by “hiding” some proportion of the known features and seeing how well the model can predict them. Closely related work looks at *WALS* as an input data set for statistical models to either propose candidate typological implications (Daumé & Campbell 2007; Lu 2013), detect areal features (Daumé 2009), or reconstruct language family trees (Georgi et al. 2010).

On the other hand, there are projects which aim to predict *WALS* values based on running text enriched with some other kind of information. Lewis & Xia (2008) use IGT from ODIN (Lewis & Xia 2010), parse the translation line using resources for English and project that structure to the source language line. The projected structures as well as the glosses in the gloss line are then used to estimate values for *WALS* features concerning constituent order, morpheme order, and constituent presence/absence (e.g., articles). Östling (2015) uses parallel texts from New Testament translations and word alignments between these to estimate values for *WALS* word order features. In these cases, since *WALS* feature values aren’t used as input, the model can be tested against existing *WALS* values for all languages present in the training data.

The AGGREGATION Project (Bender, Goodman et al. 2013; Bender, Crowgey et al. 2014) is combining the methods of Lewis & Xia (2008) to extract typological information from IGT with the typological knowledge encoded in the Grammar Matrix customization system. Here the basic idea is to use structure projected from the English translation through the gloss to the source language line (as well as the glosses themselves) to extract not *WALS*-style features but rather the finer-grained information required by the Grammar Matrix customization system questionnaire. When sufficient information can be gleaned in this fashion, the customization system can then be run, creating a small (but functioning) grammar for parsing and generating sentences of the language in question.

3.4 Summary

This section has provided an overview of different ways in which typological information is used in natural language processing. It is not meant to be comprehensive (there is much further relevant work not cited here), but rather to give a general sense of the state of the field. Also not discussed in detail here is the evaluation of NLP systems. It is in the nature of the field and the problems approached that NLP systems don't achieve perfect performance; they all make errors some of the time (e.g., incorrect parses from parsers, incorrect values from systems predicting *WALS*-style feature values, etc.). Thus, for example, a system filling in missing values in *WALS* is not meant to provide definitive answers, but rather to work in tandem with other systems that can use such information. Crucially, the quality of all of these systems is measurable and in fact measured. The reader is referred to the original papers for descriptions of evaluation methodology and specific numerical results.

4 Conclusion

Why does NLP need linguistic typology? NLP researchers strive to develop systems that work with a wide variety of human languages (or perhaps any possible human language), for both practical reasons and scientific ones: Practically, it isn't feasible to spend as much time developing resources for every one of the world's ~6,000 languages as has been spent for English and the handful of other resource-rich languages. Scientifically, a system that works across multiple languages is thought to tell us something deeper about either the nature of language or the machine learning process deployed than one that is closely adapted to a single language. In order to succeed in this multilingual program, however, NLP needs to be informed about the ways in which languages vary, the extent of that variation, and the positions that particular languages occupy in that high dimensional space. In short, the products of linguistic typological research are key to multilingual natural language processing and the field of NLP has begun to develop methodologies for effectively incorporating these results.

Acknowledgments: I would like to thank Antske Fokkens, Gina-Anne Levow, and Olga Zamaraeva for helpful discussion in the preparation of this paper. All remaining errors and infelicities are my own.

References

Ackema, Peter, Patrick Brandt, Maaike Schoorlemmer & Fred Weerman (eds.). 2006. *Arguments and agreement*. Oxford: Oxford University Press.

Ammar, Waleed, George Mulcaire, Miguel Ballesteros, Chris Dyer & Noah A. Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics* 4. 431–444. https://www.transacl.org/ojs/index.php/tacl/article/view/892

Baldwin, Timothy & Valia Kordoni (eds.). 2011. The interaction between linguistics and computational linguistics: Virtuous, vicious or vacuous? Special issue of *Linguistic Issues in Language Technology* 6. http://journals.linguisticsociety.org/elanguage/lilt/issue/view/330.html

Bandyopadhyay, Sivaji, Pushpak Bhattacharya, Vasudeva Varma, Sudeshna Sarkar, A. Kumaran & Raghavendra Udupa (eds.). 2009. *Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies (CLIAWS3)*, June 4, 2009, *Boulder, Colorado*. Madison, WI: Omnipress. http://www.aclweb.org/anthology/W09-16

Bender, Emily M. 2008. Grammar engineering for linguistic hypothesis testing. *Texas Linguistics Society* 10. 16–36.

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?*, 26–32. Vrilissia, Greece: Tehnografia Digital Press. http://www.aclweb.org/anthology/W09-0106

Bender, Emily M. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology* 6(3). 1–26. http://journals.linguisticsociety.org/elanguage/lilt/article/view/2624.html

Bender, Emily M. 2014. Language CoLLAGE: Grammatical description with the LinGO Grammar Matrix. *International Conference on Language Resources and Evaluation* 9. 2447–2451. http://www.lrec-conf.org/proceedings/lrec2014/pdf/639_Paper.pdf

Bender, Emily M., Joshua Crowgey, Michael Wayne Goodman & Fei Xia. 2014. Learning grammar specifications from IGT: A case study of Chintang. In Good et al. (eds.) 2014, 43–53. http://www.aclweb.org/anthology/W14-2206

Bender, Emily M., Scott Drellishak, Antske Fokkens, Laurie Poulson & Safiyyah Saleem. 2010. Grammar customization. *Research on Language and Computation* 23–72.

Bender, Emily M., Dan Flickinger & Stephan Oepen. 2002. The grammar matrix: An open-source starter-kit for the rapid development of crosslinguistically consistent broad-coverage precision grammars. *International Conference on Computational Linguistics* 19 (Workshop on Grammar Engineering and Evaluation). 8–14. http://www.aclweb.org/anthology/W02-1502

Bender, Emily M., Michael Wayne Goodman, Joshua Crowgey & Fei Xia. 2013. Towards creating precision grammars from interlinear glossed text: Inferring large-scale typological properties. *Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* 7. 74–83. http://www.aclweb.org/anthology/W13-2710

Böhmová, Alena, Jan Hajič, Eva Hajičová & Barbora Hladká. 2003. The Prague Dependency Treebank. In Anne Abeillé (ed.), *Treebanks: Building and using parsed corpora*, 103–127. Dordrecht: Kluwer.

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer & Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics* 16. 79–85.

Buchholz, Sabine & Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. *Conference on Computational Natural Language Learning* 10. 149–164. http://www.aclweb.org/anthology/W06-2920

Büring, Daniel. 2010. Towards a typology of focus realization. In Malte Zimmermann & Caroline Féry (eds.), *Information structure*, 177–205. Oxford: Oxford University Press.

Bybee, Joan L., Revere Perkins & William Pagliuca. 1994. *The evolution of grammar: Tense, aspect and modality in the languages of the world*. Chicago: University of Chicago Press.

Calzolari, Nicoletta, Riccardo Del Gratta, Gil Francopoulo, Joseph Mariani, Francesco Rubino, Irene Russo & Claudia Soria. 2012. The LRE map: Harmonising community descriptions of resources. *International Conference on Language Resources and Evaluation* 8. 1084–1089. http://www.lrec-conf.org/proceedings/lrec2012/pdf/769_Paper.pdf

Comrie, Bernard. 1976. *Aspect: An introduction to the study of verbal aspect and related problems*. Cambridge: Cambridge University Press.

Comrie, Bernard. 1985. *Tense*. Cambridge: Cambridge University Press.

Comrie, Bernard. 1989. *Language universals and linguistic typology*. 2nd edn. Chicago: University of Chicago Press.

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

Corbett, Greville G. 1991. *Gender*. Cambridge: Cambridge University Press.

Corbett, Greville G. 2000. *Number*. Cambridge: Cambridge University Press.

Corbett, Greville G. 2006. *Agreement*. Cambridge: Cambridge University Press.

Crowgey, Joshua. 2012. *The syntactic exponence of sentential negation: A model for the LinGO Grammar Matrix*. Seattle: University of Washington MA thesis. http://hdl.handle.net/1773/22454

Cysouw, Michael. 2003. *The paradigmatic structure of person marking*. Oxford: Oxford University Press.

Dahl, Östen. 1979. Typology of sentence negation. *Linguistics* 17. 79–106.

Dahl, Östen. 1985. *Tense and aspect systems*. Oxford: Blackwell.

Daumé, Hal, III. 2009. Non-parametric Bayesian areal linguistics. *North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 2009(1). 593–601. http://www.aclweb.org/anthology/N09-1067

Daumé, Hal, III & Lyle Campbell. 2007. A Bayesian model for discovering typological implications. *Association of Computational Linguistics* 45(1). 65–72. http://www.aclweb.org/anthology/P07-1009

Dixon, R. M. W. 1994. *Ergativity*. Cambridge: Cambridge University Press.

Dixon, R. M. W. 2004. Adjective classes in typological perspective. In R. M. W. Dixon & Alexandra Y. Aikhenvald (eds.), *Adjective classes: A cross-linguistic typology*, 1–49. Oxford: Oxford University Press.

Drellishak, Scott. 2004. *A survey of coordination strategies in the world's languages*. Seattle: University of Washington MA thesis.

Drellishak, Scott. 2009. *Widespread but not universal: Improving the typological coverage of the Grammar Matrix*. Seattle: University of Washington doctoral dissertation.

Drellishak, Scott & Emily M. Bender. 2005. A coordination module for a crosslinguistic grammar resource. *International Conference on Head-Driven Phrase Structure Grammar* 12. 108–128. http://web.stanford.edu/group/cslipublications/cslipublications/HPSG/2005/drellishak-bender.pdf

Dryer, Matthew S. 2005. Negative morphemes. In Haspelmath et al. (eds.) 2005, 454–457.

Dryer, Matthew S. 2008. Expression of pronominal subjects. In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.), *The world atlas of language structures online*, Chapter 101. München: Max Planck Digital Library. http://wals.info/feature/101

Dryer, Matthew S. 2013a. Order of adjective and noun. In Dryer & Haspelmath (eds.) 2013, Chapter 87. http://wals.info/feature/87

Dryer, Matthew S. 2013b. Order of adposition and noun phrase. In Dryer & Haspelmath (eds.) 2013, Chapter 85. http://wals.info/chapter/85

Dryer, Matthew S. 2013c. Order of demonstrative and noun. In Dryer & Haspelmath (eds.) 2013, Chapter 88. http://wals.info/chapter/88

Dryer, Matthew S. 2013d. Order of genitive and noun. In Dryer & Haspelmath (eds.) 2013, Chapter 86. http://wals.info/chapter/86

Dryer, Matthew S. 2013e. Order of numeral and noun. In Dryer & Haspelmath (eds.) 2013, Chapter 89. http://wals.info/chapter/89

Dryer, Matthew S. 2013f. Order of subject, object and verb. In Dryer & Haspelmath (eds.) 2103, Chapter 81. http://wals.info/chapter/81

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *The world atlas of language structures online*. Leipzig: Max Planck Institut für evolutionäre Anthropologie. http://wals.info/

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

Féry, Caroline & Manfred Krifka. 2009. Information structure: Notional distinctions, ways of expression. In Piet van Sterkenburg (ed.), *Unity and diversity of languages*, 123–135. Amsterdam: Benjamins.

Georgi, Ryan, Fei Xia & William D. Lewis. 2010. Comparing language similarity across genetic and typologically-based groupings. *International Conference on Computational Linguistics* 23. 385–393. http://www.aclweb.org/anthology/C10-1044

Georgi, Ryan, Fei Xia & William D. Lewis. 2012. Improving dependency parsing with interlinear glossed text and syntactic projection. *International Conference on Computational Linguistics* 24(Posters), 371–380. http://www.aclweb.org/anthology/C12-2037

Giannakopoulos, George & Georgios Petasis (eds.). 2013. *Proceedings of the workshop "Multilingual multi-document summarization" (MultiLing* 2013), August 9, 2013, *Sofia, Bulgaria*. Madison, WI: Omnipress. http://www.aclweb.org/anthology/W13-31

Givón, T. 1994. The pragmatics of de-transitive voice: Functional and typological aspects of inversion. In T. Givón (ed.), *Voice and inversion*, 3–44. Amsterdam: Benjamins.

Good, Jeff, Julia Hirschberg & Owen Rambow (eds.). 2014. *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages* (ComputEL 2014), June 26, 2014, *Baltimore, Maryland, USA*. http://www.aclweb.org/anthology/W14-22

Hajič, Jan, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jann Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue & Yi Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. *Conference on Computational Natural Language Learning* 13(2: Shared Task). 1–18. http://www.aclweb.org/anthology/W09-1201

Haspelmath, Martin, Matthew Dryer, David Gil & Bernard Comrie (eds.). 2005. *The world atlas of language structures*. Oxford: Oxford University Press.

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

Jagarlamudi, Jagadeesh, Sujith Ravi, Xiaojun Wan & Hal Daumé III (eds.). 2012. *Proceedings of the First Workshop on Multilingual Modeling, July 13, 2012, Jeju, Republic of Korea*. http://www.aclweb.org/anthology/W12-39

Kurimo, Mikko, Sami Virpioja, Ville Turunen & Krista Lagus. 2010. Morpho Challenge competition 2005–2010: Evaluations and results. *ACL Special Interest Group on Computational Morphology and Phonology* 11. 87–95. http://www.aclweb.org/anthology/W10-2211

Lewis, William D. 2006. ODIN: A model for adapting and enriching legacy infrastructure. *IEEE International Conference on E-Science* 2. 137.

Lewis, William D. & Fei Xia. 2008. Automatically identifying computationally relevant typological features. *International Joint Conference on Natural Language Processing* 3(2). 685–690. http://www.aclweb.org/anthology/I08-2093

Lewis, William D. & Fei Xia. 2010. Developing ODIN: A multilingual repository of annotated language data for hundreds of the world's languages. *Journal of Literary and Linguistic Computing* 25. 303–319.

Lu, Xia. 2013. Exploring word order universals: A probabilistic graphical model approach. *Association for Computational Linguistics* 51(3: Student research workshop). 150–157. http://www.aclweb.org/anthology/P13-3022

Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Marcus, Mitchell P., Beatrice Santorini & Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19. 313–330.

McDonald, Ryan, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló & Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. *Association for Computational Linguistics* 51(2: Short papers). 92–97. http://www.aclweb.org/anthology/P13-2017

Naseem, Tahira, Regina Barzilay & Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. *Association for Computational Linguistics* 50(1: Long papers). 629–637. http://www.aclweb.org/anthology/P12-1066

Nivre, Joakim, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel & Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. *Joint Conference on Empirical Methods in Natural Language Processing & Computational Natural Language Learning* 2007. 915–932. http://www.aclweb.org/anthology/D/D07/D07-1096

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryigit, Sandra Kübler, Svetoslav Marinov & Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering* 13. 95–135.

Östling, Robert. 2015. Word order typology through multilingual word alignment. *Association for Computational Linguistics* 53(2: Short papers). 205–211. http://www.aclweb.org/anthology/P15-2034

Payne, John R. 1985. Complex phrases and complex sentences. In Timothy Shopen (ed.), *Language typology and syntactic description*, Vol. 2: *Complex constructions*, 3–41. Cambridge: Cambridge University Press.

Petrov, Slav, Dipanjan Das & Ryan McDonald. 2012. A universal part-of-speech tagset. *International Conference on Language Resources and Evaluation* 8. 2089–2096. http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf

Pollard, Carl & Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Poulson, Laurie. 2011. Meta-modeling of tense and aspect in a crosslinguistic grammar engineering platform. *University of Washington Working Papers in Linguistics* 28. http://http://depts.washington.edu/uwwpl/vol28/poulson_2011.pdf

Rama, Taraka & Prasanth Kolachina. 2012. How good are typological distances for determining genealogical relationships among languages? *International Conference on Computational Linguistics* 24(Posters). 975–984. http://www.aclweb.org/anthology/C12-2095

Saleem, Safiyyah. 2010. *Argument optionality: A new library for the grammar matrix customization system*. Seattle: University of Washington MA thesis.

Saleem, Safiyyah & Emily M. Bender. 2010. Argument optionality in the LinGO Grammar Matrix. *International Conference on Computational Linguistics* 23(Posters). 1068–1076. http://www.aclweb.org/anthology/C10-2123

Schultz, Tanja & Katrin Kirchhoff (eds.). 2006. *Multilingual speech processing*. Burlington, MA: Academic Press.

Siewierska, Anna. 2004. *Person*. Cambridge: Cambridge University Press.

Søgaard, Anders. 2011. Data point selection for cross-language adaptation of dependency parsers. *Association for Computational Linguistics: Human Language Technologies* 49(2). 682–686. http://www.aclweb.org/anthology/P11-2120

Song, Sanghoun. 2014. *A grammar library for information structure*. Seattle: University of Washington doctoral dissertation. http://hdl.handle.net/1773/25372

Stassen, Leon. 2000. AND-languages and WITH-languages. *Linguistic Typology* 4. 1–54.

Stassen, Leon. 2003. *Intransitive predication*. Oxford: Oxford University Press.

Stassen, Leon. 2013. Predicative adjectives. In Dryer & Haspelmath (eds.) 2013, Chapter 118. http://wals.info/feature/118

Täckström, Oscar, Ryan McDonald & Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers. *North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 2013(1). 1061–1071. http://www.aclweb.org/anthology/N13-1126

Teh, Yee W., Hal Daumé III & Daniel M. Roy. 2007. Bayesian agglomerative clustering with coalescents. In John C. Platt, Daphne Koller, Yoram Singer & Sam T. Roweis (eds.), *Advances in neural information processing systems* 20. 1463–1480. Cambridge, MA: MIT Press.

Trimble, Thomas James. 2014. *Adjectives in the LinGO Grammar Matrix*. Seattle: University of Washington MS thesis. http://hdl.handle.net/1773/27512

Xia, Fei, William D. Lewis, Michael Wayne Goodman, Glenn Slayden, Ryan Georgi, Joshua Crowgey & Emily M. Bender. 2016. Enriching a massively multilingual database of interlinear glossed text. *Language Resources and Evaluation* 50. 321–349.

Yarowsky, David, Grace Ngai & 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*, 1–8. http://www.aclweb.org/anthology/H01-1035

Zeman, Daniel & Philip Resnik. 2008. Cross-language parser adaptation between related languages. *International Joint Conference on Natural Language Processing* 3(Workshop on NLP for Less Privileged Languages). 35–42. http://www.aclweb.org/anthology/I08-3008

Zhang, Yuan & Regina Barzilay. 2015. Hierarchical low-rank tensors for multilingual transfer parsing. *Conference on Empirical Methods in Natural Language Processing* 2015. 1857–1867. http://aclweb.org/anthology/D15-1213