# A Language-Independent Feature Schema for Inflectional Morphology

**John Sylak-Glassman*, Christo Kirov*, David Yarowsky**, Roger Que****
*Center for Language and Speech Processing
**Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218
jcsg@jhu.edu, ckirov@gmail.com, yarowsky@jhu.edu, query@jhu.edu

## Abstract

This paper presents a universal morphological feature schema that represents the finest distinctions in meaning that are expressed by overt, affixal inflectional morphology across languages. This schema is used to universalize data extracted from Wiktionary via a robust multidimensional table parsing algorithm and feature mapping algorithms, yielding 883,965 instantiated paradigms in 352 languages. These data are shown to be effective for training morphological analyzers, yielding significant accuracy gains when applied to Durrett and DeNero's (2013) paradigm learning framework.

## 1 Introduction

Semantically detailed and typologically-informed morphological analysis that is broadly cross-linguistically applicable and interoperable has the potential to improve many NLP applications, including machine translation (particularly of morphologically rich languages), parsing (Choi et al., 2015; Zeman, 2008; Mikulová et al., 2006), $n$-gram language models, information extraction, and co-reference resolution.

To do large-scale cross-linguistic analysis and translation, it is necessary to be able to compare the meanings of morphemes using a single, well-defined framework. Haspelmath (2010) notes that while morphological categories will never map with perfect precision across languages and can only be exhaustively defined within a single language, practitioners of linguistic typology have typically recognized that there is sufficient similarity in these categories across languages to do meaningful comparison. For this purpose, Haspelmath (2010) proposes that typologists precisely define dedicated language-independent comparative concepts and identify the presence of these concepts in specific languages. In this spirit, we present a universal morphological feature schema, in which features that have a status akin to those of comparative concepts are used to represent the finest distinctions in meaning that are expressed by inflectional morphology across languages. This schema can in turn be used to universalize morphological data from the world's languages, which allows for direct comparison and translation of morphological material across languages. This greatly increases the amount of data available to morphological analysis tools, since data from any language can be specified in a common format with the same features.

Wiktionary constitutes one of the largest available sources of complete morphological paradigms across diverse languages, with substantial ongoing growth in language and lemma coverage, and hence forms a natural source of data for broadly multilingual supervised learning. Wiktionary paradigm table formats, however, are often complex, nested, 2-3 dimensional structures intended for human readability rather than machine parsing, and are broadly inconsistent across languages and Wiktionary editions. This paper presents an original, robust multidimensional table parsing system that generalizes effectively across these languages, collectively yielding significant gains in supervised morphological paradigm learning in Durrett and DeNero's (2013) framework.

## 2 Universal Morphological Feature Schema

The purpose of the universal morphological feature schema is to allow any given overt, affixal (non-root) inflectional morpheme in any language to be given a precise, language-independent definition. The schema is composed of a set of features that represent semantic "atoms" that are never decomposed into more finely differentiated meanings in any natural language. This ensures that the meanings of all inflectional morphemes are able to be represented either through single features or through multiple features in combina-



*Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, pages 674–680, Beijing, China, July 26-31, 2015. 

tion. These features capture only the semantic content of morphemes, but can be integrated into existing frameworks that precisely indicate morpheme form (Sagot and Walther, 2013) or automatically discover it (Dreyer and Eisner, 2011; Hammarström, 2006; Goldsmith, 2001). The fact that the schema is meant to capture only the meanings of overt, non-root affixal morphemes restricts the semantic-conceptual space that must be captured by its features and renders an interlingual approach to representing inflectional morphology feasible.

The universal morphological feature schema is most similar to tagset systematization efforts across multiple languages, such as the Universal Dependencies Project (Choi et al., 2015) and Interset (Zeman, 2008). While these efforts encode similar morphological features to the current schema, their goal is different, namely to systematize pre-existing tagsets, which include lexical and syntactic information, for 30 specific languages. The goal of the schema presented here is to capture the most basic meanings encoded by inflectional morphology across all the world's languages and to define those meanings in a language-independent manner. Because of its wide-scope, our universal morphological feature schema will likely need to include other features and even other dimensions of meaning, for which the authors invite suggestions.

### 2.1 Construction Methodology

The first step in constructing the universal morphological feature schema was to identify the dimensions of meaning (e.g. case, number, tense, mood, etc.) that are expressed by inflectional morphology in the world's languages. These were identified by surveying the linguistic typology literature on parts of speech and then identifying the kinds of inflectional morphology that are typically associated with each part of speech.

For each dimension, we identified the finest distinctions in meaning made within that dimension by a natural language. Some higher-level 'cover features' representing common cross-linguistic groupings were also included. For example, features such as indicative (IND) and subjunctive (SBJV) represent groupings of basic modality features which occur in multiple languages and show similar usage patterns (Palmer, 2001).

Each dimension has an underlying semantic basis used to define its features. To determine the underlying semantic basis for each dimension, the literature in linguistic typology and in description-oriented linguistic theory was surveyed for explanations of each dimension that offered ways to precisely define the observed features.

### 2.2 Contents of the Schema

The universal morphological feature schema represents 23 dimensions of meaning with 212 features. Because space limitations preclude a detailed discussion of the semantic basis of each dimension and the definitions of each feature, Table 1 presents each dimension of meaning, the labels of its features, and citations for the main sources for the semantic bases of each dimension. To the extent possible, feature labels conform to the Leipzig Glossing Rules (Comrie et al., 2008) and to the labels in the sources used to define the semantic basis for each dimension of meaning. A substantially expanded exploration and analysis of these dimensions and schema framework may be found in Sylak-Glassman et al. (To appear).

Note that because gender categories are not necessarily defined by semantic criteria and rarely map neatly across languages, this schema treats gender features as open-class.[1]

## 3 Wiktionary Data Extraction and Mapping

Wiktionary contains a wealth of training data for morphological analysis, most notably inflectional paradigm tables. Since its pages are primarily written by human authors for human readers, and there are no overarching standards for how paradigms should be presented, these tables contain many inconsistencies and are at best semi-structured. Layouts differ depending on the *edition language* in which a word is being defined and within an edition depending on the word's language and part of speech. The textual descriptors used for morphological features are also not systematically defined. These idiosyncrasies cause numerous difficulties for automatic paradigm extraction, but the redundancy of having data presented in multiple ways across different editions gives us an opportunity to arrive at a consensus description of an inflected form, and to fill in gaps when the coverage of one edition diverges from

[1] To limit feature proliferation, the schema encodes gender categories as features that may be shared across languages within a phylogenetic stock or family, in order to capture identical gender category definitions and assignments that result from common ancestry, as may be possible for the 25 historical noun classes in the Bantu stock (Demuth, 2000).

| *Dimension* | *Features* | *Semantic Basis* |
|---|---|---|
| Aktionsart | ACCMP, ACH, ACTY, ATEL, DUR, DYN, PCT, SEMEL, STAT, TEL | Cable (2008), Vendler (1957), Comrie (1976a) |
| Animacy | ANIM, HUM, INAN, NHUM | Yamamoto (1999), Comrie (1989) |
| Aspect | HAB, IPFV, ITER, PFV, PRF, PROG, PROSP | Klein (1994) |
| Case | ABL, ABS, ACC, ALL, ANTE, APPRX, APUD, AT, AVR, BEN, CIRC, COM, COMPV, DAT, EQU, ERG, ESS, FRML, GEN, INS, IN, INTER, NOM, NOMS, ON, ONHR, ONVR, POST, PRIV, PROL, PROPR, PROX, PRP, PRT, REM, SUB, TERM, VERS, VOC | Blake (2001), Radkevich (2010) |
| Comparison | AB, CMPR, EQT, RL, SPRL | Cuzzolin and Lehmann (2004) |
| Definiteness | DEF, INDEF, NSPEC, SPEC | Lyons (1999) |
| Deixis | ABV, BEL, DIST, EVEN, MED, NVIS, PROX, REF1, REF2, REM, VIS | Bhat (2004), Bliss and Ritter (2001) |
| Evidentiality | ASSUM, AUD, DRCT, FH, HRSY, INFER, NFH , NVSEN, QUOT, RPRT, SEN | Aikhenvald (2004) |
| Finiteness | FIN, NFIN | Binary finite vs. nonfinite |
| Gender+ | BANTU1-23, FEM, MASC, NAKH1-8, NEUT | Corbett (1991) |
| Info. Structure | FOC, TOP | Lambrecht (1994) |
| Interrogativity | DECL, INT | Binary declarative vs. interrogative |
| Mood | ADM, AUNPRP, AUPRP, COND, DEB, IMP, IND, INTEN, IRR, LKLY, OBLIG, OPT, PERM, POT, PURP, REAL, SBJV, SIM | Palmer (2001) |
| Number | DU, GPAUC, GRPL, INVN, PAUC, PL, SG, TRI | Corbett (2000) |
| Parts of Speech | ADJ, ADP, ADV, ART, AUX, CLF, COMP, CONJ, DET, INTJ, N, NUM, PART, PRO, V, V.CVB, V.MSDR, V.PTCP | Croft (2000), Haspelmath (1995) |
| Person | 0, 1, 2, 3, 4, EXCL, INCL, OBV, PRX | Conventional person, obviation and clusivity |
| Polarity | NEG, POS | Binary positive vs. negative |
| Politeness | AVOID, COL, FOREG, FORM, FORM.ELEV, FORM.HUMB, HIGH, HIGH.ELEV, HIGH.SUPR, INFM, LIT, LOW, POL | Brown and Levinson (1987), Comrie (1976b) |
| Possession | ALN, NALN, PSSD, PSSPNO+ | Type of possession, characteristics of possessor |
| Switch-Reference | CN-R-MN+, DS, DSADV, LOG, OR, SEQMA, SIMMA, SS, SSADV | Stirling (1993) |
| Tense | 1DAY, FUT, HOD, IMMED, PRS, PST, RCT, RMT | Klein (1994), **?**) |
| Valency | DITR, IMPRS, INTR, TR | Number of verbal arguments from zero to three |
| Voice | ACFOC, ACT, AGFOC, ANTIP, APPL, BFOC, CAUS, CFOC, DIR, IFOC, INV, LFOC, MID, PASS, PFOC, RECP, REFL | Klaiman (1991) |

Table 1: Dimensions of meaning and their features, both sorted alphabetically

that of another.

To make these data available for morphological analysis, we developed a novel multidimensional table parser for Wiktionary to extract inflected forms with their associated descriptors. Although we describe its function in Wiktionary-specific terms, this strategy can be generalized to extract data tuples from any HTML table with correctly marked-up header and content cells. We extracted additional descriptors from HTML headings and table captions, then mapped all descriptors to features in the universal schema.

### 3.1 Extraction from HTML Tables

In its base form, the table parser takes advantage of HTML's distinction between header and content cells to identify descriptors and potential inflected forms, respectively, in an arbitrary inflection table. Each content cell is matched with the headers immediately up the column, to the left of the row, and in the "corners" located at the row and column intersection of the previous two types of headers. Matching headers are stored in a list ordered by their distance from the content cell. Figure 1 shows an example where *prenais* is assigned the following descriptors:

– Directly up the column: **tu**, **second**, **singular**, **simple**.

– Directly to the left of the row: **imperfect**, **simple tenses**.

– In corners located at the row and column intersection of any headers identified by the previous two methods: **indicative**, **person**.

– Important structured fields found outside the table, including **French** and **Verb**.

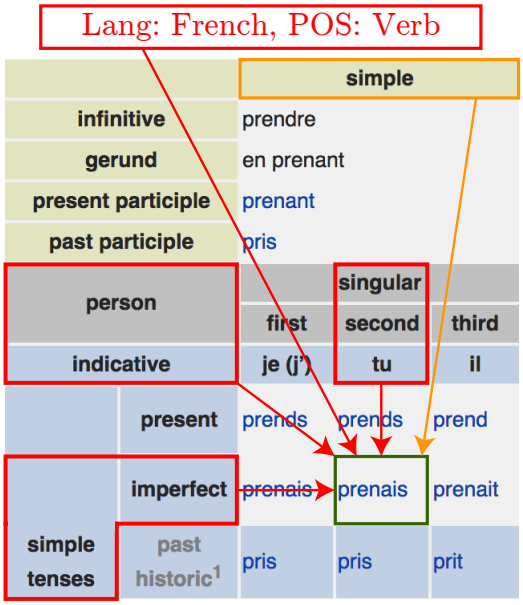


Figure 1: A portion of the English-edition Wiktionary conjugation table for the French verb *prendre* 'take.' The inflected form *prenais* and its row, column, and corner headers are highlighted.

Further, when additional content cells intervene between headers, as they do between **simple** and **singular**, the more distant header is marked as "distal." This labeling is important for proper handling of the column header **simple** in this exam-

ple: It only applies to the top half of the table, and should be left out of any labeling of the inflected forms in the lower half. This distance information, and a hierarchy of positional precedence, is used in Section 3.4 to discount these and other potentially irrelevant descriptors in the case of conflicts during the subsequent mapping of descriptors to features in the universal schema. In general, the positionally highest ranking header value for each schema dimension are utilized and lower-ranking conflicting values are discarded.

### 3.2 Extraction from Parenthetical Lists

For some languages, inflected forms are presented inline next to the headword, instead of in a separate table, as shown for the German noun *Haus* 'house':

**Haus** *n* (*genitive* **Hauses**, *plural* **Häuser**, *diminutive* **Häuschen** *n* or **Häuslein** *n*)

Here, the italic *n* indicates a neuter noun. The inflection data inside the parentheses are extracted as simple tuples containing the lemma, inflected form, and inflectional relationship (e.g. **Haus**, **Häuser**, *plural*).

### 3.3 Improving Extraction Accuracy

The approach described above is sufficient to parse most Wiktionary data, but a large percentage of Wiktionary inflection tables do not use the correct tags to distinguish between header and content cells, an important component of the parsing procedure. In particular, table authors frequently use only the content cell tag to mark up all of a table's cells, and create "soft" headers with a distinct visual appearance by changing their styling (as with Czech verbs, such as *spadat* 'to be included, fall off'). This is indistinguishable to human viewers, but a naïve parse mistakes the soft headers for inflected forms with no descriptors. Hence we investigated several methods for robustly identifying improperly marked-up table headers and overriding the HTML cell-type tags in a preprocessing step.

*Visual identification.* Since most of the soft headers on Wiktionary have a distinct background color from the rest of their containing tables, we initially added a rule that treated content cells that defined a background color in HTML or inline CSS as header cells. However, the mere presence of this attribute was not a reliable indicator since some tables, such as those for Latin nouns (e.g. *aqua* 'water'), gave every cell a background color. This caused them to be erroneously considered to consist entirely of headers, resulting in missing data. Other tables used background color for highlighting, as with Faroese nouns (e.g. *vatn* 'water') and the **past historic** row in Figure 1, whose inflected forms were considered to be headers. For these reasons, visual cues were assessed as an unreliable method of identification.

*Frequency-based methods.* Another, more successful strategy for header discrimination header discrimination utilized the frequency characteristics of cell text, regardless of the cell's type. Although Wiktionary's inflection tables have many different layouts, words with the same language and part of speech pair often share a single template with consistent descriptors. In addition, many simple descriptors, such as **singular**, occur frequently throughout a single edition. Each inflected form, however, can be expected to appear on only a few pages (and in most cases just one). We exploited this tendency by counting the number of pages where each distinct cell text in a Wiktionary edition appeared, and, for each language, manually determined a cutoff point above which any cell with matching text was considered a header. Cells containing only punctuation were excluded from consideration, to avoid problems with dashes that occurred in many tables as a content cell indicating that no such form existed. This strategy surmounted all the problems identified thus far, including both the improper tagging of headers as content cells and the overspecification of background colors.

### 3.4 Mapping Inflected Forms to Universal Features

Using the results of the frequency-based preprocessing step to the table parsing algorithm, the first two authors manually inspected the list of parsed cells and their frequencies within each language, and then determined both a threshold for inclusion as a header feature (descriptor) and a universal representation for each header feature. When possible header features were above the threshold, but judged not to be contentful, they were not given a universal schema representation.

All inflected forms found by our scrape of Wiktionary were assigned complete universal representation vectors by looking up each of their Wiktionary descriptors using the mapping described in the above paragraph and then concatenating the results. Any conflicts within a dimension were resolved using a positional heuristic that favored de-

scriptors nearer to the inflected form in its original HTML table, with column headings assigned higher precedence than row headings, which had higher precedence to corner headings, based on an empirical assessment of positional accuracy in case of conflict.

Ultimately, the process of extraction and mapping yielded instantiated paradigms for 883,965 unique lemmas across 352 languages (of which 130 had more than 100 lemmas), with each inflected form of the lemma described by a vector of features from the universal morphological feature schema.

## 4 Seeding Morphological Analyzers

To test the accuracy, consistency, and utility of our Wiktionary extraction and feature mappings, the fully mapped data from the English edition of Wiktionary were used as input to Durrett and DeNero's (2013) morphological paradigm learner. While the results were comparable to those obtained by the hand-tooled and language-specific table parsers of Durrett and DeNero (2013) given an equivalent quantity of training data, the number of language and part of speech combinations which could be subjected to analysis using data from our general-purpose Wiktionary parser and mapping to features in the universal schema was far greater: 123 language-POS pairs (88 distinct languages) versus Durrett and DeNero's 5 pairs (3 languages).[2] In addition, when the available training data were increased from 500 lemmas to the full amount (a number that varied per language but was always $> 2000$), $\chi^2$ tests demonstrated that the gain in wordform generation accuracy was statistically significant ($p < 0.05$) for 44% (14/32) of the tested language-POS pairs. In the language-POS pairs without significant gains, wordforms were predictable using smaller amounts of data. For example, nearly half (8/18) of the language-POS pairs in this category were nouns in Romance languages, whose pluralization patterns typically involve simply adding /-s/ or some similar variant. Some of the language-POS pairs with significant gains contained multiple inflection classes and/or morpheme altering processes such as vowel harmony, umlaut, or vowel shortening. These linguistic characteristics introduce complexity that reduces the number of exemplars of any given morpheme form, which increases the value of additional data. Figure 2 shows the influence of additional training data on paradigm and wordform generation accuracy for the four languages in which the addition of the full amount of training data provided the most significant improvement (all $p < 0.001$).

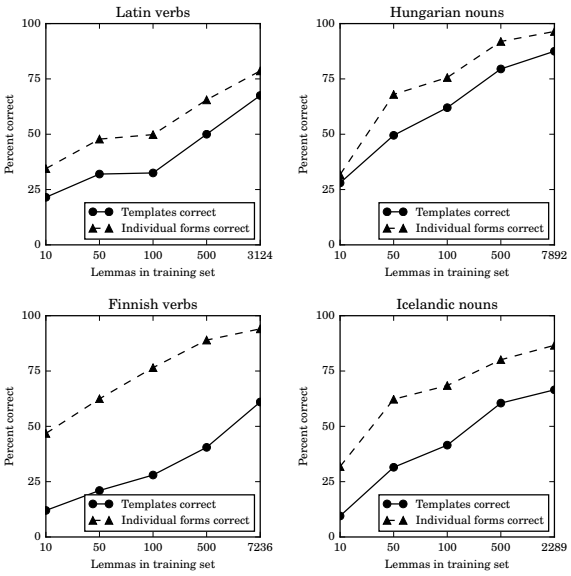


Figure 2: Examples of significant improvements in per-lemma paradigm and wordform generation accuracy with varying amounts of training data

## 5 Conclusion

The proposed universal morphological feature schema incorporates findings from research in linguistic typology to provide a cross-linguistically applicable method of labeling inflectional morphemes according to their meaning. The schema offers many potential benefits for NLP and machine translation by facilitating direct meaning-to-meaning comparison and translation across language pairs. We have also developed original, robust and general multidimensional table parsing and feature mapping algorithms. We then applied these algorithms and universal schema to Wiktionary to generate a significant sharable resource, namely standardized universal feature representations for inflected wordforms from 883,965 instantiated paradigms across 352 languages. We have shown that these data can be used to successfully train morphological analysis tools, and that the increased amount of data available can significantly improve their accuracy.

[2] Language-POS pairs were considered to be suitable for analysis if they possessed 200 or more lemmas that exhibited the maximal paradigm possible.

## References


Alexandra Y. Aikhenvald. 2004. *Evidentiality*. Oxford University Press, Oxford.

D. N. Shankara Bhat. 2004. *Pronouns*. Oxford University Press, Oxford.

Balthasar Bickel and Johanna Nichols. 2005. Inclusive-exclusive as person vs. number categories worldwide. In Elena Filimonova, editor, *Clusivity*, pages 49–72. John Benjamins, Philadelphia.

Barry J. Blake. 2001. *Case*. Cambridge University Press, Cambridge, UK, 2nd edition.

Heather Bliss and Elizabeth Ritter. 2001. Developing a database of personal and demonstrative pronoun paradigms: Conceptual and technical challenges. In Steven Bird, Peter Buneman, and Mark Lieberman, editors, *Proceedings of the ICRS Workshop on Linguistic Databases*. Institute for Research in Cognitive Science, Philadelphia.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Studies in Interactional Sociolinguistics. Cambridge University Press, Cambridge, UK.

Seth Cable. 2008. Tense, aspect and Aktionsart. Unpublished handout from "Proseminar on Semantic Theory" for *Theoretical Perspectives on Languages of the Pacific Northwest*. Available at: http://people.umass.edu/scable/PNWSeminar/handouts/Tense/Tense-Background.pdf, Fall.

Shobhana L. Chelliah and Willem J. de Reuse. 2011. *Handbook of Descriptive Linguistic Fieldwork*. Springer, Dordrecht, Netherlands.

Jinho Choi, Marie-Catherine de Marneffe, Tim Dozat, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Joakim Nivre, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Dan Zeman. 2015. Universal Dependencies. Accessible at: http://universaldependencies.github.io/docs/, January.

Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2008. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. http://www.eva.mpg.de/lingua/resources/glossing-rules.php, February.

Bernard Comrie. 1976a. *Aspect: An Introduction to the Study of Verbal Aspect and Related Problems*. Cambridge University Press, Cambridge, UK.

Bernard Comrie. 1976b. Linguistic politeness axes: Speaker-addressee, speaker-referent, speaker-bystander. *Pragmatics Microfiche*, 1.7(A3). Department of Linguistics, University of Cambridge.

Bernard Comrie. 1989. *Language Universals and Linguistic Typology*. Basil Blackwell, Oxford, 2nd edition.

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

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

William Croft. 2000. Parts of speech as language universals and as language-particular categories. In Petra M. Vogel and Bernard Comrie, editors, *Approaches to the Typology of Word Classes*, pages 65–102. Mouton de Gruyter, New York.

Pierluigi Cuzzolin and Christian Lehmann. 2004. Comparison and gradation. In Geert Booij, Christian Lehmann, Joachim Mugdan, and Stavros Skopeteas, editors, *Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung / An International Handbook on Inflection and Word-Formation*, volume 2, pages 1212–1220. Mouton de Gruyter, Berlin.

Katherine Demuth. 2000. Bantu noun classes: Loanword and acquisition evidence of semantic productivity. In G. Senft, editor, *Classification Systems*, pages 270–292. Cambridge University Press, Cambridge, UK.

Markus Dreyer and Jason Eisner. 2011. Discovering morphological paradigms from plain text using a Dirichlet process mixture model. In *Proceedings of EMNLP 2011*, pages 616–627, Edinburgh. Association for Computational Linguistics.

Greg Durrett and John DeNero. 2013. Supervised learning of complete morphological paradigms. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1185–1195. Association for Computational Linguistics, Atlanta.

John Goldsmith. 2001. Unsupervised learning of the morphology of natural language. *Computational Linguistics*, 27(2):153–198.

Harald Hammarström. 2006. A naive theory of morphology and an algorithm for extraction. In Richard Wicentowski and Grzegorz Kondrak, editors, *SIGPHON 2006: Proceedings of the 8th Meeting of the ACL Special Interest Group on Computational Phonology*, pages 79–88, New York. Association for Computational Linguistics.

Martin Haspelmath. 1995. The converb as a cross-linguistically valid category. In Martin Haspelmath and Ekkehard König, editors, *Converbs in Cross-Linguistic Perspective: Structure and Meaning of Adverbial Verb Forms – Adverbial Participles, Gerunds*, Empirical Approaches to Language Typology, pages 1–56. Mouton de Gruyter, Berlin.

Martin Haspelmath. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, 86(3):663–687, September.

M. H. Klaiman. 1991. *Grammatical Voice*. Cambridge University Press, Cambridge, UK.

Wolfgang Klein. 1994. *Time in Language*. Routledge, New York.

Knud Lambrecht. 1994. *Information Structure and Sentence Form: Topic, Focus and the Mental Representations of Discourse Referents*. Cambridge University Press, Cambridge, UK.

Christopher Lyons. 1999. *Definiteness*. Cambridge University Press, Cambridge.

Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štepánek, Zdeňka Urešová, Kateřina Veselá, and Zdeněk Žabokrtský. 2006. Annotation on the tectogrammatical level in the Prague Dependency Treebank: Annotation manual. Technical report, ÚFAL/CKL, Prague. Technical Report TR-2006-30.

Frank R. Palmer. 2001. *Mood and Modality*. Cambridge University Press, Cambridge, UK, 2nd edition.

Nina V. Radkevich. 2010. *On Location: The Structure of Case and Adpositions*. Ph.D. thesis, University of Connecticut, Storrs, CT.

Benoît Sagot and Géraldine Walther. 2013. Implementing a formal model of inflectional morphology. In Cerstin Mahlow and Michael Piotrowski, editors, *Systems and Frameworks for Computational Morphology*, pages 115–134. Springer, Berlin.

Lesley Stirling. 1993. *Switch-Reference and Discourse Representation*. Cambridge Studies in Linguistics. Cambridge University Press, Cambridge, UK.

John Sylak-Glassman, Christo Kirov, Matt Post, Roger Que, and David Yarowsky. To appear. A universal feature schema for rich morphological annotation and fine-grained cross-lingual part-of-speech tagging. In *Proceedings of the Fourth International Workshop on Systems and Frameworks for Computational Morphology*, Communications in Computer and Information Science. Springer-Verlag, Berlin.

Zeno Vendler. 1957. Verbs and times. *The Philosophical Review*, 66(2):143–160, April.

Mutsumi Yamamoto. 1999. *Animacy and Reference*. John Benjamins, Amsterdam.

Daniel Zeman. 2008. Reusable tagset conversion using tagset drivers. In *Proceedings of LREC 2008*, pages 213–218.