The LinGO Grammar Matrix Customization System

Antske Fokkens

Department of Computational Linguistics
Saarland University

03 November 2009

Outline

Acknowledgments

- This talk represents joint work with:

  Emily M. Bender, Scott Drellishak, Michael Goodman, Safiyyah Saleem and Laurie Poulson
- This material is based upon work supported by the National Science Foundation under Grant No. 0644097. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Outline

The Matrix Customization System

The LinGO Matrix Customization System is a tool that provides start-up implementations for linguistically motivated precision grammars

- From an engineering point of view
  - it supports code-sharing leading to a significant reduction of effort in grammar engineering, and more consistency across grammars
- From a scientific point of view
  - it supports syntactic research for hypothesis testing
  - it encourages research that combines typological research with formal syntactic analysis

“Deep” or “Precision” Grammars

- Deep grammars: parsing leads to, and generation comes from a semantic representation
- Precision grammars: they are linguistically based and aim at getting (only) right analyses
- They are constraint based, resulting in relatively low ambiguity, and less robustness
- Linguistic encoding requires manual effort by an expert: they are expensive to build

Multilingual Grammar Engineering

Main Ideas:

- Reduce the efforts of creating new grammars by using knowledge from those already created
- Create consistency between grammars of different languages
- Research on crosslinguistic similarity

These aims also form the main motivation for the Grammar Matrix Customization Project

Why grammar engineering?

- Broad-coverage precision grammars can be used for implementations
  - These grammars provide more elaborate analyses and are more domain independent than statistically trained parsers
- Hypothesis testing in syntactic research
- Multi-lingual grammar engineering can support typological research

Grammar engineering

- What computer scientists must imagine syntacticians do
  - We say we study rule systems assigning structure to natural language, and mapping between surface forms and semantic representations
- The rule systems are formal and the modeling domain is complex...
  - If we make our analyses machine readable:
  - computers can verify that the systems work as intended
  - and validate against far more data

Grammar engineering for hypothesis testing

Some phenomena that have been tested in Matrix based grammars:

- Morphologically induced tone changes in Hausa (Crysmann 2009)
- Second position auxiliary clusters in Wambaya (Bender 2008b)
- Suspended affixation in Turkish (Fokkens, Poulson and Bender 2009)

Grammar Matrix Context: DELPH-IN

- Delph-in (www.delph-in.net) is a collaboration effort for researchers working on deep linguistic processing.
- The Delph-in member sites contribute open-source software and linguistic resources
- The reference formalism used in Delph-in is based on HPSG (Pollard and Sag 1994) and use MRS (Copestake et al. 2005) as parse output and basis for generation
- (Most) grammars are written in tdl (type description language) — interpreted by LKB (Copestake 2002) and PET (Callmeier 2002)
- [incr tsdb()] (Oepen 2001) for regression testing and treebanking
- Large and medium scale grammars: ERG, JACY, GG, NorSource, Modern Greek, Spanish, French

Some applications using DELPH-IN grammars

- Machine Translation (Oepen et al. 2007)
- Question answering from structured knowledge sources (Frank et al. 2006)
- Robust textual entailment (Bergmair 2008)
- Knowledge extraction from scientific text
- Ontology construction (Nichols et al. 2006)

Grammar Matrix, History

- 2001: First-pass cross-linguistic core grammar (Bender et al 2002)
  - Context: EU Project DeepThought, which included multilingual grammar development
  - Source: English Resource Grammar (Flickinger 2000), with reference to JACY Japanese Grammar (Siegel and Bender 2002)
  - “Bottom up approach to linguistic universals”: Incremental refinement of core grammar as it gets deployed in different languages

The Grammar Matrix

- The core grammar is encoded in a set of files that can be shared by all Matrix grammars
- The files provide basic implementations of types that are inherited by the individual grammars
- Its contributions are: Feature geometry, semantic compositionality, headedness, head-argument and head-modifier constructions; collateral files for software interaction
- 2002-: Used in development of Norwegian (Hellan and Haugereid 2003), Modern Greek (Kordoni and Neu 2005), Spanish (Marimon et al 2007) and Italian grammars

The Matrix Core

- The Core Grammar *matrix.tdl* is meant to be used as the basis of all Matrix Grammars. It provides:
  1. Basic features and devices used in HPSG grammars (e.g. phrase, word, category, lists)
  2. Basic grammar rules (e.g. unary/binary rules, head-subject/head-complement/head-specifier, head-final/head-initial)
  3. Basics for semantics: respects principle of semantic compositionality, supports Minimal Recursion Semantics
  4. Some more advanced features (e.g. simple part of speech inventory, argument extraction, coordination)
  5. Language specific grammars can inherit implementations from *matrix.tdl*

The Matrix Core, Example

Implementation for a language with word order **S**ubject **O**bject **V**erb:

comp-head-rule := basic-head-compl-phrase & head-final.

subj-head-rule := basic-head-subj-rule & head-final &
[SYNSEM.LOCAL.VAL.COMPS < >].

The basic properties of these rules are defined in *matrix.tdl*.

For comparison: the basic-head-comp-phrase

basic-head-comp-phrase := head-nexus-phrase & basic-binary-headed-phrase &
[SYNSEM *phr-synsem-min &*
[LOCAL [CAT [VAL [SUBJ #*subj*,
SPR #*spr*],
POSTHEAD #*ph*,
HC-LIGHT #*light*],
CONT.HOOK #*hook*],
LIGHT #*light*,
NON-LOCAL.SLASH #*slash*]
INFLECTED +,
HEAD-DTR.SYNSEM [local.cat [VAL [SUBJ #*subj*,
SPR #*spr*],
HC-LIGHT #*light*,
POSTHEAD #*ph*]],
NON-LOCAL.SLASH #*slash*
NON-HEAD-DTR.SYNSEM *canonical-synsem &*
[LOCAL.COORD -],
C-CONT [RELS <! !>,
HCONS <! !>,
HOOK #*hook*],
ARGS < [INFLECTED +],
[INFLECTED +] >].

Grammar Matrix: History

- 2004: First annual multilingual grammar engineering course
- Each student works with a different language
- Extend core grammar to different languages, covering:
- Case, agreement, modification, sentential negation, yes-no questions, sentential complements, modals
- Lab instructions outline analyses for known variations

Grammar Matrix: History

- 2005: First pass customization system (Bender and Flickinger 2005)

  Lab instructions were becoming specific enough that a machine could follow them: for some parts only a typological description of a language was necessary

- 2005-2009: Refinements to customization system (Drellishak and Bender 2005, Drellishak 2009)

Matrix Libraries

- The Matrix Libraries provide implementations of grammar fragments of phenomena that vary cross-linguistically (e.g. word order, case)
- A web-based questionnaire elicits typological descriptions, which evoke specific implementations from the Matrix Libraries
- With the libraries, the customization system output grammar fragments: these can be evaluated!

Library development methodology

- Define the phenomenon to be analyzed, e.g.:
  - Dependent marking of core grammatical functions (Drellishak)
  - Morphosyntactically expressed tense/aspect contrasts (Poulson)
- Discover range of variation in phenomenon (from typological literature)
- Read previous syntactic analyses
- Develop and implement analyses for each variant
- Develop questionnaire section to elicit choices along this dimension
- Develop unit tests

Outline

Grammar Matrix Components

- Shared core grammar (matrix.tdl, collateral files): Cross-linguistically useful types and constraints
- Libraries: Analyses of cross-linguistically variable phenomena
- Customization system:
  - Web-based questionnaire to elicit choices among libraries
  - Validation to check that answers were coherent
  - Back-end script to output grammars
  - MatrixTDB: Independently generated test suites

Overview of the Matrix System

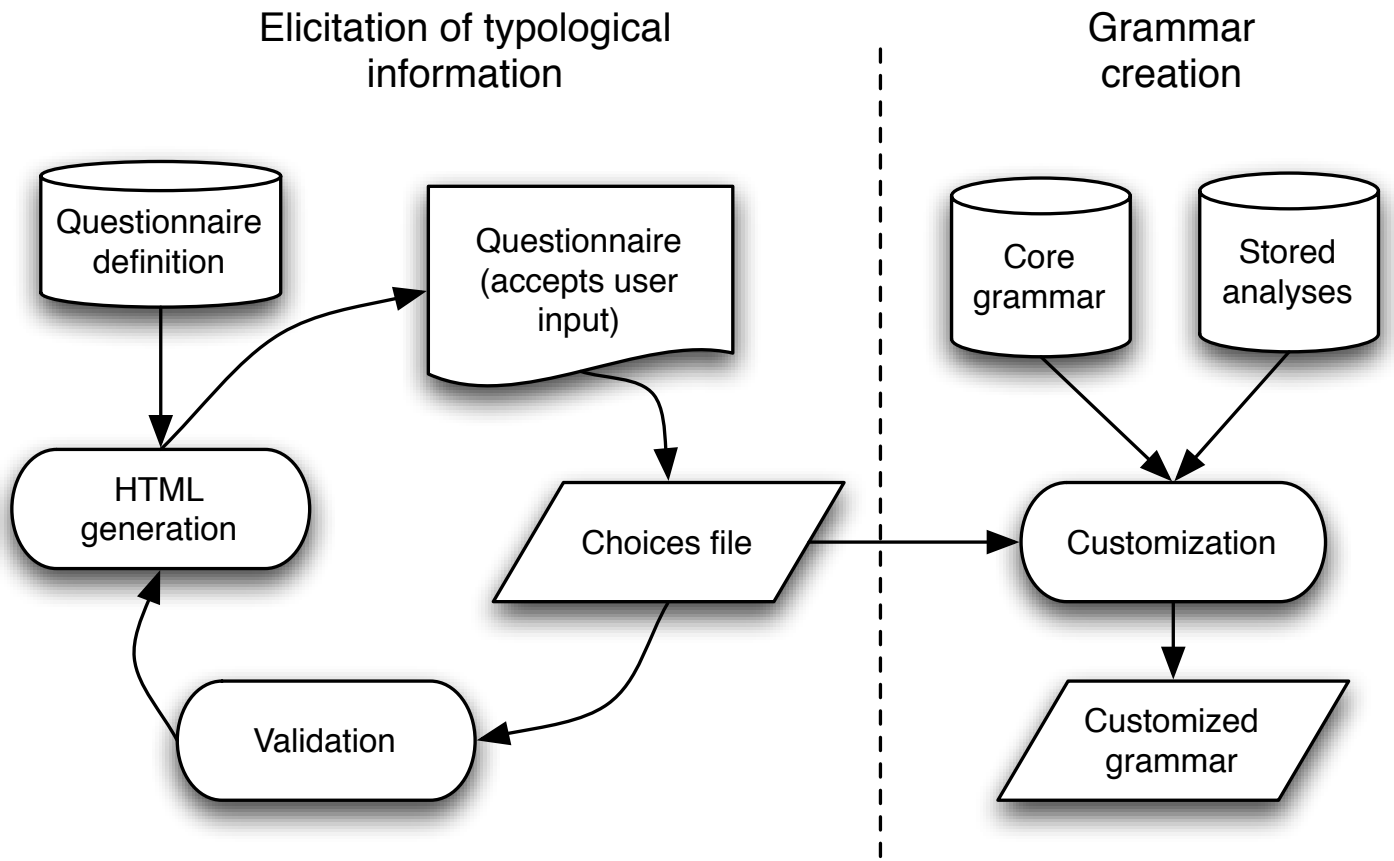

Figure: Schematic system overview

Overview of the Matrix System

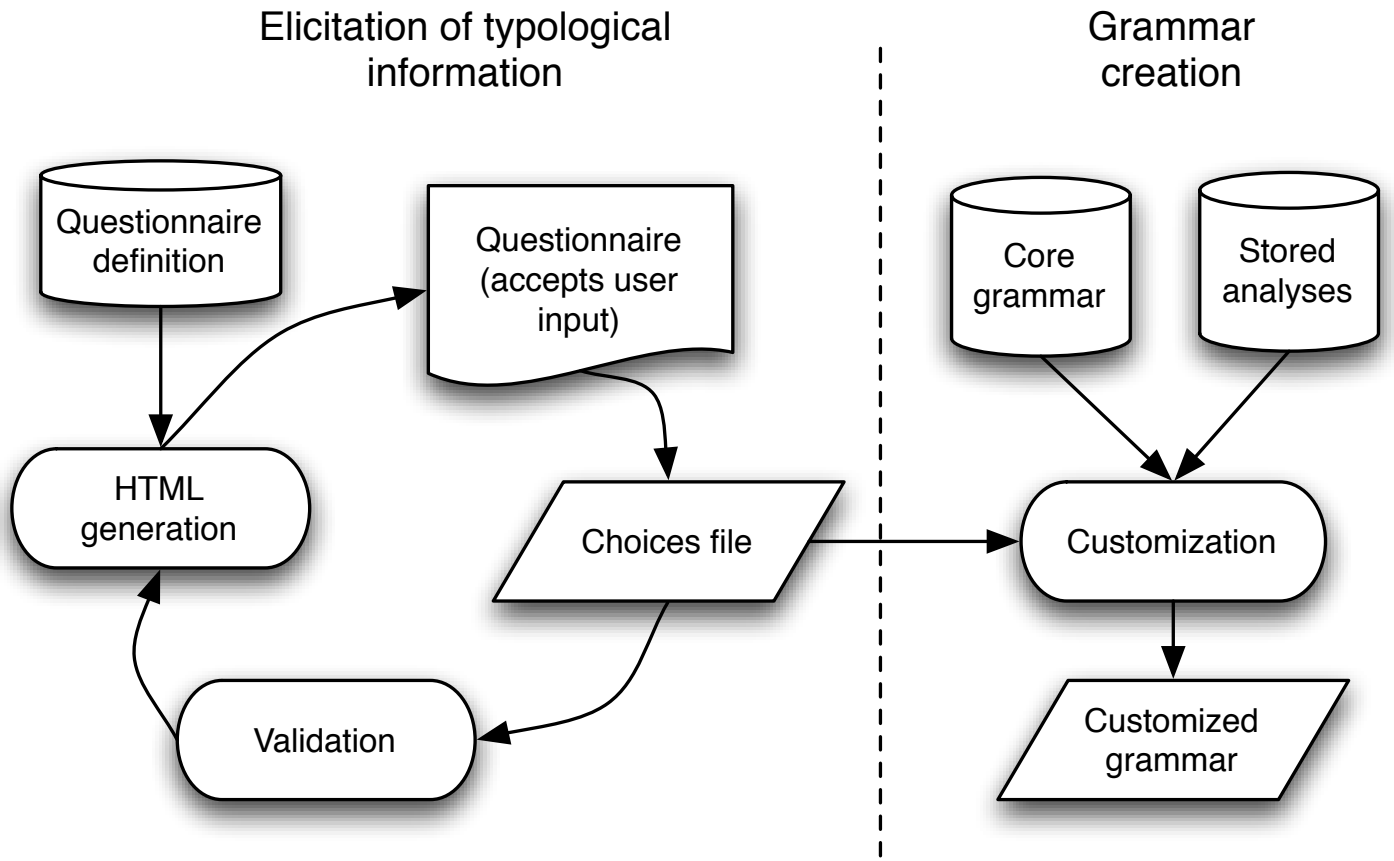

Figure: Schematic system overview

DEMO

Outline

Typology

- The study of:
  - crosslinguistic variation in grammatical structures
  - limits on said variation
  - correlation in different dimensions of variation
- Examples:
  - VO order tends to correlate with P-NP order
  - Morphological simplicity tends to correlate with rigid word order

Formal Syntax

- The study of:
  - well-formedness of sentences
  - combinatorics of words/phrases
  - structures supporting semantic composition
- Grammar engineering:
  - test interaction of analyses
  - test analyses against a broad range of data

Syntax + Typology = ?

- Much typological work focuses on syntactic features
  - but doesn't involve "deep" or "generative" analyses
- Much syntactic works purports to take a typological perspective
  - but doesn't usually consider a broad typological sample
  - typical argumentation predicts a grid or set of language types from the analysis and looks for languages to populate the predicted types

Can Grammar Engineering Help?

- Monolingual grammar engineering lets up scale-up syntactic investigations to much larger grammar fragments and test data sets.
- Can we do the same thing for cross-linguistic investigations?
- If so, what will it tell us about typology?
- About syntax?

Grammar Matrix and Typology

- The Matrix libraries aim at providing analyses that reflect typological diversity of language
- Implementations require exact syntactic analysis
- Analyses have limitations: these can be seen as predictions
- Phenomena (and matrix libraries) interact: more predictions?

Observations from the Implementations

- Some word order phenomena require more complex implementation than others:
  - Simplest word order: all head-complement structures (verb-object, adposition-complement, particle-complement, auxiliary-verb) are in the same order
  - In particular 'disharmonic' order between auxiliary-verb and verb-object may result in a complex analysis.
  - Free word order with more than one freely positioned auxiliary in the phrase cannot be implemented without introducing spurious ambiguity
- Differences in head-subj, head-complement and head-specifier order do not influence the size of the grammar.

Example completely free word order:

John may have been chasing Mary.

- John-subj **may** Mary-obj **have been chasing**.
- **May** John-subj **have** Mary-obj **been chasing**.
- **May** Mary-obj **have** John-subj **been chasing**.
- **May have** Mary-obj **been** John-subj **chasing**.
- **May** Mary-obj **have been chasing** John-subj.
- **May have** John-subj **been** Mary-obj **chasing**.
- ...

Are there natural languages that reveal such behavior?

Auxiliaries in free word order languages:

Working Hypothesis:

Either the verbal group forms a cluster, or the language allows for maximally one auxiliary per clause.

Clustering, grammatical possibilities for preceding auxiliaries:

John-subj **may have been chasing** Mary-obj.

John-subj Mary-obj **may have been chasing**.

Mary-obj **may have been chasing** John-subj.

Mary-obj John-subj **may have been chasing**.

May have been chasing John-subj Mary-obj.

May have been chasing Mary-obj John-subj.

Auxiliary cluster hypothesis

Main question:

How can we test this hypothesis?

→ For now: completely free word order is unattested...

Sahaptin (Drellishak 2009)

- Pacific NW language, varieties spoken in Washingon and Oregon (Rigsby and Rude, 1996)
- Complex argument marking and agreement: case, agreement, direct-inverse all working together
  - Case marking on verbal arguments
  - Argument marking sensitive to a scale, with proximate and obviative third-person nominals.
  - Two loci of agreement (verbal prefix and second-position enclitic), agreement with both the subject and the object
  - Number: sg/du/pl on nominals, but sg/pl in agreement morphology
  - Inclusive/exclusive distinction in person, but only on the second-position enclitic

Sahaptin Grammar

- Filled out the questionnaire for a fragment of Sahaptin
- About 80 hours of work (constructing test sentences, analyzing, filling out questionnaire, and debugging)
- The system can't handle second position enclitics: so described Sahaptin as VSO, with prefixes and enclitics as verb morphology (produces legal word order)

Sahaptin Evaluation

- Created sentences from the intransitive and transitive patterns, and ungrammatical sentences by permutation 89 grammatical, 6076 ungrammatical
- 8 of the ungrammatical sentences actually parsed: they corresponded to unfilled cells in the paradigm from Rigsby and Rude (1996)

Sahaptin Evaluation

- Created sentences from the intransitive and transitive patterns, and ungrammatical sentences by permutation 89 grammatical, 6076 ungrammatical
- 8 of the ungrammatical sentences actually parsed: they corresponded to unfilled cells in the paradigm from Rigsby and Rude (1996)
- **It turned out the customization system got it right**: the implementation lead to a more accurate description of the analysis, and a more straight-forward analysis

Evaluation of the system

- The customization system has been evaluated in two ways:
  - Evaluation on efforts to develop a medium scale grammar with help of the matrix customization system (Bender 2008a)
  - Evaluation on the coverage of the customization system itself (unpublished)

Wambaya Grammar (Bender 2008a)

- Wambaya: Non-Pama-Nguyen language from the Barkly Tablelands region of Australia
- Initial test suite: all 801 examples from Nordlinger 1998
- Initial grammar: Grammar Matrix grammar start
- Goals: Test Grammar Matrix against typologically distant language, measure development time
- Development time: 91% dev set, 76% held-out test set (narrative) in 210 hours (5.5 weeks)
- The original fieldwork and analysis by Nordlinger (1998) is the lion's share of the work ($>$95%)

Customization evaluation

- 7 languages:
  - No Indo-European languages
  - No languages that have been used for the development of the customization system
  - No lanuages that have been used in Emily's grammar engineering class
  - No languages that are closely related to languages that have influenced the system
  - Languages should not be related to each other
  - Different regions of the world should be represented
- Test suites should be designed by someone who is not familiar with the implementations of the customization system

Evaluation Languages

Language name	ISO code	Family	Country of origin	# speakers (year of estimate)
Abkhaz	abk	NW Caucasian	Georgia	117,350 (1993)
Chemehuevi	ute	Uto-Aztecan	USA	1,980 (2000)
Hausa	hau	Afro-Asiatic	Nigeria	25,988,000 (1991)
Jingulu	jig	Australian	Australia	10 (1997)
Malayalam	mal	Dravidian	India	35,893,990 (1997)
Nkore-Kiga	nyn	Niger-Congo	Uganda	2,330,000 (2002)
West Greenlandic	kal	Eskimo-Aleut	Greenland	57,800 (1995)

Location of languages used in Evaluation

Figure: Locations of languages used in evaluation

Test suites sizes

Language	Positive examples	Negative examples	Total
Abkhaz	36	52	88
Chemehuevi	29	27	56
Hausa	38	30	68
Jingulu	29	25	54
Malayalam	39	36	75
Nkore-Kiga	28	52	80
West Greenlandic	33	45	78

Results for preliminary choices files

Language	Coverage	Overgeneration	Ambiguous examples
Abkhaz	72.2%	11.5%	8.3%
Chemehuevi	31.0%	0%	6.9%
Hausa	2.6%	0%	0%
Jingulu	57.1%	7.7%	0%
Malayalam	25.6%	5.6%	0%
Nkore-Kiga	0%	0%	0%
West Greenlandic	6.1%	0%	3.0%

Results for final choices files

Language	Coverage raw	Coverage treebanked	Overgeneration	Spurious ambiguity	Average readings
Abkhaz	100%	94.4%	0%	2.8%	1.08
Chemehuevi	82.8%	75.9%	0%	3.4%	1.04
Hausa	42.1%	36.8%	6.7%	5.3%	1.31
Jingulu	100%	100%	0%	46.7%	2.00
Malayalam	89.7%	87.2%	2.8%	2.8%	1.09
Nkore-Kiga	78.6%	78.6%	11.5%	0%	1.00
West Greenlandic	93.9%	93.9%	0%	0%	1.00

Conclusion

- The LinGO Grammar Matrix customization system is a web-based interface that provides implemented grammar fragments based on typological properties of a language
- From an engineering point of view
  - it supports code-sharing leading to a significant reduction of effort in grammar engineering, and more consistency across grammars
- From a scientific point of view
  - it supports syntactic research for hypothesis testing
  - it encourages research that combines typological research with formal syntactic analysis
- Tutorials on how to fill out the questionnaire, and how to continue your grammar development efforts will be offered

Bibliography I

Bender, E. M. (2008a).

Evaluating a crosslinguistic grammar resource: A case study of Wambaya.
In *Proceedings of ACL08:HLT*.

Bender, E. M. (2008b).

Radical non-configurationality without shuffle operators.
In Müller, S., editor, *Proceedings of HPSG 2008*, Stanford, CA. CSLI Publications ONLINE.

Bender, E. M. and Flickinger, D. (2005).

Rapid prototyping of scalable grammars: Towards modularity in extensions to a language-independent core.
In *Proceedings of the 2nd International Joint Conference on Natural Language Processing IJCNLP-05 (Posters/Demos)*, Jeju Island, Korea.

Bergmair, R. (2008).

Monte Carlo semantics: McPIET at RTE4.
In *Text Analysis Conference (TAC 2008) Workshop-RTE-4 Track. National Institute of Standards and Technology*, pages 17–19.

Callmeier, U. (2002).

Preprocessing and encoding techniques in pet.
In Oepen, S., Flickinger, D., Tsujii, J., and Uszkoreit, H., editors, *Collaborative Language Engineering. A Case Study in Efficient Grammar-based Processing*. CSLI Publications, Stanford, CA.

Copestake, A. (2002).

Implementing Typed Feature Structure Grammars.
CSLI Publications, Stanford, CA.

Bibliography II


Copestake, A., Flickinger, D., Pollard, C., and Sag, I. (2005).
Minimal recursive semantics. an introduction.
Journal of Research on Language and Computation, 3(2–3):281 – 332.


Crysmann, B. (2009).
Autosegmental representations in an HPSG for Hausa.
In *Proceedings of the Workshop on Grammar Engineering Across Frameworks 2009*, Singapore.


Drellishak, S. (2009).
Widespread But Not Universal: Improving the Typological Coverage of the Grammar Matrix.
PhD thesis, University of Washington.


Drellishak, S. and Bender, E. M. (2005).
A coordination module for a crosslinguistic grammar resource.
In Müller, S., editor, *The Proceedings of the 12th International Conference on Head-Driven Phrase Structure Grammar, Department of Informatics, University of Lisbon*, pages 108–128, Stanford. CSLI Publications.


Flickinger, D. (2000).
On building a more efficient grammar by exploiting types.
Natural Language Engineering, 6 (1) (Special Issue on Efficient Processing with HPSG):15 – 28.


Fokkens, A., Poulson, L., and Bender, E. M. (2009).
Inflectional morphology in Turkish VP-coordination.
Paper to be presented at HPSG 2009.

Bibliography III

Frank, A., Krieger, H.-U., Xu, F., Uszkoreit, H., Crysmann, B., Jörg, B., and Schäfer, U. (2006).

Question answering from structured knowledge sources.
Journal of Applied Logic.

Hellan, L. and Haugereid, P. (2003).

NorSource: An exercise in Matrix grammar-building design.
In Bender, E. M., Flickinger, D., Fouvry, F., and Siegel, M., editors, *Proceedings of the Workshop on Ideas and Strategies for Multilingual Grammar Development, ESSLLI 2003*, pages 41–48, Vienna, Austria.

Kordoni, V. and Neu, J. (2005).

Deep analysis of Modern Greek.
In Su, K.-Y., Tsujii, J., and Lee, J.-H., editors, *Lecture Notes in Computer Science*, volume 3248, pages 674–683. Springer-Verlag, Berlin.

Marimon, M., Bel, N., and Seghezzi, N. (2007).

Test-suite construction for a Spanish grammar.
In King, T. H. and Bender, E. M., editors, *Proceedings of the GEAF 2007 Workshop*, Stanford, CA. CSLI Publications.

Nichols, E., Bond, F., Tanaka, T., Fujita, S., and Flickinger, D. (2006).

Multilingual ontology acquisition from multiple mrds.
In *Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 10–17, Sydney, Australia. Association for Computational Linguistics.

Bibliography IV

Nordlinger, R. (1998).

A Grammar of Wambaya, Northern Australia.
Research School of Pacific and Asian Studies, The Australian National University, Canberra.

Oepen, S. (2001).

[incr tsdb()] — competence and performance laboratory.
Technical report, DFKI, Saarbrücken, Germany.

Oepen, S., Velldal, E., LÃÿnning, J. T., Meurer, P., RosÃľn, V., and Flickinger, D. (2007).

Towards hybrid quality-oriented machine translation. On linguistics and probabilities in MT.
SkÃűvde, Sweden.

Pollard, C. and Sag, I. (1994).

Head-Driven Phrase Structure Grammar.
University of Chicago Press, Chicago, USA.

Rigsby, B. and Rude, N. (1996).

Sketch of sahaptin, a sahaptian language.
In Goddard, I., editor, *Languages*, page 666âĂŞ692, Washington DC. Smithsonian Institution.

Rupp, C., Copestake, A., Corbett, P., and Waldron, B. (2007).

Integrating general-purpose and domain-specific components in the analysis of scientific text.
In *Proceedings of the UK e-Science Programme All Hands Meeting*.

Bibliography V

Siegel, M. and Bender, E. M. (2002).
Efficient deep processing of Japanese.
In *Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization at the 19th International Conference on Computational Linguistics*, Taipei, Taiwan.