Final Report on the project *Typology and Theory of the Word*

Submitted November 2009

1 Project details

Project Nr.:	BI 799/2-3
Applicants/P.I.s:	Prof. Dr. B. Bickel (2003-2007) Prof. Dr. T. A. Hall (2003-2005)
Team members:	Kristine Hildebrandt, Ph.D. (2003-2005) Dr. René Schiering (2006-2007) *Student assistants (at various times):* Alena Witzlack-Makarevich, Kathi Stutz, Thomas Goldammer, Josh Wilbur, Franziska Crell, Jenny Seeg, Sven Siegmund, Taras Zakharko, Sabine Günther
Topic:	Typology and Theory of the Word
Institution:	Department of Linguistics, University of Leipzig
Period:	August 1, 2003 – July 31, 2007 (4 years)

2 Publications

2.1 Articles and book chapters

Bickel, B., 2003. Prosodic tautomorphemicity in Sino-Tibetan. In Bradley, D., R. J. LaPolla, B. Michailovsky, & G. Thurgood (eds.) *Variation in Sino-Tibetan and South East Asian languages*, 89 – 99. Canberra: Pacific Linguistics.

Bickel, B., G. Banjade, M. Gaenszle, E. Lieven, N. Paudyal, I. P. Rai, M. Rai, N. K. Rai, & S. Stoll, 2007. Free prefix ordering in Chintang. *Language* 83, 43 – 73.

Bickel, B., K. Hildebrandt, & R. Schiering, 2009. The distribution of phonological word domains: a probabilistic typology. In Grijzenhout, J. & B. Kabak (eds.) *Phonological domains: universals and deviations*, 47–75. Berlin: Mouton de Gruyter.

Bickel, B. & J. Nichols, 2005. Inflectional synthesis of the verb. In Haspelmath, M., M. S. Dryer, D. Gil, & B. Comrie (eds.) *The world atlas of language structures*, 94 – 97. Oxford: Oxford University Press.

Bickel, B. & J. Nichols, 2007. Inflectional morphology. In Shopen, T. (ed.) *Language typology and syntactic description*, 169 – 240. Cambridge: Cambridge University Press (Revised second edition).

Hall, T. A. & K. Hildebrandt, 2008. Phonological and morphological domains in Kyirong Tibetan. *Linguistics* 46, 215–248.

Hall, T. A., K. Hildebrandt, & B. Bickel, 2008. Introduction: theory and typology of the word. *Linguistics* 46, 183 – 192.

Hildebrandt, K., 2007. Prosodic and Grammatical Domains in Limbu. *Himalayan Linguistics Journal* 8, 1 – 34.

Schiering, R., 2006. Morphologization in Turkish: implications for phonology in grammaticalization. In *Proceedings of the 13th International Conference on Turkish Linguistics*. [submitted; `http://www.uni-leipzig.de/~autotyp/projects/wd_dom/ICTL_Schiering.pdf`].

Schiering, R., 2007. The phonological basis of linguistic rhythm: cross-linguistic data and siachronic Interpretation. *Sprachtypologie und Universalienforschung* 60, 337–359.

Schiering, R., 2009. Review of Szczepaniak, Renata (2007). *Der phonologisch-typologische Wandel des Deutschen von einer Silben- zu einer Wortsprache. Linguistic Typology* 13, 463–471.

Schiering, R., in press. Reconsidering erosion in grammaticalization: evidence from cliticization. In König, E., E. Gehweiler, & K. Stathi (eds.) *What's new in grammaticalization?* Amsterdam: Benjamins.

Schiering, R., B. Bickel, & K. Hildebrandt, 2009. Stress-timed = word-based? Testing a hypothesis in Prosodic Typology. Ms. under review [`http://www.uni-leipzig.de/~autotyp/download/Schieringetal2009Stress-timed.pdf`].

Schiering, R., B. Bickel, & K. Hildebrandt, in press. The prosodic word is not universal, but emergent. *Journal of Linguistics* [pre-print available at `http://www.uni-leipzig.de/~autotyp/download/Schieringetal2009The-prosodic.pdf`].

Schiering, R. & H. G. van der Hulst, in press. Word accent systems in the languages of Asia. In Goedemans, R., H. G. van der Hulst, & E. van Zanten (eds.) *Stress patterns of the world, part II: the Data*. Berlin: Mouton de Gruyter.

Note: Hall et al. (2008) and Hall & Hildebrandt (2008) are part of a special issue of *Linguistics* on the phonological word that was guest-edited by Hall, Hildebrandt and Bickel on the basis of a workshop held in Leipzig in April 2004.

2.2 Theses

The following MA theses were completed at the University of Leipzig on topics of the project:

Wilbur, Joshua, 2007. *Syllable Structures and Stress Patterns in Kildin Saami.*
Advisors: B. Bickel, M. Kuzmenko

Voll, Rebecca. 2006. *Prosodic Domains in Richtersveld Nama.*
Advisors: B. Bickel, T. Güldemann

Luo, Yan, 2006. *Serielle Verbkonstruktionen im Xiang-Chinesischen.*
Advisors: B. Bickel, M. Haspelmath

2.3 Language reports

In addition to publications in regular outlets, the project produced reports on individual languages that summarize the major patterns of phonological word domains. These reports are available for download at `http://www.uni-leipzig.de/~autotyp/reports/register.php`:

Language	*Report author*	*Year*
Vietnamese	René Schiering	2007
Mon	René Schiering	2006
Hixkaryana	Kristine Hildebrandt & Kathi Stutz	2005
Methei	Kristine Hildebrandt	2005
Martuthunira	Kristine Hildebrandt & Alena Witzlack-Makarevich	2005

2.4 Database

The project has made available the database on phonological words that was compiled by team members. The database can be downloaded as `csv`-files from `http://www.uni-leipzig.de/~autotyp/available.html#downloadable`. Bibliographical references are available in BibTeX format at the same site.

2.5 Questionnaires

As part of the database compilation, we also developed two questionnaires for fieldworkers, one on phonological and one on grammatical word domains. These questionnaires are available at `http://www.uni-leipzig.de~autotyp/projects/wd_dom/wd_dom_que.html`

3 Report

The project originally targeted problems in the definition of both grammatical and phonological words. Early on it became clear, however, that definitions of phonological words pose many more and more severe challenges than definitions of grammatical words. Most theories offer elaborate machinery to represent and explain mismatches in grammatical wordhood, ranging from fairly standard conceptions of morphological vs. syntactic wordhood to the assumption of a multitude of sub- and supra-terminal projection levels (X^{-n}, X^{0}, X^{n}). We concluded that while the analysis of particular grammatical word domains is always a real descriptive challenge in individual languages, there are no large-scale or fundamental theoretical problems. This includes the analysis of split or bipartite stems, which can be accounted for by prosodic or morphological specification of the domain for inflection (Bickel & Nichols 2007, Hildebrandt 2007, Bickel et al. 2007).

In many cases, phenomena that at first sight look like bipartite stems can be successfully analyzed as regular affixation mechanisms with prosodic domain specifications (see Donohue (2008) for a recent case study). The issue of prosodic domain specification of affixes is also taken up by Bickel et al. (2007), who show that prefixes in Chintang

(Sino-Tibetan, Kiranti) attach to the left edge of phonological words. Since each grammatical word contains multiple phonological words, including one phonological word per prefix, prefixes can appear at various places, leading to the surface phenomenon of free prefix permutation.

The main goal of the project however was an understanding of the range of diversity in how languages define their phonological words. The main research tool we used to achieve this goal was the development and subsequent analysis of a typological database.

3.1 Database structure

The following summarizes the fields contained in the database. There are a total of 881 records from 76 languages.

3.1.1 Phonological pattern

`full_description`: A full description of the phonological pattern in prose, with standard rule formalization
`ppattern1`: A taxonomy of phonological patterns (first level)
`ppattern1new`: An alternative taxonomy of phonological patterns (first level)
`ppattern2`: A taxonomy of phonological patterns (higher level)
`ppattern3`: A taxonomy of phonological patterns (highest level, very abstract)
`stresspattern`: Is this a stress-related pattern?
`Mishear`: Is there evidence that the sound pattern arose from perceptual ambiguity (e.g. an assimilation pattern)?
`unit`: Is this a phonological or a grammatical pattern?
`min_wd`: Is the pattern a minimal word pattern?
`LexTyp`: Extent to which a process is structure-preserving (lexical)
`structure_preserve`: Binary recode of LexTyp
`AlignID1`: Position in the morphosyntactic structure where a phonological domain has effects (edge, spanning all, etc.)
`StratmID`: Exceptional specifications (lexical, category, loanword status, etc.) of phonological pattern

3.1.2 Domain definition

`DomainType`: Formulaic description of domain
`Definition`: Prose definition of the domain
`plevel`: Is the domain phrasal or subphrasal?
`DomMrg`: A simplified taxonomy of domains
`PrefixStatus`: Does the domain include or exclude prefixes (calculated on the basis of the following)

Domains are not defined in terms like 'prefix' or 'enclitic' because these are not directly suitable for cross-linguistic comparison: the exact boundary between affixes, clitics and particle differ strongly across languages and there are many language-particular units that do not fit preconceived universal definitions (Bickel 2009, Luís & Spencer 2004,

Spencer 2006, Bérmudez-Otero & Payne in press). Instead, we define the constituents ('morpheme types') of domains in terms of four independent variables, i.e. quadruples of specific properties:

`Pos`: Phonological position behavior of domain constituent: pre, post, in etc.
`Typ`: Constituent type: stem projecting a regular part of speech category vs. grammatical marker (formative in Bickel & Nichols' 2007 sense)
`Restr`: Selectivity of domain constituent: does the constituent select for a single type of host ('restricted'), or a subset of hosts ('semi-restricted'), or is it unrestricted?
`Stratum`: Lexical specification of morphology targeted by phonological pattern

For analyzing the isomorphisms of domains within single languages, without any direct cross-linguistic comparison of individual domains (as e.g. in Schiering et al. 2009), it is also useful to operate with simplified definitions of domains. For this purpose we defined a notion of 'affix' as a restricted formative and a notion of 'clitic' as an unrestricted or semi-restricted formative. This allows simplified domain definitions in terms of whether or not a phonological pattern applies to a constituent or at the boundaries of two constituents:

`stemsuffix`: Application of phonological pattern to a stem-suffix boundary
`prefixstem`: Application of phonological pattern to a prefix-stem boundary
`prefixprefix`: Application of phonological pattern to a prefix boundary
`suffixsuffix`: Application of phonological pattern to a suffix boundary
`procliticstem`: Application of phonological pattern to a proclitic-stem boundary
`stemenclitic`: Application of phonological pattern to a stem-enclitic boundary
`procliticprefix`: Application of phonological pattern to a proclitic-prefix boundary
`suffixenclitic`: Application of phonological pattern to a suffix-enclitic boundary
`procliticproclitic`: Application of phonological pattern to a proclitic boundary
`encliticenclitic`: Application of phonological pattern to an enclitic boundary
`stem`: Application of phonological pattern to a stem alone (i.e. within the stem)
`prefix`: Application of phonological pattern to a prefix alone (i.e. wthin the prefix)
`suffix`: Application of phonological pattern to a suffix alone (i.e. within the suffix)
`stemendoclitic`: Application of phonological pattern to a stem-endoclitic boundary
`suffixendoclitic`: Application of phonological pattern to a suffix-endoclitic boundary
`enclitic`: Application of phonological pattern to an enclitic alone (i.e. within the enclitic)
`steminfix`: Application of phonological pattern to a stem-infix boundary
`proclitic`: Application of phonological pattern to a proclitic alone (i.e. within the proclitic)
`prefixendoclitic`: Application of phonological pattern to a prefix-endoclitic boundary

3.1.3 Pre-coded calculations in the database

`Size:` Total size of the domain (in number of morpheme types included, as defined in 3.1.2)
`Size_pf:` Total size of prefixal part of domain
`Size_sf:` Total size of suffixal part of domain

All other calculations are done on the fly, embedded in the quantitative analyses performed.

3.1.4 Housekeeping fields

`survey_status:` Extent to which an aggregation of patterns and domains can be reasonably done per language (i.e. only when all domains and patterns are sufficiently well surveyed)
`Reliability:` Reliability of source and data provenance
`Notes:` Notes
`Example:` Examples

3.1.5 Available morpheme types per language

In addition to this database we also tracked the total number of morpheme types that are available in a language. This makes it possible to compute the size of phonological word domains relative to what is logically possible in each language (the degree of 'coherence' as we call it in Bickel et al. 2009). The definition of morpheme types is done in terms of the same set of four variables listed in 3.1.2 above. The relevant database covers again 76 languages.

3.2 Analyses

The main insight resulting from the data collection in the database is that most languages define more than one domain between the foot and the phrase, i.e. more than one 'word domain', thereby violating foundational claims of the Prosodic Hierarchy theory (as recently defended e.g. by Vogel 2009). In Hildebrandt (2007) and Schiering et al. (in press) we demonstrate this through an in-depth study of Limbu (Sino-Tibetan, Kiranti). Schiering et al. (in press) show in addition that there are also languages that completely lack a level of the phonological word, the example studied in the paper being Vietnamese.

There are two possible responses to such a finding: one is to revise phonological theory in a such a way that languages with multiple word domains and languages lacking word domains can be analyzed and described without violating simultaneously held claims of the theory. However, while working on the 76 languages that eventually went into our database, it became clear that our empirical knowledge of the true diversity in phonological domains is still far too narrow and developing an alternative to the Prosodic Hierarchy theory therefore seems premature. (In general, our impression is that linguistic theory development has suffered in the past from a lack of empirical groundwork and we do not wish to repeat this mistake in our own work). Instead, we chose an alternative,

typological approach. This approach consists of refining the set of variables that is descriptively needed in order to capture all known domains. Such a set of variables can be seen as a theoretical framework of analysis in its own right, but it differs from theories in a more narrow sense in that it is left open-ended and does not attempt to delimit what is a possible human language (Bickel 2007). Its virtue lies elsewhere, namely in making it possible to perform quantitative analyses of the observed distribution of each variable and their interactions. Such analyses rely on the same methodology as most other sciences, combining exploratory data-mining techniques and the statistical testing of hypotheses grounded in explanatory (often functional) theories (cf. Newmeyer 2005).

To date, we have completed two such analyses: (i) Bickel et al. (2009) use Multidimensional Scaling (Cox 1994) in order to detect probabilistic clusters of word domains across languages. This led to the discovery of a possible statistical universal stating that domains based on stress patterns are always larger than domains based on any other phonological patterns. Such a universal can be theoretically grounded in pregenerative views of phonological structure, such as Pike's (1945), who strictly separates prosody in a narrow sense from morphophonological structures. It is also in line with a view of tone as patterning more with standard segmental rather than with stress phonology. Empirically, the hypotheses receives strong empirical support in an areally and genealogically stratified sample of 40 Sino-Tibetan, Indo-European, and Austro-Asiatic languages. (ii) Schiering et al. (2009) elaborate on the traditional typology of 'stress-timed' vs. 'syllable-timed' languages (critically reviewed in Schiering 2007), which predicts, among other things, that in 'stress-timed' languages, the phonological word domain is more salient than in other languages. We test this hypothesis by explicating the notion of domain saliency in terms of the number of distinct phonological patterns that identify the most frequently referenced domain in the language. The quantitative analysis of our database, however, reveals only a weak, statistically non-significant trend in direction of the hypothesis.

Both analyses also suggest that the major determinant of the kind of phonological word domains found in a language is the genealogical family to which the language belongs. In other words, phonological word domains seem be inherited over time fairly faithfully, admitting only limited diversification and only weak areal diffusion (thus challenging the extent of 'prosodic diffusibility' expected in such areas as Southeast Asia: Matisoff 2001). We explore this hypothesis in an in-depth case study of Austro-Asiatic languages, first presented by Schiering and Bickel at a Workshop on Austro-Asiatic in Leipzig in April 2007 and planned to be submitted for publication in a journal shortly. Another case study on Sino-Tibetan was presented by Hildebrandt at the 2009 Manchester Phonology Meeting Further and will be ready for submission soom. Further research on the diachronic stability of word domains is in progress.

The quantitative analyses all rely on detailed research on the languages in the database. This research gave ample opportunity for training students (either as student assistants or seminar participants). In two cases, this lead to M.A. theses (cf. Section 2.2 above) and the publication of jointly-written language reports (Section 2.3). The language-specific research also involved detailed research on some languages that

went far beyond the needs of the database per se. Apart from Limbu and Chintang mentioned earlier, this concerned most of all work on the Kyirong dialect of Tibetan. In Hall & Hildebrandt (2008) we demonstrate that Kyirong is typologically unusual in that (i) most suffixes are prosodically incoherent, i.e. impose a phonological word boundary between stem and suffix, whereas most in most languages phonological incoherence affects prefixes and not suffixes (Bickel et al. 2007, among others), and (ii) most compounds are integrated into single words and do not constitute two separate phonological words, which is probably the default option cross-linguistically.

Additional references (not listed in Section 2.1):

Bickel, B., 2007. Typology in the 21st century: major current developments. *Linguistic Typology* 11, 239 – 251.

Bickel, B., 2009. Typological patterns and hidden diversity. Plenary talk at the 8th Bi-Annual Meeting of the Asssociation for Linguistic Typology, July 24 [`http://www.uni-leipzig.de/~bickel/research/presentations/alt2009bickel-plenary.pdf`].

Bérmudez-Otero, R. & J. R. Payne, in press. There are no special clitics. In Galani, A., G. Hicks, & G. Tsoulas (eds.) *Morphology and its interfaces*. Amsterdam: Benjamins [preprint available at `http://www.bermudez-otero.com/bermudez-otero&payne.pdf`].

Cox, T. F., 1994. *Multidimensional scaling*. New York: Chapman & Hall.

Donohue, M., 2008. Complex predicates and bipartite stems in Skou. *Studies in Language* 32, 279–335.

Luís, A. & A. Spencer, 2004. A paradigm function account of 'mesoclisis' in European Portuguese. *Yearbook of Morphology* 2004, 177–228.

Matisoff, J. A., 2001. Genetic vs. areal linguistics in Southeast Asia: prosodic diffusibility in Southeast Asian languages. In Aikhenvald, A. Y. & R. M. Dixon (eds.) *Areal diffusion and genetic inheritance*, 291 – 327. Cambridge: Cambridge University Press.

Newmeyer, F. J., 2005. *Possible and probable languages: a generative perspective on linguistic typology*. New York: Oxford University Press.

Pike, K. L., 1945. *The intonation of American English*. Ann Arbor: University of Michigan Press.

Spencer, A., 2006. Morphological universals. In Mairal, R. & J. Gil (eds.) *Lingustic universals*, 101–129. Cambridge University Press.

Vogel, I., 2009. Universals of prosodic structure. In Scalise, S., E. Magni, & A. Bisetto (eds.) *Universals of language today*, 59–82. Berlin: Springer.

4 Summary

This project surveyed the structure of phonological word domains in more than 70 languages, including in-depth case studies of several Sino-Tibetan and Austro-Asiatic languages. The main discovery of the project is that most languages violate the traditional expectation that languages have exactly one phonological domain larger than the foot and smaller than the phrase, i.e. exactly one type of 'phonological word'. Most languages have more than one such domain; some languages (like Vietnamese) lack evidence for any such domain.

Our survey work led to the development of a typological database (available for download at `www.uni-leipzig.de/~autotyp`). The attested diversity in the database is substantial, suggesting that the actual diversity beyond our sample is even larger. Because of this, the project did not spend energy on developing a new theory constraining the set of possible language types but instead focused on (a) the exploration of probabilistic trends in an areally and genealogically stratified sample of 40 Sino-Tibetan, Indo-European and Austro-Asiatic languages and (b) detailed case studies of typologically rare patterns.

The quantitative analyses suggest that structures in phonological word domains are remarkably stable within families over time and do not spread easily between languages. We also find evidence that universally, word domains tend to be larger for stress patterns than for any other pattern (including tone) and a weak but statistically non-significant trend for phonological patterns to converge on unified phonological word domains to a larger extent in what are traditionally called 'stress-timed' than in 'syllable-timed' languages.

The case studies reveal typologically unusual languages with noncohering suffixes (instead of prefixes: Kyirong Tibetan), languages with prefixes that can freely attach to variable phonological word edges (Chintang), languages with a multitude of non-isomorphic phonological word domains (Limbu), and languages with no evidence for the existence of phonological words (Vietnamese).