University of California, Santa Barbara

Lexical flexibility in discourse:

A quantitative corpus-based approach

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Linguistics

by

Daniel W. Hieber

Committee in Charge:

Professor Marianne Mithun, Chair

Professor Benard Comrie

Professor Stefan Th. Gries

Professor William Croft (University of New Mexico)

December 2020

The dissertation of Daniel W. Hieber is approved.

Bernard Comrie

Stefan Th. Gries

William Croft

Marianne Mithun, Committee Chair

December 2020

Lexical flexibility in discourse:

A quantitative corpus-based approach

by

Daniel W. Hieber

Typeset using LaTeX software and the Linux Libertine family of fonts.

This thesis may be downloaded at:
https://files.danielhieber.com/publications/dissertation.pdf

The source code, data, and accompanying scripts for this thesis are available on GitHub:
https://github.com/dwhieb/dissertation

Dedication

Acknowledgments

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Curriculum Vitae

Daniel W. Hieber

Education

June 2020	Ph.D. in Linguistics, University of California, Santa Barbara
March 2016	M.A. in Linguistics, University of California, Santa Barbara
June 2008	B.A. in Linguistics & Philosophy, The College of William & Mary (magna cum laude)

Professional Experience

2018–2019	Editor, Custom Language Products, Rosetta Stone
2015–2017	Teaching Assistant, Department of Linguistics, University of California, Santa Barbara
2014–2015	Research Assistant (under Prof. Carol Genetti), Department of Linguistics, University of California, Santa Barbara
2011–2013	Associate Researcher, Research Labs, Rosetta Stone
2008–2011	Editor, Endangered Languages Program, Rosetta Stone
2007–2008	Intern, Endangered Languages Program, Rosetta Stone
2006	Spanish Instructor, Nielsen Builders
2004–2006	Lab Assistant, Language Lab, The College of William & Mary
2003–2004	Latin Instructor, Bridgewater Home School Unit

Publications

2019	The Chitimacha language: A history. In Nathalie Dajko & Shana Walton (eds.), *Languages in Louisiana: Community & culture* (America's Third Coast Series). University Press of Mississippi.
2019	Semantic alignment in Chitimacha. *International Journal of American Linguistics* 85(3): 313–363. DOI:10.1086/703239
2018	Category genesis in Chitimacha: A constructional approach. In Kristel Van Goethem, Muriel Norde, Evie Coussé, & Gudrun Vanderbauwhede (eds.), *Category change from a constructional perspective* (Constructional Approaches to Language 20), 15–46. John Benjamins. DOI:10.1075/cal.20.02hie
2016	*The cohesive function of prosody in Ékegusií (Kisii) narratives: A functional-typological approach.* M.A. thesis, University of California, Santa Barbara.

2013 On linguistics, language, and our times: A linguist's narrative reviewed. *Linguistic Typology* 17(2): 291–321. Review article of *I am a linguist* by R. M. W. Dixon (Brill, 2010). DOI:10.13140/RG.2.2.13238.96329

2013 (with Sharon Hargus & Edward Vajda, eds.) *Working papers in Athabaskan (Dene) languages 2012.* Alaska Native Language Center Working Papers 11. ANLC.

Awards

2019 SSILA Best Student Presentation Award

2015 National Science Foundation (NSF) Graduate Student Research Fellowship (GRFP)

2015 2nd place, University of California Grad Slam

2015 Winner, University of California, Santa Barbara Grad Slam

2013 Chancellor's Fellowship, University of California, Santa Barbara

2006 Boren Scholarship, National Security Education Program (NSEP)

Fields of Study

Major Fields: Linguistic Typology, Language Documentation & Description, Language Revitalization, Prosody, Discourse, Language Change, Language Contact, Digital Linguistics, Corpus Linguistics

Linguistic Typology with Professor Bernard Comrie & Professor Marianne Mithun

Language Documentation & Description with Professor Eric Campbell, Professor Carol Genetti, & Professor Marianne Mithun

Language Revitalization with Professor Carol Genetti

Prosody with Professor Carol Genetti, Professor Matthew Gordon, & Professor Marianne Mithun

Discourse with Professor Patricia Clancy, Professor John W. DuBois, Professor Carol Genetti, & Professor Marianne Mithun

Language Change with Professor Marianne Mithun

Language Contact with Professor Marianne Mithun

Digital Linguistics with Professor Eric Campbell & Professor Stefan Th. Gries

Corpus Linguistics with Professor Eric Campbell, Professor John W. DuBois, & Professor Stefan Th. Gries

Abstract

Lexical flexibility in discourse:

A quantitative corpus-based approach

by

Daniel W. Hieber

This thesis is a quantitative corpus-based study of lexical flexibility in English (Indo-European) and Nuuchahnulth (Wakashan). *Lexical flexibility* is the capacity of lexical items to serve in more than one discourse function—reference, predication, or modification (or more traditionally, noun, verb, or adjective). In this thesis I develop a procedure and metric for quantifying the lexical flexibility of words in a corpus, and apply that metric to English and Nuuchahnulth. I find that the two languages differ drastically in not only their degree of lexical flexibility, but the way in which that flexibility is realized. This study advances the discussion of lexical flexibility—as well as parts of speech more generally—by adding a new kind of empirical evidence to the discussion (quantitative corpus-based data), and in doing so provides answers to several longstanding and much-debated questions about how lexical categories operate in English and Nuuchahnulth.

The abstract should include 1) a brief statement of the problem; 2) a description of the methods and procedures used to gather data or study the problem; 3) a condensed summary of the findings. The abstract should be double-spaced. The recommended length is 1–2 pages. (add Abstract)

Table of Contents

List of Figures

List of Tables

List of Abbreviations

The following table provides the meaning of each abbreviation used in interlinear glossed examples throughout this thesis.

1	first person
2	second person
3	third person
SUBJ	subject

Conventions

This note documents the conventions I have adopted regarding linguistic data, terminology, and presentation of data throughout this thesis.

It is well known that the world's languages realize widely different sets of morphosyntactic categories (Whaley 1997: 58; Haspelmath 2007). Moreover, even when these categories bear the same name, they may differ drastically in their behavior (Dixon 2010: 9). It is the subject of much debate whether these language-specific categories can be mapped onto each other or compared in any useful way (Croft 1995; Song 2001: 10–15; Croft 2003: 13–19; Haspelmath 2010a,b; Newmeyer 2010; Stassen 2011; Hieber 2013: 308–310; Croft 2014; Plank 2016; Song 2018: 44–58). Recognizing these difficulties, I have made no attempt to standardize the linguistic terminology used in examples from different languages. I have, however, standardized the abbreviations used to refer to those terms. For example, even though one researcher may abbreviate Subject as SUBJ and another researcher abbreviate it as SUB, I nonetheless gloss all Subject morphemes as SUBJ. See the List of Abbreviations (p. xii) for a complete list of glossing abbreviations.

I have not attempted to standardize the transcription systems and orthographies used in examples. All examples are given as transcribed in their original source. The reader should consult those original sources for further details regarding orthography.

The source of each example is always provided after the example itself.

In all interlinear glossed examples, I follow the formatting conventions (but not necessarily the recommended abbreviations) of the Leipzig Glossing Rules (Bickel, Comrie & Haspel-

math 2015).

It is increasingly common in typological studies to write language-particular terms and categories with an initial capital letter, and to write terms that refer to language-general or semantic/functional concepts (e.g. the crosslinguistic notion of subject) in lowercase (Comrie 1976: 10; Bybee 1985: 47 (fn. 3), 141; Croft 2000: 66; Haspelmath 2010a: 674; Croft 2014: 535). For example, the English Participle suffix *-ing* is, obviously, specific to English, and does not exist in any other language; therefore it capitalized and written as *Participle*. If, however, a writer is discussing the category of participles generally and crosslinguistically, not specific to any particular languages, the term is written in lowercase as *participle*. I follow these same capitalization conventions in this thesis.

The first mention of a language within each chapter is followed by its genealogical affiliation (following the format `family > phylum`) and the location where it is spoken. For example, Central Alaskan Yup'ik would appear as "Central Alaskan Yup'ik (Eskimo-Aleut > Eskimo; Alaska)". Language information is taken from the Glottolog database (Hammarström, Forkel & Haspelmath 2019). Language names are given in English following Haspelmath (2017). A complete list of languages mentioned in this thesis, along with their ISO 639-3 codes and Glottolog codes, is in the List of Languages.

Within quotations, *italics* indicates emphasis in the original, while **boldface** indicates my emphasis.

After each graphical representation of data, I have included the file path within the accompanying GitHub repository for this thesis to the script which will generate that figure.

Chapter 1

Introduction

This chapter motivates the need for research on lexical flexibility by situating it within broader concerns regarding linguistic categories more generally, and categories in human cognition. The specific problem that this study seeks to address is our lack of understanding regarding what lexical flexibility looks like, and how it varies across languages. This thesis contributes to answering these questions via a quantitative corpus-based study of lexical flexibility in English (Indo-European) and Nuuchahnulth (Wakashan). It is the first study to examine lexical flexibility using natural discourse from corpus data. This chapter provides an overview of the thesis, including the specific research questions addressed, the data and methods used, a concise summary of the results, and a preview of the conclusions.

Word classes such as noun, verb, and adjective (traditionally called *parts of speech*) were once thought to be universal, easily identifiable, and easily understood. Today they are one of the most controversial and least understood aspects of language. While language scientists generally agree that word classes exist, there is much disagreement as to whether they are categories of individual languages, categories of language generally, categories of human cognition, categories of language science, or some combination of these possibilities (CITE: Mithun 2017: 166; Haspelmath 2018; Hieber forthcoming: 1). Lexical categorization—how languages separate words into categories—is of central importance to theories of language because it is tightly interconnected with linguistic categorization generally, which in turn informs (and is informed by) our understanding of cognition. Categorization is a fundamental

feature of human cognition (CITE: Taylor 2003: xi), and lexical categorization is perhaps the most foundational issue in linguistic theory (CITE: Croft 1991: 36; Vapnarsky & Veneziano 2017: 1).

One challenge for traditional theories of word classes is the existence of *lexical flexibility*—the use of a word in more than one discourse function, whether to refer (as a noun), to predicate (as a verb), or to modify (as an adjective). In traditional terms, flexible words are those which may be used for more than one part of speech. (A more precise definition of lexical flexibility is given in Chapter 2.) Examples of flexible words in several languages are shown below.

Give examples here. Discuss all the examples together in the lead out. Examples: English; Nuuchahnulth: Kingfisher 202

Flexible words like those in the examples above create an analytical problem for traditional theories of parts of speech. Traditional theories assume that words can be partitioned into mutually exclusive categories on the basis of a clear set of criteria, an approach that has its roots in the Aristotelian tradition of defining a category via its necessary and sufficient conditions. Flexible words would seem to violate this assumption because they appear to be members of more than one category at once, and the criteria for classification yield conflicting results.

Researchers have proposed numerous solutions to this problem. The most common response is to adjust the selectional criteria so that only certain features are considered definitional of the class, allowing these researchers to dismiss other, potentially contradictory evidence as irrelevant (CITE: Baker 2003; Dixon 2004; Floyd 2011 for Quechua; Chung 2012 for Chamorro; Palmer 2017). It is also common to analyze different uses of a putatively flexible word as instances of *heterosemy*—that is, entirely distinct words which share the same form but belong to different word classes (CITE: Lichtenberk 1991). In this view, heterosemous words are related only historically, via a process of conversion or functional shift, in essence

denying the existence of lexical flexibility (CITE: Evans & Osada 2005). Another approach is to claim that, while all words can be neatly categorized, some words in some languages may nonetheless be used for functions typically associated with other categories. A notable example of this is Launey's (CITE: 1994?, 2001?) analysis of Classical Nahuatl, which he calls an *omnipredicative* language. In this analysis, Classical Nahuatl has the traditional, clearly-delineated word classes of noun and verb, but allows for any word to function as a verb regardless of its category (hence the term *omnipredicative*). The reverse is not true however; only some verbs may function as nouns. This difference in behavior is taken as the basis for a categorical[1] distinction between nouns and verbs.

Make sure that this is an accurate description of the empirical facts of Classical Nahuatl.

Some researchers enthusiastically embrace the existence of lexical flexibility and abandon a commitment to the traditional categories of noun, verb, and adjective. Instead they analyze flexible lexemes as belonging to a broader, flexible word classes such as "flexibles", "contentives" or "non-verbs", etc. (CITE: Hengeveld & Rijkhoff 2005; Luuk 2010?). Other researchers abandon the commitment to word classes entirely. Mandarin, Tagalog, Tongan, and Riau Indonesian have each been analyzed as lacking parts of speech (CITE: Simon [1937], McDonald [2013], and Sun [2020] for discussions of early analyses of Mandarin; Gil [XXXX] for Tagalog; Broschart [XXXX] for Tongan; Gil [XXXX] for Riau Indonesian). Within generative linguistics, the Distributed Morphology framework takes it as an assumption that all word roots are category-neutral (CITE: Siddiqi 2018).

Note that these differences in perspective do not arise from disagreements about the empirical facts of each language. Researchers mostly agree on the empirical data, but disagree on the relative importance of various pieces of evidence, and on which criteria should be taken as diagnostic of a category (CITE: Croft & van Lier 2012: 58). It is rare that an argument for flexibility is refuted on the basis of the linguistic facts alone (CITE: cite Mithun's response to Sasse's analysis of Cayuga as a flexible language).

Add example from Stassen [1997: 32] about two researchers coming to different conclusions about Sudanese. Also the example of Quechua or perhaps Iroquoian

Since analyses of lexical flexibility depend more on the particular theoretical commitments

[1] Throughout this thesis, I use the term *categorical* to mean 'without exception; unconditional' and the term *categorial* to mean 'having to do with categories'.

of the researchers involved rather than any particular crucial pieces of evidence, this leads to an intractable problem: researchers cannot agree on the criteria that should be considered diagnostic for a given category in a specific language (let alone crosslinguistically). Instead they partake in *methodological opportunism* (CITE: Croft 2003?: ??), choosing the evidence and criteria which best support their theoretical commitments. Discussions in the literature about the existence of a particular category in a particular language are therefore often unproductive, and devolve into debates about theoretical assumptions or the relevance or importance of various pieces of evidence, which are ultimately unresolvable (CITE: Croft 2005: 435).

This is particularly unfortunate because lexical flexibility is by no means an isolated or minor problem. Additional examples like those above could be provided for many or perhaps even all of the world's languages. Lexical flexibility is not as rare or marginal as traditional approaches to word classes lead one to believe. In a survey of word classes in 48 indigenous North American languages (CITE: Hieber forthcoming), every one of the languages surveyed exhibited lexical flexibility in at least some area of the grammar (although not all authors analyzed these cases as such). In my own experience researching lexical flexibility over the last decade, I have yet to encounter a language that does not exhibit a degree of flexibility in at least some words, however marginally. The prevalence with which different areas of the grammars of the world's languages lack sensitivity to the distinctions between reference (nouns), predication (verbs), and modification (adjectives) suggests that lexical flexibility may not be so much of a problem as it is a design feature of language (CITE: Hieber forthcoming). Indeed, as discussed in Chapter 2, existing research shows that lexical flexibility is constrained and shaped by the very principles that give rise to the crosslinguistic categories of noun, verb, and adjective in the first place (CITE: Croft 2000; Croft 2005; Croft & van Lier 2012).

It is only recently that lexical flexibility has become an object of study in itself, rather than a problem to be solved. As explained above, most prior studies aim to advance a particular analysis rather than to expand empirical coverage of the phenomenon. While they often provide numerous examples, they neither quantitative nor comprehensive. As yet, there are

only a small number of empirical investigations into the extent and nature of lexical flexibility in individual languages (let alone crosslinguistically). The existing studies are of two types: lexicon-based studies which examine dictionaries to determine whether words may be used for multiple functions, and corpus-based studies which examine whether and how often words are used for multiple functions in discourse.

An early lexicon-based study, though not explicitly focused on lexical flexibility, is Croft's (CITE: 1984) study of categories of Russian word roots (summarized in (CITE: Croft 1991: 66)). Croft finds that Russian roots are unmarked, or among the least marked forms, when their semantic category (object, action, or property) aligns with their discourse function (reference, predication, or modification respectively). When roots are used for discourse functions that are atypical for their meaning—in other words, when they are used flexibly—they are marked in some way (or at least as marked as their prototypical uses). These data suggest that lexical flexibility is constrained in a principled way, by what Croft calls the *typological markedness of parts of speech* (explained in detail in Chapter 2).

In arguing that Mundari is *not* a flexible language, Evans & Osada (CITE: 2005) conduct a dictionary analysis using a focused 105-word sample as well as a larger 5,000 word-sample. In the 105-word sample, 74 words (72%) could be used as either noun or verb. In the larger sample, 1,953 words (52%) could be used as both noun and verb. The complete figures for the large sample are shown in Figure X. Evans & Osada argue on the basis of these data that, because not all the words in the Mundari lexicon are flexible, Mundari cannot be considered a flexible language. As with any whole-language typology, however, this is an oversimplification. To overlook the flexibility of these words ignores the behavior of a vast portion of the lexicon. It is exactly this behavior which is of interest in this thesis. Evans & Osada's study constitutes an important contribution to our knowledge of the empirical extent of lexical flexibility across languages.

Add the table from Evans & Osada 2005: 383—large lexicon-based sample of Mundari.

Creissels (CITE: 2017) is a careful lexicon-based of flexibility in Mandinka (Mande; West

Africa). While Mandinka has nominal and verbal constructions that allow the predicative and referring functions of words to be distinguished unambiguously, it is not as easy to separate word stems themselves into similar classes, owing to the fact that no Mandinka lexemes are used exclusively in verbal constructions—all Mandinka lexemes may occur in nominal constructions as well. While Creissels does not dispute this fact, he shows that there is a crucial distinction to be made between two classes of word stems: 1) those whose nominal use is predictable and therefore analyzable as a case of "morphologically unmarked nominalization" (zero-marked conversion) from one category (verb) to another (noun)—these are always event nominalizations; and 2) those whose meaning in nominal constructions is idiosyncratic and therefore not predictable. Creissels calls the former *verbal* words and the latter *verbo-nominal*. He states that both word classes exhibit categorial flexibility, just of different natures. There is also a small set of nominal words used marginally as verbs. These cases are always semantically predictable. Even individual senses of a word can sometimes show varying behavior as to their flexibility. Although Creissels' study unfortunately does not provide counts of the different stem types, it nonetheless adds to our understanding of lexical flexibility by showing how it may have varied realizations, within a single language or even a single word.

Mithun (CITE: 2017: 163) also conducts a lexicon-based analysis of words roots? stems? in Central Alaskan Yup'ik (Alaska; Eskaleaut) using Jacobson's (CITE: year?) exhaustive dictionary, and shows that only a small minority of roots? stems? (12%) exhibit flexibility and can be used as both nouns and verbs. The results of this study are shown in Table X. The words in these groups cannot be characterized in any general or semantic way (CITE: Mithun 2017: 163). Mithun's finding that flexibility in Yup'ik is rather marginal is surprising given that Yup'ik was the focus of an extensive debate about whether the language distinguished nouns and verbs (CITE: just cite Jacobson here). The fixation with these marginal cases in the literature seems disproportionate to their actual frequency of occurrence, again illustrating the disconnect between research advancing a particular analysis and research aiming to improve empirical coverage of the phenomenon. Just as with Mundari, however, it would be an over-

sight to simply ignore these flexible cases. What accounts for the large difference in the extent of flexibility in the lexicons of Mundari versus Yup'ik?

add table from Mithun 2017: 163—lexicon-based sample of Yup'ik

In summary, existing lexicon-based studies have yielded a range of results, each contribution to our understanding of lexical flexibility, but there are still too few such studies to draw any general conclusions as of yet.

Miscellaneous

There is however still much to discover about lexical flexibility.

Acknowledging the existence of flexibility instead raises fascinating questions: How extensive is the phenomenon in the languages of the world? Are there constraints on flexibility? (We have already shown that the answer here is "yes".) Are there commonalities among words which exhibit greater flexibility than others? Even if flexible uses of words are truly heterosemous, related only historically, there remains the question of how such rampant heterosemy arises. Again, are there patterns or principles to the emergence of heterosemous forms?

Chapter 2

Background

Chapter 3

Data & Methods

Chapter 4

Results

Chapter 5

Conclusion

References

Sources of Literature

The references listed in this section are literature on the topic of this thesis that have been cited in the text.

Bickel, Balthasar, Bernard Comrie & Martin Haspelmath. 2015. *The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses*. Max Planck Institute for Evolutionary Anthropology. Leipzig: Department of Linguistics. https://www.eva.mpg.de/lingua/resources/glossing-rules.php.

Bybee, Joan L. 1985. *Morphology: A study of the relation between meaning and form* (Typological Studies in Language 9). Amsterdam: John Benjamins. DOI:10.1075/tsl.9.

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

Croft, William. 1995. Modern syntactic typology. In Masayoshi Shibatani & Theodora Bynon (eds.), *Approaches to language typology*, 85–144. Oxford: Oxford University Press.

Croft, William. 2000. Parts of speech as typological universals and language particular categories. In Petra M. Vogel & Bernard Comrie (eds.), *Approaches to the typology of word classes* (Empirical Approaches to Language Typology 23), 65–102. Berlin: Mouton de Gruyter.

Croft, William. 2003. *Typology and universals*. 2nd edn. (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press. DOI:10.1017/CBO9780511840579.

Croft, William. 2014. Comparing categories and constructions crosslinguistically (again): The diversity of ditransitives. *Linguistic Typology* 18(3). 533–551. DOI:10.1515/lingty-2014-0021.

Dixon, R. M. W. 2010. *Basic Linguistic Theory*. Vol. 1: *Methodology*. Oxford: Oxford University Press.

Hammarström, Harald, Robert Forkel & Martin Haspelmath. 2019. *Glottolog 4.0*. Max Planck Institute for the Science of Human History. https://glottolog.org.

Haspelmath, Martin. 2007. Pre-established categories don't exist: Consequences for language description and typology. *Linguistic Typology* 11(1). 119–132. DOI:10.1515/LINGTY.2007.011.

Haspelmath, Martin. 2010a. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3). 663–687. DOI:10.1353/lan.2010.0021.

Haspelmath, Martin. 2010b. The interplay between comparative concepts and descriptive categories (Reply to Newmeyer). *Language* 86(3). 696–699. DOI:10.1353/lan.2010.0021.

Haspelmath, Martin. 2017. Some principles for language names. *Language Documentation & Conservation* 11. 81–93. DOI:10125/24725.

Hieber, Daniel W. 2013. On linguistics, linguists, and our times: A linguist's personal narrative reviewed. *Linguistic Typology* 17(2). 291–321. DOI:10.1515/lity-2013-0013.

Newmeyer, Frederick J. 2010. On comparative concepts and descriptive categories: A reply to Haspelmath. *Language* 86(3). 688–695. DOI:10.1353/lan.2010.0000.

Plank, Frans (ed.). 2016. *Linguistic Typology* 20(2): *Of categories: Language-particular – comparative – universal.*

Song, Jae Jung. 2001. *Linguistic typology: Morphology and syntax* (Longman Linguistics Library). London: Routledge.

Song, Jae Jung. 2018. *Linguistic typology* (Oxford Textbooks in Linguistics). Oxford: Oxford University Press.

Stassen, Leon. 2011. The problem of cross-linguistic identification. In Jae Jung Song (ed.), *The Oxford handbook of linguistic typology* (Oxford Handbooks in Linguistics), 90–99. Oxford: Oxford University Press. DOI:10.1093/oxfordhb/9780199281251.013.0006.

Whaley, Lindsay J. 1997. *Introduction to typology: The unity and diversity of language.* Thousand Oaks, CA: SAGE Publications.

Language Index

Todo list