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

Draft

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

Copyright © 2020

by

Daniel W. Hieber

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

Published under a Creative Commons Attribution 4.0 License (CC BY 4.0):
https://creativecommons.org/licenses/by/4.0/

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

December 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 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.

In all interlinear glossed examples, I follow the formatting conventions (but not necessarily the recommended abbreviations) of the Leipzig Glossing Rules (Bickel, Comrie & Haspelmath 2015). The source of each example is always provided after the example itself.

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.

1.1 The "problem" of lexical flexibility

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 Sec. XX [of Ch. 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, Riau Indonesian, and Proto-Indo-European have each been analyzed as lacking parts of speech in some area of the grammar (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; Kastovsky [1996] for Proto-Indo-European). Within generative linguistics, the Distributed Morphology framework takes it as an assumption that all word roots are category-neutral (CITE: Siddiqi 2018). In a more functionalist orientation, Farrell (CITE: cite) argues that *all* instances of flexible words ("functional shift") involve roots underspecified for category.

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

[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'.

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 (Chafe). Also Mundari. Chamorro: Topping and Chung. Among many others.

Since analyses of lexical flexibility depend more on the particular theoretical commitments 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 phenomenon. 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 the existence of lexical categories in a language is not necessarily a given (CITE: Hieber forthcoming).

Indeed, given what we know from both cognitive science and diachronic linguistics, it would be surprising if clear-cut categories *did* exist. Word meanings, lexical categories, and mental categories are all prototypal (CITE: Taylor 2003), and language change is both gradual

add footnote about prototypical vs. prototypal

and gradient (CITE: Gradualness and gradience in grammaticalization; Grammaticalization (Hopper & Traugott); Diachronic Construction Grammar). There will be more or less central members of any given category, and at any given point in time a word might be in a stage of transition or expansion from one category into another, meaning that it will show attributes of both. Given these facts, the real curiosity is how discourse functions come to be grammaticalized in language over time, not why it is that some languages lack such distinctions in certain areas of their grammars. Lexical flexibility is not so much of a problem as it is a design feature of language. It is precisely the liminal categorial status of flexible words that makes them interesting:

> In the functionalist view, linguists should recognize the boundary status of the cases in question and try to understand why they are boundary cases. The major empirical fact that has led to concrete results for typology is the discovery that the cross-linguistic variation in such things as the basic grammatical distinctions is patterned. ((CITE: Croft 1991: 23))

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 are 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). What follows is a brief synopsis of the existing studies of this latter type.

1.2 Previous research

The existing studies on the empirical extent of lexical flexibility 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 Sec. XX [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 Table 1.1. 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.

Table 1.1: Percentage of words used as nouns, verbs, or both in Mundari (Austroasiatic > Munda; India) (Evans & Osada 2005: 383)

noun only	772	20%
verb only	1,099	28%
noun and verb	1,953	52%
Total	3,824	100%

(CITE: Add Evans & Osada [2005: 383] citation to caption in this table.)

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 1.2. 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

oversight to simply ignore these flexible cases. Instead we should ask what accounts for the large difference in the extent of flexibility in the lexicons of Mundari versus Yup'ik.

Table 1.2: Percentage of words used as nouns, verbs, or both in Central Alaskan Yup'ik (Eskaleaut > Eskimo; Alaska) (Mithun 2007: 163)

noun only	35%
verb only	53%
noun and verb	12%
Total	100%

(CITE: Add Mithun [2017: 163] citation to caption in this table.)

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.

Corpus-based studies of lexical flexibility are also scarce. In a study of the discourse functions of property words in English and Mandarin (Sino-Tibetan > Sinitic; China), Thompson (CITE: 1989) reports that predicative uses of adjectives are in fact more common than attributive (modifying) uses of adjectives in conversation. The resulting figures from this study are shown in Table 1.3.

Table 1.3: Distribution of functions of property words in English (Indo-European > Germanic; England) and Mandarin (Sino-Tibetan > Sinitic; China) (Thompson 1989: 253, 257)

	English		Mandarin	
predicative adjectives	209	86%	243	71%
attributive adjectives	34	14%	97	29%

Some of the attributive adjectives reported in Table 1.3 have "anaphoric head nouns" (CITE: Thompson 1989: 258), meaning that they are adjectives functioning to refer, so the figures presented are not entirely representative of the pragmatic functions of these words. The study also does not discuss the extent to which *individual* words exhibit this predicate-modifier flexibility—we only have the data in aggregate—and it also excludes any prototypical nouns

being used to modify. These methodological choices are appropriate for a study of the discourse uses of prototypical adjectives, but the result is that we cannot infer much about the extent of lexical flexibility in English or Mandarin from this study.

Nonetheless, Thompson's study suggests a functional underpinning to the observed flexibility in prototypical property words. She finds that property words have primarily two functions in discourse: 1) to introduce new referents; and 2) to predicate an attribute about a referent. It is therefore no surprise that property words in some languages have their own specialized constructions, since they represent a unique mix of referring and predicating functions. Likewise it is unsurprising that languages would encode property concepts using either referring or predicating constructions, since prototypical adjectives exhibit behavior related to both functions.

A similar study is Croft's (CITE: 1991: Sec. 2.5) investigation of *textual markedness*, which refers to the fact that prototypical uses of a word are more frequent than non-prototypical uses of a word in texts (as is generally predicted by prototype theory; (CITE: Taylor 2003: ??)). Croft counts the frequency with which object, action, and property words are used for each of the pragmatic functions of reference, predication, and modification, and the resulting counts give confirmation to textual markedness theory. Moreover, the data partially elucidate how frequently words of different semantic classes are used for multiple pragmatic functions. Like Thompson's (CITE: 1989) study, however, we do not know these distributions for individual words. Additionally, Croft's data include cases of overtly marked uses of words in non-prototypical functions, which would not be considered instances of lexical flexibility.

There are also some studies which count the proportion of nouns vs. verbs. vs. adjectives in texts (CITE: Hudson 1994; Polinsky & Magyar 2020), but again the data are not disaggregated to the word level, so no firm conclusions can be draw about the extent of lexical flexibility.

In sum, no existing studies examine the distribution of pragmatic functions for individual words, or limit themselves to only flexible (morphologically unmarked) cases. To my knowl-

edge, the studies just reviewed exhaust those that take an empirical approach to determining the extent of lexical flexibility in or across languages. There are numerous additional studies of lexical flexibility, but these either a) focus on particular analyses or theories of flexible words rather than attempt to expand the empirical coverage of lexical flexibility, as mentioned earlier; or b) focus on various dimensions of the *behavior* of flexible words rather than studying the overall *prevalence* of flexibility. This point is not a criticism, but simply a recognition of a lacuna in existing research. The emergent literature which treats lexical flexibility as a phenomenon of interest in its own right and applies empirical data to the task of understanding its behavior has advanced our knowledge of the various ways lexical flexibility can be realized, and what the constraints on that variation are. Existing research shows, for example, 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). This literature and its many findings are reviewed in Sec. XX [in Chapter 2].

There is however still much to discover about lexical flexibility. Most significantly, we do not yet know the overall prevalence of the phenomenon. Most grammatical descriptions of flexibility present a relatively small set of handpicked examples, so that we do not know how representative these examples are. Croft (CITE: Croft 2001: 70) makes this point nicely:

> How do we know that when we read a grammar of an obscure "flexible" language X that the author of the grammar has systematically surveyed the vocabulary in order to identify what proportion is flexible? If English were spoken by a small tribe in the Kordofan hills, and all we had was a 150 page grammar written fifty years ago, might it look like a highly flexible language? ((CITE: Croft 2001: 70))

Equally as significant (and equally as unknown) is whether there are any commonalities among words or languages which exhibit greater flexibility than others. These questions are relevant even if one adopts the position that flexible uses of words are truly heterosemous, related only historically. There remains the question of how such rampant heterosemy arises in the first place. Are there patterns or principles to the emergence of heterosemous forms?

Whether one prefers to analyze this phenomenon as conversion, zero derivation, functional shift, polycategoriality, heterosemy, acategoriality, or something else, the fact is we do not yet have a strong empirical grasp of just how this phenomenon is realized in the world's languages. This thesis is a first foray into filling that empirical gap. The following section describes the contribution made by this thesis to addressing this gap, and gives an overview of the present study.

1.3 Overview of this study

This thesis is a quantitative corpus-based study of lexical flexibility in English (Indo-European > Germanic; England) and Nuuchahnulth (Wakashan; Pacific Northwest). It is exploratory and descriptive, with the primary goal of describing the prevalence of lexical flexibility within and across languages. The specific research questions investigated are as follows:

R1: How flexible are words in English and Nuuchahnulth?

R2: Is there a correlation between degree of lexical flexibility for a word and frequency (or corpus dispersion)?

R3: How do the semantic properties of words pattern with respect to their flexibility?

I explore each of these questions from several angles. R1, "How flexible are words in English and Nuuchahnulth?" is the core focus of this thesis. To answer it, I count the frequency with which lexical words are used for each of the three functions of reference, predication, and modification in a corpus of spoken texts for each language. Each word is given a flexibility rating from 0 to 1 based on how evenly its uses are distributed across the three functions. A rating of 0 indicates that the word is highly inflexible, with all its occurrences being used for a single function; a rating of 1 indicates that the word is maximally flexible, with its occurrences evenly distributed across the three functions. By quantifying the flexibility of each word in this way, it then becomes possible to look for statistical correlations between the flexibility of

a word and other factors, such as those addressed by the other two research questions. It also enables us to answer the question of just how pervasive flexibility is in the two languages.

R2, "Is there a correlation between degree of lexical flexibility for a word and frequency (or corpus dispersion)?", uses the flexibility ratings calculated in R1 to consider whether the flexibility of a word correlates with either its overall frequency or with its *corpus dispersion* (how evenly/regularly the word appears in a corpus), a measure which is thought to more accurately capture the notion of frequency of exposure (CITE: Gries and accompanying citations therein). This question two motivations: First, some researchers have claimed or implied that all words may exhibit flexibility if you examine enough tokens of the word (CITE: citations). If true, this would lend some empirical support to the claim that all words are to some degree flexible, or perhaps even acategorial. Second, higher-frequency words often preserve irregular or atypical forms or functions (CITE: Bybee, others?), such that words with higher frequencies might be more likely to retain their non-prototypical, flexible uses. Both these potential factors invite inquiry into the relationship between frequency and flexibility.

R3, "How do the semantic properties of words pattern with respect to their flexibility?", is investigated using a mix of quantitative and qualitative methods. Unlike the other two research questions, which are intended to capture the extent of flexibility in and across languages, R3 is an inquiry into the semantic *behavior* of flexible (and inflexible) words. This research question is directly motivated by Croft's (CITE: all the usual citations for this) typological markedness theory of lexical categories, which claims among other things that words used in non-prototypical functions (for example, a property word being used to refer, as a noun) will always show a semantic shift in the direction of the meaning typically associated with that function. So, if a property word is used to refer, its meaning should be more object-like than property-like; that is, it should mean something like 'an entity with the property X' rather than 'the abstract property X'. Croft's (CITE: 1991) seminal work in this area provides strong empirical evidence for this semantic markedness principle, but is nonetheless somewhat preliminary. Croft himself has in various places implored linguists to investigate

the lexical semantics of these functional shift further (CITE: Croft 2005: 440; Croft & van Lier 2012: 70), but as yet little research has responded to this call (though see (CITE: citations)). Investigating the semantic patterns that appear in cases of lexical flexibility is therefore another contribution of this thesis, addressed by question R3.

The preceding notes are simply a high-level summary of the principal research questions investigated in this thesis. A complete description of the methods used in answering each question is given in Chapter 3.

This study aims to be framework neutral in the sense of Haspelmath (CITE: 2008). Its findings should be interpretable and of interest to researchers working in a range of linguistic theories and with different approaches to lexical categories. As mentioned in §??, the results of this study do not depend on whether one analyzes lexical flexibility as polycategoriality, conversion, or something else. While my own perspective on language is decidedly functional, this is of little relevance to how I coded the data, the procedures for which are described in detail in Chapter 3. The relevant factors in this study are operationalized in a theory-neutral way (to the extent such a thing is possible), and I expect that my coding decisions for individual data points will be found largely unobjectionable.

Several principles guided the choice of data used for this study. First, a self-imposed requirement for this project is that of empirical accountability and replicability. It should be possible for other researchers to apply the measure of lexical flexibility defined in Chapter 3 to new corpora, or to replicate the results of the present study on the existing dataset. As such, I only used data that were publicly available and, if possible, open access. Second, since the aim of this study is to investigate lexical flexibility in actual language *use*, I rely solely on naturalistic data from spoken texts. This has the additional advantage of abetting comparison between other, less well documented languages, since the majority of corpora of minority languages consist mainly of spoken texts. Third, I sought to examine data from languages that have featured prominently in discussions of lexical flexibility in the literature, with the intention of offering a more expansive empirical foundation for future discussions. With these

principles in mind, I chose to focus this study on English and Nuuchahnulth.

English has at various times been described as both a highly flexible language with fluid category membership (CITE: Crystal 1967: 47-48; Vonen 1994; Farrell 2001: 111; Cannon 2009) and a fairly rigid language with clearly-delineated categories (CITE: Rijkhoff 2007; Schachter & Shopen 2007; Velupillai 2012). It is used as a point of comparison for nearly every discussion of lexical flexibility, but we do not have a clear idea of just how flexible English words are. Its inclusion in this study is therefore well justified. The data for English are from the Open American National Corpus (OANC), a 15-million word corpus of American English comprising numerous generes of both spoken and written data, all of which is open access (CITE: OANC). This study uses just the spoken portion of the corpus, consisting of approximately 3.2 million words, which is itself composed of two distinct subcorpora—the Charlotte Narrative & Conversation Collection (or simply "the Charlotte corpus") (CITE: Charlotte corpus) and the Switchboard Corpus (CITE: Switchboard corpus).

Nuuchahnulth (formerly referred to in the liteature as Nootka) is a Wakashan language presently spoken by a hundred or so people on and around Vancouver Island, British Columbia, in the Pacific Northwest. Nuuchahnulth, together with the other members of the Wakashan family (especially Makah and Kwak'wala / Kwakiutl) is one of the most discussed languages in the literature on lexical flexibility (CITE: many, many citations, of both Nuuchahnulth and the other Wakashan languages). This is due largely to the following examples of flexible words from (CITE: Sapir & Swadesh?).

add famous Nuuchahnulth examples

Hardly a single typological survey of lexical categories or study of lexical flexibility has failed to include these examples since . Yet we still do not know how representative these examples are of Nuuchahnulth in general. What is more, lexical flexibility is an areal feature of the entire Pacific Northwest. The nearby Salishan, Chimakuan, Tsimshianic, Chinookan, and Sahapatian families as well as the isolate Kutenai each exhibit lexical flexibility to a presumably strong degree, since they have caught the attention of so many researchers in this regard

add year of famous Nuuchahnulth examples

(CITE: citations for each family/isolate). Again, we do not actually know whether this literature is truly representative of the pervasiveness of the phenomenon, or whether its "exotic" nature as compared to Indo-European languages has simply garnered undue attention to the topic in this geographic region. Nuuchahnulth, being the most discussed of these languages, is therefore nearly obligatory to include in a study such as this one.

The data used for the investigation of Nuuchahnulth comes from a corpus of texts collected and edited by Toshihide Nakayama and published in Nakayama (CITE: 2003a, 2003b). The corpus consists of 24 texts dictated by two speakers, containing 2,081 utterances and 8,366 tokens (comprising 4,216 types). The texts cover a variety of genres, including procedural texts, personal narratives, and traditional stories. I manually retyped these texts as scription files (a simple text format for representing interlinear glosses in a way that is both familiar to linguists and computationally parseable; (CITE: scription)) for analysis. The resulting digitally-searchable corpus is available on GitHub at `https://github.com/dwhieb/Nuuchahnulth`.

Other languages that would have been obvious choices for inclusion in this study are Riau Indonesian (CITE: Gil), Mundari (CITE: Evans & Osada; Rijkhoff & Hengeveld), Classical Nahuatl (CITE: Launey), and Yup'ik (CITE: Thalbitzer; Jacobson; Mithun). Each of these has generated contested claims about their flexibility and the existence of flexibility more generally. However, practicalities have limited me to examining just English and Nuuchahnulth for the time being. I leave investigations of other languages to future research and researchers.

Both the English and Nuuchahnulth corpora were converted into to the Data Format for Digital Linguistics (DaFoDiL) (a JSON format for representing linguistic data; (CITE: DaFoDiL)) for tagging and scripting purposes. This made it possible to use the Digital Linguistics (DLx) ecosystem of tools and software to more quickly tag and analyze the data. More information about Digital Linguistics may be found at `https://digitallinguistics.io`.

All of the datasets, scripts, and source files for this thesis are publicly available on GitHub

at https://github.com/dwhieb/dissertation.

Turning now to results:

Regarding R1, "How flexible are words in English and Nuuchahnulth?", I find that English and Nuuchahnulth differ significantly not only in their overall degree of flexibility, but also in how that flexibility is realized. In English, the majority of words surveyed are flexible, but only to a small degree. Most lexical words of English can be used as nouns, verbs, or adjectives, but there is a strong tendency for each word to be used for primarily one function. English thus shows a consistent but somewhat marginal degree of flexibility. In constrast, most words in Nuuchahnulth are highly flexible, but primarily along the noun-verb axis; Nuuchahnulth words are very freely used as both nouns and verbs, but only infrequently used as adjectives. Nuuchahnulth thus shows a consistently high degree of flexibility, but primarily in just one dimension.

final results: flexibility vs. frequency

For R2, "Is there a correlation between degree of lexical flexibility for a word and frequency (or corpus dispersion)?", I found that higher frequency words are more flexible than lower frequency words, but that the effect was very small.. The same facts held when comparing degree of lexical flexibility with corpus dispersion. Words that are more evenly dispersed in a corpus have a slight tendency to be more flexible than those that are less evenly dispersed. These findings suggest that the degree of flexibility exhibited by a word does depend in part on how regularly speakers use it.

Lastly, R3 asks "How do the semantic properties of words pattern with respect to their flexibility?". With respect to Nuuchahnulth, I find that property words, especially numerals and quantifiers, are the most flexible semantic class of words. Nearly all of the most flexible words denote property concepts. Deictic expressions such as *this*, *that*, *here*, *there* also rank very high in their flexibility. I also find that there are strong correlations between morphologically marked aspect (durative, continuative, inceptive, etc.) and discourse function. In Nuuchahnulth, aspect markers may be used with either predicates or referents; they are not an exclusively verbal category. However, I find that the presence of any aspect marker

does correlate strongly with predication, lending additional empirical evidence to Hopper & Thompson's (CITE: 1984) claim that words items used in their prototypical function will show the inflectional behaviors typical of that function. The momentaneous and telic aspect markers are the only ones in Nuuchahnulth which show any sort of tendency towards use with referents, while the durative was the only aspect marker to show any sort of tendency towards use with modifiers. Since aspect is a grammatical category that expresses how speakers construe the temporal structure of an event, these data suggest that flexibility has a great deal to do with how speakers are construing a word—as an action, object, or property—as has been suggested by Croft (CITE: 1991; also some of the cognitive literature).

Nuuchahnulth also has a definite suffix *-ʔiː* used with referents. Nakayama (CITE: 2001: 48) states that this suffix is used with action words being construed as objects. This observation suggests that the definite suffix may have a clarifying function, appearing whenever a predicate is used for the atypical role of reference (as predicted by Croft's structural coding hypothesis; see §?? for more details). One hypothesis that arises from applying typological markedness theory to Nuuchahnulth is that aspect markers which correspond to more object-like construals of a word (durative, telic, momentaneous) are more likely to be marked with the definite suffix. This turns out to be true, but only trivially so—only a tiny percentage (7.98%) of words with definite markers also had aspect markers. However, this leads to the far more interesting observation that the definite marker and the aspect markers in Nuuchahnulth are *almost* entirely mutually exclusive. They only rarely co-occur. These facts demonstrate that even in a language with rampant flexibility, as this study shows Nuuchahnulth to be, that flexibility is nonetheless bound by universal typological constraints.

To summarize, this thesis makes contributions in several areas. The first is methodological: this thesis lays out a procedure for quantifying lexical flexibility for individual words in a corpus that can be replicated for other languages and corpora (Chapter 3). The second contribution is empirical and descriptive: I describe the extent of lexical flexibility and the manner in which it operates in English and Nuuchahnulth (Chapter 4). The final area is analytical

and theoretical: I argue that the data and statistical analysis presented in this thesis support Croft's typological markedness theory of word classes, in which lexical categories such as noun, verb, and adjective are not in fact categories of particular languages as has been historically assumed, but instead are emergent patterns that arise from how speakers use object, action, and property words for different functions in discourse (reference, predication, and modification). Words used for functions that are not prototypical of their meaning *tend* to be more marked (morphologically, behaviorally, semantically, and/or frequentially) than prototypical uses, but this is not an absolute universal. Lexical flexibility is the natural and expected result of the fact that these non-prototypical uses are *not* always *morphologically* marked, even when they are marked in other ways (Chapter 5).

The remainder of this thesis is organized as follows: Chapter 2: Background summarizes previous definitions of lexical flexibility and discusses their shortcomings. I propose an alternative, functionally-oriented definition that is consistent with cognitive and typological approaches to word classes instead. Chapter 3: Data & Methods describes in detail how the data were coded and analyzed for each of the major research questions (and contributing subquestions) in this study. I discuss factors that influenced how the data were coded, and outline the various coding decisions that were made. I present and explain a measure of corpus dispersion that is used partly in place of, and partly as a complement to, raw frequencies of words. Lastly, I set forth a procedure for operationalizing and quantifying lexical flexibility in a crosslinguistically comparable way. Chapter 4: Results presents the empirical findings from this study. I demonstrate how the methodolgical techniques from Chapter 3 are applied to individual words, and then present aggregated views of the data for English and Nuuchahnulth respectively. Chapter 5: Discussion & Conclusion considers the implications of the results in Chapter 4 for theories of lexical categories. I argue that the data support a typological-universal theory of word classes, and that lexical flexibility should be viewed as a natural result of the cognitive and diachronic processes at work in language, rather than as an exceptional phenomenon. I conclude by discussing some limitations of the present study

and avenues for future research, followed by closing remarks.

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.

To Do