

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 Bernard Comrie

Professor Stefan Th. Gries

Professor William Croft (University of New Mexico)

January 2021

The dissertation of Daniel W. Hieber is approved.

---

Bernard Comrie

---

Stefan Th. Gries

---

William Croft

---

Marianne Mithun, Committee Chair

January 2021

Lexical flexibility in discourse:  
A quantitative corpus-based approach

Copyright © 2021

by

Daniel W. Hieber

Typeset using L<sup>A</sup>T<sub>E</sub>X 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

March 2021	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
2015–2019	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

2023	Chitimacha. In Carmen Jany, Marianne Mithun, & Keren Rice (eds.), <i>The languages and linguistics of indigenous North America: A comprehensive guide</i> (The World of Linguistics 13). Mouton de Gruyter.
2023	Word classes. In Carmen Jany, Marianne Mithun, & Keren Rice (eds.), <i>The languages and linguistics of indigenous North America: A comprehensive guide</i> (The World of Linguistics 13). Mouton de Gruyter.
2019	The Chitimacha language: A history. In Nathalie Dajko & Shana Walton (eds.), <i>Languages in Louisiana: Community &amp; culture</i> (America's Third Coast Series). University Press of Mississippi.
2019	Semantic alignment in Chitimacha. <i>International Journal of American Linguistics</i> 85(3): 313–363. doi: <a href="https://doi.org/10.1086/703239">10.1086/703239</a> .

- 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](https://doi.org/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](https://doi.org/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 2<sup>nd</sup> 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 by analyzing the discourse function of nearly 400,000 tokens of English and nearly 9,000 tokens of 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

Acknowledgments . . . . .	v
Curriculum Vitae . . . . .	vi
Abstract . . . . .	ix
Table of Contents . . . . .	x
List of Figures . . . . .	xi
List of Tables . . . . .	xii
List of Abbreviations . . . . .	xiii
List of Languages . . . . .	xiv
Conventions . . . . .	xvii
<b>1 Introduction</b>	<b>1</b>
1.1 The “problem” of lexical flexibility . . . . .	1
1.2 Previous research . . . . .	9
1.3 Overview of this study . . . . .	14
<b>2 Background</b>	<b>25</b>
2.1 Introduction: Approaches to lexical flexibility . . . . .	25
2.2 Traditional approaches . . . . .	27
2.2.1 Universalism . . . . .	27
2.2.2 Relativism . . . . .	30
2.2.3 Structuralism . . . . .	31
2.3 Flexible approaches . . . . .	36
2.3.1 Key concepts . . . . .	37
2.3.1.1 Lexical flexibility . . . . .	38
2.3.1.2 Polycategoriality . . . . .	40
2.3.1.3 Multifunctionality / Polyfunctionality . . . . .	40
2.3.1.4 Precategoriality / Acategoriality . . . . .	42
2.3.1.5 Monocategoriality . . . . .	45
2.3.1.6 Transcategoriality . . . . .	45
2.3.1.7 Conversion / Zero derivation . . . . .	46
2.3.1.8 Functional shift / Functional expansion . . . . .	47
2.3.2 Key findings . . . . .	47
2.3.2.1 Parts-of-speech hierarchy . . . . .	48
2.3.2.2 Reference-predication asymmetries . . . . .	49
2.3.2.3 Locus of categoriality . . . . .	51
2.3.2.4 Item-specificity . . . . .	53

2.3.3	Problems & critiques . . . . .	56
2.3.3.1	Methodological opportunism . . . . .	57
2.3.3.2	Semantic shift . . . . .	58
2.3.3.3	Lexical gaps . . . . .	60
2.3.3.4	Counterarguments . . . . .	61
2.4	Functional approaches . . . . .	67
2.4.1	Prototype theory . . . . .	67
2.4.2	Typological markedness theory . . . . .	73
2.5	Lexical flexibility: A functional definition . . . . .	79
<b>3</b>	<b>Data &amp; Methods</b>	<b>84</b>
3.1	Introduction . . . . .	84
3.2	Data . . . . .	86
3.3	Methods . . . . .	89
3.4	Analysis . . . . .	95
3.4.1	Measuring lexical flexibility . . . . .	95
3.4.2	Flexibility and corpus size . . . . .	97
3.4.3	Frequency vs. dispersion . . . . .	98
3.5	Summary . . . . .	101
<b>4</b>	<b>Results</b>	<b>103</b>
4.1	Introduction . . . . .	103
4.2	Interpreting the results . . . . .	104
4.3	R1: Degree of lexical flexibility . . . . .	106
4.4	R2: Lexical flexibility and corpus size . . . . .	111
4.5	R3: Lexical flexibility and frequency / dispersion . . . . .	111
4.6	R4: The semantics of lexical flexibility . . . . .	111
<b>5</b>	<b>Conclusion</b>	<b>112</b>
 <b>Appendices</b>		
<b>A</b>	<b>Early grammatical descriptions</b>	<b>114</b>
<b>B</b>	<b>100-item samples</b>	<b>116</b>
B.1	English . . . . .	116
B.2	Nuuchahnulth . . . . .	120
<b>C</b>	<b>Sample annotations</b>	<b>124</b>
C.1	English . . . . .	124
C.2	Nuuchahnulth . . . . .	125

Backmatter	
References	108
Language Index	125

# List of Figures

2.1	Timeline of early grammatical descriptions of European vs. American languages	29
2.2	Monocentric vs. polycentric categories . . . . .	44
2.3	Hengeveld's (1992) typology of parts-of-speech systems . . . . .	48
2.4	Semantic map of English <i>run</i> . . . . .	70
2.5	Crosslinguistic differences in the encoding of spatial relationships . . . . .	71
2.6	Typological prototypes for noun, verb, and adjective . . . . .	77
3.1	The relation between word frequency and dispersion (DP) . . . . .	99

# List of Tables

1.1	Types of conversion in English (Indo-European > Germanic) . . . . .	9
1.2	Percentage of lexical items used as nouns, verbs, or both in Mundari (Aus- troasiatic > Munda) . . . . .	10
1.3	Percentage of roots used as nouns, verbs, or both in Central Alaskan Yup'ik (Eskimo-Aleut > Yupik) . . . . .	11
1.4	Distribution of functions of property words in English (Indo-European > Ger- manic) and Mandarin (Sino-Tibetan > Sinitic) . . . . .	12
2.1	Distribution of English Verbs and Adjectives . . . . .	32
2.2	Distributional analysis of English (Indo-European > Germanic) temporal nouns	33
2.3	Distributional analysis of Russian (Indo-European > Slavic) numerals . . . . .	34
2.4	Distributional analysis of French (Indo-European > Romance) verbs . . . . .	35
3.1	Distribution of discourse functions for a perfectly flexible lexical item . . . . .	95
3.2	Distribution of discourse functions for a perfectly rigid/inflexible lexical item .	96
A.1	Some first grammatical descriptions of European vs. American languages . . .	114
B.1	Corpus statistics for the 100-item English sample . . . . .	116
B.2	Corpus statistics for the 100-item Nuuchahnulth sample . . . . .	120

# List of Abbreviations

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

1	first person	MOM	momentaneous
2	second person	NAME	proper name
3	third person	NEG	negative
ACC	accusative	NEUT	neutral position
AGR	agreement	NF	non-first person
AGT	agent	PAST	past
CAUS	causative	PFV	perfective
COMPL	completive	PL	plural
COND	conditional	PLACT	pluractional
CONN	connective	POSS	possessive
DEF	definite	PRED	predicate
DUB	dubitive	PRES	present tense
DUP	reduplication	PURP	purposive
EMPH	emphatic	QUOT	quotative
EP	epenthetic	REF	referent
FIN	finite	REFL	reflexive
FUT	future	REL	relative
HAB	habitual	SG	singular
HUM	human	SPEC	specific
IMP	imperative	SS	same subject
INCEP	inceptive	SUBJ	subject
IND	indicative	SUBORD	subordinate
INDEF	indefinite	TR	transitive
INSTR	instrumental	VZR	verbalizer
INTER	interrogative		
IPFV	imperfective		
LINK	linker		
MOD	modifier		

# List of Languages

The following table provides information about each language mentioned in this thesis: the name of the language in English (following Haspelmath [2017]), the [International Standards Organization \(ISO\) 639-3 language code](#), and the [Glottolog](#) code (Hammarström, Forkel & Haspelmath 2019). Genealogical information follows the format family > phylum.

Language Name (English)	ISO 639-3	Glottocode	Genetic Affiliation
Basque	eus	basq1248	isolate
Castilian Spanish	spa	cast1244	Indo-European > Romance
Cayuga	cay	cayu1261	Iroquoian > Northern Iroquoian
Chamorro	cha	cham1312	Austronesian > Malayo-Polynesian
Cherokee	chr	cher1273	Iroquoian > Southern Iroquoian
Chitimacha	ctm	chit1248	isolate
Central Alaskan Yup'ik	esu	cent2127	Eskimo-Aleut > Yupik
Classical Greek	grc	anci1242	Indo-European > Hellenic
Classical Nahuatl	nci	clas1250	Uto-Aztecan > Nahuan
Dutch	nld	mode1257	Indo-European > Germanic
English	eng	stan1293	Indo-European > Germanic
French	fra	stan1290	Indo-European > Romance
German	deu	uppe1397	Indo-European > Germanic
Gooniyandi	gni	goon1238	Bunuban
Indonesian	ind	indo1316	Austronesian > Malayan



Irish (Gaelic)	gle	iris1253	Indo-European > Celtic
Latin	lat	lati1261	Indo-European > Italic
Kuikuro	kui	kuik1245	Cariban > Nahukwa
Kutenai	kut	kute1249	isolate
Mandarin Chinese	cmn	mand1415	Sino-Tibetan > Sinitic
Mandinka	mnk	mand1436	Mande > Manding
Middle English	enm	midd1317	Indo-European > Germanic
Mixtec	various	mixt1427	Oto-Manguean > Mixtecan
Mundari	unr	mund1320	Austroasiatic > Munda
Munya (Muya)	mvm	muya1239	Sino-Tibetan > Qiangic
Narragansett	xnt	narr1280	Algic > Eastern Algonquian
Navajo	nav	nava1243	Na-Dene > Athabaskan
North Efate (Nguna)	llp	nort2836	Austronesian > Oceanic
Nuuchahnulth (Nootka)	nuk	nuuc1236	Wakashan > Southern Wakashan
Occitan	oci	occi1239	Indo-European > Romance
Old English	ang	olde1238	Indo-European > Germanic
Quechua	qwe	quec1387	Quechuan
Quiché Maya	quc	kich1262	Mayan > Quichean
Russian	rus	rus1263	Indo-European > Balto-Slavic
Soddo	gru	kist1241	Afroasiatic > Ethiopic
Spanish	spa	stan1288	Indo-European > Romance
Standard Arabic	ara	arab1395	Afroasiatic > Semitic
Sundanese	sun	sund1251	Austronesian > Malayo-Polynesian
Tagalog	tgl	taga1280	Austronesian > Philippine
Tarascan (Purépecha)	tsz	tara1323	isolate
Timucua	tjm	timu1245	isolate
Tongan	ton	tong1325	Austronesian > Polynesian

Tuscan (Italian)	cay	cayu1261	Indo-European > Romance
Tzeltal Maya	tzh	tze11254	Mayan > Cholan > Tzeltalan
Ute	ute	utee1244	Austronesian > Polynesian
Wambon	wms	ketu1239	Trans-New Guinea > Awyu-Dumut
Welsh	cym	wels1247	Indo-European > Celtic
Wolof	wol	wolo1247	Niger-Congo > Senegambian
Yucatec Maya	yua	yuca1254	Mayan > Yucatec
Zapotec	zap	zapo1437	Oto-Manguean > Zapotecan

## CONVENTIONS

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

### Interlinear Examples

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,c; 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.

## Prose

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 is capitalized and written as *Participle*. If, however, a writer is discussing the category of participles generally and crosslinguistically, not specific to any particular language, the term is written in lowercase as *participle*. I follow these same capitalization conventions in this thesis.

## Quotations

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

# 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 addressed 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 > Germanic) and Nuuchahnulth (Wakashan > Southern Wakashan). I analyze approximately 400,000 tokens of English and 9,000 tokens of Nuuchahnulth for their discourse function to determine the overall prevalence of lexical flexibility in these languages. This is the first study to examine lexical flexibility using natural discourse data from corpora. 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 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 (Mithun 2017: 166; Haspel-

math 2019; Hieber forthcoming). Lexical categorization—how languages assign lexical items<sup>1</sup> to 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 (Taylor 2003: xi; van Lier & Rijkhoff 2013: 2–3), and lexical categorization is perhaps the most foundational issue in linguistic theory (Croft 1991: 36; Vapnarsky & Veneziano 2017a: 1).

One challenge for traditional theories of word classes is the existence of *lexical flexibility*—the use of a lexical item in more than one discourse function with no overt derivational morphology, whether it is used to refer (like a noun), to predicate (like a verb), or to modify (like 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 §2.5.) Examples of flexible lexical items in several languages are shown below. In the examples, **Ref** stands for a lexical item being used for reference, **Pred** for a lexical item being used for predication, and **Mod** for a lexical item being used for modification. The flexible item in each set of examples is shown with **emphasis**. Here and throughout this thesis, I use the terms *reference*<sup>2</sup>, *predication*, and *modification* so as to focus on the functions of lexical items and avoid committing to any analysis regarding their part-of-speech classification.

(1) English (Indo-European > Germanic)

**Ref:** And the spots of **paint** would change every hundred degrees.

(Ide & Suderman 2005: FrancisClem)

**Pred:** One story does come to my mind though where you **painted** the foundation coating on the house and got tar all over you.

(Ide & Suderman 2005: BorelRaymondHydellIII)

---

<sup>1</sup>I use the term *lexical item* as a convenient cover term for root, stem, or fully inflected word. This term does not here refer to the phonological word, syntactic word, or any other concept of word. The reason for this vague usage is because languages vary as to which linguistic level bears category information. This issue is discussed more fully in Section 2.3.2.3. I use *lexical item* instead of *lexeme* because the concept of a lexeme implies lexical unity, that is, that we are discussing a single polysemous item rather than two homophonous ones. Use of the term *lexical item* is intended to bypass this distinction in favor of a focus on form. However, I also avoid the term (*lexical*) *form* because some lexical items have multiple forms (in the case of suppletion).

<sup>2</sup>*Reference* is different than *referentiality*. Reference is a pragmatic function, having to do with discourse. Referentiality is ontological, having to do with the state of reality. Thus a *referent* can be *non-referential*, such as with indefinites (*a nap would be a good idea*) or when referring to entities which do not exist (*the unicorn*).

**Mod:** And it happened to be one of the rare **paint** jobs.

(Ide & Suderman 2005: sw2236)

(2) Mandinka (Mande > Manding)

**Ref:** **Kuuráŋ**-o mǎŋ díyaa.  
**sick**-DEF PFV.NEG pleasant  
‘Sickness is not pleasant.’

(Creissels 2017: 46)

**Pred:** Díndíŋ-o mǎŋ **kuraŋ**.  
child-DEF PFV.NEG **sick**  
‘The child is not sick.’

(Creissels 2017: 46)

(3) Mundari (Austroasiatic > Munda)

**Ref:** **buru**=ko bai-ke-d-a.  
**mountain**=3PL.SUBJ make-COMPL-TR-IND  
‘They made the mountain.’

(Evans & Osada 2005: 354)

**Pred:** saan=ko **buru**-ke-d-a.  
firewood=3PL.SUBJ **mountain**-COMPL-TR-IND  
‘They heaped up the firewood.’

(Evans & Osada 2005: 355)

(4) Nuuchahnulth (Wakashan > Southern Wakashan)

**Ref:** watqšičł ʔaλimt ...  
watq-ši(λ) **ʔaλa**-imt ...  
swallow-MOM **two**-PAST ...  
completely.swallowed two ...  
‘He swallowed two of them [...].’

(Louie 2003: Qawiqaałth 57)

**Pred:** wikaλ haʔukšičł ʔaλičičł  
wik-’aλ haʔuk-ši(λ) **ʔaλa**-’i-čičł  
not-FIN eat-MOM **two**-INCEP  
didn’t ate became.two  
‘He (Mink) didn’t eat them and the crabs became two.’

(Louie 2003: Mink 266)

**Mod:** hiitqyaapup ʔaλa q<sup>w</sup>ayačičł  
hił-tqya-’pi-up **ʔaλa** q<sup>w</sup>ayačičł  
there-back-MOM.CAUS **two** wolf  
put.on.the.back two wolf  
‘Two wolves put (the dead wolf) on their back.’

(Louie 2003: FoodThief 46)

(5) Quechua (Quechuan)

**Ref:** rikaška: **hatun**-(kuna)-ta  
I.saw **big**-(PL)-ACC  
‘I saw the big one(s).’

(Schachter & Shopen 2007: 17)

Pred: chay runa **hatun** (kaykan)  
 that man **big** is  
 ‘that man is big’

(Schachter & Shopen 2007: 17)

Mod: chay **hatun** runa  
 that **big** man  
 ‘that big man’

(Schachter & Shopen 2007: 17)

(6) Tongan (Austronesian > Polynesian)

Ref: na’e lele e kau **feḿé**  
 PAST run SPEC PL.HUM **woman**.DEF  
 ‘The women were running.’

(Broschart 1997: 134)

Pred: na’e **feḿé** kotoa e kau lelé  
 PAST **woman** all SPEC PL.HUM run.DEF  
 ‘The ones running were all female.’

(Broschart 1997: 134)

(7) Central Alaskan Yup’ik (Eskimo-Aleut > Yup’ik)

a. *iqa-* ‘dirt’; ‘be dirty’  
*-ngtak* ‘very’  
 Ref: ***iqa-ngtak*** ‘one that is very dirty’  
 Pred: ***iqa-ngtaq-*** ‘be very dirty’

(Mithun 2017: 159)

b. *tangerr-* ‘see’  
*-uaq* ‘imitation, inauthentic’; ‘pretend to, without serious purpose’  
 Ref: ***tangerr-uaq*** ‘movie, vision, hallucination’  
 Pred: ***tangerr-uar-*** ‘hallucinate, watch a movie’

(Mithun 2017: 159)

c. *iqeq-* ‘corner of mouth’  
*-mik* ‘thing held in one’s mouth’; ‘to put in one’s’  
 Ref: ***iq-mik*** ‘chewing tobacco’  
 Pred: ***iq-mig-*** ‘put in one’s mouth’

(Mithun 2017: 160)

In the English example in (1), the predicative use of *paint* takes the English Past Tense suffix *-ed* like any prototypical verb in English, but there is no morpheme present that explicitly converts the stem from noun to verb (or vice versa). The remaining examples illustrate the same situation for a variety of language families around the world. Even though in some cases there is inflectional morphology indicating the function of the word, none of these examples have explicit derivational morphology converting the target lexical items from one function



to another.

Flexible items like those in the examples above create an analytical problem for traditional theories of parts of speech. Traditional theories assume that lexical items can be partitioned into mutually exclusive categories based on 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 items would seem to violate this assumption because they appear to be members of more than one category at once, and the criteria for classifying them yield conflicting results.

Researchers have proposed numerous solutions to this problem. One response is to analyze different uses of a putatively flexible item as instances of *heterosemy*—a special case of homonymy in which two lexemes share the same form but belong to different word classes (Lichtenberk 1991). In this view, heterosemous items are related only historically, via a process of conversion or functional shift, in essence denying the existence of lexical flexibility (Evans & Osada 2005). However, this perspective fails to answer why functional shift is rampant in some languages but not others, or why it happens to some lexemes but not others, or what motivates this functional conversion. Moreover, as will be discussed in [Section 2.3.3.4](#), it is difficult to maintain a principled distinction between polysemy and heterosemy. Semantic, distributional, and formal similarity between words are continua, meaning that questions like “are uses X and Y of a form instances of the same or different lexemes?” cannot be answered categorically. Questions about multifunctional uses of the same form—call it lexical flexibility, conversion, or something else—merit empirical investigation irrespective of one’s analytical position on the matter.

A more common approach to lexical flexibility 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 (M. C. Baker [2003]; Dixon [2004]; Palmer [2017]; Floyd [2011] for Quechua; Chung [2012] for Chamorro). Another approach is to say that languages exhibiting flexibility have only some of the traditional categories. A notable example of this is Launey’s (1994; 2004) analysis of Classical Nahuatl, which he calls

an *omnipredicative* language. In this analysis, all lexical items are predicates, so there is just one giant class of verbs.

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 class such as “flexibles”, “contentives” or “non-verbs”, etc. (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 by some researchers (see Simon [1937], McDonald [2013], and Sun [2020] for discussions of early analyses of Mandarin; Gil [1995] for Tagalog; Broschart [1997] for Tongan; Gil [1994] for Riau Indonesian; Kastovsky [1996] for Proto-Indo-European). Within generative linguistics, the Distributed Morphology framework takes it as an assumption that all roots are category-neutral (Siddiqi 2018). Farrell (2001) argues that *all* instances of flexible items (which he describes as cases of “functional shift”) involve roots underspecified for category.

Note that these differences in perspective do not arise from disagreements about the empirical facts. 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 (Wetzer 1992: 235; Stassen 1997: 32; Croft & van Lier 2012: 58). Examples of contested languages include those of the Iroquoian family (Chafe 2012), Mundari (Evans & Osada 2005; Hengeveld & Rijkhoff 2005), Quechua (Schachter & Shopen 2007: 17; Floyd 2011), and Sundanese (Robins 1968: 352; Hardjadibrata 1985: 62–63), with many others that could be cited as well. It is rare that an argument for flexibility is refuted by linguistic facts alone (though see Mithun’s [2000] response to Sasse [1988] regarding Cayuga).

Since analyses of lexical flexibility depend more on the theoretical commitments of the researchers involved rather than any 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* (Croft 2001b: 30), 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 (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 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 (Hieber forthcoming), every language surveyed exhibits lexical flexibility in at least some area of the grammar (although not all authors analyzed these cases as such). In my own experience studying 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 lexical items, however marginally. The prevalence with which different areas of the grammars of the world's languages lack sensitivity to the distinctions between reference, predication, and modification suggests that the existence of lexical categories in a language is not necessarily a given (Hieber forthcoming).

Given what we know from both cognitive science and diachronic linguistics, it would be surprising if clear-cut categories *did* exist. Cognitive science tells us that mental categories, word meanings, and lexical categories are all prototypical<sup>3</sup> (Taylor 2003). What it means for a category to be *prototypical* is that category membership is graded so that some members of the category are perceived as better representatives of that category than others. The prototypical meaning or concept within a category is the one that speakers conceive of as the most basic. The fact that mental categories are prototypical leads to various *prototype effects* in

---

<sup>3</sup>In this thesis, I use the term *prototypical* to mean 'having the properties of the prototype, exemplar, or central member of a category' and the term *prototypal* to mean 'having a prototype structure, with central and less central members'. The term *prototypal* is borrowed from the programming community, where it is used to describe programming languages (such as JavaScript) in which objects inherit properties from shared prototypes. Word classes may be described as prototypal, and their members as prototypical or non-prototypical.

both everyday life and language. More prototypical members of a category are learned earlier in development and acquisition, are used more frequently, can be recalled more quickly, are more likely to be represented using a simple lexical item rather than a complex word or compound, and are more strongly primed by the name of the category itself (Croft & Cruse 2004: 78–79). Exactly which of these observed effects best picks out the most prototypical meaning of a category is an open question and an area of active research (Gries 2006: 75; Gries & Divjak 2009: 58–59). Regardless, given the prototypical nature of mental categories, it would be quite surprising if lexical categories did not also exhibit prototype effects.

We also know from diachronic linguistics that language change is both gradual and gradient (Hopper & Traugott 2003; Traugott & Trousdale 2010). At any given point in time a lexical item might be in a stage of transition or expansion from one category into another, meaning that it will show attributes of both. Likewise, languages develop constructions dedicated to signaling the discourse functions of reference, predication, and modification over time, but at any given point in time, a language may have few or many of these constructions, and they may be at various stages of development (Vogel 2000). 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 a problem as it is a design feature of language. It is precisely the liminal categorial<sup>4</sup> status of flexible items 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. (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

---

<sup>4</sup>In this thesis, I use the term *categorial* to mean ‘without exception; unconditional’ and the term *categorical* to mean ‘having to do with categories’.

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 lexical items may be used for multiple functions, and corpus-based studies which examine whether and how often lexical items are used for multiple functions in discourse.

Cannon (1985) is an early lexicon-based study of functional shift in the history of English. Functional shift became an especially common pattern of word formation in early Middle English as inflectional paradigms were leveled (Cannon 1985: 414). Cannon examines 13,805 lexical items from three English dictionaries with etymological information, and finds that just 541 entries (3.92%) were created via conversion. Conversion from noun > verb is the most common, adjective > noun conversion the second most common, and verb > noun conversion the third most common. The full results from the study are shown in Table 1.1.

Table 1.1: Types of conversion in English (Indo-European > Germanic)  
(Cannon 1985: 416)

from	to	count
noun	verb	189
adjective	noun	121
verb	noun	114
noun	adjective	77
verb	adjective	19
adjective	verb	11
adverb	adjective	10
Total		541

Another lexicon-based study, though not explicitly focused on lexical flexibility, is Croft’s (1984) study of categories of Russian roots (summarized in 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 §2.4).

In a study of Mundari, Evans & Osada (2005) conduct a dictionary analysis using a focused 105-entry sample as well as a larger 5,000-entry-sample. In the 105-entry sample, 74 stems (72%) could be used as either noun or verb. In the larger sample, 1,953 stems (52%) could be used as both noun and verb. The complete figures for the large sample are shown in Table 1.2. Evans & Osada argue on the basis of these data that, because not all the items in the Mundari lexicon are flexible, Mundari is *not* a flexible language. As with any whole-language typology, however, this is an oversimplification. To overlook the flexibility of these items ignores the behavior of a vast portion of the lexicon. It is exactly this flexible behavior which is of interest in this thesis. Evans & Osada’s study nonetheless constitutes an important contribution to our knowledge of the empirical extent of lexical flexibility across languages.

Table 1.2: Percentage of lexical items used as nouns, verbs, or both in Mundari (Austroasiatic > Munda) (Evans & Osada 2005: 383)

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

Mithun (2017: 163) also conducts a lexicon-based analysis of roots in Central Alaskan Yup’ik using Jacobson’s (2012) exhaustive dictionary, and shows that only a small minority of roots (12%) are flexible, and can be used as both nouns and verbs. The results of this study

are shown in Table 1.3. Mithun reports that the words in these groups cannot be characterized in any general or semantic way. 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 (Sadock 1999). The fixation with these marginal cases in the literature seems disproportionate to their actual type 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 1.3: Percentage of roots used as nouns, verbs, or both in Central Alaskan Yup’ik (Eskimo-Aleut > Yupik) (Mithun 2017: 163)

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

In summary, existing lexicon-based studies have yielded differing results, each contributing to our understanding of lexical flexibility, but there are still too few such studies to draw any general conclusions. Since lexicon-based studies report only type frequencies, we do not know whether the flexible lexemes in these studies account for a greater or lesser portion of tokens in a corpus.

Corpus-based studies of lexical flexibility are also scarce. In a study of the discourse functions of property words in English and Mandarin, Thompson (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 1.4. Some of the attributive adjectives reported in Table 1.4 have “anaphoric head nouns” (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 items. The study also does not discuss the extent to which *individual* lexical items 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.

**Table 1.4:** Distribution of functions of property words in English (Indo-European > Germanic) and Mandarin (Sino-Tibetan > Sinitic) (Thompson 1989: 253, 257)

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

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. However it is equally unsurprising that some languages encode property concepts using either referring or predicating constructions, since prototypical adjectives exhibit behavior related to both functions.

A similar study to Thompson’s is Croft’s (1991: §2.5) investigation of the frequencies with which different semantic classes of lexical items (object words, action words, and property words)<sup>5</sup> are used for different discourse functions (reference, predication, and modification) in four languages: Quiché Maya (Mayan), North Efate (Austronesian), Soddo (Austroasiatic), and Ute (Uto-Aztecan). In all four languages, the most frequent use of lexical items is when their discourse function aligns with their semantic class. Object words are most frequently used to refer, action words are most frequently used to predicate, and property words are most frequently used to modify. Together with data from morphological markedness, se-

<sup>5</sup>I use the terms *object word*, *action word*, and *property word* when referring to the semantic class of a word rather than its discourse function. Object words are object-denoting, action-words are action-denoting, and property words are property-denoting.



semantic shifts, and combinatorial possibilities, Croft takes this as evidence that these are the most prototypical discourse functions for those semantic classes. As with other prototype categories, then, lexical categories display prototype effects in grammar. This fact is a key component of Croft's typological-markedness theory of lexical categories, to be explained fully in [Section 2.4.2](#). Like Thompson's (1989) study, however, Croft's study does not tell use the distributions for individual lexemes. Additionally, Croft's data include cases of overtly marked uses of lexical items in non-prototypical functions, which would not be considered instances of lexical flexibility.

Finally, there are some studies which count the proportion of nouns vs. verbs. vs. adjectives in English texts (Hudson 1994; Polinsky & Magyar 2020). Again, the data are not disaggregated to the item level, so no firm conclusions can be drawn about the extent of lexical flexibility.

In sum, no existing studies examine the distribution of pragmatic functions for individual items and limit themselves to only flexible (morphologically unmarked) cases. To my knowledge, 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 items 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 items 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 (Croft 2000; 2005; Croft & van Lier 2012). This literature and its many findings are reviewed in [Section 2.3](#).

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 (2001b: 70) makes this point nicely:

Does English have too few N/V lexemes to qualify as a flexible N/V language? If not, then how many is enough? [...] 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? (Croft 2001b: 70)

Equally significant (and equally unknown) is whether there are any commonalities among lexical items or languages which exhibit greater flexibility than others. These questions are relevant even if one adopts the position that flexible uses of lexical items 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 that guide 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) and Nuuchahnulth (Wakashan > Southern Wakashan). 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 lexical items in English and Nuuchahnulth?

R2: Is there a correlation between degree of lexical flexibility and size of the corpus?

R3: Is there a correlation between degree of lexical flexibility for a lexical item and frequency (or corpus dispersion)?

R4: How do the semantic properties of lexical items pattern with respect to their flexibility?

I explore each of these questions from several angles. R1, “How flexible are lexical items in English and Nuuchahnulth?” is the core empirical focus of this thesis. To answer it, I count the frequency with which stems are used for each of the three functions of reference, predication, and modification in a corpus of spoken texts for each language. I annotated nearly 400,000 tokens of English and 9,000 tokens of Nuuchahnulth for their discourse function. Based on these data, each stem is then given a flexibility rating from 0 to 1 based on how evenly its uses are distributed across the three function, computed using a normalized Shannon diversity/entropy index (Shannon 1948). A rating of 0 indicates that the stem is highly inflexible, with all its occurrences being used for a single function; a rating of 1 indicates that the stem is maximally flexible, with its occurrences evenly distributed across the three functions. By quantifying the flexibility of each stem in this way, it then becomes possible to look for statistical correlations between the flexibility of a stem 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 and size of the corpus?”, is motivated by claims made by some researchers that all items exhibit flexibility if you examine enough of their tokens (Mosel & Hovdhaugen 1992: 77). If true, this would lend some empirical support to the claim that all items are (or least can be) to some degree flexible, or perhaps even acategorical.

R3, “Is there a correlation between degree of lexical flexibility for a lexical item and frequency (or corpus dispersion)?”, uses the flexibility ratings calculated in R1 to consider whether the flexibility of a stem correlates with either its overall frequency or with its corpus

dispersion. *Corpus dispersion* refers to how evenly/regularly the item appears in a corpus, a measure which is thought to more accurately capture the notion of frequency of exposure (Gries 2008; forthcoming). This question has two motivations: First, higher-frequency items often preserve irregular or atypical forms or functions (Bybee 2007: Ch. 13), such that items with higher frequencies might be more likely to retain their non-prototypical, flexible uses. Second, the fact that a lexical item is flexible means that there is a wider range of constructions it can appear in. This could reasonably result in a higher overall frequency for flexible items. Both of these potential factors invite inquiry into the relationship between frequency and flexibility.

R4, “How do the semantic properties of lexical items 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) lexical items. This research question is directly motivated by Croft’s (1991; 2000; 2001b; forthcoming) typological markedness theory of lexical categories, which claims among other things that lexical items 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 (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 shifts further (Croft 2005: 440; Croft & van Lier 2012: 70), but as yet little research has responded to this call (though see Rogers [2016] and Mithun [2017]). 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 a high-level summary of the principal research questions investi-

gated 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 (2010b). 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 §1.2, 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. Thus some researchers may choose to view this study as an empirical investigation into the frequency of conversion in languages rather frequency or degree of lexical flexibility.

While the methods used in this study are compatible with a variety of theories of lexical flexibility, I nonetheless argue in [Chapter 2](#) for a cognitively informed, typological-constructional theory of word classes and flexible items. It is cognitively informed in that it treats mental categories as *prototypical* and recognizes the existence of various prototype effects in language. I also adopt a Radical Construction Grammar approach (Croft 2001b) in which the basic categories in language are *constructions* rather than *parts of speech* (see also [Langacker 1987; Fillmore, Kay & O'Connor 1988; Goldberg 1995; 2006]). In construction grammar, language is viewed as a structured taxonomic network of constructions, whether those constructions are *substantive* (like words and morphemes) or *schematic* (like grammatical relations).

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 most 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 (Crystal 1967: 47–48; Vonen 1994; Croft 2000: 75–76; 2001b: 69; Farrell 2001; Cannon 1985) and a fairly rigid language with clearly-delineated categories (Rijkhoff 2007: 710; Schachter & Shopen 2007: 4, 11, 12; Velupillai 2012: 122, 126). 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 items in English 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-token corpus of American English comprising numerous genres of both spoken and written data, all of which is open access (Ide & Suderman 2005). This study uses just the spoken portion of the corpus, consisting of approximately 3.2 million tokens, which is itself composed of two distinct subcorpora—the [Charlotte Narrative & Conversation Collection](#) (or simply “the Charlotte corpus”) and the [Switchboard Corpus](#).

Nuuchahnulth (formerly referred to in the literature 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 widely discussed languages in the literature on lexical flexibility (Swadesh 1939b; Jacobsen 1979; Braithwaite 2015). This is due largely to the following examples of flexible items from Swadesh (1939b).

Nuuchahnulth (Wakashan > Southern Wakashan)

- (8) a. qoʔas-ma      ʔi-h-ʔi  
           man-3SG.IND    large-DEF  
           ‘The large one is a man.’ (Swadesh 1939b: 78)
- b. ʔi-h-ma      ʔoʔas-ʔi  
           large-3SG.IND    man-DEF  
           ‘The man is large.’ (Swadesh 1939b: 78)
- (9) a. mamok-ma      ʔoʔas-ʔi  
           work-3SG.IND    man-DEF  
           ‘The man is working.’ (Swadesh 1939b: 78)
- b. ʔoʔas-ma      mamok-ʔi  
           man-3SG.IND    work-DEF  
           ‘The working one is a man.’ (Swadesh 1939b: 78)

Hardly a single typological survey of lexical categories or study of lexical flexibility has failed to include these examples since (see especially the much-cited chapter by [Schachter & Shopen](#) [[1985] 2007: 12]). 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 Chimakuan, Chinookan, Coosan, Sahaptian, Salishan, and Tsimshianic 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 (Chimakuan: Andrade [1933: 179]; Chinookan: Duncan, Switzler & Zenk [forthcoming]; Coosan: Frachtenberg [1922: 318]; Sahaptian: Wetzer [1996: 142]; Salishan: Kuipers [1968], Hébert [1983], Kinkade [1983], van Eijk & Hess [1986], Jelinek & Demers [1994], Matina [1996], Beck [2002: §4.1.1], Montler [2003], Beck [2013], Davis, Gillon & Matthewson [2014]; Tsimshianic: Davis, Gillon & Matthewson [2014]; Kutenai: Morgan [1991]). 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 come from a corpus of texts collected

and edited by Toshihide Nakayama and published in Little (2003) and Louie (2003). The corpus consists of 24 texts dictated by speakers Caroline Little and George Louie, containing 2,081 utterances and 8,366 tokens (comprising 4,216 distinct wordforms). The texts cover a variety of genres, including procedural texts, personal narratives, and traditional stories. I manually retyped these texts as [scription](#) files for analysis. Scription is a simple text format for representing interlinear glosses in a way that is both familiar to linguists and computationally parseable (Hieber 2021a). 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 (Austronesian > Malayo-Polynesian) (Gil 1994), Mundari (Evans & Osada 2005; Hengeveld & Rijkhoff 2005), Classical Nahuatl (Uto-Aztecan) (Launey 1994; 2004), and Central Alaskan Yup'ik (Eskimo-Aleut > Yupik) (Thalbitzer 1922; Sadock 1999; Mithun 2017). 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 to the [Data Format for Digital Linguistics](#) (DaFoDiL) (a JSON format for representing linguistic data; Hieber [2021b]) 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>.

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 lexical items 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 items surveyed are



flexible, but only to a small degree. Most lexical items of English can be used as referents, predicates, or modifiers, but there is a strong tendency for each item to be used for primarily one function. English thus shows a consistent but somewhat marginal degree of flexibility. In contrast, most lexical items in Nuuchahnulth are highly flexible, but primarily along the reference-predication axis; Nuuchahnulth lexical items are very freely used for both reference and predication, but only infrequently used as modifiers. Nuuchahnulth thus shows a consistently high degree of flexibility, but primarily in just one dimension.

finalize the findings on flexibility vs. corpus size, especially the reliability size

In relation to R2, “Is there a correlation between degree of lexical flexibility and size of the corpus?”, I find that there is indeed a strong correlation between the two in English for sample corpus sizes up to approximately 100,000 words, after which the relative increase in flexibility drops off precipitously. In other words, a corpus of at least 100,000 words provides a fairly reliable representation of the actual flexibility of its words. The data for Nuuchahnulth are consistent with the findings for English, but the overall corpus size for Nuuchahnulth is too small to say with confidence that the same findings hold. The point of diminishing returns on flexibility for Nuuchahnulth could be quite different from that of English.

For R3, “Is there a correlation between degree of lexical flexibility for an item and frequency (or corpus dispersion)?”, I find that higher frequency items are more flexible than lower frequency items, but that the effect is very small. The same facts hold when comparing degree of lexical flexibility with corpus dispersion. Items 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 an item does depend in part on how regularly speakers use it.

Lastly, R4 asks “How do the semantic properties of lexical items 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 items. Nearly all of the most

flexible items denote property concepts. Deictic expressions such as *this*, *that*, *here*, *there* also rank very highly 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 (1984) claim that items used in their prototypical function will show the inflectional behaviors typical of that function, and Croft's (1991) behavioral potential hypothesis. 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 conceptualize or construe concepts—as an action, object, or property—as has been suggested by Croft (1991: 99; 2001b: 104).

Nuuchahnulth also has a definite suffix *-ʔi:* used with referents. Nakayama (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 an action word is used for the atypical role of reference (as predicted by Croft's structural coding hypothesis; see §2.4 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 an item (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 tokens 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 Nuuchah-

nulth to be, 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 lexical items in a corpus that can be replicated for other languages and corpora ([Chapter 3](#)). The second is empirical and descriptive: I describe the extent of lexical flexibility and the way it operates in English and Nuuchahnulth ([Chapter 4](#)). The final contribution 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). Lexical items used for functions that are not prototypical of their meaning *tend* to be more marked (morphologically, behaviorally, semantically, and/or frequently) 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 items. 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 methodological techniques from [Chapter 3](#) are ap-

plied to individual lexical items, 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

The focus of this chapter is to explain the concept of lexical flexibility, consider its criticisms, and offer a more robust, functionally grounded definition instead. I first briefly describe how flexible approaches to lexical categories developed as a response to weaknesses in traditional theories of parts of speech. I then survey the landmark studies and important findings on lexical flexibility, along with criticisms of this research. Following that, I present the typological markedness theory of lexical categories, which states that lexical categories are epiphenomenal markedness patterns regarding how different semantic classes of words are used for different discourse functions. I conclude by offering a revised formulation of lexical flexibility which is in line with typological markedness theory.

### 2.1 Introduction: Approaches to lexical flexibility

The field of linguistics as a whole and the subfield of typology in particular is undergoing a radical shift in how we understand lexical categories, along primarily two dimensions. The first dimension is our understanding of what lexical categories are a property *of*. Early researchers viewed categories as universal properties of both language generally and specific languages. I call this the *universalist* position. After Boas, many researchers then came to view categories as language-specific, with patterned similarities across languages. I call this the *relativist* approach. Most recently, some researchers view categories as typological patterns rather than properties of any particular language. This is the *typological* position, and

the one I adopt here.

The second dimension of historical change in linguistic theories of categories is in the *nature* of the categories themselves. In the Classical tradition, categories were thought to be categorical and well-defined by a set of necessary and sufficient conditions in the tradition of Aristotle. After the cognitive turn in the 1960s and 1970s, many linguists came to view categories as prototypical, with some members of a category being more central, or better exemplars, than others. Cognitive research into the nature of idioms then led to the development of construction grammar, which sees language as consisting of a network of constructions rather than monolithic categories. I adopt a constructional approach to categories in this thesis.

These theoretical paradigm shifts are summarized in (10). At each stage of development, there has not been a wholesale displacement of previous theories. There are still many who regard word classes as universal and categorical, and the typological-constructional approach is still nascent.

- (10) a. universal > language-specific > typological  
b. categorical > prototypical > constructional

[Section 2.2](#) gives a synopsis of these theoretical positions and shows how research on lexical flexibility developed in recognition of the shortcomings of traditional approaches. [Section 2.3](#) summarizes the key concepts and findings that have arisen from the research on lexical flexibility. Such research, however, is not without its own shortcomings. [Section 2.3](#) also presents the main criticisms that have been leveled against flexible analyses of word classes. [Section 2.4](#) then presents an alternate, functionally-oriented approach—the typological-constructional perspective. The final section of this chapter ([§2.5](#)) then applies this functional perspective to formulate an improved definition of lexical flexibility.

## 2.2 Traditional approaches

This section is a necessarily brief history of approaches to lexical categories up until the cognitive turn of the 1960s. It covers the universalist position that developed in the Classical tradition, the relativist position that developed as a result of Boas’ cultural relativism, and the structuralist (or “distributionalist”) position that developed in the tradition of Saussure. Depending on how one understands and applies these different perspectives, none of them are mutually exclusive. It is especially common for linguists to simultaneously hold that lexical categories must be identified on the basis of language-internal evidence alone (the relativist position) and that lexical categories are universal in some sense or another (the universalist position).

### 2.2.1 Universalism

Historically and still presently, many researchers assumed that a small set of lexical categories are basic and universal to all languages (Bolinger & Sears 1981: 81; Croft 1991: 2; Payne 1997: 32; Stassen 2011: 95). The set typically consists of some variation of the following: Noun, Verb, Adjective, Adverb, Pronoun, Adposition, Conjunction, Numeral, and Interjection (Haspelmath 2001: 16538). This list has its origins in the *Τέχνη Γραμματική* / *Tékhnē Grammatiké* (‘The art of grammar’) of the 2<sup>nd</sup> century B.C.E. grammarian Dionysius Thrax. The *Tékhnē* synthesizes the work of Dionysius’ predecessors, describing eight parts of speech for Classical Greek. These parts of speech were based largely on morphological (especially inflectional) criteria (Rauh 2010: 17–20). The *Tékhnē* was then translated and its model applied to Latin in the *Ars Grammatica* of Remnius Palaemon. The *Ars Grammatica* initiated a tradition wherein the languages of Europe and eventually the world (e.g. Mandarin [McDonald 2013]) were described using both Dionysius’ categories (with occasionally additions / subtractions) as well as his method of identifying those categories on the basis of morphological criteria (Rauh 2010: 20). Because of the strong association of the term *parts of*

*speech* with this Classical perspective, I prefer the term *lexical categories* in this thesis.

Implicit in the Classical method is the assumption that lexical categories are universal in the sense of being instantiated in all languages. However, as European scholars began to encounter non-Indo-European languages (or even non-Romance languages) in both Europe and abroad, this assumption was challenged, as early as the first grammatical descriptions of Irish in the 7<sup>th</sup> century. At first, these languages either had Classical grammar imposed upon them or were deemed grammatically deficient (Suárez 1983: 3). Nonetheless, missionary linguists in the early colonial era were aware of the significant grammatical differences between these languages and Latin and made their best attempts at describing them (Suárez 1983: 3–4). It is also important to realize that the project of describing the languages in the Americas and other zones of colonial influence was partially contemporaneous with the publication of the first grammars of the vernacular languages of Europe, as illustrated in Figure 2.1 (the data for which are given in Appendix A). Between 1524 and 1572, over 100 catechisms, manuals for confession, collections of sermons, grammars, and vocabularies were written in or about ten languages within the Viceroyalty of New Spain alone (an area smaller than present-day Mexico), mostly by Spanish Franciscan and Jesuit missionaries (Suárez 1983: 2). The task of converting the indigenous peoples to Christianity via the medium of their own languages was so important to the Spanish crown that the first bishop of Mexico, Francisco de Zumárraga, brought a printing press to Mexico in 1534 (just 15 years after the arrival of the first Spaniards in Mexico in 1519). The first book printed in Mexico was a Spanish-Nahuatl catechism by Alonso de Molina (Suárez 1983: 2). All this is merely to illustrate that language scholars in the colonial era were still in the early stages of discovering the complexities of the world's languages and how much they differed from Latin and Greek. Nonetheless, they were aware of the challenges that non-Indo-European languages posed to Classical theories at an early stage.

As documentary linguistics turned its attention to North American (as opposed to Mesoamerican) languages, lexical flexibility in particular became a more prominent issue. In fact, even



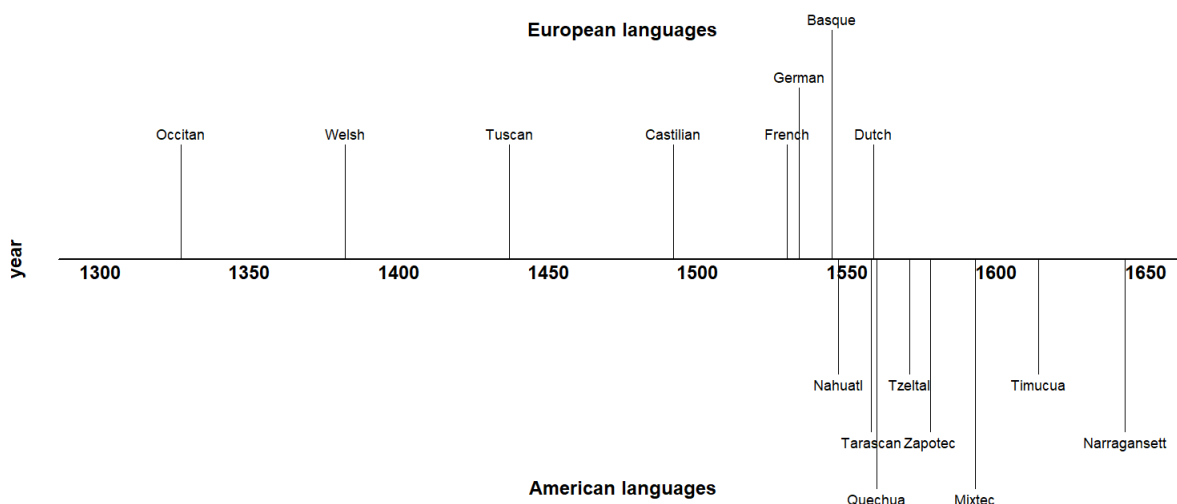


Figure 2.1: Timeline of early grammatical descriptions of European vs. American languages

the first comprehensive survey of North American languages contains an entire section on “Conversion of nouns into verbs” (Gallatin 1836: 174–177), in which Gallatin depicts lexical flexibility as a rampant feature of all languages on the continent:

It is the substantive [i.e. copula / auxiliary] verb which we [speakers of Indo-European languages] conjugate; whilst the [Native American] conjugates what we call the adjective and even the noun itself, in the same manner as [s/he] does other intransitive verbs. [...] I believe it must appear sufficiently obvious, that this general if not universal character of the [Native American] languages, the conversion into verbs and the conjugation, through all the persons, tense, and moods, of almost all the adjectives and of every noun which, without a palpable absurdity, is susceptible of it, is entirely due to the absence of the substantive verb. (Gallatin 1836: 175–176)

As evidenced by the above passage, increasing familiarity with non-Indo-European languages prompted some writers to abandon the universalist commitment. However, categorial universalism is still a widely-held position today, either in the sense of a) categories being universally instantiated in all languages (commonly assumed by most generative frameworks; although see Culicover [1999]), or b) categories being available to all languages, but only instantiated in some (sometimes called the “smörgåsbord” or “grab bag” approach, as exemplified by Dixon’s Basic Linguistic Theory framework [2010: 9, 11, 14, 27, 50; 2011: 26]; see

Hieber [2013: 298] and Croft [2001b: 10] for discussion).

### 2.2.2 Relativism

American ethnographers in the tradition of Franz Boas questioned the universalist assumption in a programmatic and comprehensive way. Writing on grammatical categories, Boas states, “Grammarians who have studied the languages of Europe and western Asia have developed a system of categories which we are inclined to look for in every language” (Boas 1911: 35). He concludes that this endeavor is a folly, and that “in a discussion of the characteristics of various languages **different fundamental categories** will be found” (Boas 1911: 35). Boas’ students all adopted his grammatical relativism, and it became a foundational principle of the American linguistics tradition. His student Edward Sapir, writing on lexical categories specifically, makes one of the best-known and strongest statements of this position in his influential textbook *Language*: “[N]o logical scheme of the parts of speech—their number, nature, and necessary confines—is of the slightest interest to the linguist. Each language has its own scheme. Everything depends on the formal demarcations which it recognizes.” (Sapir 1921: 125).

Many linguists today hold to Boas’ grammatical relativism in some fashion or another. Textbooks and typological surveys commonly state that languages have varying numbers of lexical categories, though usually with the caveat that all languages seem to differentiate at least noun and verb (e.g. Velupillai 2012: §6.2). Some researchers, especially those working in typology, argue that linguists are still not rigorous *enough* in their application of grammatical relativism; they criticize certain kinds of crosslinguistic comparisons for imposing the categories of one language onto another (Croft 2001b; Gil 2001; Haspelmath 2010a; 2012; LaPolla 2016). This position is discussed further in §2.4.

### 2.2.3 Structuralism

Developing alongside the early anthropological linguistics of Boas was the linguistic structuralism of Ferdinand de Saussure. His work informed both the Prague school under Nikolay Trubetzkoy and Roman Jakobson, and the distributional method of Leonard Bloomfield. The term *structuralism* has any number of uses (P. Matthews 2001: Ch. 1); here I refer to the idea that “language is a [...] self-contained, self-regulating system, whose elements are defined by their relationship to other elements” (P. H. Matthews 2014: 383). In particular, I am referring to the positivistic flavor of structuralism as practiced by Bloomfield, which focused on the structural relations between elements, and on establishing a set of rigorous scientific discovery procedures for linguistic structures (Bloomfield 1933). Bloomfield saw lexical categories as something to be empirically discovered in the different syntactic distributions of words, rather than imposed on a language a priori (Rauh 2010: 33). Zellig Harris later refined and expanded this methodology (Harris 1951), which was in turn incorporated into Chomsky’s early Phrase Structure Grammar (Chomsky 1957).

The signature methodological feature of this form of structuralism is the *distributional method*, a procedure for defining categories in terms of the set of contexts in which its words can appear—that is, their distributions (Harris 1951: 5; Croft 2001b: 11). As an illustration of distributional analysis applied to lexical categories, Croft (1991: 11–12) considers the distributions of the English words *cold*, *happy*, *dance*, and *sing* in two constructions: in the Predicate construction after *be*, and in the 3<sup>rd</sup> Person Singular Present Tense (-s) construction. Example data are shown below.

(11) English (Indo-European > Germanic)

- a. i. Jack is cold.  
ii. \* Jack colds.
- b. i. Jack is happy.  
ii. \* Jack happies.
- c. i. \* Jack is dance.

- ii. Jack dances.
- d. i. \*Jack is sing.
- ii. Jack sings.

We can see that *cold* and *happy* have the same distributions in these tests (both may appear in the Predicate construction but not the Person-Tense inflection construction), while *dance* and *sing* have the same distribution (the inverse situation as *cold* and *happy*). The results of these two distributional tests are summarized in [Table 2.1](#). Data like these are used to justify categories like Adjective and Verb in English.

**Table 2.1:** Distribution of English Verbs and Adjectives (adapted from Croft [2001b: 12])

	Predicate Construction	Inflectional Construction
<b>Adjective:</b> <i>cold, happy</i> , etc.	✓	✗
<b>Verb:</b> <i>sing, dance</i> , etc.	✗	✓

As applied in practice, however, the distributional method suffers from one serious drawback when used to argue for large, traditional categories like noun, verb, and adjective: distributional tests yield conflicting and overlapping results. Perhaps no two lexical items behave the same way in every distributional test. Each new test that is introduced therefore partitions the lexicon into smaller and smaller classes. This fact has been demonstrated empirically for English temporal nouns (Crystal 1967: 54), Russian numerals (Corbett 1978), and French verbs (Gross 1979). Distributional tables from each of these studies are reproduced in [Table 2.4](#), [Table 2.3](#), and [Table 2.2](#) respectively. It is clear from these studies that distributional analysis does *not* lead to large, unified categories like noun, verb, and adjective, but rather a myriad of small constructions (Crystal 1967: 27; Croft 2005: 434). Each distributional test is in fact its own construction (Croft 2005: 436). This fact is a major motivation underlying constructional approaches to language.

Many scholars nonetheless choose to retain lexical categories as a necessary component of their linguistic theories or descriptions, at the expense of consistent application of the dis-

	in a <i>N</i> or two	in that <i>N</i>	in the <i>N</i> <sub>sg.</sub> (no postmodification) <sup>65)</sup>	in a <i>N</i> (no postmodification) <sup>65)</sup>	in $\emptyset$ <i>N</i> <sub>pl.</sub> (no postmodification) <sup>65)</sup>	in $\emptyset$ <i>N</i> <sub>sg.</sub>	$\emptyset$ <i>N</i> <sub>pl.</sub> on the <i>N</i> <sub>sg.</sub>	on a <i>N</i> <sub>sg.</sub> (no postmodification) <sup>65)</sup>	on $\emptyset$ <i>N</i> <sub>pl.</sub> (no postmodification) <sup>65)</sup>	on $\emptyset$ <i>N</i> <sub>sg.</sub>	at that <i>N</i>	at the <i>N</i> <sub>sg.</sub>	at $\emptyset$ <i>N</i> <sub>sg.</sub>
afternoon	+	+	+	+	-	-	+	+	+	-	-	-	-
evening	+	+	+	+	-	-	+	+	?+	-	-	-	-
weekend	+	+	?+	+	-	-	+	+	?+	-	+	+	-
night	+	+	+	?+	-	-	?-	+	-	- <sup>66)</sup>	-	-	+
morning	+	+	+	+	-	-	+	+	-	- <sup>66)</sup>	-	-	-
Monday...	+	-	-	-	-	-	+	+	+	+	-	-	-
January...	?+	+	+	+	-	+	-	-	-	-	-	-	-
hour	+	+	+	+	?-	-	-	+	-	-	+	+	-
minute	+	+	-	+	+	-	-	+	-	-	+	+	-
second	+	+	-	+	+	-	-	+	-	-	+	+	-
day	+	+	+	+	-	-	-	+	-	- <sup>66)</sup>	-	-	-
summer	?+	+	+	?-	?--	+	?+	-	-	-	-	-	?+
winter	?+	+	+	?-	?--	+	?+	-	-	-	-	-	?+
spring	?+	+	+	?-	-	+	-	-	-	-	-	-	?+
autumn	?+	+	+	?-	-	+	-	-	-	-	-	-	?+
month	+	+	+	+	+	-	-	-	-	-	-	-	-
week	+	+	+	+	+	-	-	-	-	-	-	-	-
year	+	+	+	+	+	-	-	-	-	-	-	-	-
decade	+	+	-	+	?+	-	-	-	-	-	-	-	-
century	+	+	-	+	?-	-	-	-	-	-	-	-	-
fortnight	+	+	+	+	-	-	-	-	-	-	-	-	-
instant	+	?+	-	+	-	-	-	-	-	-	+	-	-
moment	+	?+	-	+	-	-	-	-	-	-	+	-	-
lifetime	-	?+	-	+	-	-	-	-	-	-	-	-	-
daytime	-	-	+	-	-	+	-	-	-	-	-	-	-
nighttime	-	-	+	-	-	+	-	-	-	-	-	-	-

Table 2.2: distributional analysis of English (Indo-European > Germanic)  
temporal nouns (Crystal 1967: 54)

tributional method. Rather than considering all possible distributional contexts for a word, these scholars instead treat certain constructions as definitional of the category. Other distributional tests which yield cross-cutting results are either ignored or treated as evidence

	<i>odin</i> 1	<i>dvo</i> 2	<i>tri</i> 3	<i>pjat'</i> 5	<i>sto</i> 100	<i>tysjača</i> 1,000	<i>million</i> 1,000,000
1. Agrees with N in syntactic number	+	–	–	–	–	–	–
2. Agrees in case throughout	+	–	–	–	–	–	–
3. Agrees in gender	+	(+)	–	–	–	–	–
4. Marks animacy	+	+	+	–	–	–	–
5. Has own plural	+	+	+	+	(–)	–	–
6. Takes agreeing determiner	+	+	+	+	+	–	–
7. Takes N in genitive plural throughout	+	+	+	+	+	±	–

Table 2.3: distributional analysis of Russian (Indo-European > Slavic) numerals (Corbett 1978: 359)

of subcategories instead of categories. Many researchers even prefer the term *syntactic categories* over *lexical categories* for this reason, focusing on just the syntactic evidence for categories (M. C. Baker 2003; Rauh 2010). A severe methodological problem for this approach is that there are no generally agreed-upon principles for determining which distributional tests should be considered definitional. In this regard, Schachter & Shopen (2007: 4) note, “there may be considerable arbitrariness in the identification of distinct parts of speech rather than subclasses” (see also Crystal [1967]). Different scholars choose or prioritize different kinds of evidence for lexical categories over others based on their theoretical commitments. This is the reason, as stated in Section 1.1, that disagreements about the existence of particular lexical categories in particular languages are typically *not* about the empirical facts. The results of a given distributional analysis are not usually controversial; the choice of distributional tests used to support one’s analysis are. Unsurprisingly, then, debates over how to analyze lexical categories in various languages have been largely unproductive and unresolved (Croft 2005: 435). The problem only worsens when scholars attempt to apply the same criteria across languages. Distributions of lexical items with similar meanings vary drastically across languages (Croft 2001b: §1.4.1).

i/ V Ω					Compléments directs ou indirects										Comp. indirect												
					Auxiliaire avoir	Auxiliaire être	N <sub>0</sub> est V <sub>pp</sub> Ω	N <sub>0</sub> V Prép N <sub>1</sub>		Complétives					Noms					N <sub>0</sub> V Prép N <sub>2</sub>		Noms					
										que P	que P <sub>subj</sub>	[pc z.]	V <sup>2</sup> Ω	ce (ci + la)	ppv	N <sub>hum</sub>	ppv	N <sub>hum</sub>	le fait Qu P			ppv	N <sub>hum</sub>	ppv	N <sub>hum</sub>	ppv	N <sub>0</sub> V V <sup>0</sup> Ω
+	s'agir	-	+	-	+	de	-	+	+	+	+	-	+	-	+	+	-	pour	+	-	-	-	-				
+	apparaître	-	+	-	+	0	+	-	-	-	+	+	-	-	+	-	+	-	à	+	+	-	-	-			
+	apparaître	-	-	-	-	0	-	+	-	-	-	-	-	-	-	-	-	-	0	-	-	-	-	-			
+	s'avérer	-	+	-	+	0	+	-	-	+	+	-	-	+	-	+	-	à	+	-	-	+	+	+			
+	y avoir avantage	+	-	-	+	à	-	+	-	+	+	-	-	-	-	-	-	+	pour	+	-	-	-	-			
+	y avoir lieu	+	-	-	+	de	-	+	+	+	-	-	-	-	-	-	-	pour	+	-	-	-	-	-			
+	n'empêcher	+	-	-	+	0	+	-	-	-	-	-	-	-	-	-	-	0	-	-	-	-	-	-			
+	être besoin	+	-	-	-	de	-	+	+	+	+	+	-	-	+	-	+	pour	+	-	+	-	-	-			
+	être l'heure	+	-	-	-	de	-	+	+	+	+	+	-	-	+	-	+	pour	+	-	-	-	-	-			
+	être question	+	-	-	+	de	-	+	+	+	+	+	+	-	+	+	+	pour	+	-	+	-	-	-			
+	être temps	+	-	-	+	de	-	+	+	+	+	+	-	-	+	-	+	pour	+	+	-	-	-	-			
+	faire bon	+	-	-	+	0	-	-	-	+	-	-	-	-	-	-	-	pour	+	-	-	-	-	-			
+	falloir	+	-	-	+	0	-	+	-	+	+	+	+	+	+	-	+	à	+	+	-	-	-	-			
+	s'en falloir	-	+	-	-	pour	-	+	-	-	-	-	-	-	-	-	-	de	-	-	+	-	-	-			
+	paraître	+	-	-	+	0	+	-	-	-	+	+	-	-	+	-	+	à	+	+	-	-	+	+			
+	paraître	-	-	-	+	0	+	-	-	-	-	-	-	-	-	-	-	0	-	-	-	-	-	-			
+	sembler	+	-	-	+	0	+	-	-	+	+	-	-	+	-	+	-	à	+	+	-	-	+	+			
+	souvenir	-	+	-	+	de	+	-	+	+	+	+	-	+	+	+	-	à	+	+	-	-	-	-			

Table 2.4: distributional analysis of French (Indo-European > Romance)  
verbs (Gross 1979: 860)

The real methodological problem here is *not* that we have yet to ascertain the correct principles for selecting the right distributional tests. The problem is being selective regarding which tests to apply in the first place. If we take the distributional method seriously, then we must apply it consistently, without ignoring distributional evidence that contradicts



our theoretical or pretheoretical assumptions. To do otherwise is a kind of *methodological opportunism* (Croft 2001b: 30, 41).

Other scholars treat flexible items as members of *hybrid* or *mixed* categories simultaneously possessing properties of more than one part of speech (Lois et al. 2017: 149; Malouf 1999; Nikolaeva & Spencer 2020). Adjectives are frequently described as a hybrid category (Wetzer 1996; Stassen 1997: 343; Pustet 2003: 13–16; Genetti & Hildebrandt 2004: 95; van Lier 2017), as are participles (Hopper & Thompson 1984: 704) and gerunds (Denison 2001). Lois et al. (2017: 149) also distinguish hybridity from polycategoriality, stating that polycategoriality applies to roots or stems, while hybridity is a matter of the syntactic context that a word appears in.

An analysis couched in mixed categories does not avoid the problem of methodological opportunism, however. The existence of a mixed category implies that there are other, more basic categories that the mixed category is a hybrid of. Hybrid models of parts of speech merely exacerbate the distributional problem. There is however a sense in which thinking of minor lexical categories as “mixed” categories is useful: typological markedness theory states that lexical categories are epiphenomenal patterns that arise from combinations of the semantic classes of object, action, or property words with the discourse functions of reference, predication, and modification. Categories frequently discussed as “mixed” in the literature are precisely those combinations which are non-prototypical and therefore more likely to be typologically marked. Section 2.4.2 explains this approach to lexical categories in more detail.

Partly in response to these problems, a growing cadre of linguists in the last 30 years have adopted one of various *flexible* approaches to word classes. Flexible analyses of word classes come in many flavors, some of which arguably still commit methodological opportunism, and others of which introduce new difficulties. These flexible approaches are reviewed in the following section.



## 2.3 Flexible approaches

In this section I summarize the key concepts (§2.3.1), findings (§2.3.2), and criticisms (§2.3.2) of lexical flexibility research. [Section 2.3.1](#) surveys the wide variety of definitions and theoretical perspectives on lexical flexibility. This review of the literature reveals that there is little consensus as to what exactly constitutes “lexical flexibility”; as such, there are numerous alternative terms for the phenomenon. Despite these incongruities, a few important findings do consistently surface across different empirical studies. These findings are summarized in [Section 2.3.2](#). Then [Section 2.3.3](#) looks at the arguments and evidence that researchers have presented against the notion of lexical flexibility, and some counterarguments.

### 2.3.1 Key concepts

It is only a small exaggeration to say that there are as many definitions and terms for what I am here calling “lexical flexibility” as there are scholars who research it. I use the term *lexical flexibility* in this thesis merely because it is the most widely recognized of the cluster of terms that are used, not because it is necessarily the most precise or accurate. My own choice would be *functional expansion*, for reasons discussed below. The analytical or theoretical perspective adopted by each researcher generally determines their choice of terminology. The remainder of this section is devoted to explaining these perspectives in more detail.

Generally speaking, there are two ways to analyze flexible items. The first method assigns flexible items to members of specific categories in a language, whether those categories are the canonical four major classes (Noun, Verb, Adjective, Adverb), or a new large supercategory subsuming multiple discourse functions (e.g. Contentives, Non-Verbs, Flexibles), or a smaller subcategory of an existing major lexical category (e.g. Adjectival Verbs, Verbonominals). The second method of analysis assumes that lexical items are uncategorized at some level (root, stem, or inflected word), and that items receive their categorial assignment from context. Different researchers posit different mechanisms for how lexical items receive their

categorization in context.

The traditional approaches to lexical flexibility summarized in [Section 2.2](#) are all instances of the former method of analysis, while the flexible approaches outlined in this section are a mix of categorial and acategorial analyses.

### 2.3.1.1 Lexical flexibility

Though awareness of lexical flexibility can be traced back to at least Gallatin ([1836](#): 174–177) if not earlier, the term *lexical flexibility* itself seems to have originated with Hengeveld ([1992](#): Ch. 4). This publication, perhaps because it was the first to assign a technical term to the concept, marks a shift in how scholars frame the concept of lexical flexibility. Previously, the issue was framed in terms of whether particular languages (especially those of the Pacific Northwest) distinguished noun from verb (Kuipers [1968](#); Jacobsen [1979](#); Hébert [1983](#); Kinkade [1983](#); van Eijk & Hess [1986](#); Jelinek & Demers [1994](#)). After this point, an increasing number of publications began to ask whether lexemes were *flexible* instead. Though the difference in emphasis seems subtle, this change constitutes a turning point because it fostered an increased interest in lexical flexibility as a grammatical phenomenon in its own right instead of just a problem for traditional categorization schemes.

Hengeveld’s ([1992](#): Ch. 4) typology of parts-of-speech systems is a whole-language typology wherein languages are either *specialized*, with one morphosyntactic category for each of the functions of reference (Noun), predication (Verb), referent modification (Adjective), and predicate modification (Adverb<sup>1</sup>), or *non-specialized*. Non-specialized languages deviate from the four-category canon in one of two ways: one part of speech may assume more than one function with no additional morphosyntactic marking, in which case the language is considered *flexible*; or the language may lack a dedicated part of speech for that function entirely and use other, marked constructions instead, in which case the language is considered *rigid*.

[Hengeveld](#) gives examples from Dutch and Wambon to illustrate the distinction between

---

<sup>1</sup>Note that Hengeveld’s typology only includes manner adverbs, not other semantic types of adverbs.

rigid and flexible languages. In the Dutch examples in (12), the same word *mooi* is used for both referent modification (12a) and predicate modification (12b), with no overt morphology indicating its function in either case. Wambon on the other hand uses medial verbs for manner expressions and must take the overt verbalizing suffix *-mo* shown in (13). In Hengeveld’s framework, Dutch is a flexible language because one category subsumes both the functions of referent modification and predicate modification, while Wambon is a rigid language because derivational morphology (here, the verbalizing suffix *-mo*) is required to indicate the function of predicate modification.

(12) Dutch (Indo-European > Germanic)

- a. een        **mooi**        kind  
       INDEF    **beautiful**    child  
       ‘a beautiful child’

(Hengeveld 1992: 65)

- b. het    kind    dans-t        **mooi**  
       DEF    child    dance-3SG.PRES    **beautifully**  
       ‘the child dances beautifully’

(Hengeveld 1992: 65)

(13) Wambon (Trans-New Guinea > Greater Awyu)

- jakhov-e        **matet-mo**        ka-lembo  
  they-CONN    **good-VZR.ss**    go-3PL.PAST  
  ‘did they travel well?’

(de Vries [1989: 49], cited in Hengeveld [1992: 65])

Hengeveld’s analysis is of the categorial type discussed at the beginning of Section 2.3.1, specifically the supercategory *kind*. Each lexeme is assumed to have a category, and new supercategories are introduced for lexemes which have multiple functions: *Contentives* for lexemes which perform all four functions, *Non-Verbs* for lexemes which perform all non-predicating functions, and *Modifiers* for lexemes which perform referent modifier and predicate modifier functions.

Hengeveld’s parts-of-speech typology and the subsequent research it inspired (Don & van Lier 2013; Hengeveld & Rijkhoff 2005; van Lier 2006; Hengeveld & van Lier 2012; Luuk 2010; van Lier & Rijkhoff 2013; van Lier 2016) constitute important empirical contributions to the study of lexical flexibility. However, Hengeveld’s definition of flexible languages and

his parts-of-speech typology still rely on large, language-specific categories of the kind that have been problematized by Croft (2001b: §2.2.2) and Croft & van Lier (2012), and are therefore subject to the same difficulties as traditional approaches to parts of speech. However, numerous scholars have since adopted Hengeveld’s term *lexical flexibility* to describe cases where lexical items serve more than one discourse function, regardless of their theoretical commitments or analysis of flexible items. As a convenient cover term, *lexical flexibility* is now well established.

### 2.3.1.2 Polycategoriality

Vapnarsky & Veneziano (2017b: 4) introduce the alternative term *polycategoriality* as their preferred characterization of flexible items. (The term is also used by Carter (2006), but he does not give a precise definition for it.) While Vapnarsky & Veneziano use this term mostly interchangeably with *lexical flexibility*, there are important differences between the two concepts. Hengeveld’s use of *lexical flexibility* is meant to imply the existence of large, flexible supercategories that subsume multiple discourse functions, whereas Vapnarsky & Veneziano are not committed to any particular schema for parts of speech. Central to their notion of polycategoriality is the idea that lexical categories exist, but that “there are lexical forms that are not specified for lexical category (or are not specified fully, or univocally) on some level of representation.” (Vapnarsky & Veneziano 2017b: 4). In other words, one lexeme may belong simultaneously to multiple lexical categories. Under this definition, a language could still have all four major lexical categories but nonetheless exhibit rampant polycategoriality; this is not a possibility in Hengeveld’s framework. Like Hengeveld, however, Vapnarsky & Veneziano are committed to the existence of large lexical categories in particular languages. Their analysis is therefore also of the categorial kind discussed at the beginning of Section 2.3.1.

### 2.3.1.3 Multifunctionality / Polyfunctionality

Another term for our phenomenon of interest, introduced by (van Lier 2012), is *multifunctionality*, in which a single lexical item can have multiple discourse functions. An advantage of this analysis is that it takes no theoretical position on the issue of whether lexical items are categorial or acategorial; it just focuses on their functions. The term *multifunctionality* is meant to stand in contrast with *conversion* or *zero derivation*. Van Lier takes conversion to be idiosyncratic and unproductive, producing meanings for forms in alternate discourse functions that are not predictable (see §2.3.2.4 and §2.3.3.2 for further discussion). Multifunctionality is also distinct from zero derivation from a common root. Instead, multifunctional lexemes are those whose semantic interpretation is entirely predictable from context, and whose uses in different contexts are productive. Their meanings should be *compositional*. For example: when an action word is used in a referring construction its predicted meaning is that of an *action nominalization*, ‘(the act of) X-ing’; and when an object word is used in a predicate construction its predicted meaning is that of a *predicate nominal*, ‘be an X’. Examples of these predictable, compositional meanings for flexible items in Chamorro are shown in (14).

(14) Chamorro (Austronesian > Malayo-Polynesian)

- a. para **batangka-n** karabão esti  
FUT **shed-LINK** carabao this  
‘this is going to be a carabao shed’ (Chung 2012: 8)
- b. para **gatbesa** ha’  
FUT **decoration** EMPH  
‘[she] is going to be a decoration’ (Chung 2012: 20)

In the two examples above, the meaning of the object words ‘shed’ and ‘decoration’ are predictable when used in a predicative context: ‘be a shed/decoration’. However, lexical items used in their non-prototypical functions very frequently do not have predictable meanings. Consider the example in (15).

(15) Chamorro (Austronesian > Malayo-Polynesian)

ma    **se'si'**    i    babui  
 AGR   **knife**    the    pig  
 'they stabbed the pig'

(Chung 2012: 29)

In this example, the meaning 'stab' cannot be predicted from the meaning of the object word 'knife'. It could have just as easily meant 'be a knife' or 'cut'.

Van Lier takes examples like those in (14) to be instances of genuine multifunctionality, and those in (15) to be cases of conversion. Others have also adopted a position similar to van Lier's, in which only the semantically compositional / predictable uses of a lexical item in different discourse functions are considered flexible (Croft 2001b: §2.2.2–§2.2.3; Evans & Osada 2005: §3.2).

#### 2.3.1.4 Precategoriality / Acategoriality

The various approaches which analyze flexible items as being at some level uncategorized until they receive their interpretation from context may be lumped together under the umbrella terms *precategoriality* or *acategoriality*. Hopper & Thompson's influential (1984) paper is an early application of the concept of acategoriality to the analysis of flexible items:

[L]inguistic forms are in principle to be considered as *lacking categoriality* completely unless nounhood or verbhood is forced on them by their discourse functions. To the extent that forms can be said to have an apriori existence outside of discourse, they are characterizable as *acategorial*; i.e. their categorial classification is irrelevant. Categoriality—the realization of a form as either a N or a V—is imposed on the form by discourse. (Hopper & Thompson 1984: 747)

The term *precategorial* has become a somewhat more common term for roughly the same concept (though some researchers use the term in more strictly-delineated ways) (Evans & Osada 2005: 357, 362–364; Bisang 2008; 2013). It is especially preferred by morphological models that presuppose stages of derivation, such that lexical items are precategorial before they reach a certain stage of the derivation (Halle & Marantz 1994; Arad 2005; McGinnis-Archibald 2016; Siddiqi 2018). Vapnarsky & Veneziano (2017b: 5) distinguish polycategoriality from acategoriality by defining acategoriality as implying “no primitive / original cat-

egorial marking at all”, and polycategoriality as allowing a lexical item “to be only partially unspecified for category, with possible constraints on the relevant categories”. Languages for which precategorial analyses have been advanced include Cherokee (Haag 2017), Gooniyandi (McGregor 2013), Kuikuro (Franchetto & Santos 2017), Mundari (Hengeveld & Rijkhoff 2005). Pfeiler (2017) also presents psycholinguistic evidence that the earliest utterances of L1 learners of Yucatec Maya are acategorial.

A central concern in precategorial approaches is the precise mechanism by which a lexical item receives its categorization in context (Hengeveld, Rijkhoff & Siewierska 2004: §3.7). There are two main theories of semantic indeterminacy in flexible items: *underspecification* (Farrell 2001; Rijkhoff & van Lier 2013) and *vagueness* (Tuggy 1993; Hengeveld, Rijkhoff & Siewierska 2004; Hengeveld & Rijkhoff 2005). The essential difference is that underspecification entails semantic minimalism, while vagueness entails semantic maximalism. An underspecified lexeme has a minimal, core meaning, and receives its categorial meaning from the discourse context it appears in; a vague lexeme has a maximal, broad meaning that covers all the possible discourse contexts it appears in. (There is of course quite a deal of variation in the literature as to how scholars use these terms, with many researchers conflating the two.) Hengeveld & Rijkhoff (2005: 414) offer the example of English *cousin* as a word that is semantically underspecified for gender, such that the gender of the referent must be understood from context. Denison (2018) argues that the English word *long* exhibits adjective ~ adverb underspecification in Old English and Middle English.

In contrast, Hengeveld, Rijkhoff & Siewierska (2004: 539–541) outline a theory regarding exactly how vagueness operates in the context of lexical flexibility:

[E]ach flexible lexeme has a single (vague) sense. By placing the flexible lexeme in a particular syntactic slot or by providing it with certain morphological markers, the speaker highlights those meaning components of the flexible lexeme that are relevant for a certain lexical (verbal, nominal, etc.) function. Thus we contend that the meaning of a flexible lexeme always remains the same, and that morphosyntactic and other contextual clues signal to the addressee how to interpret this lexeme in an actual utterance. In other words, it is the use of a vague lexeme in a certain context (an actual linguistic expression) that brings out certain parts of its meaning, giving the category-neutral lexeme a

particular categorial (verbal, nominal, etc.) flavour. (Hengeveld & Rijkhoff 2005: 541)

(Note that while vagueness implies a certain potential ambiguity, Hengeveld & Rijkhoff reserve the term *ambiguity* for cases of distinct, homophonous lexemes.)

Evans & Osada (2005: 363–364) and Kihm (2017) criticize both precategorial approaches for their imprecision, claiming that it would be impossible to formulate a definition for many flexible items that is broad enough to encompass all their uses. Kihm (2017: 87) illustrates this difficulty with the various meanings of the Arabic root *s-q-t*, which could arguably be glossed FALL. A selection of stems containing this root are given in (16).

(16)	<u>Standard Arabic (Afroasiatic &gt; Semitic)</u>	
	saqata	‘to fall’
	saqiit	‘hail’
	saqqaata	‘door latch’
	masqat	‘place where a falling object lands; waterfall’
	isqaat	‘overthrow; shooting down; miscarriage; subtraction’
	tasaaqut	‘fall of hair’
	saaqita	‘fallen woman; harlot’
	suquut	‘fall; crash; collapse’
	saqt	‘dew’
	siqt	‘miscarried fetus’
	suqt	‘sparks flying from a flint’
	saqat	‘offal; rubbish’
	saqta	‘tumble; slip; mistake’
	saaqit	‘fallen; mean; missing’

It is difficult to imagine a single definition of *s-q-t* which could adequately demarcate just this set of meanings. This difficulty could perhaps be overcome, however, by loosening the requirement that the meaning of a lexical item be unitary. Word meanings are *polycentric*, where different senses of an item are related through *meaning chains* rather than all through a single, central member (Taylor 2003: 110). This is often referred to as a *family resemblance* structure for categories. The difference between monocentric and polycentric categories is illustrated schematically in Figure 2.2. In both diagrams, each letter A–E represents a sense of a lexical item. In the monocentric case, all the senses of the lexical item are related through its core sense A. In the polycentric case, senses A and E are related only through their intervening



connections.<sup>2</sup>



Figure 2.2: Monocentric vs. polycentric categories

Recognizing that word meanings are polycentric addresses Evans & Osada's (2005) and Kihm's (2017) criticisms of vagueness theory because it shows that the disparate senses of a lexical item can be related without having to share any core component of their meanings. The use of a lexeme in a certain context then profiles one of these senses over others. Kihm himself hints at this solution by referring to the related Arabic stems in (16) as a *lexical family*.

### 2.3.1.5 Monocategoriality

In the extreme case where all lexical items in a language are precategoryal, the language could be considered *monocategorial*, possessing a single, open syntactic category. This is effectively the same as saying that the language lacks lexical categories altogether, the difference being primarily one of emphasis. David Gil analyzes both Tagalog (1995) and Riau Indonesian (1994) as being of this extreme monocategorial type. Moreover, he argues that monocategoriality must have been typical of an earlier stage of language evolution in which dedicated morphosyntactic constructions for different discourse functions had yet to evolve (Gil 2005; 2006; 2012). He names this abstract language type an *isolating-monocategorial-associational* (IMA) language.

---

<sup>2</sup>The terms *monothetic* and *polythetic* are sometimes used for this distinction instead (Lewandowska-Tomaszczyk 2007: 146).

### 2.3.1.6 Transcategoriality

It is also worth briefly mentioning *transcategoriality*, since the term arises occasionally in connection with lexical flexibility and is potentially easily confused with other terms mentioned above. Robert (2003) uses *transcategoriality* to describe the ability for a single form to serve both lexical and grammatical functions. This is common in grammaticalization scenarios in which the original, lexical use of a form continues to exist alongside its newer, functional use. This is commonly referred to in the grammaticalization literature as *divergence* (Hopper & Traugott 2003: 118). Since the focus of lexical flexibility is on *lexical* items and categories rather than *functional* ones, the concept of transcategoriality is not directly relevant to the study of lexical flexibility.

### 2.3.1.7 Conversion / Zero derivation

*Conversion* is the process whereby a lexical item simply changes its word class with no overt morphological marker of that change (Crystal 2008: 114). *Zero derivation* is an alternate analysis of the same phenomenon that posits the presence of a derivational marker with no phonological realization. I prefer the term *conversion* in this thesis. Since the literature on conversion and zero derivation is extensive and the concepts are well-established, I will treat them only summarily here, focusing on their relationship to lexical flexibility.

The concept of conversion is based on the premise that lexical items in a language are fully categorized for part of speech, meaning that an analysis of lexical flexibility as conversion falls under the categorial (as opposed to acategorial) analyses of lexical flexibility mentioned at the beginning of Section 2.3.1. Conversion is generally characterized as a kind of word formation, implying that a new lexeme has been created. Therefore, conversion and lexical flexibility are mutually exclusive analyses of multifunctional items; lexical flexibility implies the existence of one polysemous lexeme which can fulfill multiple discourse functions, while conversion implies the existence of two homonymous / heterosemous lexemes with different discourse functions. Remember too from Section 2.3.1.3 that van Lier (2012) distinguishes

conversion from multifunctionality, where conversion is reserved for unproductive / unpredictable derivations. Not all scholars would delimit conversion in this way, however.

Conversion also implies directionality. In cases of conversion, one of the two uses of a form is in some way basic or prior to the other (Mithun 2017: 156; Vapnarsky & Veneziano 2017b: 5). Under a flexible analysis, by contrast, the different functions of a single flexible item have equal theoretical status. If it could be shown that certain seemingly flexible uses of a lexical item were in some way marked in relation to each other, this would therefore constitute potentially disconfirming evidence against a flexible analysis. This is in fact one of the major arguments presented against flexible analyses, to be discussed in Section 2.3.3. There are at least four ways in which one member of a putatively flexible set of lexical items might be considered more basic than the others: 1) diachronically, in which one use of the lexical item appears before the others historically; 2) semantically, in which the meaning of the derived item is more semantically complex than that of the basic one; 3) morphologically, in which the more basic item is irregularly inflected but the derived item is regularly inflected; or 4) frequently, in which derived lexical items are used less frequently than their more basic counterpart (Plag 2003: 108–111). Speakers themselves also have intuitions about which member of a flexible set is basic and which are derived (Mithun 2017: 166). As will be explained in Section 2.4.2, the idea that certain uses of a lexical item are marked in relation to each other is also central to the typological markedness theory of lexical categories.

#### 2.3.1.8 Functional shift / Functional expansion

Especially in the American context, another common term for conversion is *functional shift* (Cannon 1985). In most research, the term is used essentially interchangeably with *conversion* or *zero derivation*. However, functional shift can be usefully distinguished from conversion by its emphasis on function over than category, paralleling the distinction between polycategoriality (implying language-specific categories) and polyfunctionality (with no such implication). In its literal interpretation, the term suggests a shift in the meaning of a lexical item

from one discourse function to another, an analysis amenable to a constructional approach, and one that is not committed to the existence of language-particular categories. A slight improvement on this term would be *functional expansion*, since it emphasizes the expansion of a linguistic form into new functions / contexts as opposed to the wholesale shift from one function to another implied by *functional shift*.

### 2.3.2 Key findings

The emergence of lexical flexibility as an object of study has led to a number of edited collections or journal volumes focused on flexibility and word classes more generally (Vogel & Comrie 2000, Evans & Osada 2005 (target article), Ansaldo, Don & Pfau 2010, Lois & Vapnarsky 2003, Rijkhoff & van Lier 2013, Simone & Masini 2014, Błaszczak, Klimek-Jankowska & Migdalski 2015, Vapnarsky & Veneziano 2017a, van Lier 2017 (target article), Vapnarsky & Veneziano 2017b, Cuyckens, Heyvaert & Hartmann 2019), plus any number of individual articles (see especially Farrell [2001], Rijkhoff [2007], van Lier [2012], and Mithun [2019]). Out of these collections have emerged several recurring findings, each of which is summarized in this section.

#### 2.3.2.1 Parts-of-speech hierarchy

In addition to laying out a theory of flexible categories, Hengeveld (1992) presents the results of a 30-language survey of parts of speech in which he finds that the categories which are most likely to occur as an independent class in a language are subject to an implicational hierarchy, shown in (17), which Hengeveld refers to as the *parts-of-speech hierarchy*.

(17) Verb > Noun > Adjective > Adverb

Categories to the left of the hierarchy are more likely to occur as a distinct part of speech than categories to the right. Applying this hierarchy to Hengeveld's flexible vs. rigid distinction yields the parts-of-speech typology in Figure 2.3 (adapted from Hengeveld [1992: 69] and

Rijkhoff [2007: 718]). The terms for the different categories in flexible languages are from Hengeveld, Rijkhoff & Siewierska (2004). Hengeveld points out that this is not a strict classification scheme; languages may sit at the boundaries between types and exhibit exceptions.

	<b>predication</b>	<b>reference</b>	<b>predicate modification</b>	<b>referent modification</b>
<b>flexible</b>	contentive			
	verb	non-verb		
	verb	noun	modifier	
<b>rigid</b>	verb	noun	adjective	adverb
	verb	noun	adjective	
	verb	noun		
	verb			

Figure 2.3: Hengeveld’s (1992: 69) typology of parts-of-speech systems

As mentioned in Section 2.3.1.1, Hengeveld’s typology could be criticized for its reliance on large, language-specific lexical categories instead of constructions. One could however reframe Hengeveld’s implicational hierarchy in terms of functions rather than categories, as in (18). I call this the *hierarchy of discourse functions*.

(18) predicate > referent > predicate modifier > referent modifier

In (18), functions to the left of the hierarchy are more likely to have dedicated morphosyntactic constructions than those to the right. This reformulation avoids a commitment to any language-particular categories while still capturing the implicational trend observed by Hengeveld.

This hierarchy of discourse functions has proven to be a fairly robust finding in the literature on lexical flexibility, now supported by a number of subsequent studies (Anward 2000; Rijkhoff 2000; Vogel 2000; Beck 2002; Rijkhoff 2002; 2003; Hengeveld, Rijkhoff & Siewierska 2004; van Lier 2006; Hengeveld 2007; Hengeveld & van Lier 2012; Hengeveld & Valstar 2010; Beck 2013; Bisang 2013; Hengeveld 2013).

### 2.3.2.2 Reference-predication asymmetries

The hierarchy of discourse functions also hints at another important feature of lexical categories: there is something privileged about the predicating function. A survey of the literature on lexical flexibility reveals patterned asymmetries in the behavior of lexical items with regard to predication vs. reference, even in very flexible cases. For starters, while it is quite common for languages to freely allow object words to be used as predicates with no special marking, the reverse case is much less likely Hopper & Thompson (1984: 745). The functional expansion of an item's uses from predication into reference always seems to be more marked (or at least as marked) as the shift from a referring function to a predicating function.

This fact has been observed independently by numerous researchers. For example, Stevick (1968: 251) and Marchand (1969: 373–374) both observe that conversion from noun to verb in English has always been more common than from verb to noun, and Kastovsky (1996: 98) points out that English does not even have a native noun > verb derivational suffix—any affixes of this type are borrowed from Romance languages. Central Alaskan Yup'ik is another example of a language with very many nominalizers but few verbalizers (Mithun 2017: 158).

Flexibility itself is frequently *unidirectional*, meaning that any object word may be used for predication, but that action words used for reference are marked (Croft 2001b: 69; Evans & Osada 2005: §3.3; Beck 2013). Nakayama (2001: 44) frames flexibility in Nuuchahnulth in terms of a stem's ability to predicate, reporting that “all inflectional stems are potentially predicative”, but the reverse is not true. Discussing Classical Nahuatl, Launey (1994; 2004) introduces the term *omnipredicativity* to describe languages in which all lexical items are potentially predicative, but no corresponding term *omnireferentiality* has appeared in the literature. That said, languages which have undergone *insubordination* (in which subordinate, usually nominalized clauses are reanalyzed as main clauses, and the nominal inflectional marking of the subordinate clause is reinterpreted as verbal inflectional marking [Evans 2007; Mithun 2008; Evans & Watanabe 2016]) do exhibit noun-oriented flexibility in the sense that verbal inflection mirrors nominal inflection. This famously led to the claim that all lexical items in

Eskimo languages are fundamentally nominal in nature (Sadock 1999). However, cases of insubordination do not constitute counterexamples to the predicating tendency in language. Even in these languages, the use of action words for reference is still less marked than the use of object words for predication.

Kastovsky (1996) argues that this asymmetry arises from the fact that “deverbal nouns have a much more diversified semantics than denominal verbs” (Kastovsky 1996: 96), meaning that the range of possible meanings for a deverbal noun (a noun derived from a verb) is broader than for a denominal verb (a verb derived from a noun). Examining data from English, Kastovsky shows that when an object word is used to predicate, its possible meanings are limited to combinations of BE, BE LIKE, BE IN, BECOME, HAVE, DO, DO WITH, and CAUSE. When an action word is used as a predicate, however, the range of meanings include any abstract representation of the event itself (an action nominalization), or any one of the arguments associated with the verb, which come in a variety of semantic roles.

A similar, cognitively oriented explanation for reference-predication asymmetries is given by Hopper & Thompson (1984: 745):

[A] nominalization names an event taken as an entity; however, a “verbalization” does not name an “entity taken as an event”, but simply names an event associated with some entity. In other words, a nominalization still names an event, albeit one which is being referred to rather than reported on in the discourse; it is, accordingly, still in part a [verb], and not a “bona fide” [noun]. However, a denominal [verb] no longer names an entity at all, and thus has no nominal “stains” to prevent its being a bona fide [verb]. (Hopper & Thompson 1984: 745)

Hopper & Thompson (1984: 746) analyze nominalizations as a kind of metaphor following Lakoff & Johnson (1980: 3a), in which an abstract event is conceptualized as a concrete object. However, no such metaphor exists for verbalizations, explaining the asymmetry in the directionality of lexical flexibility.

### 2.3.2.3 Locus of categoriality

The grammatical level at which a language exhibits flexibility—the root, the stem, or the fully inflected word—differs from one language to the next. In some languages, roots are strongly associated with a particular discourse function, but stems are flexible; in other languages, the reverse is true. I refer to the linguistic level at which a language associates different discourse functions as its *locus of categoriality*. Some linguistic theories include a premise that the locus of categoriality in every language always sits at a certain level (Halle & Marantz 1994; M. C. Baker 2003; 2015; Booij & Audring 2018; Siddiqi 2018), but the evidence from research on lexical flexibility gives strong empirical support to the position that locus of categoriality varies from language to language. In contrast, Błaszczak, Klimek-Jankowska & Migdalski (2015) argue that category information is distributed across different levels of representation.

As one illustration of how flexibility depends on grammatical level, we have seen that roots in Central Alaskan Yup'ik are generally categorical—except for 12% of roots, they are typically strongly associated with just one discourse function, and derivational affixes select for roots of a particular category (Mithun 2017: 162–167). While many derived stems are also strictly associated with just one discourse function, a large but indeterminate number have both referential and predicative uses. Examples of such flexible stems have already been shown in (7) in Section 1.1. Fully inflected words in Central Alaskan Yup'ik, however, never exhibit flexibility (Mithun 2019: 6). So Central Alaskan Yup'ik displays partial flexibility at the root and stem level but not the inflected word level.

As another example, in Mandinka all stems are flexible. No Mandinka stem except for *săa* ‘die’ is used in just one discourse function (Creissels 2017: 46). At the level of the inflected word, however, lexical items in Mandinka belong unambiguously to one category or another (Creissels 2017: 37). Mandinka therefore shows complete flexibility at the stem level but complete rigidity at the inflected word level. (Creissels does not include an analysis of roots in his discussion.)

Surprisingly, some languages display flexibility even at the level of the fully inflected word.



In many North American languages, it is common for fully morphological verbs to function as referents (Hieber [forthcoming](#)), as shown in the following examples.

(19) Chitimacha (isolate)

- a. dzampuyna  
dza-ma-(p)uy-na  
thrust-PLACT-HAB-NF.PL  
'they usually thrust / spear with it'  
'spear' (Swadesh [1939a](#): 56)

- b. pamtuyna  
pamte-(p)uy-na  
ford-HAB-NF.PL  
'they usually cross (it)'  
'bridge' (Swadesh [1939a](#): 17)

(20) Cayuga (Iroquoian > Lake Iroquoian)

- a. qtekhonyá?tha?  
ye-ate-khw-qni-a?t-ha?  
INDEF.AGT.REFL-meal-make-INSTR-IPFV  
'one makes a meal with it'  
'restaurant' (Mithun [2000](#): 200)

- b. kaqtanéhkwi  
ka-rqt-a-nehkwi  
NEUT.AGT-log-EP-haul-IPFV  
'it hauls logs'  
'horse' (Mithun [2000](#): 200)

(21) Navajo (Na-Dene)

- a. tsinaa'eeł  
tsi(n)-naa'eeł  
wood-it.moves.about.floating  
'ship, boat' (Young [1989](#): 316)

- b. chahałheeł  
it.is.dark  
'darkness' (Young [1989](#): 316)

Each of these flexible uses of a morphological verb sits somewhere on a continuum between being fully lexicalized as a referent, so that its predicating use is no longer available, to being a fully productive predicate, with both predicative and referential uses (Mithun [2000](#): 413).

The reason that lexical items may exhibit flexibility at one level of analysis but not another is because “*categorical* shift is often not *categorical*” (Mithun 2019: 1). When an item expands its use into new contexts, not all the morphological, syntactic, and semantic properties of the item shift to accommodate that new use at the same time. It takes time before the morphosyntactic properties of a item adjust to reflect its new use, a process referred to as *actualization* in the grammaticalization literature (De Smet 2012) and *post-constructionalization constructional changes* in the framework of diachronic construction grammar (Hopper & Traugott 2003: 27).

It is in part because the locus of categoriality can vary from language to language that I have used the vague term *lexical item* throughout this thesis, which is intended to be a convenient cover term for root, stem, or inflected word.

#### 2.3.2.4 Item-specificity

A final significant finding to emerge from the empirical research on lexical flexibility is the fact that flexibility is *item-specific* and even *sense-specific*. Individual lexical items or even individual senses of an item that are otherwise very similar in their meanings and morphosyntactic behavior can nonetheless differ in terms of their flexibility.

This fact is nicely illustrated by both Mithun’s (2017) study of lexical flexibility in Central Alaskan Yup’ik and Creissels’s (2017) study of Mandinka. Mithun (2017: 163–164), for example, considers roots for meteorological concepts, and shows that even within this small semantic domain, roots vary as to whether they exhibit flexibility. In (22a) the meteorological roots have predicative counterparts but in (22b) the meteorological roots do not.

(22) Central Alaskan Yup’ik (Eskimo-Aleut > Yupik)

a.	<i>amirlu</i>	‘cloud’	<i>amirlu-</i>	‘be cloudy’
	<i>kaneq</i>	‘frost’	<i>kaner-</i>	‘be frosted’
	<i>aniu</i>	‘snow on ground’	<i>aniu-</i>	‘to snow’
b.	<i>taituk</i>	‘fog, mist’	* <i>taitug-</i>	‘be foggy’
	<i>kavtak</i>	‘hailstone’	* <i>kavtag-</i>	‘to hail’
	<i>mecaliqaq</i>	‘sleet’	* <i>mecaliqar-</i>	‘to sleet’

(Mithun 2017: 163)

Mithun also provides similar data illustrating flexibility gaps for the domains of clothing and instruments:

(23) Central Alaskan Yup'ik (Eskimo-Aleut > Yupik)

a.	<i>taqmak</i>	'dress'	<i>taqmag-</i>	'put on a dress'
	<i>nacaq</i>	'hat, parka hood, cap'	<i>nacar-</i>	'put on a hat, hood'
	<i>atkuk</i>	'parka'	<i>atkug-</i>	'put on a parka'
b.	<i>*piluk</i>	'footwear'	<i>pilug-</i>	'put on footwear'
	<i>*at'e</i>	'clothing'	<i>at'e-</i>	'don, put on clothing'
	<i>*kive</i>	'pants'	<i>kive-</i>	'pull down pants'

(24) Central Alaskan Yup'ik (Eskimo-Aleut > Yupik)

a.	<i>ay'uytaq</i>	'hockey stick'	<i>ay'utar-</i>	'play hockey'
	<i>iqsak</i>	'fishhook'	<i>iqsag-</i>	'to jig for fish'
	<i>kapkaanaq</i>	'trap'	<i>kapkaanar-</i>	'to trap, get trapped'
	<i>keviq</i>	'plug, cork, stopper'	<i>kevir-</i>	'to plug, stuff, caulk'
	<i>kuvya</i>	'fishnet'	<i>kuvya-</i>	'fish by driftnetting'
b.	<i>*kagi</i>	'broom'	<i>kagi-</i>	'sweep'
	<i>*ipuk</i>	'ladle'	<i>ipug-</i>	'ladle, move with bow of boat high in air'
	<i>*pangeq</i>	'double-bladed paddle'	<i>panger-</i>	'paddle with a double-bladed paddle'

On the basis of data like these and discussion with speakers, Mithun observes, “Speakers simply know whether a given root functions as a noun and what its meaning is, and whether it functions as a verb and what its meaning is. Gaps are not predictable[.]” (Mithun 2017: 163). These gaps also vary from dialect to dialect. While the dialect in the above examples has no predicative counterpart for *taituk* ‘fog’, the Nunivak Island dialect does have a pair of roots *nugu* ‘fog’ and *nungu-* ‘be foggy’.

Creissels’s (2017) study of Mandinka is another good illustration of the item-specific nature of flexibility. While Mandinka has nominal and verbal constructions that allow the predicative and referring functions of inflected words to be distinguished unambiguously, it is not as easy to separate stems into similar classes. In Mandinka, all items are flexible, but the way in which items are flexible varies. Stems in Mandinka may be divided into three classes based

on their semantic behavior with regards to flexibility:

- *verbal* lexemes are those whose meaning is predictable when used to refer and therefore analyzable as a case of “morphologically unmarked nominalization”; these are always event nominalizations
- *verbo-nominal* lexemes are those whose meaning in referring constructions is idiosyncratic and therefore not predictable
- *nominal* lexemes are those whose meaning when used as predicates is predictable and limited to ‘provide someone with X’

In Mandinka, therefore, flexibility must be considered on an item-by-item basis since the behavior of each item with regard to flexibility may differ.

In fact, flexible behavior in Mandinka is not just item-specific, but sense-specific as well. Creissels (2017: 54) reports that polysemous lexemes may show different behavior for their different senses. The stem *díŋ*, for example, has two senses: ‘child, young (of an animal)’ and ‘fruit’. However, only the ‘fruit’ sense is available for predication: when used as an intransitive verb, *díŋ* may only mean ‘bear fruit’, not ‘give birth’, even though ‘give birth’ is a perfectly conceivable meaning of this stem in predication. In the sense of ‘child, young (of an animal)’, *díŋ* behaves as a nominal lexeme, but in the sense of ‘fruit’ it behaves as a verbo-nominal lexeme.

When lexical items undergo functional expansion into new discourse functions, it is also only specific senses that do so, not every one of its senses. More evidence for this comes from the diachronic development of the word *run* in English: though the word *run* when used as a predicate has numerous senses, the earliest attestations of *run* used referentially are by and large with just the prototypical sense of ‘fast pedestrian motion’ (the exceptions to this stem from just one corpus file) (Gries 2006: 76). Other referential uses of *run* did not develop until later.

The existence of dialectal differences for flexibility as well as the unpredictable meanings of lexical items when used in various discourse functions show that the development of flexibility depends on conventionalization—whether a given form has assumed a conventionalized

meaning in its role for a specific discourse function. These conventionalizations are language-specific, dialect-specific, item-specific, and even sense-specific (Croft 2000: 97). Speakers can and do playfully use existing lexical items for new discourse functions, but it is not until that combination of form and discourse function is conventionalized with a specific meaning in a community of speakers that we can say the lexical item has undergone functional expansion. An excellent illustration of this is the word *friend* in English. Prior to the widespread adoption of the social networking sites MySpace and Facebook around 2006, the use of *friend* as a predicate had not been widely conventionalized. The growth of social networking sites then led to the specific use of *friend* to mean ‘add as a connection on a social networking site’. Note that it does *not* have the more general sense of ‘be a friend’ or ‘befriend’. Like with Yup’ik and Mandinka, this shows not just that flexibility is item-specific, but that the meaning of flexible uses is often item-specific as well; in many cases it is unpredictable and must be memorized by speakers.

### 2.3.3 Problems & critiques

Despite the robust findings in [Section 2.3.2](#), researchers have challenged the very possibility of lexical flexibility and its presence in various languages. Some of these challenges stem from the fact that certain conceptions of lexical flexibility are based on traditional ideas about the existence of large, language-specific parts of speech, and therefore subject to the same set of criticisms. Other challenges stem from precisely the facts presented in the previous section, namely that both flexibility and the meaning of flexible words are item-specific and often unpredictable, such that these words are not truly “flexible”. Moreover, languages must indicate the discourse function of their lexical items *somehow*—this is basic to our ability to communicate. In a certain sense, the idea that there are items which are fully ambiguous in their discourse function is doomed at the outset. The question is really where these indications of pragmatic function live—the root, the stem, the inflected word, or the clausal context. This section summarizes the main criticisms that scholars have raised against flexible analyses. In

Section 2.4, we then look at alternative theories of word classes and their approach to lexical flexibility.

### 2.3.3.1 Methodological opportunism

A methodological problem with certain theories of flexible items is that they, like traditional theories, commit the fallacy of *methodological opportunism* (Croft 2001b: 30, 41) presented in Section 2.2.3. They do not apply the distributional method consistently. Instead, the criteria which separate lexical items into categories are determined on the basis of additional theoretical commitments. Croft (2001b: §2.2.2) criticizes Hengeveld’s parts-of-speech typology on this basis, noting that Hengeveld ignores distributional evidence for classes smaller than the ones he posits in his typology (noun, contentive, etc.). Evans & Osada (2005) raise similar concerns for Hengeveld’s theory as applied to Mundari. They state that in order for two lexical items to be members of the same lexical class, they must have *equivalent combinatorics*, which is to say that their distributions should be identical (Evans & Osada 2005: 366). Evans & Osada also state that for a language to be flexible, that flexibility must be *exhaustive* in the sense that all members of a putatively flexible class must show equal degrees of flexibility and *bidirectional* in the sense that nouns may be used as verbs and vice versa. Both these criteria are merely different ways of reframing the broader principle that items in a class should share the same distributions (Croft 2005: 434). Evans & Osada proceed to show various ways in which these criteria are not applicable to Mundari, and that Mundari is therefore not a flexible language. At the same time, however, Evans & Osada use these facts to argue for the existence of the equally problematic categories of Noun and Verb in Mundari, using just a “canonical subset of distributional facts” (Evans & Osada 2005: 434, fn. 17). Croft’s (2005) commentary on Evans & Osada’s (2005) target article is partially devoted to critiquing them on this point. The problem of methodological opportunism is present for any analysis which assumes that languages have a small set of large lexical categories—whether that analysis is flexible or traditional.

### 2.3.3.2 Semantic shift

Broadly speaking, however, the primary argument against theories of flexible word classes is that they ignore a great deal of item-specific knowledge that speakers have about lexical items and their uses in different functions (Evans & Osada 2005: §3.2; Beck 2013: 216). This issue has already been discussed in some detail in [Section 2.3.2.4](#), but it bears explaining precisely why such item-specific knowledge constitutes a problem for theories of lexical flexibility.

For starters, when a lexical item expands into a new discourse function, there is a *semantic shift* in the direction of the meaning typically associated with the new context (Croft 1991: 74–77; 2001b: 73). For example, when a property word is used in a referring expression, its meaning shifts to a person or thing possessing that property, not a reference to the abstract property itself. These semantic shifts cannot be attributed to some broader pragmatic principles—they are a matter of convention and require broader uptake in a community of speakers in order to be conventionalized (as illustrated with the English word *friend* above). Because the meaning that results from this semantic shift is conventional, language-specific, and often idiosyncratic, flexible items cannot be truly productive, as is implied by the term “flexible”. There is always a conventionalized component to their meanings.

Examples of idiosyncratic and unproductive shifts in the meaning of flexible items abound in the literature. Consider again the examples from Mundari in (3), repeated here as (25).

(25) Mundari (Austroasiatic > Munda)

- a. buru=ko                      bai-ke-d-a.  
    mountain=3PL.SUBJ    make-COMPL-TR-IND  
    ‘They made the mountain.’

(Evans & Osada 2005: 354)

- b. saan=ko                      buru-ke-d-a.  
    firewood=3PL.SUBJ    mountain-COMPL-TR-IND  
    ‘They heaped up the firewood.’

(Evans & Osada 2005: 355)

As a predicate, the stem *buru* means ‘heap up’, but this meaning is not predictable from just the combination of the nominal sense ‘mountain’ and its predicative use. The word could have just as easily meant ‘climb a mountain’ or ‘overcome’ or simply ‘be a mountain’. No

general pragmatic principles could have predicted this meaning. Likewise consider the Central Alaskan Yup'ik examples in (7c) from [Chapter 1](#). Why does the combination of *iqeq-* 'corner of mouth' + *-mik* 'thing held in one's mouth', 'to put in one's mouth' result in *iqmik* 'chewing tobacco'? Why not 'oral thermometer' or 'toothpick'? Mithun provides many more unpredictable examples, shown in (26).

(26) Central Alaskan Yup'ik (Eskimo-Aleut > Yupik)

- |    |                                   |   |
|----|-----------------------------------|---|
| a. | <i>mecur-</i><br><i>mecuq</i>     | 'get blood poisoning'<br>'liquid part of something, sap, juice, green/waterlogged wood'   |
| b. | <i>melug-</i><br><i>meluk</i>     | 'suck; eat roe directly from the fish'<br>'fish eggs, roe, fish eggs prepared by allowing them to age and become a sticky mess' |
| c. | <i>qager-</i><br><i>qageq</i>     | 'explode, to pop'<br>'blackfish which is boiled, allowed to set in its cooled, jelled broth'                                    |
| d. | <i>qumig-</i><br><i>qumik</i>     | 'hold inside (of clothing)'<br>'enclosed thing, thing inside, fetus'  |
| e. | <i>aveg-</i><br><i>avek</i>       | 'divide in half, to halve'<br>'half'; also 'half-dollar; person who is half Native'   |
| f. | <i>napa-</i><br><i>napa</i>       | 'stand upright'<br>'tree'   |
| g. | <i>yuurqar-</i><br><i>yuurqaq</i> | 'sip'<br>'hot beverage, tea'  |

Or consider the example from Cayuga in (20b), repeated here as (27).

(27) Cayuga (Iroquoian > Lake Iroquoian)

kaqtanéhkwi  
ka-rqt-a-nehkwi  
NEUT.AGT-log-EP-haul.IPFV  
'it hauls logs'  
'horse'

(Mithun 2000: 200)

Of all the possible nominal meanings that could reasonably derive from 'it hauls logs'—cart, tractor, ox—the fact that its nominal use means 'horse' is specific to Cayuga and must be



memorized by speakers.

Conventionalizations of lexical items used in new discourse functions also vary across languages. While the principle of semantic shifts still broadly holds, the specific meanings of these conventionalizations are unpredictable. Croft exemplifies this point by comparing English *school* with Tongan *ako* ‘school / study’.

English *school* used predicatively does not mean the same thing as Tongan *ako* used predicatively, namely ‘study’. Going in the opposite direction, English *study* used referentially does not mean the same thing as Tongan *ako* used referentially, namely ‘school’. Finally, English *small* used referentially does not mean the same thing as Tongan *si’i* ‘childhood’ used referentially. (Croft 2000: 71)

Since the meanings of putatively flexible items in different discourse functions are not predictable, many scholars reason that these lexical items cannot be truly “flexible” in the sense of multifunctional or precategoryal.

### 2.3.3.3 Lexical gaps

Just as unpredictable in flexible cases is which sense of a item will be co-opted into the new discourse function. In Wolof, for example, the referential use of the word *ndaw* can only mean ‘young’, whereas the predicative use may mean either ‘be young’ or ‘be little, small’ (Kihm 2017: 91). We have also already seen similar lexical gaps for Central Alaskan Yup’ik and Mandinka in Section 2.3.2.4 above. If a lexical item lacks any conventionalized use in different discourse functions, than it cannot rightly be considered flexible, or a member of a flexible word class.

### 2.3.3.4 Counterarguments

Pointing out that functional expansion involves both semantic shifts and functional gaps is generally intended to show that lexemes cannot be truly flexible in the sense of being multifunctional (§2.3.1.3) or precategoryal (§2.3.1.4), and that uses of the same lexical item

for different discourse functions should therefore be considered cases of conversion—that is, homonymy or heterosemy. There are however two major problems with this argument.

The first is that it creates a false dichotomy between homonymy and polysemy, when in fact the two phenomena are opposite endpoints on a continuum. Debates over the lexical unity of an item—that is, whether two uses of a lexical item are homonymous or polysemous—arise from an Aristotelian desire to neatly sort those uses into distinct lexemes, when in fact reality is much more complex. If this problem sounds familiar, that is because it is the same methodological problem that arises when trying to exclusively categorize lexical items into different classes. The complex adaptive nature of language makes categorical classification at either level impossible.

As discussed in [Section 2.3.1.4](#), we know from cognitive research that mental categories are prototypal, and that the meanings of words display a polycentric, family resemblance structure. Two senses of a lexical item are often related only tenuously through a network of intervening semantic connections or meaning chains. Langacker (1988) calls this the *network model* of category structure. [Taylor](#) points out that “[o]ne consequence of adopting the network model is that the question of whether a lexical item is polysemous turns out to be incapable of receiving a definite answer.” (Taylor 2003: 167)

Over time, as this lexical network expands, the meanings of a lexical item can diverge so drastically that speakers no longer have a direct cognitive association between them. [Mithun](#) exemplifies this nicely for both Cayuga and English. Discussing morphological Verbs used as referents in Cayuga, she notes the following:

If asked the meaning of *kaqtanéhkwi* [lit. ‘it hauls logs’], Cayuga speakers normally respond ‘horse’. Though it has the morphological structure of a verb, it has been lexicalized as a nominal. The literal meanings of many verbal nominals are still accessible to speakers, but the origins of others have faded, and speakers express surprise at discovering them. Similarly, when asked “What would you like for breakfast?”, most English speakers do not think about breaking their night-time fast, though they can usually be made aware of the literal meaning of *breakfast*. (Mithun 2000: 413)

Lexicalization is a process and a continuum. Words can be lexicalized in new discourse func-

tions to varying degrees. The first use of a lexical item in a new discourse function is innovative; each subsequent use then contributes further to its conventionalization in that function.

Pointing out that functional expansion often creates idiosyncratic and unpredictable meanings essentially amounts to saying that senses of lexical items can be highly divergent. This point is not in itself an argument against flexible analyses. Flexible items themselves may sit anywhere on the continuum from having closely connected, productive and predictable meanings, to having extremely divergent, idiosyncratic, and unpredictable meanings. This is not a special fact about flexible items; it is simply true of words generally.

Croft (2001b: 73) expresses concern that ignoring semantic shifts in the analysis of flexible items overlooks important insights about how such semantic shifts are patterned (specifically, the universal fact that semantic shifts are always in the direction of the item's new discourse function; see §2.3.3.2). Given that so many researchers have indeed ignored semantic shift when arguing for flexible analyses, Croft's concern is warranted. However, it is entirely possible to define lexical flexibility in a way that both allows for the meaning of a lexical item to encompass multiple discourse functions and acknowledges that such multifunctional uses involve patterned semantic shifts. The way to do this is to ground the definition of lexical flexibility in the pragmatic functions of reference, predication, and modification rather than language-specific categories like Noun, Verb, and Adjective. I offer such a definition in [Section 2.5](#).

The second significant problem with using semantic shifts as an argument against the existence of flexible lexemes is that it proves too much. If semantic shift is taken as evidence against the lexical unity of putatively flexible items, then it must also be taken as evidence against the lexical unity of non-flexible items. Put simply, semantic shift is an analytical problem for all words, not just flexible ones.

This fact becomes clear when we ask, "What counts as a semantic shift? Just how 'large' of a change in meaning (if it were even possible to quantify such a thing) does a semantic shift require?" To illustrate this problem, consider the semantic contribution of plural mark-

ing crosslinguistically. In the canonical case, plural marking is considered inflectional rather than derivational (Corbett 2000: 2), meaning that it does not create a new lexeme. Instead, it modifies the meaning of the existing lexeme slightly, in line with the classic distinction between inflection vs. derivation. However, there are numerous cases of lexical items in English with more or less drastic differences in meaning between the singular and plural, and/or senses that are only available in one of the two numbers. Consider the examples in (28).

(28) English (Indo-European > Germanic)

a.	<i>air</i> <i>airs</i>	‘atmosphere’ ‘affected manners’
b.	<i>arm</i> <i>arms</i>	‘upper limb; anything resembling a limb’ ‘weapons, firearms’
c.	<i>blind</i> <i>blinds</i>	‘unable to see’ ‘screen for a window’
d.	<i>character</i> <i>characters</i>	‘personality, mental qualities’ ‘people in a novel, play, or film’
e.	<i>custom</i> <i>customs</i>	‘tradition; socially accepted behavior’ ‘department which levies duties on imports’
f.	<i>force</i> <i>forces</i>	‘strength, energy’ ‘collection of military units’
g.	<i>good</i> <i>goods</i>	‘excellent, high quality’ ‘merchandise or possessions’
h.	<i>manner</i> <i>manners</i>	‘way of doing something’ ‘social conduct; socially acceptable conduct’
i.	<i>spectacle</i> <i>spectacles</i>	‘visually striking performance or display’ ‘pair of glasses’
j.	<i>wood</i> <i>woods</i>	‘fibrous material in the trunk of trees or shrubs’ ‘area of land covered with trees’

Semantic shifts for plural marking in English are not limited to just a handful of specific lexical items. Generic uses of the plural as in the expression *foxes are cunning* create a semantic shift away from a concrete entity (*a/the fox*) to a generic, unperceivable one—a use which

strays from the prototypical function of nouns as concrete perceptible entities (Hopper & Thompson 1984: 708).

As with flexible items, the semantic shifts that occur with plural marking can become so substantial that speakers no longer cognize the morphological singular and plural as members of the same lexeme. Such is the case in the historical development of *brother* vs. *brethren* in English. The word *brethren* became so strongly conventionalized with its religious meaning in the plural that it was independently lexicalized as a plural-only (*plurale tantum*) noun, and the original plural underwent renewal with the emergence of the form *brothers*. This is exactly the kind of lexicalization process that occurred for many morphological verbs reanalyzed as nouns in Cayuga and many other North American languages.

A similar example comes from Chitimacha, which has a pluractional marker *-ma* indicating verbal number (plural agents, plural patients, or repeated action). In some cases the use of *-ma* is purely compositional, so that it can be considered merely an inflectional marker of verbal number. In other cases *-ma* so significantly alters the meaning of the word that it must be considered derivational. Compare the uses of *-ma* in each of the pairs of verbs in (29) (note that (29b) and (29c) are phrasal verbs with a preverbal particle).

(29) Chitimacha (isolate)

- |    |                    |   |
|----|--------------------|---|
| a. | <i>kow-</i>        | ‘call’                                      |
|    | <i>kooma-</i>      | ‘call multiple people’                      |
| b. | <i>qapx cuw-</i>   | ‘come back; go about’                       |
|    | <i>qapx cuuma-</i> | ‘travel; wander’                            |
| c. | <i>qapx qiy-</i>   | ‘turn together; mix, join’                  |
|    | <i>qapx qiima-</i> | ‘give a prayer, benediction; perform magic’ |

(Swadesh 1939a)

In (29a), the use of *-ma* is entirely compositional. The presence of *-ma* indicates that the verb has a plural patient argument. In (29b), the use of *-ma* is still arguably compositional, though perhaps somewhat lexicalized given the high frequency with which the stem appears in the texts. ‘travel, wander’ could reasonably be interpreted as a continued repetition of ‘go about’.

In (29c), however, *qapx qiima-* has become lexicalized with a new meaning not directly related to that of *qapx qiy-*. The diachronic connection between the two meanings is that prayers and magical incantations were traditionally accompanied by circling gestures with the arms. *qapx qiima-* originally meant ‘turn/circle around repeatedly’, but over time lexicalized with its new religious meaning in the pluractional, ‘give a prayer, benediction’. This lexicalization process parallels that of *brethren* in English. Such a range of inflectional vs. derivational uses of pluractionals is quite common crosslinguistically (Mithun 1988; Mattioli 2020).

Finally, there are many languages which do not typically mark plurality on nouns (Dryer 2013), and yet have senses available in semantically plural contexts but not singular ones (where the semantic number can be understood from the clausal context, usually through verbal number marking). For example, the word *soq* in Chitimacha may mean ‘foot’ or ‘paw’ in a singular context and ‘feet’ or ‘paws’ in a plural context, but may also mean ‘tracks’ (e.g. animal tracks) in a plural context—a significant and idiosyncratic shift in meaning, and one that is both language-specific and item-specific and thus conventional. This use constitutes a *morphologically unmarked semantic shift* in the meaning of the word, just as idiosyncratic meanings of words in cases of functional expansion also constitute morphologically unmarked semantic shifts. If we take such unmarked semantic shifts as evidence against lexical unity in the cases of flexible items, then we must also say that the ‘foot’ and ‘tracks’ meanings of *soq* constitute two distinct lexemes as well.

One might ask, if we start splitting up lexemes based on every degree of semantic shift, where does the splitting stop? This is exactly analogous to the problem of lumping vs. splitting in the context of lexical categories. The Radical Construction Grammar solution to this problem is to abandon the commitment to larger groupings of items (the major lexical categories) and acknowledge that languages consist of an interconnected network of smaller items (constructions) instead (Croft 2001b). This approach has the major advantage of sidestepping unproductive debates about the existence or unity of lexical categories in particular languages, and shifts the focus instead to understanding the relationships and patterns among individual

constructions. This is precisely what I propose to do for lexemes as well. If we abandon the idea that all the meanings associated with a form must be in some way grouped into lexemes based on their morphosyntactic contexts of occurrence, we sidestep unproductive debates regarding homonymy vs. polysemy, and can instead focus on the relationships and patterns among the various senses associated with that form—specifically, the nature of the semantic shifts that occur between uses of the form in different discourse functions.

In sum, idiosyncratic semantic shifts are not the problem for theories of lexical flexibility that they are often taken to be. Indeed, functional expansion would not be possible if hearers were not capable of determining the meaning of a form when used in even highly unusual contexts. Innovative uses of lexical items in new functions would be all but impossible, providing no opportunity for such innovations to receive broader adoption in the linguistic community. Each time a hearer encounters a novel use of a lexical item for the first time, they must accomplish the difficult task of discerning its meaning. This is no less true for flexible items as it is for non-flexible items, or for items whose meaning is predictable vs. unpredictable. *Every* use of a word is an instance of functional expansion because every use of a word is always in a slightly different discourse and social context than the one before. The meaning of a word in a given context is highly socially and situationally dependent, and that context can change completely from one utterance to the next. Every token of a word thus necessarily appears in a new pragmatic context, and that pragmatic context slightly shapes its meaning. Language use *is* language change. Semantic shift is therefore an integral and ubiquitous part of language use; the question is simply the degree of that semantic shift.

## 2.4 Functional approaches

Functionalism as an approach to linguistic explanation is multifaceted. It looks to factors outside of the structural form of language as an explanation for that form—most especially cognition, usage effects from frequency, and information structuring in discourse (Croft 2001a:

6323–6324). In this section I present Croft’s (1991; 2000; 2001b) functional theory of lexical categories, which explains crosslinguistic patterns in the coding of reference, predication, and modification as arising from the interaction between our mental categories and the needs of discourse. I then use this theory as a framework for defining lexical flexibility in Section 2.5. I begin with a brief discussion of prototype theory as it pertains to lexical categories (§2.4.1), before expounding upon typological markedness theory (§2.4.2).

### 2.4.1 Prototype theory

It has long been recognized that the categories of human cognition are prototypical. In a series of studies, Eleanor Rosch and colleagues demonstrate that category membership is a matter of degree, and that there are better and worse representatives of any given mental category (E. Rosch 1973; E. H. Rosch 1973; E. Rosch 1975; Rosch & Mervis 1975; Rosch et al. 1976; E. Rosch 1978). Prototype theory was then popularized in linguistics by Lakoff (1987), Langacker (1987), Taylor ([1989] 2003), and Croft ([1990] 2003; 1991), among others.

The evidence for prototypical structure in mental categorization is robust (Taylor 2003: 46–47). When asked to rate whether an item is a good example of a category, participants consistently rate prototypical members as better examples of the category than non-prototypical ones. In listing experiments where participants are asked to list members of a category, prototypical members are listed earlier and more frequently than non-prototypical members. Finally, prototypical members of a category are identified by participants as being members of the category more quickly than non-prototypical members. Each of these effects is scalar, such that individual members of a category sit anywhere on a scale of more to less prototypical.

Prototype effects arise from the basic human need to interpret the world around us: “Strictly speaking, every entity and every situation that we encounter is uniquely different from every other. In order to function in the world, all creatures, including humans, need to be able to group different entities together as instances of the same kind. [...] [C]ategorization



serves to reduce the complexity of the environment.” (Taylor 2003: xi) This fact is often referred to as the *principle of cognitive economy*, whereby we group similar stimuli together in order to maximize information while minimizing cognitive effort (Evans & Green 2006: 255). The gradience within these groupings results from the fact that “concepts function as mental reference points. When we come across new phenomena, we tend to interpret them in terms of existing categories” (Lewandowska-Tomaszczyk 2007: 149).

Linguistic constructions are also subject to prototype effects (Taylor 2003: Ch. 12). Hopper & Thompson (1980), though not yet working in a prototype framework, nonetheless demonstrate that transitivity is very much a prototype category, with individual clauses showing greater or lesser degrees of transitivity depending on their features. Ross (1972) shows that lexical items are graded in their ability to undergo various transformations, with human beings being close to prototypical noun phrases, while inanimates, events, abstract concepts are quite atypical. Taylor (2003: §12.5) likewise points out that the transitive construction in English has steadily expanded its functions over time “to encode states of affairs which diverge increasingly from prototypical transitivity” (Taylor 2003: 235). The result of this diachronic development is significant gradation as to which verbs now lend themselves to transitivization. Taylor (2003: 236) gives the example of the transitive construction being used to imply a semantic path, in lieu of an explicit preposition. Compare the pairs of English sentences in (30).

(30) English (Indo-European > Germanic)

*Preposition*

He regularly **flies across** the Atlantic.

He **swam across** the Channel.

She **swam across** our new swimming pool.

We **drove across** the Alps.

The child **crawled across** the floor.

*Transitive*

He regularly **flies** the Atlantic.

He **swam** the Channel.

?She **swam** our new swimming pool.

?We **drove** the Alps.

\*The child **crawled** the floor.

(Taylor 2003: 236)

These examples illustrate that there are indeed better and worse members of the English Transitive Path construction.

Individual lexemes are also a type of construction, and therefore also subject to prototype effects. This is unsurprising, since language forces speakers to map a non-discrete cognitive representation of the world onto discrete linguistic entities—we are forced to cut up and categorize the world around us into discrete objects and events/states so that we can refer to them and predicate statements about them. Reality, however, is not so neat. The result of this mapping is a linguistic form that imperfectly demarcates a portion of our mental world, centered on a clear prototype but with imprecise boundaries. Using a topological metaphor, we typically call some portion of our mental representation of the world a *semantic space* (Finch 2003: 140), and that space can be graphically represented using a *semantic map* (Croft 2001b: §2.4.3; Haspelmath 2003). Though semantic maps are most often used to represent a *functional* space for grammatical morphemes, they are equally applicable to lexical spaces as well. Gries (2006: 74) provides one such semantic map for the meanings of the English word *run*, shown in Figure 2.4, based on a comprehensive corpus analysis. As another example, Bowerman & Choi (2001: 485) present a semantic map of spatial relations based on data from 38 languages (25 families), with a relation indicating prototypical support from below (ON) at one end and a relation indicating prototypical containment (IN) at the other. As pictured in Figure 2.5, lexical items in different languages cut up this semantic space in different ways.



Figure 2.4: Semantic map of English *run* (Gries 2006: 74)



1. English      ON      IN

2. Japanese      UE      NAKA

3. Dutch      OP      AAN      IN

4. Berber      X      DI

5. Spanish      EN

**Figure 2.5:** Crosslinguistic differences in the encoding of spatial relationships (Bowerman & Choi 2001: 485)

These examples illustrate that word meanings are polycentric and cover a range of possible uses, as mentioned in [Section 2.3.1.4](#). Some of these uses may be more prototypical than others. The English expression *apple on a twig* is a slightly less prototypical use of *on* than *apple on a table*. The fact that lexical items cover a range of uses, and that some of these uses are more prototypical than others, is an important component of the typological markedness theory of lexical categories.

Even the formal categories that linguists use to describe linguistic structure tend to be prototypical (Taylor 2003: xii, 201). Taylor (2003: §11.1) argues that linguists' conceptions of the formal labels *word*, *affix*, and *clitic* are prototypical in nature, with better and worse members of the category. Haspelmath (2005) likewise shows that simple structural definitions of these categories are inadequate and reframes the word–affix continuum in functional terms instead.

Much research in the Canonical Typology framework (Corbett 2005) also demonstrates the prototypical nature of linguists' categories. Though Corbett is careful to distinguish between a *canon* and a *prototype* / *exemplar* (Corbett 2010: 142), his accumulated work nonetheless shows that linguists view phenomena in the world's languages as better or worse instances of various descriptive categories.

What type of category are lexical categories then? Are word classes categories of human cognition, categories within particular languages, categories of languages generally, or analytic categories of linguists? Or some combination of these? Typological markedness theory posits that parts of speech like noun, verb, and adjective are not categories of particular languages. Languages have constructions, not parts of speech. Speakers, however, have *mental prototypes* of objects, actions, and properties. And although there is no *one* Noun construction in English that would correspond to the mental category of OBJECT, there are numerous constructions in English which have the function of indicating *reference to an object*, such as the Definite Article construction or the Transitive Subject construction. Likewise, there is no one construction—in English or any language—that can be definitively called the Verb construction or the Adjective construction, but there are plenty of constructions which have the function of predicating or attributing properties. Naturally, then, speakers are more likely to use referring constructions when talking about something which they mentally categorize as an object, predicating constructions when talking about something they conceive of as an action, and modifying constructions when talking about something they conceptualize as a property.

Speakers' conceptualizations, however, are fluid. Speakers often conceptualize things in non-prototypical ways. They may construe events as bounded entities that they can refer to, or objects as properties with duration. As a result, speakers often use lexical items in constructions that do not align particularly well with the item's meaning, such as the appearance of an action word like *sing* in a referring construction like the Gerund in the phrase *his singing was beautiful*. When speakers use words in this atypical manner, those uses are much

more likely to be marked in some way—whether morphologically, behaviorally, semantically, or frequently (Croft 1991: §2.2). As a consequence of this tendency, clear asymmetries emerge between the prototypical vs. non-prototypical uses of object words, action words, and property words. It is the unmarked use of these lexical items that most closely aligns with linguists’ traditional conceptions of noun, verb, and adjective. Parts of speech as traditionally conceived are nothing more than the emergent effects of our cognitive prototypes on language. They do not have any real status in grammar or individual grammars. This is the fundamental idea behind typological markedness theory. Section 2.4.2 lays out this theory in more detail.

A last clarifying point is in order. Recognizing the existence of prototype-based categories, many linguists have described parts of speech as prototypical. Dixon (2004: 1–2), for example, says that the word classes noun, verb, and adjective each have a “prototypical conceptual basis” and “prototypical grammatical functions”. Taylor (2003: 217) states, “A prototype view of NOUN entails that some nouns are better examples of the category, while others have a more marginal status.” But languages have constructions, not parts of speech, and individual constructions are not gradient (Croft 2007). What linguists are in fact observing when they say that parts of speech are prototypical is not gradation in *linguistic categories* like noun, verb, and adjective (since those are not categories of particular languages), but rather gradation in the *mental categories* of objects, actions, and properties, which do indeed exhibit prototype structures, and which therefore have emergent effects on the organization of constructions in languages.

## 2.4.2 Typological markedness theory

I have already previewed various aspects of typological markedness theory at different points in this thesis. In this section I present a concise overview of the specific claims made by this theory, and some of the evidence for those claims. The phrase *typological markedness* or *typological markedness asymmetries* simply refers to an implicational universal regarding the

behavior of basic versus non-basic members of a conceptual category. At its simplest, the theory posits that less basic or prototypical members of a category are marked in some way; basic or prototypical category members are unmarked by comparison (Greenberg 1966). This *cognitive* markedness is then realized *linguistically* in several ways. The marked member of a category *may* be literally marked with an affix or other overt morphological indicator, but this is just one of the ways an item can be a marked member of a category. The marked member of a category may also be less frequent, or have a smaller range of inflectional / distributional possibilities, or be semantically more complex. It is important to emphasize that *typological* markedness does *not* always entail *formal* markedness. Typological markedness is an *implicational* universal rather than an *absolute* universal. The more marked members of a category must be *at least as marked* as the unmarked member, but this does not preclude the possibility of all members being *equally* marked. Formal markedness is merely an emergent tendency of structures to reflect cognitive markedness.

As applied to word classes, typological markedness theory states that the most unmarked discourse functions for object, action, and property words are reference, predication, and modification, respectively. Therefore, when a lexical item is used for a function that does not align with its prototypical meaning, typological markedness theory predicts that it will be marked. Again, it must be emphasized that not *every* instance of a lexical item being used in a non-prototypical function will be marked in comparison to its prototypical function; but it will always be *at least as marked*. This theory of typological markedness for the major discourse functions is laid out in detail by Croft in various publications (Croft 1991; 2000; 2001b; Croft & van Lier 2012). It is also important to understand that typological markedness theory is *not* a theory of parts of speech in the sense of large partitionings of the lexicon into categories like noun, verb, and adjective. Instead, noun, verb, and adjective are epiphenomenal, crosslinguistic markedness patterns that arise from the interaction of semantic prototypes (object, action, property) and their use in different discourse functions (reference, predication, and modification). They are not categories of particular languages.

Throughout this thesis, I have used the term *discourse function* to refer to the functions of reference, predication, or modification. These are what Croft (1991: 51) calls *pragmatic functions* or *propositional act functions* following the tradition of pragmatics and speech act theory in philosophy (Austin 1962; Searle 1969). These three functions are taken as fundamental to human communication, arising out of the communicative intent behind what speakers are attempting to *do* with language. This perspective was articulated early on by Sapir:

There must be something to talk about and something must be said about this subject of discourse once it is selected. This distinction is of such fundamental importance that the vast majority of languages have emphasized it by creating some sort of formal barrier between the two terms of the proposition. (Sapir 1921: 87)

A similar point is made by Croft while articulating his theory of typological markedness as applied to lexical categories: “[N]o matter how complex a given situation is in terms of the number of entities involved and the number and kinds of relations that hold between them, a human being attempting to describe it in natural language must split it into a series of reference-predication pairs[.]” (Croft 1991: 124)

Modification is generally seen as less central a function than reference and predication, as illustrated by its lack of mention in the quotes above. For example, Hengeveld (1992: 55) takes the reference-predication dichotomy to be fundamental, yielding the major categories of noun and verb, while the modification function then combines with these two functions to yield the major categories of adjective and adverb, respectively. The primacy of the reference-predication distinction also appears to be reflected structurally in the world’s languages, which do not always have dedicated morphological means for encoding modification but appear to always have constructions dedicated to reference and predication.

Croft (1991: 123) defines the pragmatic functions in terms of their discourse functions, following work in the discourse-functional tradition (Chafe 1976; Hopper & Thompson 1984; Chafe 1987; Du Bois 1987). Previous research defines *referents* as “discourse-manipulable participants” (Hopper & Thompson 1984: 711; Kibrik 2011), *predicates* as reported events (Hopper & Thompson 1984: 726), and *modifiers* as a mix of these two functions (Thompson 1989). Croft



(1991: 123) synthesizes ideas from this body of research and offers the following revised definitions instead:

- the act of *reference* identifies a referent and establishes a cognitive file for that referent
- the act of *predication* ascribes something to a referent
- the act of *modification* enriches the cognitive image of the referent with an additional feature

The exact pragmatic function chosen for any given mention of a concept is then just a matter of how the speaker chooses to portray or construe that concept—whether as a referent, predicate, or modifier (Croft 1991: 100); as Croft & van Lier note, “apparent instances of ‘fuzziness’ are actually variable construals” (Croft & van Lier 2012: 63).

With this understanding of discourse functions in mind, we can restate the thesis of typological markedness theory as applied to lexical categories: Noun, verb, and adjective are epiphenomenal markedness patterns that arise from the use of different semantic prototypes (objects, actions, and properties) in different discourse functions (reference, predication, modification). Uses of these semantic classes in non-prototypical functions are typologically marked. As mentioned, there are four ways in non-prototypical uses can be marked: structurally, behaviorally, semantically, and/or frequently.

The first type of marking, *structural coding* or *formal marking*, refers to the fact that non-prototypical uses of lexical items are at least as formally marked as prototypical ones. Structural coding in this context refers specifically to “dedicated formal markers in a specific language that indicate a lexeme’s syntactic function” (Croft & van Lier 2012: 62). Figure 2.6 is a schematic representation of some of the formal realizations of these markedness patterns. It indicates the different morphosyntactic means that languages tend to develop for marking each of the non-prototypical uses of lexical items. For instance, participle constructions are one way that languages have of indicating the non-prototypical case of an action word being used for modification.

		FUNCTION		
		reference	predication	modification
MEANING	<b>object</b>	prototypical noun	predicate nominal copula	genitive adjectivalization PP on noun
	<b>action</b>	action nominal complement infinitive gerund	prototypical verb	participle relative clause
	<b>property</b>	deadjectival noun	predicate adjective copula	prototypical adjective

Figure 2.6: Typological prototypes for noun, verb, and adjective (adapted from Croft (2000: 89) and van Lier (2012: 62))

The second way in which non-prototypical uses of lexical items can be marked is in terms of their *behavioral potential*, that is, the range of combinatorial possibilities for that lexical item. This is most clearly illustrated with an example from inflection: in many languages, property words used in predicate constructions are limited in their inflectional possibilities. In Munya, for example, property words functioning as predicates cannot inflect for person and number of the subject, and cannot take the imperfective marker, perfective marker, or direct evidential marker (Bai 2019: 96–97). The only grammatical markers allowed in property predication clauses are the stative aspect marker, a clause-final particle, and an egophoric marker. Hopper & Thompson’s (1984) study of the discourse functions of different parts of speech is largely a study of behavioral potential. They conclude that “the closer a form is to signaling this prime [prototypical] function, the more the language tends to recognize its function through morphemes typical of the category—e.g. deictic markers for [Nouns], tense markers for [Verbs].” (Hopper & Thompson 1984: 703, abstract). Croft advances a cognitive explanation for these behavioral markedness patterns:

In general, only the core members of the syntactic category will display the full grammatical behavior characteristic of their category because only they have all the semantic characteristics that the characteristic inflections tap into. This is to say that the inflectional categories of the major syntactic categories have been “tailored” to their semantically core members. This is an example of a processing constraint: languages inflect only for those properties that are of relevance to core members of the category; they do not

inflect for properties of peripheral members of the category that are not of relevance to the core members of the category. (Croft 1991: 86)

Non-prototypical uses of lexical items may also be marked semantically by a *semantic shift* in their meaning towards the semantic class prototypically associated with the discourse function they are found in (Croft 2000: 96; 2001b: 73; Croft & van Lier 2012: 68). I have already discussed the semantic shifts that occur in functional expansion in some detail in Section 2.3.3.2. Croft (1991: 60–61) makes the even stronger claim that non-prototypical uses of lexical items will *always* be marked semantically, making semantic markedness an absolute rather than implicational universal.

These semantic shifts are caused by a combination of conventionalization and *coercion*, wherein the meaning of the constructional context is imposed on the meaning of the lexical item (Pustejovsky 1991; Croft 1991: 69, 108; Panther & Thornburg 2007: 252; Audring & Booij 2016). For example, predicate nominals (where an object word is used in a predicate construction) involve coercion of lexical items from denoting objects to denoting classifying or equational relations (Croft 1991: 69). In Nuuchahnulth, for instance, nominal predicates are always semantically durative and interpreted as either existential, classifying, or identifying expressions (Nakayama 2001: 47).

The final way in which lexical items used in atypical functions may be marked is in terms of their frequency. Croft (1991: 59, 87) also refers to this as *textual markedness*. Frequential markedness predicts that lexical items are used more frequently in their prototypical functions than in non-prototypical ones. This means that object words should be most frequent in their use in referring constructions, and that referring constructions should most frequently denote objects (Croft 1991: 87).

The field of linguistics has accumulated a good deal of empirical evidence in support of the typological markedness theory of lexical categories. Croft (1991) provides empirical evidence from 12 languages for each of these markedness patterns. Dixon (1977) also provides evidence of typological markedness patterns as they relate to property words, using a com-

bination of structural and behavioral evidence. As mentioned, Hopper & Thompson's (1984) study also provides empirical support from a variety of languages for markedness in terms of behavioral potential. Stassen (1997) is a massive study of intransitive predication in 410 languages, demonstrating the marked behavior of non-action words when used in predicate constructions.

Having explicated the basic tenets of typological markedness theory, I now turn to re-framing the concept of lexical flexibility in a way that utilizes this framework.

## 2.5 Lexical flexibility: A functional definition

Within the framework of typological markedness asymmetries, lexical flexibility can be understood as follows:

**lexical flexibility** The use of a lexical item (root, stem, or inflected word) in more than one discourse function (reference, predication, or modification) with no difference in structural coding.

This definition qualifies as a valid *comparative concept* in the sense of Haspelmath (2010a) because it is couched in terms of universal *functions* rather than language-specific *structures* (Croft 2016). It also has the advantage of being intentionally equivocal with respect to the lexical and cognitive unity of the item, and with respect to the linguistic level (root, stem, or inflected word) at which the flexibility is realized. In some cases when a single lexical form appears in more than one discourse function, speakers may have a close cognitive association between the two uses. This is most likely the case for the predicative and referential uses of the word *run* in the phrases *I run every morning* and *I'm going for a run* respectively. In other cases, speakers may have little to no awareness of the diachronic connection between uses of a form. For example, the use of *run* in the sense of *to run a print job* is extremely distant from the prototypical "fast pedestrian motion" sense in the semantic network for that form (Gries 2006: 74; see also Figure 2.4). It is unlikely that these two senses are closely cognitively connected by

most speakers, even though they both share a predicated function. The definition of lexical flexibility given here allows for any degree of semantic shift. Croft admits this possibility explicitly: “It of course a priori possible to construct a typological classification of parts-of-speech systems using only structural coding and allowing any degree of semantic shift.” (Croft 2001b: 68) Of course, I am not concerned here with constructing a classification of parts of speech—quite the opposite, in fact. This definition of lexical flexibility is intended to delimit exactly those cases where a language does *not* provide formal indicators of discourse function.

Allowing for any degree of semantic shift does *not* imply that semantic shift is in any way unimportant for understanding lexical flexibility. On the contrary, semantic shift is a key component of the process of functional expansion that leads to flexibility in the first place (see below). Moreover, carefully circumscribing the phenomenon of lexical flexibility without regard to the degree and type of semantic shifts involved puts us in a position to then compare the semantic shifts that occur in cases of lexical flexibility with those that occur in cases of overt derivation. This raises the intriguing question of whether semantic shifts in flexible cases differ in principled ways from overt derivation. Mithun (2017: 165) shows that for Central Alaskan Yup’ik the types of semantic relationships between flexible uses of words mirror those between basic and derived words. This would suggest that functional expansion follows the same principles as overt derivation. However, much more research is needed in this area.

As we have seen, a great abundance of evidence also shows that the meaning of any given combination of form and discourse function is a matter of convention, and often highly idiosyncratic (§2.3.2.4; §2.3.3.2). This fact suggests that flexible items are not truly “flexible” in the sense that speakers can use any lexical item for any discourse function and expect hearers to be able to infer their meaning from context. We know that item-specific gaps in usage exist. Certainly, novel cases of forms being used in new discourse functions do occur, or else it would not be possible for functional shift to happen in the first place. But these cases are necessarily restricted by the cognitive limits on our ability to deal with ambiguity.

If it were truly the case that any lexical item could be used in any discourse function at any time, it would barely be possible for hearers to interpret the intended pragmatic effects of each word. Instead, flexibility must rely on a degree of *conventionalization*. Conventionalization in turn implies *time*—conventionalization is a diachronic process. Thus, *lexical flexibility is best understood as a synchronic pattern resulting from the diachronic process of functional expansion*, where functional expansion is defined as follows:

**functional expansion** A diachronic expansion in the use of a lexical item (root, stem, or inflected word) into a new discourse function (reference, predication, or modification) with no additional structural coding.

Cases of lexical flexibility therefore arise whenever a new combination of form and discourse function is conventionalized in a community of speakers. This understanding of lexical flexibility is in line with cognitive research on lexicalization and constructional change. Functional expansion occurs because of speakers' need to construe concepts in different ways—as objects, actions, or properties. The semantic shifts that occur during functional expansion are the result of coercion by the new constructional context, and the resultant meaning then becomes conventionalized as the meaning of that particular form in that particular discourse function (Croft 1991: 108).

If lexical flexibility is the result of a diachronic process, it should be possible to enumerate some of the specific pathways which give rise to it. Here I will mention just a few. One pathway is insubordination, whereby subordinate clauses in a language are reanalyzed as main clauses (Evans 2007; Mithun 2008; Evans & Watanabe 2016) (see also §2.3.2.2). Insubordination frequently results in formal similarities between noun phrases and verb phrases, and this formal ambiguity can abet the functional expansion of lexical items from referential to predicative uses and vice versa.

A second pathway to lexical flexibility is relexicalization (or more precisely, reconventionalization). This is the process that occurred in the case of morphological verbs being reanalyzed as nouns in many North American languages (see §2.3.2.3) and certain English

plurals like *brethren* or *arms*. In these cases, the conventionalized meaning associated with the form changed (e.g. from Cayuga ‘it hauls logs’ to ‘horse’), so that the meaning more closely reflected its new discourse context.

A third pathway is topicalization, exemplified in the Wakashan family. Jacobsen (1979: 122, 142) observes the formal similarity between the Definite Article and the Third Person Singular Indicative markers in Wakashan languages, and argues for a diachronic connection between the two. It is likely that cleft constructions such as ‘it was the dog that ran’ became so common that speakers started to reanalyze the topicalized cleft as a definite noun phrase, ‘the dog’, thereby creating a formal similarity between referring expressions and predicating expressions.

Each of these pathways results in the functional expansion of lexical items into new discourse contexts with no new overt structural coding. Of course, functional expansion can also occur without any other accompanying grammatical changes. This happens in any instance where speakers simply use stems in new discourse functions. Lexical flexibility is the natural and expected result of the fact that non-prototypical uses of lexical items are *not* always structurally marked—as allowed for by the fact that typological markedness is implicational and not absolute—even if they are marked in other ways. The use of additional structural coding in cases of functional expansion is not obligatory, but merely a statistical tendency. Lexical flexibility occupies the theoretical space where structural coding asymmetries fail to apply.

When viewed in this light, *lexical flexibility is not so much a problem as it is a design feature of language*. The presence of lexical flexibility should be *expected* in every language, not treated as exotic. The cognitive-typological approach outlined in this chapter inverts the lexical flexibility question: the interesting question is not why some languages fail to make distinctions in parts of speech (thereby framing lexical flexibility as a *deficit* in a way similar to colonial researchers), but rather why languages develop specialized constructions for different discourse functions in the first place (see Gil [2012] for an attempt to answer this question for

predication). Lexical flexibility exists in any area of the grammar where specialized function-indicating morphology has yet to develop, or where such distinctions have been leveled via diachronic changes. Flexibility should therefore be considered the default state of affairs for language. Gil (2005; 2006) has in fact argued, partially on the basis of data from highly flexible languages, that early human language must have been *isolating-monocategorical-associational* before the development of dedicated function-indicating morphology.

The idea that the “natural state” of language is monocategorical or acategorical would seem to conflict with the point made above that lexical flexibility can result from diachronic processes, but the two positions are not mutually exclusive. Languages develop constructions which indicate different discourse functions, but languages are also subject to counteracting pressures. This is a classic case of competing motivations: on the one hand, the frequency with which speakers need to perform the discourse functions of reference, predication, and modification all but ensures the development of constructions dedicated to indicating those functions; on the other hand, speakers need to construe states of affairs in various ways—as objects, actions, or properties—creating pressures which have the potential to level or cut across those formally marked distinctions. Reconventionalization and the reanalysis of cleft constructions could both be viewed as diachronic processes motivated by this latter pressure.

In sum, then, lexical flexibility is a natural result of the cognitive and diachronic forces at work in language. Defining lexical flexibility in terms of typological markedness (or more accurately, the lack of formal marking for otherwise marked uses) provides a crosslinguistically applicable definition of the phenomenon which avoids methodological opportunism while still recognizing that lexical flexibility requires some degree of semantic shift and conventionalization. This cognitive-typological definition of lexical flexibility is the primary theoretical contribution of this thesis. With this definition in place, the remainder of this thesis turns to exploring the prevalence of lexical flexibility in English and Nuuchahnulth and the semantic behavior of lexical items in cases of functional expansion.



# Chapter 3

## Data & Methods

This chapter describes the data used for this study and how those data were analyzed. It covers the languages chosen, the corpora used, and how samples from each corpus were created and annotated [Section 3.2](#). I describe the methods used to annotate the data, and the factors that influenced annotation decisions [Section 3.3](#). I also discuss the specific statistical measures used in this study in [Section 3.4](#). [Section 3.4.1](#) introduces a metric for quantifying the flexibility of individual lexical items in a corpus. This formulation of lexical flexibility is a key methodological contribution of this thesis. [Section 3.4.2](#) explains how I examine the relationship between lexical flexibility and corpus size, by comparing average flexibility ratings for differently-sized portions of the same corpus. Finally, [Section 3.4.3](#) presents and motivates a measure of corpus dispersion (Deviation of Proportions, or *DP*) that is used partly in place of, and partly as a complement to, raw token frequencies.

### 3.1 Introduction

The process of collecting, annotating, and analyzing the data for this study adheres to several self-imposed principles. First and foremost, the data in this study consist of naturalistic discourse data rather than elicited data. This principle has two motivations: First, as discussed in [Section 1.2](#), few studies examine token frequencies of lexical items used for different discourse

functions, and those that do only report aggregated results. Most extant research consists of lexicon-based counts. This study therefore explores a previously unexamined aspect of lexical flexibility. Second, corpus-based methods study real-world instances of language in use, rather than made-up examples or examples produced by introspection, which are subject to various cognitive and social biases (P. Baker 2018: 168). Corpus data are also more likely to reveal prototype effects through statistical tendencies. For this study, I rely on specialized corpora of spoken narratives and conversational texts only. This ensures greater comparability between the corpora used in this study and other documentary corpora that these methods may be applied to in the future, since most documentary corpora likewise consist of spoken narratives and conversations.

The second self-imposed requirement for this study is adherence to the [Austin principles of data citation in linguistics](#) (Berez-Kroeker et al. 2018). In particular, the source for each data point discussed in this thesis is uniquely identified with its location in the corpus, and the data used in this study are made freely available on GitHub at <https://github.com/dwhieb/dissertation>. All of the data and my annotations on that data may be viewed there.

Finally, as a matter of scientific accountability, this study is designed to be replicable using the same or other datasets. All of the technical details regarding how to acquire the data, annotate it, and run statistical analyses for those data are documented in the GitHub repository for this project, which may be viewed at <https://github.com/dwhieb/dissertation>.

The remainder of this chapter details the methods used to answer each of the major research questions presented in [Chapter 1](#). The core empirical question addressed by this study is [R1](#): “How flexible are lexical items in English and Nuuchahnulth?” The other two research questions build on this one. To answer this core question, I count the frequency with which stems are used for each of the three functions of reference, predication, and modification in corpora for each language. [Section 3.2](#) describes the corpora used, where to acquire the data, and how lexical items in the corpora were selected for annotation. [Section 3.3](#) describes the

details of this annotation procedure. Finally, [Section 3.4](#) explains the specific statistical measures used in this study. [Section 3.4.1](#) describes how to use the annotated data to calculate a measure of lexical flexibility for each of the lexical items in the sample. This procedure for quantifying lexical flexibility based on corpus data is the primary methodological contribution of this thesis. [Section 3.4.3](#) then discusses some shortcomings in the use of token frequencies, and presents a measure of corpus dispersion (Deviation of Proportions, or *DP*) as an alternative.

## 3.2 Data

In [Section 1.3](#), I discussed the motivations for using English and Nuuchahnulth as the languages of focus in this study. Both languages have featured prominently in the literature on lexical flexibility, with researchers taking opposite positions as to their overall flexibility. For English, I opted to use the [Open American National Corpus](#) (OANC), a 15-million-token open access corpus of American English (Ide & Suderman 2005). I restricted my analysis to just the spoken portion of the corpus, comprising approximately 3.2 million tokens, so that the data would be comparable to the spoken corpus of Nuuchahnulth and other documentary corpora. The spoken portion of the corpus itself consists of two distinct subcorpora—the [Charlotte Narrative & Conversation Collection](#) (the “Charlotte corpus”) and the [Switchboard Corpus](#). The Open American National Corpus can be obtained for free at <http://www.anc.org/>.

The data for Nuuchahnulth come from a documentary corpus compiled by Toshihide Nakayama and published in Little (2003) and Louie (2003). The corpus consists of 24 texts by two speakers (Caroline Little and George Louie), containing 2,081 utterances and 8,366 tokens. The texts are personal narratives, traditional stories, and procedural texts. I manually retyped the corpus in [scription](#) format (Hieber 2021a), which is a simple way of formatting interlinear texts so as to make them computationally parseable. I then converted the corpus to the Data Format for Digital Linguistics (DaFoDiL) (Hieber 2021b), which is a way of repre-

senting interlinearized data in JSON, allowing programmers to easily and programmatically work with linguistic data. The resulting corpus is available in both formats on GitHub at <https://github.com/dwhieb/Nuuchahnulth>.

The sheer size of the Open American National Corpus—even when considering just the smaller, spoken portion of 3.2 million tokens—made it practically impossible to tag every token in the corpus for its discourse function for the time being. At the opposite end of the spectrum, the Nuuchahnulth corpus is small enough (~8,300 tokens) that it was possible to tag every single lexical token in the corpus. Given this size disparity, it was important to sample lexical items from each corpus in such a way as to make them reasonably comparable. I did this by extracting two kinds of samples from each corpus: 1) a 100-item sample of lexemes randomly selected from different frequency bins, and 2) a small corpus sample (<10,000 tokens) for which all lexical items in the sample were annotated.

To create the 100-item samples, I first *lemmatized* each corpus. For every lexical token in the corpus, I programmatically determined the lemma associated with that particular wordform. For example, the English wordforms *knows* and *knew* are associated with the lemma *know*. For English, lemmatization was accomplished with the [Natural Language Toolkit](#) for Python (Bird, Klein & Loper 2009), using the Wordnet lemmatizer. The OANC includes Penn tags for parts of speech, so I was able to use those part-of-speech tags with Wordnet’s `lemmatize()` method to improve lemmatization. For Nuuchahnulth, lemmatization simply involved programmatically stripping away the inflectional morphology from each token, leaving just the stem. For example, token in (31) is lemmatized as an instance of the stem *?am-umt-* ‘first-be.born’. Since the entire Nuuchahnulth corpus is interlinearized with glosses and stored in DLx JSON format (Hieber 2021b), this was accomplished with a simple Node (JavaScript) script.

- (31) Nuuchahnulth (Wakashan > Southern Wakashan)  
       ?aamumʔaʔquu  
       ?am-umt-’aʔ-qu:  
       first-be.born-FIN-COND  
       when.first.born

‘when [a baby] was born’

(Little 2003: Afterbirth 1)

It is worth mentioning that annotating the Nuuchahnulth corpus for discourse function would not have been practical without the detailed descriptive work of Toshihide Nakayama. The creation of text collections is often underappreciated as a worthwhile academic endeavor, but this is the only way we gain new corpora of naturalistic discourse for minority languages. The empirical and theoretical findings of this project would not have been possible without this important work.

After lemmatizing each corpus, I calculated the raw frequencies for each lexeme. I then grouped lexemes into 100 bins based on their frequencies, and randomly selected one lexeme from each bin. This produced a sample of lexemes from a range of different frequencies. The frequencies of lexemes in the English sample, for instance, ranged from 44,687 for the word *know* to 53 for the word *central*. Lexemes with a frequency  $< 4$  were excluded, because the lexical flexibility measure described in [Section 3.4.1](#) requires a minimum token frequency of 4 in order to return a statistically significant value.

Various other types of words were excluded from this process as well:

- words written using numeric characters (e.g. *12%* or *117*)
- obvious cases of code-switching or code-mixing (e.g. *union manči?aλ* ‘became a union man’)
- transcategorial words (those with both lexical and grammatical uses) (e.g. *be*, *do*)
- discourse markers (e.g. *uh*, *well*)

Some types of items that were *not* excluded are compounds written as a single word (e.g. *guidepost*) and proper names (e.g. *San Francisco*), although neither of these wound up in the final list.

The output of this selection process was a list of 100 lexical items in each language to be examined for lexical flexibility. The list of 100 lexical items for each corpus is given in [Appendix B](#), along with statistics about their frequencies, corpus dispersions, and flexibility. I then created a list of every instance of these 100 lexical items in each corpus. For English,

this resulted in a list of 382,512 tokens to be annotated. For Nuuchahnulth, there were just 1,632 tokens to annotate. I annotated each one of these approximately four hundred thousand tokens for discourse function by hand. This procedure is described in the following section.

Having created the 100-item samples, I next created a small corpus sample (<10,000 tokens) for each language. The smaller size of these samples allowed me to annotate every single lexical item in the sample for its discourse function. The Nuuchahnulth sample simply consists of the entirety of the corpus (8,300 tokens), while the English sample consists of the first four texts in the corpus, totaling ~9,700 tokens. These two subcorpora are both available in the GitHub repository for this study at <https://github.com/dwhieb/dissertation>.

With the two samples prepared, I next turned to the process of annotating each lexical item in the sample for its discourse function. This annotation procedure is described in the following section.

### 3.3 Methods

Within each of the samples, not every token was annotated for its discourse function. This section discusses the various reasons why tokens might be excluded from the analysis, and the factors that contribute to the determination of the discourse function for each token.

First, I only annotated lexical uses of words. Grammatical/functional words and discourse markers were ignored. Among lexical words, adverbial uses were also excluded. Ignoring adverbial uses of words sometimes results in lexical items with a very high overall corpus frequency, but very low occurrences of use for reference, predication, or modification. For example, the English word *never* has a high overall frequency (3,024 tokens), but has exactly 1 modifying use (*that's a **never** touch*). The rest of its uses are adverbial. Proper names were included, a decision which turned out to be fortuitous since proper names displayed flexible, non-referential uses in both English and Nuuchahnulth, as in (32) and (33).

(32) English (Indo-European > Germanic)  
they settled down in the **Chicago** suburbs

(Ide & Suderman 2005: JamiesonSean)

(33) Nuuchahnulth (Wakashan > Southern Wakashan)

q <sup>w</sup> aa	ýuuq <sup>w</sup> aa	wiikinanišitquu
q <sup>w</sup> a:	ýu:q <sup>w</sup> a:	wi:kinaniš-it-qu:
thus	also	NAME-PAST-COND.3
thus	also	who.was.Wiikinaniš

‘So was the one whose name was *Wiikinaniš*’

(Louie 2003: GL 19)

The function of each lexical item was determined in relation to its most immediate syntactic constituent. As an illustration, consider how to analyze the word *time* in the phrase *all time favorite*. The phrase *all time* is functioning to modify the referring expression *favorite*, with the syntactic structure  $[[all\ time]\ favorite]$ . However, within the context of *all time*, the word *time* is a referent, not a modifier. Compare this to the expression *all time slots*, which has the syntactic structure  $[all\ [time\ [slots]]]$ , and where *time* is indeed modifying the referent *slots* directly. Therefore I annotated *time* as a referent in the phrase *all time<sub>REF</sub> favorite* and as a modifier in the phrase *all time<sub>MOD</sub> slots*. As another example, when annotating tokens of the word *woman* I excluded its appearance in the phrase *anti-women statements*, because it forms one part of the complex word *anti-women*, with the structure  $[[anti-women]_{MOD}\ statements]$ . If the phrase had been just *women statements* instead, I would have analyzed *women* as a modifier.

In the remainder of this section I discuss some analytical issues specific to English and Nuuchahnulth respectively. The following points are specific to English:

- Words related through stress shifts (e.g. *con'duct* and *'conduct*) were treated as separate lexical items since their phonological forms are distinct. In the corpus, context always made it possible to determine which use was intended.
- Compound words were included in the analysis, but individual components of compound words were not. For example, when annotating tokens of the word *back*, instances within the compounds *back yard* or *hard back book* or *back burner* were excluded from the analysis. Instances of lexical items within noun-verb compounds (“noun incorporation”) were also excluded, such as *pie* in *pie baking*. However, compound words as a whole were included in the analysis. For example, the lexical item *back yard* was treated as a lexical unit and analyzed for its discourse function. Therefore I analyze *back yard* as a referent in *we were sitting in the [back yard]<sub>REF</sub>* and a modifier in *it was a [back yard]<sub>MOD</sub> party*.

- Lexicalized phrasal verbs such as *back up* were treated as a lexical unit, such that it was possible for the lexical item to appear in different discourse functions: *he doesn't [back up]<sub>PRED</sub> that point* vs. *please make a [back up]<sub>REF</sub>* vs. *you have a fairly good [back up]<sub>MOD</sub> quarterback*.
- Tokens used as gerunds, infinitives, or predicate nominals / adjectives were tagged separately and ultimately excluded from the analysis, since most researchers would consider these to be instances of morphologically marked conversion in English.
- Adverbial uses of participles similar in function to the Latin ablative absolute were excluded from analysis, e.g. *talking about the golf thing, [...]*.
- Stative (modification) versus dynamic (predicational) uses of past participle forms required special consideration. It was not always possible to discern with certainty whether a given token of a past participle form was being used statively or dynamically. Compare the use of the word *relieved* in the phrases *she was relieved of duty* vs. *she was relieved to find her car*. The first use is arguably predicative while the second seems more like a predicate adjective. In cases where the discourse context does not make the intended use clear, I opted to code the data as a predicate, since this is the more conservative, historically prior form. Stative, predicate adjective uses were excluded from the analysis.

The analysis of discourse functions in Nuuchahnulth faces a different set of issues. A first problem is that Nakayama's (2001) grammar is inconsistent in the application of the notions of reference, predication, and modification. He says, for example, that "[i]n *modification* one predicate restricts the interpretation of the other semantically main predicate." (Nakayama 2001: 113) This implies that a word in Nuuchahnulth can simultaneously be both a modifier and a predicate. This conflation arises, I believe, from the holophrastic nature of Nuuchahnulth, in which it is extremely common for a single word to constitute an entire clause (52.2% of the time according to Nakayama [2001: 149]). While an individual lexical item may be functioning as a predicate within its clause, the clause itself may be functioning to refer or to modify. Since the inflected word and the clause are coterminous, however, the potential for ambiguity arises. This problem is exacerbated by the fact that, even though Nuuchahnulth is highly polysynthetic, it is nonetheless quite common for stems to appear with no inflectional morphology indicating their discourse function. To the researcher not familiar with Nuuchahnulth morphosyntax and discourse patterns, it can seem at first glance as though determining clausal boundaries with any certainty in the language is near impossible.



Thankfully, this impression is just superficial. While there are indeed tokens that are ambiguous as to their discourse function, this is generally not the case. Converging evidence from morphology, word order, topic continuity, word-level translations, and utterance-level translations is typically sufficient to determine the discourse function of any token with a high degree of confidence. The following paragraphs briefly summarize the relevant factors for determining the discourse function of a given token.

Two features of Nuuchahnulth grammar in particular are extremely helpful in determining the discourse function of words. First, Nuuchahnulth is strongly predicate-initial. When an argument is present, the predicate precedes the argument 84.9% of the time (Nakayama 2001: 149). Arguments only precede their predicates in pragmatically marked situations like contrast or disambiguation, which is typically made clear by an accompanying topicalization construction in the English translation. Second, Nuuchahnulth speakers have a strong dispreference for using more than one argument in a clause. In a sample of 734 clauses, only 39 (3.8%) have two arguments, and none have three (Nakayama 2001: 149). This disinclination is so strong that speakers often express a single event in successive clauses, repeating the predicate (Nakayama 2001: 75). Consider the examples in (34a) and (34b).

(34) Nuuchahnulth (Wakashan > Southern Wakashan)

- a. **hinaaciʔaʔ**                      ʔaʔuukʷiʔaθ                      **hinaaciʔ**  
**hin-a·ci(ʔ)-aʔ**                      ʔaʔu:kʷiʔaθ                      **hin-a·ci(ʔ)**  
**there.MOM-go.out.to.meet-FIN**                      Clayoquot                      **there.MOM-go.out.to.meet**  
**went.out.to.meet**                      Clayoquot                      **went.out.to.meet**
- minwaaʔaθʔi  
minwa:ʔaθ-ʔi  
British.soldiers-DEF  
the.British.soldiers
- ‘The Clayoquots went [in their canoes] out to sea to meet the British soldiers.’  
(Nakayama 2001: 75)
- b. **sukʷiʔ**    ʔawituk    ʔaʔuukʷiʔaθ    [...]    **sukʷiʔ**    miimixt  
**sikʷi(ʔ)**    ʔawit-uk    ʔaʔu:kʷiʔaθ    [...]    **sukʷi(ʔ)**    mi:mixt  
**take**    chief-poss    Clayoquot    [...]    **take**    NAME  
**take**    their.chief    Clayoquot    [...]    **take**    NAME
- ‘The Clayoquot chief took Miimixt.’  
(Nakayama 2001: 75)

In (34a), the arguments *λaʔuuk<sup>wi</sup>ʔath* ‘Clayoquot’ and *minwaaʔathʔi* ‘the British soldiers’ are distributed over two clauses, with the predicate *hinaačiλ* is repeated in each clause. Example (34b) follows a similar pattern. Awareness of just these few abovementioned facts does most of the work of determining the discourse functions of words by establishing the predicate and referent in each clause.

Certain inflectional markers, when present, also unambiguously indicate the discourse function of the word they appear with. Words which take the definite suffix *-ʔi* (glossed as DEF) or one of the relative suffixes (glossed as REL) always function to refer. Except when they co-occur with either the definite or relative markers, the following kinds of mood suffixes always indicate a predicate. In Nuuchahnulth, most mood suffixes are fused with the following person suffixes, so each of the suffixes in this list has multiple realizations depending on the person and number of the clausal arguments.

- conditional (COND)
- dubitive (DUB)
- imperative (IMP)
- indicative (IND)
- interrogative (INTER)
- purposive (PURP)
- quotative (QUOT)
- subordinate (SUBORD)

In Nuuchahnulth, verb serialization is quite common, and the above mood suffixes only appear on the first (main) stem in a serial verb construction (Nakayama 2001: 42). Main predicates are also predominantly marked for person even if mood marking is not present (over 90% of main predicates in the first person) (Nakayama 2001: 29). Aspect markers, however, are not a completely reliable indicator of predication. Though it happens infrequently, aspect markers may occur with referents or modifiers as well (see Section 4.6 for more details).

Certain distributional behaviors also abet identification of the discourse function of a word. Nakayama notes the following in regard to referents: “Nominals can be modified with expressions of property concepts, quantity, or quantifiers, but not directly with qualifying expressions like *hiik<sup>wat</sup>* ‘almost’ or *?anat’uu* ‘barely’.” (Nakayama 2001: 49). Syntactic patterns are also helpful: Negation is accomplished by means of a negative predicate *wik-*, which takes another predicate as its complement. Modifiers generally precede their heads, whether the head is a referent or predicate. In serial verb constructions, only the main predicate takes person and mood marking, and the other members of the serialization immediately follow the main predicate as bare stems.

Finally, discourse-level considerations play an important role in determining the pragmatic function of each word. Most helpful is topic continuity, wherein a referent is already established in the discourse. This is accomplished either directly via an overt referent, or indirectly via inflectional affixes or features on a word that imply the existence of an referent (what Kibrik (2011) calls *referential aids*). Each successive lexical item encountered in a text must be interpreted in the context of the previously established discourse referents, so that certain interpretations of the item are much more sensible than others. Lastly, in a few particularly ambiguous cases, I consulted the audio files accompanying the corpus in order to prosodic information into account. Clear prosodic breaks in the discourse help to determine clausal boundaries.

Small annotated extracts from each corpus are given in Appendix C in order to illustrate the resulting annotations. While the actual annotations are stored in JSON format, these extracts are presented in a more human-readable format instead. The discourse function of each token is written as a subscript (REF, PRED, or MOD). Tokens without their discourse function indicated were excluded from the analysis for one of the reasons mentioned above.

## 3.4 Analysis

This section discusses the specific statistical measures used in this study. In [Section 3.4.1](#), I present the measure used to quantify the lexical flexibility of individual items in a corpus, and in [Section 3.4.3](#) I discuss the use of token frequencies versus dispersion.

### 3.4.1 Measuring lexical flexibility

Once the lexical tokens in a corpus are annotated for their discourse functions, it is possible to calculate the flexibility of each lexical item using a measure known as Shannon’s diversity index. This section summarizes the rationale for using this metric and the procedure for calculating it.

Intuitively speaking, a lexical item is most flexible when it is used with equal frequency for reference, predication, and modification. A perfectly flexible lexical item which appears 300 times a corpus would therefore have a distribution like that in [Table 3.1](#). By contrast, a perfectly rigid / inflexible lexical item with the same overall frequency would have a distribution like that in [Table 3.2](#). What is needed is a metric that captures how evenly distributed the tokens of a lexical item are across the different discourse functions. A perfectly flexible item like that in [Table 3.1](#) should receive a high rating (say, 1), while a perfectly rigid item like that in [Table 3.2](#) should receive a low rating (say, 0).

**Table 3.1:** Distribution of discourse functions for a perfectly flexible lexical item

lexical item	reference	predication	modification
<i>stem</i>	100	100	100

I elected to use Shannon’s diversity index ( $H$ ) for this purpose (Shannon [1948](#); [1951](#)). Originally devised as a measure of entropy in text (uncertainty or information content), the Shannon index has also become a popular measure of species diversity in ecology (Avolio et al. [2012](#)) and attention diversity in political science (Boydstun, Bevan & Thomas [2014](#)).

**Table 3.2:** Distribution of discourse functions for a perfectly rigid/inflexible lexical item

lexical item	reference	predication	modification
<i>stem</i>	300	0	0

Here I am using it as a measure of the functional diversity of lexical items. The normalized version of Shannon’s  $H$  yields a value between 0 (low diversity) and 1 (high diversity). For a categorical variable with  $n$  possible values,  $H_{norm}$  is calculated using the formula in (35), where  $p_i$  corresponds to the percent frequency of the  $i^{\text{th}}$  possible value of the variable.

$$(35) \quad H_{norm} = \frac{-\sum_{i=1}^n (p_i \cdot \ln p_i)}{\ln n}$$

For this study,  $n$  will always be 3 (reference, predication, and modification). Future researchers may wish to adjust this number depending on the number of discourse functions examined (for example, if the predicate modifier function were included).

Frequently there will not be any instances of a lexical item being used in one discourse function or another. Since  $\log 0$  is undefined, the above formula cannot be resolved in these cases. One common workaround to this problem is to increment the frequencies of each discourse function by 1 before performing the calculation. Another is to simply treat  $\log 0$  as equal to 0 (Gries 2013: 120–121). I use the latter procedure in this study.

Applying Shannon’s  $H$  to the fabricated data in Table 3.1 and Table 3.2 produces the desired results: a value of 1 for  $H$  in the perfectly flexible case and a value of 0 in the perfectly rigid case.

One limitation of the Shannon diversity index as applied to this study stems from the fact that there are so few discourse functions under consideration (just three: reference, predication, and modification). This means that at low frequencies there are a limited number of possible values of Shannon’s  $H$ . For example, a lexical item with a frequency of 2 will either have an  $H$  value of 0 or .63, because there are only two ways those tokens can be distributed across discourse functions (2 0 0 or 1 1 0). A lexical item with a frequency of 3 will have an  $H$

value of 0, .58, or 1, because there are only three ways those tokens can be distributed across discourse functions (3 0 0, 2 1 0, or 1 1 1), and so on.

To address this issue, I only included lexical items in the samples that had a raw frequency of at least 4. This cutoff was established based on the fact that 4 is the smallest frequency that can theoretically return a significant result for Shannon’s  $H$  when a lexical item is maximally flexible, in one of the two ways one can compute a multinomial test (probabilities vs. a  $\chi^2$  test).

Shannon’s  $H$  was calculated for each of the lexical items in the samples from both corpora to produce a flexibility rating for each item. The resulting flexibility ratings for the 100-item samples are provided in [Appendix B](#).

One final methodological point is merited: many common constructions recognized by all speakers nonetheless do not appear in even a 1.5-million-word corpus. For example, in the spoken portion of the Open American National Corpus, the word *hate* occurs as a predicate and a modifier but never as a referent. We know that referential uses of *hate* are possible (for instance in phrases like *five-minute hate* or *don’t spread hate*), but they are not attested in the OANC. As a consequence, the English stem *hate* in this study shows no flexibility in the reference dimension, even though we know such cases are possible. The flexibility ratings in this study are necessarily approximations, based on a representative sample.

### 3.4.2 Flexibility and corpus size

As discussed in [Section 1.3](#), some researchers suggest that flexibility should increase as a function of corpus size. The intuition behind this claim is that the larger the corpus, the more likely there are to be flexible uses of any given lexical item. This is the basis for [R2](#), “Is there a correlation between degree of lexical flexibility and size of the corpus?”. To test this claim, I took differently-sized samples of each corpus, calculated the flexibility of the lexical items in each sample, and then examined whether that flexibility changed as a function of corpus size.

Because of the drastically different sizes of the two corpora, I used slightly different techniques for comparing sample sizes in English and Nuuchahnulth. For English, I calculated the flexibility of words in the 100-item sample for corpus sizes in multiples of 10,000. That is, I calculated the flexibility of the 100 items for a corpus of 10,000 tokens, and then the flexibility of those same items in a corpus of 20,000 tokens, then 30,000, and so on. For Nuuchahnulth, I calculated the flexibility of corpus sizes in multiples of 1,000 instead. Additionally, for Nuuchahnulth I included every lexical token, not just the stems in the 100-item sample.

After gathering the flexibility data for different corpus sizes, I applied a generalized additive model (GAM) to examine the relationship between the two variables. I present the results of this model in [Section 4.4](#).

### 3.4.3 Frequency vs. dispersion

Research question [R3](#) asks, “Is there a correlation between degree of lexical flexibility for a lexical item and its frequency?”. The intuition behind the notion of frequency, however, can be understood and quantified in different ways. In this study I examine two different metrics and their relationship to lexical flexibility: relative token frequency and corpus dispersion. This section describes the rationale and procedures for each of these metrics.

*Token frequency* is by far the most common statistic used in corpus linguistics (Gries [2008](#): 403), and is central to usage-based theories of language (Bybee [1985](#); Tomasello [2003](#); Goldberg [2006](#); Bybee [2007](#); [2010](#); Diessel [2019](#)). It is computed by simply counting the number of instances (tokens) of a lexical item in a corpus. When working with multiple corpora it is important to normalize this statistic because the sizes of corpora vary. An item that occurs a large number of times in a million-word corpus may nonetheless be relatively infrequent compared to other items in the corpus. In order to compare the English and Nuuchahnulth corpora (which are drastically different in size), I report both the raw token frequency of lexical items as well as their *relative token frequencies*, calculated as the number of occurrences per 1,000 tokens in the corpus. Both metrics are reported for each lexical item in the 100-item

Adjust this paragraph based on what you wind up doing. You might for example take the average flexibility ratings, but show the error bars on the graph.

samples in [Appendix B](#).

Token frequencies can be misleading, however (Gries 2008; 2021; [forthcoming](#)). There is often a great deal of within-corpus and between-corpus variability in the frequency of a lexical item. Moreover, words with the same token frequencies may differ significantly in how evenly distributed or dispersed they are in a corpus. For example, while the words *enormous* and *staining* both occur 37 times in the Brown corpus, all 37 instances of *staining* are clustered within just one corpus part. By contrast, the tokens of *enormous* are distributed mostly evenly across 36 corpus parts, with 35 of those parts containing a single use of *enormous* (Gries 2021: 100).

Disparities between token frequency and dispersion are especially common for lexical items in the middle frequencies (between 1,000 and 10,000 tokens), as demonstrated in [Figure 3.1](#) from Gries (2021: 112). In this plot, word frequency is shown on the x-axis (logged to the base of 10), and dispersion is shown on the y-axis (measured using *Deviation of Proportions* (*DP*); see below for details). Each word in the corpus is represented by a gray point. Lexical items are divided into 10 bins based on frequency, and the blue whisker in each bin represents the range of dispersion values in that frequency bin. The plot makes clear just how widely words within the same frequency bin can vary in terms of their dispersion, especially in the middle frequencies.

If what we are intending to capture with these statistics is some idea of the regularity with which speakers encounter a word, it is clear that raw frequency is a deceptive measure. Instead, recent work has shown that *corpus dispersion*—how evenly an item is distributed in a corpus—more accurately represents frequency of exposure or lexical access (Gries 2008; 2010; [forthcoming](#)). Corpus dispersion correlates more strongly with reaction time data than does frequency, for example (Gries [forthcoming](#)).

Thus for this project I report a measure of corpus dispersion in addition to relative token frequency. I use a measure called *Deviation of Proportions* (*DP*), created by Gries (2008). In a review of various measures of corpus dispersion, Gries (2008) discusses shortcomings with



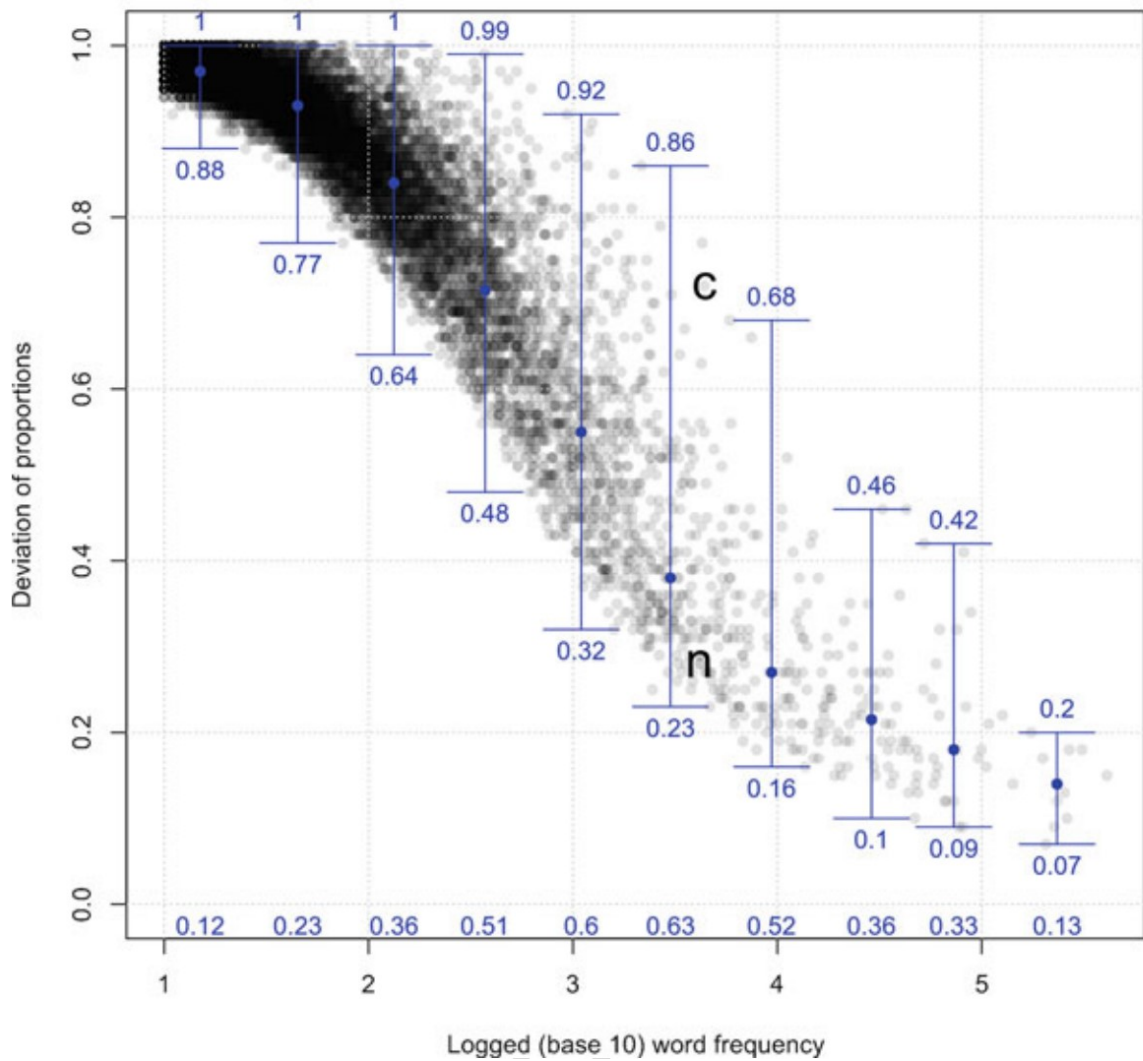


Figure 3.1: The relation between word frequency and dispersion ( $DP$ )  
(from Gries [2021: 112])

existing measures and proposes Deviation of Proportions as a conceptually simple alternative; it is also this measure which most strongly correlates with reaction time data, as mentioned above. In essence, Deviation of Proportions measures how much the frequency of an item within the various parts of a corpus deviates from what one would expect if the item were evenly distributed in the corpus. The procedure for calculating *DP* for a given lexical item is as follows:

1. Determine the sizes of each part of the corpus as a percentage of the overall corpus. These values represent the *expected* percentage of the time that one would expect the item to appear in each corpus part, if it were evenly distributed.
2. Determine the frequencies with which the target item occurs in each part, as a percentage of its overall frequency of occurrence. These values represent the *actual* or *observed* percentage of the time that the item appears in each corpus part.
3. Compute the pairwise absolute differences between the expected and observed percentages, sum them up, and divide the result by two.
4. The result is *DP*, which theoretically ranges from 0 (the item is evenly distributed across the corpus, given the size of the parts) to 1 (the item is unevenly distributed across the corpus, given the size of the parts).

The mathematical formulation of *DP* is shown in (36), where  $n$  is the number of corpus parts,  $v$  is the frequencies of the target item in each corpus part,  $f$  is the overall frequency of the target item in the corpus, and  $s$  is the percent size of each corpus part.

$$(36) \quad DP = 0.5 \times \sum_{i=1}^n \left| \frac{v_i}{f} - s_i \right|$$

A more detailed explanation of this calculation, with examples, is in Gries (2008: §3). Note that while the theoretical range of *DP* is between 0 and 1, it will never actually reach these two limits because a particular proportion of the lexical item was expected to occur in each

corpus part anyway. This issue is only noticeable in corpora with a very small number of parts.

Both the token frequencies and corpus dispersions of each lexical item in the 100-item samples are reported in [Appendix B](#).

### 3.5 Summary

This chapter has presented the methodological tools necessary for answering the research questions put forth in [Chapter 1](#). The methods adopted in this study are novel for several reasons. First, this is the first study to utilize naturalistic discourse data from corpora to examine lexical flexibility at the level of the individual lexical item. Second, this is the first study to *quantify* the lexical flexibility of individual lexical items, in a crosslinguistically applicable way. The calculation of lexical flexibility using Shannon's  $H$  is intended as the main methodological contribution of this thesis. Finally, this study incorporates findings from recent research in corpus linguistics which suggest that corpus dispersion is a better measure of frequency of exposure than just raw token frequency. As such, I report on both token frequency and corpus dispersion and examine their interaction as they relate to lexical flexibility in [Section 4.5](#). With these methodological prerequisites in place, I now turn to answering this study's research questions in [Chapter 4](#).

# Chapter 4

## Results

This chapter reports the results of applying the procedures described in [Chapter 3: Data & Methods](#). I begin by demonstrating for the reader how to interpret the ternary plots used to visually represent the degree of lexical flexibility for individual items ([R1](#)) ([Section 4.2](#)). Next I look at the flexibility of lexical items in English and Nuuchahnulth, both independently and in comparison ([Section 4.3](#)). I then I investigate whether lexical flexibility depends on corpus size ([R2](#)) ([Section 4.4](#)). Next, examine the relationship between the degree of lexical flexibility and frequency / dispersion ([R3](#)) ([Section 4.5](#)). Finally, I discuss the behavior of flexible items with respect to their semantics ([R4](#)) ([Section 4.6](#)).

### 4.1 Introduction

This chapter presents the empirical findings from this study, answering the research questions posed in [Chapter 1](#). I employ a useful visualization for displaying information about lexical flexibility called a *ternary plot* or *triangle plot*; I explain how these ternary plots are to be read in [Section 4.2](#). [Section 4.3](#) focuses on answering [R1](#), “How flexible are lexical items in English and Nuuchahnulth?”, both individually and in comparison. [Section 4.3](#) is dedicated to answering [R2](#), “Is there a correlation between degree of lexical flexibility and size of the corpus?”, and [Section 4.4](#) answers [R3](#), “Is there a correlation between degree of lexical flexibility for a lexical item and frequency (or corpus dispersion)?”. In the final section ([§4.6](#)), I look at the semantic behavior of more and less flexible items (question [R4](#)).

## 4.2 Interpreting the results

In [Section 3.4.1](#) I describe the procedure for quantifying the lexical flexibility of an item in a corpus using a Shannon diversity index ( $H$ ). While the resulting values nicely align with our intuitions about when a lexical item is more or less flexible, some information is lost in the process. Reducing the lexical flexibility of an item to a single number obscures the fact that items can be equally flexible in different ways. Consider the fictional frequency data for two different stems in Table ?? . Stem A displays a great deal of reference-predicate flexibility, but very few instances of use as a modifier. Stem B, in contrast, displays extensive reference-modifier flexibility, but few instances of use as a predicate. However, the overall flexibility ratings of the two stems are the same.

add table

One way to address this reduction in fidelity is to report frequencies and corpus dispersions for each function in addition to the overall flexibility rating for each stem. I provide this information in [Appendix B](#) alongside each item's flexibility rating. However, it is also possible to visualize the relative usage of an item for each discourse function in an intuitive way by using a *ternary plot* (also called a *triangle plot*, *simplex plot*, *Gibbs triangle*, or *de Finetti diagram*). A ternary plot depicts the ratios of three variables as points within an equilateral triangle. Each corner of the triangle corresponds to one of the three possible categories (in this case, reference, predication, or modification). The closer a data point is to a particular corner, the larger the ratio of that category is. To illustrate with an example: Figure ?? is a ternary plot for the functions of the word *difficult* in English, along with the underlying frequency data and resulting flexibility rating.

add ternary plot for 'difficult', along with underlying data

Because the word *difficult* only appears as a modifier in the corpus, it has a flexibility rating of 0. In the ternary plot, this is evident from the fact that the plot point for *difficult* sits in the

modification corner of the triangle.

Compare the plot for *difficult* in Figure ?? to that of *away* in Figure ??.

add ternary plot for 'away', along with underlying data

The stem *anything* also has a flexibility rating of 0 because all of its tokens are used for reference. Even though its flexibility rating is the same as that of *difficult*, it is plotted in a different corner of the ternary plot (reference).

Figure ?? shows a case where a stem (*childhood*) is flexible between reference and modification, but not predication.

add ternary plot for 'childhood', along with underlying data

A perfectly flexible item which has equal use as a referent, predicate, and modifier, would sit exactly in the center of the triangle. The Nuuchahnulth stem *?u-q* 'good' is one such case, shown in Figure ?? . The closer a point is towards the center of the triangle, the more flexible it is.

add ternary plot for Nuuchahnulth 'good', along with underlying data

Finally, remember from [Chapter 3](#) that corpus dispersion is a better measure of frequency of exposure than just raw frequency. Thus in addition to relative frequency data, I also report corpus dispersions for the discourse functions of each lexical item in [Appendix B](#). Note that the corpus dispersions are calculated separately for each discourse function (in addition to the overall corpus dispersion of the lexical item). A particular lexical item might be used for one function evenly throughout the corpus, and thus have a low *DP* for that function, but might only be used for another function in one or two texts, thus giving that function a high *DP*. The ratios of these corpus dispersions for each function can be plotted on a ternary plot just like frequency. Plots based on corpus dispersions are sometimes notably different from plots based on frequencies, as Figure ?? illustrates. In most cases however the plots are identical or near-identical. As such, for the remainder of this study I will use ternary plots based on

corpus dispersion rather than frequency, noting where the two diverge only when relevant.

compare ternary plots for frequency and dispersion where they differ significantly

### 4.3 R1: Degree of lexical flexibility

In this section I examine the degree of lexical flexibility for words in English and Nuuchahnulth from several angles, both independently and in comparison, using the lexical flexibility ratings calculated with the methods in [Section 3.4.1](#). The result of these calculations for the 100-item samples are shown in [Appendix B](#), and a selection of the results appear in Table ?? for English and Table ?? for Nuuchahnulth.

add English-sample table, based on the one in your colloquium talk

add Nuuchahnulth-sample table, based on the one in your colloquium talk

Figure ?? visualizes these flexibility ratings for the 100-item samples from English (left-hand side) and Nuuchahnulth (righthand side). The top portion of each figure is a histogram showing the number of lexical items at different flexibility ratings, providing an approximate representation of the distribution of the flexibility ratings. Beneath the histograms are boxplots showing the median flexibility rating for each language.

add the histograms (with boxplots) for the 100-item samples

Figure ?? shows the same visualizations for the small corpus samples.

add the histograms (with boxplots) for the small corpus samples

Tables ?? and ?? summarize basic descriptive statistics for each of the samples.

tables showing min, max, mean, median for each sample

One immediately obvious observation to be made from these flexibility ratings is that individual lexical items may vary widely in their flexibility, both within and across languages.

While this finding is entirely unsurprising, the results very well could have been otherwise. The way Nuuchahnulth is often described, one might expect all the lexical items in the language to fall within a more limited range of high-flexibility values. This is clearly not the case. Flexibility ratings for Nuuchahnulth range from the theoretical minimum of 0 to a maximum

of .920 or 0.000 depending on the sample. However, of the small corpus Nuuchahnulth sample have a flexibility rating of 0 (), potentially challenging the claim that all Nuuchahnulth stems are flexible.

Likewise, those who claim that English parts of speech are well defined must confront the fact that the range of flexibility values for English is nearly the same as for Nuuchahnulth for both samples: 0 on the lower end and .919 or 0.000 on the upper end. In fact, in both samples there are fewer English stems with a flexibility rating of 0 than there are Nuuchahnulth stems with a flexibility rating of 0. In this respect, then, English could be viewed as more flexible than Nuuchahnulth. Of course, it may be that this difference is due to the large difference in corpus sizes between English and Nuuchahnulth, an issue which is explored in [Section 4.4](#).

Thus the answer to the question, “Are some lexical items more flexible than others?” is unsurprisingly “yes”. To pose a related question, “Can it be shown empirically and quantitatively that some lexical items are more flexible than others, as many linguists have claimed?”. The answer is again, “yes”. If we want to evaluate the claim that some languages are more or less flexible than others, it must be possible to quantify that flexibility at the level of the individual lexical item and compare them in a meaningful way. The data and methods in this thesis show that this is indeed possible, and that we can provide clear empirical answers to these kinds of questions. The flexibility of individual lexical items varies widely both within and between languages.

A slightly different question than whether individual stems vary in their flexibility is whether they exhibit flexibility to any substantive degree in the first place. Or, to invert the question, is lexical flexibility a marginal / rare phenomenon which has merely been given disproportionate attention in the literature? A quick look at the flexibility data above shows



that this is not the case. When lexical items in English and Nuuchahnulth exhibit flexibility, it is typically not to a marginal degree. A stem is more likely to have a flexibility of, say, .2 than something like .002. According to a one-sided, one-sample sign test, the median flexibility in all samples differs highly significantly from zero (see the summary table in Table ??).

summarize the statistical results comparing the median flexibility in a table

This result may seem obvious, but it must be recognized that the result could have been different. Flexibility for English stems could have been so marginal as to not significantly deviate from zero. This would have supported an analysis of lexical flexibility in English as mere occasional language play, something exceptional rather than rampant or productive. The data show otherwise: lexical flexibility is a prevalent feature of both English and Nuuchahnulth, though to different degrees.

Another obvious question to ask of these data is whether English and Nuuchahnulth differ in their overall flexibility. The answer to this is clearly “yes”, as the mean and median flexibility ratings in Tables ?? and ?? show. The mean flexibility ratings are higher for Nuuchahnulth than English in both samples. But to reduce the entire lexicon of a language to a single average obscures important details. While it is certainly interesting that Nuuchahnulth stems are on net more flexible than English stems, the way in which the two languages exhibit flexibility is arguably the more interesting finding from this study.

verify this

How then is lexical flexibility realized in English and Nuuchahnulth? In addition to the histograms in Figures ?? and ??, the ternary plots in Figures ?? and ?? illustrate just how flexibility operates in the two languages. In these figures, each lexical item is represented by a single point on the ternary plot.

add ternary plots comparing English vs. Nuuchahnulth for both samples

Beginning with English, we can see that most lexical items exhibit some flexibility, but to a relatively small degree. After zero-flexibility cases, the next most frequent flexibility rating is in the 0–0.05 range. This is evident from the ternary plots, where lexical items tend to cluster

near (but not precisely on) the corners for their most prototypical functions. Interestingly, the small English corpus appears to show *more* flexibility than the 100-item sample. This could be an effect of the specific words chosen, but it could also be the case that it takes a certain number of tokens for the prototypical function of an item to become evident. This possibility is examined further in [Section 4.4](#).

Nuuchahnulth differs from English in several notable ways. First, a much higher proportion of items display no flexibility whatsoever. However, for those items which do exhibit flexibility, the average flexibility rating is generally higher than that of English stems. In both samples, the biggest cluster of items with non-zero flexibility ratings have ratings around .6. English items with non-zero flexibility, by comparison, generally have ratings closer to .2. Thus for Nuuchahnulth lexical items are either totally inflexible or generally strongly flexible.

This bifurcation of the data is very likely due to the small size of the Nuuchahnulth corpus, as the discussion in [Section 4.4](#) suggests. It may be that Nuuchahnulth words are generally highly flexible, but that more tokens are needed to see this trend. Alternatively, it may be that certain Nuuchahnulth stems are strongly associated with a specific discourse function and thus inflexible, while others are generally flexible. This would suggest a probabilistic division of Nuuchahnulth stems into two classes: those that are productively flexible, and those that are not.

This second possibility would challenge existing analyses of Nuuchahnulth. The existence of a productively flexible class of stems would be counterevidence to the many claims that Nuuchahnulth word classes can in fact be clearly defined using selectional criteria such as ability to take possession or the definite suffix [\(CITE: citations\)](#). Similarly, Nakayama (2001: ???) characterizes word classes in Nuuchahnulth as “strong statistical tendencies”. For many Nuuchahnulth stems, however, there is no clear prototypical use. Many stems are used roughly as equally for predication as they are for reference, making it difficult to assess which use is more basic or unmarked.

As the ternary plots for Nuuchahnulth in Figure ?? and Figure ?? make clear, the distribution of lexical items across functions in Nuuchahnulth differs strongly from that of English. For starters, there is very little clustering around prototypical functions in the corners, in direct contrast to English. Secondly, Nuuchahnulth shows very little flexibility in the modification direction, but rampant flexibility along the reference-predication axis. For the small corpus sample in particular, there is a smooth cline of values between reference and predication. Nuuchahnulth stems sit anywhere on a continuum from prototypical referents to prototypical predicates, but none show prototypical modifier behavior.

These findings nicely reflect the intuitions of many researchers about these two languages. English is mostly rigid, but most words exhibit a marginal degree of flexibility. English words are *primarily* associated with one discourse function, but not exclusively so. Nuuchahnulth, by contrast, shows a very high degree of reference-predicate flexibility. However, Nuuchahnulth stems are not frequently used for modification. This is in line with the analysis of most researchers regarding lexical categories in Nuuchahnulth. Nakayama (2001: 50), for example, says that the categories noun and verb must be recognized for Nuuchahnulth, but that there is not sufficient evidence to justify an adjective category, even as a statistical tendency. He instead treats “adjectivals” as a subclass of verbs; the central location of the points in the Nuuchahnulth plot in Figure ??, however, suggest that Nuuchahnulth modifiers are as nounlike as they are verblike. The low frequency with which stems are used for modification also mirrors the results from Croft’s (1991: 88–89) four-language survey of the textual frequency of different lexical classes. He also finds that “the overall frequency of roots denoting properties and occurrences of modifiers is extremely low compared to the frequencies of object and action roots and of referring expressions and predications” (Croft 1991: 88–89).

**4.4 R2: Lexical flexibility and corpus size**

**4.5 R3: Lexical flexibility and frequency / dispersion**

**4.6 R4: The semantics of lexical flexibility**

## Chapter 5

### Conclusion

This chapter summarizes the methods and main findings of this study, and then considers the implications of those results for theories of lexical categories. I argue that the data provide compelling evidence in favor of functional approaches to lexical categorization, most especially cognitive prototype theory and Croft's theory of lexical categories as typological markedness patterns. I also argue for a reversal of the canonical position on parts of speech: instead of working from the default assumption that all languages have clearly-defined or even loosely-defined parts of speech, we should begin from the understanding that dedicated referring, predicating, or modifying constructions develop diachronically, and that even when they do, they do not do so for the entire lexicon, or in all areas of the grammar equally. Even languages like English whose lexemes pattern strongly with the standard prototypes of noun, verb, and adjectives nonetheless exhibit varying degrees of flexibility for different lexemes. Lexical categories are not a given in grammar. I conclude by discussing some limitations of the present study and avenues for future research, followed by closing remarks.

# Appendices

# Appendix A

## Early grammatical descriptions

The table below contains publication information on some of the first grammatical descriptions of languages other than Latin and Greek.

**Table A.1:** Some first grammatical descriptions of European vs. American languages

Language	Year	Title	Author
Irish	600s	<i>Auraicept na n-Éces</i> 'The scholars' primer'	Longarad
Occitan	1327	<i>Leys d'amors</i> 'Laws of love'	Guilhèm Molinièr
Welsh	1382–1410	<i>Llyfr Coch Hergest</i> 'Red book of Hergest'	unknown
Tuscan	1437–1441	<i>Grammatica della lingua toscana</i> 'Grammar of the Tuscan language'	Leon Battista Alberti
Castilian	1492	<i>Gramática de la lengua castellana</i> 'Grammar of the Castilian language'	Antonio de Nebrija
French	1530	<i>L'Éclaircissement de la langue francoyse</i> 'Explication of the French language'	John Palsgrave
German	1534	<i>Ein Teutsche Grammatica</i> 'A German grammar'	Valentin Ickelsamer
Basque	1545	<i>Linguæ Vasconum Primitiæ</i> 'First fruits of the Basque language'	Bernard Etxepare

**Table A.1:** Some first grammatical descriptions of European vs. American languages

Language	Year	Title	Author
Totonac	1539–1554	<i>Arte de la lengua totonaca</i> ‘Grammar of the Totonac language’	Andrés de Olmos
Nahuatl	1547	<i>Arte para aprender la lengua mexicana</i> ‘Grammar for learning the Mexican language’	Andrés de Olmos
Tarascan	1558	<i>Arte de la lengua tarasca de Michoacán</i> ‘Grammar of the Tarascan language of Michoacán’	Maturino Gilberti
Dutch	1559	<i>Den schat der Duytsscher Talen</i> ‘The treasure of the Dutch language’	John III van de Werve
Quechua	1560	<i>Grammatica o arte de la lengua general de los Indios de los Reynos del Peru</i> ‘Grammar or Art of the General Language of the Indians of the Royalty of Peru’	Domingo de Santo Tomás
Tzeltal Maya	1571	<i>Ars Tzeldaiica</i> ‘Tzeltal Grammar’	Fray Domingo de Hara
Zapotec	1578	<i>Arte en lengua Zapoteca</i> ‘Grammar in the Zapotec language’	Juan de Córdova
English	1586	<i>Pamphlet for Grammar</i>	William Bullokar
Mixtec	1593	<i>Arte de lengua Mixteca</i> ‘Grammar of the Mixtec language’	Antonio de los Reyes
Timucua	1614	<i>Gramatica de la lengua Timuquana de Florida</i> ‘Grammar of the Timucua language of Florida’	Francisco Pareja
Narragansett	1643	<i>A key into the language of America</i>	Roger Williams



# Appendix B

## 100-item samples

[Table B.1](#) and [Table B.2](#) list various statistics for each lexical item in the 100-item samples of English and Nuuchahnulth, respectively. See [Chapter 3](#) for how these samples were selected and their accompanying statistics calculated.

### B.1 English

**Table B.1:** Corpus statistics for the 100-item English sample

Stem	Frequencies		Flexibility (Shannon's $H$ )	Dispersion ( $DP$ )	Frequencies			Dispersions ( $DP$ )		
	Raw	Relative			Reference	Predication	Modification	Reference	Predication	Modification
able	1207	0.371	0.000	0.650	0	0	5	—	—	0.998
anything	2458	0.755	0.000	0.449	2081	0	0	0.488	—	—
area	1544	0.474	0.027	0.652	1526	0	7	0.656	—	0.997
away	1208	0.371	0.000	0.636	5	0	0	0.997	—	—
back	3757	1.154	0.844	0.411	272	54	143	0.886	0.974	0.941
believe	1014	0.312	0.014	0.709	0	953	2	—	0.723	0.999
best	777	0.239	0.537	0.742	201	0	526	0.915	—	0.807
big	2701	0.830	0.046	0.474	21	0	2381	0.991	—	0.489
bill	328	0.101	0.127	0.920	310	10	0	0.923	0.997	—
business	629	0.193	0.517	0.818	460	0	158	0.852	—	0.945

Table B.1: Corpus statistics for the 100-item English sample

Stem	Frequencies		Flexibility (Shannon's $H$ )	Dispersion ( $DP$ )	Frequencies			Dispersions ( $DP$ )		
	Raw	Relative			Reference	Predication	Modification	Reference	Predication	Modification
central	53	0.016	0.091	0.977	1	0	48	0.999	—	0.978
certain	733	0.225	0.018	0.767	2	0	706	0.999	—	0.775
child	2551	0.784	0.326	0.677	2165	0	283	0.682	—	0.954
come	5446	1.673	0.255	0.330	64	4528	263	0.969	0.351	0.881
consider	474	0.146	0.058	0.834	0	336	4	—	0.874	0.998
day	3082	0.947	0.329	0.478	2577	0	343	0.491	—	0.945
decide	789	0.242	0.097	0.752	3	652	10	0.999	0.780	0.996
different	2130	0.654	0.012	0.524	3	0	1705	0.999	—	0.580
difficult	380	0.117	0.000	0.868	0	0	54	—	—	0.978
door	430	0.132	0.093	0.866	420	0	9	0.868	—	0.996
down	3369	1.035	0.277	0.457	1	0	10	0.999	—	0.995
end	1368	0.420	0.721	0.614	604	693	34	0.781	0.766	0.987
enjoy	1565	0.481	0.005	0.677	0	1485	1	—	0.686	0.999
everybody	1228	0.377	0.302	0.635	1044	120	0	0.671	0.939	—
everything	1971	0.606	0.000	0.518	1960	0	0	0.520	—	—
fan	217	0.067	0.067	0.942	211	0	3	0.942	—	0.998
father	447	0.137	0.040	0.867	401	0	3	0.875	—	0.999
feel	2707	0.832	0.135	0.529	73	2382	5	0.971	0.549	0.997
figure	577	0.177	0.378	0.788	50	384	5	0.977	0.840	0.998
first	2130	0.654	0.328	0.498	178	0	1345	0.924	—	0.609
five	2222	0.683	0.302	0.512	141	0	1226	0.938	—	0.631
four	1757	0.540	0.301	0.578	140	0	1223	0.939	—	0.660
friend	1270	0.390	0.012	0.653	1237	1	1	0.658	0.999	1.000
fun	913	0.280	0.594	0.761	267	0	149	0.895	—	0.943
good	6868	2.110	0.175	0.355	196	0	3888	0.925	—	0.410
grow	1227	0.377	0.310	0.701	3	885	96	0.999	0.757	0.950
hard	1583	0.486	0.000	0.587	0	0	380	—	—	0.868
hate	455	0.140	0.026	0.840	0	442	2	—	0.845	0.999
here	4859	1.493	0.055	0.425	451	5	0	0.841	0.997	—
house	2159	0.663	0.182	0.668	1994	3	98	0.682	0.998	0.964
husband	1381	0.424	0.011	0.668	1281	0	2	0.679	—	0.999
idea	826	0.254	0.009	0.736	823	0	1	0.737	—	1.000
important	590	0.181	0.063	0.839	2	0	151	0.999	—	0.942

Table B.1: Corpus statistics for the 100-item English sample

Stem	Frequencies		Flexibility (Shannon's $H$ )	Dispersion ( $DP$ )	Frequencies			Dispersions ( $DP$ )		
	Raw	Relative			Reference	Predication	Modification	Reference	Predication	Modification
job	1283	0.394	0.239	0.693	1185	0	94	0.701	—	0.971
know	44687	13.729	0.030	0.214	7	11496	51	0.998	0.265	0.977
large	509	0.156	0.042	0.845	2	1	428	0.999	0.999	0.868
like	3768	1.158	0.003	0.447	1	3105	0	0.999	0.505	—
little	5657	1.738	0.511	0.362	1345	0	4062	0.610	—	0.409
live	3399	1.044	0.349	0.480	132	2626	148	0.946	0.529	0.938
look	3614	1.110	0.263	0.432	89	2713	103	0.965	0.470	0.945
make	5712	1.755	0.129	0.352	25	4029	91	0.991	0.387	0.960
man	933	0.287	0.101	0.765	752	1	16	0.805	0.999	0.994
manage	150	0.046	0.195	0.943	1	119	5	0.999	0.953	0.997
money	2293	0.704	0.066	0.622	2220	1	29	0.625	1.000	0.988
more	6191	1.902	0.545	0.355	1294	12	3500	0.606	0.994	0.429
move	1267	0.389	0.236	0.678	41	948	18	0.982	0.738	0.994
much	5470	1.680	0.429	0.330	3240	15	640	0.406	0.993	0.765
name	922	0.283	0.485	0.755	755	66	70	0.781	0.971	0.969
need	2711	0.833	0.220	0.501	164	2475	3	0.943	0.517	0.999
never	3024	0.929	0.725	0.441	2	10	2	0.999	0.995	0.999
old	1977	0.607	0.054	0.565	5	3	838	0.998	0.998	0.749
one	13052	4.010	0.571	0.245	8384	1	3944	0.310	1.000	0.368
order	300	0.092	0.691	0.893	92	65	3	0.969	0.973	0.999
other	4845	1.488	0.441	0.338	895	0	3841	0.683	—	0.370
paint	490	0.151	0.919	0.954	131	139	47	0.981	0.976	0.986
pay	2979	0.915	0.322	0.648	89	1789	80	0.971	0.701	0.975
person	1171	0.360	0.013	0.690	1111	1	1	0.698	1.000	1.000
pretty	3808	1.170	0.170	0.440	1	1	51	1.000	1.000	0.978
problem	2429	0.746	0.016	0.548	2422	0	6	0.549	—	0.997
put	3571	1.097	0.109	0.458	16	2288	36	0.994	0.522	0.980
real	3500	1.075	0.010	0.434	0	1	655	—	0.999	0.765
right	9104	2.797	0.680	0.308	312	7	305	0.918	0.997	0.880
run	1662	0.511	0.509	0.608	97	1038	106	0.964	0.702	0.955
say	8784	2.699	0.095	0.303	45	7385	96	0.981	0.324	0.956
see	8267	2.540	0.056	0.343	46	5563	11	0.991	0.395	0.994
seven	695	0.214	0.361	0.767	54	0	344	0.976	—	0.868

Table B.1: Corpus statistics for the 100-item English sample

Stem	Frequencies		Flexibility (Shannon's $H$ )	Dispersion ( $DP$ )	Frequencies			Dispersions ( $DP$ )		
	Raw	Relative			Reference	Predication	Modification	Reference	Predication	Modification
six	1185	0.364	0.288	0.648	81	0	760	0.960	—	0.733
something	5418	1.665	0.000	0.341	5092	0	0	0.347	—	—
sound	1150	0.353	0.268	0.679	54	1048	24	0.974	0.702	0.992
stick	377	0.116	0.670	0.853	33	182	28	0.985	0.922	0.987
take	6186	1.900	0.066	0.341	21	4570	34	0.991	0.374	0.985
talk	3308	1.016	0.247	0.420	58	1974	70	0.975	0.528	0.969
thing	10666	3.277	0.003	0.267	10649	4	0	0.267	0.998	—
think	21082	6.477	0.060	0.262	162	20089	58	0.930	0.267	0.970
three	2560	0.786	0.338	0.477	269	1	1964	0.897	0.999	0.524
time	7523	2.311	0.071	0.309	7310	11	93	0.313	0.996	0.957
transfer	80	0.025	0.662	0.968	7	49	9	0.996	0.978	0.997
try	3814	1.172	0.195	0.407	28	2764	109	0.986	0.454	0.951
two	4232	1.300	0.419	0.383	648	0	3106	0.764	—	0.424
understand	896	0.275	0.053	0.724	4	752	3	0.998	0.756	0.999
want	5053	1.552	0.037	0.374	7	4899	23	0.997	0.379	0.989
watch	2134	0.656	0.226	0.730	36	1329	40	0.986	0.793	0.980
way	3962	1.217	0.004	0.376	3730	1	1	0.388	1.000	0.999
week	1493	0.459	0.013	0.627	1476	0	3	0.629	—	0.999
whole	1753	0.539	0.102	0.546	41	0	1682	0.980	—	0.551
woman	1112	0.342	0.146	0.827	969	0	38	0.837	—	0.993
wonder	669	0.206	0.194	0.793	26	589	4	0.988	0.814	0.998
work	6368	1.956	0.707	0.423	1381	3698	323	0.645	0.475	0.905
worst	152	0.047	0.484	0.937	32	0	111	0.986	—	0.955
year	6773	2.081	0.006	0.355	6728	2	3	0.355	0.999	0.999

## B.2 Nuuchahnulth

Table B.2: Corpus statistics for the 100-item Nuuchahnulth sample

Stem	Gloss	Frequencies		Flexibility (Shannon's $H$ )	Dispersion ( $DP$ )	Frequencies			Dispersions ( $DP$ )		
		Raw	Relative			Reference	Predication	Modification	Reference	Predication	Modification
ʔa:q	fat	7	0.837	0.000	0.830	7	0	0	0.830	—	—
ʔayix	swift	6	0.717	0.000	0.773	0	6	0	—	0.773	—
ʔaʔu:	another	19	2.271	0.835	0.322	11	6	2	0.505	0.614	0.784
ʔaqmis	oil	5	0.598	0.000	0.843	5	0	0	0.843	—	—
ʔatwa	paddling.steadily	4	0.478	0.000	0.710	0	4	0	—	0.710	—
ʔawa	hear	7	0.837	0.000	0.909	0	7	0	—	0.909	—
ʔut	nice	7	0.837	0.545	0.631	2	5	0	0.895	0.736	—
ʔu:csa:mi:h	women	5	0.598	0.000	0.613	5	0	0	0.613	—	—
ʔahku·	right.here	6	0.717	0.631	0.688	3	3	0	0.881	0.807	—
ʔac	go.out.hunting	6	0.717	0.000	0.614	0	6	0	—	0.614	—
ʔac-yu·	go.out.hunting-done	14	1.673	0.000	0.693	0	14	0	—	0.693	—
ʔana	only	12	1.434	0.410	0.462	0	10	2	—	0.482	0.863
ʔana-ʔi·c	only-eat	4	0.478	0.000	0.640	0	4	0	—	0.640	—
ʔath	night	4	0.478	0.000	0.678	0	4	0	—	0.678	—
ʔaya	many	34	4.064	0.652	0.424	2	23	6	0.801	0.496	0.669
ʔi:qh	telling	19	2.271	0.000	0.494	0	19	0	—	0.494	—
ʔu-(w)aλ	it-find	4	0.478	0.000	0.788	0	4	0	—	0.788	—
ʔu-hta·	it-doing.to	7	0.837	0.000	0.776	0	7	0	—	0.776	—
ʔu-ca-hta	it-go.to-apart	11	1.315	0.000	0.659	0	11	0	—	0.659	—
ʔu-kla·	it-having.as.name	21	2.510	0.000	0.542	0	21	0	—	0.542	—
ʔu-na·k	it-having	28	3.347	0.140	0.448	1	27	0	0.887	0.478	—
ʔu-yi·ha	it-because.of	7	0.837	0.000	0.589	0	7	0	—	0.589	—
ʔu:s	some	20	2.391	0.920	0.556	9	8	3	0.647	0.603	0.857
ʔu:s-ck <sup>w</sup> i·	some-remains.of	5	0.598	0.000	0.634	0	5	0	—	0.634	—
ʔuh	being.it	28	3.347	0.000	0.312	0	28	0	—	0.312	—
ʔucq	foggy	4	0.478	0.000	0.957	0	4	0	—	0.957	—
ʔunic	how.much.time.spent	9	1.076	0.625	0.601	4	5	0	0.794	0.693	—
ʔunwi:λ	there.is.a.reason	13	1.554	0.000	0.620	0	13	0	—	0.620	—
ʔaʔuk	lake	4	0.478	0.512	0.755	3	1	0	0.887	0.868	—

Table B.2: Corpus statistics for the 100-item Nuuchahnulth sample

Stem	Gloss	Frequencies		Flexibility	Dispersion	Frequencies			Dispersions ( $DP$ )		
		Raw	Relative	(Shannon's $H$ )	( $DP$ )	Reference	Predication	Modification	Reference	Predication	Modification
ħa:ɬ	there	21	2.510	0.761	0.506	5	14	2	0.803	0.559	0.759
ħa:ħu:p	instructing	8	0.956	0.000	0.827	0	8	0	—	0.827	—
ħa:k <sup>w</sup> a:λ	girl	6	0.717	0.000	0.721	6	0	0	0.721	—	—
ħacwəħs	bowl	5	0.598	0.000	0.978	5	0	0	0.978	—	—
ħamip	knowing	7	0.837	0.000	0.650	0	7	0	—	0.650	—
ħawɪɬ	chief	35	4.184	0.417	0.549	29	6	0	0.594	0.649	—
ħu:	over.there	25	2.988	0.400	0.489	4	21	0	0.778	0.477	—
ħu:ɬ	over.there	15	1.793	0.455	0.385	3	12	0	0.887	0.398	—
ħumi:s	red.cedar	7	0.837	0.373	0.859	6	1	0	0.859	0.891	—
čaʔak	water	4	0.478	0.000	0.759	4	0	0	0.759	—	—
čəħ	adze	4	0.478	0.000	0.891	0	3	0	—	0.891	—
čəpac	canoe	38	4.542	0.113	0.579	36	1	0	0.576	0.891	—
či:q	sing	11	1.315	0.277	0.728	1	10	0	0.990	0.728	—
ču	move	5	0.598	0.000	0.920	0	5	0	—	0.920	—
camaqλ	take.time	4	0.478	0.000	0.891	0	4	0	—	0.891	—
ciq	speak	16	1.913	0.213	0.467	1	15	0	0.957	0.499	—
haʔuk <sup>w</sup>	eat	4	0.478	0.000	0.842	0	4	0	—	0.842	—
hapt	hide	12	1.434	0.261	0.597	1	11	0	0.957	0.597	—
hiɬ	there	52	6.216	0.606	0.393	20	32	0	0.512	0.374	—
hiɬ-ʼa-ʔa	here-on.the.rock	6	0.717	0.579	0.628	2	4	0	0.954	0.651	—
hiʔi:s	there.on.the.ground	10	1.195	0.296	0.514	1	9	0	0.868	0.503	—
hi:hi:q-šəħap	various-doing	4	0.478	0.000	0.912	0	4	0	—	0.912	—
hi:nip	obtain	4	0.478	0.000	0.967	0	4	0	—	0.967	—
hi:tkin	strange	6	0.717	0.790	0.773	1	4	1	0.920	0.853	0.887
hicnup	couple	5	0.598	0.613	0.699	3	2	0	0.863	0.836	—
hin-ʔaɬ	there-aware.of	5	0.598	0.000	0.867	0	5	0	—	0.867	—
hin-in	there-come	9	1.076	0.000	0.577	0	9	0	—	0.577	—
hini:p	take.long	6	0.717	0.000	0.821	0	6	0	—	0.821	—
his	beat	8	0.956	0.000	0.643	0	8	0	—	0.643	—
his-i-k	there-going.along	9	1.076	0.625	0.743	4	5	0	0.755	0.743	—
his-taq	there-come.from	19	2.271	0.188	0.432	1	18	0	0.886	0.470	—
hiš-umɬ	all-in.a.bunch	4	0.478	0.000	0.960	0	4	0	—	0.960	—
hu:ʔak	long.ago	29	3.466	0.137	0.421	0	28	1	—	0.449	0.868

Table B.2: Corpus statistics for the 100-item Nuuchahnulth sample

Stem	Gloss	Frequencies		Flexibility	Dispersion	Frequencies			Dispersions ( $DP$ )		
		Raw	Relative	(Shannon's $H$ )	( $DP$ )	Reference	Predication	Modification	Reference	Predication	Modification
huḥtak	know	32	3.825	0.000	0.356	0	32	0	—	0.356	—
k <sup>w</sup> is	different	10	1.195	0.296	0.627	1	9	0	0.920	0.605	—
k <sup>w</sup> is-tu·p	different-thing	8	0.956	0.000	0.899	0	8	0	—	0.899	—
k <sup>w</sup> ac	hit.the.right.spot	6	0.717	0.000	0.561	0	6	0	—	0.561	—
kəḥ-k <sup>w</sup> a	split-in.pieces	5	0.598	0.000	0.886	0	5	0	—	0.886	—
kamatq	running	23	2.749	0.000	0.792	0	23	0	—	0.792	—
ku:ciḥ	filleting.fish	4	0.478	0.000	0.809	0	4	0	—	0.809	—
ku:kuh <sup>w</sup> 'isa	hair.seal	15	1.793	0.223	0.831	14	1	0	0.826	0.977	—
ṛuksyi	stone	11	1.315	0.000	0.763	11	0	0	0.763	—	—
ma-maḥ-ni·	dwel-move-come	6	0.717	0.000	0.664	6	0	0	0.664	—	—
ma:ʔak	gray.whale	12	1.434	0.000	0.852	12	0	0	0.852	—	—
ma:ma:ti	bird	6	0.717	0.410	0.708	5	1	0	0.708	0.977	—
mamaḥni	white.man	16	1.913	0.213	0.476	15	1	0	0.465	0.886	—
mu:-ci-ḥ	four-days.long	9	1.076	0.602	0.721	3	5	0	0.755	0.853	—
mu:-qʔich	four-year	6	0.717	0.613	0.875	2	3	0	0.988	0.886	—
ṇa:s	day	7	0.837	0.000	0.511	7	0	0	0.511	—	—
ṇup-ci-ḥ	one-days.long	10	1.195	0.613	0.610	2	3	0	0.759	0.744	—
naʔa:	hear	21	2.510	0.000	0.532	0	21	0	—	0.532	—
nani·qsu	grandparent	10	1.195	0.556	0.873	7	3	0	0.873	0.907	—
nas	try.in.vain	6	0.717	0.000	0.793	0	6	0	—	0.793	—
nunu:k	singing	13	1.554	0.562	0.824	4	9	0	0.944	0.790	—
piḥ	observe	9	1.076	0.000	0.734	0	9	0	—	0.734	—
q <sup>w</sup> ayači:k-štaqumḥ	wolf-groups	13	1.554	0.000	0.901	13	0	0	0.901	—	—
q <sup>w</sup> is	do.so	60	7.172	0.476	0.294	13	47	0	0.517	0.372	—
qaḥa:tik	younger.brother	5	0.598	0.000	0.671	4	0	0	0.671	—	—
qaḥ	dead	29	3.466	0.228	0.491	2	27	0	0.834	0.512	—
qi:-sasa	for.a.long.time-precisely	5	0.598	0.000	0.566	0	5	0	—	0.566	—
qu:ʔas	person	81	9.682	0.106	0.341	78	2	0	0.355	0.834	—
quṇa:	amount	8	0.956	0.602	0.481	5	3	0	0.565	0.704	—
si:ḥ-iḥ	you.all-to	4	0.478	0.000	0.943	0	4	0	—	0.943	—
suk <sup>w</sup> iḥ	take	8	0.956	0.000	0.785	0	8	0	—	0.785	—
sut-(c)iḥ	you-doing.to	13	1.554	0.000	0.469	0	13	0	—	0.469	—
ṭa:ṭna	children	9	1.076	0.318	0.653	8	1	0	0.764	0.868	—

Table B.2: Corpus statistics for the 100-item Nuuchahnulth sample

Stem	Gloss	Frequencies		Flexibility (Shannon's $H$ )	Dispersion ( $DP$ )	Frequencies			Dispersions ( $DP$ )		
		Raw	Relative			Reference	Predication	Modification	Reference	Predication	Modification
faña	child	9	1.076	0.000	0.659	9	0	0	0.659	—	—
tupał	sea	4	0.478	0.000	0.750	4	0	0	0.750	—	—
wa·	say	273	32.632	0.120	0.261	6	199	0	0.844	0.306	—
wa·ł-’aqstuł	word-inside	4	0.478	0.000	0.929	0	4	0	—	0.929	—
wik	not	139	16.615	0.039	0.190	1	138	0	0.920	0.188	—



# Appendix C

## Sample annotations

This appendix contains small sample annotations from the English and Nuuchahnulth corpora. Each token that was included in this study is given an annotation indicating its discourse function—REF, PRED, or MOD for reference, predication, or modification, respectively. Lexical items that are excluded from this study for one of the reasons discussed in [Section 3.3](#) are not given an annotation. Recall from [Chapter 3](#) that for English, predicate adjectives and non-finite verbs are *not* included. For English, the annotation is given as a subscript after the word. For Nuuchahnulth, the discourse function is indicated as a 5<sup>th</sup> line in the interlinear gloss.

### C.1 English

Well life<sub>REF</sub> there in the country<sub>REF</sub> is nice and tranquil. I lived<sub>PRED</sub> working<sub>MOD</sub> all of my life<sub>REF</sub> with livestock<sub>REF</sub>. I always had to get<sub>PRED</sub> up early milk<sub>PRED</sub> the cows<sub>REF</sub> and uh run<sub>PRED</sub> run<sub>PRED</sub> them as we say<sub>PRED</sub> because it's a— to the pastures<sub>REF</sub> until times<sub>REF</sub> got<sub>PRED</sub> pretty bad and one<sub>MOD</sub> day<sub>REF</sub> I sent<sub>PRED</sub> my daughter<sub>REF</sub> to to the pasture<sub>REF</sub> to bring in the cows<sub>REF</sub>. We brought<sub>PRED</sub> them back in the afternoon<sub>REF</sub> when I saw<sub>PRED</sub> that behind her there came<sub>PRED</sub> a big<sub>MOD</sub> group<sub>REF</sub> of they looked<sub>PRED</sub> like soldiers<sub>REF</sub> but in street<sub>MOD</sub> clothes<sub>REF</sub>. Then she came<sub>PRED</sub> my daughter<sub>REF</sub> came<sub>PRED</sub> almost green pale and she said<sub>PRED</sub> to me "Mama" she said<sub>PRED</sub> to me "Those are guerillas<sub>REF</sub>!" That was the first<sub>MOD</sub> time<sub>REF</sub> I saw<sub>PRED</sub> them the gue— the guerillas<sub>REF</sub>. (Ide & Suderman 2005: ArguetaBertila-ENG)

## C.2 Nuuchahnulth

qiiʔaλ	qiifañaλ	q <sup>w</sup> iyuck <sup>w</sup> iʔitq	q <sup>w</sup> is	ʔaḥ
qi:-'aλ	qi:-faña-'aλ	q <sup>w</sup> iyu-ck <sup>w</sup> i-ʔi·tq	q <sup>w</sup> is	ʔaḥ
for.a.long.time-FIN	for.a.long.time-slightly-FIN	time-done-REL.3	happen.thus	this
happened.long.ago	quite.a.while.ago	when.it.occurred	happen.thus	this
PRED	PRED	REF	PRED	REF

‘This happened a long time ago.’

siikc̣inλ	siikaa	hitaçinλ	maaqtusiis
si:k-çinλ	si:k-(y)a·	hita-çinλ	ma:qtusi:s
sailing-into.a.bay	sailing-CONT	there.MOM-into.a.bay	NAME
sail.into.a.bay	sailing	entered.into.a.bay	NAME
PRED	PRED	PRED	REF

‘They sailed into the bay of Maaqtusiis.’

yuupick <sup>w</sup> imatak	yuupi	yuksaaʔa
yu:pi-ck <sup>w</sup> i-·matak	yu:pi	yu-ksa·ʔa
breeze-done-probably	breeze	blowing-come.to.land
probably.there.was.a.breeze	breeze	breeze.along.the.shoreline
PRED	REF	PRED

‘There probably was a little wind, blowing towards the land.’

q <sup>w</sup> iyimtii	ñaas	hitaçinλ
q <sup>w</sup> iyu-imt-(y)i:	ña:s	hita-çinλ
when-PAST-INDEF.3	day	there.MOM-in.a.bay
whenever.it.was	day	entered.a.bay
PRED	REF	PRED

‘They came into the bay one day.’

hitaçu	ʔukʔaakʔakna	ýuuq <sup>w</sup> aa
hitaçu	ʔu-kʔa-·ak-ʔa·k-na·	ýu:q <sup>w</sup> a:
NAME	it-called-DUR-POSS-1PL	also
NAME	we.also.call.it	also
	PRED	PRED

‘We also call it (the bay) “hitaçu”.’

waʔyuu	maaqtusiis	wiiḥaaqsusiis
waʔ-yu·	ma:qtusi:s	wi:ḥa:qsusi:s
go.home-done	NAME	NAME
gone.home	NAME	NAME
PRED	REF	

‘They went to Maaqtusiis – [to be exact,] *Wiiḥaaqsusiis*.’

ʔuʔiiačištck <sup>wi</sup>	wiiḥaaqsusiis	ʔayuuk <sup>wi</sup> λ	kuunaa
ʔu-ʔi·ya-āčīšt-ck <sup>wi</sup> ·	wi:ḥa:qsusi:s	ʔayu:-k <sup>wi</sup> (λ)	ku:na:
it-reach-on.the.sea-done	NAME	anchored-MOM	schooner
reached	NAME	anchored	schooner
PRED	REF	PRED	REF

‘The schooner reached *Wiiḥaaqsusiis* and dropped anchor.’

wik	ʔiiḥ	wikck <sup>wii</sup>	ʔiiḥ
wik	ʔi:ḥ	wik-ck <sup>wi</sup> ·	ʔi:ḥ
not	large	not-done	large
not	large	was.not	large
PRED	PRED	PRED	PRED

‘It (the schooner) was not so big.’

ʔaλa	ʔaλista	qac̣c̣istamitquu
ʔaλa	ʔaλa-ista	qac̣c̣a-ista-mit-qu:
two	two-people.on.board	three-people.on.board-PAST-COND.3
two	two.people.on.board	there.could.have.been.three.people.on.board
PRED	REF	PRED

‘There were two crewmen, or there could have been three, on the ship.’

hinaačiλaʔ	yaqitii
hin-a-či(λ)-ʔa·ʔ	yaq-it-(y)i:
there.MOM-go.out.to.meet-PL	who-PAST-INDEF.3
they.go.out.to.meet	whoever.it.was
PRED	REF

‘Some people went out to meet them (the people on the schooner).’

ʔin	ʔutwiick <sup>wi</sup> ʔaaʔ	hinaačiλ	wiʔakʔi
ʔin	ʔutwi:-ck <sup>wi</sup> ·-ʔa·ʔ	hin-a-či(λ)	wiʔak-ʔi·
since	first-done-always	there.MOM-go.out.to.meet	brave-DEF
since	they.were.the.first.one	go.out.to.meet	the.brave.one
	PRED	PRED	REF

wiiʔaksaʔi	ḥaaʔakatʔi	ʔimaqsti
wiʔak-sa-ʔi·	ḥa:ʔak-ʔat-ʔi·	ʔimaqsti
brave-real-DEF	strong-POSS-DEF	mind
the.bravest.one	the.one.with.strong.one	mind
REF	REF	REF

‘The first ones to go out were the bravest ones, the ones with strong minds.’

ʔin	naʔaack <sup>w</sup> aλ	ʔaya	mamaʔni	hisiick <sup>w</sup> iʔitqʔaʔ	hiistiλ
ʔin	naʔa:-ck <sup>w</sup> i--'aλ	ʔaya	mamaʔni	hisi:-ck <sup>w</sup> i--ʔi-tq-ʔa-ʔ	hi:stiλ
??	hear-done-FIN	many	white.man	??-done-REL.3-PL	from
??	understood	many	white.man	the.way.they.spoke	from
	PRED	MOD	REF	REF	PRED

ciqʔak	čiinuukʔath
ciq-ʔak <sup>w</sup>	či:nu:k-'ath
speak-instrument	Chinook-belonging.to
language	Chinook
REF	PRED

‘Many white men could understand Chinook Jargon.’

čiīčiinuk <sup>w</sup> ack <sup>w</sup> aλ	ʔuuš
DUP-či:nu:k-(y)a-ck <sup>w</sup> i--'aλ	ʔu:š
DISTR-speak.Chinook-REP-done-FIN	some
spoke.Chinook.Jargon	some
PRED	REF

‘Some of them spoke Chinook Jargon.’

histathck <sup>w</sup> aλukʔaʔ	ʔah	[Hudson Bay]
hist-'ath-ck <sup>w</sup> i--'aλ-uk-ʔa-ʔ	ʔah	
there-belonging.to-done-FIN-POSS-PL	this	
they.got.theirs.from.there	they	
PRED	REF	

‘They got theirs (= their knowledge of Chinook Jargon) from Hudson Bay Company.’


yaq <sup>w</sup> iyyii	naʔaaʔaλ	Captainmitquu	yaq <sup>w</sup> acʔitq
yaq <sup>w</sup> -wi--(y)i:	naʔa:-'aλ	Captain-mit-qu:	yaq <sup>w</sup> -ac-ʔi-tq
who-first-INDEF.3	hear-FIN	captain-PAST-COND.3	who-belonging.to-REL.3
the.ones.who.were.first	understood	one.who.was.Captain	owner.of
REF	PRED	PRED	REF

šipʔii
šip-ʔi-
ship-DEF
the.ship
REF

‘Among the first ones that [learned to] understand the language might have been the Captain who was taking command of the ship.’

(Louie 2003: Kingfisher)

update page  
number



# References

- Andrade, Manuel J. 1933. Quileute. In Franz Boas (ed.), *Handbook of American Indian languages*, vol. 3, 151–292. Columbia University Press.
- Ansaldi, Umberto, Jan Don & Roland Pfau (eds.). 2010. *Parts of speech: Empirical and theoretical advances* (Benjamins Current Topics 25). John Benjamins. doi:[10.1075/bct.25](https://doi.org/10.1075/bct.25).
- Anward, Jan. 2000. A dynamic model of part-of-speech differentiation. In Petra M. Vogel & Bernard Comrie (eds.), *Approaches to the typology of word classes* (Empirical Approaches to Language Typology 23), 3–46. Mouton de Gruyter. doi:[10.1515/9783110806120](https://doi.org/10.1515/9783110806120).
- Arad, Maya. 2005. *Roots and patterns: Hebrew morpho-syntax* (Studies in Natural Language & Linguistic Theory 63). Springer. doi:[10.1007/1-4020-3244-7](https://doi.org/10.1007/1-4020-3244-7).
- Audring, Jenny & Geert Booij. 2016. Cooperation and coercion. *Linguistics* 54(4). 617–637. doi:[10.1515/ling-2016-0012](https://doi.org/10.1515/ling-2016-0012).
- Austin, J. L. 1962. *How to do things with words*. Clarendon Press.
- Avolio, Meghan L., Jeremy M. Beaulieu, Eugenia Y. Y. Lo & Melinda D. Smith. 2012. Measuring genetic diversity in ecological studies. *Plant Ecology* 213(7). 1105–1115. doi:[10.1007/s11258-012-0069-6](https://doi.org/10.1007/s11258-012-0069-6).
- Bai, Junwei. 2019. *A grammar of Munya*. James Cook University. (Ph.D. thesis). doi:[10.25903/2shv-x307](https://doi.org/10.25903/2shv-x307).
- Baker, Mark C. 2003. *Lexical categories: Verbs, nouns, and adjectives* (Cambridge Studies in Linguistics 102). Cambridge University Press. doi:[10.1017/CBO9780511615047](https://doi.org/10.1017/CBO9780511615047).
- Baker, Mark C. 2015. Nouns, verbs, and verbal nouns: Their structures and their structural cases. In Joanna Błaszczak, Dorota Klimek-Jankowska & Krzysztof Migdalski (eds.), *How categorical are categories?: New approaches to the old questions of noun, verb, and adjective* (Studies in Generative Grammar 122), 13–46. Mouton de Gruyter. doi:[10.1515/9781614514510-003](https://doi.org/10.1515/9781614514510-003).
- Baker, Paul. 2018. Corpus methods in linguistics. In Lia Litosseliti (ed.), *Research methods in linguistics*, 2nd edn., 167–192. Bloomsbury.
- Beck, David. 2002. *The typology of parts of speech systems: The markedness of adjectives* (Outstanding Dissertations in Linguistics). Routledge. doi:[10.4324/9780203475201](https://doi.org/10.4324/9780203475201).
- Beck, David. 2013. Unidirectional flexibility and the noun-verb distinction in Lushootseed. In Jan Rijkhoff & Eva van Lier (eds.), *Flexible word classes: Typological studies of under-*

*specified parts of speech*, 185–220. Oxford University Press. doi:[10.1093/acprof:oso/9780199668441.003.0007](https://doi.org/10.1093/acprof:oso/9780199668441.003.0007).

Berez-Kroeker, Andrea L., Helene N. Andreassen, Lauren Gawne, Gary Holton, Susan Smythe Kung, Peter Pulsifer, Lauren B. Collister & The Data Citation & Attribution in Linguistics Group. 2018. *The Austin principles of data citation in linguistics*. <https://site.uit.no/linguisticsdatacitation/austinprinciples>.

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>.

Bird, Steven, Ewan Klein & Edward Loper. 2009. *Natural language processing with Python: Analyzing text with the Natural Language Toolkit*. O'Reilly.

Bisang, Walter. 2008. Precategoriality and syntax-based parts of speech: The case of Late Archaic Chinese. *Studies in Language* 32(3). 568–589. doi:[10.1075/sl.32.3.05bis](https://doi.org/10.1075/sl.32.3.05bis).

Bisang, Walter. 2013. Word class systems between flexibility and rigidity: An integrative approach. In Jan Rijkhoff & Eva van Lier (eds.), *Flexible word classes: Typological studies of underspecified parts of speech*, 275–303. Oxford University Press. doi:[10.1093/acprof:oso/9780199668441.003.0010](https://doi.org/10.1093/acprof:oso/9780199668441.003.0010).

Błaszczak, Joanna, Dorota Klimek-Jankowska & Krzysztof Migdalski (eds.). 2015. *How categorical are categories?: New approaches to the old questions of noun, verb, and adjective* (Studies in Generative Grammar 122). Mouton de Gruyter. doi:[10.1515/9781614514510](https://doi.org/10.1515/9781614514510).

Bloomfield, Leonard. 1933. *Language*. George Allen & Unwin.

Boas, Franz. 1911. Introduction. In Franz Boas (ed.), *Handbook of American Indian languages*, vol. 1 (Bureau of American Ethnology Bulletin 40), 1–84. Smithsonian Institution.

Bolinger, Dwight & Donald A. Sears. 1981. *Aspects of language*. 3rd edn. Harcourt Brace.

Booij, Geert & Jenny Audring. 2018. Category change in construction morphology. In Kristel Van Goethem, Muriel Norde, Evie Coussé & Gudrun Vanderbauwhede (eds.), *Category change from a constructional perspective* (Constructional Approaches to Language 20), 209–228. John Benjamins. doi:[10.1075/cal.20.08boo](https://doi.org/10.1075/cal.20.08boo).

Bowerman, Melissa & Soonja Choi. 2001. Shaping meanings for language: Universal and language-specific in the acquisition of spatial semantic categories. In Melissa Bowerman & Stephen C. Levinson (eds.), *Language acquisition and conceptual development* (Language, Culture & Cognition 3). Cambridge University Press. doi:[10.1017/CBO9780511620669.018](https://doi.org/10.1017/CBO9780511620669.018).

Boydston, Amber E., Shaun Bevan & Herschel F. Thomas. 2014. The importance of attention diversity and how to measure it. *Policy Studies Journal* 42(2). 173–196. doi:[10.1111/psj.12055](https://doi.org/10.1111/psj.12055).

Braithwaite, Ben. 2015. Nuu-chah-nulth nouns and verbs revisited: Root allomorphy and the structure of nominal predicates. In Joanna Błaszczak, Dorota Klimek-Jankowska & Krzysz-

- tof Migdalski (eds.), *How categorical are categories?: New approaches to the old questions of noun, verb, and adjective* (Studies in Generative Grammar 122), 47–74. Mouton de Gruyter. doi:[10.1515/9781614514510-003](https://doi.org/10.1515/9781614514510-003).
- Broschart, Jürgen. 1997. Why Tongan does it differently: Categorical distinctions in a language without nouns and verbs. *Linguistic Typology* 1(2). 123–165. doi:[10.1515/lity.1997.1.2.123](https://doi.org/10.1515/lity.1997.1.2.123).
- Bybee, Joan L. 1985. *Morphology: A study of the relation between meaning and form* (Typological Studies in Language 9). John Benjamins. doi:[10.1075/tsl.9](https://doi.org/10.1075/tsl.9).
- Bybee, Joan L. 2007. *Frequency of use and the organization of language*. Oxford University Press. doi:[10.1093/acprof:oso/9780195301571.001.0001](https://doi.org/10.1093/acprof:oso/9780195301571.001.0001).
- Bybee, Joan L. 2010. *Language, usage and cognition*. Cambridge University Press. doi:[10.1017/CBO9780511750526](https://doi.org/10.1017/CBO9780511750526).
- Cannon, Garland. 1985. Functional shift in English. *Linguistics* 23(3). 411–432. doi:[10.1515/ling.1985.23.3.411](https://doi.org/10.1515/ling.1985.23.3.411).
- Carter, Richard. 2006. Polycategoriality and predictability: Problems and prospects. In Ximena Lois & Valentina Vapnarsky (eds.), *Lexical categories and root classes in Amerindian languages*, 343–390. Peter Lang.
- Chafe, Wallace L. 1976. Givenness, contrastiveness, definiteness, subjects, topics and points of view. In Charles Li (ed.), *Subject and Topic*, 25–56. Academic Press.
- Chafe, Wallace L. 1987. Cognitive constraints on information flow. In Russell S. Tomlin (ed.), *Coherence and grounding in discourse* (Typological Studies in Language 11), 21–52. John Benjamins. doi:[10.1075/tsl.11.03cha](https://doi.org/10.1075/tsl.11.03cha).
- Chafe, Wallace L. 2012. Are adjectives universal?: The case of Northern Iroquoian. *Linguistic Typology* 16(1). 1–39. doi:[10.1515/lingty-2012-0001](https://doi.org/10.1515/lingty-2012-0001).
- Chomsky, Noam. 1957. *Syntactic structures*. 1st edn. Mouton & Co.
- Chung, Sandra. 2012. Are lexical categories universal?: The view from Chamorro. *Theoretical Linguistics* 38(1-2). 1–56. doi:[10.1515/tl-2012-0001](https://doi.org/10.1515/tl-2012-0001).
- Comrie, Bernard. 1976. *Aspect: An introduction to the study of verbal aspect and related problems* (Cambridge Textbooks in Linguistics). Cambridge University Press.
- Corbett, Greville G. 1978. Universals in the syntax of cardinal numerals. *Lingua* 46. 355–368. doi:[10.1016/0024-3841\(78\)90054-2](https://doi.org/10.1016/0024-3841(78)90054-2).
- Corbett, Greville G. 2000. *Number* (Cambridge Textbooks in Linguistics). Cambridge University Press. doi:[10.1017/CBO9781139164344](https://doi.org/10.1017/CBO9781139164344).
- Corbett, Greville G. 2005. The canonical approach in typology. In Zygmunt Frajzyngier, Adam Hodges & David S. Rood (eds.), *Linguistic diversity and language theories* (Studies in Language Companion Series 72), 25–50. John Benjamins. doi:[10.1075/slcs.72.03cor](https://doi.org/10.1075/slcs.72.03cor).
- Corbett, Greville G. 2010. Canonical derivational morphology. *Word Structure* 3(2). 141–155. doi:[10.3366/word.2010.0002](https://doi.org/10.3366/word.2010.0002).



- Creissels, Denis. 2017. The flexibility of the noun/verb distinction in the lexicon of Mandinka. In Valentina Vapnarsky & Edy Veneziano (eds.), *Lexical polycategoriality: Cross-linguistic, cross-theoretical and language acquisition approaches* (Studies in Language Companion Series 182). John Benjamins. doi:[10.1075/slcs.182.02cre](https://doi.org/10.1075/slcs.182.02cre).
- Croft, William. 1984. Semantic and pragmatic correlates to syntactic categories. *Chicago Linguistic Society* 20. 53–71.
- Croft, William. 1991. *Syntactic categories and grammatical relations: The cognitive organization of information*. University of Chicago Press.
- Croft, William. 1995. Modern syntactic typology. In Masayoshi Shibatani & Theodora Bynon (eds.), *Approaches to language typology*, 85–144. 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. Mouton de Gruyter. doi:[10.1515/9783110806120.65](https://doi.org/10.1515/9783110806120.65).
- Croft, William. 2001a. Grammar: Functional approaches. In *International encyclopedia of the social & behavioral sciences*, 6323–6330. Elsevier. doi:[10.1016/b0-08-043076-7/02946-6](https://doi.org/10.1016/b0-08-043076-7/02946-6).
- Croft, William. 2001b. *Radical Construction Grammar: Syntactic theory in typological perspective*. Oxford University Press. doi:[10.1093/acprof:oso/9780198299554.001.0001](https://doi.org/10.1093/acprof:oso/9780198299554.001.0001).
- Croft, William. 2003. *Typology and universals*. 2nd edn. (Cambridge Textbooks in Linguistics). Cambridge University Press. doi:[10.1017/CBO9780511840579](https://doi.org/10.1017/CBO9780511840579).
- Croft, William. 2005. Word classes, parts of speech, and syntactic argumentation. *Linguistic Typology* 9(3). 431–441. doi:[10.1515/lity.2005.9.3.391](https://doi.org/10.1515/lity.2005.9.3.391).
- Croft, William. 2007. Beyond Aristotle and gradience: A reply to Aarts. *Studies in Language* 31(2). 409–430. doi:[10.1075/sl.31.2.05cro](https://doi.org/10.1075/sl.31.2.05cro).
- 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](https://doi.org/10.1515/lingty-2014-0021).
- Croft, William. 2016. Comparative concepts and language-specific categories: Theory and practice. *Linguistic Typology* 20(2). 377–393. doi:[10.1515/lingty-2016-0012](https://doi.org/10.1515/lingty-2016-0012).
- Croft, William. Forthcoming. Word classes in Radical Construction Grammar. In Eva van Lier (ed.), *The Oxford handbook of word classes*. Oxford University Press.
- Croft, William & D. Alan Cruse. 2004. *Cognitive linguistics* (Cambridge Textbooks in Linguistics). Cambridge University Press.
- Croft, William & Eva van Lier. 2012. Language universals without universal categories. *Theoretical Linguistics* 38(1-2). 57–72. doi:[10.1515/t1-2012-0002](https://doi.org/10.1515/t1-2012-0002).

- Crystal, David. 1967. English. *Lingua* 17(3-4). 24–56. doi:[10.1016/0024-3841\(66\)90003-9](https://doi.org/10.1016/0024-3841(66)90003-9).
- Crystal, David. 2008. *A dictionary of linguistics and phonetics*. 6th edn. (The Language Library). Blackwell. doi:[10.1002/9781444302776](https://doi.org/10.1002/9781444302776).
- Culicover, Peter W. 1999. *Syntactic nuts: Hard cases, syntactic theory, and language acquisition* (Foundations of Syntax 1). Oxford University Press.
- Cuyckens, Hubert, Liesbet Heyvaert & Stefan Hartmann (eds.). 2019. *Language Sciences 73: Categorical shift: From description to theory and back again*.
- Davis, Henry, Carrie Gillon & Lisa Matthewson. 2014. How to investigate linguistic diversity: Lessons from the Pacific Northwest. *Language* 90(4). e180–e226. doi:[10.1353/lan.2014.0076](https://doi.org/10.1353/lan.2014.0076).
- De Smet, Hendrik. 2012. The course of actualization. *Language* 88(3). 601–633. doi:[10.1353/lan.2012.0056](https://doi.org/10.1353/lan.2012.0056).
- Denison, David. 2001. Gradience and linguistic change. In Laurel J. Brinton (ed.), *Historical linguistics 1999* (Current Issues in Linguistic Theory 215), 119–144. John Benjamins. doi:[10.1075/cilt.215.10den](https://doi.org/10.1075/cilt.215.10den).
- Denison, David. 2018. Why would anyone *take long*?: Word classes and Construction Grammar in the history of *long*. In Kristel Van Goethem, Muriel Norde, Evie Coussé & Gudrun Vanderbauwhede (eds.), *Category change from a constructional perspective* (Constructional Approaches to Language 20), 119–148. John Benjamins. doi:[10.1075/cal.20.05den](https://doi.org/10.1075/cal.20.05den).
- Diessel, Holger. 2019. *The grammar network: How linguistic structure is shaped by language use*. Cambridge University Press. doi:[10.1017/9781108671040](https://doi.org/10.1017/9781108671040).
- Dixon, R. M. W. 1977. Where have all the adjectives gone? *Studies in Language* 1(1). 19–80. doi:[10.1075/sl.1.1.04dix](https://doi.org/10.1075/sl.1.1.04dix).
- Dixon, R. M. W. 2004. Adjective classes in typological perspective. In R. M. W. Dixon & Alexandra Y. Aikhenvald (eds.), *Adjective classes: A cross-linguistic typology* (Explorations in Linguistic Typology 1), 1–49. Oxford University Press.
- Dixon, R. M. W. 2010. *Basic Linguistic Theory, Vol. 1: Methodology*. Oxford University Press.
- Dixon, R. M. W. 2011. *I am a linguist*. Brill. doi:[10.5749/j.ctv6zd95b.13](https://doi.org/10.5749/j.ctv6zd95b.13).
- Don, Jan & Eva van Lier. 2013. Derivation and categorization in flexible and differentiated languages. In Jan Rijkhoff & Eva van Lier (eds.), *Flexible word classes: Typological studies of underspecified parts of speech*. Oxford University Press.
- Dryer, Matthew S. 2013. Coding of nominal plurality. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology. <http://wals.info/chapter/33>.
- Du Bois, John W. 1987. The discourse basis of ergativity. *Language* 63(4). 805–855. doi:[10.2307/415719](https://doi.org/10.2307/415719).

- Duncan, Philip T., Valerie (Lamxayat) Switzler & Henry B. Zenk. Forthcoming. Chinookan. In Carmen Jany, Marianne Mithun & Keren Rice (eds.), *The languages and linguistics of indigenous North America: A comprehensive guide* (The World of Linguistics 13). Mouton de Gruyter.
- van Eijk, Jan P. & Thom Hess. 1986. Noun and verb in Salish. *Lingua* 69(4). 319–331. doi:[10.1016/0024-3841\(86\)90061-6](https://doi.org/10.1016/0024-3841(86)90061-6).
- Evans, Nicholas. 2007. Insubordination and its uses. In Irina Nikolaeva (ed.), *Finiteness: Theoretical and empirical advances*, 366–431. Oxford University Press.
- Evans, Nicholas & Toshiki Osada. 2005. Mundari: The myth of a language without word classes. *Linguistic Typology* 9(3). 351–390. doi:[10.1515/lity.2005.9.3.351](https://doi.org/10.1515/lity.2005.9.3.351).
- Evans, Nicholas & Honoré Watanabe. 2016. The dynamics of insubordination: An overview. In Nicholas Evans & Honoré Watanabe (eds.), *Insubordination* (Typological Studies in Language 115), 1–38. John Benjamins. doi:[10.1075/tsl.115.01eva](https://doi.org/10.1075/tsl.115.01eva).
- Evans, Vyvyan & Melanie Green. 2006. *Cognitive linguistics: An introduction*. Edinburgh University Press.
- Farrell, Patrick. 2001. Functional shift as category underspecification. *English Language & Linguistics* 5(1). 109–130. doi:[10.1017/S1360674301000156](https://doi.org/10.1017/S1360674301000156).
- Fillmore, Charles J., Paul Kay & Mary Catherin O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language* 64(3). 501–538. doi:[10.2307/414531](https://doi.org/10.2307/414531).
- Finch, Geoffrey. 2003. *How to study linguistics: A guide to understanding language*. 2nd edn. (Palgrave Study Guides). Palgrave Macmillan. doi:[10.1007/978-0-230-80213-1](https://doi.org/10.1007/978-0-230-80213-1).
- Floyd, Simeon. 2011. Re-discovering the Quechua adjective. *Linguistic Typology* 15(2011). 25–63. doi:[10.1515/LITY.2011.003](https://doi.org/10.1515/LITY.2011.003).
- Frachtenberg, Leo J. 1922. Coos. In Franz Boas (ed.), *Handbook of American Indian languages*, vol. 2 (Bureau of American Ethnology Bulletins 40), 297–430. Smithsonian Institution.
- Franchetto, Bruna & Mara Santos. 2017. The ontology of roots and the emergence of nouns and verbs in Kuikuro: Adult speech and children's acquisition. In Valentina Vapnarsky & Edy Veneziano (eds.), *Lexical polycategoriality: Cross-linguistic, cross-theoretical and language acquisition approaches* (Studies in Language Companion Series 182), 275–306. John Benjamins. doi:[10.1075/slcs.182.10fra](https://doi.org/10.1075/slcs.182.10fra).
- Gallatin, Albert. 1836. *A synopsis of the Indian tribes within the United States* (Transactions of the American Antiquarian Society 2).
- Genetti, Carol & Kristine Hildebrandt. 2004. The two adjective classes in Manange. In R. M. W. Dixon & Alexandra Y. Aikhenvald (eds.), *Adjective classes: A cross-linguistic typology* (Explorations in Linguistic Typology 1), 74–76. Oxford University Press.
- Gil, David. 1994. The structure of Riau Indonesian. *Nordic Journal of Linguistics* 17(2). 179–200. doi:[10.1017/S0332586500003000](https://doi.org/10.1017/S0332586500003000).

- Gil, David. 1995. Parts of speech in Tagalog. *Southeast Asian Linguistics Society* 3. 67–90.
- Gil, David. 2001. Escaping Eurocentrism: Fieldwork as a process of unlearning. In Paul Newman & Martha Ratliff (eds.), *Linguistic fieldwork*. Cambridge University Press.
- Gil, David. 2005. Isolating-monocategorical-associational language. In Henri Cohen & Claire Lefebvre (eds.), *Handbook of categorization in cognitive science*, 348–380. Elsevier. doi:[10.1016/B978-008044612-7/50070-6](https://doi.org/10.1016/B978-008044612-7/50070-6).
- Gil, David. 2006. Early human language was isolating-monocategorical-associational. In Angelo Cangelosi, Andrew D. M. Smith & Kenny Smith (eds.), *The evolution of language*, vol. 6, 91–98. World Scientific. doi:[10.1142/9789812774262\\_0012](https://doi.org/10.1142/9789812774262_0012).
- Gil, David. 2012. Where does predication come from? *Canadian Journal of Linguistics* 57(2). 32–36. doi:[10.1353/cjl.2012.0030](https://doi.org/10.1353/cjl.2012.0030).
- Goldberg, Adele E. 1995. *Constructions: A Construction Grammar approach to argument structure* (Cognitive Theory of Language & Culture). University of Chicago Press.
- Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language*. Oxford University Press. doi:[10.1093/acprof:oso/9780199268511.001.0001](https://doi.org/10.1093/acprof:oso/9780199268511.001.0001).
- Greenberg, Joseph H. 1966. *Language universals, with special reference to feature hierarchies*. Mouton de Gruyter. doi:[10.1515/9783110899771](https://doi.org/10.1515/9783110899771).
- Gries, Stefan Th. 2006. Corpus-based methods and cognitive semantics: The many senses of *to run*. In Stefan Th. Gries & Anatol Stefanowitsch (eds.), *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis* (Trends in Linguistics: Studies & Monographs 172), 57–100. Mouton de Gruyter. doi:[10.1515/9783110197709.57](https://doi.org/10.1515/9783110197709.57).
- Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403–437. doi:[10.1075/ijcl.13.4.02gri](https://doi.org/10.1075/ijcl.13.4.02gri).
- Gries, Stefan Th. 2010. Dispersions and adjusted frequencies in corpora: Further explorations. In Stefan Th. Gries, Stefanie Wulff & Mark Davies (eds.), *Corpus-linguistic applications: Current studies, new directions* (Language & Computers: Studies in Practical Linguistics 71), 197–212. Rodopi. doi:[10.1163/9789042028012](https://doi.org/10.1163/9789042028012).
- Gries, Stefan Th. 2013. *Statistics for linguistics with R*. 2nd edn. Mouton de Gruyter. doi:[10.1515/9783110216042](https://doi.org/10.1515/9783110216042).
- Gries, Stefan Th. 2021. Analyzing dispersion. In Magali Paquot & Stefan Th. Gries (eds.), *A practical handbook of corpus linguistics*, 99–118. Springer. doi:[10.1007/978-3-030-46216-1\\_5](https://doi.org/10.1007/978-3-030-46216-1_5).
- Gries, Stefan Th. Forthcoming. On, or against?, (just) frequency. In Hans C. Boas (ed.), *Applications of cognitive linguistics*. Mouton de Gruyter.
- Gries, Stefan Th. & Dagmar Divjak. 2009. Behavioral profiles: A corpus-based approach to cognitive semantic analysis. In Vyvyan Evans & Stéphanie Pourcel (eds.), *New directions in cognitive linguistics* (Human Cognitive Processing 24). John Benjamins.

- Gross, Maurice. 1979. On the failure of generative grammar. *Language* 55(4). 859–885. doi:[10.2307/412748](https://doi.org/10.2307/412748).
- Haag, Marcia. 2017. What determines constraints on the relationships between roots and lexical categories?: Evidence from Choctaw and Cherokee. In Valentina Vapnarsky & Edy Veneziano (eds.), *Lexical polycategoriality: Cross-linguistic, cross-theoretical and language acquisition approaches* (Studies in Language Companion Series 182). John Benjamins. doi:[10.1075/slcs.182.07haa](https://doi.org/10.1075/slcs.182.07haa).
- Halle, Morris & Alec Marantz. 1994. Some key features of Distributed Morphology. *MIT Working Papers in Linguistics* 21. 275–288.
- Hammarström, Harald, Robert Forkel & Martin Haspelmath. 2019. *Glottolog 4.0*. Max Planck Institute for the Science of Human History. <https://glottolog.org>.
- Hardjadibrata, R. R. 1985. *Sundanese: A syntactical analysis* (Pacific Linguistics Series D 65). Australian National University.
- Harris, Zellig S. 1951. *Methods in structural linguistics*. University of Chicago Press.
- Haspelmath, Martin. 2001. *Word classes and parts of speech*. doi:[10.1016/B0-08-043076-7/02959-4](https://doi.org/10.1016/B0-08-043076-7/02959-4).
- Haspelmath, Martin. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In Michael Tomasello (ed.), *The new psychology of language*, vol. 2, 211–242. Lawrence Erlbaum.
- Haspelmath, Martin. 2005. Defining vs. diagnosing linguistic categories: A case study of clitic phenomena. In Joanna Błaszczak (ed.), *How categorical are categories?: New approaches to the old questions of noun, verb, and adjective* (Studies in Generative Grammar 122). Mouton de Gruyter. doi:[10.1515/9781614514510-009](https://doi.org/10.1515/9781614514510-009).
- 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](https://doi.org/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](https://doi.org/10.1353/lan.2010.0021).
- Haspelmath, Martin. 2010b. Framework-free grammatical theory. In Bernd Heine & Heiko Narrog (eds.), *The Oxford handbook of linguistic analysis* (Oxford Handbooks in Linguistics), 341–366. Oxford University Press. doi:[10.1093/oxfordhb/9780199544004.013.0014](https://doi.org/10.1093/oxfordhb/9780199544004.013.0014).
- Haspelmath, Martin. 2010c. The interplay between comparative concepts and descriptive categories (Reply to Newmeyer). *Language* 86(3). 696–699. doi:[10.1353/lan.2010.0021](https://doi.org/10.1353/lan.2010.0021).
- Haspelmath, Martin. 2012. Escaping ethnocentrism in the study of word-class universals. *Theoretical Linguistics* 38(1-2). 91–102. doi:[10.1515/t1-2012-0004](https://doi.org/10.1515/t1-2012-0004).
- Haspelmath, Martin. 2017. Some principles for language names. *Language Documentation & Conservation* 11. 81–93. doi:[10.1255/24725](https://doi.org/10.1255/24725).



- Haspelmath, Martin. 2019. How comparative concepts and descriptive linguistic categories are different. In Daniël Van Olmen, Tanja Mortelmans & Frank Brisard (eds.), *Aspects of linguistic variation* (Trends in Linguistics: Studies & Monographs 324). Mouton de Gruyter. doi:[10.1515/9783110607963-004](https://doi.org/10.1515/9783110607963-004).
- Hébert, Yvonne M. 1983. Noun and Verb in a Salishan language. *Kansas Working Papers in Linguistics* 8(2). 31–82. doi:[10.17161/KWPL.1808.478](https://doi.org/10.17161/KWPL.1808.478).
- Hengeveld, Kees. 1992. *Non-verbal predication: Theory, typology, diachrony* (Functional Grammar Series 15). Mouton de Gruyter. doi:[10.1515/9783110883282](https://doi.org/10.1515/9783110883282).
- Hengeveld, Kees. 2007. Parts-of-speech systems and morphological types. *ACLCL Working Papers* 2(1). 34–48. doi:[11245/1.270937](https://doi.org/10.11245/1.270937).
- Hengeveld, Kees. 2013. Parts-of-speech systems as a basic typological determinant. In Jan Rijkhoff & Eva van Lier (eds.), *Flexible word classes: Typological studies of underspecified parts of speech*. Oxford University Press. doi:[10.1093/acprof:oso/9780199668441.003.0002](https://doi.org/10.1093/acprof:oso/9780199668441.003.0002).
- Hengeveld, Kees & Eva van Lier. 2012. An implicational map of parts of speech. *Linguistic Discovery* 8(1). 129–156. doi:[10.1349/ps1.1537-0852.a.348](https://doi.org/10.1349/ps1.1537-0852.a.348).
- Hengeveld, Kees & Jan Rijkhoff. 2005. Mundari as a flexible language. *Linguistic Typology* 9(3). 406–431. doi:[10.1515/lity.2005.9.3.391](https://doi.org/10.1515/lity.2005.9.3.391).
- Hengeveld, Kees, Jan Rijkhoff & Anna Siewierska. 2004. Parts-of-speech systems and word order. *Journal of Linguistics* 40(3). 527–570. doi:[10.1017/S0022226704002762](https://doi.org/10.1017/S0022226704002762).
- Hengeveld, Kees & Marieke Valstar. 2010. Parts-of-speech systems and lexical subclasses. *Linguistics in Amsterdam* 3(1). 1–24. doi:[11245/1.325466](https://doi.org/10.11245/1.325466).
- 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](https://doi.org/10.1515/lity-2013-0013).
- Hieber, Daniel W. 2021a. *Scriptio*. doi:[10.5281/zenodo.2595548](https://doi.org/10.5281/zenodo.2595548).
- Hieber, Daniel W. 2021b. *The Data Format for Digital Linguistics (DaFoDiL)*. doi:[10.5281/zenodo.1438589](https://doi.org/10.5281/zenodo.1438589).
- Hieber, Daniel W. Forthcoming. Word classes. In Carmen Jany, Marianne Mithun & Keren Rice (eds.), *The languages and linguistics of indigenous North America: A comprehensive guide* (The World of Linguistics 13). Mouton de Gruyter.
- Hopper, Paul J. & Sandra A. Thompson. 1980. Transitivity in grammar and discourse. *Language* 56(2). 251. doi:[10.2307/413757](https://doi.org/10.2307/413757).
- Hopper, Paul J. & Sandra A. Thompson. 1984. The discourse basis for lexical categories in Universal Grammar. *Language* 60(4). 703–752. doi:[10.2307/413797](https://doi.org/10.2307/413797).
- Hopper, Paul J. & Elizabeth Closs Traugott. 2003. *Grammaticalization*. 2nd edn. (Cambridge Textbooks in Linguistics). Cambridge University Press. doi:[10.1017/CBO9781139165525](https://doi.org/10.1017/CBO9781139165525).
- Hudson, Richard. 1994. About 37% of word-tokens are nouns. *Language* 70(2). 331–339. doi:[10.2307/415831](https://doi.org/10.2307/415831).

- Ide, Nancy & Keith Suderman. 2005. *Open American National Corpus*. <http://www.anc.org/>.
- Jacobsen, Jr., William H. 1979. Noun and verb in Nootkan. In Barbara S. Efrat (ed.), *The Victoria conference on northwestern languages* (Heritage Record 4), 83–153. British Columbia Provincial Museum.
- Jacobson, Steven A. 2012. *Yup'ik Eskimo dictionary*. 2nd edn. Alaska Native Language Center.
- Jelinek, Eloise & Richard A. Demers. 1994. Predicates and prenominal arguments in Straits Salish. 70(4). 68–111. doi:[10.4324/9780203068236](https://doi.org/10.4324/9780203068236).
- Kastovsky, Dieter. 1996. Verbal derivation in English: A historical survey, or, much ado about nothing. In Derek Britton (ed.), *English historical linguistics 1994* (Current Issues in Linguistic Theory 135), 93–118. John Benjamins. doi:[10.1075/cilt.135.09kas](https://doi.org/10.1075/cilt.135.09kas).
- Kibrik, Andrej. 2011. *Reference in discourse* (Oxford Studies in Typology & Linguistic Theory). Oxford University Press. doi:[10.1093/acprof:oso/9780199215805.001.0001](https://doi.org/10.1093/acprof:oso/9780199215805.001.0001).
- Kihm, Alain. 2017. Categorical flexibility as an emergent phenomenon: A comparison of Arabic, Wolof, and French. In Valentina Vapnarsky & Edy Veneziano (eds.), *Lexical polycategoriality: Cross-linguistic, cross-theoretical and language acquisition approaches* (Studies in Language Companion Series 182), 79–99. John Benjamins. doi:[10.1075/slcs.182.04lio](https://doi.org/10.1075/slcs.182.04lio).
- Kinkade, M. Dale. 1983. Salish evidence against the universality of ‘noun’ and ‘verb’. *Lingua* 60(1). 25–39. doi:[10.1016/0024-3841\(83\)90045-1](https://doi.org/10.1016/0024-3841(83)90045-1).
- Kuipers, Aert H. 1968. The categories verb-noun and transitive-intransitive in English and Squamish. *Lingua* 21. 610–626. doi:[10.1016/0024-3841\(68\)90080-6](https://doi.org/10.1016/0024-3841(68)90080-6).
- Lakoff, George. 1987. *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago Press.
- Lakoff, George & Mark Johnson. 1980. *Metaphors we live by*. Chicago: University of Chicago Press.
- Langacker, Ronald W. 1987. Nouns and verbs. *Language* 63(1). 53–94. doi:[10.2307/415384](https://doi.org/10.2307/415384).
- Langacker, Ronald W. 1988. A usage-based model. In B. Rudzka-Ostyn (ed.), *Topics in cognitive linguistics*, 127–161. John Benjamins. doi:[10.1075/cilt.50.06lan](https://doi.org/10.1075/cilt.50.06lan).
- LaPolla, Randy J. 2016. On categorization: Stick to the facts of the languages. *Linguistic Typology* 20(2). 365–375. doi:[10.1515/lingty-2016-0011](https://doi.org/10.1515/lingty-2016-0011).
- Launey, Michel. 1994. *Une grammaire omniprédicative: Essai sur la morphosyntaxe du nahuatl classique* (Sciences du Langage). CNRS.
- Launey, Michel. 2004. The features of omnipredicativity in Classical Nahuatl. *STUF* 57(1). 49–69. doi:[10.1524/stuf.2004.57.1.49](https://doi.org/10.1524/stuf.2004.57.1.49).
- Lewandowska-Tomaszczyk, Barbara. 2007. Polysemy, prototypes, and radial categories. In Dirk Geeraerts & Hubert Cuyckens (eds.), *The Oxford handbook of cognitive linguistics*, 139–169. Oxford University Press. doi:[10.1093/oxfordhb/9780199738632.013.0006](https://doi.org/10.1093/oxfordhb/9780199738632.013.0006).

- Lichtenberk, Frantisek. 1991. Semantic change and heterosemy in grammaticalization. *Language* 67(3). 475–509. doi:[10.1353/lan.1991.0009](https://doi.org/10.1353/lan.1991.0009).
- van Lier, Eva. 2006. Parts-of-speech systems and dependent clauses: A typological study. *Folia Linguistica* 40(3-4). 239–304. doi:[10.1515/flin.40.3-4.239](https://doi.org/10.1515/flin.40.3-4.239).
- van Lier, Eva. 2012. Reconstructing multifunctionality. *Theoretical Linguistics* 38(1-2). 119–135. doi:[10.1515/tl-2012-0006](https://doi.org/10.1515/tl-2012-0006).
- van Lier, Eva. 2016. Lexical flexibility in Oceanic languages. *Linguistic Typology* 20(2). 197–232. doi:[10.1515/lingty-2016-0005](https://doi.org/10.1515/lingty-2016-0005).
- van Lier, Eva. 2017. The typology of property words in Oceanic languages. *Linguistics* 55(6). 1237–1280. doi:[10.1515/ling-2017-0027](https://doi.org/10.1515/ling-2017-0027).
- van Lier, Eva & Jan Rijkhoff. 2013. Flexible word classes in linguistic typology and grammatical theory. In Jan Rijkhoff & Eva van Lier (eds.), *Flexible word classes: Typological studies of underspecified parts of speech*, 1–30. Oxford University Press. doi:[10.1093/acprof:oso/9780199668441.001.0001](https://doi.org/10.1093/acprof:oso/9780199668441.001.0001).
- Little, Caroline. 2003. *Caroline Little's Nuu-chah-nulth (Ahousaht) texts with grammatical analysis*. Toshihide Nakayama (ed.) (Endangered Languages of the Pacific Rim A2-27). Nakanishi Press.
- Lois, Ximena & Valentina Vapnarsky. 2003. *Polyvalence of root classes in Yukatekan Mayan languages* (LINCOM Studies in Native American Linguistics). LINCOM Europa.
- Lois, Ximena, Valentina Vapnarsky, Cédric Becquey & Aurore Monod Becquelin. 2017. Polycategoriality and hybridity across Mayan languages: Action nouns and ergative splits. In Valentina Vapnarsky & Edy Veneziano (eds.), *Lexical polycategoriality: Cross-linguistic, cross-theoretical and language acquisition approaches* (Studies in Language Companion Series 182), 101–154. John Benjamins. doi:[10.1075/slcs.182.05loi](https://doi.org/10.1075/slcs.182.05loi).
- Louie, George. 2003. *George Louie's Nuu-chah-nulth (Ahousaht) texts with grammatical analysis*. Toshihide Nakayama (ed.) (Endangered Languages of the Pacific Rim A2-028). Nakanishi Press.
- Luuk, Erkki. 2010. Nouns, verbs and flexibles: Implications for typologies of word classes. *Language Sciences* 32(3). 349–365. doi:[10.1016/j.langsci.2009.02.001](https://doi.org/10.1016/j.langsci.2009.02.001).
- Malouf, Robert P. 1999. *Mixed categories in the hierarchical lexicon* (Studies in Constraint-Based Lexicalism). CSLI Publications.
- Marchand, Haus. 1969. *The categories and types of present-day English word-formation: A synchronic-diachronic approach*. 2nd edn. Verlag C. H.
- Matthews, Peter. 2001. *A short history of structural linguistics*. Cambridge University Press. doi:[10.1017/CBO9780511612596](https://doi.org/10.1017/CBO9780511612596).
- Matthews, Peter H. 2014. *The concise Oxford dictionary of linguistics*. 3rd edn. Oxford University Press. doi:[10.1093/acref/9780199675128.001.0001](https://doi.org/10.1093/acref/9780199675128.001.0001).



- Mattina, Nancy J. 1996. *Aspect and category in Okanagan word formation*. Simon Fraser University. (Ph.D. thesis).
- Mattiola, Simone. 2020. Pluractionality: A cross-linguistic perspective. *Language & Linguistics Compass* 14(3). 1–35. doi:[10.1111/lnc3.12366](https://doi.org/10.1111/lnc3.12366).
- McDonald, Edward. 2013. The creation of ‘parts of speech’ for Chinese: ‘Translingual practice’ across Graeco-Roman and Sinitic traditions. *History & Philosophy of the Language Sciences*. <https://hiphilangsci.net/2013/06/12/the-creation-of-parts-of-speech-for-chinese-translingual-practice-across-graeco-roman-and-sinitic-traditions/>.
- McGinnis-Archibald, Martha. 2016. Distributed Morphology. In Andrew Hippisley & Gregory Stump (eds.), *The Cambridge handbook of morphology* (Cambridge Handbooks in Language & Linguistics). Cambridge: Cambridge University Press. doi:[10.1017/9781139814720.015](https://doi.org/10.1017/9781139814720.015).
- McGregor, William B. 2013. Lexical categories in Gooniyandi, Kimberley, Western Australia. In Jan Rijkhoff & Eva van Lier (eds.), *Flexible word classes: Typological studies of underspecified parts of speech*, 221–246. Oxford University Press. doi:[10.1093/acprof:oso/9780199668441.003.0008](https://doi.org/10.1093/acprof:oso/9780199668441.003.0008).
- Mithun, Marianne. 1988. Lexical categories and the evolution of number marking. In Michael Hammond & Michael Noonan (eds.), *Theoretical morphology: Approaches in modern linguistics*, 211–234. Academic Press.
- Mithun, Marianne. 2000. Noun and verb in Iroquoian: Multicategorisation from multiple criteria. In Petra M. Vogel & Bernard Comrie (eds.), *Approaches to the typology of word classes*, 397–420. Mouton de Gruyter. doi:[10.1515/9783110806120.397](https://doi.org/10.1515/9783110806120.397).
- Mithun, Marianne. 2008. The extension of dependency beyond the sentence. *Language* 84(1). 69–119. doi:[10.1353/lan.2008.0054](https://doi.org/10.1353/lan.2008.0054).
- Mithun, Marianne. 2017. Polycategoriality and zero derivation: Insights from Central Alaskan Yup’ik Eskimo. In Valentina Vapnarsky & Edy Veneziano (eds.), *Lexical polycategoriality: Cross-linguistic, cross-theoretical and language acquisition approaches* (Studies in Language Companion Series 182), 155–176. John Benjamins. doi:[10.1075/slcs.182.06mit](https://doi.org/10.1075/slcs.182.06mit).
- Mithun, Marianne. 2019. Categorical shift: Foundations, extensions, and consequences. *Language Sciences* 73. 10–31. doi:[10.1016/j.langsci.2018.08.014](https://doi.org/10.1016/j.langsci.2018.08.014).
- Montler, Timothy. 2003. Auxiliaries and other categories in Straits Salishan. *International Journal of American Linguistics* 69(2). 103–134. doi:[10.1086/379680](https://doi.org/10.1086/379680).
- Morgan, Lawrence. 1991. *A description of the Kutenai language*. University of California, Berkeley. (Ph.D. thesis). <https://escholarship.org/uc/item/0f76g7f2>.
- Mosel, Ulrike & Even Hovdhaugen. 1992. *Samoan reference grammar*. Scandinavian University Press.
- Nakayama, Toshihide. 2001. *Nuuchahnulth (Nootka) morphosyntax* (University of California Publications in Linguistics 134). University of California Press.

- 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](https://doi.org/10.1353/lan.2010.0000).
- Nikolaeva, Irina & Andrew Spencer. 2020. *Mixed categories: The morphosyntax of noun modification* (Cambridge Studies in Linguistics). Cambridge University Press. doi:[10.1017/9781108233903](https://doi.org/10.1017/9781108233903).
- Palmer, Bill. 2017. Categorical flexibility as an artefact of the analysis: Pronouns, articles, and the DP in Hoava and Standard Fijian. *Studies in Language* 41(2). 408–444. doi:[10.1075/s1.41.2.05pal](https://doi.org/10.1075/s1.41.2.05pal).
- Panther, Klaus-Uwe & Linda L. Thornburg. 2007. Metonymy. In Dirk Geeraerts & Hubert Cuyckens (eds.), *The Oxford handbook of cognitive linguistics*, 236–263. Oxford University Press.
- Payne, Thomas E. 1997. *Describing morphosyntax: A guide for field linguists*. Cambridge University Press. doi:[10.1017/CBO9780511805066](https://doi.org/10.1017/CBO9780511805066).
- Pfeiler, Barbara. 2017. The acquisition of action nouns in Yucatec Maya. In Valentina Vapnarsky & Edy Veneziano (eds.), *Lexical polycategoriality: Cross-linguistic, cross-theoretical and language acquisition approaches* (Studies in Language Companion Series 182), 443–466. John Benjamins. doi:[10.1075/slcs.182.15pfe](https://doi.org/10.1075/slcs.182.15pfe).
- Plag, Ingo. 2003. *Word-formation in English* (Cambridge Textbooks in Linguistics). Cambridge University Press. doi:[10.1017/9781316771402](https://doi.org/10.1017/9781316771402).
- Plank, Frans (ed.). 2016. *Linguistic Typology* 20(2): *Of categories: Language-particular – comparative – universal*.
- Polinsky, Maria & Lilla Magyar. 2020. Headedness and the lexicon: The case of verb-to-noun ratios. *Languages* 5(1). 1–25. doi:[10.3390/languages5010009](https://doi.org/10.3390/languages5010009).
- Pustejovsky, James. 1991. The generative lexicon. *Computational Linguistics* 17(4). 409–441. doi:[10.5555/176321.176324](https://doi.org/10.5555/176321.176324).
- Pustet, Regina. 2003. *Copulas: Universals in the categorization of the lexicon* (Oxford Studies in Typology & Linguistic Theory). Oxford University Press. doi:[10.1093/acprof:oso/9780199258505.001.0001](https://doi.org/10.1093/acprof:oso/9780199258505.001.0001).
- Rauh, Gisa. 2010. *Syntactic categories: Their identification and description in linguistic theories* (Oxford Surveys in Syntax & Morphology 7). Oxford University Press.
- Rijkhoff, Jan. 2000. When can a language have adjectives?: An implicational universal. In Petra M. Vogel & Bernard Comrie (eds.), *Approaches to the typology of word classes* (Empirical Approaches to Language Typology 23), 217–258. Mouton de Gruyter. doi:[10.1515/9783110806120.217](https://doi.org/10.1515/9783110806120.217).
- Rijkhoff, Jan. 2002. Verbs and nouns from a cross-linguistic perspective. *Italian Journal of Linguistics* 14(1). 115–147.
- Rijkhoff, Jan. 2003. When can a language have nouns and verbs? *Acta Linguistica Hafniensia* 35(1). 7–38. doi:[10.1080/03740463.2003.10416072](https://doi.org/10.1080/03740463.2003.10416072).

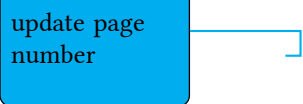
- Rijkhoff, Jan. 2007. Word classes. *Language & Linguistics Compass* 1(6). 709–726. doi:[10 . 1111/j.1749-818x.2007.00030.x](https://doi.org/10.1111/j.1749-818x.2007.00030.x)Word.
- Rijkhoff, Jan & Eva van Lier (eds.). 2013. *Flexible word classes: Typological studies of underspecified parts of speech*. Oxford University Press. doi:[10 . 1093/acprof:oso/9780199668441.001.0001](https://doi.org/10.1093/acprof:oso/9780199668441.001.0001).
- Robert, Stéphane (ed.). 2003. *Perspectives synchroniques sur la grammaticalisation* (Afrique et Langage 5). Peeters.
- Robins, R. H. 1968. Basic sentence structures in Sundanese. *Lingua* 21(C). 351–358. doi:[10 . 1016/0024-3841\(68\)90061-2](https://doi.org/10.1016/0024-3841(68)90061-2).
- Rogers, Phillip. 2016. *Illustrating the prototype structures of parts of speech: A multidimensional scaling analysis*. University of New Mexico. (M.A. thesis).
- Rosch, Eleanor. 1973. Natural categories. *Cognitive Psychology* 4. 328–350. doi:[10 . 1016 / 0010-0285\(73\)90017-0](https://doi.org/10.1016/0010-0285(73)90017-0).
- Rosch, Eleanor. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General* 104(3). 192–233. doi:[10 . 1037/0096-3445.104.3.192](https://doi.org/10.1037/0096-3445.104.3.192).
- Rosch, Eleanor. 1978. Principles of categorization. In Eleanor Rosch & B. B. Lloyd (eds.), *Cognition and categorization*. Lawrence Erlbaum.
- Rosch, Eleanor H. 1973. On the internal structure of perceptual and semantic categories. In *Cognitive development and the acquisition of language*, 111–144. Academic Press. doi:[10 . 1016/b978-0-12-505850-6.50010-4](https://doi.org/10.1016/b978-0-12-505850-6.50010-4).
- Rosch, Eleanor H., Carolyn B. Mervis, Wayne D. Gray, David M. Johnson & Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive Psychology* 8. 382–439. doi:[10 . 1016/B978-0-12-505850-6.50010-4](https://doi.org/10.1016/B978-0-12-505850-6.50010-4).
- Rosch, Eleanor & Carolyn B. Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* 7(4). 573–605. doi:[10 . 1016/0010-0285\(75\)90024-9](https://doi.org/10.1016/0010-0285(75)90024-9).
- Ross, John Robert. 1972. The category squish: Endstation Hauptwort. *Chicago Linguistic Society* 8. 316–328.
- Sadock, Jerrold M. 1999. The nominalist theory of Eskimo: A case study in scientific self-deception. *International Journal of American Linguistics* 65(4). 383–406. doi:[10 . 1086 / 466400](https://doi.org/10.1086/466400).
- Sapir, Edward. 1921. *Language: An introduction to the study of speech*. Harcourt Brace.
- Sasse, Hans Jürgen. 1988. Der irokesische Sprachtyp. *Zeitschrift für Sprachwissenschaft* 7. 173–213. doi:[10 . 1515/ZFSW.1988.7.2.173](https://doi.org/10.1515/ZFSW.1988.7.2.173).
- Schachter, Paul & Timothy Shopen. 2007. Parts-of-speech systems. In Timothy Shopen (ed.), *Language typology and syntactic description, Vol. 1: Clause structure*, 2nd edn., 1–60. Cambridge University Press. doi:[10 . 1017/CBO9780511619427.001](https://doi.org/10.1017/CBO9780511619427.001).

- Searle, John R. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge University Press.
- Shannon, Claude E. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27(3). 379–423. doi:[10.1002/j.1538-7305.1948.tb01338.x](https://doi.org/10.1002/j.1538-7305.1948.tb01338.x).
- Shannon, Claude E. 1951. Prediction and entropy of printed English. *The Bell System Technical Journal* 30(1). 50–64. doi:[10.1002/j.1538-7305.1951.tb01366.x](https://doi.org/10.1002/j.1538-7305.1951.tb01366.x).
- Siddiqi, Daniel. 2018. Distributed Morphology. In Jenny Audring & Francesca Masini (eds.), *The Oxford handbook of morphological theory* (Oxford Handbooks in Linguistics), 143–165. Oxford University Press. doi:[10.1093/oxfordhb/9780199668984.013.15](https://doi.org/10.1093/oxfordhb/9780199668984.013.15).
- Simon, Walter. 1937. Has the Chinese language parts of speech? *Transactions of the Philological Society* 36(1). 99–119. doi:[10.1111/j.1467-968X.1937.tb00675.x](https://doi.org/10.1111/j.1467-968X.1937.tb00675.x).
- Simone, Raffaele & Francesca Masini (eds.). 2014. *Word classes: Nature, typology and representations* (Current Issues in Linguistic Theory 332). John Benjamins. doi:[10.1075/cilt.332](https://doi.org/10.1075/cilt.332).
- Song, Jae Jung. 2001. *Linguistic typology: Morphology and syntax* (Longman Linguistics Library). Routledge. doi:[10.4324/9781315840628](https://doi.org/10.4324/9781315840628).
- Song, Jae Jung. 2018. *Linguistic typology* (Oxford Textbooks in Linguistics). Oxford University Press.
- Stassen, Leon. 1997. *Intransitive predication* (Oxford Studies in Typology & Linguistic Theory). Clarendon 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 University Press. doi:[10.1093/oxfordhb/9780199281251.013.0006](https://doi.org/10.1093/oxfordhb/9780199281251.013.0006).
- Stevick, Robert D. 1968. *English and its history: The evolution of a language*. Allyn & Bacon.
- Suárez, Jorge A. 1983. *The Mesoamerican Indian languages* (Cambridge Language Surveys). Cambridge University Press. doi:[10.1017/CBO9780511554445](https://doi.org/10.1017/CBO9780511554445).
- Sun, Linlin. 2020. *Flexibility in the parts-of-speech system of Classical Chinese* (Trends in Linguistics: Studies & Monographs 334). Mouton de Gruyter. doi:[10.1515/9783110660791](https://doi.org/10.1515/9783110660791).
- Swadesh, Morris. 1939a. *Chitimacha-English dictionary (Copy 1)* (American Council of Learned Societies Committee on Native American Languages Mss.497.3 B63c G6.5). American Philological Society Library.
- Swadesh, Morris. 1939b. Nootka internal syntax. *International Journal of American Linguistics* 9(2). 77–102. doi:[10.1086/463820](https://doi.org/10.1086/463820).
- Taylor, John R. 2003. *Linguistic categorization*. 3rd edn. (Oxford Textbooks in Linguistics). Oxford University Press.
- Thalbitzer, William. 1922. Eskimo. In Franz Boas (ed.), *Handbook of American Indian languages*, vol. 2 (Bureau of American Ethnology Bulletins 40), 971–1069. Smithsonian Institution.

- Thompson, Sandra A. 1989. A discourse approach to the cross-linguistic category ‘Adjective’. In Roberta Corrigan, Fred R. Eckman & Michael Noonan (eds.), *Linguistic categorization* (Current Issues in Linguistic Theory 61), 245–266. John Benjamins. doi:[10.1075/cilt.61.16tho](https://doi.org/10.1075/cilt.61.16tho).
- Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.
- Traugott, Elizabeth Closs & Graeme Trousdale (eds.). 2010. *Gradience, gradualness and grammaticalization* (Typological Studies in Language 90). John Benjamins. doi:[10.1075/tsl.90](https://doi.org/10.1075/tsl.90).
- Tuggy, David. 1993. Ambiguity, polysemy, and vagueness. *Cognitive Linguistics* 4(3). 273–290. doi:[10.1515/cogl.1993.4.3.273](https://doi.org/10.1515/cogl.1993.4.3.273).
- Vapnarsky, Valentina & Edy Veneziano (eds.). 2017a. *Lexical polycategoriality: Cross-linguistic, cross-theoretical and language acquisition approaches* (Studies in Language Companion Series 182). John Benjamins. doi:[10.1075/slcs.182](https://doi.org/10.1075/slcs.182).
- Vapnarsky, Valentina & Edy Veneziano. 2017b. Lexical polycategoriality – Cross-linguistic, cross-theoretical and language acquisition approaches: An introduction. In Valentina Vapnarsky & Edy Veneziano (eds.), *Lexical polycategoriality: Cross-linguistic, cross-theoretical and language acquisition approaches* (Studies in Language Companion Series 182), 1–33. John Benjamins. doi:[10.1075/slcs.182.01val](https://doi.org/10.1075/slcs.182.01val).
- Velupillai, Viveka. 2012. *An introduction to linguistic typology*. John Benjamins. doi:[10.1075/z.176](https://doi.org/10.1075/z.176).
- Vogel, Petra M. 2000. Grammaticalisation and part-of-speech systems. In Petra M. Vogel & Bernard Comrie (eds.), *Approaches to the typology of word classes* (Empirical Approaches to Language Typology 23), 259–284. Mouton de Gruyter. doi:[10.1515/9783110806120](https://doi.org/10.1515/9783110806120).
- Vogel, Petra M. & Bernard Comrie (eds.). 2000. *Approaches to the typology of word classes* (Empirical Approaches to Language Typology 23). Mouton de Gruyter. doi:[10.1515/9783110806120](https://doi.org/10.1515/9783110806120).
- Vonen, Arnfinn M. 1994. Multifunctionality and morphology in Tokelau and English. *Nordic Journal of Linguistics* 17(2). 155–178. doi:[10.1017/S0332586500002997](https://doi.org/10.1017/S0332586500002997).
- de Vries, Lourens. 1989. *Studies in Wambon and Kombai*. University of Amsterdam. (Ph.D. thesis).
- Wetzer, Harrie. 1992. “Nouny” and “verby” adjectivals: A typology of predicative adjectival constructions. In Michel Kefer & Johan van der Auwera (eds.), *Meaning and grammar: Cross-linguistic perspectives* (Empirical Approaches to Language Typology 10), 223–263. Mouton de Gruyter. doi:[10.1515/9783110851656.223](https://doi.org/10.1515/9783110851656.223).
- Wetzer, Harrie. 1996. *The typology of adjectival predication* (Empirical Approaches to Language Typology 17). Mouton de Gruyter. doi:[10.1515/9783110813586](https://doi.org/10.1515/9783110813586).
- Whaley, Lindsay J. 1997. *Introduction to typology: The unity and diversity of language*. SAGE Publications. doi:[10.4135/9781452233437](https://doi.org/10.4135/9781452233437).

Young, Robert W. 1989. Lexical elaboration in Navajo. In Mary Ritchie Key & Henry M. Hoenigswald (eds.), *General and Amerindian ethnolinguistics: In remembrance of Stanley Newman* (Contributions to the Sociology of Language 55), 303–320. De Gruyter. doi:[10.1515/9783110862799-027](https://doi.org/10.1515/9783110862799-027).

update page  
number



# Language Index

- Arabic, [44](#), [45](#)
- Basque, [115](#)
- Castilian, *see* Spanish
- Cayuga, [6](#), [53](#), [60](#), [62](#), [65](#), [83](#)
- Central Alaskan Yup'ik, [50](#), [52](#), [54](#), [55](#), [57](#),  
[60](#), [61](#), [81](#), *see* Yup'ik
- Chamorro, [5](#), [41](#)
- Cherokee, [43](#)
- Chimakuan, [19](#)
- Chinookan, [19](#)
- Chitimacha, [53](#), [65](#), [66](#)
- Classical Greek, [27](#)
- Classical Nahuatl, [50](#), *see* Nahuatl
- Coosan, [19](#)
- Dutch, [38](#), [39](#), [116](#)
- English, [xx](#), [2](#), [4](#), [9](#), [11–13](#), [18](#), [20](#), [31–33](#),  
[43](#), [50](#), [51](#), [57](#), [61](#), [62](#), [64–66](#), [69–73](#),  
[82](#), [84](#), [87–91](#), [116–120](#)
- French, [32](#), [35](#), [115](#)
- German, [115](#)
- Gooniyandi, [43](#)
- Greek, *see* Classical Greek
- Indo-European, [6](#), [19](#), [28](#), [29](#)
- Indonesian, [20](#), [45](#)
- Irish, [28](#), [115](#)
- Iroquoian, [6](#)
- Kuikuro, [43](#)
- Kutenai, [19](#)
- Latin, [27](#), [28](#)
- Mandarin, [6](#), [11](#), [12](#), [27](#)
- Mandinka, [3](#), [52](#), [54–57](#), [61](#)
- Middle English, [43](#)
- Mixtec, [116](#)
- Mundari, [3](#), [6](#), [10](#), [11](#), [20](#), [58](#), [59](#)
- Munya, [78](#)
- Nahuatl, [5](#), [20](#), [28](#), [116](#)
- Narragansett, [116](#)
- Navajo, [53](#)
- Nguna, *see* North Efate
- Nootka, *see* Nuuchahnulth
- North Efate, [12](#)
- Nuuchahnulth, [3](#), [18–20](#), [22](#), [50](#), [79](#), [84](#), [87](#),  
[88](#), [90–95](#), [121–124](#)
- Occitan, [115](#)
- Old English, [43](#)
- Proto-Indo-European, *see* Indo-European
- Purépecha, *see* Tarascan
- Quechua, [3](#), [5](#), [6](#), [116](#)
- Quiché Maya, [12](#)
- Riau Indonesian, [6](#), *see* Indonesian
- Romance, [28](#), [50](#)
- Russian, [10](#), [32](#), [34](#)
- Sahaptian, [19](#), *see* Ichishkin
- Salishan, [19](#)
- Soddo, [12](#)
- Spanish, [28](#), [115](#)
- Standard Arabic, [44](#), *see* Arabic
- Sundanese, [6](#)
- Tagalog, [6](#), [45](#)
- Tarascan, [116](#)



Timucua, [116](#)  
Tongan, [4](#), [6](#), [61](#)  
Totonac, [116](#)  
Tsimshianic, [19](#)  
Tuscan, [115](#)  
Tzeltal Maya, [116](#)

Ute, [12](#)

Wakashan, [83](#)  
Wambon, [38](#), [39](#)  
Welsh, [115](#)  
Wolof, [61](#)  
  
Yucatec Maya, [43](#)  
Yup'ik, [4](#), [10](#), [11](#), [20](#)  
  
Zapotec, [116](#)

# To Do

■ add Dedication . . . . .	iv
■ add Acknowledgments . . . . .	v
■ add Abstract . . . . .	ix
■ finalize the findings on flexibility vs. corpus size, especially the reliability size . . . .	21
■ Adjust this paragraph based on what you wind up doing. You might for example take the average flexibility ratings, but show the error bars on the graph. . . . .	98
■ add table . . . . .	104
■ add ternary plot for ‘difficult’, along with underlying data . . . . .	104
■ add ternary plot for ‘away’, along with underlying data . . . . .	105
■ add ternary plot for ‘childhood’, along with underlying data . . . . .	105
■ add ternary plot for Nuuchahnulth ‘good’, along with underlying data . . . . .	105
■ compare ternary plots for frequency and dispersion where they differ significantly .	106
■ add English-sample table, based on the one in your colloquium talk . . . . .	106
■ add Nuuchahnulth-sample table, based on the one in your colloquium talk . . . . .	106
■ add the histograms (with boxplots) for the 100-item samples . . . . .	106
■ add the histograms (with boxplots) for the small corpus samples . . . . .	106
■ tables showing min, max, mean, median for each sample . . . . .	106
■ update these values . . . . .	107
■ XX (XX%) . . . . .	107
■ XX or XX% for the 100-item sample . . . . .	107
■ update these values . . . . .	107
■ verify that this is true . . . . .	107
■ summarize the statistical results comparing the median flexibility in a table . . . . .	108
■ verify this . . . . .	108
■ add ternary plots comparing English vs. Nuuchahnulth for both samples . . . . .	108
■ citation . . . . .	109
■ get exact phrasing and page for this . . . . .	109
■ update page number . . . . .	108
■ update page number . . . . .	124