

QUERY-BASED ANNOTATION AND THE SUMERIAN VERBAL PREFIXES

by

Eric J. M. Smith

A thesis submitted in conformity with the requirements  
for the degree of Doctor of Philosophy  
Graduate Department of Linguistics  
University of Toronto

Copyright © 2010 by Eric J. M. Smith

# Abstract

Query-Based Annotation and the Sumerian Verbal Prefixes

Eric J. M. Smith

Doctor of Philosophy

Graduate Department of Linguistics

University of Toronto

2010

The study of Sumerian has traditionally been carried out in isolation from mainstream linguistics, thus limiting our ability to understand the language and to situate it in a cross-linguistic context. This dissertation shows how the tools of corpus linguistics and modern syntactic theory can be gainfully applied to Sumerian.

Existing corpora of Sumerian texts are largely lacking in morphological annotation, with query facilities consisting only of basic string searches. Two existing corpora (one completely unannotated and one tagged for part-of-speech) are given morphological annotation using a process of query-based annotation. A query language (based on CQL and XPath) is used to query this corpus, and as queries are made, the results are tagged so that the resultant query objects can be used as the basis for subsequent queries. In this fashion a morphologically-annotated corpus is built up without having to rely on the services of a skilled annotator.

This annotated corpus is then used to provide evidence for two important problems in Sumerian morphosyntax: the dimensional prefixes and the conjugation prefixes. The dimensional prefixes, which have previously been considered to represent concord between the verb and the associated nominal phrases, are shown instead to be a system of applicative heads which serve to introduce the verb's arguments. The conjugation prefixes, whose purpose has been the subject of a century of debate, are shown to be the manifestation of inner aspect features which express the speaker's perspective on the structure of the event.

By using a corpus to provide the underlying data and by considering Sumerian morphosyntax in light of cross-linguistic evidence and modern syntactic theory, previously misanalysed aspects of Sumerian are shown to have analogues in other languages. The dimensional prefixes and conjugation prefixes are not oddities specific to Sumerian, but represent variations on morphological systems found elsewhere.

# Dedication

In memory of

Donald Morison Smith, Ph.D., U.E.

(1925-1993)

Foo



(1995-2009)

## Acknowledgements

The journey towards this Ph.D. began in the Fall of 2000. At the time I was pursuing a B.A. in linguistics with no particular intention of finishing, and enrolled in a one-time-only course in Linguistics and Archæology taught by grad student Jean Balcaen. That course unexpectedly inspired me with the goal of becoming the world's leading expert on the long-lost Elamite language. That goal remains unattained, but in the meantime you see the results of my sideline in Sumerian.

I had the privilege of having not one but two excellent supervisors. Elizabeth Cowper inspired me to believe that I could actually be a syntactician; I would go into a meeting with her demoralised and discouraged, but emerge feeling inspired and eager to work. Graeme Hirst came on board to supervise the computational side of things, but just as importantly he contributed immensely to improving the clarity and rigour of my writing.

Thanks also to the third member of my committee, Paul-Alain Beaulieu, who made sure that I stayed on track Sumerology-wise. And thanks to the other members of the defence committee: Cristina Cuervo, whose insights into applicatives were invaluable, and Steven Bird, who coaxed me into making this work more relevant to corpus linguists. Doug Frayne was not on my committee, but his contribution was essential nonetheless, both in contributing the RIM files and in being my first Sumerian teacher.

A special thanks to the taxpayers of Canada, for their gracious support in the form of an Ontario Graduate Scholarship, a Social Sciences and Humanities Research Council Doctoral Fellowship, and additional funding from the Natural Sciences and Engineering Research Council. I am proud to live in a country which has the wisdom to support research such as this.

However, the most important source of support (both financial and otherwise) is my beloved wife Sandra. There were too many occasions to count when she kept me from giving up in despair. Without her love and patience, this document would never have come to be.

I regret that two of those who would most have appreciated the completion of this thesis did not live to see it. One of them is my beloved Foo, who supervised most of this work from his

feline vantage point atop my desk. The other is my father, Donald Morison Smith, who would have greatly appreciated seeing a second Dr. Smith following in his footsteps. This makes me doubly glad that my mother is with me to celebrate this accomplishment, which is a tribute to the spirit of intellectual curiosity I owe to both my parents.

# Contents

<b>1</b>	<b>Overview</b>	<b>1</b>
1.1	Motivation . . . . .	2
1.2	Theoretical Assumptions . . . . .	3
1.3	Application . . . . .	3
1.4	A Sketch of Sumerian Morphology . . . . .	5
1.4.1	Sources of Information . . . . .	5
1.4.2	Writing System . . . . .	6
1.4.3	Noun Morphology . . . . .	8
1.4.4	The Verbal Chain . . . . .	9
1.4.5	Compound Verbs . . . . .	11
1.4.6	Modal Prefixes . . . . .	11
<b>2</b>	<b>Source Corpora</b>	<b>13</b>
2.1	Electronic Text Corpus of Sumerian Literature . . . . .	15
2.1.1	Preprocessing the ETCSL . . . . .	17
2.2	Royal Inscriptions of Mesopotamia . . . . .	18
2.2.1	General Approach . . . . .	19
2.2.2	Adapting the ePSD . . . . .	20
2.2.3	Identifying Part of Speech . . . . .	21
2.2.4	Compound Verbs . . . . .	22

2.2.5	Lexicon Appendix . . . . .	23
2.2.6	Preprocessing . . . . .	24
2.2.7	Regularising Transliterations . . . . .	25
2.2.8	Identifying Paragraphs . . . . .	27
2.2.9	Morphological Processing . . . . .	27
2.2.10	“Amissible” Consonants . . . . .	28
2.2.11	Affix Stripping . . . . .	29
2.2.12	Disambiguating . . . . .	30
2.3	Additional Corpora . . . . .	31
<b>3</b>	<b>Query Infrastructure</b>	<b>33</b>
3.1	Requirements . . . . .	35
3.2	LPath and LPath <sup>+</sup> . . . . .	37
3.2.1	Reference Implementation . . . . .	38
3.2.2	LPath <sup>+</sup> . . . . .	39
3.2.3	Limitations . . . . .	40
3.3	CQL . . . . .	41
3.4	LPattern . . . . .	42
3.5	Query-based Annotation . . . . .	45
3.5.1	Context within Annotation Science . . . . .	47
3.5.2	Implementation of Query Objects . . . . .	49
3.6	State of the Annotation . . . . .	51
<b>4</b>	<b>Dimensional Prefixes</b>	<b>56</b>
4.1	Applicatives . . . . .	59
4.2	Dimensional Prefixes as Applicative Morphology . . . . .	63
4.2.1	Dative Prefix . . . . .	65
4.2.2	Comitative Prefix . . . . .	69



4.2.3	Allative Prefix . . . . .	72
4.2.4	Ablative Prefix . . . . .	75
4.2.5	Locative Prefix . . . . .	76
4.2.6	Locative 2 Prefix . . . . .	77
4.3	Krecher’s Rule . . . . .	79
4.4	Typological Context . . . . .	81
4.5	Summary . . . . .	83
<b>5</b>	<b>Conjugation Prefixes</b>	<b>85</b>
5.1	Earlier Theories . . . . .	86
5.2	Michalowski 2004 . . . . .	91
5.3	Woods 2008 . . . . .	92
5.4	Are Conjugation Prefixes a System of Voice? . . . . .	99
5.5	Conjugation Prefixes as Inner Aspect . . . . .	106
5.6	Summary . . . . .	115
<b>6</b>	<b>Conclusion</b>	<b>117</b>
<b>A</b>	<b>Query Objects</b>	<b>120</b>
<b>B</b>	<b>LPattern Grammar</b>	<b>127</b>
<b>C</b>	<b>Lemmatiser Source Code</b>	<b>129</b>
	<b>Bibliography</b>	<b>139</b>

# List of Figures

3.1	Defining the V-DAT.3SG object . . . . .	50
4.1	Structure of high vs. low applicatives . . . . .	61
4.2	Causative and inchoative constructions . . . . .	62
4.3	Dative agreement with applicative <i>v</i> head (based on (Béjar, 2003)) . . . . .	64
4.4	Dative agreement with APPL <sub>BEN</sub> head . . . . .	65
4.5	Tree for example (4.8) with high APPL <sub>BEN</sub> . . . . .	68
4.6	Allative-case nominal with locative <i>ni-</i> agreement prefix . . . . .	77
4.7	Agreement head to account for Krecher’s Rule . . . . .	84
5.1	Organisation of primary prefixes according to prototypical usage . . . . .	93
5.2	Tree for unaccusative <i>mu</i> <sub>2</sub> . . . . .	103
5.3	Tree for transitive <i>mu</i> <sub>2</sub> . . . . .	104
5.4	Structure for agentive causative <i>mu</i> <sub>2</sub> (after (Kallulli, 2006)) . . . . .	105
5.5	Structures of accomplishments and statives (MacDonald, 2006) . . . . .	108
5.6	Tree for middle-voice <i>mu</i> <sub>2</sub> . . . . .	111
5.7	Trees for <i>šu ba-ti</i> vs. <i>šu imma-ti</i> . . . . .	113
5.8	[medial] feature on a <i>v</i> <sub>DO</sub> head . . . . .	114
5.9	Contrastive features for inner aspect . . . . .	115

# List of Tables

1	Abbreviations used in glosses . . . . .	xiii
2	Abbreviations for cited texts . . . . .	xiv
1.1	Case clitics . . . . .	8
1.2	Semantics of “dimensional” cases . . . . .	9
1.3	Modal prefixes . . . . .	12
2.1	Transformation of a typical ETCSL entry . . . . .	17
2.2	<i>kar</i> <sub>2</sub> as a compound verb . . . . .	23
2.3	Information contained in the lexicon appendix . . . . .	24
2.4	Some concatenated attributive expressions from RIM . . . . .	26
3.1	LPath operators added to XPath . . . . .	37
3.2	Example LPath queries . . . . .	38
3.3	Example CQL queries . . . . .	42
3.4	LPattern operators . . . . .	44
3.5	Functions of corpus/annotation tools (McEnery and Rayson, 1997) . . . . .	48
3.6	Storage of query objects . . . . .	49
3.7	Summary of current annotations . . . . .	52
3.8	Ambiguous queries for dimensional prefixes . . . . .	53
3.9	Defining the NP query object . . . . .	54
4.1	Dimensional clitics, prefixes, and cooccurrences . . . . .	57

4.2	Forms of dimensional prefixes . . . . .	59
4.3	Argument introducers . . . . .	60
4.4	Summary of $\phi$ -feature agreement for dimensional prefixes . . . . .	64
4.5	Dative case agreement morphology . . . . .	66
4.6	Comitative $\phi$ -feature agreement morphology . . . . .	69
4.7	Occurrences of comitative $\phi$ -feature morphology . . . . .	70
4.8	Abilitative use of comitative in the NBGT . . . . .	71
4.9	Allative $\phi$ -feature agreement morphology . . . . .	72
4.10	Occurrences of allative $\phi$ -feature agreement morphology . . . . .	73
4.11	Locative 2 prefix . . . . .	78
4.12	Evidence from the corpus for Krecher's Rule . . . . .	80
5.1	Features of conjugation prefixes . . . . .	89
5.2	Contexts for middle prefixes . . . . .	94
5.3	Cooccurrences of $\tilde{g}al_2$ and $til_3$ with conjugation prefixes . . . . .	97
5.4	Cooccurrence of conjugation prefixes with dative arguments . . . . .	98
5.5	Most frequent verbs for <i>mu-</i> . . . . .	99
5.6	Most frequent verbs for <i>imma-</i> . . . . .	100
5.7	Most frequent verbs for <i>ba-</i> . . . . .	101
5.8	Most frequent verbs for <i>i-</i> . . . . .	102
5.9	Taxonomy of event introducers . . . . .	102
A.1	Queries for defining NP objects . . . . .	121
A.2	Queries for annotating NP objects . . . . .	122
A.3	Queries for dimensional prefixes . . . . .	123
A.4	Queries for conjugation prefixes <i>ba-</i> and <i>mu-</i> . . . . .	124
A.5	Queries for conjugation prefixes <i>imma-</i> and <i>immi-</i> . . . . .	125
A.6	Queries for conjugation prefix <i>i-</i> . . . . .	126

Table 1: Abbreviations used in glosses

Abbr.	Name	Alternate names
1	1st person	
2	2nd person	
3	3rd person	
ABL	ablative case	
ABS	absolutive case	ablative-instrumental
ACC	accusative case	
ACT	active voice	
ALL	allative case	terminative, directive
AP	antipassive	
APPL	applicative	
ASP	aspect	
BEN	benefactive $\theta$ -role	
COM	comitative case (and $\theta$ -role)	
CONJ	conjugation prefix	
COP	copula	
DAT	dative case	
DEM	demonstrative	
EQU	equative case	
ERG	ergative case	
FOC	focus	
FUT	future	
GEN	genitive case	
IPFV	imperfective aspect	<i>marû</i>
INT	intensifier	
INTR	intransitive	
N	inanimate	non-human, non-personal
LOC	locative case	
LOC2	locative 2 case	locative-terminative, directive
MOOD	mood	
ORD	ordinal	
PASS	passive voice	
PD	pronoun	
PFV	perfective aspect	<i>hamtu</i>
PL	plural	
POSS	possessive	
SG	singular	
SRC	source $\theta$ -role	
STAT	stative	
SUB	subordination/nominalisation	

Table 2: Abbreviations for cited texts

Abbr.	Name	Catalogue
CAk	The cursing of Agade	ETCSL 2.1.5
EmkLA	Enmerkar & the Lord of Aratta	ETCSL 1.8.2.3
Eanatum	Royal inscription	RIME 1.9.3.5
Enanatum I	Royal inscription	RIME 1.9.4.8, 1.9.4.19
Enmetena	Royal inscription	RIME 1.9.5.1
Enšakušana	Royal inscription	RIME 1.14.17.1
EnkNh	Enki & Ninhursağa	ETCSL 1.1.1
EnlNI	Enlil & Ninlil	ETCSL 1.2.1
EnlSu	Enlil & Sud	ETCSL 1.2.2
GgAk	Gilgameš & Aka	ETCSL 1.8.1.1
GgEN	Gilgameš, Enkidu and the nether world	ETCSL 1.8.1.4
GgHw-B	Gilgameš & Huwawa (Version B)	ETCSL 1.8.1.5.1
Gudea Cyl	Gudea Cylinder	ETCSL 2.1.7, RIME 3/1.1.7
InEnk	Inana & Enki	ETCSL 1.3.1
LUr	The Lament for Ur	ETCSL 2.2.2
Nanna A	Nanna A	ETCSL 4.13.01
NinTrtl	Ninurta & the Turtle	ETCSL 1.6.3
Proverbs 5	Proverb collection 5	ETCSL 6.1.05
Proverbs 13	Proverb collection 13	ETCSL 6.1.13
Proverbs Ur	Proverbs from Urim	ETCSL 6.2.3
Šulgi C	A praise poem of Šulgi	ETCSL 2.4.2.03
Ur-Namma A	The death of Ur-Namma	ETCSL 2.4.1.1
Utu-heğal	Victory of Utu-heğal	ETCSL 2.1.6, RIME 2.13.6.3

# Chapter 1

## Overview

The research described in these pages is intended to bridge a perceived gap between Sumerology on the one hand and theoretical and corpus linguistics on the other. The theoretical tools which are provided by modern syntactic theory have been all but ignored in the study of Sumerian, and by applying them we can achieve a better understanding of Sumerian. Similarly, while corpora of Sumerian texts do exist, their usefulness could be greatly enhanced by applying recognised techniques from corpus linguistics.

Historically, the study of Sumerian has been carried out in almost complete isolation from theoretical linguistics. On the one hand, this hinders the progress of our understanding of Sumerian, by failing to take into account modern advances in linguistic theory, particularly in the study of syntax. But also it means that linguistics as a whole suffers, because data from Sumerian is ignored by theoretical linguists. This is particularly unfortunate, because Sumerian is a language isolate with many interesting features, and by studying Sumerian we help to shed light on broader cross-linguistic questions.

There is also a divide between corpus linguistics and Sumerology. While there do exist corpora of Sumerian texts, they tend to be organised around the needs of archaeologists and not those of corpus linguists. Queries against these corpora are largely limited to basic string searches against the transliterated texts, and linguistic annotation ranges from rudimentary to

non-existent.

The field is thus ripe for the introduction of new tools, and the research described here has two complementary goals. The first is to take techniques from corpus linguistics and apply them to extracting linguistic data from corpora of Sumerian texts. The second is to take this newly-accessible data and to consider it in the light of modern syntactic theory. The final result will be new theoretical analyses for two important phenomena of Sumerian morphosyntax: the dimensional prefixes and the conjugation prefixes.

## 1.1 Motivation

The original motivation for this undertaking grew out of earlier efforts to use existing corpora to study the syntax of Elamite and the phonology of Sumerian (Smith, 2006b, 2007b). Such efforts were severely hampered by the inadequacy of the search facilities provided by these corpora. In particular, the peculiarities of the cuneiform writing system made the process of extracting linguistic information from corpora of transliterated texts cumbersome, repetitive, and time-consuming.

In large part, this difficulty arises because cuneiform does an inexact job of representing a language's phonology and morphology. A particular morpheme may show up orthographically in a variety of ways. Often this is due to ordinary morphophonological processes, such as assimilation and vowel harmony, but there is also a certain amount of orthographic variation which cannot be attributed to changes in the phonology being represented. In either case, it is generally not possible to identify occurrences of a morpheme by a simple string query against the transliterated cuneiform text. More typically, multiple queries are required, and the resulting hits must then be manually sifted to extract the desired morphemes.

As linguists, we are primarily interested not in the orthography, but rather in the linguistic content which the orthography encodes. Thus, the first task was to set up a query apparatus to meet the needs of linguists, making it possible to deal with linguistic entities rather than or-



thographic ones. Once such an apparatus was in place, the investigation of linguistic questions would become much more straightforward.

## 1.2 Theoretical Assumptions

As Johnson (2004) put it, “One of the more troubling aspects of recent work on Sumerian morphosyntax is that it has remained persistently morphological in orientation and avoided, wherever possible, syntactic argumentation or investigation.” Johnson’s lament is worth echoing, and it should be added that this “troubling aspect” is hardly restricted to recent work on Sumerian, but extends rather to the very earliest studies of the language. The tradition in Sumerian has always been philological rather than linguistic, and syntax has therefore always been neglected in favour of morphology.

The formal syntactic framework employed here will be the one best described as “mainstream generative grammar”, which has also been known as “minimalism”. From a practical standpoint, the choice of framework is not critical, but this particular one was chosen because it has a track record of being applied by linguists to a wide variety of languages. There is every reason to believe that this framework is equally well-suited to describing the syntax of Sumerian.

One of the most important theoretical concepts employed in this discussion is the notion of applicative heads (Pylkkänen, 2002; Cuervo, 2003), which serve to introduce the arguments of a verb. Since much of the discussion of applicatives comes from work done in the minimalist tradition, it seemed best to continue working within the same formal framework.

## 1.3 Application

The original intention was to use the newly-established query apparatus to investigate syntactic questions in two corpora: one of Elamite-language texts, and one of Sumerian texts. This was

intended to demonstrate that the query apparatus was general-purpose and not specific to any one language. However, as the research proceeded, it became clear that the effort devoted to the Elamite corpus would be better spent improving the quality of the Sumerian corpus, so the Elamite corpus was set aside and the immediate focus was limited to Sumerian.

In order to have as broad as possible a range of Sumerian texts, a composite corpus was built, as described in Chapter 2. The construction of this composite corpus included the creation of tools to parse an unannotated corpus in order to create a corpus with part-of-speech tags. These particular tools were developed for incorporating the Royal Inscriptions of Mesopotamia, but the techniques employed should be equally applicable to any corpus of transliterated Sumerian texts.

The query apparatus itself is described in Chapter 3, and consists of a language, LPattern, along with a facility for building what are referred to as “query objects”. Each of these query objects encapsulates a set of LPattern queries, allowing those queries to be reused and to be treated as a unit for building subsequent queries. Once the linguist has created a set of query objects which represent the morphemes of interest, it becomes possible to investigate the contents of the corpus at a morphological level, without having to be constantly concerned with the details of the language’s orthography. The orthographic information is still accessible, and can be referred to when needed, but the linguist is no longer restricted to working only at the level of orthography.

The first problem to be examined using these tools is the question of Sumerian dimensional prefixes, the subject of Chapter 4. These are shown to represent not a phenomenon of agreement or concord, as has been argued in the past, but rather the morphological realisation of applicative heads corresponding to the verb’s thematic roles. In fact, the rich set of applicative heads in Sumerian extends our understanding of what it is possible for applicatives to do in a language.

The second theoretical question is the function of the conjugation prefixes, discussed in Chapter 5. Here, much of the groundwork has already been done by Woods (2008). However,

in the tradition of Sumerian studies, Woods avoids making any appeal to any sort of formal syntactic framework. This is remedied by analysing the conjugation prefixes as a system of inner-aspectual features which are carried by light-verb heads.

Sumerian is of course a human language just like any other. The tools of minimalist syntax and of corpus linguistics are just as applicable to Sumerian as they are to other languages. What has been lacking until now is the desire to bring Sumerology out of its isolation and into the broad, sunlit uplands of modern linguistics.

## **1.4 A Sketch of Sumerian Morphology**

The primary goal of this dissertation is to provide an account of the syntax of the Sumerian verb. However, before addressing the syntactic aspects of the verbal system, it is necessary to provide a description of the Sumerian verbal system within the broader context of Sumerian as a whole. This also provides the opportunity to present a brief overview of the Sumerian language for readers who may not be familiar with the notation and terminology employed by Sumerologists. For a much more complete overview, readers are referred to the excellent summary by Michalowski (2004).

### **1.4.1 Sources of Information**

Sumerian is a language isolate which is known to us from texts written using the cuneiform writing system. The earliest known cuneiform texts date to ca. 3200 BC, and it is likely (though far from certain) that these very early texts are written in the Sumerian language. The high point of classical written Sumerian is the period from 2400 BC to 2004 BC. At some point after 2000 BC the Sumerian language ceased to have native speakers, but it continued to be a prestige language which was used by scribes whose native language was Akkadian. Sumerian continued to be written until possibly as late as 200 AD (Geller, 1997).


Since Sumerian is an isolate, our knowledge of Sumerian is largely seen through the lens of Akkadian. Because Akkadian, as a Semitic language, is relatively well understood, it provides a foothold from which we can start to understand Sumerian. That being said, our reliance on Akkadian can also be a hindrance, since it often obscures details of Sumerian which were not apparent to Akkadian speakers.



Vital to our understanding of Sumerian are the various types of texts which were used by Akkadian-speaking scribes who needed to learn how to write Sumerian in the absence of any native speakers of the language. Among these materials are tablet sets known as “lexical lists”, which simply list Sumerian words with their Akkadian synonyms. This practice of creating lists also extends to “grammatical texts”, which are lists of inflected forms in Sumerian with the corresponding Akkadian forms. As well, there are “syllabaries”, which provide lists of cuneiform characters with their associated phonetic readings.

In addition to these educational texts, there are numerous bilingual inscriptions which contain parallel texts in Sumerian and Akkadian. This is particularly often the case for royal inscriptions, which continued to be written in both languages long after the demise of Sumerian as a spoken language.

## 1.4.2 Writing System

This dissertation does not concern itself directly with the cuneiform writing system. However, it is worth explaining some of the details of the cuneiform writing system as well as the conventions which are used to transcribe cuneiform.

Most cuneiform characters can be read either as logograms or as phonograms. Thus  can be read logographically as the word *UD* ‘sun, day’ or phonographically as the sound /ud/. In general, a sign is transcribed lower-case if it is being read as a phonogram and in capital letters if it is being read as a logogram, or if its phonetic value is unknown.

There are special notations for signs which are composed of other signs. So for instance, *MUNUS+NI* would refer to a sign which consists of *MUNUS*  followed by *NI*  , while

$URU \times A$  would consist of  $A$  𒀭 written inside  $URU$  𒌦.

Complicating matters, signs may have multiple readings and multiple signs may share the same reading. For example, the  $UD$  𒌵 sign can also have a phonographic reading of /u/. In such cases 𒌵 is transcribed as  $u_4$ , to indicate that it is in fact the fourth sign having a reading of /u/, sharing that phonographic value with the signs  $u$  𒌶,  $u_2$  𒌷,  $u_3$  𒌸, and a range of less common signs.

With the exception of a handful of vowel (V) signs, most phonograms represent the combination of a vowel and a single consonant (CV or VC). Although there are a large number of CVC signs, the system lacks CVC signs for many combinations of consonants and vowels. For example, since there is no *lan* sign, a closed syllable like /lan/ would either have to be written *la-an* 𒌶 𒌶, or else it would be written as *la* 𒌶, and it would be up to the reader to reconstruct the missing /n/. In addition to this tendency to avoid writing syllable codas, Sumerian had a phonological rule which dropped word-final obstruents (“amissible consonants”). Thus, the genitive case clitic *-ak* is not written with the *ak* 𒌶𒌵 sign, but usually manifests itself as a suffixed *-a*. Only when the genitive is followed by another case clitic does the /k/ appear in the writing.

In general, the accuracy with which the writing system reflected the spoken language changed over time. In the earliest texts, writing seems to have had largely a mnemonic purpose: only lexical items were written, and the written order of signs did not necessarily correspond to the spoken word order. By about 2400 BC, the written order consistently reflects the word order, and affixes are generally written. But it is only after Sumerian ceases to be a natively-spoken language that scribes take the care to fully indicate the inflectional morphology.

As with other parts of Sumerian, our knowledge is filtered through Akkadian, so the phonographic readings are based on reconstructed Akkadian pronunciations. There are a few cases where the Akkadian and Sumerian pronunciations are known to differ. For instance, in Akkadian, both *ga* 𒌶𒌵 and *ga<sub>2</sub>* 𒌶𒌶 are read as /ga/. In Sumerian phonology, a velar nasal has been reconstructed, so the 𒌶𒌵 sign is still transcribed as *ga*, but 𒌶𒌶 is transcribed as *g̃a<sub>2</sub>* to

indicate a likely pronunciation of /ŋa/.

### 1.4.3 Noun Morphology

Sumerian nouns are divided into two classes which are usually referred to as “animate” and “inanimate”. As Michalowski (2004) points out, the terms “personal” and “non-personal” might be more accurate, since actual animals fall into the “inanimate” class.

Sumerian nouns have 10 possible cases, which are indicated by case clitics. These clitics are written at the end of the noun phrase to which they apply. With the exception of the genitive, which can cooccur with any of the other cases, only one case clitic can occur on a given noun phrase. These are summarised in Table 1.1, using the terminology employed by Michalowski.

Table 1.1: Case clitics (after Michalowski (2004) and Thomsen (1984))

Case	Abbreviation	Form	Notes
Ergative	ERG	-e	
Absolutive	ABS	-∅	
Genitive	GEN	-ak	
Dative	DAT	-ra	Animate nouns only.
Comitative	COM	-da	
Ablative	ABL	-ta	Inanimates only.
Allative	ALL	-še <sub>3</sub>	Traditionally referred to as “terminative”.
Locative	LOC	-a	Inanimates only.
Locative 2	LOC2	-e	Inanimates only. Traditionally referred to as “locative-terminative”.
Equative	EQU	-gin <sub>7</sub>	Denotes comparison (e.g. <i>lugal-gin<sub>7</sub></i> ‘like a king’).

Edzard (2003) sums up the relationship between the locative, the allative, and the locative 2 in terms of motion and location, as shown in Table 1.2. Although Edzard does not mention the ablative in this context, it can be considered to be a fourth member of this system.

Table 1.2: Semantics of “dimensional” cases (after Edzard (2003))

Case	Meaning
Locative	motion into, position inside
Allative	motion towards, position in front of
Locative 2	motion arriving at, position next to
Ablative	motion out of

### 1.4.4 The Verbal Chain

The Sumerian verb is highly inflected, typically appearing with a long chain of affixes, which have long been the subject of debate among Sumerologists. At the most fundamental level, there is disagreement over the number and morphological realisation of these affixes, which is exacerbated by the fact that orthographic system does not accurately reflect the language’s morphology. Even when there is agreement over the presence of a particular prefix, there is typically dispute over its role within Sumerian grammar.

A fairly conservative view of the general order of morphemes within the verbal complex for a transitive verb is shown in (1.1). It is based largely on the summary of Sumerian grammar by Michalowski (2004), and to a lesser extent on the earlier grammar description by Thomsen (1984). Not all elements of the chain will necessarily be present, but when elements are present they always occur in the same relative order.

(1.1) Order of elements in the verbal chain (after Michalowski (2004) and Thomsen (1984))

$$\text{MOD} - \text{CONJ} - \text{DAT} - \text{COM} - \left\{ \begin{array}{c} \text{ABL} \\ \text{ALL} \end{array} \right\} - \left\{ \begin{array}{c} \text{LOC} \\ \text{LOC2} \end{array} \right\} - \text{PRO}_1 - \text{Verbal root} - \text{-ed} - \text{PRO}_2^1$$

The MOD or “modal” prefixes are fairly well delineated, and encode such things as negatives, subjunctives, and conditionals. Descriptions and examples of all the modal prefixes are provided in §1.4.6.

<sup>1</sup>Throughout the paper, glosses will use the abbreviations used by the Leipzig Glossing Rules <http://www.eva.mpg.de/lingua/resources/glossing-rules.php>, supplemented where necessary by additional abbreviations listed in [http://en.wikipedia.org/wiki/List\\_of\\_glossing\\_abbreviations](http://en.wikipedia.org/wiki/List_of_glossing_abbreviations).

The CONJ slot is filled by the prefixes which have traditionally been referred to as “conjugation prefixes”, although the name is misleading since they are unlikely to have anything to do with conjugation. The forms and functions of these prefixes are the subject of considerable debate, and are one of the main topics of this dissertation. The current views of these prefixes will be discussed in more detail in Chapter 5, and a theoretical account for them can be found in §5.5.

The DAT, COM, ABL, ALL, LOC, and LOC2 slots are collectively referred to as “dimensional infixes” in the literature, although they are really prefixes rather than infixes. In general, if the prefix for a given case appears on the verb, the clause will also contain a noun phrase in the corresponding case. Gragg (1973) provides the primary study of these prefixes, but the work is dated, and takes no account of modern syntactic theory. These prefixes will be described further in Chapter 4, and a theoretical account for them will be presented in §4.2.

The verbal root reflects the verb’s aspect, which can be either perfective or imperfective. In the literature, these are often referred to as *hamṭu* and *marû* respectively (literally ‘quick’ and ‘fat’) after the Akkadian terms which are found alongside these forms in ancient grammatical texts. The *marû* is typically formed by reduplicating the stem, but for many verbs there are suppletive *marû* forms.

The PRO<sub>1</sub> and PRO<sub>2</sub> slots are what Thomsen (1984) calls the “pronominal” prefix and suffix. These slots contain morphemes which show agreement in person, number and animacy with the subject and object of the verb. In the perfective aspect, PRO<sub>1</sub> agrees with the person, number, and animacy features of the ergative subject, while PRO<sub>2</sub> typically agrees with the person and number of the absolutive object.<sup>2</sup> In the imperfective aspect, the verb shows nominative/accusative agreement, with PRO<sub>1</sub> agreeing with the accusative-case object, while PRO<sub>2</sub> agrees with the nominative-case subject.

The suffix *-ed* is also the source of considerable disagreement. It does not appear to be

---

<sup>2</sup>However, there are at least some cases where PRO<sub>2</sub> agrees with the subject rather than the object. This may be related to the verb involved.



an agreement marker of any sort, but rather seems to be associated with future events. This extends to events which involve obligation, so *-ed* may act as a sort of modal suffix.

### 1.4.5 Compound Verbs

One special type of verbs is the “compound verb”, which consists of a verb preceded by an uninflected noun, often with a very idiomatic meaning. So for instance the verb *sa<sub>2</sub>* ‘to equal’ preceded by *si* ‘horn’ becomes *si sa<sub>2</sub>* ‘to set in order’. In such verbs, the uninflected noun serves notionally as the direct object, so the actual direct object has to appear in one of the oblique cases, typically dative or locative 2. Johnson (2004) describes these compound verbs as being analogous to English particle verbs such as “hammer out” and “tie up”, where the particle adds the notion of telicity to the base verb along with some sort of idiomatic semantics.

### 1.4.6 Modal Prefixes

In an ordinary indicative sentence, the leftmost position in the verbal prefix chain, the slot for modal prefixes, is empty. As their name suggests, it is generally agreed that when modal prefixes are present, they alter the mood of the verb. There are prefixes for subjunctives, conditionals, some imperatives, and a variety of “wish” forms. In addition, the negative prefix *nu-* falls into this category; although negation is not traditionally considered to be a mood, negation can still be said to reflect a relation between the real world and the proposition in the sentence, just as modals do.

For several of the modal prefixes there is general agreement on their interpretation. For instance, *ga-* is agreed to be a cohortative or first-person wish form. Similarly, *u-* is seen to be the “prospective” marker or “prefix of anteriority”, indicating a sequence of events in narratives or instructions.

One place where interpretations of the prefixes diverge are in the cases where the meaning seems to be quite different depending on whether the verb is in the perfective or imperfective aspect. This is the mainstream traditional view among Sumerologists, as exemplified by

Thomsen (1984) and Edzard (2003). Civil (2000/2005) disputes this, arguing that the relation is the other way around, and that it is modality which largely determines the aspect of the verb. His account relies on a distinction between prefixes which express epistemic modality (alternative worlds which could exist instead of the present one) and prefixes which express deontic modality (alternative worlds which could develop out of the current one); since deontic modality typically deals with events which have not yet been completed, they are naturally associated with verbs in the imperfective.

These differing views of the possible readings for the modal prefixes are summarised in Table 1.3. It shows that while there is general agreement on which prefixes are present, for several of the prefixes there is no consensus on their semantics.

Table 1.3: Modal prefixes (after Civil (2000/2005), Thomsen (1984), and Edzard (2003))

Prefix	Civil (2000/2005)	Thomsen (1984); Edzard (2003)
∅-	indicative	
<i>nu-</i>	negative indicative	
<i>ga-</i>	cohortative (1st person wish)	
<i>u-</i>	prospective	
<i>ha-</i>	deontic optative (with imperfective) epistemic subjunctive (with perfective)	with imperfective: precative with perfective: affirmative
<i>bara-</i>	negative epistemic subjunctive	with imperfective: negative precative with perfective: negative affirmative
<i>na-</i>	negative deontic optative	with imperfective: prohibitive with perfective: affirmative
<i>ša-</i>	<i>meaning disputed</i>	<i>meaning disputed</i>
<i>iri-</i>	<i>not mentioned</i>	<i>meaning unknown</i>
<i>nuš-</i>	<i>not a prefix</i>	frustrative (“if only X”)

However, the actual details of the semantics of the modal prefixes is peripheral to the theoretical goals of this dissertation. From a theoretical standpoint, these morphemes can all be accounted for as manifestations of a MOOD head. The exact features of such a head will differ depending on whether one accepts Civil’s view or the traditional view, but the basic mechanism is the same. This dissertation is concerned with projections lower than MOOD, starting with VOICE and preceding downwards.

## Chapter 2

### Source Corpora

For Sumerian there are a number of available electronic corpora. However, none of these existing corpora has quite the desired range of chronological coverage and richness of linguistic annotation which is required for the sorts of syntactic analysis being undertaken here.

Thus, the corpus to be used for the actual research would have to be a derived corpus, built upon existing corpora of Sumerian texts. Each base corpus required a pre-processing step in order to convert it to the format of the working corpus, and as part of the pre-processing step, a limited amount of morphological annotation, notably lemma and part-of-speech information, was also compiled. However, the intent was that the bulk of the morphosyntactic annotation of the corpus would be created as part of the query-based annotation mechanism described in Chapter 3.

As pointed out by Sinclair (2005), it is important for a corpus to be as representative as possible of the linguistic material being studied. To this end, the ideal base for the working corpus would consist of texts in a range of genres, drawn from a broad range of chronological periods. However, the available resources were limited, and the task of creating a general-purpose lemmatiser and part-of-speech tagger for unannotated Sumerian texts was deemed impractical.

By far the largest existing corpus of Sumerian texts is the Cuneiform Digital Library Initia-

tive (CDLI) from UCLA and the Max Planck Institute (Englund and Damerow, 2000–2005). It has a broad range of texts from all periods, but the focus of the project is archaeological rather than linguistic. Consequently, the entry for each text contains catalogue information, provenance, and images, but the texts themselves are only provided in transliteration with no translation or morphological markup. This is also true of a number of other smaller corpora associated with the CDLI, such as the Digital Corpus of Cuneiform Lexical Texts (Veldhuis, 2003) and the Database of Neo-Sumerian Texts (Molina, 2002–2010).

Of great interest was the Pennsylvania Parsed Corpus of Sumerian (Tinney and Karahashi, 2003-2004), which was conceived as a hand-parsed treebank in the style of the English-language Penn Treebank. Such a corpus would have been close to ideal for the purposes of identifying the morphosyntactic phenomena being studied in this thesis. Unfortunately, work on the corpus seems to have stopped, and the corpus has never been publicly released.

Of all the broadly available corpora, Oxford's Electronic Text Corpus of Sumerian Literature (ETCSL) was selected as being the most suitable base corpus. In addition to transliterations, the corpus provides English translations, and the Sumerian text has already been lemmatised and tagged for part-of-speech. Although the corpus consists only of literary texts and is drawn largely from the Old Babylonian period, the limited amount of pre-processing made it an appealing choice.

In addition, it was possible to acquire the original Microsoft Word documents which were used to create the Sumerian volumes of the Royal Inscriptions of Mesopotamia (RIM) project. The RIM texts include English translations, but no morphological markup. Due to its relatively small size, and due to the accompanying English translation, the RIM texts were more practical than the CDLI as a target for lemmatising, so a tagger/lemmatiser for the RIM was developed as a useful experimental effort. Using this software, described in §2.2, the first three RIM volumes, consisting of royal inscriptions from the earliest periods up to the end of the Ur III period, were lemmatised and tagged with part-of-speech information.

## 2.1 Electronic Text Corpus of Sumerian Literature

The ETCSL consists of 394 texts from genres which Sumerologists classify as “literary”: mythological epics, royal praise poems, literary letters, laws, hymns, cult songs, and proverbs. The corpus totals approximately 170 000 words of text. While 170 000 words is not a large corpus by the standards of corpus linguistics, for Sumerian it is quite substantial.

One significant advantage the ETCSL has over other corpora of Sumerian texts is that the ETCSL has been lemmatised to the extent that the words in the corpus have all been tagged for part of speech. This process was described at an ATALA conference (Ebeling and Cunningham, 2005) and on the ETCSL web-site (Black et al., 1998–2006).

The majority of the texts date from a fairly narrow period (ca. 2200–1600 BCE), so the corpus is quite cohesive. Where variants exist they have been edited by the team at Oxford into a standardised form.

The XML source files for the corpus were made available by Jarle Ebeling and his colleagues. The corpus is organised as shown in (2.1), with the top level being the `<text>`, which represents a self-contained document, possibly several hundred lines long. Below the `<text>`, some of the documents are further subdivided using `<div1>` tags (used when there are lacunæ in the text) and `<lg>` tags (to group lines in certain genres, such as proverbs within a proverb collection). In addition, a small number of compound verbs (see §1.4.5) have been enclosed in `<phr>` tags. However, these intermediate groupings are not reliably present, so it is not safe to depend on them.

### (2.1) Hierarchical structure within ETCSL

**Top-level** `<text>`

**Intermediate groupings** `<div1>`, `<lg>`, `<phr>`

**Lines** `<l>`

**Words** `<w>`

The one grouping which is reliably present is the line, <1>. Unfortunately, in cuneiform texts there is no consistent correlation between lines and sentence boundaries. The line is purely a scribal unit and may only incidentally correspond to a linguistic unit. The lack of phrase or sentence boundaries is a significant disadvantage for investigating syntactic questions, since the phenomena being explored are expected to be scoped to a single clause or sentence.

Some typical word entries from the ETCSL are shown in (2.2). At first glance, the ETCSL provides a fair bit of morphological annotation. The bound attribute seemed particularly promising, since it promises a morpheme-by-morpheme breakdown of each word. Unfortunately, the bound attribute is only present on a handful of nouns. Similarly, the form-type attribute is not as useful as it might be because it too is used for only a limited range of forms.

## (2.2) Sample word entries from the ETCSL

```
<w form="nu-gi4-gi4" lemma="gi4" pos="V" label="to return"
form-type="RR">nu-gi4-gi4</w>
<w form="iri&ki;-za" lemma="iri" pos="N" label="town"
bound="L,zu.a" det="&ki;">iri&ki;-za</w>
<w form="ki-en-gi-ra" lemma="ki-en-gi" pos="N" type="GN"
label="Sumer">ki-en-gi-ra</w>
```

**form** orthography

**lemma** standardised citation form/lexeme

**pos** part of speech

**type** further sub-grouping of *pos* (e.g. PN, DN)

**label** English gloss

**form-type** morphological information on word (e.g. reduplicated)

**bound** segmentational information (e.g. ergative-case suffix)

**det** determinative

### 2.1.1 Preprocessing the ETCSL

As discussed by Leech (1997), the choice of a tagset for a new corpus is always a tradeoff between what is linguistically desirable and what is computationally feasible. From the standpoint of computational feasibility, the best approach would have been to directly employ the ETCSL's schema as the schema for the working corpus. However, ETCSL's tagset is not ideal for the sorts of syntactic queries being performed, so a small amount of additional work was done during the preprocessing stage in order to make the corpus more accessible.

In particular, given the importance of part-of-speech information in any syntactic analysis, this information was given prominence by making the word-level tags directly reflect the part of speech. As well, prefixes, suffixes, and reduplication were identified and stored as separate attributes of each word. Early experimentation with the corpus suggested that it was helpful to be able to refer to these items separately. LPattern queries are easier to write given the knowledge that a particular grapheme is in the prefix or the suffix.

The transformations between the original ETCSL entries are not major, as shown in Table 2.1. It is this transformed format which defines the target format for importing the RIM data. Similarly, if data from other corpora, such as the CDLI, were to be incorporated into our corpus, this would be the target format.

Table 2.1: Transformation of a typical ETCSL entry

Before:	<code>&lt;w form="nu-gi4-gi4" lemma="gi4" pos="V" label="to return" form-type="RR"&gt;nu-gi4-gi4&lt;/w&gt;</code>
After:	<code>&lt;V prefix="nu-" lemma="gi4" english="to return" stem="gi" orth="nu-gi4-gi4" reduplication ="+"/&gt;</code>

The other significant preprocessing done to the ETCSL files was to group the text into `<para>` entries. As described above, the ETCSL's `<text>` are too large to be useful for scoping syntactic phenomena, while its `<l>` elements only occasionally correspond to syntactic units.

Although true clause boundaries are not available, it is still useful to have some sort of structural units within `<text>` elements, in order to narrow the scope of queries. Fortunately, the ETCSL does indicate which lines of the transliteration correspond to a block of complete sentences in the English translation. In some cases, these translation blocks consist of a single English sentence, and it is safe to assume that the `<para>` corresponds to a single Sumerian sentence. More often, the end of a sentence fails to line up neatly with the end of a Sumerian line, so a translation block consists of multiple English sentences, which means that the `<para>` element contains multiple Sumerian sentences. While this is not ideal, it is still much better than having no structure at all below the `<text>`.

## 2.2 Royal Inscriptions of Mesopotamia

The texts in the ETCSL have the disadvantage of being drawn largely from the Old Babylonian period, which post-dates the demise of Sumerian as a natively-spoken language. For the purpose of understanding the syntax of Sumerian as employed by native speakers, it is necessary to include texts from earlier periods. One collection of such texts is contained in the Early Series of the volumes compiled by the Royal Inscriptions of Mesopotamia (RIM) project (Frayne, 1993; Edzard, 1997; Frayne, 1997, 2005). These include texts from the pre-Sargonic period (ca. 2700 BC) through to the end of the Ur III period (2004 BC).

While the RIM texts include English translations, they have nothing in the way of morphological markup. Most significantly, the RIM contains no part-of-speech information. A lemmatiser had to be developed to take the RIM texts and add the same level of morphological markup and part-of-speech information found in the ETCSL. The general approach of the lemmatiser is a straightforward one, taking Sumerian words from the RIM texts and matching them up with entries in the Electronic Pennsylvania Sumerian Dictionary (ePSD) (Sjöberg et al., 2004).

Because the ultimate goal is syntactic analysis, the single most important task of the lem-



matizing process is to correctly assign words to their part of speech. While it is also important to identify a word's affixes, its stem, and its English gloss, without knowing a word's part of speech these secondary attributes are much less useful.

### 2.2.1 General Approach

Since the goal is to create a composite corpus which is compatible in structure to the ETCSL, the existing structure used for the ETCSL texts (§2.1.1) provides the target format for lemmatising the RIM. Earlier work with the ETCSL had shown the benefit of a certain amount of preprocessing in order to make the ETCSL data easier to work with using LPattern.

The RIM texts are taken from their original form as Word documents, edited and converted to XML. The first stage of the lemmatisation process takes the XML input and runs it through a SAX (Simple API for XML) parser, converting the input text into a stream of Sumerian words. These Sumerian words are organised into “paragraphs” based on the associated English translation.

The second stage takes the stream of Sumerian words, strips them of inflectional morphology and attempts to match them up against entries in the lexicon. For each word, a set of possible lexemes is determined. Where there is only one possible lexeme, that is recorded as being the correct one. If there are no matching lexemes, an error is logged and the word is flagged for manual processing. The relevant source code for this stage can be found in Appendix C as the `EPSDLoder::createWordNodes` method.

The third stage of processing takes all the words which have been identified as having multiple possible lexemes, and attempts to resolve the ambiguity. For each of the lexemes, all of the available English glosses are checked against the English translation of the associated paragraph. If an exact match is found in the English translation, that lexeme is recorded as being the correct one. If no exact match is found, the next step depends on whether the ambiguous lexemes are all from the same part of speech. If they are, the word is recorded with the known part of speech and with a range of possible glosses and lemmata. If they are from different

parts of speech, an error is logged and the word is recorded as having an indeterminate part of speech. This stage corresponds to the body of the `EPSDLoder::attachWord` method in Appendix C.

This whole process is executed repeatedly, with the error logs from each execution being used to improve the results of the next execution. In particular, the errors are used to refine the lexicon appendix. This appendix contains a variety of information missing from the Electronic Pennsylvania Sumerian Dictionary (ePSD) proper, notably named entities, words not found in the ePSD, additional glosses to help guide the lemmatisation process, and complete lists of all known prefixes and suffixes.

The final result of the lemmatisation process is an XML file which contains all the Sumerian-language RIM texts with all the words marked for part of speech, lemma, prefix, suffix, and English gloss. To make the corpus easier to search, the words are grouped into logical paragraphs.

### 2.2.2 Adapting the ePSD

The Electronic Pennsylvania Sumerian Dictionary or ePSD (Sjöberg et al., 2004) is the online version of the printed Pennsylvania Sumerian Dictionary (Sjöberg and Behrens, 1992–2010) which has been under publication since 1974. The ePSD is currently available online, but can only be accessed through a manual query interface, and is not available in a form suitable for processing. However, an earlier edition of the ePSD was published as a single large HTML file, and this formed the basis for the lexicon which was used during the lemmatisation process.

A typical ePSD entry is shown in (2.3). The first `<p>` element provides the canonical form of the word along with a canonical translation, followed by the word's possible transliterations. Transliterations which the editors of the ePSD considered current are shown in bold (i.e. surrounded by `<b>` and `</b>` tags), while older transliterations are shown in plain text. Having older transliterations is important because the RIM volumes tend to follow older traditions of transliteration. Following the headword is a list which contains a series of English glosses,

followed by Akkadian glosses (where known).

(2.3) A typical ePSD entry (Sjöberg et al., 2004) with HTML source

**a[water]: a, e<sub>4</sub>, ea.**

1. "water"
2. "semen"
3. "progeny"

Akk. *mû, rihûtu.*

HTML Source:

```
<p class="item"><b>a[water]:</b> <b>a</b>, e<sub>4</sub>, ea.</p>
<ul>
  <p>1. "water"</p>
  <p>2. "semen"</p>
  <p>3. "progeny"</p>
  <p>Akk. <i>mû</i>, <i>rihûtu</i>.</p>
</ul>
```

### 2.2.3 Identifying Part of Speech

Since the primary goal of this process is to produce a corpus which is tagged for parts of speech, it is important to extract this information as best as possible from the ePSD. Unfortunately, ePSD entries do not actually indicate the part of speech, so we had to rely on the English gloss to provide a best guess as to which part of speech was indicated.

The general rule was that an entry whose gloss was of the form ‘to X’ would be categorised as a verb. In Sumerian, many verbs can also function as adjectives. Any entry whose gloss started with ‘(to be) X’ would be recorded in the lexicon as both a verb ‘to be X’ and as an adjective ‘X’.

Any item whose English gloss contained ‘an exclamation’ or ‘an interjection’ was classified as an interjection. This is important because a number of interjections are small words (e.g. *e*) whose interpretation as nouns would confuse the tagging process.

Anything else was considered to be a noun. As a side-effect, this would tend to misclassify a number of words which should properly be considered conjunctions, pronouns, or other parts of speech. Fortunately, these categories are all relatively small closed sets, so it was possible to assemble a list of overrides to specify part-of-speech for these cases. This list of overrides is stored in the lexicon appendix (§2.2.5).

## 2.2.4 Compound Verbs

One group of ePSD entries which causes particular complications is the class of compound verbs. These consist of a normally-inflected verb which is preceded by a noun. These verbs are semantically bleached, sometimes to the extent of having no identifiable meaning in the absence of a nominal element. Indeed, there are a handful of verbs which never appear independently, the most important of which is *ru*, which is very frequent in the RIM corpus as the compound *a ru* ‘to dedicate’. Similarly, there are nouns which never occur independently, such as *en<sub>3</sub>*, which is only ever found as the nominal element of *en<sub>3</sub> tar* ‘to ask’.

The nominal element of the compound verb serves to narrow down the semantic range of the verb. Many of these nominal elements are body parts such as *šu* ‘hand’, *gu<sub>2</sub>* ‘neck’, *ḡiri<sub>3</sub>* ‘foot’, or *igi* ‘eye’. Often the meaning of the nominal element gives a clue as to the meaning of the compound. For instance, compound verbs with *igi* are usually somehow related to vision<sup>1</sup> while verbs with *ḡiri<sub>3</sub>* are typically related to locomotion.

Often however, the meaning of the compound is quite idiosyncratic. So for instance, *sa<sub>2</sub>* on its own means ‘to equal’, but when combined with *si* ‘horn’, *si sa<sub>2</sub>* means ‘to straighten’ or ‘to put in order’. Treating the sequence *si sa<sub>2</sub>* as ‘to equal the horn’ would be misleading. Another

---

<sup>1</sup>With notable exceptions, such as *igi ru-gu<sub>2</sub>* ‘to oppose’ which is formed from the verb *ru-gu<sub>2</sub>* ‘to withstand’, ‘to sail upstream’.

example is given in Table 2.2, which shows the various compounds of the verb *kar<sub>2</sub>* ‘to shine’ or ‘to illuminate’.

Table 2.2: *kar<sub>2</sub>* as a compound verb

Nominal Element	Meaning
(none)	<i>kar<sub>2</sub></i> ‘to shine’, ‘to illuminate’
<i>aga</i> ‘crown’, ‘tiara’	<i>aga kar<sub>2</sub></i> ‘to defeat’, ‘to conquer’
<i>igi</i> ‘eye’	<i>igi kar<sub>2</sub></i> ‘to examine’
<i>šu</i> ‘hand’	<i>šu kar<sub>2</sub></i> ‘to denigrate’

Initially it was hoped that it would be possible to treat compound verbs as a simple sequence of a noun and a verb, and ignore the idiosyncratic meanings of some of the combinations. After all, the priority in this whole process is to correctly assign parts of speech, while assigning meanings and lemmata is only of secondary importance. However, there were a number of problematic “nouns” which can also do double duty as “verbs”, most notably the *a* of *a ru* and the *si* of *si sa<sub>2</sub>*. For instance, *a* on its own could be a noun meaning ‘father’, ‘house’, ‘water’, or ‘a cry of woe’, but it could equally well be a verb meaning ‘to cry’ or ‘to do’. Early versions of the lemmatiser were misclassifying far too many of these words, so the algorithm was revised to take compound verbs into account, with a corresponding increase in accuracy.

## 2.2.5 Lexicon Appendix

In addition to the ePSD proper, there is an additional file which contains information not found in the ePSD. It was originally intended to record named entities found in the RIM text, but it soon became useful for a variety of purposes in helping the lemmatisation process. The range of entries found in the lexicon appendix is shown in Table 2.3.

In its initial state, the lexicon appendix contained only the prefixes and suffixes listed in Thomsen’s grammar of Sumerian (Thomsen, 1984). Of particular usefulness was Thomsen’s comprehensive list of attested orthographies for verbal prefixes. This was particularly helpful since a Sumerian verb can have a long string of prefixes, and the actual orthographic manifesta-

Table 2.3: Information contained in the lexicon appendix

Example				Explanation
giš	ĝeš			Substitution to replace RIM sign-reading with newer ePSD reading
a-ni	ene	he, she	PD	Override to indicate proper part of speech for a lexeme
-ak	N			Indication of a suffix
e-ma-ni-	V			Indication of a prefix
du11	dug	to speak	V	RIM transliteration which differs from that used by ePSD
bad	bad/r	to open	V	Lemma which is recorded in ePSD with the wrong final consonant
du3	du	built	V	Additional gloss to assist disambiguation
ZIZ2.AN	ziz	emmer wheat	N	Morphogram used by RIM where ePSD has an ordinary reading
gu2 ĝar-ĝar	gu ĝar	to submit	V	Reduplicated form
šu-ur6	šur	angry	AJ	RIM word not found in ePSD
gu3-de2-a	Gudea	Gudea	N	Named entity

tions of prefixes are often not obvious. For instance, for the sequence of morphemes  $ha+i+bi$ <sup>2</sup> Thomsen lists possible orthographies of  $he_2-mi-$ ,  $he_2-em-mi-$ , and  $he_2-me-$ .

## 2.2.6 Preprocessing

The Royal Inscriptions of Mesopotamia are ordinarily available only in printed form;<sup>3</sup> however, Douglas Frayne, the editor of several of the volumes from the RIM Early Series, was kind enough to provide the original Microsoft Word documents which were used to create the printed volumes.<sup>4</sup> These Word files were edited together into a single large Word document

<sup>2</sup>This is Thomsen's analysis of the prefixes involved. Other authors, such as Michalowski (2004), would differ in their interpretation of the prefixes, but for the purposes of lemmatising, these distinctions are not important. All that matters here is that the given sequence of graphemes represents some form of prefix.

<sup>3</sup>A number of the volumes are available online via eBrary, but not in a form which gives direct access to the volumes' text.

<sup>4</sup>Prof. Frayne also provided the source materials for RIME 4 (Frayne, 1990), but these were in troff format; the extra work of importing this format was not deemed to be justified, since the Sumerian texts in that volume

containing the relevant Sumerian transliterations and translations from those volumes.

Descriptive text, historical notes, and provenance information were all excised at this point. In addition, a number of Akkadian-language texts are included in the RIM Early Series volumes, and these were removed as well, along with the Akkadian portions of any bilingual texts. This left a file containing only Sumerian texts in transliteration and their corresponding English translations.

This Word document was then exported to HTML, producing a file which could, in theory at least, be used as the input for an XML parser. However, the HTML generated by Microsoft Word 2004 is far from being XHTML-compliant, so it had to undergo further processing to make it into a valid XML document.

Fortunately for the purposes of building a corpus, the RIM project held to a very strict style-sheet, so the HTML exported from Word possesses a great deal of regularity. In particular, the text of interest is all contained within HTML `<table>` elements, with each `<tr>` (table row) element containing two `<td>` (table cell) elements: one with the transliterated Sumerian text, and the other with the English translation of that text.

Once in XML form, this HTML file was processed by a SAX-based parser which breaks the text into words and paragraphs. This stream of words is then passed to the following stages of the lemmatisation process to determine their part of speech, their lemma, their affixes, and their English gloss.

### 2.2.7 Regularising Transliterations

Because most of the RIM volumes were compiled more than ten years ago, they tend to follow an older tradition of transliteration than that used by the ePSD. The RIM uses acute and grave accents instead of subscripted sign indices <sub>2</sub> and <sub>3</sub>, and most determinative signs are transliterated in capital letters, separated from the word by a period. Thus, in an example with the  $\check{g}e\check{s}$  determinative (indicating an object made of wood), the RIM would have *GIŠ.gígir* ‘chariot’

---

date to the Old Babylonian period and were probably not written by native speakers of the language.

where the ePSD would have  $\tilde{g}e\check{s}$  *giġir*<sub>2</sub>. It was possible to implement most of these transformations within the SAX parser, but some manual editing was also required.

One place where the RIM reflects an older tradition is in the transliteration of the velar nasal / $\tilde{g}$ /. The RIM editors would often write *gá* for the sign which is now generally accepted to be  $\tilde{g}a_2$ .<sup>5</sup> As well, there are a number of cases where the RIM's reading of a vowel differs from that used in the ePSD, most notably transliterating *giš* or  $\tilde{g}i\check{s}$  where the ePSD would have  $\tilde{g}e\check{s}$ . These differences are dealt with by a lookup table which is stored as part of the lexicon appendix.

One unexpected difficulty with the RIM texts was that the editors had a tendency to concatenate together attributive expressions. This included not just adjectives, but genitives and even subordinate clauses, as shown in Table 2.4. Being unable to identify these as separate words, the lemmatiser was unable to process them. Because these expressions are so varied, there was no easy way to systematically identify them all, so they had to be separated manually into their component words. Although it seems like a minor problem, the manual nature of this task meant that it turned out to be the single most time-consuming stage of the entire lemmatisation process.

Table 2.4: Some concatenated attributive expressions from RIM

RIM Form	Expected Form
<i>amar-bànda-<sup>d</sup>en-líl-ka</i>	<i>amar</i> <i>banda</i> <sub>3</sub> <i><sup>d</sup>en-lil</i> <sub>2</sub> - <i>ka</i> calf      impetuous      Enlil-GEN 'impetuous calf of Enlil' (an epithet of Sîn)
<i>kur-gú-<math>\tilde{g}</math>ar-<math>\tilde{g}</math>ar-<sup>d</sup>nin-<math>\tilde{g}</math>ír-su-ka</i>	<i>kur</i> <i>gu</i> <sub>2</sub> <i><math>\tilde{g}</math>ar-<math>\tilde{g}</math>ar</i> <i><sup>d</sup>nin-<math>\tilde{g}</math>ir</i> <sub>2</sub> - <i>su-ka</i> foreign land      neck      to lay down      Ningirsu-GEN 'who conquers foreign lands for Ningirsu'

<sup>5</sup>In practice, the ePSD actually renders this as  $\hat{g}a_2$  rather than  $\tilde{g}a_2$ . This is probably a compromise necessitated by the absence of  $\tilde{g}$  from the Unicode standard.



## 2.2.8 Identifying Paragraphs

In order to parallel the structure of the ETCSL, the RIM texts had to be organised into <para> elements. As with the ETCSL, the RIM contains an indication of which lines of Sumerian text correspond to which lines of English translation, so it was a straightforward matter to match these up into “paragraphs”. In the case of processing the RIM, this division into <para> elements serves an extra purpose, because the English translation is being used to help disambiguate Sumerian words (§2.2.12). Thus, the English translation attached to the <para> element provides the scope over which disambiguation is applied.

Paragraphs in the ETCSL are identified based on complete sentences, so the same definition of a <para> element was used for processing the RIM texts. That is, a <para> element is identified as a segment of the English translation which ends in a period, an exclamation mark, or a question mark, or else with an indication of a break in the text (e.g “Lacuna” or “(broken)”).

## 2.2.9 Morphological Processing

The output of the first stage of the lemmatisation process is a stream of Sumerian words, organised into “paragraphs”. At this point, these words are purely orthographic strings, with no indication of lemma or part of speech. As well, these words are fully inflected, the verbs having suffixes and prefixes, and the nouns having suffixed clitics. Before being able to determine the root for an orthographic word, it was first necessary to strip off this extra morphology.

A list of prefixes and suffixes was assembled in order to recognise what strings needed to be stripped off. The initial list consisted of the prefixes and suffixes provided in Thomsen’s Sumerian grammar (Thomsen, 1984). Thomsen’s list is not comprehensive, since it does not attempt to account for verbal forms which include dimensional prefixes or subject prefixes. These prefixes can occur in a broad range of combinations whose orthography often interacts in unexpected ways. Identifying all the possibilities was an iterative process, running the lemmatiser repeatedly to determine which prefixes were unaccounted for, and adding these to the

lexicon appendix. In addition to the 132 prefix combinations identified by Thomsen, another 171 have been added to the lexicon appendix.

The situation with nominal affixes is simpler than that with verbs, but involves similar problems. Nominal morphology is exclusively suffixing, and the range of affix combinations is slightly narrower than what is found for verbs.

In the case of nouns, what have traditionally been called suffixes are better described as clitics attached to the noun phrase. One consequence of this is that the full range of nominal suffixes can also appear on an adjective or on the participial form of a verb, should that adjective or participle happen to occur at the end of a noun phrase. Another consequence is that the same clitic can potentially appear twice on the same noun, so a noun-phrase like “son of the ruler of Lagaš” would have two genitive-case markers on the word “Lagaš”.

### 2.2.10 “Amissible” Consonants

Suffixes pose an additional complication due to the nature of Sumerian phonology. When a suffix starts with a vowel, it is often written with a sign starting in a consonant, and the consonant which is written depends on the root to which the suffix is attached. So for instance, the locative case suffix *-a* will be written as *-da* after the noun *u<sub>4</sub>* ‘day’, as *-la* after the noun *ti* ‘life’, and as *-ga* after the word *ša<sub>3</sub>* ‘heart’. The regularity of these consonants has long been recognised as evidence that these words do end with the corresponding final consonant, but that Sumerian has a phonological rule which drops certain word-final consonants. These have traditionally been referred to as “amissible consonants”.

The challenge for the lemmatiser is to recognise that a suffixed *-a* can potentially be realised in a variety of different ways, depending on the stem to which it is being attached. Fortunately, when the ePSD records a lemma, it includes the final consonant. Thus *u<sub>4</sub>* will be listed in the ePSD as one of the orthographies of the stem *ud*, which is enough information to tell the lemmatiser that the *-da* on a form like *u<sub>4</sub>-da* could actually represent a locative case *-a*.

Due to one additional peculiarity of Sumerian phonology, there are a few forms whose

stems end in /d/, but which unexpectedly are written using suffixes starting with /r/. For example, when the verb *kud* ‘to cut’ is followed by a suffixed *-a*, the suffix will be written *-ra<sub>2</sub>*. This has been used as evidence that these stems end in an unknown phoneme which is neither /d/ nor /r/.<sup>6</sup> Whatever the phonological details, the lemmatiser needs to know which stems exhibit this sort of behaviour. Entries are added to the lexicon appendix to help the lemmatiser recognise that a sequence like *ku<sub>5</sub>-ra<sub>2</sub>* actually represents the verb *kud* with a suffixed *-a*.

### 2.2.11 Affix Stripping

Given a raw orthographic word from the first stage of the lemmatisation process, the second stage iterates through all the possible prefixes to find any which match the beginning of the word’s orthography. Whenever a match is found, the prefix is stripped from the orthography, and then any potential suffixes are checked in similar fashion.

At each step in this process, a check is made to see if the current stripped orthographic form corresponds to an entry in the lexicon. If it does, the lexeme is recorded as a possible match. At the end of this stage, the word has a collection of possible matching lexemes, along with the affixes which had to be stripped off in order to find the match.

If a matching lexeme happens to be the verbal portion of a compound verb, this provides an opportunity to look for the preceding nominal element. If the expected nominal element is found in the preceding two words, the combination is considered to be unambiguously identified as the compound verb, and the word is added to the corpus.

At this point, if there is only one possible matching lexeme, it is accepted as being the correct one, and the word is added to the corpus. If there is more than one matching lexeme, the word is passed on to the final stage where it is disambiguated using the English translation.

---

<sup>6</sup>Possibly an alveolar flap /r/, although this is disputed by Black (1990).

### 2.2.12 Disambiguating

One of the problems reported during the lemmatisation of the ETCSL was that some forms are inherently ambiguous, a good example being the form *mu-zu* (Ebeling and Cunningham, 2005). This could be the noun *mu* meaning ‘year’ or ‘name’ with a second person possessive suffix, but it could equally well be the verb *zu* ‘to know’ with the prefix *mu*.<sup>7</sup> While ‘your year’ is an unlikely reading for *mu-zu*, both ‘your name’ and ‘he knew’ are quite plausible. In the absence of more information, there is no way of deciding between the two readings; this is a serious obstacle for the primary task of tagging the parts of speech, since this form is ambiguously either a noun or a verb.

After the conclusion of the morphological processing stage, each word has been narrowed down to a small number of possible lexemes. Each of the possible lexemes has a set of English glosses from the ePSD, supplemented by additional glosses from the lexicon appendix. The English translation of the paragraph being processed is also available, and it is a simple matter to search the translation looking for a match. If an exact match is found, the corresponding lexeme is considered to be the correct one, and the word is added to the corpus.

Currently the processing considers only stems. So in the previous example of *mu-zu*, it would be sufficient to find a match in the translation for ‘name’, rather than ‘your name’. This appears to be adequate for the purposes of assigning parts of speech, and the additional complexity of trying to find matches for affixes is difficult to justify.

It soon became apparent that relying solely on the glosses from the ePSD was not adequate. For example, *-a* is a valid suffix on both nouns and verbs, so the form *sum-ma* could be a form of the verb meaning ‘to give’ or of the noun meaning ‘garlic’.<sup>8</sup> However, the English translation will seldom explicitly contain ‘to give’ or even ‘give’, but more often will employ a past-tense form like ‘gave’, a near-synonym like ‘granted’ or ‘endowed’, or even an unrelated verb like

---

<sup>7</sup>This prefix is called “ventive” by Ebeling and Cunningham (2005), but is more likely part of the Sumerian system of voice prefixes (Woods, 2008).

<sup>8</sup>Given that the texts are royal inscriptions one might expect ‘to give’ to be a much more plausible reading than ‘garlic’, but in fact ‘garlic’ is attested five times in the Royal Inscriptions of Mesopotamia.

‘to set up’. To allow these translations to be recognised as a correct match for the verb ‘to give’, appropriate entries are added to the lexicon appendix to serve as secondary glosses.

There are occasions when a false match is made, largely for words which have very short English glosses. So for instance, in one of Gudea’s statue inscriptions the lemmatiser misidentified  $ni_2\text{-}\tilde{g}al_2$  as a prefixed form of the verb  $\tilde{g}al$  ‘to be’. This was done on the strength of a match between “be” and the first two letters of the word “best” in the English translation. The lemmatiser missed properly identifying the adjective  $ni_2\text{-}\tilde{g}al_2$  because the ePSD glosses  $ni_2\text{-}\tilde{g}al_2$  as “awe-inspiring” while the RIM translates it as “unpleasant to look at”.

There are also a handful of words which are problematic because they often fail to appear within the English translation. For example, *a-ba* functions as an interrogative or relative pronoun meaning “who”. However, in a translation like the one in (2.4), “who” never actually appears in the English. The word *a-ba* is especially problematic because *a-* happens to be a known verbal prefix while *-ba* is a nominal suffix. If there is no “who” in the translation, this results in *a-ba* matching eight different lexemes: *aba* ‘who’, *aya* ‘father’, *a* ‘house’, *a* ‘a cry of woe’, *a* ‘water’, *ba* ‘to distribute’, *bad* ‘to open’, and *be* ‘to cut off’.

(2.4) Translation with *a-ba* without “who”

$\langle^d en\text{-}lil_2 \quad lugal\text{-}mu\text{-}ra \quad a\text{-}ba \quad du_{11}\text{-}ga\text{-}na \quad a\text{-}ba \quad \check{s}ar_2\text{-}ra\text{-}na \rangle$   
 Enlil    master=1SG.POSS=DAT    who    say-PTCP    who    reiterate-PTCP  
 ‘After what he has declared and has reiterated to my master the god Enlil,’

All of this underscores the fact that, despite the best efforts of an automatic lemmatiser, a certain amount of human input is also involved. Errors like *a-ba* are easy to identify and can be fixed manually. Misclassification errors of the sort exemplified by  $ni_2\text{-}\tilde{g}al_2$  are more insidious because there is every indication that the lemmatiser has successfully identified the lexeme.

## 2.3 Additional Corpora

At this point, the only two corpora supported are the ETCSL and the RIM corpus. This is entirely a consequence of limited resources and limited time. It has always been intended that

it should be easy to apply the LPattern tools to other corpora with a modicum of extra effort.

In fact, the original intention had been to use LPattern with both a Sumerian-language corpus and with an Elamite-language corpus, in order to establish that the methodology was not specific to one language. Preliminary work was done in preprocessing the Electronic Corpus of Elamite Texts, a small corpus developed as part of earlier research on Elamite syntax (Smith, 2006b). However, it was decided that this effort would be a distraction, and it has been shelved for the time being.

While the methodology described in §2.2 was developed with particular reference to the Royal Inscriptions of Mesopotamia, much of the same methodology could also be applied to the much larger CDLI corpus. The CDLI is by far the largest single corpus of Sumerian-language texts, but the corpus lacks part-of-speech information and English translations. For those reasons, the CDLI corpus was ignored in favour of the ETCSL and the RIM.

However, the technique of using the ePSD to guide the lemmatisation process is hardly specific to the RIM. The one major difficulty posed by the CDLI is that there are no accompanying English translations. This would eliminate the lemmatiser's ability to disambiguate word forms using the English translation (as described in §2.2.12). This would not make the task of lemmatisation impossible, but it would certainly require considerably more human intervention than was needed for the RIM.


While the effort involved in bringing the CDLI texts into the same corpus as the ETCSL and the RIM would be significant, the benefits would be commensurate. Sumerologists and linguists would have a single corpus with the broadest possible coverage of Sumerian texts, annotated with morphological information contributed by the corpus's own users, and searchable using the LPattern tools described in the next chapter.

## Chapter 3

# Query Infrastructure

The particular motivation for this part of the research came from needs which arose during earlier research into Sumerian phonology (Smith, 2006a,b). Existing corpora are often ill-suited for the linguist who is attempting to locate particular linguistic phenomena.

In corpora of languages which, like Sumerian, are written using a cuneiform script, searching is made more difficult by the fact that the cuneiform system permits a considerable amount of orthographic variation. A particular morpheme may be represented in the corpus in a variety of ways, sometimes due to morphophonological processes within the language, but sometimes due to peculiarities of the scribal tradition.

Furthermore, the transcription itself is subject to variation beyond that found in the text itself. Different scholars may transcribe the same cuneiform text in quite different ways. For instance, when transcribing the  $\check{s}e_3$   sign in Sumerian, the Electronic Pennsylvania Sumerian Dictionary (Sjöberg et al., 2004) would use  $\check{s}e_3$ , the Electronic Text Corpus of Sumerian Literature (Black et al., 1998–2006) would use  $ce_3$  (because  $c$  is easier to type than  $\check{s}$ ), the Royal Inscriptions of Mesopotamia (Frayne, 2005) would use  $\beta\grave{e}$  (rendered in a custom font to make  $\beta$  look like  $\check{s}$ )<sup>1</sup>, and the Cuneiform Digital Library Initiative (Englund and Damerow, 2000–2005) would use  $sz\acute{e}_3$  (to ensure that words with  $sz$  alphabetise between  $s$  and  $t$ ).

---

<sup>1</sup>In Assyriology, using a grave accent is another (more traditional) way of referring to the third grapheme with a particular reading. Thus,  $\grave{e} = e_3$  and  $\acute{e} = e_2$ .

In addition to these largely cosmetic differences in transcription, there are also more fundamental differences in the choice of transcriptions to represent a given character. For instance, Sumerian has a regular phonological process that deletes word-final stops. The ePSD, being concerned with the canonical forms of lexical items, might record  $\langle \text{𒄀} \rangle$  (the adjective meaning ‘good’) as *kug*. The RIM, which is more traditionalist in its transcriptions and not concerned with canonical forms, would record the same word as *kù*, reflecting the way it is believed to have been pronounced.

My earlier work on vowel harmony in Sumerian (Smith, 2007b) involved studying the allomorphs of the conjugation prefix *i-*. Although the focus of that paper was phonology rather than syntax, many of the problems which arose there also face the syntactician. A large portion of the hours spent working on the paper involved searching through corpora of Sumerian texts, such as the CDLI and RIM, in order to locate particular instances of the conjugation prefix *i-* and its [–ATR] allomorph *e-*. Or more precisely, searching for the written forms  $i_3$   $\langle \text{𒄀} \rangle$  and  $e$   $\langle \text{𒄀} \rangle$ . Fortunately, the conjugation prefix generally occurs word-initially and not sentence-initially so it was possible to search the corpora for strings with a leading space, such as “ $\langle \text{𒄀} \rangle$ ”, “ $\langle \text{𒄀} \rangle_3$ ”, or “ $\langle \text{𒄀} \rangle$ ”, which yielded a large number of incorrect hits, but not so many as to make it completely impractical to filter the results manually.

Tedious as it was to locate instances of the *i-* prefix, it was at least feasible to perform the necessary queries manually. It is far more difficult to put together queries to locate instances of a relation between non-contiguous elements, such as the agreement relations described in more detail in §4.

Briefly, the Sumerian verbal stem is prefixed with a variety of morphemes (typically referred to as “dimensional infixes”) which may indicate agreement with oblique arguments of the verb. Thus for instance the verb will have an allative prefix *še-* (written as  $\langle \text{𒄀} \rangle_3$   $\langle \text{𒄀} \rangle$  or  $\langle \text{𒄀} \rangle$ ) to indicate agreement with a preceding noun in the allative case. These dimensional prefixes occur between the conjugation prefixes and the verbal root, so they do not occur word-initially. An attempt to search one of the existing corpora for a string like “ $\langle \text{𒄀} \rangle$ ” would result in a huge



number of hits, the vast majority of which would have nothing to do with the allative prefix *ši-*.

Even if a linguist is able to winnow the thousands of occurrences of the *še<sub>3</sub>* and *ši* graphemes down to those which actually represented an allative prefix, there would still be the daunting task of trying to match those occurrences up with the corresponding allative-case nouns. Thus, a syntactician who was attempting to use a corpus to establish the syntactic rules governing allative-case agreement would have a difficult time doing so with any existing corpus. Linguists who wish to explore this aspect of Sumerian syntax must currently rely on the painstaking work done by Gragg (1973).

### 3.1 Requirements

The intention was to implement a query language which provides a sufficient level of access to the syntactic information buried within the corpus, without being so complex that it would be unappealing to the Sumerologists and linguists who are its intended audience. The requirements for querying our corpora are fourfold: simplicity, pattern matching, morphological awareness, and an embeddable implementation.

The first requirement is that the query language must be simple. A query language such as `tgrep` (Rohde, 2005) is certainly powerful enough to perform any of the queries which a Sumerologist could conceivably be interested in. However the syntax of `tgrep` is sufficiently recondite that actual Sumerologists would be extremely unlikely to use such a tool.

The second requirement is that the queries should be pattern-based. Since the corpus is largely lacking in hierarchical structure (as described in Chapter 2), most queries are likely to be looking for a particular linear arrangement of tokens. So whatever the capabilities of the language, it should be easy to perform queries for sequences or patterns of tokens.

And thirdly, it must be possible to make queries which look for particular morphemes within the corpus. As much as possible, the query language should insulate the user from having to cope with the details of Sumerian orthography. That being said, it should still be

possible to make queries against orthographic strings as well. Some Sumerologists might have their own notions of morphological markup or lemmatisation, and they should still be able to use the corpus to perform queries directly against the language's orthography.

As part of this third requirement, it should also be possible to add morphological information to the corpus. As queries are made, the results of the query can be used to add additional morphological markup to the corpus. This approach has a number of benefits. First, by allowing the markup to be created as a side-effect of linguists doing their own research, it allows the corpus to acquire morphological markup without relying on the services of dedicated annotator. Second, it allows linguists to contribute to the annotation of the corpus based upon their own areas of expertise. This ongoing process of query-based annotation (Smith, 2007a, 2008) is discussed in greater detail in §3.5.

The fourth requirement, embeddability, arose from the expectation that query-based annotation would require a user interface for analysing query results. Thus the query language would have to be able to be embedded within the user interface of a larger querying/concordancing program. In the initial selection of query language, this requirement was given the most weight, the thought being that any existing successful query language would at least adequately meet the other three requirements.

Rather than inventing a query language entirely from scratch, it would be more efficient to build the query infrastructure around an existing query language implementation. The survey of query languages provided by Lai and Bird (2004) proved to be a useful starting point. However, Lai and Bird (2004) are concerned only with query languages for treebanks, so much of their argumentation was not strictly relevant to the selection of a query-language for working with a non-hierarchical corpus. Nonetheless, drawing on the discussion in Lai and Bird (2004), LPath was selected as the query language to be used for working with the corpus. One strong argument in its favour was the existence of a Python-language implementation of LPath. This meant that LPath could easily be embedded within a user interface which would support query-based annotation.

## 3.2 LPath and LPath<sup>+</sup>

One of the goals of the survey by Lai and Bird (2004) is to present the LPath language developed by Steven Bird and his colleagues at the University of Pennsylvania (Bird et al., 2005, 2006; Lai and Bird, 2010). Bird’s work with query languages started with the investigation of query languages for annotation graphs (Bird et al., 2000). In the past few years, he has turned to tree-structured data, and enhancements to a standard XML search language called XPath (Clark and DeRose, 1999). The XPath language is intended for locating nodes within tree-structured XML documents, so it is a natural match for the task of locating elements within tree-structured linguistic data.

LPath extends XPath by adding a variety of search operators which are intended to be useful for the kinds of searches done in linguistics. These are shown in Table 3.1.

Table 3.1: LPath operators added to XPath (Lai and Bird, 2005)

->	immediate-following
<-	immediate-preceding
=>	immediate-following-sibling
<=	immediate-preceding-sibling
^	left-edge alignment
\$	right-edge alignment
{	subtree-scoping
}	

Some sample LPath queries are shown in Table 3.2. The first one searches for a sentence, S, and that sentence must contain some entity (indicated by the underscore) which has a *lex* attribute with the value of “saw”. This query would return all sentences containing any form of the word “see”. The second query is straightforward, locating nouns which follow a verb which is itself the child of a verb-phrase. However, the second query might not be what a linguist was looking for, since it will find matches where the verb and the noun lie in separate verb phrases; the third query gives an example of the braces used to restrict the scope of a search to a subtree, thus returning only verb-noun sequences which are contained within the

same verb phrase. The fourth through sixth demonstrate how the  $\wedge$  and  $\$$  edge-alignment operators can be used to search for particular structural configurations.

Table 3.2: Example LPath queries (Lai and Bird, 2005)

LPath	Explanation
//S[//_[@lex=saw]]	A sentence containing the word “saw”.
//VP/V-->N	Nouns that follow a verb which is a child of a VP.
//VP{V-->N}	Within a verb phrase, nouns that follow a verb which is a child of the given verb phrase.
//VP{/NP\$}	Noun phrases which are the rightmost child of a VP.
//VP{//NP\$}	NPs which are rightmost descendants of a VP.
//VP[{//^V->NP->PP\$}]	Verb phrases composed of a verb, a noun phrase, and a prepositional phrase.

### 3.2.1 Reference Implementation

A reference implementation of LPath written in Python is provided as part of the Natural Language Toolkit (NLTK) (Bird et al., 2001-2007), an open-source collection of Python-language tools for computational linguists. Since Steven Bird is involved with both the NLTK project and with LPath, the NLTK is an appropriate place for LPath to be made publicly available. That being said, LPath is not considered to be an integral part of NLTK, being relegated to the `nltk_contrib` branch of the NLTK distribution.

The NLTK’s LPath implementation stores the corpus in an SQL database, using Python’s built-in database classes to make the code largely independent of which specific database back-end is used. The reference implementation has support for Postgres and Oracle databases. As part of the process of testing LPath with the ETCSL corpus, MySQL support was added to LPath. The choice of MySQL as a back-end was motivated entirely by the fact that MySQL is free software. For linguists working on desktop computers, the ability to use LPath with a MySQL back-end is a significant benefit.

The LPath distribution provides a number of sample programs which provide a starting point for working with LPath. One program, `tb2tbl`, shows how to import a file in Penn Tree-

bank format into the SQL tables used by LPath, and it was straightforward to modify `tb2tbl` to import the XML-based format used by the ETCSL. In addition, the LPath distribution includes a demonstration program called `qba`, which uses a graphical “Query By Annotation” model (Bird and Lee, 2007) to let users assemble an LPath query string by pointing and clicking. The `qba` program also serves as an example of how other programs can interface with the NLTK’s LPath parser, and it served as the basis for writing an LPath query interface to the ETCSL.

Although LPath is intended as an extension of XPath, this is not strictly true of the NLTK’s LPath implementation. In particular, XPath includes a large number of built-in utility functions for string operations, type conversions, and other operations. The LPath implementation lacks these functions, which is unfortunate since some of the basic functions (e.g. `substring`) would have been very useful in certain queries against the ETCSL corpus. Fortunately, the LPath implementation does include undocumented support for wild-card access using the SQL `like` operator, which provides a stand-in for some of the missing XPath string functions.

As well, the LPath implementation supplied as part of NLTK has major issues when it comes to performance. A query like `//N[@orth like "%-ce3"]`, intended to locate all allative-case nouns, executes in a reported time of 0.01s. However, a structurally equivalent query, `//N[@ALL="+"]`, takes upwards of 10 minutes to execute. Examining the underlying SQL queries generated by the two LPath queries indicates that they are in fact equivalent, so it is unclear why the performance suffers in the second case.

### 3.2.2 LPath<sup>+</sup>

In addition to the original LPath language described by Bird et al. (2005), the NLTK implementation also includes undocumented support for a dialect called LPath<sup>+</sup>. LPath<sup>+</sup> is described in a paper by Lai and Bird (2005), and extends LPath by adding a Kleene star operator to express closures in the language. For example, assuming that an `ALL="+` attribute has been used to mark corpus words which have allative-case morphology, a query such as `//para[/N[@ALL] (=>N)*=>V[@ALL]]` would be necessary to find all allative-case nouns fol-

lowed by any number of other nouns, followed by a verb with an allative-case agreement prefix.

### 3.2.3 Limitations

The original intent when starting work with the ETCSL had been to use LPath<sup>+</sup> as the query language. However, after working with the NLTK's LPath<sup>+</sup> implementation for several months, it became clear that the tool was turning out to be more of a hindrance than a help.

Typical queries performed against the ETCSL involved searching for a linear arrangement of tagged words scoped within a single <para> element.<sup>2</sup> A very simple and very common query, such as one which looks for matches between an allative-case noun and a verb with an allative agreement marker requires a query such as `//para/N[@ALL]->V[@ALL]`, which seems rather complex for such a simple query. The fact that such a basic query required such a verbose query string strongly suggested that LPath would fail to meet the goal of providing a query language suitable for Sumerologists.

As well, the LPath implementation's lack of full support for XPath's built-in functions proved to be more of a limitation than first expected. With a bit of ingenuity, LPath's undocumented `like` operator is able to fill in for missing XPath string functions like `contains`, `starts-with`, and `ends-with`. However, an expression like (3.1), which is intended to perform some of the genitive-case queries described below in (3.3), is far beyond the capabilities of the `like` operator.

#### (3.1) XPath query to locate genitive-case suffixes

```
N[substring(@suffix,2,2)=concat(substring(@stem,string-length(@stem),1),"a")]
```

Outside of string processing, there are many other useful XPath functions which are simply unavailable in the LPath implementation. For instance, it is often helpful to know whether a given part-of-speech is absent from a particular context. In XPath, the built-in `count` function can be employed for this purpose, in a query like `//V[count(preceding-sibling::N)=0]`,

---

<sup>2</sup>The use of <para> elements in the corpus is described in §2.1.1.

which finds all verbs which are not preceded by a noun. In the reference LPath implementation, such a query would simply be impossible to express.

Finally, the unpredictable performance characteristics of the NLTK's LPath implementation made it unsuitable for productive work. It is quite possible that some of these performance issues derived from the choice of MySQL as the database back-end. However, these issues were impossible to resolve without delving deep into the details of LPath's SQL-generation code and database schema. Storing the corpus inside a database also had one significant disadvantage, namely that (short of writing one's own custom SQL queries) there was no easy way to browse the raw data in order to determine why a particular LPath query was not working as expected.

After working with LPath for a considerable amount of time, it was apparent that these limitations were seriously hindering progress on the actual syntactic research. Despite the appeal of using an existing query language implementation, it was clear that it would be more productive to write a new query language which was better suited to the task. The CQL (Corpus Query Language) of the British National Corpus seemed to provide the desired level of simplicity, so it was chosen as the model for the new language.

### 3.3 CQL

One query language designed for a non-hierarchical corpus is CQL (Corpus Query Language), a simple query language provided by the British National Corpus to serve as an interface to their SARA (SGML Aware Retrieval Application) software (Dodd, 2005). The syntax is quite straightforward, as can be seen from the examples in Table 3.3.

In recent editions of the BNC, the corpus's content has transitioned from SGML to XML. As part of that transition, SARA has become Xaira (XML Aware Indexing and Retrieval Architecture), and a new XXQ (Xaira XML Query) language has been introduced (Dodd, 2006). For instance, the first CQL query from Table 3.3 would be expressed in XXQ as shown in (3.2). Clearly, XXQ sacrifices the simplicity of CQL, without adding any additional expressiveness,

Table 3.3: Example CQL queries (Dodd, 2005)

CQL	Explanation
cat _ dog	Find three-word phrases of which the first word is “cat” and the last is “dog”.
!cat dog	Find occurrences of “dog” not preceded by “cat” within the same document.
cat*dog	Find occurrences of “cat” followed anywhere within the same document by “dog”.
cat#dog	Find occurrences of “cat” followed or preceded by “dog” anywhere within the same document.
cat*dog/10	Find occurrences of “cat” followed by “dog” within ten words.
cat*dog/	Find occurrences of “cat” followed by “dog” within a single <head> element.

making XXQ too cumbersome to be practical as an end-user query language.

(3.2) XXQ equivalent of CQL query `cat _ dog` (Dodd, 2005)

```
<seq>
  <lemma>cat</lemma>
  <gap/>
  <lemma>dog</lemma>
</seq>
```

### 3.4 LPattern

In view of the limitations of LPath described in §3.2.3, it was finally decided to create a new query language specifically designed for the task at hand. The goal was to create a language with the simplicity of CQL, but which would still be powerful enough to make any useful query that would be possible using LPath. Since the language’s focus was to be on locating patterns of tokens, it was named LPattern.

An efficient XPath implementation is available as part of Nokia’s Qt toolkit, and this provided the basis for LPattern’s query implementation. In effect, LPattern provides just a thin



veneer which simplifies commonly-used XPath queries. The use of XPath as an underlying query implementation has the advantage of providing access to all the built-in functions which are available in XPath.

A number of simplifying assumptions made the implementation of LPattern somewhat more manageable. Since the corpus is being designed to explore syntactic phenomena, it seemed reasonable to restrict the scope of all queries to the sentence level. In practice, this means that LPattern only looks for sequences which are fully contained within a given `<para>` element. However, if a technically-sophisticated user is interested in exploring cross-sentential phenomena, it is possible to specify XPath queries directly, bypassing LPattern completely.

Although LPattern is intended for syntactic research, there is no reason why it would not also be useful for a linguist studying other aspects of the language. For instance, while the original intent of the query-based annotation feature (§3.5) is for adding morphological markup, it could equally well be used for adding phonological information to the corpus. Such a facility would have been invaluable for the vowel harmony study described by Smith (2006a).

Likewise, there is nothing which ties LPattern specifically to Sumerian. Preliminary work was carried out using LPattern to query the Electronic Corpus of Elamite Texts, a corpus which was assembled for research into Elamite syntax (Smith, 2006b). In practice, LPattern could be applied against any XML-based corpus for which hierarchical structure is not relevant. All that would be required to support another corpus would be a small amount of preprocessing, similar to that described in §2.1.1.

Given that simplicity is one of the main goals of LPattern, the query language should do the simplest reasonable thing wherever possible. If a user types in just the string `su-a` without quotes, the interface should be smart enough to expand it to the query `"su-a"`, which is surely what the user intended.

In the same vein, when matching string literals, the software attempts to anticipate the needs of working Sumerologists. So a string like `"su-a"` will look for matches in both the orthographic text and lemmata of words. Similarly, LPattern recognises that a query string

starting with a hyphen, like "-še3", is intended to locate suffixes, so it will look for the text in the `suffix` attribute of words.

Just as often, a query will be interested in the relation between particular parts of speech, rather than the particular words involved. Since part-of-speech queries are a large part of the work for which LPattern is being designed, they are given prominence. The most basic LPattern query is one such as `N`, which will simply find all nouns.

The operators defined in the LPattern language are inspired by CQL. The actual implementation is built on top of Qt's XPath module. A preprocessor written using `flex` and `bison` expands LPattern queries into the equivalent XPath queries, as shown in Table 3.4. In addition, the use of XPath as the underlying query language provides access to all of XPath's built-in functions. So for instance, one of the complex genitive-case queries discussed below in (3.3) can be expressed as `N"-ba"[ends-with(@stem, "b")]`, combining LPattern syntax with an XPath predicate expression.

Table 3.4: LPattern operators

LPattern	Explanation (with XPath expansion)
<code>N _ V</code>	Find three-word phrases of which the first word is a noun and the last word is a verb. <code>//N/following-sibling::*[1]/self::*[1]/following-sibling::*[1]/self::V</code>
<code>!N V</code>	Find occurrences of a verb not preceded by any nouns within the same <code>&lt;para&gt;</code> . <code>//V[count(preceding-sibling::N)=0]</code>
<code>N*V</code>	Find occurrences of a noun followed anywhere within the same <code>&lt;para&gt;</code> by a verb. <code>//N/following-sibling::V</code>
<code>su-a</code> or <code>"su-a"</code>	Find all occurrences of "su-a" in the corpus. <code>//*[contains(@orth, "su-a") or contains(@lemma, "su-a")]</code>
<code>N"su-a"</code>	Find all occurrences of the noun "su-a" in the corpus. <code>//N[contains(@orth, "su-a") or contains(@lemma, "su-a")]</code>
<code>N"-ra"</code>	Find all nouns suffixed with <i>-ra</i> . <code>//N[contains(@suffix, "-ra")]</code>
<code>V"ra-"</code>	Find all verbs prefixed with <i>ra-</i> . <code>//V[contains(@prefix, "ra-")]</code>
<code>V-DAT.2SG</code>	Find all verbs which have been marked as having a 2nd person singular dative prefix. <code>//V[@DAT.2SG]</code>

### 3.5 Query-based Annotation

As mentioned earlier, there is a significant mismatch between the morphology of Sumerian and the orthography. Due to the lack of morphological annotation in the ETCSL and RIM corpora which provide the basis for the working corpus, all queries must be made against the language’s orthography. In many cases, a desired query for a simple piece of morphology can require a rather convoluted query (or series of queries) when referring to the language’s orthography.

A case in point is a query for a genitive-case noun. The genitive-case suffix is *-ak*, but it is never actually written as  $\text{𒀭𒀗}$  ⟨ak⟩. Instead, it manifests itself using orthographic rules such as the ones shown in (3.3).

#### (3.3) Orthography of genitive case suffix *-ak*

- Word-final vowel assimilates to /a/ (e.g. written  $\text{𒀭𒀗}$  ⟨g̃a<sub>2</sub>⟩ after words ending in /g̃u/).
- Sometimes written as  $\text{𒀭}$  ⟨a⟩.
- Only reflects the /k/ before another suffix (e.g.  $\text{𒊕𒀭𒀗𒀭𒀗}$   $\text{𒀭𒀗}$   $\text{𒀭𒀗}$  ⟨lugal-la-ke<sub>4</sub>⟩ ‘king-GEN-ERG’).
- $\text{𒀭𒀗}$  ⟨ba⟩ after stems ending in /b/.
- $\text{𒀭𒀗}$  ⟨da⟩ after stems ending in /d/.
- *similar queries for other stem-final consonants...*
- $\text{𒀭𒀗}$  ⟨za⟩ after stems ending in /z/.
- $\text{𒀭𒀗}$  ⟨ra<sub>2</sub>⟩ after certain stems ending in /r/.
- $\text{𒀭𒀗}$  ⟨la<sub>2</sub>⟩ after certain stems ending in /l/.

In practice, searching for a genitive-case noun is likely to be a fairly common operation. Hence, it would be useful to allow direct querying for the genitive-case morphology, instead of having each time to perform a series of queries against the language’s orthography. To this end, the LPattern software allows the user to define “query objects”, which provide shorthand for referring to the results of a particular query. Internally, query objects are stored as attributes attached to word-level XML nodes, as described in §??.

In the example of the genitive case noun suffix, a user might want to define a query object called N-GEN. As the user performs the various queries listed in (3.3), the definition of N-GEN is built up. At each step, the user can look at the results of the query to verify that it is returning the expected results; if so, the results can be added to the definition of N-GEN. When all the relevant

queries have been made, the corpus has effectively been annotated to identify all genitive-case nouns. From that point on, N-GEN acts as a first-class member of the corpus, and can in turn be used as the basis for other queries.

In a similar fashion, query objects representing phrases can be added to the corpus. An NP object can be built up by starting with a query such as N, since all nouns are inherently noun phrases. A noun (or noun phrase) modified by an adjective is also itself a noun phrase, so NP ADJ can be added to the definition of the NP query object. If an earlier set of queries has identified verbs with the subordinating suffix *-a* as a query object V-SUB, then the query NP V-SUB can also be added to the definition of the NP object.

By taking this approach, annotation can be added to the corpus incrementally. As a side-effect of performing the queries which happen to be necessary for their own work, users define the query objects which are relevant to them. These query objects are then available to subsequent users of the corpus. The hope is that gradually the corpus will build up its annotation without the need for any one person to explicitly devote themselves to the task of annotation. Instead, the annotation will be a cooperative effort by all the scholars using the corpus.

One of the advantages of this process of annotation is that it avoids the need to impose a single unified model of Sumerian morphological markup. Given that there is considerable disagreement over so many aspects of Sumerian morphology, it is impossible to expect that any single model would be acceptable to all Sumerologists. With the query-based annotation approach, if a user has a theoretical disagreements regarding the query objects which have already been defined then they are free to define their own objects. For instance, a linguist might have their own view about the conjugation prefixes which conflicts with the query objects defined in Appendix A. In such a case the user could simply ignore the existing query objects and define new objects.

Of course, the usefulness of such a cooperatively-constructed corpus would be greatly increased if it were available over the Internet. If work on LPattern is to proceed beyond the completion of this dissertation, making the corpus Internet-accessible would be one of the

most useful enhancements. Making a static version of the corpus available would be a simple first step, but it would be far more useful if users were able to contribute their own annotations over the Internet. Creating a collaborative version of the corpus would of course raise new issues such as concurrency control and the need to attribute changes to particular users. However, these are problems which have already been solved for software such as wikis, and there is no reason to believe that these issues would provide a significant technical obstacle.

### 3.5.1 Context within Annotation Science

In recent years, the field of linguistic annotation has been receiving increased attention, with the creation of the SIGANN special interest group within the Association for Computational Linguistics, and with a series of annual Linguistic Annotation Workshops held in conjunction with ACL meetings. Large parts of the annotation process have been successfully automated, most notably the task of part-of-speech tagging. Linguistic annotation is even tentatively being described as a “science” (Hovy, 2006). Nonetheless, significant parts of the annotation process continue to be carried out in the same sorts of labour-intensive manual processes which have been prevalent since the earliest days of electronic linguistic corpora (Leech, 2005).

There are many existing annotation-support tools which are intended to expedite the process of manual annotation, so query-based annotation does not represent a formal change in the annotation process. However, it does blur the stages in the functional taxonomy described by McEnery and Rayson (1997), and shown in Table 3.5. In particular, the workflow of query-based annotation feeds the output of stages 3(g) and 3(j) back into stages 1(b) and 2(d).

In the corpus linguistics literature, the closest approximation to query-based annotation is the approach proposed by Smith et al. (2008). They stress the desirability of allowing linguists working with a corpus to add their own annotations to the results returned by concordancing (stage 3(g) in Table 3.5). They refer to this as a “top-down” approach to annotation, as contrasted to the “bottom-up” approach which characterises traditional manual corpus annotation techniques.

Table 3.5: Functions of corpus/annotation tools (McEnery and Rayson, 1997)

1. Corpus development (*the input of annotated information into a corpus*):
  - (a) Text encoding
  - (b) Annotation
  - (c) Encoding of annotation
2. Corpus editing (*changing annotation information in a corpus*):
  - (d) Correction (including correction of annotations)
  - (e) Disambiguation of annotations
  - (f) Conversion/transduction of annotations
3. Extraction of information (*the output of annotation information from a corpus, whether raw or annotated*):
  - (g) Concordancing
  - (h) Frequency analysis
  - (i) Input to lexicons, grammars, etc.
  - (j) Information retrieval
  - (k) Bilingual/multilingual variants of (g)-(j)

Their paper includes a review of existing work on the area of annotating concordancing results, with the conclusion that the current state of affairs is unsatisfactory. In cases where a corpus tool itself allows for annotation of concordancing results, such support tends to be an afterthought with limited functionality. The alternative is to export the concordancing results to an external database or spreadsheet for subsequent annotation. This approach is more flexible, but has serious drawbacks, most notably the severing of the connection between the exported results and the original corpus data.

Recognising that nothing currently satisfies their requirements, they envisage three possible solutions to this problem: 1) the corpus tool and the external database share the same data

source, 2) manual annotations added in the external database are re-imported into the corpus tool, and 3) a common reference system between the two tools to maintain the correspondence between the exported data and the corpus data. Query-based annotation is effectively an implementation of the first of these solutions, since the annotations created by analysing concordance results are stored into the corpus itself. It also goes somewhat beyond this vision, since query-based annotation is intended to be the primary mode of annotation, rather than simply an adjunct for allowing end-users to add additional annotations to the corpus.

### 3.5.2 Implementation of Query Objects

The storage of query-based annotations is deliberately made as simple as possible. When creating a new morpheme gloss, the new morpheme is stored simply as an attribute of the part-of-speech node. This is shown in the first example in Table 3.6.

Table 3.6: Storage of query objects

Query Object	Sample XML representation
V-DAT.3SG	<code>&lt;V orth="mu-na-sum" prefix="mu-na-" lemma="šum2" english="gave" DAT.3SG="+"/&gt;</code>
NP	<code>&lt;N orth="nam-lugal" lemma="namlugal" english="kingship" START_NP="2" END_NP="1"&gt; &lt;N orth="lagaš[ki]" lemma="Lagaš" english="Lagaš" GEN="+" START_NP="1" END_NP="2"&gt;</code>

The screen shot shown in Figure 3.1 shows the process of defining the V-DAT.3SG object. A query has just located all verb instances which have the sequence “mu-na-” in their prefix. The linguist can then filter the results to verify that these are all 3rd person singular datives, removing the checkmark next to any which fail to qualify. By pressing the *Define* button, the DAT.3SG attribute is added to all the checkmarked verbs in the query’s result set (and removed from all the unchecked verbs). The *Toggle* button inverts the state of all the checkmarks, making it easy to use a particular query to subtract annotations which were added by an earlier

query. The *Export* button exports the checkmarked query results as a tab-limited text file which can be imported into an external application such as Microsoft Excel or Open Office.

Figure 3.1: Defining the V-DAT.3SG object

The screenshot shows the 'LPattern/XPath Query' interface. At the top, the query is 'V"mu-na-"' and the context is '3 words'. Below this is a 'Query Results' section with a table of results. Each row has a checkbox, a reference ID, and a match string. The 'Define Construct' section at the bottom has a 'Name' field containing 'V-DAT.3SG' and a 'Define' button. At the very bottom, it says '1782 hits (query took 4602ms)'.

Reference	Match
<input checked="" type="checkbox"/> E1.1.7.2.1	ensi2-ĠAR adab(ki) e2-mah mu-na-du3 ur2-bi ki-še3 temen
<input checked="" type="checkbox"/> E1.1.9.2001.3	[d]nin-šubur diġir-ra-ni a mu-na-ru
<input checked="" type="checkbox"/> E1.9.1.17.1	[d]nanše sangax a mu-na-A+KU4 eš3-ir mu-tu ur-nimin
<input checked="" type="checkbox"/> E1.9.3.1.9	[d]nin-hur-saġ-ke4 ubur zi-da-ne2 mu-na-la2
<input checked="" type="checkbox"/> E1.9.3.1.12	nam-ga-hul2-da nam-lugal lagaš(ki) mu-na-sum
<input checked="" type="checkbox"/> E1.9.3.1.21	na2-a-ra na2-a-ra saġ-ġa2 mu-na-gub e2-an-na-tum2 na2-a-ra lugal
<input checked="" type="checkbox"/> E1.9.3.1.21	ki-aġ2-ni [d]nin-ġir2-su2 saġ-ġa2 mu-na-gub
<input checked="" type="checkbox"/> E1.9.3.1.47	[ġeš]KUŠU2(ki)-ke4 e2-an-na-tum2-ra nam mu-na-ku5-de6 zi [d]en-lil2 lugal
<input checked="" type="checkbox"/> E1.9.3.1.57	[ġeš]KUŠU2(ki)-ke4 e2-an-na-tum2-ra nam mu-na-ku5-de6 zi [d]nin-hur-saġ-ka a-ša3
<input checked="" type="checkbox"/> E1.9.3.1.65	[ġeš]KUŠU2(ki)-ke4 e2-an-na-tum2-ra nam mu-na-ku5-de6 zi [d]en-ki lugal
<input checked="" type="checkbox"/> E1.9.3.1.75	[ġeš]KUŠU2(ki)-ke4 e2-an-na-tum2-ra nam mu-na-ku5-de6 zi [d]EN.ZU amar
<input checked="" type="checkbox"/> E1.9.3.1.85	[ġeš]KUŠU2(ki)-ke4 e2-an-na-tum2-ra nam mu-na-ku5-de6 zi [d]utu lugal
<input checked="" type="checkbox"/> E1.9.3.1.95	[ġeš]KUŠU2(ki)-ke4 e2-an-na-tum2-ra nam mu-na-ku5-de6 zi [d]nin-ki-ka a-ša3
<input checked="" type="checkbox"/> E1.9.3.1.111	[d]nin-ġir2-su-ka-ke4 [d]nin-ġir2-su-ra mu-na-ru2-a-e
<input checked="" type="checkbox"/> E1.9.3.1.114	[d]nin-ġir2-su-ra šu-na mu-ni-ni4-a mu-na-ru2 e2-an-na-tum2 kur-ru2

The storage of a phrasal query object is also quite straightforward. Whenever a phrase is defined (i.e. it is checkmarked in the query results list when the *Define* button is clicked), the start and end nodes of the phrase are marked with numeric attributes which indicates the level of nesting for the phrase. This is best understood by considering the second example in Table 3.6: both *namlugal* ‘kingship’ and *Lagaš* are noun phrases on their own, but *namlugal Lagaš* is also a noun phrase meaning ‘kingship of Lagaš’.



### 3.6 State of the Annotation

Using this query-based annotation approach, a significant amount of markup has already been added to the corpus. Since the focus of the research, described in Chapters 4 and 5, is on the dimensional prefix system and the conjugation prefix system, those prefixes have received the bulk of the attention. A complete list of the queries used to annotate the current corpus is contained in Appendix A. For those not interested in the full details of the queries being used, the current state of the annotation is summarised in Table 3.7. It should be stressed that the query objects described here are only one linguist's view of Sumerian morphology, and are dedicated to one particular line of research. Other linguists and other Sumerologists are free to ignore these query objects and use LPattern to define ones which are appropriate to their own theoretical positions and research needs.

While LPattern does make the job of annotation easier, it is not a panacea. In particular, there are many instances where identical orthographies could represent two different underlying morphologies. LPattern is unable to tease apart such distinctions, so manual intervention is required. For instance, the NP clitics for the ergative and locative 2 cases both consist of a suffixed *-e*, and the two cases cannot be distinguished by looking at a word's orthography. The situation is similar for the genitive and locative clitics. Although the genitive clitic is *-ak* and the locative is *-a*, due to amissible final consonants, they usually end up being indistinguishable in the orthography.

In the case of ambiguities, the general approach has been to over-annotate, taking the position that good recall is more important than good precision. Thus, the corpus currently contains 6806 noun phrases which are marked as both NP-ERG and NP-LOC2, and 3292 noun phrases which are annotated as both NP-GEN and NP-LOC. Accurately annotating the corpus for these case clitics would require analysing the context of each of the occurrences in order to determine which case is actually present. Until that happens, any query involving these objects will return a large number of excess hits, which will need to be thinned out manually.

Given the focus of the research being undertaken, it is naturally the verbs which have

Table 3.7: Summary of current annotations

		Query object		Notes
Phrases		NP		To identify NP clitics
Verbs	Modal prefixes	V- <i>bara</i>	<i>ba-ra-</i>	To isolate DAT.2SG <i>ra-</i>
	Conjugation prefixes	V- <i>mu</i> V- <i>i</i> V- <i>imma</i> V- <i>immi</i> V- <i>ba</i>	<i>mu-</i> <i>i<sub>3</sub>-</i> <i>im-ma-</i> <i>im-mi-</i> <i>ba-</i>	Also allomorphs like <i>ma-</i> Also allomorphs like <i>e-</i> Also <i>e-ma-</i> , <i>am-ma-</i> , and others Also <i>e-mi-</i> , <i>am-mi-</i> , and others
	Dimensional prefixes	V-DAT.1SG V-DAT.2SG	<i>a-</i> <i>ra-</i>	With conjugation prefix <i>ma-</i> Often <i>ma-ra-</i> with conjugation prefix <i>ma-</i>          Only isolated in certain contexts
		V-DAT.3SG V-DAT.3PL	<i>na-</i> <i>ne-</i>	
		V-COM	<i>da-</i>	
		V-ALL	<i>ši-</i>	
V-ALL.3SG		<i>nši-</i> , <i>mši-</i>		
V-ALL.3N		<i>bši-</i>		
V-ABL		<i>ta-</i>		
V-LOC V-LOC2		<i>ni-</i> <i>e-</i>		
Unclassified prefixes	V- <i>bi</i> V- <i>im</i> V- <i>mi</i> V- <i>mini</i>	<i>bi<sub>2</sub>-</i> <i>i<sub>3</sub>-im-</i> <i>mi-</i> <i>mi-ni-</i>	An allomorph of <i>i<sub>3</sub></i> ? An allomorph of <i>mu-</i> or <i>bi-</i> ? Derived from V- <i>bi</i> + V-LOC?	
Suffixes	V-SUB	<i>-a</i>	Needed for defining NP	
Nouns	Dimensional clitics	NP-DAT NP-COM NP-ALL NP-ABL NP-LOC	<i>-ra</i> <i>-da</i> <i>-še<sub>3</sub></i> <i>-ta</i> <i>-a</i>	Needs to be manually separated from NP-GEN Needs to be manually separated from NP-ERG
		NP-LOC2	<i>-e</i>	
	Other case clitics	NP-ERG	<i>-e</i>	Needs to be manually separated from NP-LOC2
		NP-GEN	<i>-ak</i>	Needs to be manually separated from NP-LOC
Other clitics	NP-EQU	<i>-gin<sub>7</sub></i>		
	NP-3SG.POSS	<i>-ani</i> , <i>-ni</i>		
	NP-3N.POSS NP-3SG.COP	<i>-bi</i> <i>-am</i>		

received most of the attention. In particular, only the conjugation prefixes and dimensional prefixes have been fully annotated. Other morphemes have been marked up as well, but only where necessary. For instance, a query object was created for the modal prefix *bara-* because of possible confusion with both the *ba-* conjugation prefix and the DAT.2SG prefix *ra-*. Other elements of the verbal morphology, such as the subject and object agreement prefixes, which were not relevant to the research at hand, were not assigned query objects.

Query objects were defined for the conjugation prefixes *mu-*, *i-*, *imma-*, and *ba-*. This includes allomorphs such as *ma-* (the *mu-* prefix followed by a DAT.1SG prefix) and *e-* (the allomorph of *i-* before a [−ATR] vowel). The *immi-* prefix was annotated separately from *imma-*; although *immi-* generally behaves very much like *imma-* (Woods, 2008), V-*immi* merited its own query object in order to allow the prefix to be studied separately.

All the dimensional prefixes have had query objects defined for them. In many instances, there is a certain degree of orthographic ambiguity, but this is less of an issue than with the case clitics. For instance, *ra-* could be either a DAT.2SG prefix or an ABL prefix, but the location of the *ra-* relative to other prefixes is generally enough to determine whether the prefix is dative or ablative. Examples of some of these inherently ambiguous dimensional prefixes are given in Table 3.8.

Table 3.8: Ambiguous queries for dimensional prefixes

Query	Possibilities	Comments
V"da-"	V-COM V-ABL	Requires manual disambiguation
V"ni-"	V-LOC V-DAT.3SG-LOC2 V-LOC2	DAT.3SG <i>na-</i> followed by LOC2 <i>i-</i> LOC2 <i>i-</i> written with the $i_3$ sign $\overline{\text{ni}}$ , also read <i>ni</i>
V"ra-"	V-DAT.2SG V-ABL	Requires manual disambiguation
V"ri-"	V-DAT.2SG-LOC2 V-ABL-LOC2	<i>ra-</i> followed by LOC2 <i>i-</i>

Other prefixes, whose classification is unclear, also had query objects defined for them. The most important of these is *bi-*; while the actual status of *bi-* may be in dispute, it is easy enough to identify its occurrences. Annotations were also created for other prefixes discussed in the literature, such as *a-*, *mini-*, and *im-*, as part of determining whether they might also be conjugation prefixes themselves.

For the nouns, the task of annotating has one significant additional complication. Many of the nominal suffixes are actually clitics which attach to noun phrases rather than just nouns. Consequently, these clitics are often found attached to an adjective or to a non-finite verb. It was thus necessary to build up a query object for noun phrases, using the queries shown in Table 3.9, where the += operator indicates the union of a query's result set with an existing query object. It is worth pointing out that the definition of query objects is not recursive. That is, when the query *NP ADJ* is used as part of the definition of the NP query object, it only located *NP ADJ* sequences where the NP has previously been defined as an NP. Hence the need for the query *NP ADJ ADJ*, which would be unnecessary if the process were truly recursive. In theory, the lack of recursion would mean that further queries such as *NP ADJ ADJ ADJ* and *NP V-SUB ADJ* are also necessary, but in practice they are unnecessary since they produce no additional hits against the actual corpus.

Table 3.9: Defining the NP query object

1.	NP = N
2.	NP += PD
2.	NP += NP ADJ
3.	NP += NP V-SUB
4.	NP += NP ADJ ADJ

The order of queries is significant when defining the NP object, and this is true for certain other queries as well. As mentioned above, it was important to first define the *V-bara* query object for the *bara-* modal prefix prior to being able to accurately retrieve the DAT.2SG *ra-* prefix. Likewise the query objects for the NP-3SG.POSS and NP-3SG.COP clitics both needed

to be defined before the NP-LOC object, since those two clitics (*-ani* and *-am*) can be easily confused with the suffixed *-a* representing the locative case.

The process of defining query objects is an ongoing one. A large number of additional queries have been mapped out, based upon a synthesis of Thomsen (1984), Edzard (2003), and Michalowski (2004), all of whom provide differing accounts of which morphemes are present and how they are represented in Sumerian orthography. So far the only query objects which have been defined are those which are necessary to support the sorts of queries described in Chapters 4 and 5, but the intention is that as other questions of Sumerian morphosyntax are explored, the system of query objects would gradually extend until it eventually becomes a full annotation of the corpus.

# Chapter 4

## Dimensional Prefixes

Like many other languages, Sumerian has verbal morphology to indicate agreement with the verb's subject and direct object, but it also has a set of verbal prefixes which correspond to the presence of oblique NPs. These prefixes are described as “taking up” or “resuming” the nominal, or as indicating “coreference” or “concord”.

Traditionally referred to as the “dimensional infixes”, these prefixes appear in the verbal chain immediately following the conjugation prefix, and immediately before the subject agreement prefixes, as shown in (4.1), which shows an example with several dimensional prefixes cooccurring.

(4.1) Verbal complex with multiple dimensional prefixes (Thomsen, 1984)

*⟨mu-na-ra-ni-e<sub>3</sub>-eš⟩*  
mu-**na-ra-ni**-e<sub>3</sub>-eš  
CONJ-DAT.3SG-ABL-LOC-come.out-ABS.3PL  
'They came out for him from there'

While these dimensional prefixes often reflect the presence of an NP with the corresponding case, Gragg (1973) notes that there are frequent mismatches between the verbal prefixes and the NP suffixes with which they notionally correspond. Indeed, for any of the prefixes, the number of exceptions to agreement in the corpus significantly outnumbers the number of instances where the expected agreement does take place.

The degree of mismatches can be seen in Table 4.1, which uses LPattern queries based on the query objects described in §3.6. Each row summarises the totals for a particular case.<sup>1</sup> So for instance, the first row shows the results for dative case, accumulated using the LPattern queries *NP-DAT*, *V-DAT*, and *NP-DAT\*V-DAT*. Some broad trends are apparent. First, it is clear that for all the cases, in only a minority of the sentences does the nominal clitic cooccur with the verbal prefix.<sup>2</sup> Second, the verbal prefixes for the dative and comitative cases are considerably more common than the nominal clitics for those cases. Third, the clitics for the allative and ablative cases significantly outnumber the prefixes.

Table 4.1: Dimensional clitics, prefixes, and cooccurrences

<i>x</i>	NP- <i>x</i>	V- <i>x</i>	NP- <i>x</i> *V- <i>x</i>
DAT	1356	3131	518
COM	806	2277	245
ALL	3835	967	334
ABL	1946	677	300
LOC	n/a	3264	n/a
LOC2	n/a	3556	n/a

Any given verb will typically only have one or two of the dimensional prefixes, although it is possible to have several, as was shown in (4.1). Note that (4.1) also shows one of the most common reasons that the dimensional prefixes appear without a nominal bearing the corresponding case clitic: being a pro-drop language, Sumerian tends to omit pronouns except when they are needed for emphasis. This helps to explain why the V-DAT and V-COM entries in Table 4.1 are so much higher than the corresponding NP-DAT and NP-COM entries.

Sentence (4.1) could have explicitly included the pronoun as in (4.2), but this would suggest a particular emphasis on the goal/benefactive. Likewise, if it had been desirable to emphasise the source or the location of the action, nouns with the ablative or locative case clitics could

<sup>1</sup>Totals for the NP-LOC and NP-LOC2 objects are omitted because they would be grossly inflated due to the large numbers of misidentified GEN and ERG clitics respectively.

<sup>2</sup>In fact, the rightmost column probably overstates the actual number of within-sentence cooccurrences, since the LPattern searches are scoped to a <para> element, which may contain several sentences.

have explicitly been included.

(4.2) Verbal complex with multiple agreement prefixes and explicit dative

⟨*e-ne-ra*    *mu-na-ra-ni-e<sub>3</sub>-eš*⟩  
 ene=**ra**    mu-**na**-ra-ni-e<sub>3</sub>-eš  
 3SG=**DAT**    CONJ-**DAT.3SG**-ABL-LOC-come.out-ABS.3PL  
 ‘They came out *for him* from there’

For several of the oblique cases, the form of the case prefix bears a phonological resemblance to the corresponding NP clitic, as can be seen in Table 4.2. This has led to the suggestion that the prefixes are etymologically connected to the NP clitics, but due to the lack of supporting data, this can be no more than a speculation.

In addition to reflecting the presence of an NP in the appropriate case, some of the verbal prefixes also have morphology to show agreement with  $\phi$ -features of the corresponding NP. This is clearest with the dative case, which uncontroversially represents the person and number features of the corresponding argument. For the other oblique cases, the situation is less straightforward.

Edzard (2003) goes the farthest, suggesting that all the cases have corresponding person and number morphology. Thomsen (1984) takes a more moderate position, stating that  $\phi$ -feature agreement extends to the comitative and allative cases in addition to the dative. If  $\phi$ -feature agreement is actually present for cases other than the dative, it is not consistently represented in the orthography. It should be noted that even for the structural cases, which are uniformly accepted to show person, number and noun-class agreement, the actual orthography does not reliably reflect the subject and direct object agreement morphology. One explanation for the apparently sporadic indication of person and number morphology for the dimensional prefixes is “Krecher’s Rule”, which will be discussed in §4.3.

There are also numerous instances where the verbal agreement fails to appear even though the sentence contains a nominal in the appropriate case to trigger agreement. Gragg (1973) analyses such instances, and hypothesises that certain verbal roots are incompatible with certain



Table 4.2: Forms of dimensional prefixes (Thomsen, 1984)

Case	Verbal prefix	Corresponding NP clitic	Agreement
Dative	<i>a-/ra-/na-/me-/ne-</i>	<i>-ra</i>	person, number
Comitative	<i>da-/e-da-/n-da-/b-da-</i>	<i>-da</i>	person, number, animacy
Allative	<i>še<sub>3</sub>-/ši-</i>	<i>-še<sub>3</sub></i>	person, animacy
Ablative	<i>ra-/ta-</i>	<i>-ta</i>	
Locative	<i>ni-</i>	<i>-a</i>	
Locative 2	<i>e-</i>	<i>-e</i>	

agreement morphemes. If so, any model of the Sumerian dimensional prefixes will have to account for these anomalies.

## 4.1 Applicatives

The term “applicative” originally comes from the study of the morphology of Bantu languages. For instance, in a Bantu language such as Kichaga, an applicative morpheme appears on the verb to introduce benefactives, locatives, instrumentals, and motives as arguments of the verb, as shown by the contrast in (4.3). In effect, the applicative morpheme has taken a secondary thematic role and made it into one of the core arguments of the verb.

(4.3) Applicative morphemes in Kichaga (Bresnan and Moshi, 1990)<sup>3</sup>

*N-ǎ-í-ly-à*                      *k-élyà.*  
 FOC-1.SBJ-PRS-eat-FV 7-food  
 ‘He/She is eating food.’

*N-ǎ-í-lyì-í-à*                      *m-kà k-élyà.*  
 FOC-1.SBJ-PRS-eat-APPL-FV 1-wife 7-food  
 ‘He is eating food for/on his wife.’

Recent work on applicatives, most notably by Pylkkänen (2002) and Cuervo (2003), has suggested that these heads have a much broader role, and are relevant to languages beyond

<sup>3</sup>In glosses: FV = final vowel; numerals indicate noun classes.

Table 4.3: Argument introducers (Pylkkänen, 2002)

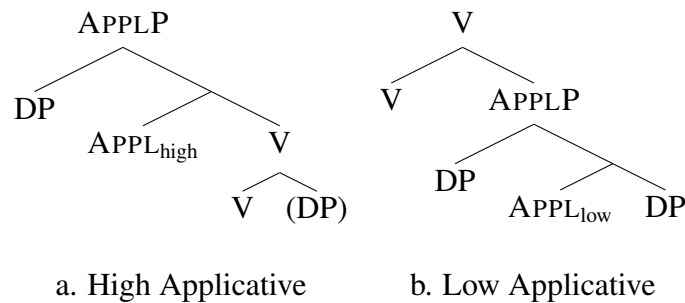
Head	Meaning
(1) High Applicative	Thematic relation between an applied argument and the event described by the verb.
(2) Low Recipient Applicative	A transfer of possession relation between two individuals: asserts that the direct object is <i>to</i> the possession of the indirect object.
(3) Low Source Applicative	A transfer of possession relation between two individuals: asserts that the direct object is <i>from</i> the possession of the indirect object.
(4) Root-selecting CAUSE	Relates a causing event to a category-free root.
(5) Verb-selecting CAUSE	Relates a causing event to a verb.
(6) Phase-selecting CAUSE	Relates a causing event to a phase, i.e. is able to combine with a constituent to which an external argument has been added.
(7) VOICE	Thematic relation between the external argument and the event described by the verb.

those which have traditionally been considered to have applicative constructions. Constructions such as the English and Japanese double object constructions can be analysed as examples of applicatives. Just like the benefactive in the Kichaga applicative construction from (4.3), the recipient of a double object construction becomes one of the immediate arguments of the verb, instead of being relegated to a PP complement. The only difference is that English lacks an overt applicative morpheme.

Pylkkänen argues that applicative heads are of great importance cross-linguistically, forming a significant part of the inventory of heads whose purpose is to introduce the arguments of a verb. The particular type of head used to introduce an argument determines the relationship between that argument and the event being described by the verb, as summarised in Table 4.3. In her analysis, double object constructions would be classified as low recipient applicatives.

In structural terms, the distinction between high and low applicatives is shown by the trees in Figure 4.1. Semantically, the essential difference is that the high applicative establishes a relation between an individual (the DP in Spec/APPL<sub>high</sub>) and an event (the VP complement of

Figure 4.1: Structure of high vs. low applicatives (McGinnis, 2005)

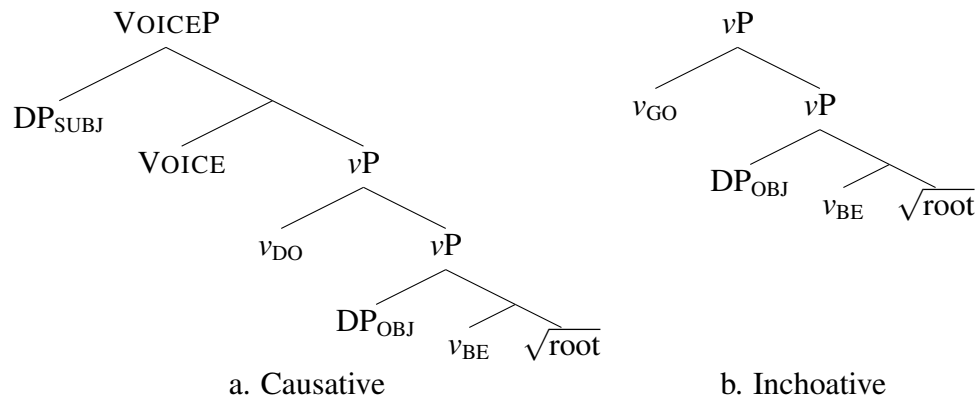


APPL<sub>high</sub>) while a low applicative establishes a semantic relationship between two individuals (one in the specifier of APPL<sub>low</sub> and one in its complement).

In order to constrain the scope of this dissertation, we will not be considering the CAUSE heads identified by Pylkkänen, so the discussion will be restricted to Pylkkänen's high and low applicatives (as well as the VOICE head). Indeed, Cuervo (2003) casts doubt on the need for CAUSE heads. In her analysis, instead of requiring a CAUSE head, the causal semantics are a consequence of the bieventive structure of the causative itself, which consists of one *v*P inside another, as in Figure 4.2. The same internal structure is also applicable to inchoatives, except that the *v*<sub>DO</sub> of the causative is replaced by a *v*<sub>GO</sub> in the inchoative, and the inchoative also has no need for a VOICE projection since it lacks an Agent.

The rest of this chapter shows how applicative heads of the sort proposed by Pylkkänen and Cuervo can be extended to account for the verbal prefixes encountered in Sumerian. It should be mentioned that this is not the first time that applicatives have been discussed in reference to Sumerian. Johnson (2004) analyses the *bi-* prefix as a low source applicative, an analysis which is not compatible with the one being presented here. However, we have followed the mainstream view that *bi-* does not actually exist as a separate prefix, but consists merely of the *ba-* conjugation prefix followed by a LOC2 prefix *i-*.

Figure 4.2: Causative and inchoative constructions (Cuervo, 2003)



While Pyllkänen and Cuervo both retain the distinction between the VOICE heads and the applicative heads, they are clearly part of the same argument-introduction mechanism. The essential difference is that the VOICE head introduces the topmost argument, while the APPL heads introduce oblique arguments which are lower in the structure. The extra salience one associates with the VOICE head is merely a consequence of the special importance the Agent  $\theta$ -role has in the event structure of a verb. Dividing the VOICE head from the rest of the applicative system is largely a matter of tradition; it would be more accurate to refer to the VOICE head as  $\text{APPL}_{\text{AGENT}}$ , that is, as the applicative head which introduces the Agent. This would make it clearer that the VOICE head is only *primus inter pares* among the applicative heads.

Following proposals from Pesetsky (1995) and Marantz (1997), verbs can be understood as consisting of a category-neutral root which gets its verbal status by being placed in an environment where a verb would be appropriate. Given this, it is hard to argue for the traditional view where the verb has an associated lexical entry which defines the argument structure around it. Instead, the argument structure is external to the verb; when the root is merged into the structure it is constrained to being a root which is compatible with the environment into which it is

being inserted.

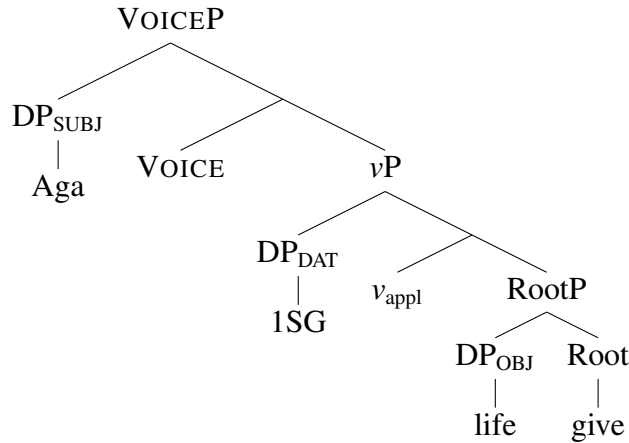
This represents an inversion of the traditional way in which an event is understood to get its argument structure. Rather than the verb imposing the argument structure, it is the structure which imposes constraints on the choice of roots which will serve as the verb. Such a position is essentially the same as the “exo-skeletal” model described by Borer (2005).

## 4.2 Dimensional Prefixes as Applicative Morphology

The original inspiration for considering the dimensional prefixes as applications was the work by Béjar (2003) on the mechanisms of agreement. In her account, verbal agreement is caused by the need for uninterpretable features to match and value. In the case of the ergative and absolutive cases, this agreement can be accounted for by means of uninterpretable features on the  $T^\circ$  and  $v^\circ$  heads. What is relevant for Sumerian is that Béjar uses a  $v_{appl}$  head to account for indirect-object agreement in Georgian and Choctaw. This seems to directly parallel the behaviour of the Sumerian dative case, so such an account could be applied to Sumerian as well. This analysis is shown in Figure 4.3.

Following Pykkänen (2002) and Cuervo (2003), it seems likely that any satisfactory account for applicatives requires the presence of separate APPL heads, rather than following Béjar and placing the applicative features on a  $v$  head. For one thing, placing the applicative features on  $v$  effectively entails that the construction is a high applicative, yet, as will be shown in §4.2.1, the semantics of the verb suggest that here we have a low source applicative. Figure 4.4 reformulates Figure 4.3 to incorporate APPL as a separate head. Also, the head is renamed to  $APPL_{BEN}$  in order to give a better indication that here it corresponds specifically to the benefactive  $\theta$ -role.

While the dative case is the one which most clearly displays  $\phi$ -feature agreement, several of the other dimensional cases show some level of agreement as well. The extent of the agreement for these other cases depends on whether one accepts the range of agreement claimed by Edzard

Figure 4.3: Dative agreement with applicative  $v$  head (based on (Béjar, 2003))

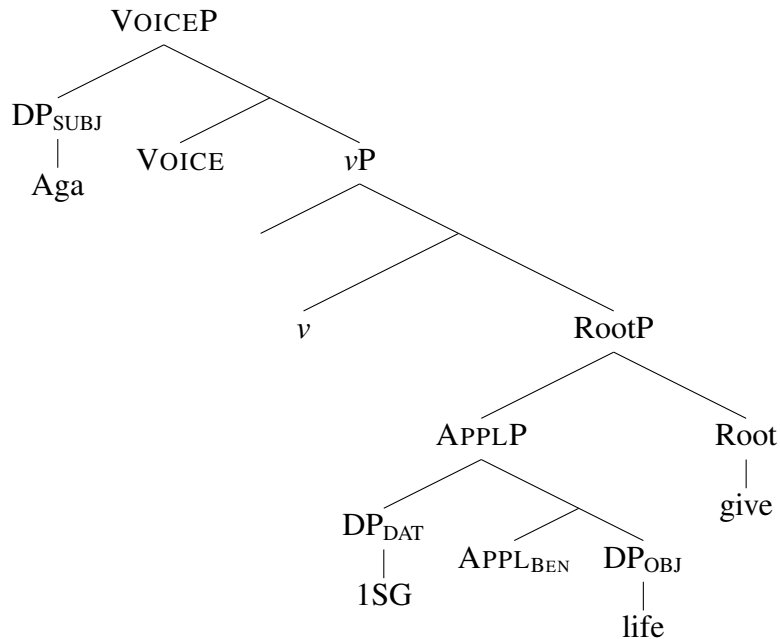
⟨ag-ga      nam-til<sub>3</sub>      ma-an-šum<sub>2</sub>⟩  
 Aga=e      namtil-∅      mu-a-n-šum  
 Aga=ERG    life=ABS    CONJ-DAT.1SG-ERG.3SG-give  
 ‘Aga gave me life’ (GgAk)

(2003), the more conservative position of Michalowski (2004), or the intermediate position taken by Thomsen (1984). Table 4.4 summarises Thomsen’s position. Michalowski agrees about the dative and comitative cases, but does not describe any agreement for the allative case.

Table 4.4: Summary of  $\phi$ -feature agreement for dimensional prefixes

Case	1SG	2SG	3SG	3N	1PL	2PL	3PL
Dative	<i>a-</i>	<i>ra-</i>	<i>na-</i>	—	<i>me-</i>	—	<i>ne-</i>
Comitative	<i>da-</i> or <i>eda-</i>	<i>eda-</i>	<i>nda-</i>	<i>bda-</i>	—	—	<i>PI-da-</i>
Allative	<i>muši-</i>	<i>eši-</i>	<i>nši-</i>	<i>baši-</i> , <i>bši-</i> or <i>mši-</i>	—	—	—

There is no reason to believe that the syntactic explanation for the dative prefix is any different from the other dimensional prefixes. The exact details differ for the other prefixes, but the same general mechanism of applicative heads which accounts for the dative can account for the other cases as well.

Figure 4.4: Dative agreement with APPL<sub>BEN</sub> head

⟨ag-ga      nam-til<sub>3</sub>    ma-an-šum<sub>2</sub>⟩  
 Aga=e      namtil=∅    mu-a-n-šum  
 Aga=ERG   life=ABS   CONJ-DAT.1SG-ERG.3SG-give  
 ‘Aga gave me life’ (GgAk)

### 4.2.1 Dative Prefix

The dative case is the one whose prefixes most clearly show agreement for person and number features. The dative does not refer to inanimate nominals, and it happens not to be attested in the second person plural. These prefixes are summarised in Table 4.5. The dative prefix appears before any of the other dimensional prefixes. Since the first person dative almost always follows the conjugation prefix *mu-*, it usually appears as *ma-*, derived from /mu/+a/. The second person dative also has a strong tendency to cooccur with the *mu-* conjugation prefix, while the third person can appear following any of *mu-*, *imma-*, or *ba-*. The datives are never found after the prefixes *immi-* and *bi-*, nor after the stative prefix *al-*.

The core meaning of the dative prefix is of course the notion of giving, and it occurs partic-

Table 4.5: Dative case agreement morphology

	Singular	Plural
First	<i>a-</i>	<i>me-</i>
Second	<i>ra-</i>	—
Third	<i>na-</i>	<i>ne-</i>

ularly frequently with the verbs *šum*<sub>2</sub> ‘to give’ and *ba* ‘to allot’, as in (4.4). The dative extends beyond the narrow notion of transferring a physical object to broader senses of giving, such as speech acts. The verbs *dug*<sub>4</sub> ‘to speak’ and *gu*<sub>3</sub> *de*<sub>2</sub> ‘to say’ are almost always found with a dative-case prefix indicating the addressee, as in (4.5).

## (4.4) Dative with verb of giving

*⟨nam-sipad kalam-ma an-ne<sub>2</sub> ma-ra-an-šum<sub>2</sub>⟩*  
 namsipad kalam=ak An=e mu-**ra**-n-šum  
 shepherdship land=GEN An=ERG CONJ-**DAT.2SG**-ERG.3SG-give  
 ‘An has given you the shepherding of the Land’ (Nanna A)

## (4.5) Dative for addressed speech

*⟨nun-e sukkal-a-ni <sup>d</sup>isimud-ra gu<sub>3</sub> mu-na-de<sub>2</sub>-e⟩*  
 nun=e sukkal=ani Isimud=ra gu mu-**na**-de  
 prince=ERG minister=3SG.POSS Isimud=DAT voice CONJ-**DAT.3SG**-pour  
 ‘The prince spoke to his minister, Isimud’ (InEnk)

The semantic range of the dative prefix also extends to ethical datives, where an action is being done for the benefit of some person. In royal inscriptions, the dative prefix appears on the verb *a ru* ‘to dedicate’ but it also appears on the verbs *du*<sub>3</sub> ‘to build’ and *ak* ‘to do/make’ when the action is directed to the benefit of the king or of a deity.

In general, since the range of arguments introduced by the dative prefix is not restricted to any sort of physical change of location, the Goal  $\theta$ -role seems like a poorer fit than the



Benefactive, so we will refer to the head as  $\text{APPL}_{\text{BEN}}$ . In almost all cases, the dative argument can be seen as the beneficiary of the verb's action. That being said, there are a cases such as (4.6) where the dative argument is affected but not in a beneficial way. Gragg (1973) categorises these as using the dative “to indicate that the subject has some affect on the emotions/sensitivity of an animate object.”

(4.6) Dative prefix with *gig* ‘to pain, to trouble’

<i>a-na-zu</i>	<i>a-ra-gig</i>	<i>zu<sub>2</sub>-mu</i>	<i>ma-gig</i>
ana=2SG.POSS	a- <b>ra</b> -gig	zu <sub>2</sub> =mu	mu- <b>a</b> -gig
what=2SG.POSS	CONJ- <b>DAT.2SG</b> -pain	tooth=1SG.POSS	CONJ- <b>DAT.1SG</b> -pain
‘What pains you? My tooth pains me.’ (EnkNh)			

The question remains as to whether the  $\text{APPL}_{\text{BEN}}$  head is a high applicative, a low applicative, or an affected applicative. The tree given above in Figure 4.3 takes the applicative head to be a high applicative rather than a low one. In contrast, the tree in Figure 4.4 is based on the analysis of the  $\text{APPL}_{\text{BEN}}$  head as a low applicative rather than a high one.

(4.7) Diagnostics for high vs. low applicatives (Pylkkänen, 2002)

DIAGNOSTIC 1: TRANSITIVITY RESTRICTIONS

Only high applicative heads should be able to combine with unergatives. Since a low applicative head denotes a relation between the direct and indirect object, it cannot appear in a structure that lacks a direct object.

DIAGNOSTIC 2: VERB SEMANTICS

Since low applicatives imply a transfer of possession, they make no sense with verbs that are completely static: for example, an event of holding a bag does not plausibly result in the bag ending up in somebody's possession. High applicatives, on the other hand, should have no problem combining with verbs such as *hold*: it is perfectly plausible that somebody would benefit from a bag-holding event.

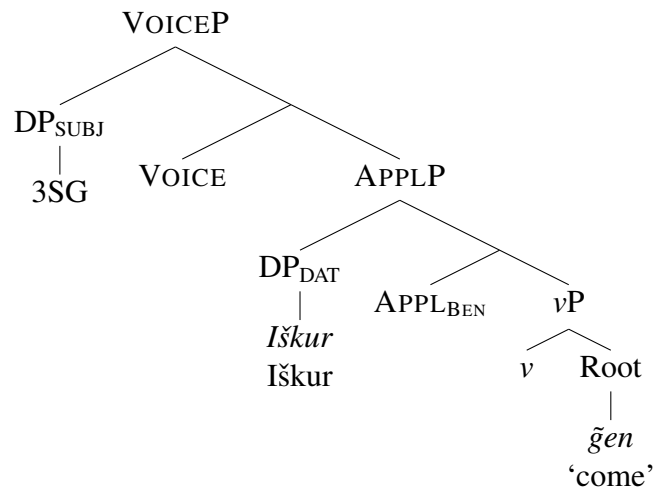
While Figure 4.4 diagrammed a sentence where  $\text{APPL}_{\text{BEN}}$  was low, the evidence from both diagnostics indicates that  $\text{APPL}_{\text{BEN}}$  can at least sometimes be a high applicative. For the first diagnostic, there are examples of the dative prefix with verbs which clearly lack a direct object. This is illustrated with *ḡen* ‘to come’ in (4.8), which appears with a dative prefix when the individual is coming with the purpose of praying to a god.

(4.8) Dative prefix with unergative *ḡen* ‘to come’

$\langle {}^d i\check{s}kur \ mu-na-an-\check{g}en \quad \check{s}udu_3 \ mu-na-\check{s}a_4 \rangle$   
 Iškur mu-**na**-n-ḡen                      šudu    mu-na-ša  
 Iškur    CONJ-DAT.3SG-ERG.3SG    prayer    CONJ-DAT.3SG-pray?  
 ‘He came for Iškur; he prayed to him.’ (Utu-heḡal)

Example (4.8) also provides evidence for the second diagnostic since it involves no transfer of possession. Indeed, the vast majority of dative prefixes in Sumerian occur in contexts where a transfer of possession is clearly absent. For such cases, a tree like Figure 4.5 seems more plausible than Figure 4.4.

Figure 4.5: Tree for example (4.8) with high APPL<sub>BEN</sub>



‘He came for Iškur’

The fact that the dative case is restricted to animate arguments may be related to its role as a Benefactive marker rather than a Goal marker; by their nature it is difficult for inanimates to be the beneficiaries of an action. When an inanimate nominal is at the receiving end of an action, it is typically marked with the locative 2 case, which is clearly connected to physical movement and a Locative  $\theta$ -role (§4.2.5).

### 4.2.2 Comitative Prefix

The comitative prefix actually has a broader range of  $\phi$ -feature marking than the dative, since it is marked for animacy as well as for person and number features, as shown in Table 4.6. The comitative can cooccur with any of the other dimensional prefixes.

All forms of the comitative prefix include the same *da* morpheme which is also found as the comitative clitic. The first person singular form is often abbreviated to *da-* because the initial vowel is assimilated into the vowel of a preceding conjugation prefix (typically *mu-*). The third person plural form is written with the *PI*  $\text{𐎧𐎠}$  sign, but the reading is uncertain. The correct reading might be *be*<sub>3</sub>, *ta*<sub>x</sub>, or *da*<sub>x</sub> (Gragg, 1973; Thomsen, 1984), but there is not enough evidence to make a decision.

Table 4.6: Comitative  $\phi$ -feature agreement morphology

	Singular	Plural
First	<i>da-</i> or <i>eda-</i>	—
Second	<i>eda-</i>	—
Third	<i>nda-</i>	<i>PI-da-</i>
Inanimate	<i>bda-</i> or <i>mda-</i>	

The distribution of  $\phi$ -feature agreement morphology in the corpus is summarised in Table 4.7. For the comitative case, the  $\phi$ -feature morphology is not consistently indicated, and often only *da-* is written. In older texts (before the Ur III period), this discrepancy may be due to a tendency for scribes to not write syllable-closing consonants (Edzard, 2003). However, the discrepancy may also be due to the operation of Krecher’s Rule (§4.3).

The semantic range associated with the comitative prefix is broad, but centres around the notion of accompaniment. When two subjects are engaged in the same activity, the comitative prefix can occur to indicate that they are undertaking the activity together. So while *dug*<sub>4</sub> with the dative prefix might mean ‘to speak to’, with the comitative prefix it means ‘to speak with’ or ‘to converse’. Similarly, the comitative prefix is found on verbs like *sa*<sub>2</sub> ‘to equal’ whose meaning inherently involves a shared action, state, or property.

Table 4.7: Occurrences of comitative  $\phi$ -feature morphology

Query object	Count	Percent
V-COM. 1SG	461	20.2%
V-COM. 2SG	256	11.2%
V-COM. 3SG	509	22.4%
V-COM. 3N	206	9.0%
V-COM (unmarked)	845	37.1%
V-COM (total)	2277	100.0%

This notion of accompaniment extends to verbs which denote activities where the agent and the comitative argument are engaged in a joint activity. For instance, *zu* ordinarily means ‘to know’, but becomes ‘to learn from’ with the addition of a comitative prefix. Some verbs, such as *a<sub>2</sub> aḡ<sub>2</sub>* ‘to instruct’ and *ad gi<sub>4</sub>* ‘to advise’, always denote this sort of joint activity, so they uniformly appear with the comitative prefix.

Beyond the core meaning of actual accompaniment, the comitative prefix is also used for many verbs of emotion. For verbs such as *hul<sub>2</sub>* ‘to rejoice’, *šag<sub>5</sub>* ‘to be pleasing to’, and *ni<sub>2</sub> te* ‘to fear’, the object of the emotion is indicated by the comitative prefix. When the emotion is not directed at or associated with any particular object, the comitative prefix will be absent.

There are a number of verbs whose use of the comitative prefix can best be described as idiosyncratic. For instance, *mu<sub>2</sub>* ‘to grow’ often appears with the comitative prefix when it is used transitively. Similarly, transitive uses of other verbs such as *si* ‘to fill’ and *sim* ‘to filter’ are also accompanied by the comitative prefix.

Also outside the usual semantic range of the comitative are uses of the prefix in an “abilitative” sense. Here the presence of the prefix seems to express a capability of performing the associated action. For instance in (4.9) the presence of the comitative prefix reflects the ability or inability of the agent to perform the verb.

(4.9) Abilitative use of comitative prefix with compound verb *šu gi<sub>4</sub>* ‘to return’

⟨*kiḡ<sub>2</sub>-gi<sub>4</sub>-a ka-ni*                      *dugud šu*  
*kiḡgia ka-ani*                      *dugud šu*  
 messenger mouth-3SG.POSS heavy hand

*nu-mu-un-da-an-gi<sub>4</sub>-gi<sub>4</sub>*⟩

*nu-mu-nda-n-gi-gi*

NEG-CONJ-COM.3SG-ERG.3SG-return-INT

‘The messenger, whose mouth was heavy, was not able to repeat it.’ (EmkLA)

Akkadian scribes were certainly aware of the abilitative use of the comitative prefix. Gragg (1973) notes that in the Neo-Babylonian Grammatical Texts (Hallock and Landsberger, 1956), we find the equation of the Sumerian comitative forms with the Akkadian verb *le’û* ‘to be able’, as shown in Table 4.8. The forms *mu-da*, *e-da*, and *an-da* would appear to be their attempt to represent the comitative prefixes *muda-* (COM.1SG, with *mu-* conjugation prefix), *eda-* (COM.2SG), and *nda-* (COM.3SG).<sup>4</sup> Elsewhere in the same text, the comitative form *e-da* is equated with the Akkadian *ittika* ‘with you’, so the Akkadian scribes were also familiar with the prefix’s core meaning.

Table 4.8: Abilitative use of comitative in the NBGT (Hallock and Landsberger, 1956)

Sumerian	Akkadian
<i>mu-da</i>	<i>e-li-i</i> ‘I am able to’
<i>e-da</i>	<i>te-li-i</i> ‘you are able to’
<i>an-da</i>	<i>i-li-i</i> ‘he is able to’

Gragg (1973) argues that this abilitative use of the comitative prefix is etymologically derived from the comitative’s primary function of accompaniment or joint action. In his view ‘X is able to do Y’ would derive from something like ‘it is with X to do Y’. For example, referring to the sentence in (4.9), he suggests the paraphrase ‘it is not with him to repeat it’.

<sup>4</sup>Note that there is also a verbal element *da* in some compound verbs which appears to have a base meaning of ‘to be able to’. This is probably etymologically connected to the abilitative use of the *da-* prefix. However, the verb *da* does not appear to be what the NBGT scribes are trying to represent, since the forms in Table 4.8 do not correspond to 1SG, 2SG, and 3SG verbal forms at all. As well, it should be noted that the NBGT texts were written more than a millenium after the demise of the last native speaker of Sumerian, so the data should be accepted with some caution.

Despite these divergences from the core meaning, the comitative prefix does appear to be essentially connected to the notion of accompaniment. The same sort of applicative head which accounts for the  $\phi$ -feature agreement for the dative prefix and a Benefactive  $\theta$ -role can also explain the comitative case and a Comitative  $\theta$ -role. Although “Comitative” is not usually considered to be a  $\theta$ -role, it appears to be functioning as such in Sumerian. Since in some contexts the comitative prefix shows animacy-feature agreement, in those contexts the  $\text{APPL}_{\text{COM}}$  head must also have an uninterpretable animacy feature which needs to be matched and valued.

### 4.2.3 Allative Prefix

“Allative” is the name given by Michalowski (2004) to the case which is generally referred to in the literature as the “terminative”. The allative prefix is generally written *ši*, although *še<sub>3</sub>* is also found in earlier texts. The allative can cooccur with any of the other dimensional prefixes except the ablative.

Thomsen (1984) describes the allative case as having  $\phi$ -feature agreement morphology as shown in Table 4.9, but Michalowski (2004) makes no mention of this sort of morphology for the allative prefix. According to Thomsen, plural forms are not attested.

Table 4.9: Allative  $\phi$ -feature agreement morphology (Thomsen, 1984)

	Singular	Plural
First	<i>muši-</i>	—
Second	<i>eši-</i>	—
Third	<i>nši-</i>	—
Inanimate	<i>bši-, mši-, or baši-</i>	

The distribution of this  $\phi$ -feature agreement morphology is summarised in Table 4.10. Like the comitative case, the person and number features are often omitted, and the bare prefix is written. In fact, bare *ši-* prefixes might be more common than Table 4.10 indicates; although

Thomsen (1984) analyses *muši-* and *baši-* as having  $\phi$ -feature morphology, these prefixes could equally well be analysed as the conjugation prefixes *mu-* and *ba-* followed by a bare *ši-* prefix.

Table 4.10: Occurrences of allative  $\phi$ -feature agreement morphology

Query object	Count	Percent
V-ALL . 1SG	96	9.9%
V-ALL . 2SG	152	15.7%
V-ALL . 3SG	236	24.4%
V-ALL . 3SN	323	33.4%
V-ALL (unmarked)	160	16.5%
V-ALL (total)	967	100.0%

Like several of the other prefixes, the number of cooccurrences between the allative prefix and the allative noun clitic is actually quite small. The corpus contains 967 occurrences of the allative verb prefix and 3835 occurrences of the allative clitic, but only 334 cooccurrences. In part this disparity is because the allative clitic  $-še_3$  very commonly appears in an adverbial sense which is not matched by a corresponding prefix on the verb. In fact, Michalowski (2004) identifies these occurrences of  $-še_3$  as representing a completely separate adverbial morpheme, unrelated to the allative.

In general, the allative clitic has a broader semantic range than the allative prefix. As noted above in Table 1.2, the allative clitic can refer both to motion towards and to location in front of, but the allative verbal prefix is largely confined to the notion of action towards a particular direction. The allative prefix is thus a good fit for the Goal  $\theta$ -role. This goal can include actual motion, as well as less literal actions such as directing one's gaze, or other goals in a more metaphorical sense.

When added to a verb with no inherent directional semantics, the presence of the *ši-* prefix adds the notion of direction towards a goal. Thus, *dal* ordinarily means 'to fly', but with the allative prefix it means 'to fly towards'. The base meaning of *ku\_4* is 'to enter', but prefixed with *ši-* it becomes 'to enter into the presence of'.

Gragg (1973) identifies a number of verbal roots which never cooccur with the *ši-* prefix,

most notably *ed*<sub>2</sub> ‘to go/bring out’ and *ed*<sub>3</sub> ‘to go up/down’. However, Gragg’s discussion is concerned with locating mismatches between the allative prefix and an NP bearing the allative clitic, so many of the examples he provides are actually instances of the adverbial *-še*<sub>3</sub> rather than the allative *-še*<sub>3</sub>. In such cases, the absence of *ši-* prefix on the verb can be accounted for by the absence of a Goal  $\theta$ -role in the verb’s argument structure. Even so, there remains a small residue of examples such as (4.10) where there would seem to be a Goal  $\theta$ -role, but there is no *ši-* prefix on the verb. Evidently, the lexical entries for verbs such as *ed*<sub>2</sub> and *ed*<sub>3</sub> are so closely associated with a particular directionality that they are incompatible with the *ši-* prefix.

(4.10) Goal  $\theta$ -role with no allative prefix

*⟨ur-gir<sub>15</sub> ur<sub>3</sub>-ra-še<sub>3</sub> mu-un-ed<sub>3</sub>⟩*  
 urgir ur-še mu-n-ed<sub>3</sub>  
 dog roof-ALL CONJ-ABS.3SG-go.up/down  
 ‘The dog climbed onto the roof.’ (Proverbs 5)

The actual distinction between the allative case as Goal and the dative case as Benefactive can be quite subtle, as seen in sentences such as (4.11). The actual beneficiary of the dedicatory action is the god Ningirsu, but the goal and purpose of the dedication is the life of the king.

(4.11) Allative and dative in combination with *a ru* ‘to dedicate’

*⟨<sup>d</sup>nin-ḡir<sub>2</sub>-su e<sub>2</sub>-ninnu-ra ... nam-ti lugal-ni*  
 Ningirsu Eninnu=ak=**ra** ... namtil lugal=ani  
 Ningirsu Eninnu=GEN=**DAT** ... life king=3SG.POSS  
*en-an-na-tum<sub>2</sub>-ma-še<sub>3</sub> a mu-na-še<sub>3</sub>-ru⟩*  
 Enanatum=ak=**še** a mu-**na-ši**-ru  
 Enanatum=GEN=**ALL** a CONJ-**DAT.3SG-ALL**-dedicate  
 ‘He dedicated it to Ningirsu of E-ninnu ... for the life of his king Enanatum.’  
 (Enanatum I)

Given the allative prefix’s close association with the Goal  $\theta$ -role, the prefix can be accounted for by the existence of an APPL<sub>GOAL</sub> head whose purpose is to insert a Goal into the verb’s argument structure. Since there does seem to be  $\phi$ -feature marking on the allative prefix,



the head must have uninterpretable features for person and animacy. Whether the  $\text{APPL}_{\text{GOAL}}$  head has number features is an open question, given the lack of plural data in the corpus.

#### 4.2.4 Ablative Prefix

The ablative prefix *ta-* occupies the same slot in the verbal prefix chain as the allative prefix, so the two prefixes never cooccur. The ablative prefix is generally written with the same *ta* 𐎠𐎢𐎣 sign which is used for the ablative noun clitic. In some contexts it is written with the *da* 𐎠𐎢𐎣𐎠 or *ra* 𐎠𐎢𐎣𐎠 signs, which has led some to suggest that the prefix may phonetically be /ra/ with an alveolar tap or trill.

The semantic range of the ablative prefix is centred around the notion of motion away from or separation. Not surprisingly, the prefix is frequently found with the verbs *ed*<sub>2</sub> ‘to go/bring out’ and *ed*<sub>3</sub> ‘to go up/down’. When added to a verb with no inherent directionality, the prefix adds the idea of motion away. Thus, *sar* ‘to chase’ becomes ‘to chase away’ when the ablative prefix is present; *ḡar* ordinarily means ‘to place’, but it means ‘to remove’ with the ablative prefix.

(4.12) *sar* ‘to chase’ with and without ablative prefix

*lugal ḡis̄kiri<sub>6</sub> ib<sub>2</sub>-ta-an-sar-re*  
 lugal kiri i-**ta**-n-sar-e  
 master orchard CONJ-**ABL**-ERG.3SG-chase-*ed*  
 ‘(When a dog goes into an orchard to get dates,) the owner of the orchard chases him away.’ (Proverbs 5)

*<sup>d</sup>nu-nam-nir i<sub>3</sub>-ḡen ki-sikil mu-un-sar-re*  
 Nunamnir i-ḡen kisikil mu-n-sar-e  
 Nunamnir CONJ-go maiden CONJ-ERG.3SG-chase-e  
 ‘Nunamnir went; the maiden chased him.’ (EnlNI)

The core semantic range of the ablative prefix centres around the notion of motion away from, so it seems like a perfect fit for the Source  $\theta$ -role. Since the ablative can only refer to

an inanimate nominal, it is not unexpected that the  $\text{APPL}_{\text{SRC}}$  head lacks  $\phi$ -feature agreement morphology.

Beyond the core directional meaning the ablative prefix sometimes adds a negative or destructive meaning to a verb. So *gul* means ‘to destroy’, but with the *ta-* prefix it means ‘to destroy utterly’. In such cases, the action can be considered to proceed so completely that the object is removed from existence. In a similar metaphorical sense of removal, the ablative prefix appears on the verb *zal*, which is the verb used to express the passing of time. In such cases, the Source argument introduced by the  $\text{APPL}_{\text{SRC}}$  head would be a generic one, indicating movement in a general sense away “from here” or “from now”.

In addition, the ablative prefix also occasionally has an instrumental reading. While this instrumental use is difficult to reconcile with the Source  $\theta$ -role, there are other languages, such as Latin, where the ablative case can also be used with an instrumental meaning.

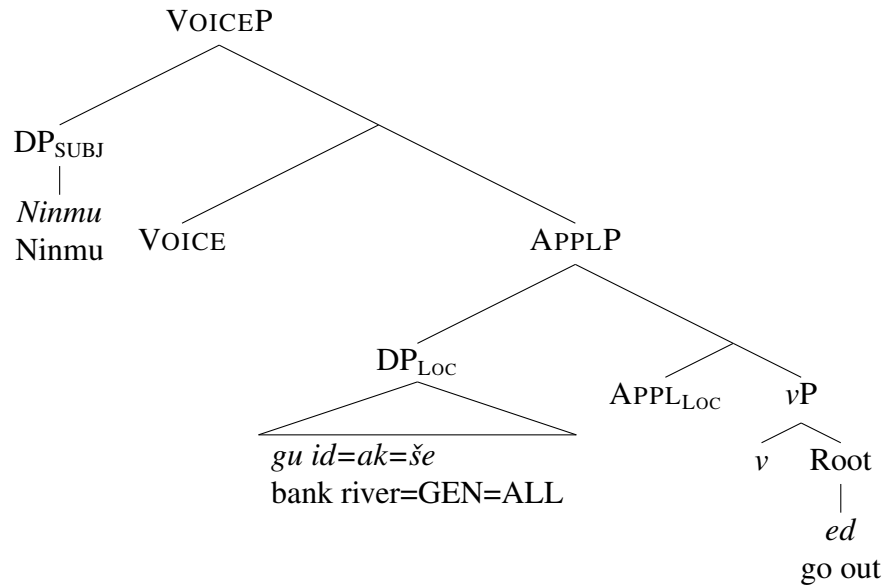
#### 4.2.5 Locative Prefix

The locative case is one of the most straightforward of the cases in terms of its orthography. It is uniformly marked with the verbal prefix *ni-*. The corresponding locative noun clitic is *-a*.

The semantic range is very closely confined to specifying the location of the associated action. Consequently, a verb with the locative prefix can cooccur with a noun bearing one of the other locational clitics when it is important to emphasise the location of an action, as in (4.13) and Figure 4.6. While the allative *-še* clitic usually indicates motion towards, it can also indicate stationary location adjacent to a place, versus the ordinary locative *-a* enclitic, which refers to location directly within a place. So in (4.13) the *ni-* prefix merely reflects the Locative  $\theta$ -role, while the *-še* on the noun indicates that the action is taking place in front of the bank of the river, rather than directly in or on it. Here, instead of its usual role of indicating motion towards a reference point, the *-še* clitic indicates stationary location relative to a reference location, as described above in Table 1.2.

(4.13) Allative-case nominal with locative *ni-* agreement prefix (Gragg, 1973)

*<sup>d</sup>nin-mu<sub>2</sub> gu<sub>2</sub> id<sub>2</sub>-da-ke<sub>4</sub>-še<sub>3</sub> mi-ni-ib-e<sub>3</sub>*  
 Ninmu gu id=ak=še CONJ-ni-b-ed  
 Ninmu bank river=GEN=ALL CONJ-LOC-ERG.3SG-go out  
 ‘Ninmu went out on the bank of the river’ (EnkNh)

Figure 4.6: Allative-case nominal with locative *ni-* agreement prefix

#### 4.2.6 Locative 2 Prefix

The term “locative 2” is the name Michalowski (2004) gives to what has traditionally been referred to as the “locative-terminative” case. This case is one of the most elusive in terms of determining its presence or absence. The noun clitic form is *-e*, which is often assimilated to a stem-final vowel or mistaken for the ergative case clitic (also *-e*).

The form of the verbal prefix is usually given as /e/ or /i/, although the possibility also exists that it might be /j/ to account for some unusual orthographies (Zólyomi, 1999; Karahashi, 2000/2005). Whatever the prefix’s underlying form, it is generally observed orthographically as a change of the quality of a preceding vowel. For instance, the comitative *da-* might be written as *di-* or *de<sub>3</sub>-* when followed by the LOC2 prefix.

Karahashi (2000/2005) provides a detailed discussion of the various orthographic manifestations of this prefix. This is summarised in Table 4.11. Since the actual phonological form of the prefix is uncertain, she refers to it as /I/ to avoid committing herself to any one phonological form. Note that there is no separate entry for the ALL-LOC2 sequence, presumably because the sequence /ši+/I/ would still be written with the *ši* sign.

One of the difficulties with the LOC2 prefix is that several of the expected forms are easily confused with other morphemes. In particular, since *i*<sub>3</sub> and *ni* are both readings of the  $\text{𐌹𐌺}$  sign, a prefix recorded as *mu-ni-* could actually consist of the *mu-* prefix followed by any of three different morpheme sequences: 1) the LOC2 prefix *i*<sub>3</sub>-, 2) the LOC prefix *ni-*, or 3) the DAT.3SG prefix *na-* followed by the LOC2 prefix.

Table 4.11: Locative 2 prefix

Context	Prefix		Query	Count
After conjugation prefix	<i>mu-i</i> <sub>3</sub>	< /mu+/I/	V"mu-i3-	0
	<i>i</i> <sub>3</sub> - <i>i</i>	< /i+/I/	V"i3-i-	0
	<i>im-mi</i>	< /imma+/I/	V-immi	897
	<i>bi</i> <sub>2</sub>	< /ba+/I/	V-bi	1866
After dative	<i>mu-e</i>	< /mu+/a+/I/	V"mu-e-"[not (@COM. 2SG)][not (@ALL. 2SG)]	328
	<i>ri</i>	< /ra+/I/	V-DAT. 2SG"ri-	0
	<i>ni</i>	< /na+/I/	n/a	n/a
After comitative	<i>di</i>	< /da+/I/	V-COM"de3-	243
	<i>de</i> <sub>3</sub>	< /da+/I/	V-COM[not (@LOC)]"di-	4
After ablative	<i>te</i>	< /ta+/I/	V-ABL"te-	22
	<i>ri</i>	< /ra+/I/	V-ABL"ri-	2

The semantic distinction between the locative and locative 2 is not always clear. Edzard (2003) is probably accurate when he describes the locative as expressing motion into or position inside, while the locative 2 expresses motion arriving at or positioned next to a reference location. Consequently, they must represent two slightly different Locative-like  $\theta$ -roles, with two slight correspondingly different applicative heads: APPL<sub>IN</sub> for the Locative and APPL<sub>AT</sub> for the Locative 2.

Both the locative and locative 2 cases have an important secondary use for expressing the logical objects of compound verbs. Since the nominal element of a compound verb appears

in the absolutive case, the actual arguments have to appear in some other case, typically the locative or locative 2. However, such “locative” noun phrases are not actually filling a Locative  $\theta$ -role. In such cases, the verb does not have the corresponding APPL<sub>IN</sub> or APPL<sub>AT</sub> head, and so it does not appear with the LOC or LOC2 prefix. When one of these prefixes is present with a compound verb, it indicates that the corresponding  $\theta$  role must be present. This contrast is shown in (4.14) with the compound verbs *gu<sub>3</sub> de<sub>2</sub>* ‘to call’ (lit. ‘to pour the voice’) and *igi du<sub>8</sub>* ‘to see’ (lit. ‘to spread the eye’). In both sentences the a logical object appears in one of the locative cases, but only the second sentence has an actual locative  $\theta$ -role and the corresponding verbal prefix.

(4.14) Applicative heads on compound verbs

*<<sup>d</sup>en-lil<sub>2</sub>-le ud-de<sub>3</sub> gu<sub>3</sub> ba-an-de<sub>2</sub>>*  
 Enlil-e ud-e gu ba-n-de  
 Enlil-ERG storm-LOC2 voice CONJ-ERG.3SG-pour  
 ‘Enlil called the storm.’ (LUr)

*<lu<sub>2</sub>-zuh-a an-bar<sub>7</sub>-a igi mu-ni-in-du<sub>8</sub>-uš>*  
 luzuh-a anbar-a igi mu-ni-n-du-eš  
 thief-LOC noon-LOC eye CONJ-LOC-ERG.3PL-spread-ERG.3PL  
 ‘They saw a thief at noon.’ (Proverbs 13)

### 4.3 Krecher’s Rule

As noted throughout the preceding section, while  $\phi$ -feature morphology is always present on the dative prefix, the  $\phi$ -features are not consistently represented on the other prefixes. In particular, as seen in Table 4.7 and Table 4.10, while the comitative and allative prefixes often have  $\phi$ -feature morphology, they often do not. One explanation, proposed by Krecher (1985) and discussed further by Attinger (1993) and Zólyomi (1999), is that for any verb, only the first of the dimensional prefixes will be marked to show  $\phi$ -feature agreement. All subsequent dimensional prefixes lack agreement marking and are restricted to 3rd-person inanimate reference. This helps to account for the apparent lack of  $\phi$ -feature agreement on some of the dimensional prefixes.

This can be seen in examples such as (4.11), where the presence of a preceding dative prefix results in the allative prefix being manifested as the bare *ši-* rather than *mši-* or *bši-*. The extent of this phenomenon in the corpus is shown in Table 4.12. The comitative prefix almost never shows  $\phi$ -feature morphology when it follows the dative prefix. Likewise, the allative prefix very rarely shows  $\phi$ -feature morphology when it follows either the dative or comitative prefixes. The exceptions to this rule, all of which are drawn from the Old Babylonian period, are rare enough that they can be attributed to errors made by a non-native speaker of the language.

Table 4.12 includes a row for the ablative case. Although Thomsen (1984) does not posit ablative-case  $\phi$ -feature morphology, Edzard (2003) proposes *bta-* and *mta-* as third-person inanimate versions of the ablative prefix. Using his definition, we find that the ablative prefix functions just the same as the others, displaying  $\phi$ -feature morphology when it appears as the first dimensional prefix, but no such morphology when it follows another dimensional prefix.

Table 4.12: Evidence from the corpus for Krecher’s Rule

Prefix	Total	After DAT	After COM
COM.1SG	442	0	
COM.2SG	256	2	
COM.3SG	509	0	
COM.3N	206	0	
ALL.1SG	96	0	0
ALL.2SG	151	0	0
ALL.3SG	236	0	5
ALL.3N	323	0	1
ABL.3N	330	0	1

One could certainly conceive of a real-world situation where one would expect to find more than one of the applicatives bearing  $\phi$ -features. For instance, it is easy enough to come up with a sentence where the combination of DAT.3SG and COM.1SG would be expected to occur (e.g. “He dedicated the statue to Ningirsu with me.”). However, due to Krecher’s Rule, such a combination could not be expressed. The situation is analogous to the Person-Case Constraint proposed by Bonet (1994), where languages have constraints against certain combinations of

$\phi$ -features. This is shown in example (4.15), which is ungrammatical because the Person-Case Constraint restricts the accusative case to being 3rd person if the dative-case element has any person features.

(4.15) Person-Case Constraint in Catalan (Bonet, 1994)

\**Me*      *li*            *ha recomanat*      *la senyora Bofill*.  
 ACC.1SG DAT.3SG has recommended the Mrs.      Bofill  
 ‘Mrs. Bofill has recommended me to him/her.’

Bonet formulates her constraint as being a restriction on accusative-case marking given the presence of  $\phi$ -features on the dative: “If DAT<sub>PERS</sub> then ACC-3rd”. However, it could equally well be expressed the other way around. That is, if the accusative element has any  $\phi$ -features beyond 3rd person, the dative element is constrained to having no  $\phi$ -features whatsoever. Formulated this way, the Person-Case Constraint is more clearly parallel to Krecher’s Rule. The exact details of the  $\phi$ -feature restrictions imposed by Krecher’s Rule are different from Person-Case Constraint, but the principle is the same: the presence of  $\phi$ -features on one of the heads places restrictions on the  $\phi$ -feature marking of the lower heads.

From a syntactic standpoint, Krecher’s Rule can be explained by the existence of an agreement head with uninterpretable  $\phi$ -features at the top of the applicative complex. When the head is spelled out, the values of those uninterpretable features are taken from the topmost applicative noun phrase. A structural representation of this is shown in Figure 4.7, which corresponds to the sentence given above as (4.11).

## 4.4 Typological Context

Given the complexity of the applicative system proposed for Sumerian, it is worth asking how exotic such a system might be from a typological standpoint. Considered relative to the various applicative systems described by Peterson (2007), Sumerian is not atypical. In particular, while Sumerian’s inventory of six applicative heads is large, none of the applicatives is particularly

unusual. As described by Peterson, languages with applicative systems generally contain at least a BEN/GOAL applicative. Locative applicatives are less common, but are found in many Bantu languages. Comitatives are also uncommon, but are found in languages such as Tepehua (Totonacan, Mexico) and Nez Perce (Sahaptian, United States). In fact, the applicative inventory of Nez Perce is roughly comparable to that of Sumerian, with applicatives for the recipient, beneficiary, directional, instrumental, allative, and comitative roles.

Peterson uses a narrow definition of applicatives, which excludes constructions such as the datives described by Pylkkänen (2002) and Cuervo (2003). In his view, applicatives serve to promote an oblique argument into a core argument position, usually the direct object position (using Relational Grammar terminology, the applicative replaces one of the verb's terms with a non-term). Under this account, the applicative argument becomes very object-like, to the extent that the verb may show object agreement with the applicative argument rather than with the actual direct object. This is not the case in Sumerian, where the presence of applicatives does not interfere with the verb's ordinary absolutive-case agreement.

In Sumerian, if the applicative argument is expressed as an overt nominal, it will receive oblique case-marking. This differs from the applicatives described by Peterson, which generally receive object-like case-marking. This would suggest that the overt nominals corresponding to applicative arguments should have null (absolutive-case) suffixes. However, the behaviour of Sumerian is closer to the Spanish datives described by Cuervo (2003), which remain marked with the dative-case clitic.

In Peterson's account, Sumerian is also unusual in having multiple applicative morphemes on a single verb, since only one oblique argument is being promoted to core argument status. In fact, this provides typological support for Krecher's Rule: the topmost applicative argument is the only one permitted to have  $\phi$ -features because it is the only one which is a core argument. The lower applicative heads indicate the presence of an argument, but not of a core argument.



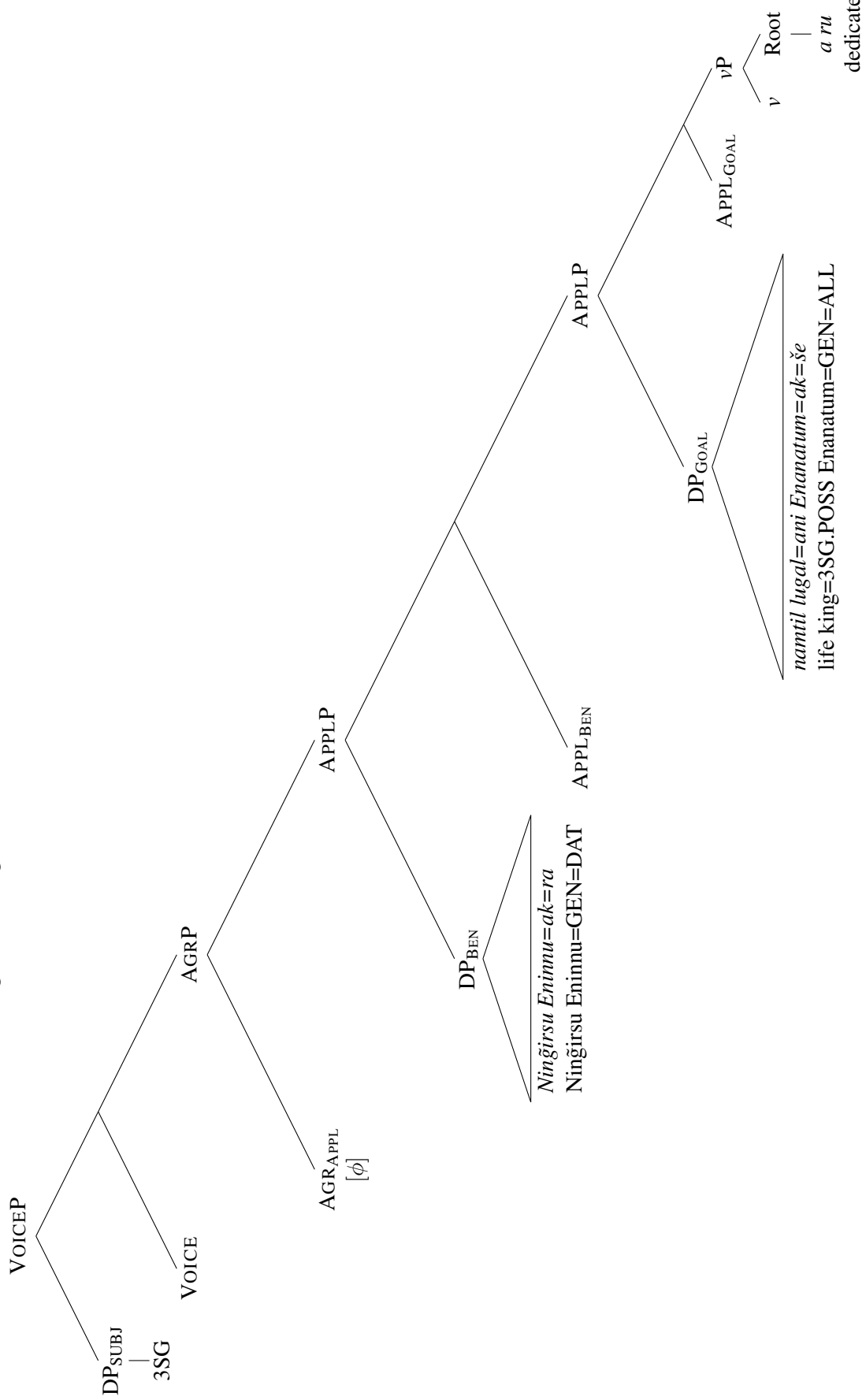
## 4.5 Summary

The end result is an extensive superstructure of applicative heads:  $APPL_{BEN}$ ,  $APPL_{COM}$ ,  $APPL_{GOAL}$ ,  $APPL_{SRC}$ , and  $APPL_{IN}$ , and  $APPL_{AT}$  which are reflected in the verb's prefix chain.

Above this complex of applicative phrases is an agreement phrase whose head has uninterpretable  $\phi$ -features which need to be matched and valued against an argument of the verb.

In traditional accounts of the dimensional prefixes, which rely on agreement between the prefix and a nominal argument with the corresponding case, there are numerous mismatches which are difficult to explain except as being exceptions. However, if these prefixes correspond to thematic roles, many of these apparent exceptions turn out to be a consequence of the relatively loose association between a noun's case clitic and the  $\theta$ -roles in the verb's argument structure.

Figure 4.7: Agreement head to account for Krecher's Rule




'He dedicated it to Ningirsu of E-ninnu for the life of his king Enanatum.'

## Chapter 5

# Conjugation Prefixes

The chain of morphemes on the Sumerian verb includes a set of prefixes which are referred to as “conjugation prefixes”. The status and function of these prefixes has long been a topic of debate in Sumerian studies. The prefixes are unlikely to have any connection to conjugation *per se*, but the term “conjugation prefix” remains in use for lack of a better term. The first methodical explanation of the conjugation prefixes was by Thureau-Dangin (1907), and numerous theories have since been proposed, with the most recent (as of this writing) being that of Woods (2008).

There is general agreement that at least three prefixes (*mu-*, *ba-*, and *i-*) are conjugation prefixes in good standing. Beyond that, there is little consensus. Scholars disagree about which prefixes are present, and also about which prefixes are basic and which ones are derived. Among the secondary conjugation prefixes which have been identified are *bi-*, *mini-*, *imma-*, *immi-*, *nga-*, *a-*, and *m-*. Often these secondary prefixes are accounted for as being etymologically derived from smaller morphemes, so *imma-* is described by Thomsen (1984) as two morphemes  $\tilde{i}+ba$ , by Falkenstein (1978) as the three-morpheme sequence  $i+b+a$ , and by Attinger (1993) as the four-morpheme sequence  $i+b+m+a$ .

The *bi-* prefix is particularly problematic. The prefix itself is quite easy to identify, appearing first on the prefix chain (usually written with the  $bi_2$   sign). The most extensive study of *bi-* is by Johnson (2004), who classifies it as an applicative morpheme. Zólyomi

(1999) classifies *bi-* as not a separate prefix at all, but rather a manifestation of the “directive” prefix *i-* preceded by an impersonal marker *b-*. Thomsen (1984) includes *bi-* as a separate prefix in her inventory of conjugation prefixes. However, the mainstream view, as expressed by Michalowski (2004) and Karahashi (2000/2005), is that the *bi-* is not itself a conjugation prefix, but consists rather of the conjugation prefix *ba-* followed by a LOC2 *i-* prefix.

While there is disagreement about the inventory of conjugation prefixes, there is even greater disagreement about their function. Theories fall into a number of different strands, which are summarised in the next section.

## 5.1 Earlier Theories

The content of this section derives largely from the summary by Woods (2008) of previous work on the conjugation prefixes given, and to a lesser extent, on the description of earlier theories by Thomsen (1984). The various theories about the conjugation prefixes can be divided into three general strands: “directional” theories, “voice” theories, and “focus” theories. Some theories share elements of different theoretical strands, and there are also some which do not fall into any of these categories.

The earliest and longest-lived theoretical strand consists of the various “directional” theories, the first of which dates to Thureau-Dangin (1907). “Directional” theories interpret the essential distinction being the relation of the event to some conceptual centre-point, with *mu-* and *ba-* being opposite ends of a spectrum and *i-* having some sort of intermediate interpretation.

The details vary, but essentially the *mu-* prefix involves motion towards a centre-point, while *ba-* does not. In the terminology typically employed in the field, *mu-* (or possibly just *m-*) has a “ventive” meaning while *ba-* (or possibly just *b-*) has a “separative” meaning.

The clearest evidence for the directional interpretation comes from contrasts such as the one involving the verb *de<sub>6</sub>* ‘to carry’ shown in (5.1). The prefix *mu-* gives *de<sub>6</sub>* the sense of ‘to

bring’, while the presence of *ba-* indicates ‘to take’. Woods (2008) provides the example in (5.1), where the same verbal root is used to describe both ends of the transaction.

(5.1) *de*<sub>6</sub> as ‘to bring’ vs. ‘to take’ (Woods, 2008)

⟨1	<i>igi-nu-du</i> <sub>8</sub>	15	<i>gin</i> <sub>2</sub> - <i>kam</i>	<b>mu-de</b> <sub>6</sub>	<i>dingir-a-mu</i>	<i>nu-kiri</i> <sub>6</sub> - <i>ke</i> <sub>4</sub>	<b>ba-de</b> <sub>6</sub> ⟩
1	<i>iginudu</i>	15	<i>gin</i> =am	<b>mu-de</b>	Dingiramu	<i>nukirik</i> =e	<b>ba-de</b>
1	<i>iginudu</i> -worker	15	shekel=COP	<b>mu-bring</b>	Dingiramu	horticulturalist=ERG	<b>ba-take</b>

‘(Uremuš) brought back 1 *iginudu*-worker<sup>1</sup>, costing 15 shekels of silver; Dingiramu, the gardener, took him away.’

This “directional” interpretation need not literally be spatial, but can be extended metaphorically, so this directionality could apply to time as well as space. For instance, Jacobsen (1965) describes *mu-* and *mi-* as referring to an event close to the speech situation, while *ba-* and *bi-* refer to an event which is remote either physically or temporally from the speech situation. To demonstrate this, he provides contrasts such as the one in (5.2). The first sentence uses *mi-* (manifested orthographically as ⟨im-mi⟩) to describe an event which occurred shortly before the speech situation, while the second sentence uses *bi-* to describe the same event, but occurs later in the same narrative.<sup>2</sup>

(5.2) Temporal reference of *m-* vs. *b-* (Jacobsen, 1965)

⟨ <i>dub</i>	<i>mul-an</i>	<i>dug</i> <sub>3</sub> - <i>ga</i>	<b>im-mi-ḡal</b> <sub>2</sub> ⟩
<i>dub</i>	<i>mulan</i>	<i>dug</i> =a	<b>m-i-ḡal</b>
tablet	heavenly.star	good=LOC	<b>near</b> -ALL-place

‘... she placed it (recently) on a tablet (with) propitious heavenly stars’ (Gudea Cyl)

⟨ <i>dub</i>	<i>mul</i>	<i>dug</i> <sub>3</sub> - <i>ga</i>	<b>bi</b> <sub>2</sub> -ḡal <sub>2</sub> -la-a⟩
<i>dub</i>	<i>mul</i>	<i>dug</i> =a	<b>b-i-ḡal-a</b>
tablet	star	good=LOC	<b>far</b> -ALL-place-SUB

‘... (she) who had placed it (a while ago) on a tablet (with) propitious stars.’ (Gudea Cyl)

<sup>1</sup>Literally, *iginudu* means ‘blind’, but in this context it would seem to refer to some unknown subtype of workers.

<sup>2</sup>Jacobsen considers the /i/ in *mi-* and *bi-* to be an allative marker. As well, he describes the *mi-* prefix as *m-mi* rather than just *mi-*. However, both these details are peripheral to the contrast here between *m-* as ‘near’ and *b-* as ‘far’.

Although contrasts like *mu-de*<sub>6</sub> vs. *ba-de*<sub>6</sub> provide evidence for a directional interpretation of the conjugation prefixes, there are other contrasts which provide evidence that the distinction might be one of voice. For instance, the verb *uš*<sub>2</sub> with the prefix *mu-* means ‘to kill’, while with the *ba-* prefix it means ‘to die’, as shown in (5.3).

(5.3) *uš*<sub>2</sub> as ‘to die’ vs. ‘to kill’

⟨*uru-az*<sup>ki</sup> **mu-hul**      *ensi*<sub>2</sub>-*bi*      **mu-uš**<sub>2</sub>⟩  
 Uruaz    **mu-hul**      ensik=bi      **mu-uš**  
 Uruaz    **ACT?**-destroy ruler=3N.POSS **ACT?**-kill  
 ‘(Eanatum) destroyed Uruaz and killed its ruler.’ (Eanatum)

⟨*unug*<sup>ki</sup>-*ga*    *lu*<sub>2</sub>      **ba-uš**<sub>2</sub>      *ur*<sub>5</sub>    **ba-sag**<sub>3</sub>⟩  
 Unug=a    lu      **ba-uš**      ur    **ba-sag**  
 Unug=LOC person **PASS?**-die liver **PASS?**-beat  
 ‘In Unug, people are dying and souls are full of distress.’ (GgHw-B)

A similar contrast is shown in (5.4), which shows two different ways of referring to the third regnal year of Amar-Suen. The “voice” theories are the second major strand of thought on the conjugation prefixes, most recently elaborated by Woods (2008), and discussed in greater detail in §5.3.

(5.4) The prefixes *mu-* and *ba-* in year names of Amar-Suen (Thomsen, 1984)

⟨*mu* <sup>d</sup>AMAR.<sup>d</sup>SUEN-*ke*<sub>4</sub>    *ur-bi-lum*<sup>ki</sup>    **mu-hul**⟩  
 mu    Amar-Suen=e      Urbilum    **mu-hul**  
 year   Amar-Suen=ERG    Urbilum    **ACT?**-destroy  
 ‘The year in which Amar-Suen destroyed Urbilum’

⟨*mu*    *ur-bi-lum*<sup>ki</sup>    **ba-hul**⟩  
 mu    Urbilum    **ba-hul**  
 year   Urbilum    **PASS?**-destroy  
 ‘The year in which Urbilum was destroyed’

It is clear from other contexts, such as (5.5), that *ba-* cannot strictly be described as a “passive” marker, since it occurs in what appears to be a straightforward active-voice sentence. Still, the fact that the active *mu-* is so often contrasted with a passive-like *ba-* does suggest that the conjugation prefixes are somehow connected to voice.

(5.5) Active sentence with *ba-* prefix.

$\langle lu_2$ -IM *kug* *ba-an-zuh*  $lu_2$  *gen*<sub>6</sub>-*na* *giḡ*<sub>4</sub>  
 lu’IM *kug* *ba-n-zuh* *lu* *gen* *giḡ*  
 criminal silver **ba**-ERG.3SG-steal man honest weight  
*mu-ni-in-ak-de*<sub>3</sub>  
 mu-ni-n-ak-ed  
 mu-LOC-ERG.3SG-do-FUT  
 ‘The dishonest man stole silver, the honest man will earn his pay.’ (Proverbs 13)

The third main strand of thought tries to explain the conjugation prefixes in terms of focus. The most recent proponent of the position that the conjugation prefixes control focus is Vanstiphout (1985), who proposes that the essential contrast is that *mu-* focusses on the person, while *ba-* is focussed on the locus of the event. In his account, the neutral *i-* prefix acts as a backgrounding device, with no associated focus. His analysis is summarised in Table 5.1.

Table 5.1: Features of conjugation prefixes (Vanstiphout, 1985)

	+focus		–focus
<i>mu-</i>	+person	–locus	<i>i-</i>
<i>ba-</i>	–person	+locus	

What “focus” actually means may vary from scholar to scholar. So for instance, while explaining the contrast between *mu-* and *i-* in terms of topicality, Yoshikawa (1979) includes social class as one of the factors in determining the choice of conjugation prefix. This can be justified on the strength of a well-known administrative text from Lagaš detailing the exchange of gifts between the wives of the rulers of Adab and Lagaš. In this text, with relevant excerpts in (5.6), the actions of giving gifts to the lady of Lagaš are marked with *mu-*; actions where gifts are given to the lady from the lesser city are marked with *i-*.

(5.6) Contrast between *mu-* and *i-* (Woods, 2008)<sup>3</sup>

⟨(various gifts) *nin-giškim-til*<sub>3</sub> *dam ensi*<sub>2</sub> *adab*<sup>ki</sup>-*ka-ke*<sub>4</sub> *barag-nam-tar-ra*  
 (various gifts) Ningiškimtil *dam ensik* Adab=ak=ak=e Baragnamtara  
 (various gifts) Ningiškimtil *wife ruler* Adab=GEN=GEN=ERG Baragnamtara  
*dam lugal-an-da ensi*<sub>2</sub> *lagaš*<sup>ki</sup> *2-kam-ma-ka* *šu mu-na-taka*<sub>4</sub>⟩  
*dam Lugalanda ensik Lagaš*=ak=ak *2-kam*=ak=a *šu mu-na-taka*  
*wife Lugalanda ruler Lagaš*=GEN=GEN *2-ORD*=GEN=LOC **mu**-DAT.3SG-send  
 ‘On the second (delivery), Ningiškimtil, the wife of the ruler of Adab, sent (various gifts) to Baragnamtara, wife of Lugal-Anda, ruler of Lagaš.’

⟨*a-ne-da-nu-me-a lu*<sub>2</sub>-*ni* *ma-al-ga-sud-da mu-da-ḡen-na-a* **mu-de**<sub>6</sub>⟩  
 Anedanumea *lu*=ni *Malgasuda mu-da-ḡen-a* **mu-de**  
 Anedanumea *person*=3SG.POSS *Malgasuda mu-COM.3SG-go-SUB* **mu-bring**  
 ‘Anedanumea, her servant, who came with Malgasud, delivered them.’

⟨(a garment) *nin-giškim-til*<sub>3</sub>-*e* *ma-al-ga mu-na-sum*⟩  
 (a garment) Ningiškimtil=e *Malga mu-na-sum*  
 (a garment) Ningiškimtil=ERG *Malga mu-DAT.3SG-give*  
 ‘Ningiškimtil gave (a garment) to Malga(sud).’

⟨(metals) *barag-nam-tar-ra dam lugal-an-da ensi*<sub>2</sub> *lagaš*<sup>ki</sup>-*ka-ke*<sub>4</sub>  
 (metals) Baragnamtara *dam Lugalanda ensik Lagaš*=ak=ak=e  
 (metals) Baragnamtara *wife Lugalanda ruler Lagaš*=GEN=GEN=ERG  
*2-kam-ma-ka nin-giškim-til*<sub>3</sub> *dam ensi*<sub>2</sub> *adab*<sup>ki</sup>-*ka-ra* *šu e-na-taka*<sub>4</sub>⟩  
*2-kam*=ak=a *Ningiškimtil dam ensik* Adab=ak=ak=ra *šu e-na-taka*  
*2-ORD*=GEN=LOC *Ningiškimtil wife ruler* Adab=GEN=GEN=DAT **e**-DAT.3SG-send  
 ‘On the second (delivery), Baragnamtara, the wife of Lugal-Anda, ruler of Lagaš, sent (metals) to Ningiškimtil, wife of the ruler of Adab.’

⟨*ma-al-ga e-da-ḡen*⟩  
 Malga **e**-da-ḡen  
 Malga **e**-COM.3SG-go  
 ‘Malga(sud) came with her (Anedanumea).’

⟨(garments and scented oil) *barag-nam-tar-ra a-ne-da-nu-me-a e-na-sum*⟩  
 (garments and scented oil) Baragnamtara *Anedanumea e-na-sum*  
 (garments and scented oil) Baragnamtara *Anedanumea e-DAT.3SG-give*  
 ‘Baragnamtara gave (garments and scented oil) to Anedanumea.’

<sup>3</sup>*šu taka*<sub>4</sub> = ‘to send’. Also note that in this dialect, vowel harmony causes the *i-* prefix to become *e-* before a [–ATR] vowel (Smith, 2007b).



Of course, the contrast displayed in (5.6) has also been used as evidence to support a “directional” interpretation of the conjugation prefixes (Thureau-Dangin, 1907). The author of the text is in Lagaš, so gifts coming to Lagaš would naturally be marked with the ventive *mu-*, while gifts going away from Lagaš are marked with *i-*.

There are other scholars, most notably Falkenstein (1978), who claim that the conjugation prefixes have no intrinsic semantic function. In Falkenstein’s view, their primary purpose is to index nominal arguments of the verb. The *mu-* prefix tends to occur with persons, while *i-* tends to be associated with non-persons. In his system, the *ba-* prefix is not a conjugation prefix at all; Falkenstein is one of those scholars who likes to slice morphemes as finely as possible, so *ba-* is accounted for as an inanimate marker *b-* combined with a locative *a-*.

## 5.2 Michalowski 2004

For a recent overview of the subject, we can turn to Michalowski (2004), whose capsule summary of Sumerian syntax stresses repeatedly that the interpretation of the conjugation prefixes remains extremely contentious. He describes his own position as “minimalist”, in that he recognises only three basic conjugation prefixes: *mu-*, *ba-*, and *i-*, with *imma-* being a derived prefix.

Michalowski’s interpretation of the prefixes falls into the general “focus” school. He argues that the prefixes serve to delineate the amount of control that the agent has over the event. Thus, *mu-* marks “focus on control over an action that is within the control and propinquity of the agent”. For actions with a lesser degree of control, *ba-* is employed, which explains why *ba-* can sometimes have a passive-like interpretation. The remaining basic prefix, *i-*, is neutral.

In Michalowski’s view, *imma-* is a reduplicated form of *mu-*. This reduplication serves to “intensify” the focus, so *imma-* is used in cases such as with movement towards the agent or manipulation of the object by the agent. Unfortunately this position is only sketched out in a book chapter which summarises all aspects of Sumerian, so Michalowski provides no examples

to clarify exactly what “intensify” means in this context.

Example (5.7) shows a three-way contrast involving the verb *dab*<sub>5</sub> ‘to hold, to seize’ which might help to clarify Michalowski’s position. In the example with *mu-*, the emphasis is on the agent, king Enšakušana. In the second example, *imma-* is used to give additional emphasis to the action of Aruru’s grasping of the hand. In the final example, *ba-* is used because the subject is a turtle, which is less agentive than the other participant in the action, the god Ninurta.

(5.7) Contrast of *mu-*, *imma-*, and *ba* with *dab*<sub>5</sub> ‘to seize’

⟨*en-bi*<sub>2</sub>-*iš*<sub>8</sub>-*tar*<sub>2</sub> *lugal kiš*<sup>ki</sup> **mu-dab**<sub>5</sub>⟩  
 Enbi-Ištar      lugal    Kiš    **mu-dab**  
 Enbi-Ištar      king     Kiš    **mu-seize**  
 ‘He captured Enbi-Ištar, the king of Kiš.’ (Enšakušana)

⟨*šu-ni*                    **im-ma-an-dab**<sub>5</sub>                    <sup>d</sup>*a-ru-ru eš*<sub>3</sub>-*mah-še*<sub>3</sub>⟩  
 šu-ni                    **imma-n-dab**                    Aruru    Ešmah-še  
 hand-3SG.POSS **imma-ERG.3SG-seize**    Aruru    Ešmah-ALL  
*mi-ni-in-kar*⟩  
 mi-ni-n-kar  
 mu-LOC-ERG.3SG-remove  
 ‘Aruru grasped her by the hand and led her away into the Eš-mah’ (EnlSu)

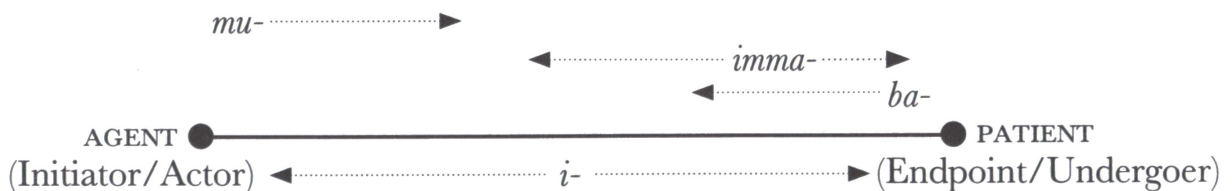
⟨*ba-al-gu*<sub>7</sub> *eđer-ra-ni*                    *sa-bi*                    **ba-da-an-dab**<sub>5</sub>⟩  
 balgu                    eđer-ani                    sa-bi                    **ba-da-n-dab**  
 turtle                    back-3SG.POSS    sinew-3N.POSS    **ba-COM.3SG-ERG.3SG-seize**  
 ‘The turtle seized his (Ninurta’s) tendon from behind him.’ (NinTrtl)

### 5.3 Woods 2008

The most recent work on the conjugation prefixes is by Woods (2008), which describes *mu-*, *i-*, *imma-*, and *ba-* as voice markers which serve to distinguish the speaker’s perspective on the event. According to Woods, rather than distinguishing between active and passive voices, Sumerian provides morphemes which make voice distinctions on a spectrum running from the active, Agent-focussed *mu-*, through a middle voice *imma-* which “opens up” the structure

of the event, to the medio-passive endpoint-focussed *ba-*. The fourth prefix, *i-*, he describes as being neutral with respect to voice, and is typically used for backgrounding information. Woods summarises the range of the prefixes in the diagram included here as Figure 5.1.

Figure 5.1: Organisation of primary prefixes according to prototypical usage (Woods, 2008)



The *mu-* prefix expresses what Woods calls the “marked” active voice, and it focusses on the initiator of the event, corresponding to situations which are prototypically transitive, with associations of control, agency, animacy, volition, and emphasis on the affectedness of the object (rather than the subject). Prototypically transitive examples of *mu-* with high transitivity are shown in (5.8).

(5.8) Prototypically transitive examples of *mu-* (Woods, 2008)

<i>&lt;uru-az<sup>ki</sup></i>	<b>mu-hul</b>	<i>ensi<sub>2</sub>-bi</i>	<b>mu-uš<sub>2</sub></b>
Uruaz	<b>mu-hul</b>	ensik=bi	<b>mu-uš</b>
Uruaz	<b>mu-destroy</b>	ruler=3SG.POSS	<b>mu-kill</b>

‘(Eanatum) destroyed Uruaz and killed its ruler’ (Eanatum)

<i>&lt;ama-ḡeštín-ra</i>	<i>e<sub>2</sub></i>	<i>saḡ-ub<sub>x</sub>-ka-ni</i>	<b>mu-na-du<sub>3</sub></b>
Ama-ḡeštín=ra	e	Saḡub=ak=ani	<b>mu-na-du</b>
Ama-ḡeštín=DAT	temple	Saḡub=GEN=3SG.POSS	<b>mu-DAT.3SG-build</b>

‘For Amaḡeštín he built her temple of Saḡub.’ (Eanatum)

The prefixes *imma-* and *ba-* are both described by Woods as middle voice markers. He calls *imma-* a “middle” marker while *ba-* is a “medio-passive” marker. For the most part, they cover much of the same semantic range, both functioning in contexts such as those listed in Table 5.2.

Table 5.2: Contexts for middle prefixes (Woods, 2008)

<i>imma-</i>	<i>ba-</i>
Body-action events	
Motion events	
Self-benefactive events	
Mental/emotion events	
	Spontaneous events
	Passives

According to Woods, the distinction is that *imma-* represents an elaboration of the event, focussing more on its internal structure. Thus *imma-* is typically used rather than *ba-* for events which are plural or collective. But *imma-* also has increased emphasis, as can be seen by the contrast between *šu imma+ti* ‘to seize’ vs. the less agentive *šu ba+ti* ‘to receive’.

While *ba-* and *imma-* do overlap to a large extent, the range of *ba-* extends farther in the non-agentive dimension. This means that *ba-* is used for spontaneous events and for true passives. As well, Woods claims that *ba-* prefix focusses on the “set-in-motion” phase of the event, which is the reason why verbs with *ba-* often have a separative sense. In this manner, Woods manages to incorporate the entire “directional” theory of the conjugation prefixes. The “directional” semantics are merely a side effect of the “set-in-motion” semantics associated with *ba-*. In Woods’s analysis, the distinction between *mu-de<sub>6</sub>* ‘to bring’ and *ba-de<sub>6</sub>* ‘to take away’ seen in (5.1) is not fundamentally one of directionality. Rather, *mu-* focusses on the event as a whole while *ba-* focusses on the “set-in-motion” phase of the event. Woods makes a similar argument in favour of the distinction between *ni<sub>2</sub> mu+te* ‘to be afraid’ and *ni<sub>2</sub> ba+te* ‘to become fearful’, as also being a consequence of the “set-in-motion” focus of *ba-*.

The semantic shifts created by the conjugation prefixes are often quite subtle, so it is worth examining some of the three-way contrasts which occur. The most straightforward are ones such as the ‘kill’ vs. ‘die’ contrast which was shown in (5.3), where the *mu-* form is active and

transitive while the *ba-* form is passive and intransitive.<sup>4</sup>

One verb which does occur with all three prefixes is *mu*<sub>2</sub> ‘to grow’. The contrast can be seen in (5.9). In the first sentence, the ruler is causing the temple to grow. In the second sentence, the goddess Ninhursaġa is growing plants as part of her revenge against the god Enki, and *imma-* is used to stress the self-benefactive nature of the event. In the third sentence, grass is growing unattended after the fall of Akkad.

(5.9) Three-way contrast involving *mu*<sub>2</sub> ‘to grow’ (Woods, 2008)

⟨*ensi*<sub>2</sub>-*ke*<sub>4</sub> *e*<sub>2</sub> *mu-du*<sub>3</sub> **mu-mu**<sub>2</sub> *kur* *gal-gin*<sub>7</sub> **mu-mu**<sub>2</sub>⟩  
 ensik=e e mu-du **mu-mu** kur gal=gin **mu-mu**  
 ruler=ERG temple *mu*-build **mu**-grow mountain great=EQU **mu**-grow  
 ‘The ruler built the temple; he made it grow; he made it grow like a great mountain.’  
 (Gudea Cyl)

⟨<sup>u</sup><sub>2</sub>[*am-ha-ru*] **im-ma-an-mu**<sub>2</sub>⟩  
 amharu **imma**-n-mu  
*amharu* **imma**-ERG.3SG-grow  
 ‘She grew the *amharu*-plant (for her own use).’ (EnkNh)

⟨*gu*<sub>2</sub> <sup>giš</sup>*ma*<sub>2</sub> *gid*<sub>2</sub>-*da* *id*<sub>2</sub>-*da-ba* *u*<sub>2</sub> *gid*<sub>2</sub>-*da* **ba-an-mu**<sub>2</sub>⟩  
 gu ma gid=a id=bi=a u gida **ba-mu**  
 bank boat tow=SUB canal=POSS.3N=LOC grass long **ba**-grow  
 ‘On its canal banks, where boats were towed, the grass grows long.’ (CAk)

Although *mu-* is generally associated with prototypically transitive events, it can occur with a low-transitivity verb like *zu* ‘to know’ to indicate that the Agent (or perhaps more accurately, the Experiencer) has a degree of control over the act of knowing, as in the first sentence of (5.10). The second and third sentences show how the presence of the *imma-* and *ba-* prefixes reduces the level of control. In Woods’s view, *mu-zu* indicates “static” state of knowing (i.e. the event of knowing as-a-whole), while *ba-* focusses on the “dynamic and inceptive phase of

<sup>4</sup>The verb *uš*<sub>2</sub> occurs only once in the corpus with the *imma-* prefix, in the phrase *ur*<sub>5</sub> *im-ma-uš*<sub>2</sub> ‘people died’. The sense is passive but the collective nature of the direct object prompts the use of *imma-* rather than *ba-*.

the event”. Thus *ba-zu* often has the meaning of “become known” or “learn”; *imma-zu* has a similar meaning, but adds a certain degree of intensity.

(5.10) Contrasts involving *zu* ‘to know’ (Woods, 2008)

$\langle {}^d\text{gilgameš}_2\text{-gin}_7 \text{ zid-du} \quad \mathbf{mu-zu} \quad \text{erim}_2\text{-du} \quad \mathbf{mu-zu} \rangle$   
 Gilgameš-gin ziddu **mu-zu** erimdu **mu-zu**  
 Gilgameš-EQU righteous **mu-know** wicked **mu-know**

‘Like Gilgameš, I can recognise the righteous and I can recognise the wicked.’ (Šulgi C)

$\langle \tilde{g}e_{26}\text{-e} \quad \mathbf{im-ma-zu-a} \quad ni_2 \quad \mathbf{im-ma-an-zu-a} \rangle$   
 $\tilde{g}e=e \quad \mathbf{im-ma-zu-a} \quad ni \quad \mathbf{im-ma-zu-a}$   
 1SG=ERG **imma-know-SUB** fear **imma-know-SUB**

‘I, who have experienced (lit. come to know), who have experienced fear!’ (Ur-Namma A)

$\langle ki \quad \text{inim-ma-ka} \quad \text{nam-gu}_5\text{-li} \quad \mathbf{ba-an-zu-zu} \rangle$   
 ki inim=ak=a namguli **ba-n-zu-zu**  
 place word=GEN=LOC friendship **ba-ERG.3SG-know IPFV**

‘At the place of testimony, friendship becomes known.’ (Proverbs Ur)

Like *zu* ‘to know’, the compound verb *igi bar* ‘to look at’ is semantically low in transitivity. Since the event of looking requires the subject’s volition, and since the subject is not usually affected by the event of looking, *igi bar* generally occurs with the prefix *mu-*. However, in example (5.11), *imma-* is used to indicate that the subject is emotionally or mentally affected by the act of seeing. In this case, upon seeing Ninlil by the bank of the river, Enlil is consumed by the desire to rape her. For this reason, Woods proposes that here ‘eyed’ would be a better translation than merely ‘look at’.

(5.11) Use of *imma-* to indicate subject affectedness (Woods, 2008)

$\langle [\text{sipad} \quad n]a\text{-a}\tilde{g}_2 \text{ tar-tar-re} \quad \text{igi} \quad \text{kug-ga-am}_3 \quad \text{igi} \quad \mathbf{im-ma-ši-in-bar} \rangle$   
 sipad naḡ tar-tar-a igi kuga-COP igi **imma-ši-n-bar**  
 shepherd fate cut IPFV-SUB eye bright eye **imma-ALL-ERG.3SG-look**  
 ‘The shepherd who determines destinies, the bright-eyed one, eyed her there.’ (Enlil)

At the farthest extreme, the examples in (5.12) are actually quite intransitive, both syntactically and semantically, involving verbs like *til*<sub>3</sub> ‘to live’ and *ḡal*<sub>2</sub> ‘to be’. This seems rather at

odds with the notions of control, animacy, and agency which are normally associated with the *mu-*, but Woods argues that in such cases *mu-* serves to emphasise the initiator of the event, or to “simply [underscore] the verbally denoted action or state”. Since *mu-* emphasises the entire event, while *ba-* focusses only on the “set-in-motion” phase, it is understandable that static events such as the ones shown in (5.12) would employ *mu-* rather than *ba-*.

(5.12) Examples of *mu-* providing emphasis on initiator/action-as-a-whole (Woods, 2008)

⟨20 *la*<sub>2</sub> 3 *giš-ur*<sub>3</sub> <sup>*giš*</sup>*til-lu-ub*<sub>2</sub> *kiri*<sub>6</sub> *e*<sub>2</sub>-*ku-ka* **mu-*ḡal***<sub>2</sub>⟩  
 20 *la* 3 *gišur* *tilub* *kirik* *Eku=ak=a* **mu-*ḡal***  
 20 minus 3 plank Oriental.plane.tree garden *Eku=GEN=LOC* **mu-be**  
 ‘17 planks of Oriental plane tree wood were available in the garden of Eku.’

⟨*a-bu-ni* *kaskal-a* **mu-*til***<sub>3</sub>-*la-am*<sub>3</sub> *bi*<sub>2</sub>-*dug*<sub>4</sub>⟩  
 Abuni *kaskal=a* **mu-til-a-am** *bi-dug*  
 Abuni journey=LOC **mu-live-SUB-COP** CONJ-say  
 ‘He declared that Abuni was, in fact, on a journey.’

As it turns out, despite their lack of transitivity, both *ḡal*<sub>2</sub> and *til*<sub>3</sub> are more likely to appear in the corpus with *mu-* than with any of the other prefixes, as shown in Table 5.3. Furthermore, it is not clear that the emphasis on the initiator or the action is any different for *ḡal*<sub>2</sub> and *til*<sub>3</sub> when they appear with *imma-* or *ba-* rather than *mu-*. This remains one aspect of Woods’s argument which is not entirely convincing.

Table 5.3: Cooccurrences of *ḡal*<sub>2</sub> and *til*<sub>3</sub> with conjugation prefixes

	% <i>mu-</i>	% <i>imma-</i>	% <i>ba-</i>
<i>ḡal</i> <sub>2</sub> ‘to be’	24.2	4.0	3.6
<i>til</i> <sub>3</sub> ‘to live’	17.4	2.1	1.4

The fourth conjugation prefix, *i-*, stands outside the the system articulated by the prefixes *mu-*, *imma-*, and *ba-*. The *i-* prefix serves to neutralise the voice distinctions which can be provided by the other prefixes.

In this respect, Woods is following the earlier work of Vanstiphout (1985), who describes *i-* as having a backgrounding or defocussing function. In general, *i-* serves to provide additional information, secondary to the events described using *mu-*. One use of *i-* is in subordinate clauses which provide background information. In the same vein, *i-* is often used in royal inscriptions for the actions of the king’s enemies, in contrast with the king’s own actions which are prefixed with *mu-*.

It has long been noted that when the verb has dative-case arguments, the choice of the prefixes *mu-*, *i-*, and *ba-* correlates with the position of the recipient on the Nominal Hierarchy. That is, *mu-* is always present with a DAT.1SG recipient, usually present with DAT.2SG prefix, and often present with DAT.3SG. The *i-* prefix is never found with DAT.1SG recipients, is rare with DAT.2SG, and is often found with DAT.3SG. If *ba-* is found with a dative at all, it is only with 3rd person, and even that is very rare. This is summarised in Table 5.4. Under Woods’s account, this sort of correlation is only natural, since *mu-* will correlate best with recipients whose animacy is more salient (i.e. higher on the Nominal Hierarchy).

Table 5.4: Cooccurrence of conjugation prefixes with dative arguments

	<i>mu-</i>	<i>i-</i>	<i>imma-</i>	<i>ba-</i>
DAT.1SG	677	0	0	0
DAT.2SG	426	12	2	7
DAT.3SG	1519	39	17	32

While Woods’s account of the conjugation prefixes is not without its small flaws, it still represents a significant advance over any earlier theory of the prefixes. In particular, by identifying the “set-in-motion” semantics of the *ba-* prefix, Woods is able to subsume the competing strand of “directional” theories within a coherent voice-based theory of the conjugation prefixes. The task of the next section is to take Woods’s descriptive account and place it into a theoretical framework.



## 5.4 Are Conjugation Prefixes a System of Voice?

Woods (2008) musters over 300 exemplar sentences in support of his claim that the conjugation prefixes represent a system of voice. Given that the corpus provides a considerably broader sample-set, it is worth considering how well Woods's claims about the conjugation prefixes stand up in the light of that data.

As described by Woods, the *mu-* prefix tends to occur in situations where the agent is emphasised, and the action is prototypically transitive. With the aid of the query objects described in §3.6 one can get a sense of the sorts of verbal roots which tend to cooccur with *mu-*. Table 5.5 shows the top ten verbal stems identified by the *V-mu* object, ranked by the proportion of cooccurrences with the *mu-* prefix relative to all occurrences of the verb. So for instance, the verb *ru* (in the compound *a ru* 'to dedicate') occurs 133 times in the corpus, and 131 of those occurrences are with the *mu-* prefix. For purposes of comparison, the percent of occurrence with the other conjugation prefixes is also shown. To avoid having rare verbs skew the data, this table (and subsequent ones like it) only consider verbal roots which occur at least 20 times in the corpus. The totals do not add to 100% because the table includes neither the prefix *i-* nor the large numbers of participles and infinitives in the corpus.

Table 5.5: Most frequent verbs for *mu-*

Verb	% <i>mu-</i>	% <i>imma-</i>	% <i>ba-</i>
<i>ru</i> 'to dedicate'	98.5	0.0	0.0
<i>du</i> <sub>12</sub> 'to perform (music)'	73.0	2.7	0.0
<i>ba-al</i> 'to dig'	62.9	0.0	5.7
<i>hur</i> 'to scratch, draw'	61.8	0.0	5.7
<i>dun</i> 'to dig'	61.4	6.8	6.8
<i>kug</i> 'to purify, clean'	53.6	14.3	0.0
<i>du</i> <sub>3</sub> 'to build'	51.9	2.2	4.2
<i>de</i> <sub>2</sub> 'to pour' (incl. <i>gu</i> <sub>3</sub> <i>de</i> <sub>2</sub> 'to call')	44.6	5.0	11.9
<i>dim</i> <sub>2</sub> 'to create'	42.7	9.4	9.8
<i>šum</i> <sub>2</sub> 'to give'	42.2	4.3	14.5
All verbs	16.8	4.0	7.9

As can be seen from Table 5.5, the verbs which cooccur most commonly with *mu-* are indeed all canonically transitive verbs where the agent role is clearly prominent. In particular, royal actions such as building temples, dedicating statues, and digging canals are almost always associated with *mu-*. It is also worth noting that of the three prefixes shown in Table 5.5, *mu-* is by far the commonest. In fact, *i-* is commoner still (at 18.1% of all verb forms), but it does not participate in the distinction between *mu-*, *imma-*, and *ba-*.

Table 5.6 shows the ten verbs with the highest tendency to cooccur with the *imma-* prefix. As predicted by Woods, there are a large number of body-action and reflexive events (rubbing, bathing, dressing, and cleaning oneself). The verb *ti* appears as both a motion verb ‘to approach’ and as the compound *šu imma+ti* mentioned in the previous paragraph. The verb *tag* most commonly appears as *ḡiṣ tag* ‘to sacrifice’, which clearly falls into Woods’s self-benefactive category. The verb *sal* ‘to be thin’ at first appears out of place, but in the contexts where it appears with *imma-*, it involves the process of thinning out grain, so *imma-* is expected due to the collective nature of the direct object.

Table 5.6: Most frequent verbs for *imma-*

Verb	% <i>imma-</i>	% <i>ba-</i>
<i>su-ub</i> ‘to rub’	22.0	1.7
<i>lu</i> <sub>3</sub> ‘to mix’	20.0	5.0
<i>tu</i> <sub>5</sub> ‘to bathe’	20.0	6.7
<i>sal</i> ‘to be thin’	16.7	6.1
<i>mur</i> <sub>10</sub> ‘to dress’	16.1	19.4
<i>tag</i> ‘to touch’	15.9	12.1
<i>ti</i> ‘to approach’	14.7	29.0
<i>kug</i> ‘to purify, clean’	14.3	0.0
<i>šeš</i> <sub>2</sub> ‘to weep’	11.9	13.9
<i>dab</i> <sub>5</sub> ‘to seize’	11.7	15.7
All verbs	4.0	7.9

The contrast between Table 5.7 and Table 5.6 is instructive. The only verb which also appears in both tables is *ti* ‘to approach’. Verbs of motion are represented in both tables, but verbs like washing and rubbing, which involve intentional activities, tend to occur with *imma-*

rather than with *ba-*. The range of *ba-* extends beyond *imma-* to contexts which are truly agentless, such as *su* ‘to be flooded’ and *uš<sub>2</sub>* ‘to die’. The appearance of *sag<sub>3</sub>* ‘to beat’ in this list at first seems odd, but its occurrence with *ba-* can be explained by the occurrence of two compounds *g<sup>iš</sup>tukul sag<sub>3</sub>* ‘to be defeated’ (literally, ‘to beat the weapons’) and *šag<sub>4</sub> sag<sub>3</sub>* ‘to be depressed’ (‘to beat the heart’), both of which tend to appear as passives.

Table 5.7: Most frequent verbs for *ba-*

Verb	% <i>ba-</i>	% <i>imma-</i>
<i>su</i> ‘to sink, to be flooded’	48.4	3.1
<i>gib</i> ‘to lie across (obstructively)’	40.9	4.5
<i>u<sub>5</sub></i> ‘to ride’	40.3	2.5
<i>ti</i> ‘to approach’	29.0	14.7
<i>de<sub>6</sub></i> ‘to bring’	28.2	2.6
<i>kar</i> ‘to flee’	26.8	5.5
<i>uš<sub>2</sub></i> ‘to die’	26.0	0.6
<i>nu<sub>2</sub></i> ‘to lie down’	24.1	1.2
<i>sag<sub>3</sub></i> ‘to beat’	23.9	4.0
<i>ha-za</i> ‘to hold’	23.3	0.0
<i>ze<sub>2</sub>-er</i> ‘to tear out’	20.0	8.9
All verbs	7.9	4.0

As seen in Table 5.8, *i-* tends to occur on a wide range of verbal stems, rather than having any tendency to cooccur with only certain types of verbs. This is not unexpected, since *i-* covers the entire range of contexts shown in Figure 5.1.

Thus, the corpus data does provide support for Woods’s position that the conjugation prefixes represent a system of voice. However, there is one significant difficulty with this analysis, which is that while the *imma-* and *ba-* prefixes may have the semantics associated with middle-voice constructions in other languages, the prefixes do not behave syntactically like middles.

The syntactic framework employed for this analysis is based on the inventory of *vP* shells proposed by Cuervo (2003). In her taxonomy, *v* heads serve as event introducers, and these *v* heads come in three different basic flavours, as shown in Table 5.9. These basic *v* heads can in turn be combined to create composite bieventive structures such as causatives. Of the

Table 5.8: Most frequent verbs for *i-*

Verb	% <i>i-</i>	% <i>mu-</i>	% <i>imma-</i>	% <i>ba-</i>
<i>silig</i> ‘to be forceful’	81.3	0.0	0.0	9.4
<i>še</i> ‘to agree’	68.2	3.0	3.0	3.0
<i>me</i> ‘to be’	54.7	1.3	2.6	0.2
<i>šum</i> ‘to slaughter’	50.0	5.0	0.0	15.0
<i>i</i> ‘to bring out’	45.8	20.6	3.8	4.6
<i>ak</i> ‘to do’	44.5	15.3	3.0	7.0
<i>sam</i> <sub>2</sub> ‘to buy, to barter’	44.3	3.3	0.0	1.6
<i>tuku</i> ‘to have’	43.2	6.5	1.4	8.3
<i>ku</i> ‘to lay down’	39.2	17.6	2.0	2.0
<i>tud</i> ‘to give birth, to fashion’	38.2	17.2	4.0	2.2
All verbs	18.1	16.8	4.0	7.9

three basic types of heads, only  $v_{DO}$  projects a VOICE head, so it is the only one which has a structural position available for an Agent argument.

Table 5.9: Taxonomy of event introducers (Cuervo, 2003)

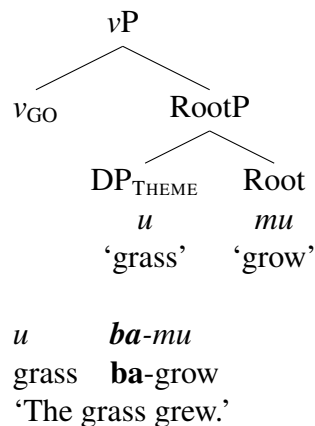
	Event	Examples
$v_{DO}$	activities	dance, sweep, run
$v_{GO}$	changes	fall, go, die, grow (intr.)
$v_{BE}$	states	like, admire, lack
$v_{DO}+v_{DO}$	causatives	make wash, make laugh
$v_{DO}+v_{GO}$	causatives	make fall, make grow
$v_{DO}+v_{BE}$	causatives	break, burn, close
$v_{GO}+v_{BE}$	inchoatives	break (intr.), burn (intr.), close (intr.)

We will start with the simplest contrast, the straightforward equation of *mu-* with active and *ba-* with (medio)passive. This contrast can be seen in sentences such as the examples involving *mu*<sub>2</sub> ‘to grow’ (5.9) and *uš*<sub>2</sub> ‘to kill, die’ (5.3).

Starting with the simplest case, that of a simple unaccusative verb, Figure 5.2 shows the syntactic tree for the verb *mu*<sub>2</sub> ‘to grow’ in its unaccusative sense with *ba-*. The event is an unaccusative event of change, represented in Cuervo’s framework by a  $v_{GO}$  head, which does

not project a VOICE phrase. Since there is no VOICE projection, there is nowhere for an Agent argument to be merged.

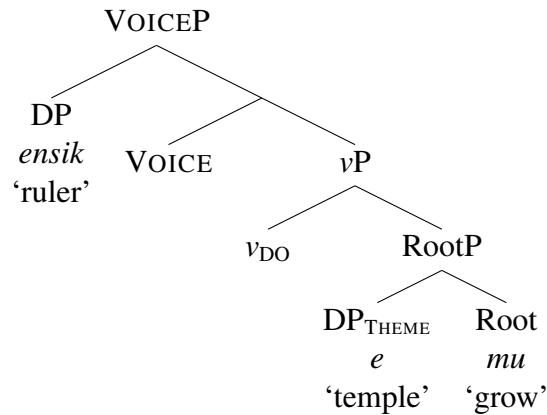
Figure 5.2: Tree for unaccusative  $mu_2$



When an overt Agent is added, this is represented by using a transitive light verb,  $v_{DO}$ , in place of the intransitive  $v_{GO}$ , as in Figure 5.3. The Root itself is agnostic about whether it is transitive or intransitive, and it is the syntactic structure into which it is merged which determines the Root’s argument structure. It is the choice of a  $v_{DO}$  rather than a  $v_{GO}$  which causes a VOICE phrase to be projected, and thus provides a location to merge an Agent argument.

The distinction between the unaccusative *ba-* and the transitive (*mu-*, *imma-*, and *i-*) variants of  $mu_2$  can easily be accounted for by the structural difference between Figure 5.2 and Figure 5.3. However, the contrast between *mu-*, *imma-*, and *i-* is difficult to account for by appealing only to structural differences. All three of these prefixes occur with an Agent argument, and the sentences involving *imma-* and *i-* would also be structurally represented by Figure 5.3. The contrasts introduced by using *imma-* or *i-* rather than *mu-* are at a discourse level, not represented in this syntactic structure.

Although Woods (2008) refers to *imma-* as a middle marker, traditional analyses of middle constructions are not much help in finding a structural explanation. As described by Schäfer (2007), middles crosslinguistically share the following characteristics:

Figure 5.3: Tree for transitive  $mu_2$ 

*ensik*=*e*      *e*      ***mu***-*mu*  
 ruler=ERG    temple    **mu**-grow  
 'The ruler grew the temple.'

A: The subject of the sentence corresponds to the internal argument (the understood or notional object).

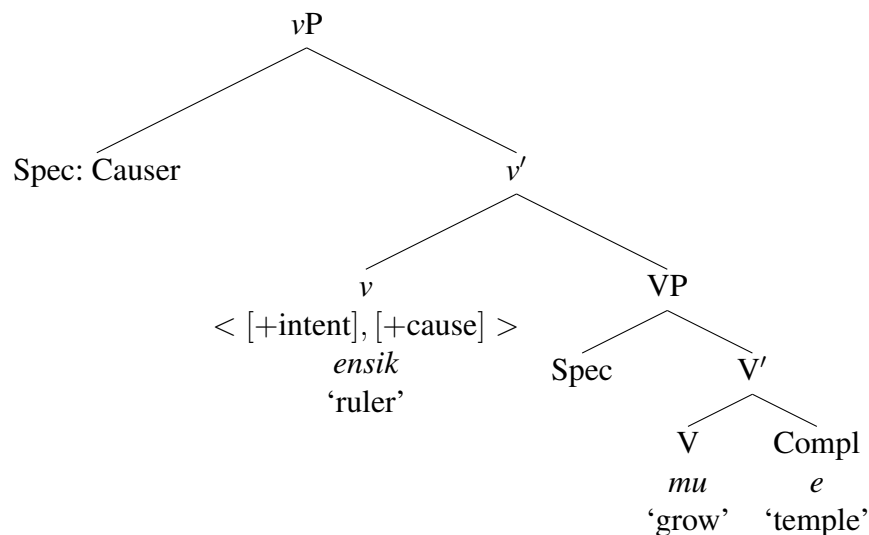
B: The agent is demoted and receives an arbitrary interpretation.

C: The interpretation of the sentence is non-episodic. Middles do not make reference to an actual event having taken place; rather, they report a property of the grammatical subject. The otherwise eventive verb becomes a derived stative and, more precisely, receives a generic modal interpretation.

In a typical sentence with Woods's middle marker *imma-*, none of Schäfer's three criteria hold true. Clearly Woods has a somewhat different interpretation of what constitutes a "middle". To be fair, while Woods provides ample evidence to associate *imma-* with the semantics of middle constructions, he makes no attempt to tie them to syntactic middles. In particular, while Woods's "middle" prefixes have the effect of reducing the semantic Agency of the subject, it is only in a restricted set of cases that they completely remove the Agent from the argument structure.

One possible mechanism would draw upon the analysis of anticausative constructions given by Kallulli (2006). In her view each  $v$  head has a tuple of features which controls what sorts of arguments can be merged into Spec/ $v$ P. Her version of Figure 5.3 might look something like Figure 5.4. Since *ensik* is animate, it is compatible with [+intent] and [+cause], and thus meets the requirements to be merged into Spec/ $v$ P as an Agent.

Figure 5.4: Structure for causative  $mu_2$  (after (Kallulli, 2006))



Kallulli's framework is able to provide for the unaccusative reading of *ba-mu* by suppressing the [+intent] feature of the  $v$  head, which means that no Agent can be merged into Spec/ $v$ P, so instead some other argument has to be moved into Spec/ $v$ P to saturate the [+cause] feature. While this provides an adequate alternative to Figure 5.2 for explaining the unaccusative *ba-mu<sub>2</sub>* sentence, it fails to provide a mechanism for deriving the differing semantics of *mu-mu<sub>2</sub>* and *imma-mu<sub>2</sub>* seen above in (5.9). The subjects of both the *mu-* and *imma-* sentences are equally qualified for [+intent] and [+cause], and equally ineligible to be merged into Spec/ $v$ P when [+intent] is suppressed.

Indeed, any attempt to account for the difference between *mu-* and *imma-* as a voice distinction is bound to fail, since the two prefixes normally appear on verbs with identical argument structures. Clearly, something other than structural differences must be involved. While *imma-*

may carry middle-voice semantics, it is not syntactically a middle construction. The following section will show that these prefixes are best analysed as a morphological manifestation of inner aspect.

## 5.5 Conjugation Prefixes as Inner Aspect

It is worth reexamining the semantic tendencies that Woods associates with *imma-* and *ba-*, in order to determine what actually constitutes the essence of the distinction between *mu-*, *imma-*, and *ba-*. Woods uses the overarching term “perspective”, but in specific terms the prefixes express the speaker’s perspective on the event: *mu-* is used to refer to the event as seen from the Agent’s perspective, *imma-* is used (in Woods’s words) to “open up” the structure of the event, while *ba-* emphasises the end result of the event. In other words, these prefixes are functioning as aspectual operators.

Sumerian already expresses aspect in the verb stem itself, which provides a well-known distinction between perfective (*hamṭu*) and imperfective (*marû*). However, this is “viewpoint aspect” or “outer aspect”. The type of aspect expressed by the *hamṭu/marû* distinction is concerned with the real-world relationship between the time of the speech-act and the endpoint of the event.

There is however, also a separate but related notion of “inner aspect” or “situation aspect”. This type of aspect is “concerned with the way in which a predicate **describes** real world events, not the actual structure of the real world.” (MacDonald, 2006). Consider the English-language contrast provided in (5.13). Both sentences could describe the same real-world event, and both sentences are identical in terms of outer aspect. However, there is a significant difference in how these predicates describe that real-world event, namely a difference in terms of inner or situation aspect. In sentence (5.13a), the theme is completely affected by the action (i.e. the pitcher is fully consumed), but in sentence (5.13b) there is no such implication of complete affectedness.



(5.13) Example of inner aspect (MacDonald, 2006)

- a. Rufus drank **a pitcher of beer** at the local pub.
- b. Rufus drank **beer** at the local pub.

Cross-linguistically, voice and inner aspect are often associated. For instance, Johns (2006) argues that the Inuktitut *-si* morpheme can be either an antipassive morpheme or an inner aspectual morpheme, depending on where it is merged. This contrast is shown by the examples in (5.14). Earlier analyses of the morpheme had considered it to be exclusively an antipassive, and were unable to account for its range of behaviours.

(5.14) Inuktitut *-si* morpheme (Johns, 2006)<sup>5</sup>

- a. *Peter pisu-si-juq*  
Peter walk-INCEPT-INTR.3SG  
'Peter starts to walk'
- b. *Peter surak-si-juq            anautar-mik*  
Peter break-AP-INTR.3SG stick-MIK  
'Peter broke the stick'

In English, distinctions of inner aspect typically depend on the nature of the internal argument, most importantly on whether the internal argument is quantised or not. However, in Sumerian, it seems that it is the conjugation prefixes which have the primary role in the expression of inner aspect.

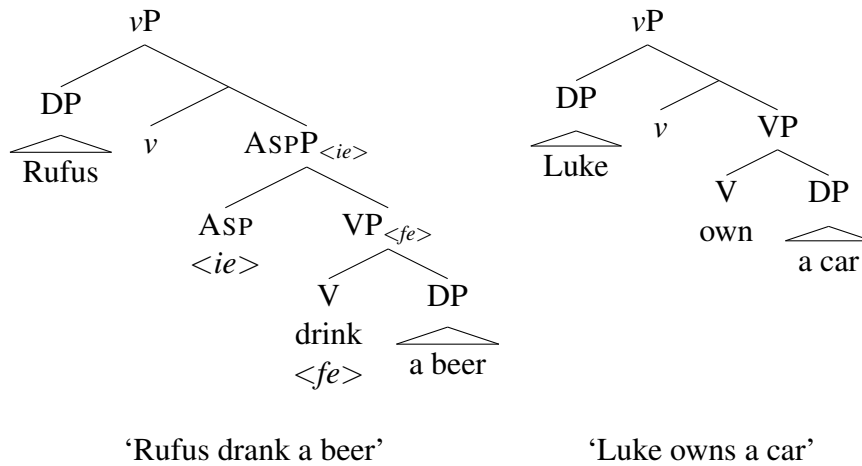
The architecture of inner aspect described by MacDonald (2006) relies on ASP heads which can carry the features <ie> ("initial subevent") and <fe> ("final subevent") to distinguish between the Vendlerian classes of statives, activities, and accomplishments. For example, Figure 5.5 shows an accomplishment, which is expressed by an initial subevent feature on the ASP head and a final subevent feature on the V head.

In MacDonald's view, a stative differs from an eventive in that its structure has no ASPP projection. Lacking an ASPP, the stative has nowhere to put the <ie> or <fe> features so

---

<sup>5</sup>In Johns's glosses, MIK is a case suffix whose meaning is in dispute, so she prefers not to assign it a particular gloss.

Figure 5.5: Structures of accomplishments and statives (MacDonald, 2006)



the state has neither initial nor final subevents. This contrasts with the conception of statives presented by Cuervo (2003), where the stativity is simply a consequence of the presence of  $v_{BE}$  as the outermost event introducer. This suggests that in her framework, inner aspect can be described as a feature associated with the  $v$  head itself, rather than requiring the introduction of a new type of head.<sup>6</sup>

The type of inner aspect being expressed by the Sumerian conjugation prefixes is also somewhat different from the sort of inner aspect being described by MacDonald (2006). MacDonald’s discussion is primarily concerned with telicity, and with whether an event can be bounded and quantised. The distinctions we have seen with *imma-* and *ba-* are not concerned with inner aspect in the narrow sense of telicity, but rather with the broader question of the speaker’s perspective relative to the structure of the event.

Like MacDonald (2006), other modern accounts for inner aspect have centred around telicity and the discussion of how aspect is tied to the quantisation of the direct object. Borer (2005) even names her aspectual head  $ASP_Q$  to indicate that she is referring to “quantity aspect”. The

<sup>6</sup>If independent evidence could be found for an ASP projection in Sumerian, the features described below could equally well reside there rather than on the  $v$ .

account of inner aspect given by Travis (2010) revolves around the event being “measured out” over the direct object. In the case of Sumerian, such an analysis might explain some of the inchoative and ingressive uses of the *imma-* and *ba-* prefixes, but it would have difficulty accounting for other factors which are associated with the conjugation prefixes, most notably subject affectedness.

A closer fit for inner aspect as observed in Sumerian would be the theories of Voorst (1988, 1993). While telicity and quantisation do enter into his discussion of aspect, Voorst also identifies a number of other factors, most notably the notion of “power relation”. A “power relation” exists if the subject exerts control over the event and over the direct object. This seems particularly relevant to the discussion of the Sumerian conjugation prefixes: the *mu-* prefix denotes a relationship where the subject has complete power over the object, while the *imma-* prefix expresses a relation where the distribution of power and affectedness is somewhat blurred. The type of quantisation-associated inner aspect described by Borer (2005) and Travis (2010) can be seen as just a particular case of a broader phenomenon which offsets the affectedness of the object against other factors such as agentivity and subject affectedness.

Thus, it seems reasonable to refer to the contrast between *mu-* on the one hand versus *imma-* and *ba-* on the other as not being one of voice, but rather one of inner aspect. The fourth prefix, *i-*, does not enter into this opposition. Since *i-* can occur across the whole spectrum shown above in Figure 5.1, it seems that when *i-* is present the aspectual opposition is absent. There must then be some feature which *mu-*, *imma-*, and *ba-* have in common, which is missing from *i-*. Using the features proposed by Vanstiphout (1985), shown above in Table 5.1, this feature would be [ $\pm$ focus]. The *mu-*, *imma-*, and *ba-* prefixes would be [+focus] feature while the *i-* prefix would be [–focus].<sup>7</sup> Woods (2008) also uses the term “focus” when referring to the contrast between *mu-* and *i-*. However adopting the term “focus” here would be somewhat misleading, since “focus” is typically used by mainstream linguists in a rather different sense.

---

<sup>7</sup>Vanstiphout (1985) is only concerned with the opposition between *mu-* and *i-*, but it seems reasonable to extend his use of [+focus] to *imma-* and *ba-* as well.

A better name for the feature, which describes its role equally well, is [foreground]. The *mu-*, *imma-*, and *ba-* prefixes are associated with the [foreground] feature, while the *i-* prefix is unmarked for [foreground].<sup>8</sup>

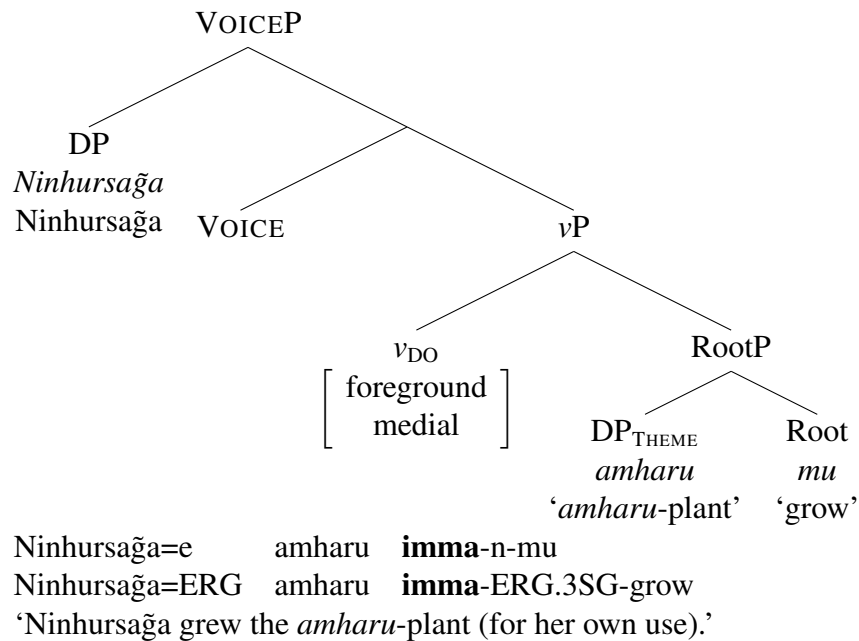
To mark the distinction between the Greek active voice on one hand and the middle and passive voices on the other, Embick (1998) uses the feature [NonAct]. As the name for a feature which marks the distinction between *mu-* and *imma-*, “NonAct” is not ideal, since many occurrences of *imma-* are unquestionably active. Alternative names like [agentivity], [transitivity], or [subject affectedness] would also be possible, but again these names focus on only one aspect of the contrast. Lacking an appropriate name for the feature which marks the contrast between *imma-* and *mu-*, we could simply use Woods’s terminology and refer to it as [middle]. However, the term “middle” could be somewhat misleading since what we have in Sumerian is not a classic middle-voice construction. Instead, we will adopt the term [medial] as a suitable cover-term for the bundle of transitivity- and agency-related tendencies represented by the spectrum of Figure 5.1. The term “medial” has the advantage of suggesting the middle and medio-passive functions of the *imma-* and *ba-* prefixes, without implying that a syntactic middle construction is involved.

The *v* head corresponding to the *imma-* prefix will be marked with the [foreground] and [medial] features while the *v* head corresponding to the *mu-* head will have only the [foreground] feature. Hence the structure of the *imma-mu* sentence from (5.9) will be as shown in Figure 5.6. Structurally, Figure 5.6 is identical to Figure 5.3, but the  $v_{DO}$  head for *imma-* has the addition of a [medial] feature.

A single feature like [medial] can resolve the distinction between *mu-mu* ‘X grows Y’ and *imma-mu* ‘X grows Y (with some effect on X)’. However, there are also cases where there is a contrast between *imma-* and *ba-* in sentences which are structurally parallel. For instance, consider the contrast between the sentences in (5.15), both of which involve the compound

---

<sup>8</sup>We employ monovalent features here rather than binary ones because they give a clearer indication of the markedness of any given feature geometry. The more features present, the more marked it is.

Figure 5.6: Tree for middle-voice  $mu_2$ 

verb *šu ti* ‘to take’ (literally, ‘to approach the hand’). Both would appear to be activity-type events introduced by a  $v_{DO}$  head, and both would have identical argument structures.

(5.15) Contrast between *šu ba-ti* and *šu imma-ti*

⟨anzud<sup>mušen</sup>-de<sub>3</sub> amar-bi      šu    **ba**-an-ti      hur-saĝ-še<sub>3</sub>  
 Anzud=e      amar=bi      šu    **ba**-n-ti      hursaĝ=še  
 Anzud=ERG      calf=3N.POSS    hand    **ba**-ERG.3SG-approach    mountain=ALL  
*ba-an-kur*<sub>9</sub>)  
 ba-n-kur  
 CONJ-ERG.3SG-enter

‘The *Anzud* (a type of mythological bird) took up its young and went to the mountains.’ (GgEN)

⟨il<sub>2</sub>-le    nam-ensi<sub>2</sub>    umma<sup>ki</sup>      šu    **e**-ma-ti)  
 Il=e    namensik    Umma=ak      šu    **imma**-ti  
 Il=ERG    rulership    Umma=GEN    hand    **imma**-approach  
 ‘Il seized the rulership of Umma.’ (Enmetena)

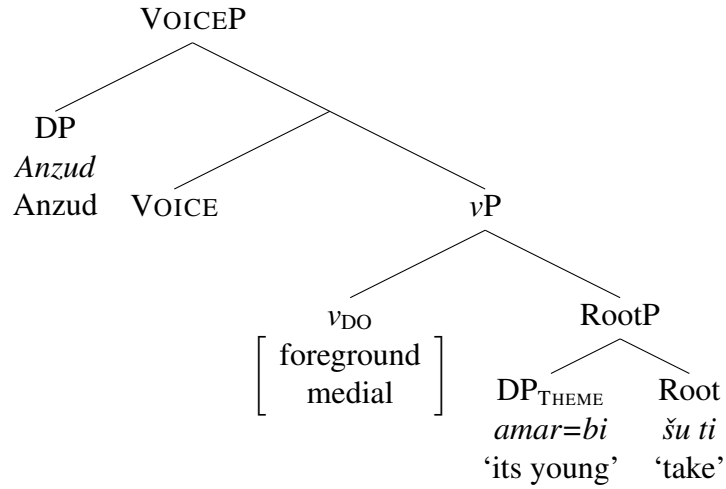
Woods (2008) speaks of *imma-* as representing an “elaboration” or “opening up” of the event, relative to *ba-*. He also makes an appeal to the fact that *imma-* is phonologically heavier than *ba-* to justify the case that *imma-* is also the semantically heavier of the two prefixes. Michalowski (2004) speaks of *imma-* as representing an intensification of the focus on the action. We will adopt Michalowski’s term and refer to the contrastive feature as [intense]. The [intense] feature is dependent on [medial]; since *mu-* is not marked for [medial], it is unspecified for the [intense] feature. The contrast between the two sentences in (5.15) can be seen in Figure 5.7.

The features proposed here are not independent of the choice of light verb. The  $v_{BE}$  light verb, which introduces statives, does not occur with any of the conjugation prefixes. Instead, a stative is indicated by the stative prefix *al-* as shown in (5.16). The inventory of aspectual features described here simply does not appear on stative  $v_{BE}$  heads. This is compatible with the observation by MacDonald (2006) that statives do not have an ASP projection; since the framework being adopted here does not posit a separate ASP head, the equivalent statement for us would be that these features are absent from  $v_{BE}$ .

- (5.16)  $\langle$  *erin<sub>2</sub>-bi*      *al-tur*      *a-ga-bi-ta*      *al-bir-re*  $\rangle$   
 erin=bi      **al-tur**      aga=bi=ta      **al-bir**  
 troops=DEM    **STAT**-small    rear=3N.POSS=ABL    **STAT**-scatter  
 ‘That army is small and scattered from the rear.’ (GgAk)

The  $v_{DO}$  prefix appears to be compatible with any of the aspectual features; with  $v_{DO}$  the choice of [medial] or [intense] is a pragmatic one, governed by the speaker’s perspective towards the event being described. In Figure 5.7 we observed the [medial] feature appearing on a transitive  $v_{DO}$  light verb. The same occurs in Figure 5.8, where the action of stealing is intrinsically self-benefactive and hence has a correspondingly high level of subject-affectedness. Although ‘to steal’ is ordinarily thought of as being a prototypically transitive event, in Sumerian the fact that *zuh* is self-benefactive would appear to be sufficient to justify the use of the *ba-* prefix despite the inherent transitivity of the action.

Figure 5.7: Trees for *šu ba-ti* vs. *šu imma-ti*



⟨anzud<sup>mušen</sup>-de<sub>3</sub> amar-bi šu **ba**-an-ti hur-saḡ-še<sub>3</sub> ba-an-kur<sub>9</sub>⟩

Anzud=e

Anzud=ERG

‘The *Anzud* took up its young’

amar=bi

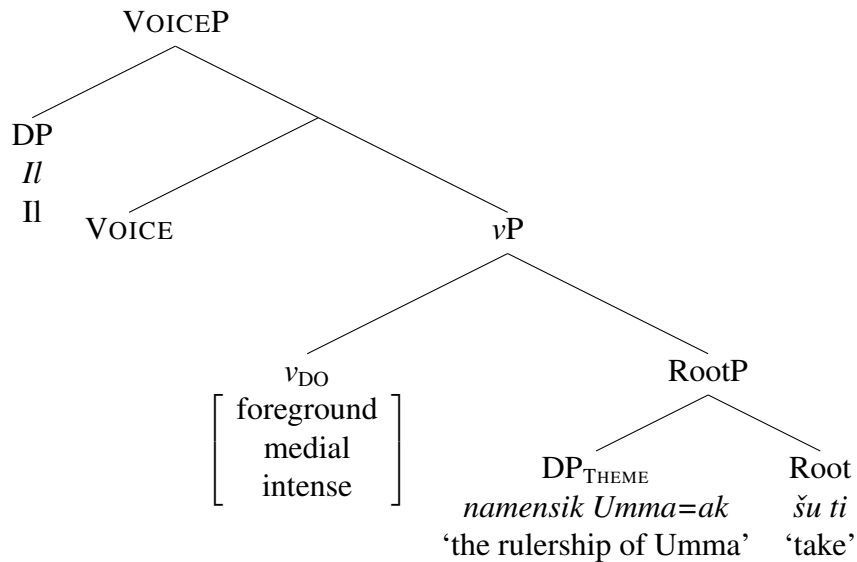
calf=3N.POSS

šu

hand

**ba-ti**

**ba**-approach



⟨il<sub>2</sub>-le nam-ensi<sub>2</sub> umma<sup>ki</sup> šu **e-ma-ti**⟩

Il=e

Il=ERG

‘Il seized the rulership of Umma.’

namensik Umma=ak

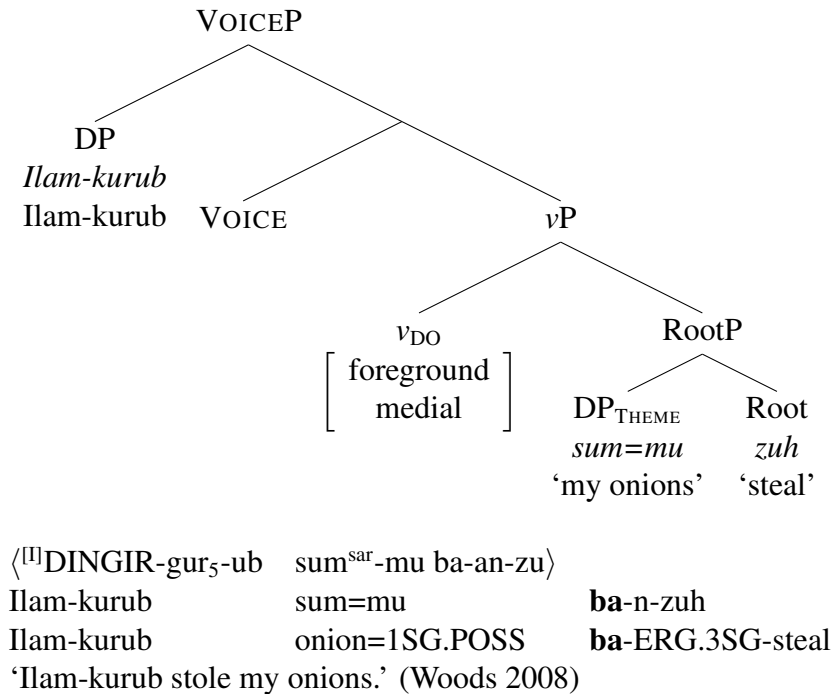
rulership Umma=GEN

šu

hand

šu **imma-ti**

**imma**-approach

Figure 5.8: [medial] feature on a  $v_{DO}$  head

At the other end of the transitivity spectrum,  $v_{GO}$  introduces simple unaccusative verbs of change, so it naturally correlates most often with the low-transitivity *ba-* prefix. Nonetheless, under certain circumstances it can occasionally be found with the *mu-* prefix, as in (5.17), which suggests that  $v_{GO}$  might also be compatible with a range of aspectual features. However, Woods (2008) points out that *mu-* only appears with verbs like *ḡen* ‘to go’ and *ku<sub>4</sub>* ‘to enter’ when accompanied by a manner adverbial. In such cases, the presence of the adverbial suggests that the verb is better analysed as an activity rather than as a simple change. That is, in (5.17), Nanna should be the agent of the verb ‘to go’ rather than simply the undergoer, and the enclosing light verb should be  $v_{DO}$  rather than  $v_{GO}$ .

In contrast, the *ba-* and *imma-* prefixes can combine with a simple  $v_{GO}$  without the presence of a manner adverbial. Thus, it seems that the  $v_{GO}$  head inherently bears the [medial] feature. This is probably a matter of semantics, since the inherent intransitivity of the  $v_{GO}$  head



corresponds to the reduced agency associated with the [medial] feature.

(5.17) *mu-* prefix with simple verb of change *ḡen* ‘to go’

<i>⟨a-a</i>	<i><sup>d</sup>nanna</i>	<i>iri<sup>ki</sup>-ni</i>	<i>urim<sub>5</sub><sup>ki</sup>-ma</i>	<i>saḡ</i>	<i>il<sub>2</sub>-la</i>	<b><i>mu-un-ḡen</i></b>
a’a	Nanna	iri=ni	Urim=a	saḡ	il-a	<b>mu-n-ḡen</b>
father	Nanna	city=3SG.POS	Urim=LOC	head	hold-SUB	<b>mu-ERG.3SG-go</b>

‘With head held high, father Nanna enters his city of Urim.’

## 5.6 Summary

In summary, the conjugation prefixes are the morphological manifestation of a system of inner aspect. The feature geometries shown in Figure 5.9 can account for the semantic contrasts observed by Woods (2008). The full range of aspectual features can appear on a  $v_{DO}$  head. The  $v_{BE}$  head is used for statives, and is consequently incompatible with any of these aspectual features. The  $v_{GO}$  head is inherently intransitive, so it implicitly carries the [medial] feature, which explains why the *mu-* prefix can never appear on a simple unaccusative.

Figure 5.9: Contrastive features for inner aspect

<i>i-</i>	<i>mu-</i>	<i>ba-</i>	<i>imma-</i>
	[foreground]	[foreground]	[foreground]
		[medial]	[medial]
			[intense]

The prefix which has the simplest feature geometry, *i-*, also happens to be the one which appears most often, on 18.1% of all verbs in the corpus. We would expect to see *mu-* only when the scribe has a particular reason to emphasise the agency or transitivity of the verb. In fact, the texts used to build the corpus likely overstate the occurrence of *mu-* (at 16.8%), since royal inscriptions and mythological texts are exactly the sorts of texts where the agents are likely to

be particularly prominent. Vanstiphout (1985) notes that other genres, such as recipes and legal texts, tend to make heavier use of *i-* rather than *mu-*. This is exactly what we would expect if *i-* functions as the default prefix, used whenever there is no rhetorical need to emphasise the inner aspect of the event. It would be interesting to extend the corpus to cover other genres to see whether the distributional patterns of the conjugation prefixes are in fact different.

The *ba-* and *imma-* prefixes are generally similar to each other, but *imma-* has the additional complexity of the added [intense] feature. The presence of this feature explains why *imma-* cannot be used with true passives. Also, the increased feature complexity may help to explain why, at 4.0%, *imma-* is the rarest of the conjugation prefixes.

# Chapter 6

## Conclusion

The research described here was intended to establish that the techniques of modern corpus and theoretical linguistics could productively be applied to the study of Sumerian. At the outset of this endeavour, it was not clear that this effort would result in any practical results.

The first step was to make Sumerian-language resources accessible using more sophisticated tools than are currently available. The construction of an electronically-accessible corpus version of the RIME texts established that existing Sumerological resources, even ones which were never intended for electronic use, could be adapted into the form of a searchable electronic corpus. The techniques developed, most notably the use of an existing lexicon and English-language translation to improve the quality of part-of-speech tagging and affix identification, will also be applicable to annotating other corpora.

For the study of Sumerian, the same techniques developed to import the RIME texts into the corpus will be applicable to other corpora of Sumerian texts, most notably the CDLI. The creation of a morphologically-annotated version of the CDLI will be a tremendous boon to linguists and other scholars working in Sumerian, providing access to texts from the broadest possible range of genres and chronological periods.

The LPattern search language is designed to provide a simple and intuitive query language which is not so complex as to discourage non-technical users. For basic users, it provides a

significant advance over the basic string searches which are typically all that is available with existing corpora of Sumerian. At the same time, LPattern is designed to allow more ambitious users to utilise XPath whenever necessary, and thus have access to a more sophisticated query language. This flexibility is of particular importance when defining a base set of query objects which end-users can work with. The query objects insulate the linguist from having to continually deal with raw orthography, and allow them to work with morphology instead.

The key aspect of this process is the demonstration that the technique of query-based annotation can take a corpus with no morphological mark-up and produce something which has a useful and usable level of annotation. The creation of such a morphologically annotated corpus makes it possible to study aspects of Sumerian on a scale which would otherwise be impractical.

More importantly, the same query-based approach used for annotating Sumerian is equally applicable to other low-resource languages. For many less-studied languages, the resources to create a manually-annotated corpus are simply never going to be available, and query-based annotation provides a shortcut by which a community of linguists can build up annotation as they work. A first demonstration of this will be the use of query-based annotation to create a morphologically-annotated corpus of Elamite-language texts. However, there is nothing in the nature of query-based annotation which restricts its use specifically to ancient texts. The approach should be equally useful in building up morphologically-annotated corpora of modern languages for which the resources for manually annotating a corpus are unavailable.

The syntactic explorations which are made possible by the existence of the annotated Sumerian corpus help to show how theoretical linguistics can clarify aspects of Sumerian syntax which have hitherto remained obscure. As well, these explorations show how the study of Sumerian can extend our knowledge of the range of the human language faculty.

The account of the Sumerian dimensional prefixes as applicative heads puts to rest the long-held notion that the dimensional prefixes represent a system of concord. The range of  $\theta$ -roles which are represented by applicatives in Sumerian is significantly broader than that found

elsewhere, so the study of Sumerian applicatives gives us a better understanding of what is possible with respect to how languages can encode argument structure. While other languages may encode indirect objects as applicatives, Sumerian seems to be unique in using applicative morphology to also encode comitatives, allatives, ablatives, and locatives within the verbal complex. A study of the range of applicative heads observed in other languages will serve to situate Sumerian's applicative morphology in a cross-linguistic context.

The corpus provides support for Woods's theory that the conjugation prefixes are semantically a voice-like system. However, the system does not appear to be syntactically a system of voice, but rather one of inner aspectual features on the  $v$  head. By articulating a system of inner aspect features on the  $v$  head, the system observed in Sumerian diverges from the inner aspect systems described in other languages. This matter will be pursued further, both to expand our knowledge of Sumerian, and to better understand cross-linguistically the range of phenomena which can be encoded as inner aspect.

The research described here fell under three separate fields of study, and has succeeded in advancing the progress of knowledge in each of these fields. The technique of query-based annotation should be a boon to corpus linguists working in low-resource languages anywhere. The new account of the dimensional and conjugation prefixes should help to resolve some long-disputed facets of Sumerian grammar. And finally, our improved understanding of Sumerian morphosyntax expands our knowledge of what is possible in human language.

# Appendix A

## Query Objects

This appendix describes the LPattern queries which were used to build the query objects employed in this research. In many cases, the results of one query were based on the results of a previous query, but such dependencies are not indicated here. As well, in practice, the raw queries were always examined for accuracy, and hits which were clearly erroneous were manually excluded.

Table A.1 shows the queries which were used to define NP objects. Once the NP objects were identified, they were further categorised using the queries shown in Table A.2. As discussed in §3.6 and shown in Table 3.9, the noun phrases were only annotated to the extent that it was useful for the research at hand. In particular, this means that the NP-ERG and NP-GEN queries have not been fully fleshed-out. Also, the task of properly separating noun forms suffixed with *-e* and *-a* remains to be completed.

Table A.3 shows the queries which were employed to define the objects referred to in Chapter 4's discussion of dimensional prefixes. The range of forms was drawn largely from Thomsen (1984).

Tables A.4, A.5, and A.6 show the queries which were used in the discussion of conjugation prefixes in Chapter 5. As can be seen in Table A.5, the *imma-* and *immi-* prefixes were classified separately, although they probably represent the same underlying prefix. The *immi-*

prefix is understood to be the *imma-* followed by a LOC2 prefix (*e-* or *i-*). The large number of queries included in Table A.6 is because identifying the *i-* conjugation prefix is particularly problematic, since it tends to be manifested as a simple vowel of varying quality. It often ends up assimilating completely to a preceding modal prefix. As a rule of thumb, if a finite verb is not marked with *mu-*, *imma-*, or *ba-*, then it is understood to be prefixed with some phonological variant of the *i-* prefix.

Table A.1: Queries for defining NP objects

Name	Underlying Queries	Comments
NP	N	
NP	PD	All pronouns are noun phrases
NP	N ADJ	
NP	NP V-SUB	
NP	N ADJ ADJ	Queries with additional ADJ elements yielded no additional results
V-SUB	V"-a"	For identifying verbs with the participial suffix <i>-a</i>
V-SUB	V[substring(@suffix,2,2)= concat(substring(@stem, string-length(@stem),1),"a")]	A more sophisticated query for identifying participles like <i>kug<sub>3</sub>-ga</i> , where the stem-final consonant is doubled.

Table A.2: Queries for annotating NP objects

Name	Underlying Queries	Comments
NP-3SG.COP	NP"-a-ni-im"	With 3SG.POSS <i>-ani</i> .
NP-3SG.COP	NP"-am3"	
NP-3SG.COP	NP"-am6"	
NP-3SG.POSS	NP"-a-ni"	
NP-3SG.POSS	NP"-a-na"	With LOC <i>-a</i> .
NP-ABL	NP"-ta"	
NP-ALL	NP"-še3"	
NP-ALL	NP"-aš"	
NP-ALL	NP"-eš"	
NP-COM	NP"-da"	
NP-DAT	NP"-ra" [not(@3SG.POSS)] [not(@GEN)]	
NP-DAT	NP"-a-ni-ir"	With 3SG.POSS <i>-ani</i> .
NP-EQU	NP"-gin7"	
NP-ERG	NP"-e"	Indistinguishable from LOC2 <i>-e</i> .
NP-ERG	NP"-ke4"	With GEN <i>-ak</i> .
NP-GEN	NP[substring(@suffix,2,2)= concat(substring(@stem, string-length(@stem),1),"a")]	Indistinguishable from LOC <i>-a</i> .
NP-GEN	NP"-ke4"	With ERG <i>-e</i> .
NP-GEN	NP"-kam"	
NP-LOC	NP"-a" [not(@3SG.POSS) and not(@3SG.COP)]	Usually indistinguishable from GEN <i>-ak</i> .
NP-LOC2	NP"-e"	Indistinguishable from ERG <i>-e</i> .



Table A.3: Queries for dimensional prefixes

Name	Underlying Queries	Comments
V-ABL	V"ta-"	
V-ABL	NP-ABL V-DAT.2SG	To identify ablative <i>ra-</i> prefixes which have been misidentified as DAT.2SG <i>ra-</i> .
V-ABL	V"te-"	With LOC2 <i>e-</i> .
V-ABL	V"ti-"	With LOC2 <i>e-</i> .
V-ABL.3N	V-ABL[contains(@prefix, "b-t")]	
V-ABL.3N	V-ABL[contains(@prefix, "m-t")]	
V-ABL.3SG	V-ABL[contains(@prefix, "n-t")]	
V-ALL	V"ši-"	
V-ALL.1SG	V-ALL[contains(@prefix, "mu-š")]	
V-ALL.2SG	V-ALL[contains(@prefix, "e-š")]	
V-ALL.2SG	V-ALL[contains(@prefix, "-a-š")]	
V-ALL.3N	V-ALL[contains(@prefix, "b-š")]	
V-ALL.3N	V-ALL[contains(@prefix, "m-š")]	
V-ALL.3N	V-ALL[contains(@prefix, "ba-š")]	
V-ALL.3SG	V-ALL[contains(@prefix, "n-š")]	
V-COM	V"da-"	
V-COM	V"da5-"	
V-COM	V"de3-"	With LOC2 <i>e-</i> .
V-COM	V"de4-"	With LOC2 <i>e-</i> .
V-COM	V"di3-ni-"	With LOC <i>ni-</i> .
V-COM	V"di-ni-"	With LOC <i>ni-</i> .
V-COM.1SG	V-mu-COM[not(@COM.3SG)] [not(@COM.2SG)]	
V-COM.2SG	V"e-de3-"	With LOC2 <i>e-</i> .
V-COM.3N	V-COM[contains(@prefix, "m-d")]	
V-COM.3N	V-COM[contains(@prefix, "b-d")]	
V-COM.3SG	V-COM[contains(@prefix, "n-d")]	
V-DAT.1SG	V"ma-"	[not(@imma)]
V-DAT.2SG	V"ma-ra-"	[not(@imma)]
V-DAT.2SG	V"ra-"	[not(@bara)]
V-DAT.2SG	V"ri-"	With LOC2 <i>e-</i> .
V-DAT.3PL	V"ne-"	
V-DAT.3SG	V"na-"	
V-LOC	V"ni-"	
V-LOC2	V-bi	Following Michalowski (2004) and Karahashi (2000/2005) that the <i>bi</i> <sub>2</sub> - prefix originates as <i>ba+i</i> .
V-LOC2	V-immi	
V-LOC2	V"mu-e-"	[not(@COM.2SG)] [not(@ALL.2SG)]
V-LOC2	V-DAT.2SG"ri-"	
V-LOC2	V-ABL"ri-"	
V-LOC2	V-ABL"te-"	
V-LOC2	V-COM"de3-"	
V-LOC2	V-COM[not(@LOC)] "di-"	

Table A.4: Queries for conjugation prefixes *ba-* and *mu-*

Name	Underlying Queries	Comments
V-ba	V"ba-" [not(@bara)]	
V-bara	V"ba-ra-"	Query object for modal prefix <i>bara-</i> defined solely to limit erroneous hits for the <i>ba-</i> prefix.
V-bi	V"bi2-"	Classified separately from V-ba, but treated as V-ba-LOC2 in the current research.
V-mi	V"mi-" [not(@immi)]	The <i>mi-</i> prefix is probably a variant of <i>mu-</i> where the vowel has assimilated to a subsequent prefix.
V-mini	V"mi-ni-" [not(@immi)]	Johnson (2004) makes specific claims about <i>mini-</i> , so a separate query object was defined to investigate that.
V-mu	V"mu-"	

Table A.5: Queries for conjugation prefixes *imma-* and *immi-*

Name	Underlying Queries	Comments
V-imma	V"im-ma-"	
V-imma	V"am3-ma-"	
V-imma	V"em-ma-"	
V-imma	V"um-ma-"	
V-imma	V"e-ma-"	
V-imma	V"i-ma-"	
V-imma	V"i3-ma-"	
V-imma	V"u3-ma-"	After modal prefix <i>u-</i> .
V-imma	V"he2-ma-"	After modal prefix <i>ha-</i> .
V-imma	V"nam-ma-"	After modal prefix <i>na-</i> .
V-imma	V"nu-ma-"	After modal prefix <i>nu-</i> .
V-imma	V"še3-ma-"	After modal prefix <i>ša-</i> .
V-immi	V"im-mi-"	
V-immi	V"am3-mi-"	
V-immi	V"em-mi-"	
V-immi	V"um-mi-"	
V-immi	V"e-mi-"	
V-immi	V"i-mi-"	
V-immi	V"i3-mi-"	
V-immi	V"u3-mi-"	After modal prefix <i>u-</i> .
V-immi	V"he2-mi-"	After modal prefix <i>ha-</i> .
V-immi	V"nam-mi-"	After modal prefix <i>na-</i> .
V-immi	V"nu-mi-"	After modal prefix <i>nu-</i> .
V-immi	V"še3-mi-"	After modal prefix <i>ša-</i> .
V-immi	V"im-me-"	
V-immi	V"am3-me-"	
V-immi	V"em-me-"	
V-immi	V"um-me-"	
V-immi	V"um-mi-"	
V-immi	V"e-me-"	
V-immi	V"i-me-"	
V-immi	V"i3-me-"	
V-immi	V"u3-me-"	After modal prefix <i>u-</i> .
V-immi	V"he2-me-"	After modal prefix <i>ha-</i> .
V-immi	V"nam-me-"	After modal prefix <i>na-</i> .
V-immi	V"nu-me-"	After modal prefix <i>nu-</i> .
V-immi	V"še3-me-"	After modal prefix <i>ša-</i> .

Table A.6: Queries for conjugation prefix *i-*

Name	Underlying Queries	Comments
V-i	V[@prefix="he2-"]	
V-i	V"he2-da-"	
V-i	V[not(@immi)] [not(@imma)] "he2-em-"	
V-i	V"he2-en-"	
V-i	V[starts-with(@prefix,"i-ib-")]	
V-i	V[starts-with(@prefix,"i-im-")]	
V-i	V[starts-with(@prefix,"i-in-")]	
V-i	V[starts-with(@prefix,"i-ni-")]	
V-i	V"i-ra-"	
V-i	V"i-ri-"	
V-i	V[@prefix="ib-"]	
V-i	V[@prefix="ib2-"]	
V-i	V[@prefix="im-"]	
V-i	V[@prefix="in-"]	
V-i	V"in-na-"	
V-i	V"in-ši-"	
V-i	V[starts-with(@prefix, "na-ab-")]	
V-i	V[starts-with(@prefix, "na-an-")]	
V-i	V[@prefix="na-"]	
V-i	V[@prefix="nu-"]	
V-i	V"nu-um-" [not(@imma)] [not(@immi)]	
V-i	V[@prefix="ši-"]	
V-i	V[starts-with(@prefix,"ši-ib-")]	
V-i	V[starts-with(@prefix,"ši-ib2-")]	
V-i	V"ši-im-" [not(@imma)] [not(@immi)]	
V-i	V[starts-with(@prefix,"ši-in-")]	
V-i	V[prefix="u3-"]	
V-i	V"u3-na-"	
V-i	V[@prefix="ub-"]	
V-i	V[starts-with(@prefix, "um-")]	
V-i	V[@prefix="un-"]	
V-i	V-a	
V-i	V"e-" [not(@immi)] [not(@imma)] [not(@DAT.3PL)]	
V-i	V"e-ga-"	
V-i	V"i3-" [not(@immi)] [not(@imma)]	
V-i	V"nu-ub-"	
V-i	V"nu-un-"	
V-i	V"nu-ši-"	

# Appendix B

## LPattern Grammar

The LPattern grammar is implemented using Gnu Flex as a lexical analyser and Gnu Bison to generate the parser. The following grammar description is incomplete because the content of a PREDICATE can be any valid XPath predicate expression, as described in <http://www.w3.org/TR/xpath20/#doc-xpath-Predicate>.

```
LPattern ::= Pattern
Pattern  ::= (Pattern Pattern) | Basic | Option | Disjunction
          | Join | Preabsence | Postabsence | "_"
Basic    ::= Ident | String
Option   ::= ("?" Ident) | ("?(" Pattern ")")
Disjunction ::= Pattern "|" Pattern
Join     ::= Pattern "*" Pattern
Preabsence ::= ("!" Basic Pattern) | ("!(" Pattern ")" Pattern)
Postabsence ::= (Pattern "!" Basic) | (Pattern "!(" Basic ")")
Ident    ::= IDENT | Construct | (Ident PREDICATE)
          | StrungIdent
StrungIdent ::= Ident STRING
           A combination of an identifier expression followed by a string, such as
           V-DAT"šum2".
String   ::= STRING | QUOTELESS_STRING
Construct ::= (IDENT CONSTRUCT) | (Construct | CONSTRUCT)
           Allows for multiple construct specifications, such as V-COM-ABL.
IDENT    ::= [A-Z] [a-zA-Z]*
```

CONSTRUCT	::=	<code>-[a-zA-Z1-3.\-]+</code>	The construct-specification part of a constructed object, such as the <code>-3SG.POSS</code> portion of <code>NP-3SG.POSS</code> .
PREDICATE	::=	<code>"[" [^]" ]"</code>	An XPath predicate expression, enclosed in square-brackets.
STRING	::=	<code>'"'" [^"]*' '"'</code>	
QUOTELESS_STRING	::=	<code>[a-zšǧ][a-z0-9\-\šǧ]*</code>	Provided for convenience, so that an end-user does not have to type in any superfluous quotation marks when making a simple string query. The characters <code>š</code> and <code>ǧ</code> are of course specific to Sumerian, and this definition would have to be changed if LPattern were to be extended for use with other languages.

# Appendix C

## Lemmatiser Source Code

The following C++ source code is extracted from the implementation of the EPSDLloader class, and is responsible for bulk of the morphological processing described in §2.2. This code is by no means complete, and is intended only for illustrative purposes.

```
// Utility function which determines whether a possible verb is better matched
// as an adjective.
static bool matchingAdjective(QDomElement& w1, const QDomElement& w2) {
    if (w1.tagName() == "V" || w2.tagName() == "V") {
        if (w1.attribute("lemma") == w2.attribute("lemma")) {
            if (w1.tagName() == "V") {
                w1 = w2;
            }
            return true;
        }
    }
    return false;
}

// Main entry point for lemmatiser. This method is called for each
// orthographic word in each <para> node.
void EPSDLloader::attachWord(const QString& word, QDomElement& parentNode, QString& english,
                             QStringList& failureList) const {
    QDomElement wordNode;
    QString w = word;
    mRescued = false;

    // Strip out any spelling notes.
    QString actual;
    if (w.contains('(')) {
        actual = w;
        w.remove(QRegExp("\\([^\(\\)]*\\)"));
    }

    // Strip out any markers of reconstructed spellings. We'll assume Doug and Dietz knew
    // what they were doing.
    if (w.contains('<')) {
        actual = w;
        w.remove(QRegExp("<>"));
    }
}
```

```

}

int bracePos = w.indexOf('{');
if (bracePos > 0) {
    actual = w;
    w.remove(bracePos - 1, w.indexOf('}') - bracePos + 2);
}

// The loop below implements the matching process, including affix-stripping
// and processing of compound verbs and amissable consonants.
QList<QDomElement> accumulatedPossibilities;
QList<PartOfSpeech> possiblePartsOfSpeech;
if (!w.isEmpty()) {
    for (int i = UNKNOWN; i < NUM_PARTS_OF_SPEECH; ++i) {
        QList<QDomElement> localPossibilities;
        wordNode = createWordNodes((PartOfSpeech)i, w, parentNode, english,
                                   localPossibilities);

        if (wordNode.isNull()) {
            if (localPossibilities.size() > 0) {
                possiblePartsOfSpeech.append((PartOfSpeech)i);
                accumulatedPossibilities += localPossibilities;
            }
        } else {
            // Found our exact match.
            break;
        }
    }
} else {
    wordNode = parentNode.ownerDocument().createElement("X");
}

// The following code corresponds to the disambiguation stage of the
// lemmatisation process.
if (wordNode.isNull()) {
    if (accumulatedPossibilities.size() == 0) {
        if (word.contains("-x") || word.startsWith("x") || word.contains("...") ||
            !word.contains(QRegExp("[a-z]"))) {
            // A damaged form (or a standalone sign) which we probably
            // won't be able to lemmatise.
            wordNode = parentNode.ownerDocument().createElement("X");
            ++numX;
        } else {
            // The form looks legitimate, but it's not one we can find.
            wordNode = parentNode.ownerDocument().createElement("W");
            ++numW;
        }
        failureList.append(word + "\tnot found\t" + parentNode.attribute("id"));
    } else {
        wordNode = accumulatedPossibilities[0];
        if (accumulatedPossibilities.size() > 1) {
            QStringList glosses;
            // If the conflict is between something and an unprefix verb, the unprefix
            // verb will lose.
            bool verbLoses = possiblePartsOfSpeech.size() == 2 &&
                wordNode.attribute("prefix").isNull() &&
                matchingAdjective(wordNode, accumulatedPossibilities[1]);
            if (possiblePartsOfSpeech.size() > 1 && !verbLoses) {
                // Multiple inexact glosses, of multiple parts of speech.
                QStringList lemmata;
                wordNode = parentNode.ownerDocument().createElement("W");
                QString f = QString("%1 ambiguous").arg(word);
                for (int i = 0; i < accumulatedPossibilities.size(); ++i) {
                    const QDomElement& elem = accumulatedPossibilities[i];
                    QString lemma = elem.attribute("lemma");
                    if (!lemmata.contains(lemma)) {
                        lemmata.append(lemma);
                    }
                }
                QString gloss = elem.attribute("english");
            }
        }
    }
}

```



```

        if (!glosses.contains(gloss)) {
            glosses.append(gloss);
        }
        f += '\t' + gloss;
    }
    failureList.append(f + '\t' + parentNode.attribute("id"));
    wordNode.setAttribute("lemma", lemmata.join(", "));
    ++numW;
} else {
    // There are ambiguous meanings, but they're all the same part-of-speech, so
    // they're just recorded as alternate glosses.
    for (int i = 0; i < accumulatedPossibilities.size(); ++i) {
        QString gloss = accumulatedPossibilities[i].attribute("english");
        if (!glosses.contains(gloss)) {
            if (!(verbLoses && gloss.startsWith("to "))) {
                glosses.append(gloss);
            }
        }
    }
}
wordNode.setAttribute("english", glosses.join(", "));
}
++numInexact;
}
} else {
    ++numExact;
    if (mRescued) {
        failureList.removeLast();
        --numW;
    }
}
wordNode.setAttribute("orth", w);
if (!actual.isEmpty()) {
    wordNode.setAttribute("actual", actual);
}
parentNode.appendChild(wordNode);
++numWords;
}

// Utility function which determines whether a word can be matched as a noun.
static bool matchingNominal(const QDomElement& lastNoun, const QString& nominalOrth,
                           const QString& nominalLemma) {
    if (lastNoun.attribute("orth") == nominalOrth) {
        return true;
    }
    // An ambiguous parse might potentially have many lemmata.
    if (lastNoun.attribute("lemma").split(", ").contains(nominalLemma)) {
        return true;
    }
    return false;
}

// For this part of speech, determine all the possible lexical entries
// which could match the given word.
QDomElement EPSDLoader::createWordNodes(PartOfSpeech partOfSpeech, const QString& word,
                                         QDomElement& parentNode, QString& english,
                                         QList<QDomElement>& accumulatedPossibilities) const {
    QDomElement result;

    if (word.indexOf(QRegExp("[0-9]")) == 0) {
        // It's a number. Skip the usual processing.
        result = parentNode.ownerDocument().createElement("NU");
        int suffixPos = word.indexOf(QRegExp("[^(0-9\\-\\|\\.\\.\\.)]"));
        QString form = word;
        if (suffixPos > 0) {
            result.setAttribute("suffix", word.mid(suffixPos - 1));
        }
    }
}

```

```

    form = word.left(suffixPos - 1);
}
result.setAttribute("lemma", form);
result.setAttribute("english", form);
} else {
    int bestPos = -1;
    int bestLen = 0;
    QString bestGloss;
    QString bestLemma;

    if (mOrthographyLookup[partOfSpeech].contains(word)) {
        const QList<Lexeme>& matches = mOrthographyLookup[partOfSpeech][word];

        // Found one or more matches. See if we can pin down the exact one.
        for (int i = 0; i < matches.size(); ++i) {
            const Lexeme& l = matches[i];
            if (!l.mNominal.isEmpty()) {
                bool foundCompound = false;
                // It's a compound verb, so look for the expected nominal element.
                QDomElement lastNoun = parentNode.lastChildElement();
                QString nominalLemma = l.mLemma.split(" ")[0];
                if (matchingNominal(lastNoun, l.mNominal, nominalLemma)) {
                    foundCompound = true;
                } else {
                    // Try the penultimate word.
                    lastNoun = lastNoun.previousSibling().toElement();
                    if (matchingNominal(lastNoun, l.mNominal, nominalLemma)) {
                        foundCompound = true;
                    }
                }
            }
            if (foundCompound) {
                if (lastNoun.tagName() == "N") {
                    qDebug("Recognised compound verb %s %s.", qPrintable(l.mNominal),
                        qPrintable(word));
                } else {
                    // The nominal we're looking for has been misdiagnosed as an ambiguous
                    // part of speech.
                    QString orth = lastNoun.attribute("orth");
                    QDomElement newNoun = parentNode.ownerDocument().createElement("N");
                    parentNode.insertAfter(newNoun, lastNoun);
                    newNoun.setAttribute("orth", orth);
                    newNoun.setAttribute("english", "nominal element of " + l.mLemma);
                    parentNode.removeChild(lastNoun);
                    qDebug("Rescued compound verb %s %s.", qPrintable(l.mNominal),
                        qPrintable(word));
                    mRescued = true;
                }
            }

            if (!l.findBestGloss(english, bestGloss, bestPos, bestLen)) {
                // Didn't actually find the gloss, but we know we have an exact match.
                bestGloss = l.mGlosses[0];
                bestPos = 0;
                bestLen = 0;
            }
            bestLemma = l.mLemma;
        }
    } else {
        if (l.findBestGloss(english, bestGloss, bestPos, bestLen)) {
            bestLemma = l.mLemma;
        }
    }
}

if (!bestLemma.isEmpty()) {
    result = parentNode.ownerDocument().createElement(partOfSpeechLookup[partOfSpeech]);
    result.setAttribute("lemma", bestLemma);
    result.setAttribute("english", bestGloss);
    if (bestLen > 0) {

```

```

        english.remove(bestPos, bestLen);
    }
    return result;
}

// No exact match, so accumulate all the possibilities and return a null element.
for (int i = 0; i < matches.size(); ++i) {
    const Lexeme& l = matches[i];

    // Skip compound verbs, since these will already have been processed (if possible).
    // Also skip proper names (identified by capital letter in gloss), since these
    // only count if they match exactly in the text.
    if (l.mNominal.isEmpty() && !l.mGlosses[0][0].isUpper()) {
        QDomElement elem = parentNode.ownerDocument().createElement(
            partOfSpeechLookup[partOfSpeech]);
        elem.setAttribute("lemma", l.mLemma);
        elem.setAttribute("english", l.mGlosses[0]);
        accumulatedPossibilities.append(elem);
    }
}

// No exact match, so also try the form with suffixes and prefixes.
result = testAffixes(partOfSpeech, word, parentNode, english, accumulatedPossibilities);
}
return result;
}

// For a given word and part-of-speech combination, performing affix
// stripping in order to find a match.
QDomElement EPSDLoader::testAffixes(PartOfSpeech partOfSpeech, const QString& word,
                                     QDomElement& parentNode, QString& english,
                                     QList<QDomElement>& accumulatedPossibilities) const {

    // First, try suffixes just by themselves.
    QDomElement result = testSuffixes(partOfSpeech, word, parentNode, english,
                                       accumulatedPossibilities);

    if (result.isNull()) {
        for (int i = 0; i < mPrefixes[partOfSpeech].size(); ++i) {
            const QString& prefix = mPrefixes[partOfSpeech][i];
            if (word.startsWith(prefix)) {
                QList<QDomElement> prefixPossibilities;
                QString w = word.mid(prefix.size());
                if (mOrthographyLookup[partOfSpeech].contains(w)) {
                    result = createWordNodes(partOfSpeech, w, parentNode, english,
                                             prefixPossibilities);
                    result.setAttribute("prefix", prefix);
                } else {
                    // Started with the right prefix, but failed to find a match in the lexicon.
                    // Try in combo with suffixes.
                    result = testSuffixes(partOfSpeech, w, parentNode, english, prefixPossibilities);
                }
            }
            if (result.isNull()) {
                // No exact match; note our prefix on each of the possibilities.
                for (int j = 0; j < prefixPossibilities.size(); ++j) {
                    prefixPossibilities[j].setAttribute("prefix", prefix);
                }
                accumulatedPossibilities += prefixPossibilities;
            } else {
                // Exact match in the English text.
                result.setAttribute("prefix", prefix);
                break;
            }
        }
    }
}
}
}

```

```

    return result;
}

// Called by testAffixes to dealing with suffixes. More complicated
// than prefixes because it has to deal with amissable consonants.
QDomElement EPSDLoader::testSuffixes(PartOfSpeech partOfSpeech, const QString& word,
                                     QDomElement& parentNode, QString& english,
                                     QList<QDomElement>& accumulatedPossibilities) const {
    QDomElement result;

    for (int i = 0; i < mSuffixes[partOfSpeech].size(); ++i) {
        QString suffix = mSuffixes[partOfSpeech][i];
        bool matched = false;
        QString w;

        if (word.size() > suffix.size()) {
            w = word.left(word.size() - suffix.size());
            if (suffix[1] == 'C') {
                QChar c = word[w.size() + 1];
                if (QRegExp("-[bdg\\x011dhklmnprs\\x016itz]" +
                           suffix.mid(2)).exactMatch(word.right(suffix.size()))) {
                    if (mOrthographyLookup[partOfSpeech].contains(w)) {
                        // Suffix which involves reconstructing an amissable consonant.
                        const QList<Lexeme>& matches = mOrthographyLookup[partOfSpeech][w];
                        for (int j = 0; j < matches.size(); ++j) {
                            if (matches[j].mLemma.endsWith(c)) {
                                // Record the actual suffix, with the real character instead of the C.
                                suffix = word.right(suffix.size());
                                matched = true;
                                break;
                            }
                        }
                    }
                }
            } else {
                if (word.endsWith(suffix)) {
                    // Simple case. Matches when the bare suffix is added.
                    matched = mOrthographyLookup[partOfSpeech].contains(w);
                }
            }
        }

        if (matched) {
            QList<QDomElement> suffixPossibilities;
            result = createWordNodes(partOfSpeech, w, parentNode, english, suffixPossibilities);
            if (result.isNull()) {
                // No exact match for english; record our suffix on each of the possibilities.
                for (int i = 0; i < suffixPossibilities.size(); ++i) {
                    suffixPossibilities[i].setAttribute("suffix", suffix);
                }
                accumulatedPossibilities += suffixPossibilities;
            } else {
                // Exact match in the English text.
                result.setAttribute("suffix", suffix);
                break;
            }
        }
    }

    return result;
}

```

# Bibliography

- Pascal Attinger. *Éléments de linguistique sumérienne : la construction de du<sub>11</sub>/e/di "dire"*. Editions universitaires; Vandenhoeck & Ruprecht, Fribourg, Suisse Göttingen, 1993.
- Susana Béjar. *Phi-syntax: a theory of agreement*. Ph.D., University of Toronto, 2003.
- Steven Bird and Haejoong Lee. Graphical query for linguistic treebanks. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 22–30, 2007.
- Steven Bird, Peter Buneman, and Wang-Chiew Tan. Towards a query language for annotation graphs. In *Second International Conference on Language Resources and Evaluation*, pages 807–814, 2000.
- Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing in Python*. University of Pennsylvania, 2001-2007.
- Steven Bird, Yi Chen, Susan Davidson, Haejoong Lee, and Yifeng Zheng. Extending XPath to support linguistic queries. In *Proceedings of Programming Language Technologies for XML (PLANX)*, pages 35–46, Long Beach, January 2005. ACM.
- Steven Bird, Yi Chen, Susan Davidson, Haejoong Lee, and Yifeng Zheng. Designing and evaluating an XPath dialect for linguistic queries. In *22nd International Conference on Data Engineering (ICDE)*, pages 52–61, Atlanta, April 2006.
- J.A. Black, G. Cunningham, J. Ebeling, E. Flückiger-Hawker, E. Robson, J. Taylor, and G. Zólyomi. The Electronic Text Corpus of Sumerian Literature. <http://www-etcsl.orient.ox.ac.uk/>, 1998–2006. Accessed January 2009.
- Jeremy Black. The alleged ‘extra’ phonemes of Sumerian. *Revue d’Assyriologie*, 84:107–118, 1990.
- Eulàlia Bonet. The person-case constraint: A morphological approach. *MIT Working Papers in Linguistics*, 22:33–52, 1994.
- Hagit Borer. *The Normal Course of Events*. Oxford University Press, Oxford, 2005.
- Joan Bresnan and Lioba Moshi. Object asymmetries in comparative Bantu syntax. *Linguistic Inquiry*, 21(2):147–185, 1990.
- Miguel Civil. Modal prefixes. *Acta Sumerologica*, 22, 2000/2005.

- James Clark and Steve DeRose. XML Path language (XPath). <http://www.w3.org/TR/xpath>, 1999.
- María Cristina Cuervo. *Datives at Large*. PhD thesis, Massachusetts Institute of Technology, 2003.
- Tony Dodd. Xaira: the reference manual. <http://www.oucs.ox.ac.uk/rts/xaira/Doc/oldrefman.xml>, 2005.
- Tony Dodd. XXQ - an informal introduction. <http://www.oucs.ox.ac.uk/rts/xaira/Doc/XXQdoc.xml>, 2006.
- Jarle Ebeling and Graham Cunningham. Lemmatising the Electronic Corpus of Sumerian Literature. Journée ATALA : Traitement automatique des langues et langues anciennes, May 2005.
- Dietz Otto Edzard. *Sumerian Grammar*. Brill, Boston, MA, 2003.
- Dietz Otto Edzard. *Gudea and his Dynasty*. Royal Inscriptions of Mesopotamia. Early periods; v. 3/1. University of Toronto Press, Toronto, 1997.
- David Embick. Voice systems and the syntax/morphology interface. *MIT Working Papers in Linguistics*, 32:41–72, 1998.
- R. K. Englund and Peter Damerow. Cuneiform Digital Library Initiative. <http://cdli.ucla.edu/>, 2000–2005. Accessed January 2009.
- Adam Falkenstein. *Grammatik der Sprache Gudeas von Lagaš*. Number 28-29 in *Analecta Orientalia*. Biblical Institute Press, Rome, 1978.
- Douglas Frayne. *Pre-Sargonic Period (2700-2350 BC)*. Royal Inscriptions of Mesopotamia: Early Periods, v. 1. University of Toronto Press, Toronto, 2005.
- Douglas Frayne. *Old Babylonian Period (2003-1595 BC)*. Royal Inscriptions of Mesopotamia: Early Periods, v. 4. University of Toronto Press, Toronto, 1990.
- Douglas Frayne. *Sargonic and Gutian Periods (2334-2113 BC)*. Royal Inscriptions of Mesopotamia: Early Periods, v. 2. University of Toronto Press, Toronto, 1993.
- Douglas Frayne. *Ur III Period (2112-2004 BC)*. Royal Inscriptions of Mesopotamia. Early periods ; v. 3/2. University of Toronto Press, Toronto, 1997.
- Mark Geller. The last wedge. *Zeitschrift für Assyriologie*, 87:43–95, 1997.
- Gene B. Gragg. *Sumerian dimensional infixes*. Butzon und Bercker, Kevelaer, 1973.
- Richard Treadwell Hallock and Benno Landsberger. *Neo-Babylonian Grammatical Texts*, volume 4 of *Materielen zum Sumerischen Lexikon*, pages 129–202. Pontificio Istituto Biblico, Rome, 1956.

- Eduard Hovy. Toward a ‘science’ of annotation: Experience from ontonotes. Fifth International Conference on Language Resources and Evaluation, May 2006.
- Thorkild Jacobsen. About the Sumerian verb. In H. G. Güterbock and T. Jacobsen, editors, *Studies in Honor of Benno Landsberger on His Seventy-Fifth Birthday*, volume 16 of *Assyriological Studies*, pages 71–102. University of Chicago Press, Chicago, 1965.
- Alana Johns. *Ergativity: Emerging Issues*, chapter Ergativity and Change in Inuktitut. *Studies in Natural Language and Linguistic Theory*. Springer, Dordrecht, 2006.
- J. Cale Johnson. *In the Eye of the Beholder: Quantificational, pragmatic and aspectual features of the \*bi- verbal prefix in Sumerian*. PhD thesis, U.C.L.A., 2004.
- Dalina Kallulli. A unified analysis of passives and anticausatives. *Empirical Issues in Syntax and Semantics 6*, 6:201–226, 2006.
- Fumi Karahashi. The locative-terminative verbal infix in Sumerian. *Acta Sumerologica*, 22, 2000/2005.
- Joachim Krecher. Die /m/-Präfixe des sumerischen Verbums. *Orientalia*, 54:133–181, 1985.
- Catherine Lai and Steven Bird. Querying and updating treebanks: A critical survey and requirements analysis. In *Proceedings of the Australasian Language Technology Workshop*, pages 139–146, 2004.
- Catherine Lai and Steven Bird. LPath<sup>+</sup>: A first-order complete language for linguistic tree query. In *Proceedings of the 19th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, pages 1–12, 2005.
- Catherine Lai and Steven Bird. Querying linguistic trees. *Journal of Logic, Language and Information*, 19:53–73, 2010.
- Geoffrey Leech. Adding linguistic annotation. In Martin Wynne, editor, *Developing Linguistic Corpora: a Guide to Good Practice*. Oxbow Books, Oxford, 2005.
- Geoffrey Leech. Grammatical tagging. In Roger Garside, Geoffrey Leech, and Anthony McEnery, editors, *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, New York, 1997.
- Jonathan MacDonald. *The Syntax of Inner Aspect*. Ph.D. thesis, Stony Brook University, 2006.
- Alec Marantz. No escape from syntax: Don’t try morphological analysis in the privacy of your own lexicon. *Penn Working Papers in Linguistics*, 4.2, 1997.
- Tony McEnery and Paul Rayson. A corpus/annotation toolbox. In Roger Garside, Geoffrey Leech, and Anthony McEnery, editors, *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, New York, 1997.
- Martha McGinnis. UTAH at merge: Evidence from multiple applicatives. *MIT Working Papers in Linguistics*, 49:183–200, 2005.

- Piotr Michalowski. Sumerian. In Roger D. Woodard, editor, *The Cambridge Encyclopedia of the World's Ancient Languages*. Cambridge University Press, Cambridge, 2004.
- Manuel Molina. Base de Datos de Textos Neosumerios. <http://bdts.filol.csic.es>, 2002–2010.
- David Michael Pesetsky. *Zero syntax : experiencers and cascades*. MIT Press, Cambridge, Mass., 1995.
- David A. Peterson. *Applicative Constructions*. Oxford University Press, New York, 2007.
- Liina Pykkänen. *Introducing Arguments*. Ph.D. thesis, Massachusetts Institute of Technology, 2002.
- Douglas L. T. Rohde. *TGrep2 User Manual*, 2005.
- Florian Schäfer. *On the nature of anticausative morphology: External arguments in change-of-state contexts*. Ph.D. thesis, Universität Stuttgart, 2007.
- John Sinclair. Corpus and text - basic principles. In Martin Wynne, editor, *Developing Linguistic Corpora: a Guide to Good Practice*. Oxbow Books, Oxford, 2005.
- Åke W. Sjöberg and Hermann Behrens, editors. *The Sumerian Dictionary of the University Museum of the University of Pennsylvania*. University of Pennsylvania, Philadelphia, 1992–2010.
- Åke W. Sjöberg, Erle Leichty, and Steve Tinney. ePSD: Electronic Pennsylvania Sumerian Dictionary. <http://psd.museum.upenn.edu/epsd>, 2004. Accessed July 2004.
- Eric J. M. Smith. *Harmony and the Vowel Inventory of Sumerian*. Generals paper, University of Toronto, 2006a.
- Eric J. M. Smith. *A Unified Account of Elamite Class-markers*. Generals paper, University of Toronto, 2006b.
- Eric J. M. Smith. Using LPath queries to annotate corpora: A case study of Elamite and Sumerian. In *Electronic Corpora of Ancient Languages: Proceedings of the International Conference, Prague, November 16-17*, Chatreššar, pages 121–134. Institute of Comparative Linguistics, Charles University, Prague, November 2007a.
- Eric J. M. Smith. [–ATR] harmony and the vowel inventory of Sumerian. *Journal of Cuneiform Studies*, 59, 2007b.
- Eric J. M. Smith. Using a query language as an annotation tool. 2008 Meeting of the American Association for Corpus Linguistics, March 2008.
- Nicholas Smith, Sebastian Hoffman, and Paul Rayson. Corpus tools and methods, today and tomorrow: Incorporating linguists' manual annotations. *Literary and Linguistic Computing*, 23(2):163–180, 2008.



- Marie Louise Thomsen. *The Sumerian language: an introduction to its history and grammatical structure*. Mesopotamia: Copenhagen studies in Assyriology; v. 10. Akademisk Forlag, Copenhagen, 1984.
- François Thureau-Dangin. Sur les préfixes du verbe sumérien. *Zeitschrift für Assyriologie*, 20: 308–404, 1907.
- Steve Tinney and Fumi Karahashi. Pennsylvania Parsed Corpus of Sumerian. <http://psd.museum.upenn.edu/ppcs/>, 2003-2004.
- Lisa Travis. *Inner Aspect*. Springer, Dordrecht, 2010.
- Herman L. Vanstiphout. On the verbal prefix /i/ in Standard Sumerian. *Revue d'Assyriologie*, 79:1–15, 1985.
- Niek Veldhuis. Digital Corpus of Cuneiform Lexical Texts. <http://cuneiform.ucla.edu/dcclt>, 2003.
- Jan van Voorst. *Event Structure*. J. Benjamins, Amsterdam; Philadelphia, 1988.
- Jan van Voorst. A localist model for event semantics. *Journal of Semantics*, 10:65–111, 1993.
- Christopher E. Woods. *The Grammar of Perspective: The Sumerian conjugation prefixes as a system of voice*. Brill, Leiden, The Netherlands, 2008.
- Mamoru Yoshikawa. The Sumerian verbal prefixes mu-, ì and topicality. *Orientalia Nova Series*, 48:185–206, 1979.
- Gábor Zólyomi. Directive infix and oblique object in Sumerian: An account of the history of their relationship. *Orientalia*, 68:215–253, 1999.