



Using a Query  
Language as an  
Annotation Tool

Eric J. M. Smith

Introduction

Corpora

Queries

Practical Example

Future Research

Discussion

Summary

# Using a Query Language as an Annotation Tool

**Eric J. M. Smith**

Department of Linguistics  
University of Toronto

American Association for Corpus Linguistics  
Provo, March 2008



# Introduction

Using a Query  
Language as an  
Annotation Tool

Eric J. M. Smith

Introduction

Motivation

Methodology

Corpora

Queries

Practical Example

Future Research

Discussion

Summary

- Existing corpora of cuneiform texts ill-suited for use in linguistic analysis
- This paper describes methodology for making linguistic data from such corpora more accessible



# Motivation

Using a Query  
Language as an  
Annotation Tool

Eric J. M. Smith

Introduction

Motivation

Methodology

Corpora

Queries

Practical Example

Future Research

Discussion

Summary

- Ultimate goal: studying the morphosyntax of agreement in Elamite and Sumerian
- Existing corpora for these languages provide orthography, not morphology
- How best to extract morphological information in the absence of annotation?



# Methodology

Using a Query  
Language as an  
Annotation Tool

Eric J. M. Smith

Introduction

Motivation

**Methodology**

Corpora

Queries

Practical Example

Future Research

Discussion

Summary

- Based on *LPath* query language (Bird et al., 2005, 2006)
- Enhance *LPath* by allowing *query objects* to be defined
- Construct library of reusable query objects to reflect language's morphology
- Query objects serve as a substitute for actual annotation



# Limitations of existing cuneiform corpora

Using a Query  
Language as an  
Annotation Tool

Eric J. M. Smith

Introduction

Corpora

Queries

Practical Example

Future Research

Discussion

Summary

- Lack of morphological markup
- Inadequate search functionality
- Lack of linguistic structure
- Orthography poorly reflects morphology
- Inconsistencies in transcription



# Electronic Text Corpus of Sumerian Literature (Black et al., 1998–2006)

Using a Query  
Language as an  
Annotation Tool

Eric J. M. Smith

Introduction

Corpora

Queries

Practical Example

Future Research

Discussion

Summary

- 394 literary texts, 170k words from ca. 2200–1600 BCE
- Transliterations, translations, bibliographical material, and some linguistic annotation
- XML source files made available by Oxford
- <http://www-etcs1.orient.ox.ac.uk>



# Structure of ETCSL

Using a Query  
Language as an  
Annotation Tool

Eric J. M. Smith

Introduction

Corpora

Queries

Practical Example

Future Research

Discussion

Summary

- Top-level `<text>`
- Intermediate groupings `<div1>`, `<lg>`
- Lines `<l>`
- Words `<w>`



# ETCSL word attributes

Using a Query  
Language as an  
Annotation Tool

Eric J. M. Smith

Introduction

Corpora

Queries

Practical Example

Future Research

Discussion

Summary

```
<w form="nu-gi4-gi4" lemma="gi4" pos="V"  
label="to return" form-type="RR">nu-gi4-gi4</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



# The LPath query language

(Bird et al., 2005, 2006)

Using a Query  
Language as an  
Annotation Tool

Eric J. M. Smith

Introduction

Corpora

Queries

LPath Query Language

Query Objects

Practical Example

Future Research

Discussion

Summary

- Based on XPath XML search language (Clark and DeRose, 1999)
- Adds linguistically-motivated operators:
  - 1  $->$  (immediate-following) and  $<-$  (immediate-preceding)
  - 2  $=>$  (immediate-following-sibling) and  $<=$  (immediate-preceding-sibling)
  - 3  $\wedge$  (left-edge alignment) and  $\$$  (right-edge alignment)
  - 4  $\{$  and  $\}$  (subtree-scoping)



# Sample LPath queries

(from Lai and Bird 2006)

Using a Query  
Language as an  
Annotation Tool

Eric J. M. Smith

Introduction

Corpora

Queries

LPath Query Language

Query Objects

Practical Example

Future Research

Discussion

Summary

① //S [//\_ [lex=saw]]

A sentence containing the word 'saw'.

② //VP/V->N

Nouns that immediately follow a verb which is a child of a verb phrase.

③ //VP{/V->N}

Within a verb phrase, nouns that immediately follow a verb which is a child of the given verb phrase.

④ //VP{/NP\$}

Noun phrases which are the rightmost child of a VP.

⑤ //VP{/NP\$}

NPs which are rightmost descendants of a VP.

⑥ //VP [{//^V->NP->PP\$}]

Verb phrases composed of a verb, an NP, and a PP.



# Converting ETCSL for use by LPath

Using a Query  
Language as an  
Annotation Tool

Eric J. M. Smith

Introduction

Corpora

Queries

LPath Query Language

Query Objects

Practical Example

Future Research

Discussion

Summary

- Pre-processing step (implemented in Python)
- Approximating paragraph boundaries from the English translation
- Extracting prefixes and suffixes

```
form="ba-an-ci-gir5-gir5-e", lemma="gir5"
```



```
@prefix="ba-an-ci-", @lemma="gir5",  
@suffix="-RED-e"
```



# Building on LPath

Using a Query  
Language as an  
Annotation Tool

Eric J. M. Smith

Introduction

Corpora

Queries

LPath Query Language

Query Objects

Practical Example

Future Research

Discussion

Summary

- In practice, LPath queries can get quite cumbersome
- e.g. the genitive case suffix *-ak*, which is never actually written as  $\rightarrow\rightarrow\rightarrow\rightarrow$   $\langle ak \rangle$ :
  - stem-final vowel assimilates to /a/ (e.g.  $\rightarrow\rightarrow\rightarrow$   $\langle \tilde{g}a \rangle$  after stems ending in / $\tilde{g}u$ /)
  - $\rightarrow\rightarrow$   $\langle la \rangle$  after stems ending in /l/.
  - $\rightarrow\rightarrow$   $\langle na \rangle$  after stems ending in /n/.
  - $\rightarrow\rightarrow\rightarrow$   $\langle ra \rangle$  after stems ending in /r/.
  - sometimes written as  $\rightarrow\rightarrow$   $\langle a \rangle$
  - only reflects the /k/ when before another suffix (e.g.  $\rightarrow\rightarrow\rightarrow\rightarrow\rightarrow\rightarrow\rightarrow\rightarrow\rightarrow\rightarrow$   $\rightarrow\rightarrow\rightarrow$   $\rightarrow\rightarrow\rightarrow\rightarrow$   $\langle lugal-la-ke_4 \rangle$  ‘of the king-ERG’)
  - etc.
- Not practical to perform these queries every time we want to refer to a genitive-case noun



# Defining query objects

Using a Query  
Language as an  
Annotation Tool

Eric J. M. Smith

Introduction

Corpora

Queries

LPath Query Language

Query Objects

Practical Example

Future Research

Discussion

Summary

- Incrementally building up N-gen using a series of queries:
  - 1 //N[@lemma like "%ju" and @form like "%ja"]
  - 2 //N[@lemma like "%l" and @suffix like "-la%"]
  - 3 //N[@lemma like "%n" and @suffix like "-na%"]
  - 4 //N[@lemma like "%r" and @suffix like "-ra%"]
  - 5 //N[@suffix like "%a-ke4"]
  - 6 etc.
- The results of each query can be verified for correctness before adding it to the definition.
- At the end of the process, N-gen is a first-class member of the corpus.



# Complex query objects

Using a Query  
Language as an  
Annotation Tool

Eric J. M. Smith

Introduction

Corpora

Queries

LPath Query Language

Query Objects

Practical Example

Future Research

Discussion

Summary

- More-complex query objects can be built up from simpler ones:
  - 1 N-erg defined as `//N[@suffix = "-e"]`
  - 2 NP-erg as N-erg
  - 3 NP-erg also as `//N <-- N-gen[@suffix like "%-ke4"]`
- Once NP-erg has been fully defined, we have effectively added a level of hierarchy to the corpus.



# A limitation of LPath

Using a Query  
Language as an  
Annotation Tool

Eric J. M. Smith

Introduction

Corpora

Queries

LPath Query Language

Query Objects

Practical Example

Future Research

Discussion

Summary

- XPath is designed to locate individual nodes
- A query like `//N <-- N-gen[@suffix like "%-ke4"]` finds an N, not an NP
- LPath implementation does provide access to all nodes identified in the query
- New nodes for higher-level structures have to be “spliced” into database representation



# Example: Terminative case agreement

Using a Query  
Language as an  
Annotation Tool

Eric J. M. Smith

Introduction

Corpora

Queries

Practical Example

Future Research

Discussion

Summary

- Sumerian verbs have prefixes which agree with verb's oblique arguments:

MODAL - CONJ - DAT - COM -  $\left\{ \begin{array}{l} \text{ABL} \\ \text{TERM} \end{array} \right\}$  - LOC - ERG - verb - ABS

- Original study of “dimensional infixes” by Gragg (1973)
- Goal: using queries to match up noun-suffixes with verb-prefixes

𒂗𒂗𒂗𒂗𒂗𒂗𒂗𒂗𒂗𒂗𒂗𒂗

*saḡ-ki*  
forehead

𒂗𒂗𒂗𒂗𒂗𒂗𒂗𒂗𒂗𒂗𒂗

*zalag-ga-ni*  
shining-3SG.POSS

𒂗𒂗𒂗𒂗𒂗𒂗𒂗𒂗𒂗𒂗𒂗

*ḡa<sub>2</sub>-a-še<sub>3</sub>*  
1SG-TERM

𒂗𒂗𒂗𒂗𒂗𒂗𒂗𒂗𒂗𒂗𒂗

*hu-mu-ši-in-zig<sub>3</sub>*  
hu-mu-TERM-ERG.3SG-lift

‘she lifted her radiant forehead to me’



# Query objects for terminative case

Using a Query  
Language as an  
Annotation Tool

Eric J. M. Smith

Introduction

Corpora

Queries

Practical Example

Future Research

Discussion

Summary

- N-term defined as `//N[@suffix like "%-ce3"]`
- V-term defined as `//V[@prefix like "%ci-%"`  
or `@prefix like "%ce3-%"]`
- Ideally: `//S{N-term <-- V-term}`
- Realistically: `//PARA{N-term <-- V-term}`
  - For this problem, good enough
  - 100% recall more important than precision
- No need to annotate MODAL, CONJ, or other prefixes not relevant to the problem at hand



# Future Research

Using a Query  
Language as an  
Annotation Tool

Eric J. M. Smith

Introduction

Corpora

Queries

Practical Example

Future Research

Discussion

Summary

- 1 Sumerian oblique-case agreement  
How can we explain exceptions to usual agreement patterns?
- 2 Elamite object and indirect object agreement  
How do verbal prefixes agree with object and/or indirect object?
- 3 Elamite possessive constructions  
How is choice of construction motivated by semantics and/or phonology?
- 4 Sumerian conjugation prefixes  
What are semantic motivations for choice of conjugation prefixes?



# Discussion

Using a Query  
Language as an  
Annotation Tool

Eric J. M. Smith

Introduction

Corpora

Queries

Practical Example

Future Research

Discussion

Summary

- Aim for 100% recall (precision can be sacrificed)
- Problem-specific (only create annotations which are actually used)
- Insulates linguist from peculiarities of orthography
- Approach can be applied to other low-resource languages



# Query Languages as Quick & Dirty Annotation

Using a Query  
Language as an  
Annotation Tool

Eric J. M. Smith

Introduction

Corpora

Queries

Practical Example

Future Research

Discussion

Summary

**By defining a library of reusable query objects, it is possible to get many of the advantages of annotation without actually having to annotate.**