2008-03-15

Compiling and Annotating a Syriac Corpus

George Busby
bazubii@gmail.com

James Carroll

Marc Carmen

Carl Griffin

Robbie Haertel

See next page for additional authors

Original Publication Citation

Eric Ringger, Kristian Heal, Carl Griffin, Kevin Seppi, Deryle Lonsdale, David Taylor. "Compiling and Annotating a Syriac Corpus." Presentation given at the Conference of the American Association for Corpus Linguistics (AACL 28). Provo, Utah. March 15, 28.

BYU ScholarsArchive Citation

Busby, George; Carroll, James; Carmen, Marc; Griffin, Carl; Haertel, Robbie; Heal, Kristian; Heaton, Joshua; Lonsdale, Deryle W.; McClanahan, Peter; Ringger, Eric K.; Seppi, Kevin; and Taylor, David, "Compiling and Annotating a Syriac Corpus" (2008). *Faculty Publications*. 1297.
https://scholarsarchive.byu.edu/facpub/1297

Compiling and Annotating a Syriac Corpus

Eric Ringger

Kristian Heal, Carl Griffin (BYU CPART)

Peter McClanahan, Robbie Haertel, James Carroll, George Busby, Kevin Seppi (CS)

Deryle Lonsdale , Marc Carmen, Joshua Heaton (Linguistics)

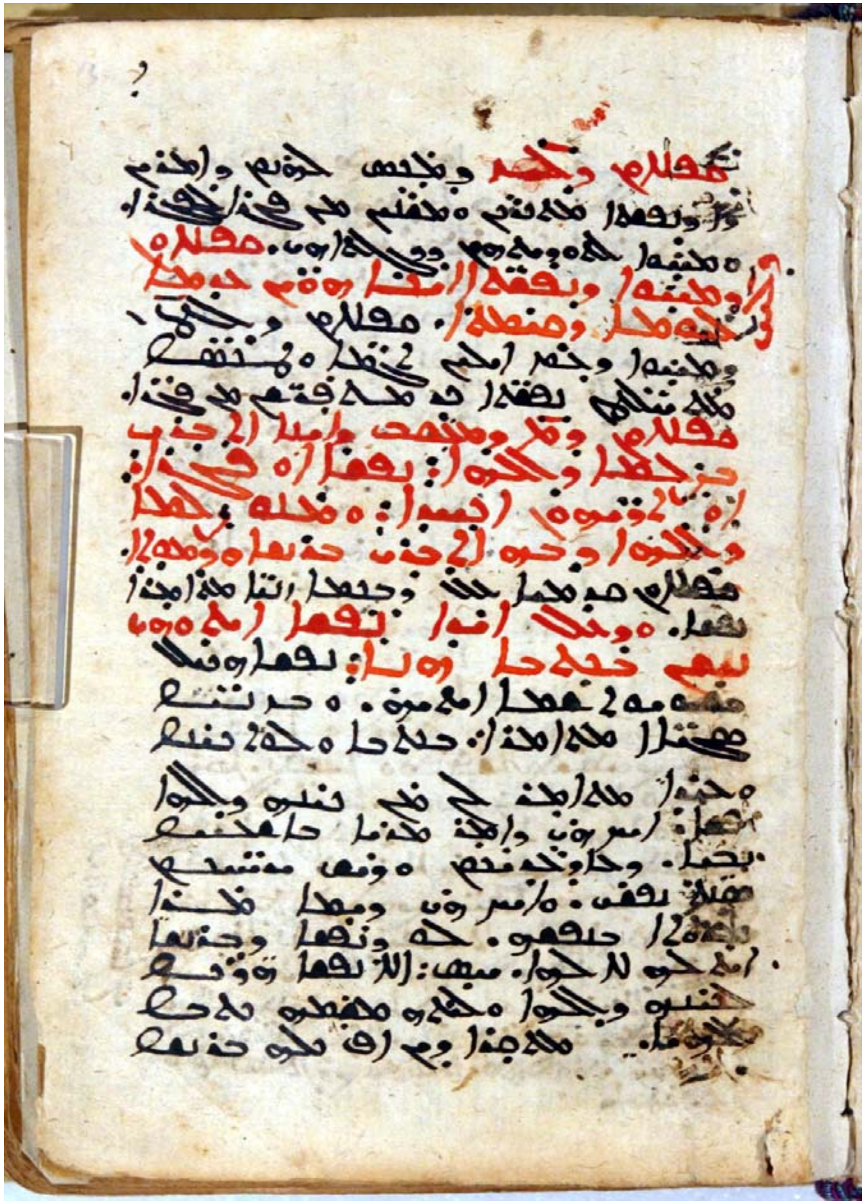
David Taylor (Oxford)

March 15, 2008 – AACL

Borgia 13

13th century manuscript, a Melkite Euchologion

Vatican Syriac 147

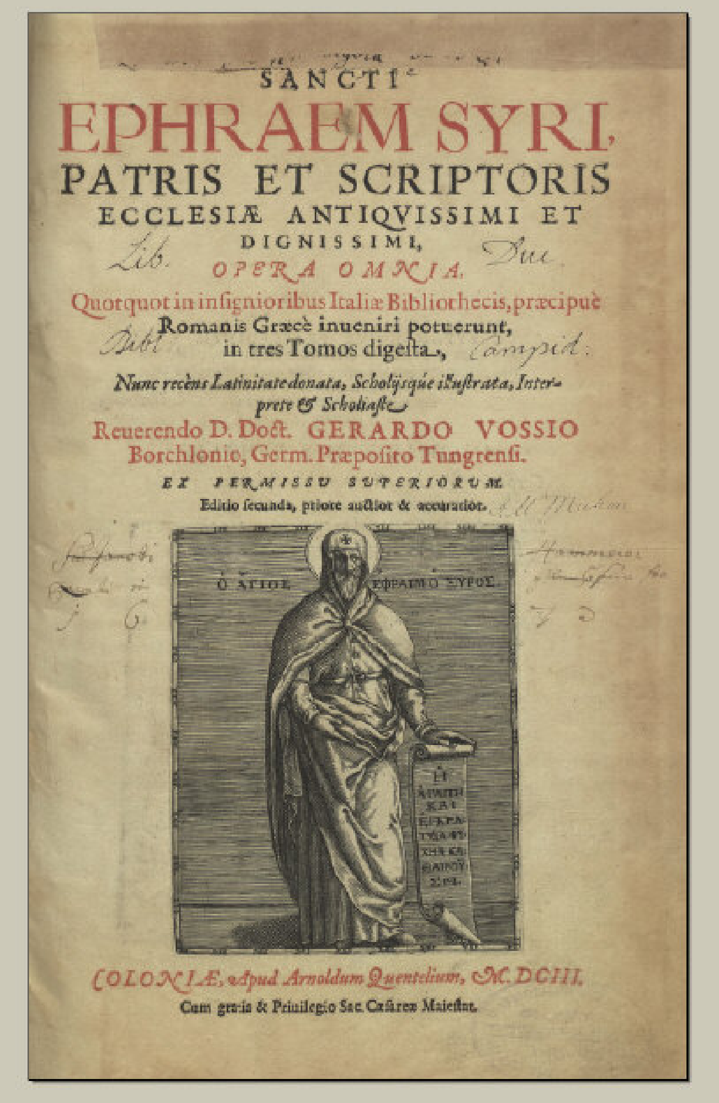


Overview

- Project Objectives
- Corpus and Lexical Resources
- Morphological Tools and Markup
- Reduction of Annotation Costs
- Review Process
- Conclusions

Project Objectives (I)

- Create a digital and print concordance of all of the works of Ephrem the Syrian (d. 373 AD)
- 0.5 million word corpus



SANCTI

EPHRAEM SYRI,

PATRIS ET SCRIPTORIS

ECCLESIÆ ANTIQVISSIMI ET

DIGNISSIMI,

Lib. OPERA OMNIA.

Quotquot in inſignioribus Italiæ Bibliothecis, præcipuè

Bibl Romanis Græcè inueniri potuerunt,

in tres Tomos digeſta, Campid.

Nunc recèns Latinitate donata, Scholijsque illuſtrata, Interprete & Scholiaſte

Reuerendo D. Doct. GERARDO VOSSIO

Borchlonio, Germ. Præpoſito Tungrenſi.

EX PERMISSU SUPERIORUM.

Editio ſecunda, priore auctior & accuratior.

COLONIÆ, Apud Arnoldum Quentelium, M. DCIII.

Cum gratia & Priuilegio Sac. Cæſareæ Maieſtat.

Project Objectives (2)

- Create an annotated digital corpus of all Syriac literature
  - From the 2nd Century
  - To the 20th Century
  - On the order of 50 million words
- Transform Syriac scholarship
  - Enable new insights
  - Discover new literary , theological, and historical connections

Near-term Objectives

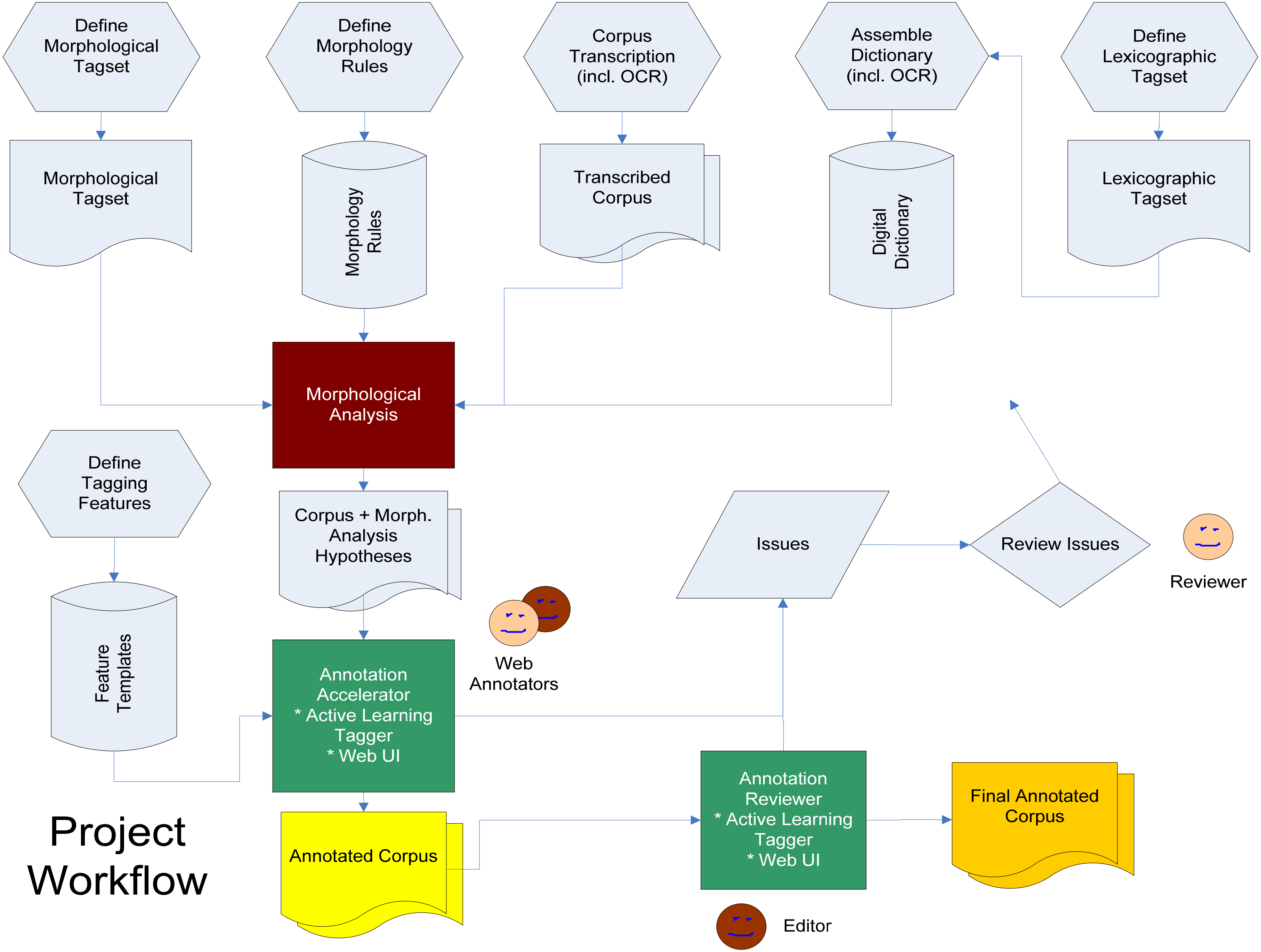
- Develop infrastructure for Syriac corpus development
  - Digital text acquisition
  - Lexical resources
  - Linguistic annotations
  - Morphological analysis and disambiguation
  - User interface
- Provide motivation for cost-conscious active learning for annotation

Syriac in One Slide

- Northwest Semitic
- Dialect of Aramaic
- Three scripts
- Reads right to left
- Highly inflective
- Texts are largely unvocalized
- Primarily a literary and ecclesiastical language beginning in the 9th century

ܒܪܺܫܺܝܬ ܐܺܝܬܰܘܗܝ݈ ܗ݈ܘܳܐ ܡܶܠܬܳܐ
. ܘܗܽܘ ܡܶܠܬܳܐ . ܐܺܝܬܰܘܗܝ݈ ܗ݈ܘܳܐ
ܠܘܳܬ ܐܰܠܳܗܳܐ . ܘܰܐܠܳܗܳܐ
ܐܺܝܬܰܘܗܝ݈ ܗ݈ܘܳܐ ܗܽܘ ܡܶܠܬܳܐ .

Project Workflow



Define Morphological Tagset

Morphological Tagset

Define Morphology Rules

Morphology Rules

Corpus Transcription (incl. OCR)

Transcribed Corpus

Assemble Dictionary (incl. OCR)

Digital Dictionary

Define Lexicographic Tagset

Lexicographic Tagset

Morphological Analysis

Define Tagging Features

Feature Templates

Corpus + Morph. Analysis Hypotheses

Annotation Accelerator
* Active Learning Tagger
* Web UI

Web Annotators

Annotated Corpus

Issues

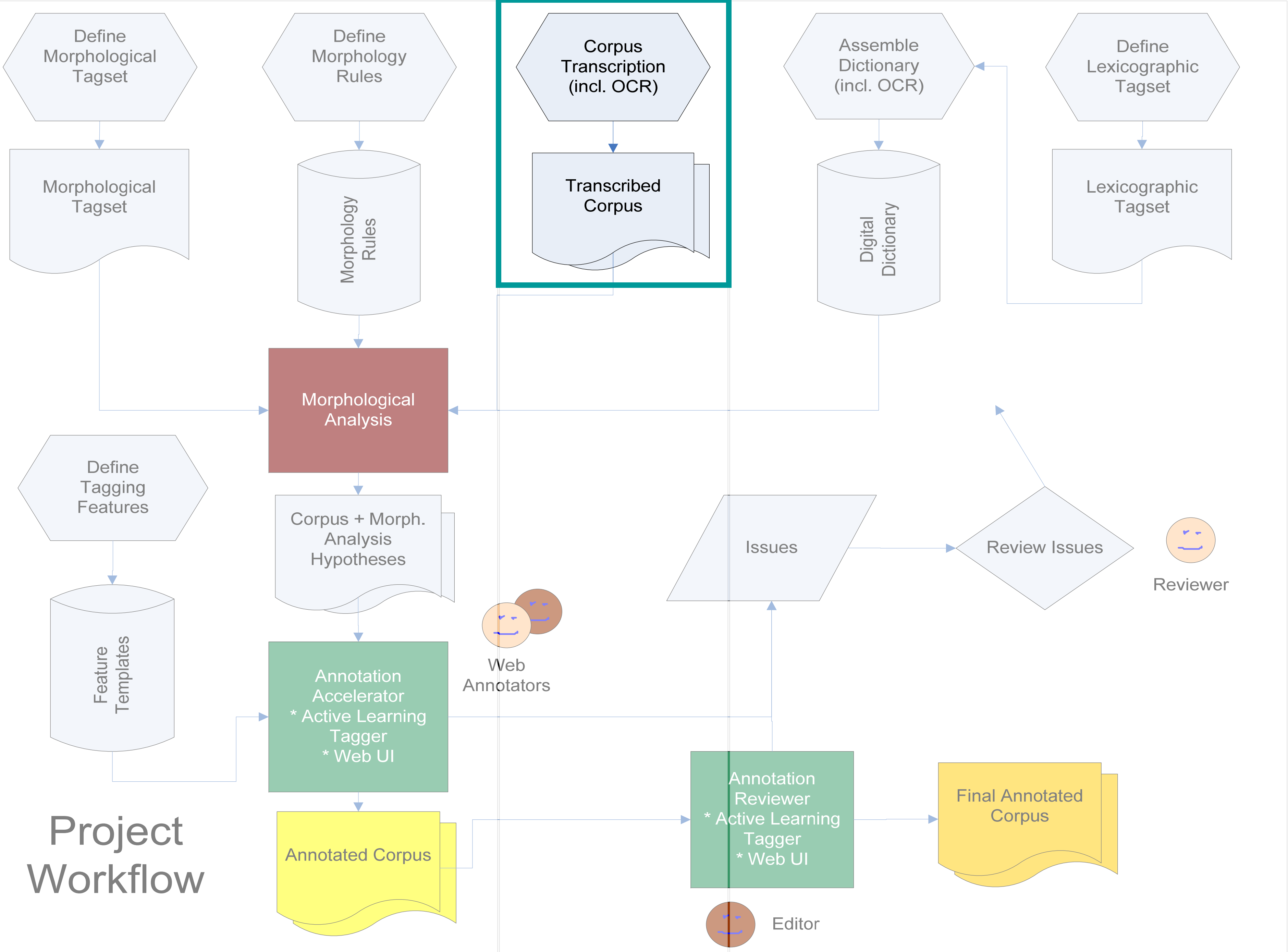
Review Issues

Reviewer

Annotation Reviewer
* Active Learning Tagger
* Web UI

Editor

Final Annotated Corpus


Define Morphological Tagset

Define Morphology Rules

Corpus Transcription (incl. OCR)

Assemble Dictionary (incl. OCR)

Define Lexicographic Tagset

Morphological Tagset

Morphology Rules

Transcribed Corpus

Digital Dictionary

Lexicographic Tagset

Morphological Analysis

Define Tagging Features

Corpus + Morph. Analysis Hypotheses

Issues

Review Issues

Reviewer

Feature Templates

Web Annotators

Annotation Accelerator
* Active Learning Tagger
* Web UI

Annotation Reviewer
* Active Learning Tagger
* Web UI

Final Annotated Corpus

Project Workflow

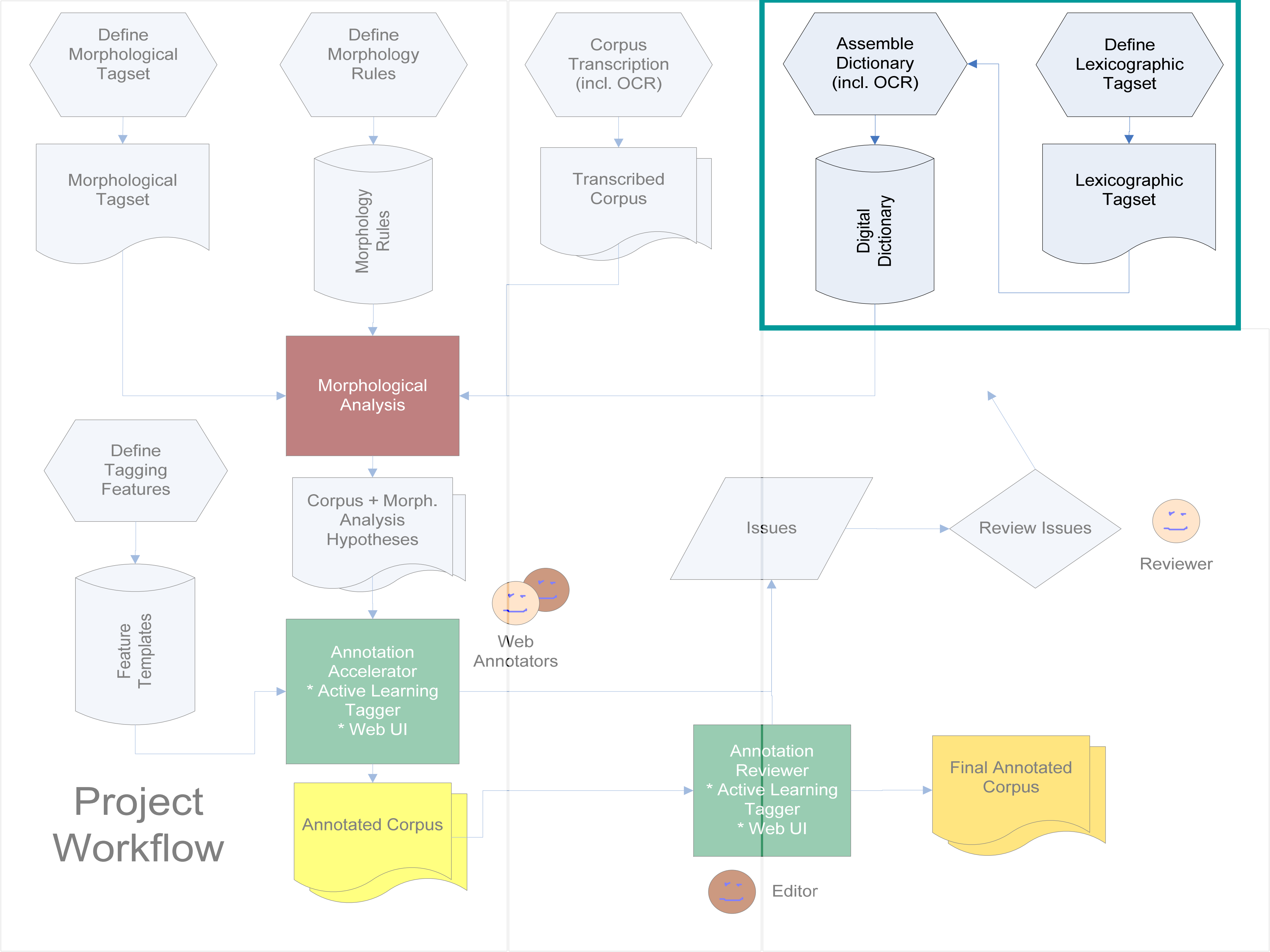
Annotated Corpus

Editor

Corpus Transcription

- Digitization of Syriac-script texts is in progress
  - By human transcription
  - By Syriac OCR (Clocksin)
  - Post-editing also in progress
- Works of Ephrem the Syrian are complete
- 5 million total words transcribed to date

Project Workflow



Define Morphological Tagset

Morphological Tagset

Define Morphology Rules

Morphology Rules

Corpus Transcription (incl. OCR)

Transcribed Corpus

Assemble Dictionary (incl. OCR)

Digital Dictionary

Define Lexicographic Tagset

Lexicographic Tagset

Morphological Analysis

Define Tagging Features

Feature Templates

Corpus + Morph. Analysis Hypotheses

Annotation Accelerator
* Active Learning Tagger
* Web UI

Web Annotators

Annotated Corpus

Issues

Review Issues

Reviewer

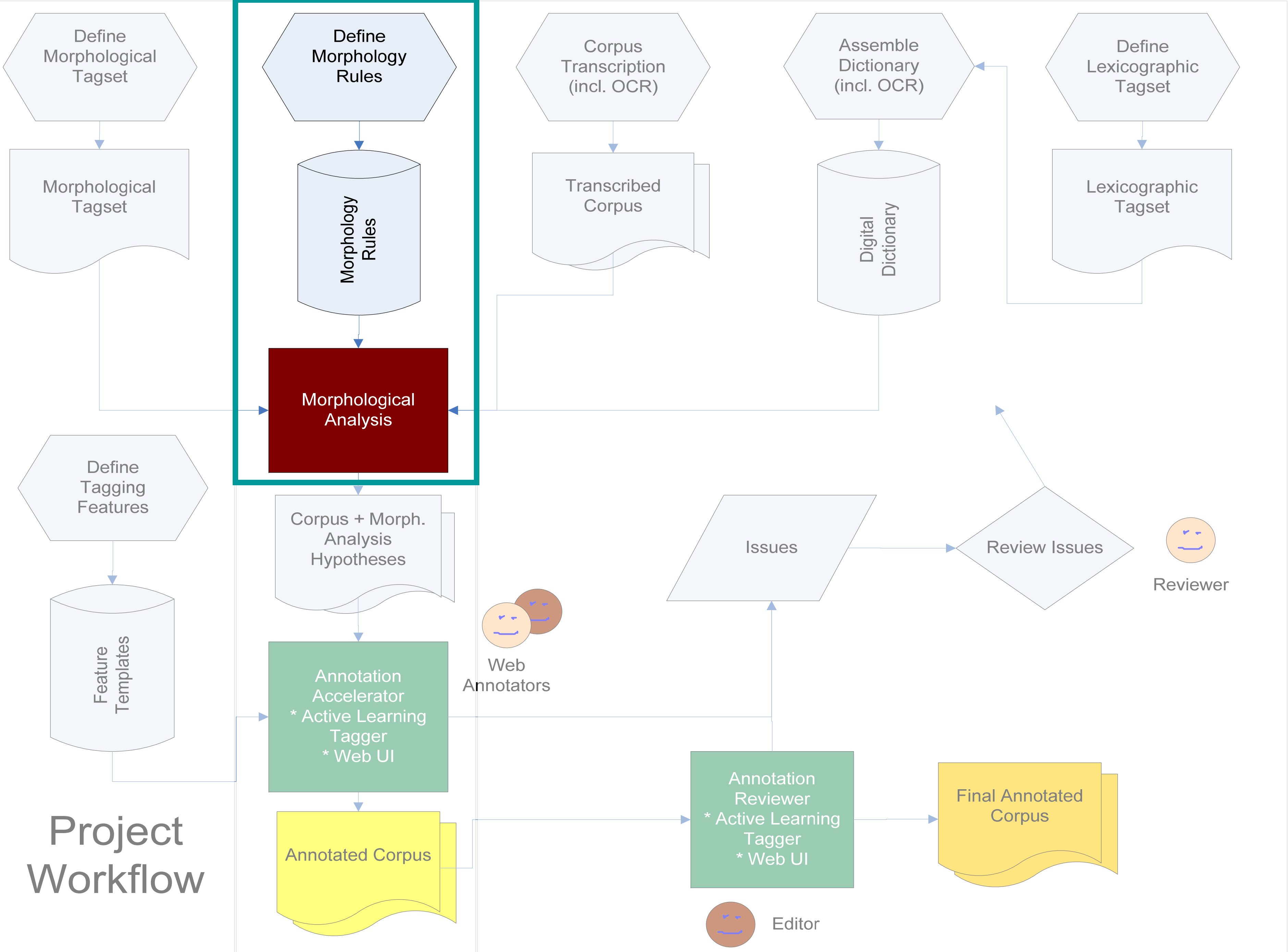
Annotation Reviewer
* Active Learning Tagger
* Web UI

Editor

Final Annotated Corpus

Lexical Resources

- Comprehensive digital dictionary in-progress
  - Based on the print dictionary of Payne-Smith
  - Augmented by other print dictionaries
  - Coverage will grow from traditional texts to newly acquired corpora
- Common resource both for computational tools and human consultation
- Encoded with XML markup (TEI)
- GUI for online access



Define Morphological Tagset

Define Morphology Rules

Corpus Transcription (incl. OCR)

Assemble Dictionary (incl. OCR)

Define Lexicographic Tagset

Morphological Tagset

Morphology Rules

Transcribed Corpus

Digital Dictionary

Lexicographic Tagset

Morphological Analysis

Define Tagging Features

Corpus + Morph. Analysis Hypotheses

Issues

Review Issues

Reviewer

Feature Templates

Web Annotators

Annotation Accelerator
* Active Learning Tagger
* Web UI

Annotation Reviewer
* Active Learning Tagger
* Web UI

Final Annotated Corpus

Annotated Corpus

Editor

Project Workflow

Morphological Analysis

- Input: Syriac text
  - Currently romanized
- Output: all possible morphological parse(s)
- Method: Finite-state morphology

Finite-State Morphology

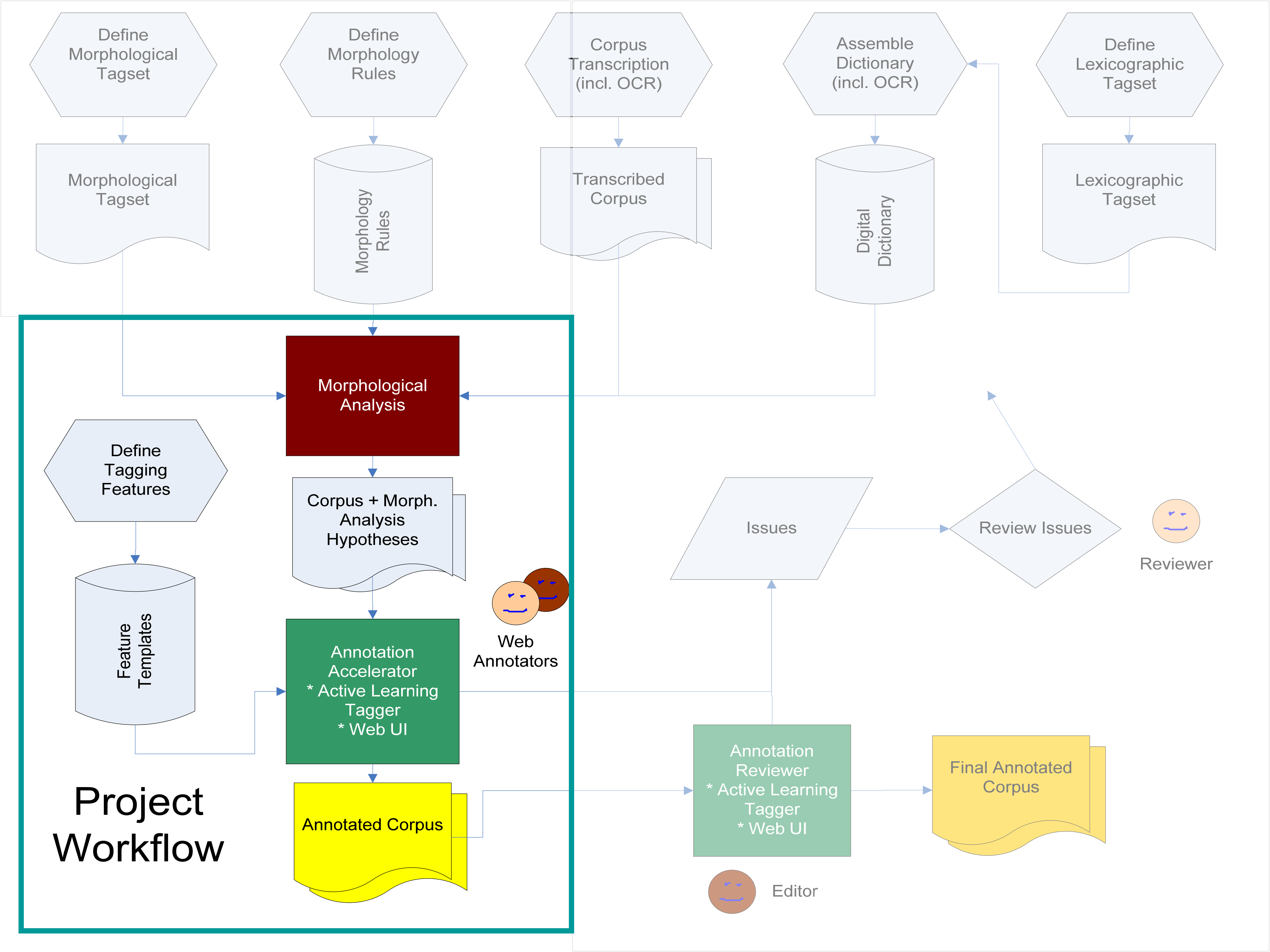
- Word formation viewed as generative process
  - From morphemes to words
  - Produced by a finite-state transducer
- Auto-segmental approach
  - Root tier
  - Consonant-Vowel tier
  - Vocalization tier
- Knowledge-engineered
  - Lexicons for roots, morphemes
  - Rules for word formation, interdigitation
- Xerox XFST toolkit and techniques (Beesley & Karttunen, 2003)
- Prior work by Kiraz (1993)
  - Currently using the Kiraz categories and attributes

Parsing morphological structure

```
xfst[1]: up mono
[PronQu+impers-4]

xfst[1]: up 1ayleyn
[PronQu+wh+pl]

xfst[1]: up 1awkelDDyenhy
[^1kl-P3+Aphel+Perf+pl+3+f-3=PronSubj+enc+3sg+f]

xfst[1]: up qTal
[^qTl-P1a+Ethpeel+Perf+pl+3+f-2]
[^qTl-P1a+Peal+Perf+sg+3+m]
[^qTl-P1a+Peal+Perf+pl+3+f-2]

xfst[1]: up 1ekal
[^1kl-P3+Peal+Perf+sg+3+m]
[^1kl-P3+Peal+Perf+pl+3+f-1]

xfst[1]: up ne1kuwl
[^1kl-P3+Peal+Imperf+sg+3+m]
[^1kl-P3+Peal+Imperf+pl+1]
```

Current Status of Morphology

- About 1500 lexical items
- Several hundred rules (mostly verbal)
- Remaining issues:
  - Working directly with Unicode
  - Some derivational patterns
  - Verb object-suffixes and effects on vowels
  - Diacritics
  - Partial vocalization



Define Morphological Tagset

Morphological Tagset

Define Morphology Rules

Morphology Rules

Corpus Transcription (incl. OCR)

Transcribed Corpus

Assemble Dictionary (incl. OCR)

Digital Dictionary

Define Lexicographic Tagset

Lexicographic Tagset

Morphological Analysis

Define Tagging Features

Corpus + Morph. Analysis Hypotheses

Feature Templates

Annotation Accelerator
* Active Learning Tagger
* Web UI

Web Annotators

Annotated Corpus

Project Workflow

Issues

Review Issues

Reviewer

Annotation Reviewer
* Active Learning Tagger
* Web UI

Editor

Final Annotated Corpus

Linguistic Annotation

- Linguistic information associated with each word (token):
  - Maximally disambiguated morphological analysis(es)
  - Including grammatical category
  - Vocalization (to varying degree)
    - Depends on metrical demands
- Not a trivial task, even for trained annotators

Accelerating Corpus Annotation

- Reduce the total cost of human annotation efforts without compromising accuracy
- Use probabilistic models for computer-aided tagging
    - In particular, for morphological disambiguation
- Use active learning
    - (Seung et al, 1992; Thrun et al., 1992)
- Still requires human expertise for selected examples
- More details in:
    - LAW 2007
    - LREC 2008

Our Tagging Approach

- Use a state-of-the art tagger:
  - Maximum Entropy tagger (Rathnaparkhi, 1995)
  - aka Maximum Entropy Markov Model (MEMM)
  - aka Conditional Mark Model (CMM) trained locally by Maximum Entropy learner
- Requirements: Syriac morphological tag set, annotated data, “feature” templates for classification, human oracles

Features for Tagging

- Combination of lexical, orthographic, contextual, morphological, and frequency-based information
- For each word:
  - The textual form of the word itself
  - Tags of the preceding two words
  - The textual form of the following word
  - Diacritics
  - Arbitrary variable-length word prefixes and suffixes
- Following Toutanova & Manning (2000)

Syriac Labels

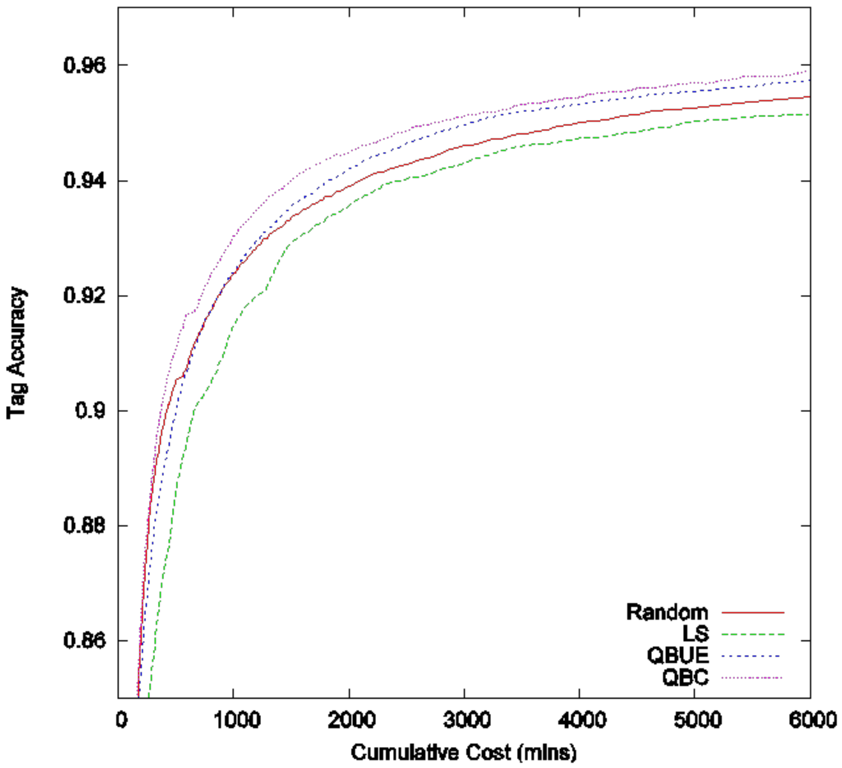
Tags	Features
Enclitic	Vocalized word (vowels)
Suffix – gender, person, number, suffix/contraction	Word – seyame
Word – gender, person, number, state, tense, form	Word – lexeme flag
Lexeme – grammatical category	Lexeme – lexeme
Lexeme – first, second, third, fourth suffix	Lexeme – seyame
Lexeme – prefix	Word type
Lexeme – form	Lexeme – vowel pattern
Root – root type	Lexeme – number of vowels
	Lexeme – radical type
	Root

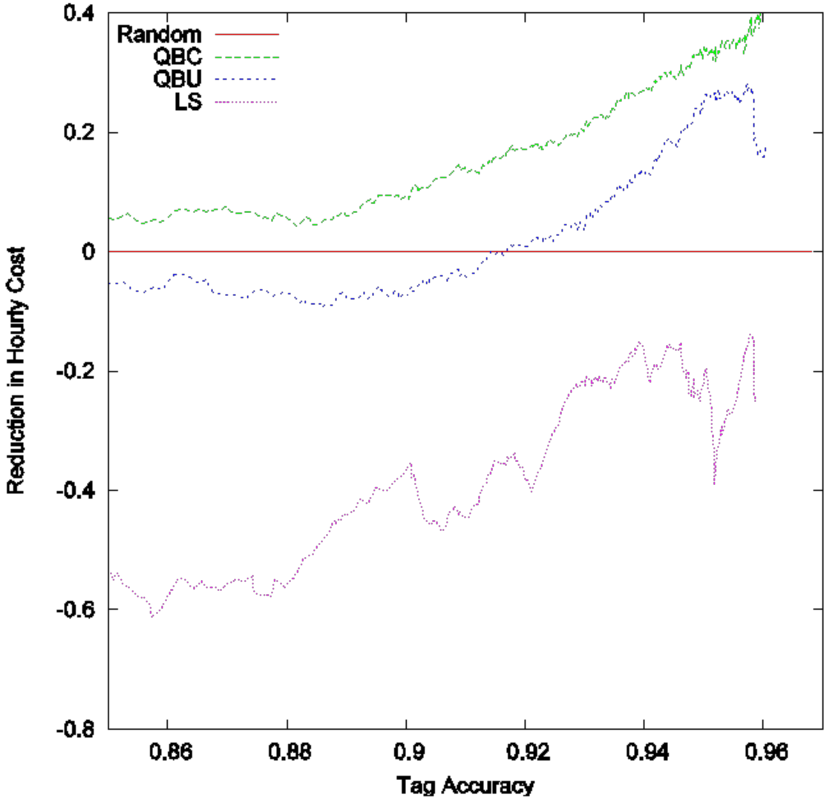
3403 distinct tags (not including features)

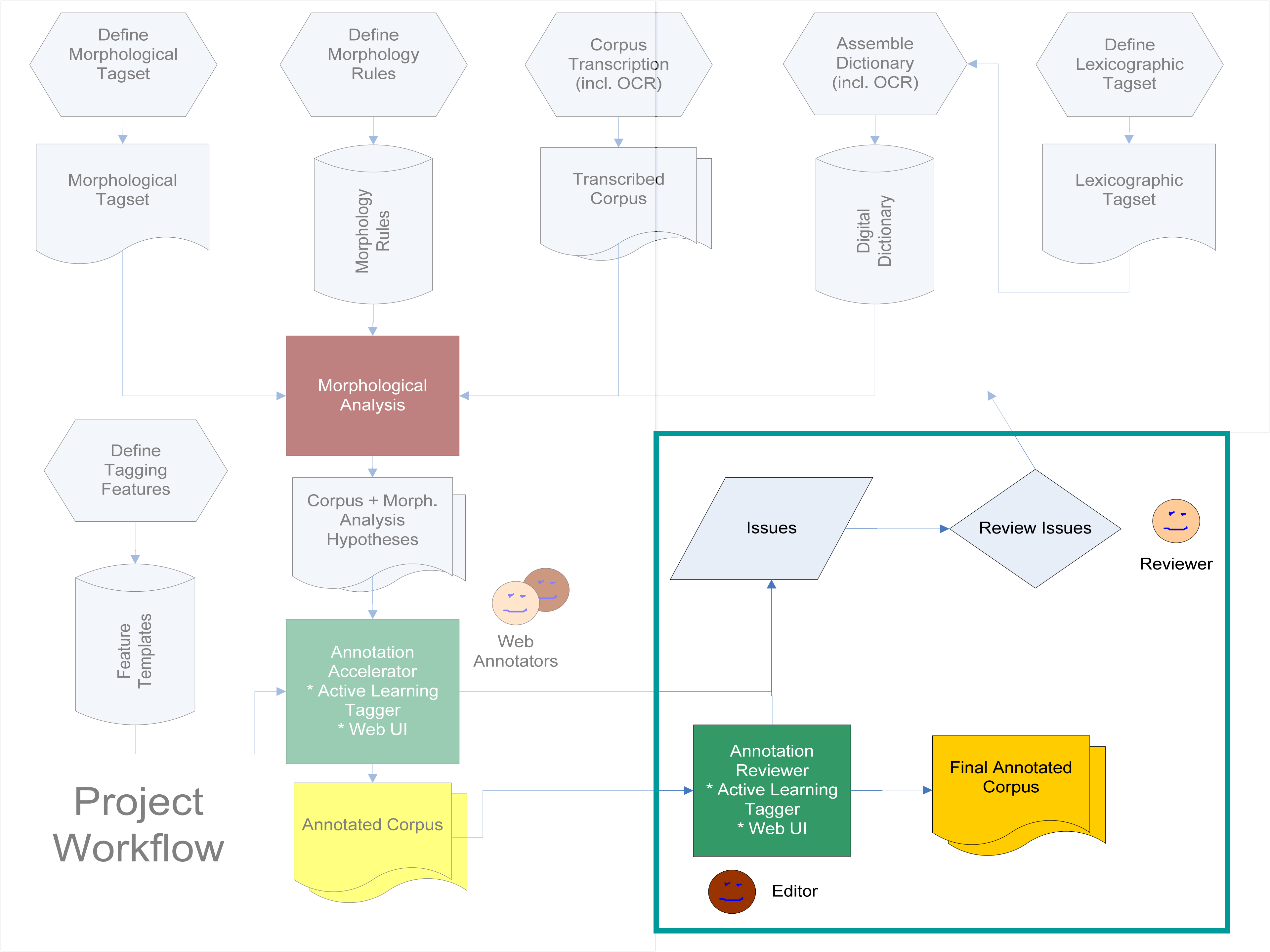
Active Learning

- Goal: produce annotated corpora with least possible time and annotator effort
- Method
  - Use probabilistic tagger to annotate new data
  - Find most informative sentences/words
  - Ask oracle (human annotator) for answer
  - Use the answer to retrain the tagger
  - Repeat the process until cost limit reached
- Developed for English, now applying to Syriac
  - Details and extensive results presented for the group earlier this afternoon by Peter McClanahan

Active Learning Results (Short Version!)







Define Morphological Tagset

Define Morphology Rules

Corpus Transcription (incl. OCR)

Assemble Dictionary (incl. OCR)

Define Lexicographic Tagset

Morphological Tagset

Morphology Rules

Transcribed Corpus

Digital Dictionary

Lexicographic Tagset

Morphological Analysis

Define Tagging Features

Corpus + Morph. Analysis Hypotheses

Feature Templates

Annotation Accelerator
* Active Learning Tagger
* Web UI

Web Annotators

Annotated Corpus

Issues

Review Issues

Reviewer

Annotation Reviewer
* Active Learning Tagger
* Web UI

Final Annotated Corpus

Editor

Project Workflow

Review Process

- Use active learning framework for editorial review of transcriptions and annotations
- Review issues raised during annotation for feedback to upstream components

Conclusions

- Now developing tools and resources for Syriac language processing
- Accelerating corpus annotation in novel ways
  - Therefore, minimizing cost
- Deliverables of interest to Syriac scholars:
  - Digital and print concordance of the works of Ephrem the Syrian
  - Large annotated Syriac corpus
- Interface specifics still undetermined
  - Seeking best practices and advice

Questions?

Eastern Christian Texts

Brigham Young University

Natural Language NLP Processing Lab