MorphAdorner v2.0

Philip R. Burns

February 5, 2014

NORTHWESTERN UNIVERSITY

MorphAdorner Web Site

- http://morphadorner.northwestern.edu/
- Documentation in HTML, PDF, EPUB, and Kindle MOBI format
- Many online examples of MorphAdorner facilities using MorphAdorner v2 server

Why “Adornment”

- Terms like “annotation” and “tagging” have too many different meanings
- “Adornment” harkens back to medieval sense of adorning or illuminating a manuscripts: attaching notes and images to words, lines, paragraphs, and pages
- Morphological adornment is the process of adorning with parts of speech, lemmata, etc.

Audience

- Intended as a programmer's construction kit
- Can be used to create customized end-user programs for morphological adornment
- Main driver and utilities are command line programs
- Server can be accessed through the web

History

- MorphAdorner used in WordHoard, Monk, and Virtual Orthographic Normalization projects
- First public release in April 2009 under NCSA style open source license
- Updated training data during 2010 and 2011
- Added MorphAdorner server in 2013
- Many other improvements in 2013 academic year

Advantages

- More aware of potential problems and pitfalls than other existing software for adornment of literary texts
- NUPos tag set allows adornment of English texts from Middle English to present (diachronic corpora)
- Licensed under a very non-restrictive NCSA style open source license
- Already integrated with Abbot (TEI normalization) and Philologic (search)

Basic Adornment Processes

- Tokenization
- Sentence boundary recognition
- Spelling normalization
- Part of speech assignment
- Lemmatization

If you get these correct, you can do a LOT of other things.

Other Facilities

- Hyphenation
- Language Recognition
- Name Extraction and Standardization
- Parsing
- Pluralization
- Stemming

Other Facilities

- Syllabification
- Text Segmentation
- Text Summarization
- Thesaurus (synonyms and antonyms)
- Verb Conjugation

V2 Enhancements

- TEI specific SGML to XML converter while waiting for Abbot update.
- Change log system for tracking token-based changes from one edition of a TEI file to another.
- Improved detection and regularization of soft hyphens.
- Corpus Indexer for adorned files using BlackLab.
- Partial element-aware adornment.

V2 Enhancements

- MorphAdorner Server which exposes MorphAdorner facilities as HTTP-based web services.
- Improved language detection for text fragments.
- Better tokenization and sentence breaking.
- Tokenization and sentence breaking for non-English languages.
- Improved spelling standardization.
- Better TEI P5 compliance for adorned output.

V2 Enhancements

- Abbreviation extraction using PUNKT algorithm.
- Improved “guessing” of part of speech values for unknown words.
- More output formats (Sketch engine, etc.)
- Updated training data for early modern English.
- Miscellaneous internal improvements and optimizations.

MorphAdorner Server

- Exposes some MorphAdorner facilities as web services
- Accessible from any program/language which can send and receive text over an HTTP connection
- Supports CORS
- Sample server at http://devadorner.northwestern.edu/maserver/

Book of English

- Large, growing, collaboratively curated, and public domain corpus of written English since its earliest modern form
- With full bibliographical detail
- And light but consistent structural and linguistic annotation.

Comprehensive Word Lexicon

- Spelling and variants with date information
- Frequencies of occurrence across centuries
- Frequencies of occurrence for specific genres
- Frequencies of occurrences by part of speech
- Lemmata by part of speech
- Allows morphological adornment processes to use a standardized lexicon ID

TCP Texts Processed

- 44,420 EEBO (Early English Books Online) texts (first 25,000 to be publicly available in 2015)
- 4,585 Evans texts (to be publicly available in June 2014)
- 2,091 ECCO texts (already publically available)
- Nebraska has the ECCO texts available in adorned and unadorned format: http://abbot.unl.edu/abbot-morphadorner/index.html

Early Modern Drama

- EEBO contains transcriptions of about 800 plays written by contemporaries and near-contemporaries of Shakespeare, e.g., from before 1660.
- 631 of those TEI transcriptions have been manually corrected by Martin and his students over the past couple of years.
- 400th anniversary of Shakespeare’s death in 2016: opportunity to produce “good” collection of these plays.

Thomason Collection

- >22,000 pamplets, broadsides, books, etc. from 1640 to 1661.
- About 7,050 currently transcribed in EEBO collection.
- Collected together by bookseller George Thomason (d. 1666). Bound into 2,000 volumes.
- Major source for political, social, religious, and military history from end of reign of Charles I through the Restoration of Charles II.

Sample Text: Fair Em

- “Fair Em, the Miller's Daughter of Manchester” a comedy
- Written about 1590
- Author unknown: possible Anthony Munday or Robert Wilson

Fair Em: EPUB3

- EPUB3 is an HTML based publishing format popular for handheld devices (phones, tablets) as well as desktop systems.
- MorphAdorned texts can be converted to EPUB3 format using XSLT stylesheets.
- EPUB3 files can be created from the original text or the standardized text.

Output Types

- Adorned and unadorned TEI
- Plain text, main/paratext selectable
- Verticalized tabular files and summaries
- HTML, Epub/EPub3, DocBook, MOBI, Latex, Markdown, PDF
- Sketch, TCF

Corpus Search

- Philologic v4.0 processes MorphAdorned files directly
- BlackLab library provides customizable and embeddable corpus search facilities
- Locally written CorpusIndexer to produce BlackLab index of MorphAdorned files
- Example BlackLab based search:

  http://devadorner.northwestern.edu/corpussearch/

Questions?