# A Morphological Analyzer for Old English Verbs

11-712 Final Project Report
Jason Adams

May 16, 2007

## 1 Introduction

After the fall of Roman influence in the British Isles at the beginning of the $5^{\text{th}}$ Century A.D., two West Germanic tribes began pouring into Britain in increasing numbers. Originally recruited as mercenaries against the Picts in modern-day Scotland, the Angles and Saxons stayed and eventually overwhelmed the Celts and other inhabitants of modern-day England and Wales who had invited them (Wikipedia, 2007a). With them came their language, an offshoot of the Germanic language family. Old English continued to be spoken in Britain until just after the Norman Invasion in 1066 A.D., when it began to transition quickly into Middle English due to the Norman French influence. There were four main dialects of Old English: Northumbrian, Mercian, Kentish, and West Saxon. In later centuries, Viking invasions increased the influence of the kingdom of Wessex, the seat of Alfred the Great, who united the Anglo-Saxons against the invaders. As a result, a large amount of written text in the West Saxon dialect survived (Wikipedia, 2007b). Therefore, this dialect is the one predominantly taught first to modern learners of Old English, who consist primarily of historians, English literature students, and hobbyists.

The purpose of this project is to create a morphological analyzer for Old English verbs. As an Indo-European language, Old English inflected many verbs using stem-vowel changes. However, it also contains a Proto-Germanic innovation for tense inflection that consists of adding a dental suffix in the preterite and past participle forms. This innovation became the productive form for Old English verbs and constitutes about three-fourths of all verbs (Quirk and Wrenn, 1994). These two methods for inflecting verbs form the basis for the categorization of Germanic verbs into *weak* (those descended from the Proto-Germanic innovation) and *strong* (those descended from Indo-European). The complexity of Old English verbal inflections makes learning Old English verbal morphology difficult for modern speakers. A morphological analyzer can assist the learning process by automatically tagging verb forms. Others who might benefit from such a tool include historical and corpus linguists in need of automatic methods for analyzing Old English corpora.

| Class | Infinitive | Preterite | Participle |
|---|---|---|---|
| I | í | á / i | i |
| II | éo or ú | éa / u | o |
| III (a) | e | æ/ u | o |
| III (b) | eo | ea / u | o |
| III (c) | e | ea / u | o |
| III (d) | ie | ea / u | o |
| III (e) | i | a / u | u |
| IV | e | æ/ ǽ | o |
| V | e | æ/ ǽ | e |
| VI | a | ó | a |
| VII | varies | é, eo or éo | a, á or ea |

Table 1: Strong verb classes (Hogg, 1992; Quirk and Wrenn, 1994).

Finite state transducers are an ideal choice for handling Old English morphology. This technique has been used to great effect for a number of different languages with rich morphology (Oflazer, 1994; Beesley and Karttunen, 2003). FSTs allow for both morphological generation and analysis without additional effort. They are also fast and efficient, even with large lexicons.

In this paper, I begin with an overview of the main features of strong and weak verbs in § 2. In § 3, I discuss several design decisions from a high-level perspective as well as which morphological features were included in this project. This discussion is followed by a description of the implementation using the Xerox Finite State Transducer toolkit (xfst) and the various problems I encountered in § 4. Before presenting my conclusions in § 6, I discuss my approach to evaluating results in § 5.

# 2 Verb Morphology

## 2.1 Strong Verbs

There are seven classes of strong verbs and one class with five subclasses. Table 1 describes the stem vowel changes that occur in these classes. In this table, the first form of the preterite, if given, is only for first and third person singular subjects. These changes occur in fairly regular orthographic contexts, but there are some exceptions. Also, some changes to the stem cause changes to the stem consonants. For example, the verb *to choose* is a class II strong verb that undergoes a consonant change in the preterite plural and past participle forms (see Table 2). Aside from the stem change in the preterite and past particle forms, strong verbs were conjugated largely the same to reflect person and number (see §2.3).

| Infinitive | Preterite | Past Participle |
|---|---|---|
| céosan | céas / curon | coren |

Table 2: An example of a strong verb undergoing consonant change.

| **Person and Number** | **Weak** | **Strong** |
|---|---|---|
| Present Indicative | | |
| 1 sg. | -e | -e |
| 2 sg. | -st | -st |
| 3 sg. | -ð | -ð |
| all pl. | -að | -að |
| Preterite Indicative | | |
| 1 sg. | -e | - |
| 2 sg. | -est | -e |
| 3 sg. | -e | - |
| all pl. | -on | -on |
| Subjunctive (Present and Preterite) | | |
| 1 sg. | -e | |
| 2 sg. | -e | |
| 3 sg. | -e | |
| all pl. | -en | |
| Imperative | | |
| 2 sg. | -e / -a / - | |
| 2 pl. | -að | |

Table 3: Conjugation of OE verbs by person and mood (Quirk and Wrenn, 1994).

## 2.2 Weak Verb

Weak verb morphology is complicated by the fact that mutated vowels are common. These vowel changes occur due to syllables in their vicinity. There are two main classes of weak verbs. Class II verbs are all weak verbs ending in *-ian* except those immediately preceded by *-r-* (Quirk and Wrenn, 1994). Class I verbs are all others. Mutation is common in class I verbs, as is gemination of the consonant after the short main vowel. Gemination is the process by which a consonant sound is pronounced longer than usual. In OE orthography, this manifests as a doubled consonant (e.g. *fremman*, 'to perform'). Consonants that are geminated in the present tense are often not geminated in the preterite.

## 2.3 Conjugation

In Old English, verbs were conjugated based on person, number, mood and tense. Inflections for tense serve as the basis for the division between strong and weak verbs in the *indicative* mood. There was also a *subjunctive* mood that has largely died out in ModE, with the exception of certain uses of the word

*were* (e.g. "If I *were* to leave..."). The subjunctive mood indicates conjecture or hypothetical situations. Finally, verbs were inflected based on the number (singular or plural) and person (first, second, or third) of their subject. With the exception of some irregular verbs, plural forms do not change based on person. Table 3 shows the various endings for strong and weak verbs based on these factors.

# 3 System Overview

For this project, I chose to address weak and strong verbs only. Irregular verbs consist mostly of exceptions. However, they also consist of some of the most commonly used verbs in OE, such as *habban* "to have"and *béon* "to be". So while important to any production system of Old English, these verbs are also the least generalizable, and for the purposes of this project, the least interesting. I also chose to ignore the imperative form due to time constraints. To summarize, this morphological analyzer covers:

- Strong Verbs
  - Preterite
  - Present
  - Past Participle
  - Present Subjunctive
  - Preterite Subjunctive
  - First, Second, and Third Persons
  - Singular and Plural
- Weak Verbs
  - Preterite
  - Present
  - Past Participle
  - Present Subjunctive
  - Preterite Subjunctive
  - First, Second, and Third Persons
  - Singular and Plural

In order for the system to be considered a complete system, it would require adding irregular verbs with all of the above forms, the imperative mood for all verbs, and the present participle for strong, weak, and irregular verbs.

| Tag | Description |
|---|---|
| `+Pres` | Present tense form |
| `+Pret` | Preterite (past tense) form |
| `+PPart` | Past participle form |
| `+Subj` | Subjunctive mood |
| `+Imp` | Imperative mood |
| `+1sg` | $1^{st}$ person singular |
| `+2sg` | $2^{nd}$ person singular |
| `+3sg` | $3^{rd}$ person singular |
| `+pl` | plural (all persons) |

Table 4: Morphological tags for person, number, tense and mood.

# 4 Implementation

One difficulty I faced was the high degree to which the different verb forms could vary. Strong verbs in Old English descended from Indo-European and behave differently than the innovated weak verbs. Old English scholars have formulated a loose set of rules that determine which class a given verb conforms to, be they strong or weak. However, there are often exceptions to these rules. For example, the verb *þringan* is a class III strong verb meaning "to press". If we modify the first letter, to make *bringan*, we now have a class I *weak* verb. So while it is somtimes possible to guess the class based on the stem vowels and their vicinity, there are numerous exceptions. As such, I felt that lexicalizing the morphological analyzer at the verb class level was the best approach to achieve accuracy while not suffering too much generality. Rules then just need to be formulated for each class of verb to handle the variation within that class. This has the downside of increasing the manual effort of adding new verbs to the lexicon but the benefit that new verbs are less likely to contradict existing rules.

## 4.1 Lexicon

The lexicon I built for this project is large enough to cover the majority of interesting phenomena that occurs in strong and weak verbs. All classes and subclasses of both strong and weak verbs were covered. This consists of eighteen subclasses for strong verbs and four subclasses for weak verbs.

## 4.2 Transliteration

For ease of development, I used a transliterated version of Old English that takes into account the macrons placed on vowels to indicate length as well as the characters not present in ME, such as æ, ð, and þ. Macrons were modeled using multi-character symbols consisting of the letter followed by an apostrophe. Special characters and their transliterated forms are given in table 5.

| Old English | Transliteration |
|---|---|
| æ | ae |
| ǣ | ae' |
| ā | a' |
| ē | e' |
| ī | i' |
| ō | o' |
| ū | u' |
| ð | th' |
| þ | th |

Table 5: Transliteration chart for special characters.

### 4.3 Methodology and System Architecture

My development methodology was targeted at creating a set of rules that can easily be integrated into a larger OE morphological analysis system. The overall system architecture, consists of the following:

1. Old English to Modern English transliteration
2. Preterite form analyzer
3. Past participle form analyzer
4. Present form analyzer
5. Subjunctive form analyzer
6. Modern English to Old English transliteration

The upper language of the finite state transducer consists of the verb in infinitive form (using Old English characters) followed by a series of morphological tags indicating person, number, tense and mood. The indicative mood is not included as a separate tag since it is assumed as default. This list of tags is given in Table 4. The imperative mood tag (`+Imp`) was included for completeness, though not actually implemented in this project.

The final system, when compiled in *xfst* consists of 575 states, 1160 arcs and 2451 paths. The upper language consists of 2431 unique forms and the lower language consists of 1358 unique forms. Compiling the finite state transducer and performing the morphological generation step on all words in the upper language takes just under 1 second on a single 3 GHz Pentium IV processor.

# 5 Evaluation

Given a lexicon size of 143 verbs and eight affixes that combine in seventeen different ways, there are a total of 2431 valid forms possible in the upper language. Several verbs have spelling variations, so there is a many-to-many mapping from the upper to lower language. The actual number of unique surface forms in the lower language is lower due to the fact that the subjunctive mood is very similar to the present tense first or second person (for weak and strong verbs respectively). I used three methods to evaluate my end results to supplement extensive checking of both the upper and lower language during debugging.

The first approach I took was to create a python script that generated all the possible upper language forms given the lexicon. From this I generated the words in the lower language associated with each form. Checking was done to confirm there were no missing entries and that entries with more than one word were due only to alternate spellings (*cȳden* versus *cȳðen*). There were twelve forms with duplicates, as expected. This evaluation strategy allowed me to ensure that all upper language forms were generating lower language forms and that they weren't overgenerating.

The second approach I took was to generate all possible words in the lower language using the *xfst* command `print lower`. From this list of all possible words in the lower language, I compared them against the list of words generated from the upper language forms in the first approach. This method allows me to ensure that there are no spurious upper language forms that are being generated that should not be. For example, one bug encountered using this method was the word *gefægen*, which is the correct preterite, subjunctive, plural form for *gefēon* 'to rejoice'. However, the upper language form it was generating was *gefēoan*, which does not exist.

The final method I employed in evaluation was manual error checking. This method is laborious and prone to error, since I am lacking an expert in Old English. However, some additional errors were caught and with rigorous debugging during development I feel confident that most errors were corrected.

# 6 Conclusions

Old English presents some interesting problems in dealing with complex morphology that has seen very little attention from the computational community. There has been extensive work done by scholars on Old English morphology, but, to my knowledge, finite state technology has never been applied to the task. The STELLA project at the University of Glasgow has produced software widely used for instructing students in Old English (Kay and Smith, 1990). However, this software's primary purpose is the incremental teaching of Old English and does not explicitly perform morphological analysis on new text,

nor does it use finite state machines. This project has shown that finite state technology may be applied effectively to the tasks of morphological generation and analysis for Old English.

Extending the lexicon for this project would also require minimal human effort. The main work would be in entering each verb according to the subclass it belongs to. As mentioned in § 3, the imperative mood, present participles, and irregular verbs must also be implemented. While this project has focused exclusively on Old English verbs, extending it to other parts of speech is posssible and would involve mainly the manual effort of deriving the necessary rules. Old English inflects nouns and adjectives in much the same way as modern German: there are three genders (masculine, feminine and neuter) and four cases (nominative, accusative, dative and genitive).

A fully functioning morphological analyzer for Old English would be helpful to students of Old English who wish to read texts written in the language but are slowed down by the complexity of the morphology, which usually conveys important information. Having a tool to automatically analyze morphological endings would accelerate this process. Corpus and historical linguists could also use such a tool to analyze bodies of work without having to resort to labor-intensive manual tagging.

## References


Beesley, K. R. and Karttunen, L. (2003). *Finite State Morphology*. CSLI Publications.

Diamond, R. E. (1970). *Old English Grammar and Reader*. Wayne State University Press.

Hogg, R. M., editor (1992). *The Cambridge History of the English Language*, volume 1. Cambridge University Press.

Kay, C. J. and Smith, J. J. (1990). Is there a teacher in this class? English Language and the Glasgow STELLA Project. *Literary and Linguistic Computing*, 5:77–80.

Oflazer, K. (1994). Two-level Description of Turkish Morphology. *Lit Linguist Computing*, 9(2):137–148.

Pollington, S. (1997). *First Steps in Old English*. Anglo-Saxon Books.

Pyles, T. and Algeo, J. (2004). *Origins and Development of the English Language*. Thomson Wadsworth.

Quirk, R. and Wrenn, C. L. (1994). *An Old English Grammar*. Northern Illinois University Press.

Wikipedia (2007a). Anglo-Saxons — Wikipedia, the Free Encyclopedia. Online; accessed 14-May-2007.

Wikipedia (2007b). Old English — Eikipedia, the Free Encyclopedia. Online; accessed 14-May-2007.

Wikipedia (2007c). Old English Morphology — Wikipedia, the Free Encyclopedia. Online; accessed 16-May-2007.