IceMorph: An Automated Morphological Analyzer and English-Language Lookup Tool for Old Icelandic[1]

Timothy R. Tangherlini
UCLA
Aurelijus Vijūnas
National Kaohsiung Normal University
Kryztof Urban
UCLA
Peter M. Broadwell
UCLA

Introduction

The advent of inexpensive computing and the creation of large machine-actionable corpora consisting of well-structured digital texts have made it possible to analyze and mark for morphosyntactic features significant amounts of text (> 1,000,000 tokens) with a high degree of accuracy (> 80 percent) rapidly and automatically. Although the problem of automatically tagging text with part-of-speech

1. Funding for IceMorph was provided through NSF #BCS-0921123; NSF #IIS-0122491 / EU IST2001–32745, with additional support from UCLA's Center for Medieval and Renaissance Studies, the UCLA Council on Research, and the UCLA Office of the Vice Chancellor for Research. Jackson Crawford (UCLA), Zoe Borovsky (UCLA), David Gabriel (UCLA), and Monit Tyagi (UCLA) all contributed to the development of IceMorph. More information about the project can be found at www.purl.org/icemorph/index.

(POS) information has been largely solved for languages with little morphonological complexity,[2] more complex languages, such as Old Icelandic (OIc) and other ancient languages, continue to pose problems for automated systems. Despite these difficulties, rich morphosyntactic markup that includes lemmatization holds great promise for both linguistic and textual scholarship. Accurate markup would enable the development of sophisticated online study environments that allow researchers to perform complex searches, make comparisons across multiple texts, and generate calculations concerning word-use and syntactical patterns. Our work, focusing on Old Icelandic, confirms that even for morphonologically complex Indo-European languages, the information gain offered by automatic morphosyntactic analysis of texts, measured as the percentage of correctly tagged tokens, sentences, and complete texts over the extant corpus, offers a marked improvement over previously available hand-marked texts (Rögnvaldsson and Helgadóttir 2008, 2011).[3]

Even the most detailed and accurate indexes produced in the past centuries—such as *Ordförrådet i de älsta isländska handskrifterna* (Larsson 1891), which provides an accurate and exhaustive word-form index for a number of the oldest Old Icelandic manuscripts (ranging from late twelfth- to mid-thirteenth-century manuscripts)—offer only minimal coverage when compared to the very large number of extant Old Icelandic texts. For a researcher interested in the study of the entire Old Icelandic corpus (or a large sub-corpus of Old Icelandic literature), these early handbooks, no matter how accurately compiled, are of limited use. Unfortunately, it is not economically feasible to extend the earlier practice of manual encoding to a greater number of manuscripts; the manual compilation of handbooks is costly and requires tremendous amounts of time, expertise, and energy. The old "paper-and-pen" approach does not, to borrow a term from computer science, "scale" well.

A dream of many researchers in Old Icelandic is to be able to work with a large number of texts (and manuscript witnesses to texts)—or even a comprehensive corpus—that include the high level of morphosyntactic detail of the early handbooks mentioned above. Similarly,

2. See, for example, Brill (1992) and Garside (1987) on tagging English.

3. For the sake of simplicity, throughout this paper, we use the term "Old Icelandic," as opposed to the less accurate "Old Norse," or the somewhat cumbersome "Old Norse-Icelandic."

historical linguists (especially syntacticians) are eager to work with a much larger parsed corpus of Old Icelandic texts than is currently available. Recent work, such as that of the Icelandic Parsed Historical Corpus group (IcePaHC) (Wallenberg et al. 2011) is a major step toward making such resources available, as it provides a considerable number of texts tagged in a semi-supervised fashion, and moves us closer to a comprehensive parsed Old Icelandic corpus. Yet, it is unlikely that IcePaHC alone will provide adequate coverage for Old Icelandic textual research, in part because it is focused on the historical development of Icelandic up through the present, and in part because it provides limited lemmatization of the texts. As such, IcePaHC diverges from our project, which has as its sole focus the morphosyntactic analysis and lemmatization of Old Icelandic texts. We believe that the computational methods developed by our group can augment those of IcePaHC and others, and have the potential to not only extend the necessarily limited scope of the earlier historical handbooks, but also increase considerably the number of richly marked texts available to researchers.[4]

Automatic morphosyntactic analysis of Old Icelandic offers an efficient method for accurately tagging millions of tokens in the growing corpus of machine-actionable texts. Rögnvaldsson and Helgadóttir, for instance, estimate the total number of tokens in their target Old Icelandic corpus at ~1.6 million (Rögnvaldsson and Helgadóttir 2011, 67). This estimate is only a fraction of the overall Old Icelandic corpus, as their corpus does not include the poetic corpus, the Kings' sagas (apart from *Heimskringla*), the courtly sagas, and the legendary sagas—works that contribute to the overall "standard" corpus of Old Icelandic literature (Rögnvaldsson and Helgadóttir 2008, 46; Clover and Lindow 1985). Rögnvaldsson and Helgadóttir's tagging makes use of a trained TnT tagger, a well-known approach to probabilistic syntactic tagging (Brants 2000). They have shown that the TnT tagger, after an initial training on a hand-corrected training set of ~95,000 words, is fast and accurate (> 90 percent) for the identification of Old Icelandic word forms in a corpus normalized to Modern Icelandic (MOIc) orthography (Rögnvaldsson and Helgadóttir 2011, 69). Our work, by contrast, does not use normalized Modern Icelandic texts,

4. Initially begun under the auspices of the Perseus project (Crane 2012) and the Cultural Heritage Language Technologies project (Rydberg-Cox 2005), our project was subsequently funded through the National Science Foundation (BCS #0921123).

but rather is intended to work with any digitized Old Icelandic text, irrespective of its original orthographic normalization. It also does not rely on a large hand-corrected tagged sub-corpus for training.[5]

Considerable progress has been made during the past decade on the creation of machine-actionable texts, either through the digitization of existing analog resources or through the creation of born-digital resources. These developments have been led by MENOTA (Medieval Nordic Text Archive) and other manuscript and text projects at the two Arnamagnæan institutes (Reykjavik and Copenhagen). The Icelandic Parsed Historical Corpus group and the ongoing development of the MENOTA texts represent an important first step toward the development of a comprehensive digital environment for the study of Old Icelandic. Our project adds automatic morphosyntactic tagging of word forms in any machine-actionable OIc text, semi-supervised resolution of unrecognized forms in those texts, English-language dictionary lookup, and comprehensive inflectional resources.

The automatic morphosyntactic tagging, lemmatization, and disambiguation of Old Icelandic texts is a non-trivial task. To approach this challenge, we divided our system, which we dubbed IceMorph, into two main parts: (1) a probabilistic morphological analyzer and disambiguation machine, and (2) a deterministic inflection engine.[6] A dictionary lookup tool provides a useful extension to the combined system, returning the available English-language definitions for any given lemma. Integrating these functions with a machine actionable text corpus completes the system. IceMorph is intended to meet two challenges, and does so with increasing accuracy: (1) given a word form from a text, IceMorph returns a lemma, its inflection, and syntactical detail for the word form in its textual context, and (2) given a lemma from the dictionary, IceMorph returns an inflectional paradigm of the word, including irregular features, and discovers attestations of the inflected forms in the corpus. In future iterations, IceMorph also will return all examples of the form in context (keyword in context).

5. Our hand-corrected tagged word list comprises two sets: "GOLD," which is a list of 462 randomly chosen high-frequency words from the corpus that have been hand-tagged, and "EXPERT," which is a list of 230 high-frequency words that have been hand-tagged. The "GOLD" corpus is used for error testing and evaluation, as described in Urban et al. (2014), while the "EXPERT" corpus is used to train the tagger. As such, our training corpus of 230 words is significantly smaller than those used in other systems.

6. As such, it shares features with the IceNLP package (Loftsson and Ingason 2009).

Recent projects in Old Icelandic have focused on developing dictionary resources for the study of early Germanic languages (Crist *GLP*), and textual resources for the study of Old Icelandic, including manuscript images (MENOTA; handrit.is; heimskringla.no; netútgáfan; Clunies-Ross et al. 2007–2012). Other important projects exploring Old Icelandic morphology include algebraic approaches to Old Icelandic morphonological development (Poschenrieder 2000); rules-based POS tagging for Modern Icelandic (Loftsson 2008); and, as noted, morphosyntactic tagging of Old Icelandic texts normalized to modern orthography (Rögnvaldsson and Helgadóttir 2008, 2011). Building upon these efforts and more general work on morphological analysis from computational linguistics, we have created a second-generation implementation of IceMorph, including the morphological analyzer and disambiguation machine, the inflection engine, and the dictionary lookup tool, all accessed through a two-part, user-friendly Web-based interface.[7]

The success of our approach relies upon the close integration of the various parts of the system: we automatically analyze and mark word forms in our corpus with morphosyntactic detail, and align these word forms with their corresponding lemmata from the concatenated dictionary resources. At the same time, we inflect all the lemmata in our dictionary using the grammatical information from the dictionary entries to trigger the appropriate inflectional paradigms derived from standard grammars of Old Icelandic (Gordon 1927; Noreen 1884). We use the observed word forms in the test corpus to limit the overgeneration of inflected forms in the inflection database (see below), and we use the unambiguously matched word forms to help disambiguate ambiguous word forms.

The two main parts of IceMorph—morphological analyzer and inflection engine—are coordinated to increase the accuracy of the system's output rapidly, in that corrections or improvements in one resource are immediately reflected in the other resource. Because "scalability" has been a concern of the project since the beginning, IceMorph has been designed to inter-operate with the constantly growing corpus of digitized Old Icelandic texts, incorporating normalization routines that allow texts, including dictionaries, to function in IceMorph even if they

7. IceMorph is available at http://www.purl.org/icemorph/index. Access directly to the user interfaces is available at www.purl.org/icemorph/gui/analyzer; www.purl.org/icemorph/gui/reader and www.purl.org/icemorph/gui/dict.

use conflicting Old Icelandic orthographies.[8] IceMorph thus represents another step toward the development of a sophisticated study environment for Old Icelandic texts. Echoing Rögnvaldsson and Helgadóttir (2008, 45), we believe that a powerful system such as this can facilitate advanced search and discovery routines that extend far beyond simple keyword or lemma searches, to include pattern searches and metrics that reveal not only patterns in word use but also in inflectional class use.[9]

The test corpus for our work is the Legendary Sagas (OIc *Fornaldarsǫgur norðrlanda*) as published by Guðni Jónsson and Bjarni Vilhjálmsson in their three volume edition (1943–1944). The corpus comprises 257,604 tokens, and 22,815 unique lexemes. The texts, normalized to a close approximation of the "standard" Old Icelandic orthography as it appears in the series *Íslenzk fornrit*, are largely tales of heroes and supernatural events set in the distant "heroic" past, and represent a typical "real-world" example of the syntactic and lexical range confronting a researcher in Old Icelandic language and literature.[10] Because IceMorph is closely integrated with the two most common Icelandic-English dictionaries (Cleasby and Vigfússon 1874; Zoëga 1910), also normalized to this same orthography, IceMorph offers English-speaking students and scholars from allied disciplines an additional valuable resource for textual study.[11]

8. The orthographic normalization of existing digital resources is an ongoing challenge. Currently, there are not many out-of-copyright resources available for broad public access, although excellent collections are becoming available (www.heimskringla.no). Many of the texts that are available live in a gray area with respect to copyright (www.snerpa.is). If one focuses entirely on orthographic normalization, the immediate problem accrues to the orthographic "modernization" used for many of the extant digital texts. The current system can work with these texts, particularly given its use of an edit-distance algorithm that allows for a degree of orthographic imprecision, albeit with a higher error rate than with texts that are normalized to the conventions of the Old Icelandic Text Society (Hið íslenzka fornritafélag) or a close approximation of that orthography. We have plans to develop normalization plug-ins that would be triggered by textual metadata, supplied by the text provider, to allow for integration of non-*Íslenzk fornrit* texts with the analyzer.

9. While these statistical analyses are not yet publicly available on our website (http://www.purl.org/icemorph/gui), as we integrate the materials and analysis engines into Perseus, the Perseus word study and pattern discovery tools will become available. For examples of these tools, see www.perseus.tufts.edu.

10. The Old Icelandic orthographical conventions are discussed below.

11. We considered including Johan Fritzner's *Ordbog over det gamle norske sprog* (1886–1896), but because of the limitations on the machine-accessible grammatical information available in our version of the dictionary, we have delayed inclusion of this resource.

On the Various Old Icelandic "Orthographies"

A quick look at several Old Icelandic dictionaries, grammars, and editions of ancient texts reveals a series of different orthographies used for the same language, most displaying certain digressions from the standardized Old Icelandic spelling, which is based on a hypothesized late twelfth-century pronunciation of the language (Noreen 1884; Heusler 1913; Iversen 1923; Steblin-Kamenskij 1955). The most divergent in this respect are those resources in which the spelling has been largely modernized; this modernized orthography remains in many respects differentiated from standard Modern Icelandic spelling. The main features of these semi-modernized editions of the Old Icelandic texts are usage of the symbol "æ" for both a historical "æ" (as in *kvæði* 'poem' < **kwāþija-*) and a historical "œ/ǿ" (as in *dœma/dǿma* 'judge' < **đōmijan-*) and the employment of the suffix *-st* for the mediopassive forms, where classical Old Icelandic would have a set of suffixes, which, depending on the form, could be *-mk* (1. sg.), *-msk* (1. pl.), *-zk* (after an underlying dental/alveolar consonant), and *-sk* (for all other cases).

In this paper, we adhere to the standardized *Íslenzk fornrit* spelling, using the symbol "ǫ" for the *u*-umlauted *a*, "ǿ" for the *i*-umlauted **ō*, and the full set of Old Icelandic mediopassive endings. One reason for preferring the symbol "ǿ" to the equally common "œ" is that the latter closely resembles the symbol "æ," particularly when italicized. However, IceMorph employs the symbol "œ," following the practice of *Íslenzk fornrit*. In the examples below, the same distinction is made, the symbol "œ" being employed in examples taken directly from the online parser or the underlying algorithms, whereas in other instances, the symbol "ǿ" is used in order to avoid confusion with "æ." Likewise, we employ "ǫ" for general purposes, but "ö" for examples taken directly from IceMorph. In IceMorph, we use the modern symbol "ö" instead of "ǫ," which represented a rounded low back vowel in classical Old Icelandic, for purely technical reasons: it is the symbol used in our target corpus as well as in our two dictionaries, Cleasby and Vigfússon and Zoëga, and it is rendered more easily in most web browsers.

Challenges in Developing IceMorph

We encountered two broad classes of challenges while developing IceMorph. The first type of challenge derived from the disconnect between the demands of developing a series of algorithms and

computer code out of resources such as grammars and dictionaries originally developed for a different approach to language study. The second type of challenge arose from the morphological complexity of the language itself.

There is a major difference between an algorithmic approach to morphological complexity and a purely analytical approach based on conventional scholarly materials such as grammars and dictionaries. In general, the purpose of a grammar is to describe regular patterns in a language. In the case of inflectional morphology, that description is largely accomplished by sorting words into well-defined classes based solely on the grammatical form of those words. A well-known example of this approach can be found in Noreen's standard grammar, *Altisländische und altnorwegische Grammatik* (1884). Similarly, the main purpose of a dictionary is to collect a comprehensive list of words for a given language, to define or translate their meaning, and to provide basic grammatical information such as the word or inflectional class. Expanded dictionaries, such as Cleasby and Vigfússon (1874), include numerous examples of usage as well as tables of irregular word forms, which largely serve as a guide for individuals as they work with texts.

The purpose of a system such as IceMorph is different: its inflection engine combines the rules-based approach of a grammar with the lexical diversity of a dictionary, while also recognizing that the observed word forms in a text corpus frequently do not align unambiguously with the generated forms from the inflection engine. In the following sections, we describe (a) an easily debugged, efficient inflection engine for generating word forms from Old Icelandic lemmata; (b) a morphological analyzer for matching observed forms with these generated forms, including routines for normalization and disambiguation; and (c) approaches to error detection and correction. This last point is of considerable importance. The methods for error detection and correction that we have implemented allow us to refine the system on an ongoing basis. As both machine-learning algorithms and language experts correct imperfect inflections or word identifications and eliminate over-generated paradigms, the refinements cascade through the corpus. By constantly feeding new information into the system—a process known as "bootstrapping"—the system "learns" in a semi-supervised fashion and is able to propagate what it has learned across both the existing corpus and any new texts that are subsequently added (Urban et al. 2014). Challenges that we have yet to tackle include

working with texts in hard-to-normalize orthography (e.g., diplomatic or facsimile transcriptions) and the classification of alternate inflections.

An Inflection Engine for Old Icelandic

Over the course of several years of early development, we discovered that various accepted approaches to morphological analysis, including two-level constraints, finite state transducers (FST), and cascading rewrite rules, were not easily adapted to producing the highly accurate results (> 90 percent) demanded by Old Icelandic scholars. Because of the complexity of the Old Icelandic morphological system, the implementation of these approaches resulted in complex programs that were very difficult to debug. Ultimately, we encountered performance barriers related to accuracy, program compile time, and debugging that were intractable for our small team given the very high performance criteria of Old Icelandic researchers. Consequently, we moved away from the original analyzer, written in Perl, to an analyzer written in Haskell, a functional language more friendly to non-programmers. We employed a particular library embedded in the Haskell computer language, referred to as FM/Haskell, which was designed specifically for morphological analysis (Forsberg and Ranta 2004).

The impact of the shift in our architecture from an inflection engine written in Perl to one written in FM/Haskell was considerable. Functional languages such as Haskell regard programming as the evaluation and application of mathematical functions. In contrast, imperative languages, such as Perl, focus on the execution of sequential commands (for morphology and the generation of inflected word forms, sequential commands are equivalent to cascading rewrite rules). Although imperative languages can adopt a "functional" style and imitate the results of functional code, there are significant differences. Functional code, due to its puritan approach to computation, is easier to maintain and debug. In addition, code written in a functional language tends to be considerably shorter and easier to read.

Functional Morphology (FM) is a systematic method of developing natural language morphologies. It is implemented in the purely functional programming language Haskell98, and is meant to give non-programmers an intuitive way to produce natural language morphologies. Forsberg and Ranta note that FM is, in a sense, part of the Grammatical Framework (GF), a functional language for defining grammars—including natural languages—that is embedded within

Haskell (Forsberg and Ranta 2004, 214). Yet despite GF's efficiencies, certain complex morphological phenomena such as umlaut or syncope are easier to write in Haskell than in GF (Forsberg and Ranta 2004, 214).[12] These concerns informed our move from Perl to FM/Haskell.

Functional Morphology is based on the idea of declension prototypes. Forsberg and Ranta note that this idea of inflection tables "has been around for over 2,000 years. . . . An inflection table captures an inflectional regularity in a language. A morphology is a set of tables and a dictionary. A dictionary consists of lemmas, or dictionary forms, tagged with pointers to tables" (Forsberg and Ranta 2004, 215). The implications of this approach are fairly clear. Given a particular word class, most words in that class tend to be declined similarly, while exceptions occur less often. Functional Morphology makes use of this fact and defines prototype declension tables for the regular word types. Although a program written in FM/Haskell is pure programming code, it nevertheless closely resembles an actual declension table, containing the morphological rules to generate word forms (see below). In IceMorph, the assignation of words to prototypes takes place largely in an automated fashion, using existing cues in the Cleasby and Vigfússon and Zoëga dictionaries (Forsberg and Ranta 2004, 219–20).

To illustrate the transparency of FM/Haskell code, we present here a sample declension table written for regular masculine *a*-stem nouns:

```
declIheimr :: DictForm -> Noun
declIheimr heimr (NounForm n c) =
  mkStr $
  case n of
    Singular -> case c of
      Nominative -> prefix ++ lexeme
      Accusative -> heim
      Genitive -> heims
      Dative -> heimi
    Plural -> case c of
      Nominative -> heimar
      Accusative -> heima
      Genitive -> heima
      Dative -> heimum
    where
      (prefix, lexeme) = splitCompound heimr
      root = (tk 1 lexeme)
```

12. This is due to Haskell's more powerful processing of lists and strings.

```
heim = prefix ++ root
heims = prefix ++ root ++ "s"
heimi = prefix ++ (syncope root ++ "i")
heimar = prefix ++ (syncope root ++ "ar")
heima = prefix ++ (syncope root ++ "a")
heimum = prefix ++ (syncope (uMutation root) ++ "um")
```

Any word in the dictionary assigned to this word class and stem and this particular sub-prototype will then inflect as *heimr* does. Although these declension prototypes are quite straightforward, the rich morphology of Old Icelandic also causes them to incorporate a number of calls to morphophonemic functions, such as syncope, umlaut, and breaking.

Programming for Morphonological Complexity

Despite the relative transparency of FM/Haskell code, morphological features of Old Icelandic complicated the development of the inflection engine and morphological analyzer. Compared to the other ancient Germanic languages, the morphonological system of Old Icelandic is complex and in many ways irregular. This complexity reflects the operation of multiple processes that began at the phonetic/phonological level in prehistoric times (such as the Verner's alternation, umlaut, breaking, syncope) but, in the course of time, greatly affected the "landscape" of the Icelandic morphological system, creating a rich allomorphy unrivaled in the Germanic language branch. This inflectional complexity was further complicated by various transfers of Icelandic words between inflectional classes, and the development of mixed inflections that arose both because of the ambiguity of word endings and the active processes of analogy. As a result, a large number of Icelandic words possess alternative forms (e.g., the plural forms *strandir/strendr* for the feminine noun *strǫnd* "shore," the past forms *flaug/fló* for *fljúga* "fly," or even entire separate paradigms for a single word, e.g. *gera/gøra* "do, make").

IceMorph's inflection engine deals with by-forms and parallel paradigms through the implementation of a hierarchy of prototypes and sub-prototypes (Forsberg and Ranta 2004, 219). This approach is an advancement over cascading rewrite rules that cannot easily produce these additional forms. Various highly irregular words, such as the noun *maðr* "man," the adjective *hár* "high," or the anomalous verb *valda* "cause," are also incorporated into the analyzer using this concept of

prototypes and sub-prototypes. Finally, irregular forms that cannot be generated through prototypes or sub-prototypes are substituted into the inflectional paradigms using a table of exceptions derived from the dictionary and grammar resources.

A large part of the complexity of Old Icelandic morphology can be attributed to the (mor)phonological processes of umlaut (also called "mutation"), breaking, and syncope.[13] These processes create the rich allomorphy of Old Icelandic. Developing accurate algorithms to deal with these processes was one of the greatest computational challenges for the IceMorph project.

Umlaut is the result of several types of vocalic assimilation that operated throughout the Germanic branch. Three types of umlaut are normally distinguished in Germanic, depending on the nature of the *assimilating* sound: *a*-umlaut (lowering), *i*-umlaut (fronting/raising), and *u*-umlaut (for the most part, rounding). Among these, *a*-umlaut operated at a very early date, long before the formation of the Icelandic language. Its effect may be observed in words such as *orð* "word" (< Proto-Scand. **wurđa*-), and *verr* "man" (< **wira*-). This umlaut is also the cause of alternations such as *sunr* "son" ~ *sonar* (gen. sg.; < **sunar* < **sunōR*), although it produces minimal allomorphy when compared to *i*-umlaut, *u*-umlaut, and "breaking."

The effects of *u*-umlaut on the development of allomorphy in Old Icelandic are best observed in the paradigms of words containing the short vowel /a/ in the root, such as *saga* "story," *barn* "child," *svartr* "black," and *taka* "take." In the inflected forms of such words, the vowel *a* changes to *ǫ* when there is either a *u* (or *v*) in the following syllable, or when *u* (or *v*) used to be there (i.e., when the *u*-umlaut was still phonological). For example, one finds OIc *sǫgu* (obl.) or *sǫgur* "saga" (nom./acc. pl.), *svǫrtum* "black" (dat. pl.), *tǫkum* "take" (1. pl. pres.), *bǫrnum* "children" (dat. pl.). The vowel *ǫ* also occurs in the nom./acc. pl. form *bǫrn*, as this form contained a vowel *u* at an earlier, reconstructible stage (Proto-Scand. **barn-u*). Here, the *u* was lost due to apocope during the Common Scandinavian era; before this *u* was lost, it rounded the root vowel *a* to *ǫ*.

13. Although the view that umlaut is purely phonological has been expressed in the critical literature (Rögnvaldsson 1981), we adhere to an alternative view, according to which, at a certain stage, umlaut stopped being a purely phonological process, transforming into a morphonological one (Árnason 1985; Harðarson 2001).

U-umlaut has one additional complication, not characteristic of the other umlauts: in words containing three (or more) syllables, the vowel *ǫ*, produced by *u*-umlaut, was normally reduced to *u* if it occurred in an unstressed syllable (the stress in Old Icelandic falls on the word-initial syllable). Thus, one finds *verndari* "protector" with the dative plural form *verndurum* (dat. pl.; < **verndǫrum*). If the initial syllable contains the vowel *a*, that vowel is also changed to *ǫ* but is not reduced to *u*. Consequently, one finds the form *kǫstulum* from *kastali* "castle." There are exceptions to this alternation. Weakening does not take place in *nd*-stems so that one has the form *gefǫndum* (dat. pl.; not **gefundum*) from *gefandi* "giver" (nom. sg.). Neither this reduction nor rounding of the preceding *a* takes place in compounds, across the boundary between its members, or when prefixes are present. Consequently, one finds the form *fjall-gǫngum* (dat. pl.; not **fjǫll-gungum* vel sim.) from *fjall-ganga* "mountain trip" (compound), and the form *athǫfnum* (dat. pl.; not **ǫthufnum*) from *athǫfn* "conduct."

I-umlaut operates in essentially the same way as *u*-umlaut, although its primary effect on the assimilated vowels is fronting. *I*-umlaut commonly occurs in the plural forms of Icelandic root nouns. Consequently, for the word *mús* "mouse," one finds the form *mýss* (nom./acc. pl.; < Proto-Scand. **mūsiR*), for *bók* "book," one has *bǿkr* (< **bōkiR*), and, in past subjunctive forms, one finds. *fór* "go" (3. sg. pret. indic.) and *fœri* (3. sg. pret. subj.; < Proto-Germ. **fōrī*), *urðu* "become" (3. pl. pret. indic.) and *yrði* (3. pl. pret. subj.; < **wurđīn-*). Along with *u*-umlaut, *i*-umlaut created numerous instances of allomorphy, but unlike *u*-umlaut, it did not cause any vowel reduction.

The historical phonological process of "breaking" added an additional degree of allomorphy to Old Icelandic. There are two types of breaking, depending on the quality of vowel that causes it. *A*-breaking is caused by the vowel *a* in an unstressed syllable, and *u*-breaking is caused by the vowel *u* in an unstressed syllable. These processes also operated during the Common Scandinavian era, and they affected the vowel *e*, changing it into *ja* (*a*-breaking) and *jǫ* (*u*-breaking). The effects of breaking can be seen throughout the morphological system of Old Icelandic. For example, one finds *bergʒ* "save" (1. sg. pres. indic.) and *bjǫrgum* (1. pl. pres. indic.) from the infinitive *bjarga*, and *fjǫrðr* "fjord" (nom. sg.) and *fjarðar* (gen. sg.).

A large number of Icelandic paradigms exhibit the effects of both umlaut and breaking, which results in very complex, irregular-looking

paradigms. An example of this can be found in the paradigm for the noun *fjǫrðr* "fjord," mentioned above:

	Singular	**Plural**
Nom.	*fjǫrðr* (< **ferþ-uR* < **-uz*)	*firðir* (< **ferþ-iR* < **-ijiz*)
Acc.	*fjǫrð* (< **ferþ-un*)	*fjǫrðu* (< **ferþ-unn* < **-unz*)
Gen.	*fjarðar* (< **ferþ-aR* < **-auz*)	*fjarða* (< **ferþ-an* < **-ōn*)
Dat.	*firði* (< **ferþ-iu* < **-iwi*)	*fjǫrðum* (< **ferþ-umm* < **-umz*)

The inflection engine addresses umlaut and breaking as language-specific functions that operate across the lexical set and in concert with the word prototypes and sub-prototypes. In most cases, the inflection engine deals quite well with this type of complexity and returns a correct paradigm for words such as *fjörðr* (i.e., *fjǫrðr* in standardized Old Icelandic):

	Singular	**Plural**
Nom.	fjörðr	firðir
Acc.	fjörð	fjörðu
Gen.	fjarðar	fjarða
Dat.	firði	fjörðum

In addition to umlaut and breaking, the Common Scandinavian protolanguage was characterized by syncope of unstressed vowels. While syncope tended to occur in words that were trisyllabic (or longer) in the protolanguage, its occurrence in Old Icelandic is not predictable based on the number of syllables alone. For example, one finds syncope in OIc. *jǫtn-ar* "giants" (nom. pl.; two syllables) < Proto-Scand. **et-un-ōR* (three syllables), but not in OIc. *skrif-ar-ar* "scribes" (three syllables). Syncope is particularly irregular among adjectives, operating in some words and not operating in others, even though they may belong to the same derivational type; compare *mál-i-gr* "talkative"—acc. sg. masc. *mál-gan*, but *kunn-i-gr* "expert"—acc. sg. masc. *kunn-i-gan*.

Defining the morphonological functions in FM/Haskell was the most difficult programming task during the creation of the inflection engine because the code cannot be easily written or debugged by a non-programmer. Accordingly, one of our goals has been to make our function library as robust as possible so that the library of inflectional prototypes and sub-prototypes rests on a stable morphological foundation. One of the least complex functions, that of describing *u*-umlaut, illustrates both the complexity of the code and the functional

nature of FM/Haskell (this function includes two helper functions "findStemVowel" and "isVowel"):

```
uMutation :: String -> String
uMutation man = m ++ mkUm a ++ n
where
(m,a,n) = findStemVowel man
mkUm v = case v of
"a" -> "ö"
_ -> v

findStemVowel :: String -> (String, String, String) findStemVowel
sprick = (reverse rps, reverse i, reverse kc) where (kc, irps) = break
isVowel $ reverse sprick (i, rps) = span isVowel $ irps

isVowel :: Char -> Bool
isVowel c = elem c "öaeiouúyýóø"
```

Given a string, the uMutation function first finds the stem vowel. If the stem vowel is an *a*, then that *a* is changed to *ö* (i.e., *ǫ*). In all other cases (indicated by the "_"), no change is performed. Although the end result is quite simple, the code necessary to reach it is not.

To illustrate the effect of these functions, consider the noun *heimr*, the prototype of which was presented above. Since its stem vowel *ei* is not *a*, *u*-mutation is not performed (and neither is syncope, for that matter), and the dative plural form is *heimum*. However, given a noun such as *afgangr* "surplus," with an *a* as the stem vowel, the dative plural form becomes *afgöngum* (i.e., *afgǫngum*). The word *afgangr* also illustrates another important feature of the language, namely the need to handle prefixes in an appropriate manner. The inflection engine is designed to accept dictionary entries as input, yet in order to apply the morphophonemic rules correctly, compound nouns must be split into the main lexeme and the prefixed elements. For example, *afgangr* consists of two parts, the prefix *af* and *gangr*, which in Old Icelandic lexica is usually signaled by a hyphen (*af-gangr*). The function "splitCompound" performs the separation of prefix and lexeme, and allows us to generate the proper forms:

```
(prefix, lexeme) = splitCompound afgangr
```

This word also highlights the key role played by the dictionary in this system, as the dictionary provides "raw material" in the form of lemmata, part-of-speech information, and inflection clues. Typically,

a dictionary provides a standard form of a given lemma along with relevant grammatical information. Due to space limitations, however, this information is normally kept at a minimum, presenting only the most practical details (and exceptional inflectional forms, if any). In our dictionaries, adjectives were labeled "adj." (Cleasby and Vigfússon 1874) or "a." (Zoëga 1910), nouns were generally labeled only with grammatical gender, and verbs were most often accompanied only with information on how to form the preterite.

Developing a concatenated machine-actionable dictionary for Old Icelandic was an important step. Because one of the main performance criteria for IceMorph was to provide an English-language lookup tool for students, we concentrated our efforts on the two best-known Old Icelandic-English dictionaries, Cleasby and Vigfússon (1874) and Zoëga (1910). We parsed the entries from the electronic versions of both dictionaries, separating out headwords, grammatical information, and definitions, identifying a total of 42,732 headwords in Cleasby and Vigfússon, 21,834 of which did not appear in Zoëga. We also identified 26,940 headwords from Zoëga, of which 6,042 did not appear in Cleasby and Vigfússon. As part of our parsing routine, words were assigned to their most likely inflectional prototype. Our final concatenated dictionary comprises 48,813 unique lexemes formed by combining the two dictionary sets and including entries for a small number of words (139) not fully resolved by the dictionary parsers.

Although Cleasby and Vigfússon and Zoëga use slightly different orthographic conventions, it became apparent, even after orthographic normalization and alignment, that only a small percentage of the Cleasby and Vigfússon-only lemmata could be attributed to variant spellings of the same word. Of the Cleasby and Vigfússon-only lemmata, approximately 64 percent were nouns, many of them compound nouns. This difference in overall vocabulary reflects not only the more comprehensive nature of Cleasby and Vigfússon, but also the different approach to the presentation of compounds in the two dictionaries. In Zoëga, compounds are usually abbreviated and embedded in the definition of the base lexeme. Consequently, they are difficult for our dictionary parser to recognize consistently and accurately. Fortunately, these compounds are presented in a more regular format in Cleasby and Vigfússon, and assigning part-of-speech and class information (POSC) to new compound words based on the known POSC of their base lexeme was by and large successful.

Interestingly, only around 9 percent of the Cleasby and Vigfússon-only lemmata were verbs.[14]

A final philosophical consideration for the implementation of the inflection engine was whether to take a "brute force" approach to Old Icelandic inflection by generating all the possible paradigms for every lemma in the concatenated dictionaries and storing these in an inflectional database, or to take a more "just-in-time" approach, generating inflections as needed by IceMorph users to populate the inflectional database.

The "brute force" approach results in a significant over-generation of both paradigms and inflectional forms, which subsequently need to be pruned. In many cases, the dictionaries do not provide sufficient limiting information to assign lemmata unambiguously to inflectional prototypes. The inflection engine currently includes ninety-six inflectional prototypes: forty noun prototypes covering nine strong and three weak declensions, fifty-five verb prototypes describing seven strong and four weak classes, and one adjective prototype. For a noun with no limiting information, for example, the engine generates forty inflectional paradigms.

Another type of over-generation occurs within the paradigms themselves. Old Icelandic has a large number of words for which the correct paradigm does not include all possible forms for a given word class. A purely mechanical inflection of the adjective *dauðr* "dead," for example, includes all the degrees of comparison (positive, comparative, superlative). In reality, only the positive degree exists since once one is dead, one cannot be more dead (although one could imagine certain Vikings speculating which of their victims is the most dead). Despite these concerns, the advantage of the brute force approach is that once the inflections have been generated, the system can be used

14. A more vexing question concerns the 6,042 headwords recognized in Zoëga not appearing in Cleasby and Vigfússon, given that Zoëga was conceived of as a student-friendly subset of Cleasby and Vigfússon. The majority of these words are (a) simple compounds made by adding prefixes (e.g., *afeigna* "to dispossess") or suffixes (e.g., *gáligr* "mindful of") (~63 percent); (b) alternate spellings (e.g., Zoëga: *myrkja* "to grow dark"; Cleasby and Vigfússon: *myrkva*) (~6 percent); (c) differences in the POS classification (e.g., Zoëga: *kné* "knee" masc. noun; Cleasby and Vigfússon: neut. noun) (~5 percent); (d) place names (e.g., *Engilsnes*) (~2 percent); and (e) certain rare words appearing in Zoëga but not in Cleasby and Vigfússon (e.g., *byrð* "birth, descent") (~4 percent). Finally, there are some words that do appear in Cleasby and Vigfússon but were not captured by our parser, either due to optical character recognition errors or other formatting errors (~20 percent).

in human interactive time and, because storage is inexpensive, the cost of calculating and storing all of the generated inflectional data is low.

The "just-in-time" approach produces a paradigm for any form matched in the target corpus and only produces the inflectional table in which that form appears. This approach has the benefit of only inflecting words as needed and reducing the number of over-generated paradigms. Despite this seeming efficiency, however, the need to produce inflections "on the fly" has little benefit for textual study. Similarly, because IceMorph also includes a comprehensive dictionary browser, which requires that the dictionary effectively be treated as an additional text, the benefits of not "inflecting everything" are limited. Consequently, we decided to implement the brute force approach. We then applied error correction, expert feedback, and disambiguation as means to reduce the over-generated forms in the inflectional database.

Morphological Analysis, Word Alignment, and Disambiguation

Form matching and grammatical disambiguation both pose significant challenges. The analyzer takes as input any form from the text corpus and returns the correct lemma as well as its dictionary definition and inflectional paradigm. For example, given as input the form *herskipa* from the first chapter of *Vǫlsunga saga*, the analyzer returns "herskip, noun, n_a, plural genitive," followed by the paradigm, which includes the discovered form highlighted:

	Singular	**Plural**
Nom.	herskip	herskip
Acc.	herskip	herskip
Gen.	herskips	herskipa
Dat.	herskipi	herskipum

Although the first pass of this matching approach provided many accurate solutions for regularly inflected words, the failure rate was still too high for use by Old Icelandic textual scholars and linguists (accuracy of about 70 percent).

Two problems became immediately apparent: First, for a considerable number of words in the corpus, despite the over-generation of paradigms, the morphological analyzer failed to find any solutions for the observed forms. Second, for many words, the analyzer proposed an improper solution, either based on grammatical ambiguity (correct word match, but incorrect POS tag as a result of too many possible

solutions in a paradigm), or lexical ambiguity (incorrect word match as a result of too many possible lemmata containing the form in their paradigms), or both.

The first problem was vexing because the analyzer's lexicon contained the data from two major dictionaries. As we explored the problem in greater detail, many of these alignment failures turned out to be a result of minor differences in spelling between the word forms in the digital corpus and the inflected dictionary.[15] To address this problem, we refined our word normalization routines. Most of the orthographic normalization problems were related to the representation of vowels (e.g., the orthographic inconsistencies surrounding *ø*, *ö*; and *ǫ*, *æ*, *œ*, *ǿ*), the use of *-sk* as opposed to *-st* in reflexive forms, and various phenomena such as vowel shortening before certain consonants (e.g., *alfr* or *álfr*). We continue to extend these normalization routines, although the number of errors due to incompatible orthographies is quickly approaching zero. Other causes of this failure are attributable to unrecognized compound words, optical character recognition (OCR) errors in the machine actionable texts, errors in parsing the dictionary resources, variant forms not in our inflectional tables, and enclitics (articles and possessive pronouns).

We developed an extension in the analyzer, a "word guesser" similar to that described in Loftsson (2007, 106–7), to propose solutions when analysis of an observed form failed. In its first incarnation, the word guesser used a straightforward Levenshtein edit-distance to find possible solutions for a given word form in the inflectional tables (Levenshtein 1966). The edit-distance algorithm calculates the lowest cost solution to an unmatched word form, returning the solution's lemma as the proposed lemma for the word. In its initial implementation, this algorithm was naïve and paid no attention to surrounding sentence context. It was also simple, choosing the first lowest cost solution that it found in the inflectional database. This precipitated a large number of incorrect assignments based entirely on the ordering of the dictionary. Another series of incorrect alignments occurred because of low-cost yet erroneous solutions that were discovered in incorrectly generated paradigms. These errors are eliminated as the paradigms become more accurate.

15. Whereas the spelling in the built-in dictionary is standard Old Icelandic (with the exception of the symbol "ǫ," which, due to technical issues, is represented by the symbol "ö" in the current version of the analyzer), the spelling of many of the digital texts available today has been partially modernized.

In our current word-form alignment implementation, a more context-aware probabilistic algorithm rejects certain solutions before making a proposed assignation (e.g., rejecting an adjective when a verb is syntactically required by the surrounding words). This approach mirrors the probabilistic approach of the TnT tagger used by Rögnvaldsson and Helgadóttir (2011), and allows the analyzer to identify words not included in our lexicon and to isolate unusual forms not covered by our prototypes (Urban et al. 2014). Given the "guessed" nature of these responses, we mark these forms in the text as such, making expert confirmation or, when needed, correction of the assignment easier to complete. Currently, there are five possible methods for solving the word-form recognition problem:

> Prototype classifier: the form matches a form generated by the inflection engine
>
> Postprocessing (Bayesian): the analyzer, using a probabilistic model, proposes a solution
>
> Postprocessing (expert similarity): the form matches another form that was solved using expert feedback
>
> Postprocessing (best match): the analyzer, using a simple edit-distance measure, proposes a low-cost match for the form.
>
> Postprocessing (unique match): the analyzer found a single unique match for the word form in the dictionary headwords

Because we mark each solved word with the solution method used to resolve it, it is easy for experts to check the accuracy of any solution. If the solution is incorrect, the expert makes a correction, which then feeds into the probabilistic algorithms during the nightly build. These expert corrections are labeled "Expert Feedback."

The inflection engine and the morphological analysis system are tightly integrated in IceMorph. Consequently, corrections in dictionary resources, underlying inflectional code, or in word matching propagate through the system during nightly "updates." This integration is a key component of the "convergence" model that underlies the system. While the system starts off returning erroneous or improperly aligned forms, it makes use of expert feedback and machine learning to continuously move toward correct solutions for every observed word form, ultimately converging on a highly accurate tagged corpus and, just as importantly, an accurately inflected dictionary.

Error Detection and Correction

There are two main classes of errors in IceMorph: paradigmatic errors that are caused by failures in the inflection engine, and morphological analysis errors that are caused by failures in the analysis and disambiguation routines. Paradigmatic errors, in turn, can be classified into two main categories: the first group of errors are caused by bugs in the underlying algorithm (algorithmic errors), and the second group of errors are philological errors.

Algorithmic errors are the most easily addressed, as these are systematic and normally occur in connection with morphonological phenomena such as umlaut, breaking, and syncope. Refining the FM/Haskell code has reduced these errors. At this point, the number of algorithmic errors is very small, and most of the early errors have been corrected by simple debugging. Philological errors are, by contrast, never systematic and cannot be corrected by modifying the inflection engine. Some of these errors stem from irregular forms that do not appear in our list of irregular forms. Detecting and correcting other philological errors is more challenging, and can only be addressed through expert input.

There are, for example, a large number of words in Old Icelandic for which the correct paradigm is "incomplete." Given the "inflect everything" approach of the inflection engine, the stored paradigms for these words are "overly complete" and include non-existent or implausible forms. Because lexical resources rarely mark for these phenomena in a standard way, there is no easy method to limit the generation of these forms. Examples of such words include nouns that have no plural (*singularia tantum*, e.g., *elli* "old age") or no singular (*pluralia tantum*, e.g., *rǫk* "reason"). Another large class of words for which the correct paradigms are incomplete is adjectives, which often miss a degree of comparison. *Dauðr* "dead," as noted, does not have comparative or superlative forms. Similarly, *óðri/óðstr* "higher/highest" lacks the positive degree. A small number of adjectives have only a "weak" inflection, such as *vinstri* "left" or *hǫgri* "right."

The main area in which implausible forms are automatically generated is in the mediopassive forms of verbs (*miðmynd*). Although the mediopassive was a common category in Old Icelandic, a large number of verbs either did not have mediopassive forms, such as the verb *fljúga* "fly," or their mediopassive forms occurred only in archaic language such as poetry (Ottósson 2008). The verbs *verða* "become"

and *vera* "be" do not have mediopassive forms in Modern Icelandic, but several obviously mediopassive forms are attested in early poetic language: "bróðr muno beriaz / oc at bǫnom verðaz" (*Vǫluspá*, st. 45) [brothers will fight / and become banes of each other]; "gumnar margir / erosc gagnollir" (*Hávamál*, st. 32) [many men / are faithful to each other]; "viðrgefendr oc endrgefendr / erost lengst vinir" (*Hávamál*, st. 41) [the giving and the receiving / are friends with each other the longest]; "lyst váromc þess lengi / at lyfia ycr elli" (*Atlamál in grǿnlenzku*, st. 78) [I have long desired / to cure the two of you of old age].[16]

Although the attestation of these forms warrants their place in the inflectional tables, generation of a full paradigm entails several problems. First, because only a very small number of forms from these putative paradigms are attested, it is not clear what the unattested forms would have looked like. Would 2. sg. indicative of the putative verb **verask* have been **ersk* (i.e., < *er* + *-sk*), **esk* (< *es* + *-sk*), **eszk* (< *est* + *-sk*), or **erzk* (< *ert* + *-sk*)? Although some of these putative forms are more plausible than others, there is no way to know for sure which one would have been the actual form used by the speakers (or whether there would have existed more than one form in the language).

On the other hand, some of the forms that may be reconstructed purely mechanically are implausible from the semantic point of view. To adduce an example, it is unlikely that 2./3. sg. mediopassive forms of the verb *vera* could have existed: unlike the 1. sg. form *erumk*, which may be plausibly rendered as "is to me" in English (cf. the example *lyst várumk lengi* above), the putative 2./3. sg. forms would have no logical meaning (*'is to you,' *'is to him/her' vel sim.). Such a sense is normally expressed in Old Icelandic with the help of periphrastic constructions, *er þér* "is to you," *er honum/henni* "is to him/her." The plural forms *erumsk* (1. pl.) and *eruzk* (2. pl.), however, are plausible, even though they are not attested. Their respective meanings would have most likely been "we are to each other" and "ye are to each other" (cf. 3. pl. *erusk* "they are to each other"). Similar observations obtain for the mediopassive forms of the verb *verða* discussed earlier: while the plural forms may well be said to have expressed reciprocity, such as "become X to each other," and the (unattested) 1. sg. form **verðumk* may have meant "becomes to me," the existence of 2./3. sg.

16. Literally, "desire was long to me to cure the two of you of old age" (i.e., "kill you"). Selections from Neckel (1983).

form * *verðsk* (vel sim.) is implausible in any mood (indicative, optative, or imperative).

An interesting by-product of our form-matching work is the identification of words that are attested in the corpus but are absent from our dictionary. Many of these "missing words" are artifacts of our dictionary parsing and can be corrected manually. Other "missing" words include those generated by simple word-formation processes, such as adding the negative prefix *ó-*, such as *ófǫrr* "impassable," *ófǫra* "impassable place; great obstacle." That said, some attested words do not appear in the dictionary in any form. These words are largely compounds, such as *konungsdóttir* "king's daughter," *konungasonr* "king's son," and *skjaldmeyjaflokkr* "band of valkyries." As more machine-actionable texts are made available, the number of such lexemes will likely increase. Consequently, we will need to extend our dictionary by adding resources such as Fritzner (1886–1896) and Ordbog over det Norrøne Prosasprog (ONP), and refine our compound word-recognition routines, perhaps including a compound word-generation module.[17]

IceMorph and Expert Systems for Old Icelandic Morphological Analysis

Our work on IceMorph has highlighted some of the challenges of working with morphonologically complex languages. We have shown that a multi-pronged approach to the problem of accurately tagging texts with rich morphosyntactic detail may offer a significant information gain over existing resources or a single-method approach. Importantly, IceMorph is a dynamic system, one that uses incremental expert feedback to rapidly increase the accuracy of the system.[18] We believe that this type of expert system can be used to address similar

17. Although most of these unmatched words are compounds built from simple words present in the dictionaries, there may also occur rare instances of non-compound nouns. One such example may be the noun **hór* "lover," which is not registered in standard dictionaries but is likely to have existed in the spoken language as a variant of the standard form *hórr* "lover" (for a fairly recent discussion, see Vijūnas 2005).

18. In our work, we hypothesized several use cases that guided the development of our tools and our user interfaces as described in the Appendix. Although we have not had the resources to develop an application programming interface (API) for IceMorph, we have made the tagged output available as downloadable resources on the main website (www.purl.org/icemorph/download). We are also eager to learn more about research workflows so that we can incorporate these into future refinements of the user interface.

problems in other morphonologically complex ancient languages for which existing resources are similarly scarce.

The convergence of fast, accurate algorithms for language analysis, low-cost storage, accessible computing, and relatively clean machine-actionable texts and dictionaries presages a potential revolution in the study of Old Icelandic language and literature. IceMorph augments many successful projects in bringing Old Icelandic language and text resources into the digital age. Our future work will focus on increasing the accuracy of our dictionary resources, adding additional texts to the target corpus, incorporating ongoing expert feedback to correct erroneous alignments, and the development of word-study tools for sophisticated analysis of word and word-class use patterns in the Old Icelandic corpus. Along with other projects focused on Old Icelandic and other early Nordic languages and corpora, we anticipate a time in the not-too-distant future when historical linguists and textual scholars alike will have access to rich, accurately tagged, comprehensive corpora.

Works Cited

Árnason, Kristján. 1985. "Morphology, Phonology, and U-Umlaut in Modern Icelandic." In *Phono-Morphology: Studies in the Interaction of Phonology and Morphology*, edited by Edmund Gussmann, 9–22. Lublin: Catholic University of Lublin.

Brants, Thorsten. 2000. "TnT: A Statistical Part-of-Speech Tagger." In *Proceedings of the Sixth Conference on Applied Natural Language Processing* (ANLC '00): 224–31. Seattle, WA.

Brill, Eric. 1992. "A Simple Rule-Based Part of Speech Tagger." In *Proceedings of the Third International Conference on Applied Natural Language Processing* (ANLC '92): 152–5. Trento, Italy.

Cleasby, Richard, and Gudbrand Vigfússon. 1874. *An Icelandic–English Dictionary*. Oxford, England: Clarendon Press.

Clover, Carol, and John Lindow, eds. 1985. *Old Norse–Icelandic Literature: A Critical Guide*. Islandica 45. Ithaca, NY: Cornell University Press.

Forsberg, Markus, and Aarne Ranta. 2004. "Functional Morphology." *ACM SIGPLAN Notices* 39 (9): 213–23.

Fritzner, Johan. 1886–1896. *Ordbog over det gamle norske Sprog*. Christiania: Feilberg og Landmarks.

Garside, Roger. 1987. "The CLAWS Word-Tagging System." In *The Computational Analysis of English: A Corpus-Based Approach*, edited by Roger Garside, Geoffrey Leech, and Geoffrey Sampson, 30–41. London: Longman.

Gordon, E. V. 1927. *An Introduction to Old Norse* (2nd ed.), revised by A. R. Taylor. Oxford, England: Clarendon Press.

Harðarson, Jón Axel. 2001. "Hvað tekur við eftir dauðann? Um u-hljóðvarp í íslenzku." Óprentaður fyrirlestur fluttur á Rask-ráðstefnu, laugardaginn 27. janúar.

Heusler, Andreas. 1913. *Altisländisches Elementarbuch*. Heidelberg: Carl Winter Universitätsverlag.

Iversen, Ragnvald. 1923. *Norrøn grammatikk*. Christiania: Aschehoug.
Jónsson, Guðni, and Bjarni Vilhjálmsson. 1943–1944. *Fornaldarsǫgur Norðurlanda*. 3 vols. Reykjavik: Bókaútgáfan Forni.
Larsson, Ludvig. 1891. *Ordförrådet i de älsta isländska handskrifterna leksikaliskt ock gramatiskt ordnat*. Lund, Sweden: P. Lindstedt.
Levenshtein, Vladimir I. 1966. "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals." *Soviet Physics Doklady* 10 (8): 707–10.
Loftsson, Hrafn. 2007. "Tagging Icelandic Text Using a Linguistic and a Statistical Tagger." *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the ACL*: 105–8. Rochester, NY.
———. 2008. "Tagging Icelandic Text: A Linguistic Rule-Based Approach." *Nordic Journal of Linguistics* 31 (1): 47–72.
Loftsson, Hrafn, and Anton Karl Ingason. 2009. *IceNLP: A Natural Language Processing Toolkit for Icelandic. User Guide*. Reykjavík: Reykjavík University.
Neckel, Gustav. 1983. *Edda: Die Lieder der Codex Regius nebst verwandten Denkmälern*. Band I. Text. 5. verbesserte Aflage von Hans Kuhn. Heidelberg: Carl Winter Universitätsverlag.
Noreen, Adolf. 1884. *Altisländische und altnorwegische Grammatik*. Halle: Max Niemeyer.
Ottósson, Kjartan. 2008. "The Old Nordic Middle Voice in the Pre-Literary Period." In *Interdependence of Diachronic and Synchronic Analyses*, edited by Folke Josephson and Ingmar Söhrman, 185–219. Amsterdam: John Benjamins.
Poschenrieder, Thorwald. 2000. "Altisländisch: Primärquellenbezug und das Greifswalder RWH-Projekt 'Töluspá.'" In *Ad Fontes!: Quellen erfassen-lesen-deuten: Was ist Computerphilologie?*, edited by Christof Hardmeier, Wolf Dieter Syring, Jochen D. Range, and Eep Talstra, 99–127. Contributions to the Computerphilologie Conference, November 5–8, 1998, Ernst-Moritz-Arndt-Universit at Greifswald. Amsterdam: VU University Press.
Rydberg-Cox, Jeffrey A. 2005. "The Cultural Heritage Language Technologies Consortium." *D-Lib Magazine* 11 (5). http://www.dlib.org/dlib/may05/rydberg-cox/05rydberg-cox.html.
Rögnvaldsson, Eiríkur. 1981. *U*-hljóðvarp og önnur *a~ö* víxl í nútímaíslensku. *Íslenskt mál og almenn málfræði* 3: 25–58.
Rögnvaldsson, Eiríkur, and Sigrún Helgadóttir. 2008. "Morphological Tagging of Old Norse Texts and Its Use in Studying Syntactic Variation and Change." In *2nd Workshop on Language Technology for Cultural Heritage Data*, 40–46. LREC 2008 workshop, Marrakech. Paris: ELRA.
———. 2011. "Morphosyntactic Tagging of Old Icelandic Texts and Its Use in Studying Syntactic Variation and Change." In *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, edited by Caroline Sporleder, Antal van den Bosch, and Kalliopi Zervanou, 63–76. Berlin: Springer.
Steblin-Kamenskij, M. I. 1953. *Drevneislandskij jazyk*. Moskva: Izdatel'stvo literatury na inostrannyx jazykax.
Urban, Kryztof, Timothy R. Tangherlini, Aurelijus Vijūnas, and Peter M. Broadwell. 2014. "Semi-Supervised Morphosyntactic Classification of Old Icelandic." *PLOS ONE* 9 (7): e102366.
Vijūnas, Aurelijus. 2005. "In Defense of a Lover (48. stanza of *Hárbarðsljóð* re-visited)." *Arkiv för nordisk filologi* 120: 221–32.
Zoëga, Geir T. 1910. *A Concise Dictionary of Old Icelandic*. Oxford, England: Clarendon Press.

Online Resources

Clunies-Ross, Margaret, Kari Ellen Gade, Guðrún Nordal, Edith Marold, Diana Whaley, and Tarrin Wills, eds. 2007–2012. The Skaldic Project. https://www.abdn.ac.uk/skaldic.

Crane, Gergory, ed. 2012. Perseus Digital Library. http://www.perseus.tufts.edu.

Crist, Sean, ed. 1998–2013. Germanic Lexicon Project. http://lexicon.ff.cuni.cz.

Handrit.is. 2009–2013. http://handrit.is.

Heimskringla. 2005–2013. Norræni textar og kvæði. http://www.heimskringla.no.

Medieval Nordic Text Archive. 2001–2013. http://www.menota.org.

Netútgáfan. 2001–2013. http://www.snerpa.is/net/netutgaf.html.

ONP: Ordbog over det norrøne prosasprog. 2010–2013. http://onp.ku.dk.

Wallenberg, Joel C., Anton Karl Ingason, Einar Freyr Sigurðsson and Eiríkur Rögnvaldsson, eds. 2011. Icelandic Parsed Historical Corpus (IcePaHC). Version 0.9. http://www.linguist.is/icelandic_treebank.

Appendix

We briefly describe two use case scenarios here, recognizing that the distinction between "student" and "researcher" is fictitious. IceMorph is not intended to be a comprehensive Old Icelandic study environment in its current state. Instead, it is intended to provide accurate word study tools, and a method for accessing increasingly accurately tagged text as part of a research or learning strategy.

> *Use Case Scenario 1*: A student is asked to provide a translation of a passage as part of his work in a graduate course in Old Icelandic. As part of the exercise, he is asked to identify and inflect the strong feminine nouns in a saga passage. Because he can access the inflected dictionary, he is able to move with more confidence through the text, easily checking his own work against a stable and authoritative resource.

> *Use Case Scenario 2*: A researcher is writing a monograph that includes a chapter on word-use patterns in the *Fornaldarsögur*. She begins the process of exploration and discovery on the IceMorph platform. Once satisfied with the inflectional accuracy and disambiguation of the target sub-corpus of texts, she moves those resources onto her desktop for further manipulation and study and integration into her monograph.