Something Old, Something New: A Computational Morphological Description of Old Swedish

Lars Borin, Markus Forsberg

Språkbanken, Department of Swedish Language
University of Gothenburg
Box 200, SE-405 30 Gothenburg, Sweden
lars.borin@svenska.gu.se, markus.forsberg@gu.se

Abstract

We present a computational morphological description of Old Swedish implemented in Functional Morphology. The objective of the work is concrete – connecting word forms in real text to entries in electronic dictionaries, for use in an online reading aid for students learning Old Swedish. The challenge we face is to find an appropriate model of Old Swedish able to deal with the rich linguistic variation in real texts, so that word forms appearing in real texts can be connected to the idealized citation forms of the dictionaries.

1. Background

1.1. Motivation

Languages with a long written tradition have accumulated over the centuries a rich cultural heritage in the form of texts from different periods in the history of the language community. In these texts, we find information on many aspects of the origins and history of our culture.

Since languages change over time, older texts can be difficult or impossible to understand without special training. Indeed, the oldest extant texts in many European languages must in fact be translated in order to be accessible to a modern reader. In Sweden, the training of professionals who can keep this aspect of our cultural heritage alive by conveying the content of, e.g., the oldest Swedish legal and religious texts (13th–15th c.) to a modern audience, is the province of the academic discipline of Swedish Language and Linguistics at Swedish universities.

Consequently, an important component of all Swedish Language university curricula in Sweden is the study of the older stages of the language. For obvious reasons, any form of any language older than a bit over a century will be accessible only in writing. Hence, our main source of historical Swedish language data is in the form of old texts, and courses in the history of the Swedish language, in comparative Scandinavian and in Old Swedish, all contain – in time-honored fashion – a module where students are required to read a certain amount of texts. In equally time-honored fashion, students have two main tools for working with Old Swedish texts: dictionaries and grammar books.

There are excellent editions of Old Swedish texts available in book form, as well as good grammatical descriptions (Noreen, 1904; Wessén, 1969; Wessén, 1971; Wessén, 1965; Pettersson, 2005) and reference dictionaries (section 1.4). Both text collections and reference works such as dictionaries and grammars tend to be out-of-print works, which does not in any way detract from their usefulness or their accuracy, but which does present some practical problems. Students are confined to using reference copies held in departmental libraries, but often departments see themselves forced to restrict access because of the excessive wear this causes.

However, historical texts are increasingly available also in digital form on the internet. Two sites with extensive collections of Old Swedish texts are *Språkbanken* (the Swedish Language Bank; <http://spraakbanken.gu.se>) at the University of Gothenburg and *Fornsvensk textbank* (the Old Swedish Text Bank; <http://www.nordlund.lu.se/Fornsvenska/Fsv%20Folder/index.html>) at Lund University. The standard reference dictionaries of Old Swedish have also been digitized (by Språkbanken) and are available for word lookup via a web interface <http://spraakbanken.gu.se/fsvldb/> and in full-text for research purposes by special agreement.

1.2. Toward a Solution

Our project aims to aid online reading of Old Swedish texts (cf. Nerbonne and Smit (1996)) by providing access to automatic morphological analysis and linking that to available lexical resources.

For our goals, we need morphological analysis, i.e. an analysis module which returns, crucially, a lemma as part of its result, since the lemma is necessary in order to access the online dictionaries. This makes the methodology of our project different from that of, e.g., Rayson et al. (2007), where a POS tagger trained on Modern English was adapted to deal with the Early Modern English of Shakespeare's time. It is doubtful whether this methodology would have helped much if the target language had been Old English (Anglo-Saxon) instead, because of the much greater linguistic distance between the two varieties, which are most appropriately seen as two entirely different, but related languages. What we have in our case are two historical language stages as far apart linguistically as Modern English and Anglo-Saxon (see section 1.3).

POS tagging is also a different kind of analysis from that provided by a morphological analyzer. The former will provide only one analysis of each text token, the most probable one in the given context, whereas the latter will provide all readings licensed by the lexicon and grammar, but will not take context into account. POS tagging will *always* provide an analysis, however, whereas a morphological analyzer may fail to do so if a lemma is lacking in its lexicon or a form is not allowed by its grammar. Thus, POS tagging and morphological analysis are complemen-

tary; POS tagging is often used to select among competing morphological analyses, the combination of the two methods thus providing both disambiguation and lemmatization. Our goal in the project described here is to provide (undisambiguated) morphological analysis including lemmatization of text words, in order to link them to the corresponding dictionary entries. A natural future extension of the work would be a disambiguation module, either a POS tagger or, e.g., a constraint grammar (Karlsson et al., 1995), but we believe that the differences between Old and Modern Swedish are great enough to preclude an easy adaptation of a Modern Swedish POS tagger or constraint grammar.

We are partway through the project. The morphological analysis module is completed for all regular paradigms and some others (section 2), and a small number of lexical entries have been provided with inflectional information (section 2.3). This is primarily what we report on in this paper. We are now in the process of providing all lexical entries with inflectional information in the form of paradigm identifiers (section 2). This task is not completely trivial, since the extant texts do not always allow us to determine to which inflectional class a particular lexical entry should belong.

However, the main challenge still remaining is the issue of how to deal in as principled a way as possible with the considerable linguistic variation present in the texts that we are working with, and which presents us with a different situation compared to working with modern texts (sections 1.3 and 3).[1]

1.3. Down the Foggy Ruins of Time: Orthographic Inconsistency, Linguistic Variation and Language Change in Old Swedish Texts

The texts are from the so-called Old Swedish period (1225–1526), conventionally subdivided into Classical Old Swedish (1225–1374) and Late Old Swedish (1375–1526), covering a time span of 300 years. The language of the extant Old Swedish texts exhibits considerable variation, for at least the following reasons:

1. The orthography was not standardized in the way that we expect of modern literary languages;
2. the language itself was not standardized, in the sense, e.g., that a deliberate choice would had been made about which of several competing forms should be used in writing; and
3. the Middle Ages was a time of rapid language change in Swedish, perhaps more so than any subsequent period of similar length.

The first shows itself in a great variation in spelling; the "same" word can be spelled in a number of different ways, even on the same page of a document. Not only is there variation in the orthography itself, but also geographical variation, because no unified standard variety had been established at the time when the texts were produced.

The second factor makes itself felt in the number of variant forms in our inflectional paradigms (section 2).

As for the third factor, during the second half of the Old Swedish period the language underwent a development from the Old Norse (or Modern Icelandic, or Old English) mainly synthetic language type to the present, considerably more analytical state. In addition (or perhaps compounding this process), the sound system of Swedish was thoroughly reorganized.

For instance, in the nouns, the case system changed profoundly during this period, from the old four-case system (nominative, accusative, dative, genitive, in two numbers) to the modern system with a basic form and a genitive clitic which is the same in all declensions (as opposed to the old system where there were a number of different genitive markers), and where most functions that the older case forms expressed by themselves have been taken over by a combination of free grammatical morphemes and a much more rigid constituent order.

In the texts that interest us, these changes are in full swing, which manifests itself as variation in inflectional endings and in the use of case and other inflectional categories and in the distribution of the corresponding forms.

It is not always easy to tease out the contributions of these different factors to the linguistic motley evinced by the texts. Without doubt, the diachronic component is important – the texts are after all from a period three centuries in length – but it is also probable that the lack of standardization simply allows normal synchronic language variation to "shine through" in the texts, as it were, rather than being eliminated as is normally the case with modern, normalized written standard languages.

1.4. The Reference Dictionaries

The three main reference dictionaries of Old Swedish are:

- Söderwall (1884) (23,000 entries);
- Söderwall (1953) (21,000 entries); and
- Schlyter (1887) (10,000 entries)

The overlap between the three dictionaries is great, so that we are actually dealing with less than 25,000 different headwords. On the other hand, compounds – whether written as one word or separately – are not listed as independent headwords, but as secondary entries under the entry of one of the compound members. Thus, a full morphological description reflecting the vocabulary of the three dictionaries will contain many more entries, possibly by an order of magnitude more.

As an example of the kind of information that is available in the dictionaries, we will briefly discuss the entries for the word *fisker* (Eng. 'fish'), as it appears in these dictionaries. The entry *fisker* in Söderwall (1884) is shown in Figure 1. From this entry we learn that *fisker* is a masculine noun (indicated by "m.", in the second line of the entry),

[1] Of course we are quite aware of the fact that spelling variation is an empirical fact of modern language as well. There would be no need of spellcheckers otherwise. In the case of Old Swedish, however, there were no spelling norms, as far as we can tell. At the most, there were local scribal practices in the monasteries (which was where most of the text production occurred), different in different places and never codified, to our knowledge.

fisker (Söderwall) (fysker Lg 3: 301; -ar BSH 5: 5067 (1512). fiisker: fiisk RK 3: 4179. -ar), m. [Isl. fiskr] L. **1)** *fisk.* han tok w fiske tolpänigh Bu 100. taka fiska KL 12. thz första han katadhe vt sin krok tha fik lhan en storan fisk ib. ib 13. Bo 240. Lg 546, 3: 9, 10, 301, 302. i slike watne äru tholka fiska GO 978. ätin the fiska oc hwitan maat Bir 4: 15. färska ällir salte fiska ib 5: 32. tw pund skarpa fisca SD NS 1: 656 (1407). - koll. han (qvarndammen) skal vara open _ree vikur vm varenä _aa fisken gaar vpp ok swa lenge vm hösten. _aa vatneth er mykith ok fiksen gar vpp FH 3: 4 (1352). ib 4: 15 (1451), 16. SD 5: 699 (1347, gammal afskr.). äta fisk oc hwitan maat Bir 4: 15. VKR 17, 62. fäghin är han som fyrme ok findher han fikh (för fiskh) a diska GO 105. tw stykke fisk Bir 5: 31. tw stykke färskan fisk ib 32. eet stykke stekan fisk Bo 234. ii pund fiisk RK 3: 4179. **2)** ?iiij (4) lösa järn bultar, item xi (11) lösa fyskar, item 1 fangabult BSH 5: 506 (1512). - Jfr arbeidis-, bnären-, flat-, horn-, hval-, skal-, skarp-, skat-, sma-, spit-, stok-fisker. — **fiska bater** (-baater: -baat Su 363), m. *fiskarebåt.* Su 363. — **fiska ben**, n. *fiskben.* eet fiska ben sath fast j hans halse Bil 900. KL 370. ST 102. — **fiska dike**, n. *fiskdamm.* ST 299. — **fiska drät**, f.L. — **fiska fiäl**, n. *fiskfjäll.* aff rutnom fiska fiällom Bir 3: 203. — **fiska fänge** (fiske-), n. *fiskafänge.* aff the fiske fängeno Lg 3: 11. aff hwario fiske fänge ib. — **fiska hovudh** (hwffwd LB 7: 265), n. *fiskhufvud.* LB 7: 265. PM XLVIII. — **fiska kyn** (-kön), n. *fiskslag.* alla handa fiska kön Al 6495. — **fiska lim**, n. *fisklim.* tak fiska lim giorth aff maghommen PM XLVII. — **fiska liver** (-leffwer: leffrenas PM XXXVIII), f. *fisklefver.* PM XXXVIII. — **fiska läghe**, n. *fiskläge.* RK 1: (Yngre red. af LRK) s. 263. Jfr fiskeläghe. — **fiska skal**, f. *musselskal, snäckskal.* trykte han ällir wredh vth aff vlla fätthen ena fiska skal äller eeth kar fwlt mz daagh (concham rore implevit) MB 2: 88. — **fiska slagh**, n. *fiskslag.* mang the fiska slagh, som aldrig fingos ther förra Lg 3: 11. — **fiska sudh** (fiiska sodh LB 7: 159), n. *fiskspad.* aff fersko fiska sudhi LB 3: 182. ib 7: 159. — **fiska thiuver**, m. L.

Figure 1: The entry *fisker* (Eng. 'fish') in Söderwall's dictionary of Old Swedish

and that it has been attested in a number of variant spellings (*fysker*, *fiisker*, *fiisk*). We also find references to occurrences of the word in the classical texts, and finally there is a listing of the compounds in which it occurs, e.g. *fiska slagh* (Eng. 'type of fish'). Söderwall (1953) is, basically, intended as a complement to Söderwall (1884), citing more forms and more attestations, originating in texts that became available after Söderwall's time.[2] Schlyter (1887) – as its title indicates – describes the vocabulary of the medieval Swedish laws, and its entries generally contain a bit less information than those in Söderwall (1884).

2. A Computational Morphology for Old Swedish

2.1. Functional Morphology

The tool we are using to describe the morphological component is Functional Morphology (FM) (Forsberg, 2007; Forsberg and Ranta, 2004). We chose this tool for a number of reasons: it provides a high-level description language (namely the modern functional programming language Haskell (Jones, 2003; Haskell, 2008)); it uses the character encoding UTF-8; it supports tasks such as (compound) analysis and synthesis; and, perhaps most importantly, it supports compilation to many standard formats, such as XML (The World Wide Web Consortium, 2000), LexC and XFST (Beesley and Karttunen, 2003), GF (Ranta, 2004), and full-form lexicons, and provides facilities for the user to add new formats.

[2] Although Söderwall is given as the author of this work, it was actually compiled after his death by members of *Svenska fornskriftsällskapet* (the Swedish Ancient Text Society).

The morphological model used in FM is *word and paradigm*, a term coined by Hockett (1954). A paradigm is a collection of words inflected in the same manner and is typically illustrated with an inflection table.

An FM lexicon consists of words annotated with paradigm identifiers from which the inflection engine of FM computes the full inflection tables.

Consider, for example, the citation form *fisker*, which is assigned the paradigm identifier *nn_m_fisker*. The paradigm identifier carries no meaning, it could just as well be any uniquely identifiable symbol, e.g. a number, but we have chosen a mnemonic encoding. The encoding is read as: "This is a masculine noun inflected in the same way as the word *fisker*" (which is trivially true in this case). If the paradigm name and the citation form is supplied to the inflection engine, it would generate the information in Table 1. To keep the presentation compact, we have contracted some word forms, i.e., the parenthesised letters are optional.

We also show (in the last column of Table 1) how this paradigm is presented in traditional grammatical treatises of Old Swedish, e.g. those by Wessén (1969) and Pettersson (2005). For a discussion of the differences between the our paradigms and those found in traditional grammatical descriptions, see section 2.3

The starting point of the paradigmatic specification, besides the dictionaries themselves, are the standard grammars of Old Swedish mentioned above, i.e., those by Noreen (1904), Wessén (Wessén, 1969; Wessén, 1971; Wessén, 1965), and Pettersson (2005). The number of paradigms in the current description by part of speech are

nn_m_fisker fisker ⇒

Lemma	fisker			**Traditional normalized form**
POS	nn			
Gender	m			
Number	**Def**	**Case**	**Word form**	
sg	indef	nom	*fisker*	*fiskẹr*
sg	indef	gen	*fisks*	*fisks*
sg	indef	dat	*fiski, fiske, fisk*	*fiski, fisk*
sg	indef	ack	*fisk*	*fisk*
pl	indef	nom	*fiska(r), fiskæ(r)*	*fiska(r)*
pl	indef	gen	*fiska, fiskæ*	*fiska*
pl	indef	dat	*fiskum, fiskom*	*fiskum*
pl	indef	ack	*fiska, fiskæ*	*fiska*
sg	def	nom	*fiskrin*	*fiskrin*
sg	def	gen	*fisksins*	*fisksins*
sg	def	dat	*fiskinum, fisk(e)num*	*fiskinum*
sg	def	ack	*fiskin*	*fiskin*
pl	def	nom	*fiskani(r), fiskæni(r)*	*fiskani(r)*
pl	def	gen	*fiskanna, fiskænna*	*fiskanna*
pl	def	dat	*fiskumin, fiskomin*	*fiskumin*
pl	def	ack	*fiskana, fiskæna*	*fiskana*

Table 1: The inflection table of *fisker*

as follows:

Part of speech	# of paradigms
Noun	38
Adjective	6
Numeral	7
Pronoun	15
Adverb	3
Verb	6

2.2. The FM Description

The paradigms of Old Swedish, which in our description amount to 75 paradigms, are defined using the tool Functional Morphology (FM). We will now give some technical details of the implementation by explaining how some of the verb paradigms in our morphology were defined. The main objective is not to give a complete description, but rather to provide a taste of what is involved. The interested reader is referred to one of the FM papers.

An implementation of a new paradigm in FM involves: a type system; an inflection function for the paradigm; an interface function that connects the inflection function to the generic lexicon; and a paradigm name. Note that if the new paradigm is in a part of speech previously defined, then no new type system is required.

A paradigm in FM is represented as a function, where the input is one or more word forms (typically the citation form or principle parts) and a set of morphosyntactic encodings, and the output of the function is a set of inflected word forms computed from the input word forms. It is a set instead of a single word form to enable treatment of variants and missing cases.

More concretely, if we represent the paradigm of regular nouns in English as a function, and only consider a morphosyntactic encoding for number, we would then define a function that expects a regular noun in nominative singular. If this function is given the word `"elephant"`, then the result would be another function. This function would, if an encoding for singular is given to the function, return `{"elephant"}`, and if an encoding for plural is given, return `{"elephants"}`. The resulting function may be translated into an inflection table given that the morphosyntactic encoding is ensured to be enumerable and finite (how this is ensured in FM will not be discussed here).

Turning now to the verb paradigms of Old Swedish, a `Verb` is a function from a morphosyntactic encoding, `VerbForm`, to a set of word forms with the abstract name `Str`.

```
type Verb = VerbForm -> Str
```

The type `VerbForm` defines the inflectional parameters of Old Swedish verbs. We only include those parameter combinations that actually exist, which will ensure, by type checking, that no spurious parameter combinations are created. A morphosyntactic encoding in FM is an algebraic data type, consisting of a list of constructors, where a constructor may have zero or more arguments. The vertical line should be interpreted as disjunction. The arguments here are also algebraic data types (only the definition of `Vox` is given here). A member of this type is, for example, `Inf Active`, where `Active` is a constructor of the type `Vox`.

```
data VerbForm =
 PresSg Modus Vox          |
 PresPl Person Modus Vox   |
 PretInd Number Person Vox |
 PretConjSg Vox            |
 PretConjPl Person Vox     |
 Inf Vox                   |
 ImperSg                   |
 ImperPl Person12
```

```
data Vox =
 Active |
 Passive
```

The `VerbForm` expands into 41 different parameter combinations. These parameter combinations may be given any string realization, i.e., we are not stuck with these rather artificially looking tags, we can choose any tag set. For example, instead of `PretConjSg Passive`, we have *pret konj sg pass*.

The next step is to define some inflection functions. We start with the paradigm of the first conjugation, exemplified by the word *ælska* (Eng. 'to love'). The inflection function `aelska_rule` performs case analysis on the `VerbForm` type. There is one input word form, which will be associated with the variable `aelska`. The function `strs` translates a list of strings to the abstract type `Str`. The function is built up with the support of a set of helper functions, such as `passive` that computes the active and passive forms, `tk` that removes the nth last characters of a string, and `imperative_pl` that computes the plural imperative forms (inflected for person).

```
aelska_rule :: String -> Verb
aelska_rule aelska p =
 case p of
 PresSg Ind Act ->
  strs [aelska++"r",aelska]
 PresSg Ind Pass -> strs [aelska ++"s"]
 Inf v -> passive v [aelska]
 ImperSg -> strs [aelska]
 ImperPl per ->
  imperative_pl per aelsk
 PresPl per m v   ->
  indicative_pl (per,m,v) aelsk
 PretInd Pl per v ->
  pret_ind_pl (per,v) aelsk
 PretConjPl per v ->
  pret_conj_pl (per,v) (aelsk++"a")
 PresSg Conj v ->
  passive v [aelsk++"i",aelsk++"e"]
 PretInd Sg _ v -> passive v [aelska++"þi"]
 PretConjSg v ->
  passive v [aelska++"þi", aelska++"þe"]
 where aelsk = tk 1 aelska
```

The inflection function `aelska_rule` computes 65 word forms from one input word form, e.g. *kalla* (Eng. 'to call').

Given that we now have defined an inflection function for a verb paradigm, we can continue by defining the other paradigms in relation to this paradigm, i.e., we first give the parameter combinations that differ from `aelska_rule` and finalize the definition with a reference to `aelska_rule`. This is demonstrated in the inflection function `foera_rule`, the paradigm of the third conjugation.

```
foera_rule :: String -> Verb
foera_rule foera p =
 case p of
 PresSg Ind Act ->
  strs [foer++"ir", foer++"i"]
 PresSg Ind Pass -> strs [foer++"s"]
 Inf v -> passive v [foera]
 ImperSg -> strs [foer]
 PretInd Pl per v ->
  pret_ind_pl  (per,v) foer
 PretConjPl per v ->
  pret_conj_pl (per,v) foer
 PretInd Sg _ v -> passive v [foer++"þi"]
 PretConjSg v ->
  passive v [foer++"þi", foer++"þe"]
 _  -> aelska_rule foera p
 where foer = tk 1 foera
```

The last inflection function we present, representing the fourth conjugation paradigm, is `liva_rule`, defined in terms of `foera_rule`. Note that we use two different forms when referring to `foera_rule`: we use `lif++"a"`, i.e., the input form where 'v' has been replaced with 'f', for the preterite (i.e., past tense) cases, and the input form for all other cases.

```
liva_rule :: String -> Verb
liva_rule liva p =
 case p of
 PresSg Ind Act  ->
  strs [liv++"er", liv++"ir", liv++"i"]
 PresSg Ind Pass -> strs [lif++"s"]
 ImperSg         -> strs [lif]
 p | is_pret p   -> foera_rule (lif++"a") p
 _               -> foera_rule liva p
 where liv = tk 1 liva
       lif = v_to_f liv
```

When the inflection functions are defined, we continue with the interface functions. An interface function translates one or more input words, via an inflection function, into an entry in the generic dictionary. This is done with the function `entry` that transforms an inflection function into an inflection table. If the current part of speech has any inherent parameters such as gender, those would be added here. The inherent parameters are not inflectional, they describe properties of a word, which is the reason why they appear at the entry level.

```
vb_aelska :: String -> Entry
vb_aelska = entry . aelska_rule

vb_foera :: String -> Entry
vb_foera = entry . foera_rule

vb_liva :: String -> Entry
vb_liva = entry . liva_rule
```

The interface functions need to be named to connect them with an external lexicon. This is done with the function `paradigm`. The names are typically the same as those of the interface functions. Every paradigm is also given a list of example word forms, which provides paradigm documentation and enables automatic generation of an example inflection table, which is done by FM applying the current interface function to its example word forms. The list of paradigm names, denoted here with `commands`, is later plugged into the generic part of FM.

```
commands = [
  paradigm "vb_aelska" ["ælska"] vb_aelska,
  paradigm "vb_foera"  ["føra"]  vb_foera,
  paradigm "vb_liva"   ["liva"]  vb_liva
          ]
```

We can now start developing our lexicon. The lexicon consists of a list of words annotated with their respective paradigm, e.g. the word *røra* (Eng. 'to touch') and *føra* (Eng. 'to move'), which is inflected according to the paradigm `vb_foera`.

```
vb_foera "røra" ;
vb_foera "føra" ;
```

The lines above are put into an external file that is supplied to the compiled runtime system of FM.

2.3. The Development of the Morphological Description and the Lexicon

In this project, we have collaborated with a linguist who is also an expert on orthographic and morphological variation in Old Swedish. In the first phase of the project, she defined the inflectional paradigms on the basis of the dictionaries and the actual variation empirically observed in the texts.

The FM description was developed in parallel with this work. The linguist selected a set of sample words from the dictionaries and annotated those with the appropriate paradigms. The full inflection tables could then be generated immediately and the result evaluated by the linguist.

At the time of writing, about 3,000 main lexical entries (headwords; see section 1.4) have been provided with inflectional information in the form of a paradigm identifier.

In our work in a parallel project to the one described here, where we are producing a large computational morphological lexicon for modern Swedish (Borin et al., forthcoming 2008a; Borin et al., forthcoming 2008b), the number of inflectional classes (paradigms) turns out to be on an order of magnitude more, i.e., around 1,000 rather than around 100.[3] Note that this holds equally for the written standard language and colloquial spoken Swedish.[4] This is something that calls for an explanation, since under the (generally accepted, at least in some form) assumption of uniformitarianism (Janda and Joseph, 2003), we would not expect to find less diversity in Old Swedish than in the modern language.

First we may note that our morphological description is not yet complete. For example, while it covers all four weak verb conjugations, as yet it accounts for only two out of the nine or so classes of strong and irregular verbs. However, even standard grammars of Old Swedish like that of Wessén (1969) list somewhere in the vicinity of 100 paradigms, and no more. We believe that the main factor here is our lack of information. For many lexical entries it is even difficult to assign an inflectional class, because the crucial forms are not attested in the extant texts, and of course, there are no native speakers on whose linguistic intuitions we could draw in order to settle the matter.

Some of the diversity built into our paradigms could thus conceivably be a case of different lexical entries now brought under the same paradigm, actually consistently using different alternatives for expressing a particular combination of morphosyntactic features; we will probably never know.

In the standard reference grammars of Old Swedish, inflectional paradigms are consistently idealized in the direction of a (re)constructed Old Swedish, arrived at on the basis of historical-comparative Indo-European and Germanic studies. In this connection, the actual variation seen in texts has been interpreted as a sign of language change, of "exceptional" usages, etc.[5] (Johnson, 2003). In our paradigms, we have endeavored to capture the actual variation encountered in the texts and in the dictionary examples (but see section 3).

3. Computational Treatment of Variation

As we have mentioned already (section 1.3), the source of variation in the texts are of three kinds: no standardized spelling; no standardized forms; and language change (diachronic drift). For our work on the computational morphological description of Old Swedish, we have found it natural and useful to make an additional distinction, namely that between *stem variation* and *ending variation*, since it has seemed to us from the outset that we need to treat stems and endings differently in this regard.

This gives us altogether six possible combinations of factors, as shown in the following table:

	spelling variation	lack of lg standardization	language change
stem variation	S_1	S_2/L_1	S_3/L_2
ending variation	M_1	M_2	M_3

(Legend: S=spelling rules; L=lexical component; M=morphological component)

[3] The distribution of inflectional patterns in the modern language is Zipfian in shape: Nearly half the paradigms are singletons, almost a fifth of them have only two members, etc.

[4] Although for slightly different reasons in the two cases: In the written standard language, it is generally the low-frequency words that have unique paradigms, e.g. learned words and loanwords. In the spoken language, high-frequency everyday words show variation in their inflectional behavior. There is some overlap, too, e.g. the strong verbs.

[5] and possibly even of carelessness or sloth on the part of the scribes; cf. the following quote, which well captures an attitude toward linguistic variation traditionally prevalent among linguists:

> Variation in Navajo pronunciation had long disturbed Haile (to Sapir, 30 March 1931: SWL): "Sometimes I do wish that the informants would be more careful in pronunciation and follow some system which would conform to theory. ... Apparently no excuse, excepting that informants are too lazy to use it correctly." Sapir responded (6 April 1931: SWL) that—at least in collecting texts—it was "not absolutely necessary to have the same words spelled in exactly the same way every time."
> (Darnell, 1990, 257)

The table reflects the fact that we have decided already to handle all inflectional ending variation – regardless of its origin – in the morphological component, i.e., our paradigms contain all attested ending variants, still a finite and in fact rather small set, which partly motivates their uniform treatment.

Representing a dead language with a finite corpus of texts, the Old Swedish stems could in theory be treated in the same way. The corpus is big enough, however, that we will need to treat it as unlimited in practice, and hence the stems as a set that cannot be enumerated.

In order not to bite off more than we can chew, we have tentatively decided to treat all stem variation as a spelling problem (with one exception; see below). It will then be natural to look to some kind of solution involving edit distance, e.g. *universal Levenshtein automata*, see, for example, Mihov and Schulz (2004).

However, the spelling is not completely anarchistic, far from it: For example, the /i:/ sound will be written <i>, <y>, <j>, <ii>, <ij>, and possibly in some other ways, but not, e.g., <a> or <m>, etc. Thus, a rule-based method may be more appropriate, or possibly a hybrid solution should be sought.

In the table above, the use of subscripts (S_1, M_3, etc.) hints at the possibility of distinguishing formally among different types of information even within a component. The present morphological description does not make a distinction between ending variation due to spelling variants of the "same" ending (from a historical-normative point of view – e.g., indef sg dat *fiski/fiske* – and ending variation whereby "different" endings occupy the same paradigmatic slot, e.g., indef sg dat *fiski/fisk*. However, there is no technical reason that we could not make this and other distinctions on the level of paradigms or even on the level of individual lexical entries. In fact, our work on the Old Swedish morphological description has clearly indicated the need for this kind of facility.[6]

There is one kind of stem variation which does not fit neatly into the picture painted so far, namely that brought about by inflectional morphological processes, in our case those of Ablaut and Umlaut. At the moment, the strong verb class paradigms do not account for variation in the realization of the Ablaut grades of the stem vowels – which of course we find in the texts – and we are still undecided as to how to treat them, by a separate normalization step or in the FM description. In the latter case we would then probably need to duplicate some information already present in the spelling rules component.

4. Summary and Conclusions

We have implemented a morphological description for Old Swedish using Functional Morphology, a tool which supports automatic morphological analysis and generation, as well as the generation of full-form lexicons. The description is intended to be used in an online Old Swedish text reading aid, where it will perform on-the-fly analysis of words in the texts in order to present the user with possible lexicon entries for the word.

The description is fairly complete, but its usefulness for this intended practical purpose is still limited by the large amount of linguistic variation found in the texts.

We have created a small test lexicon (about 3,000 entries), and we are now working on adding inflectional information to all of the headwords in the digital versions of the Old Swedish reference dictionaries (section 1.4).

We have started to look at the linguistic variation characteristic of the Old Swedish texts. Variation in inflectional endings is already uniformly handled in the morphological component, regardless of its origin, while we have still not made the final decision on a strategy to handle the various kinds of stem variation found in Old Swedish.

[6] It is not difficult to think of situations where this would be useful in modern language descriptions as well; for instance, it would be useful to be able to record the frequency of occurrence of homographs according to which lexical entry they represent.

Acknowledgements

The Old Swedish morphological description on which our computational morphology is based, was made by Rakel Johnson, Department of Swedish Language, University of Gothenburg.

The work presented here was financed in part by Swedish Research Council grant 2005-4211 (2006–2008) awarded to Aarne Ranta, Chalmers University of Technology, for the research project entitled *Library-Based Grammar Engineering*, and in part by the Faculty of Arts, University of Gothenburg, through its support to Språkbanken (the Swedish Language Bank).

We would also like to acknowledge CLT – the Centre for Language Technology, Göteborg <http://www.clt.gu.se> – for providing a creative atmosphere in which multidiscplinary collaborations such as this come naturally.

5. References

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications, Stanford University, United States,.

Lars Borin, Markus Forsberg, and Lennart Lönngren. forthcoming 2008a. The hunting of the BLARK – SALDO, a freely available lexical database for Swedish language technology. In *Festschrift to Professor Anna Sågvall Hein*. Uppsala University, Dept. of Linguistics and Philology.

Lars Borin, Markus Forsberg, and Lennart Lönngren. forthcoming 2008b. SALDO – the Swedish associative thesaurus, version 2. Technical report, Språkbanken, University of Gothenburg.

Regna Darnell. 1990. *Edward Sapir: Linguist, Anthropologist, Humanist*. University of California Press, Berkeley / Los Angeles / London.

Markus Forsberg and Aarne Ranta. 2004. Functional Morphology. *Proceedings of the Ninth ACM SIGPLAN International Conference of Functional Programming, Snowbird, Utah*, pages 213–223.

Markus Forsberg. 2007. *Three Tools for Language Processing: BNF Converter, Functional Morphology, and Extract*. Ph.D. thesis, Göteborg University and Chalmers University of Technology.

Haskell. 2008. Haskell homepage. <http://www.haskell.org>.

Charles Hockett. 1954. Two models of grammatical description. *Word*, 10:210–234.

Richard D. Janda and Brian D. Joseph. 2003. On language, change, and language change – or on history, linguistics, and historical linguistics. In Brian D. Joseph and Richard D. Janda, editors, *Handbook of Historical Linguistics*, pages 3–180. Blackwell, Oxford.

Rakel Johnson. 2003. *Skrivaren och språket*. Ph.D. thesis, Department of Swedish Language, Göteborg University.

Simon Peyton Jones. 2003. *Haskell 98 Language and Libraries: The Revised Report*. Cambridge University Press.

Fred Karlsson, Atro Voutilainen, Juha Heikkilä, and Arto Anttila, editors. 1995. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Number 4 in Natural Language Processing. Mouton de Gruyter, Berlin and New York.

Stoyan Mihov and Klaus Schulz. 2004. Fast approximate search in large dictionaries. *Computational Linguistics*, 30(4):451–477.

John Nerbonne and Petra Smit. 1996. GLOSSER-RuG: In support of reading. In *COLING–96. The 16th International Conference on Computational Linguistics. Proceedings, Vol. 2*, pages 830–835, Copenhagen. ACL.

Adolf Noreen. 1904. *Altschwedische Grammatik*. Halle. Facsimile available online: <http://lexicon.ff.cuni.cz/texts/oswed_noreen_about.html>.

Gertrud Pettersson. 2005. *Svenska språket under sjuhundra år*. Studentlitteratur, Lund, Sweden.

Aarne Ranta. 2004. Grammatical Framework: A Type-theoretical Grammar Formalism. *The Journal of Functional Programming*, 14(2):145–189.

P. Rayson, D. Archer, A. Baron, J. Culpeper, and N. Smith. 2007. Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In *Proceedings of Corpus Linguistics 2007*. University of Birmingham, UK.

C.J. Schlyter. 1887. *Ordbok till Samlingen af Sweriges Gamla Lagar. (Saml. af Sweriges Gamla Lagar 13)*. Lund, Sweden.

Knut Fredrik Söderwall. 1884. *Ordbok Öfver svenska medeltids-språket. Vol I–III*. Lund, Sweden.

Knut Fredrik Söderwall. 1953. *Ordbok Öfver svenska medeltids-språket. Supplement. Vol IV–V*. Lund, Sweden.

The World Wide Web Consortium. 2000. Extensible Markup Language (XML). <http://www.w3.org/XML/>.

Elias Wessén. 1965. *Svensk språkhistoria: Grundlinjer till en historisk syntax*. Stockholm, Sweden.

Elias Wessén. 1969. *Svensk språkhistoria: Ljudlära och ordböjningslära*. Stockholm, Sweden.

Elias Wessén. 1971. *Svensk språkhistoria: Ordböjningslära*. Stockholm, Sweden.