Old Swedish in Functional Morphology*

Markus Forsberg
Department of Computer Science and Engineering
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg, Sweden

August 27, 2007

Abstract

This document presents work in progress on a lexical resource of Old Swedish. The objective is concrete — to connect word forms in real text to entries in dictionaries available in electronic format. The connection is done through a morphological component implemented in Functional Morphology.

The challenge we face is to find an appropriate model of Old Swedish that is able to deal with the rich variation in the spelling of the word forms in real text, and connect them with the idealized citation forms of the dictionaries. The reason for the rich variation is twofold: no spelling standards were available for Swedish at that time, and the time period is three hundred years, during which many natural changes occurs.

1 Introduction

Språkbanken at Göteborg University has created searchable electronic versions of three dictionaries of Old Swedish[1]: Söderwall [10] (23k entries), Söderwall supplement [11] (21k entries) and Schlyter [9] (10k entries), with a total of 54k entries. These are the main authoritative dictionaries for Old Swedish.

The motivation behind these electronic resources was to build an infrastructure on Internet available for Old Swedish education and research. The electronic versions, as the paper versions, are structured with idealized citation forms as the entry points. However, given an arbitrary word form in a real text, it is typically difficult to guess the corresponding idealized citation form. In this situation, it is actually easier to work with the paper versions, since you may lookup an approximate position in a dictionary and browse the neighbouring entries. So the electronic resources have been largely unused. This project aims at changing this situation by creating a morphological component able to suggest appropriate dictionary entries for an arbitrary word form.

*Joint work with Lars Borin and Rakel Johnson.

[1]Dictionaries accessible at: `http://spraakbanken.gu.se/fsvldb/`

The tool we are using to describe the morphological component is Functional Morphology (FM) [3, 2], developed by M. Forsberg and A. Ranta at Chalmers University of Technology. This tool has a number of advantages: it has high-level description language; it supports (compound) analysis and synthesis; and it supports the translation to many other, more standard, formats: fullform, LexC and XFST [1], SQL, GF [8] (gives a direct connection to syntax) et cetera.

2 The dictionaries

Let us examine what kind of information that is available in the dictionaries, by looking at the entries for the word *fisker* (Eng. 'fish'). This word is interesting since it occurs in all three dictionaries.

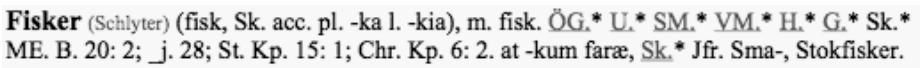
The first entry is from Söderwall. Some of the information provided here is: it is a masculine noun; its stem variations (*fysker*, *fiisker*, *fiisk*): references to occurrences of the word in the classical texts; and the compounds it occurs in, e.g. *fiska slagh* (Eng. type of fish).

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 förста 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.

The second entry is from Söderwall's supplement that supplies more examples and acts as an enrichment of the information in Söderwall.

fisker (Söderwall suppl.) (pl. dat. fiskomon MP 4: 194), m. *fisk.* i swadant vatn swadana fiska HSH 14: 25 (1525, Brask). - koll. affradith är ith hwndra fisk VKJ 151 (1447). han legies for 200 fisk vm aaredh ib. tha räntar hna 50 fisk ib. Jfr bagga-, berger-, flak-, flat-, haf-, hval-, klova-, mat-, mun-, saltavatns-, skal-, sma-, spik-, spit-, steke-, stek-, stok-, storen-, tionda-, thör-, viborgs-fisker. — ***fiska bäkker**, se fiskebäkker. — ***fiska damber** (fiske-), m. *fiskdamm.* ransaa hwath affwel kan göras . . . mädh jäkth dywngardhom fiskrj oc fiskedammom PMSkr 206. göra rwdho äller fiska damma ib 363. fiska fiäl, n. fiskfjäll. lepiotes är en sten liker fiskafyälle PMSkr 480. — ***fiska frätare**, m. *fiskätare.* piscinorus . . . fiskafrätare GU C s. 451. fiska fänge, n. fiskafänge. alla the j fiskafängena mädh honom waro MP 4: 119. — ***fiska gel** (-geell), m. fiskgäl. suencia . . . fiska geell oc drighil GU C 20 s. 566. — ***fiska hus** (fiske-), n. fiskhus, hus för förvrning av (torkad) fisk. haffwe foghten synnerliga sith wisthws oc fiskehws Arnell Brask Bil 28. Jfr fiskhus. — ***fiska karse** (fyska-), m. [Sv. dial. fisk-kars, fisk-kärse m. fl. former se Rietz 312 b, Vendell, Ordb. 193 a] *fiskkasse.* nassa . . . myärdhe ok fyska karse GU C 20. s. 343. — ***fiska kiäpte** (fiska käpte), m. [Jfr Sv. dial. käfte] *fiskgäl.* brancina (för brancia) . . . fiska käpte tu aprehende brancina (för branciam; jfr Tobias 6: 4) eius tak fiskin i käptin GU C 20 s. 61. - fiska leker, ss fiskleker. — ***fiska maghi**, m. *fiskmage.* lim giorth aff fiska maghom PMSkr 519. — ***fiska mes** (fiske-. -mees), f. fiskkasse. gurgustium . . . domus pauperum angusta wlgariter pörthe hörro hws fiska mee ok fiska näth GU C 20 s. 325. jtem skule the (_: fiskarna) haffwa . . . fiskemens, korgia Arnell Brask Bil 29. — ***fiska nät**, n. *fisknät.* gurgaustium . . . pörthe hörro hws fiska mees ok fiska näth GU C 20 s. 325. — ***fiska pakkare** (fiske-. fiskie-) ST Ämb 238 (1544)), m. fiskpackare. St Ämb 234 (1542), 238 (1544)), — ***fiska parker** (fiske-), m. fiskpark, (större) fiskdamm. i mit land ära fiske prkä hetherlikä fäm Saml 6: 66. — ***fiska saltlaki**, m. saltlake vari fisk är inlagd. Se Sdw 2: 1216. — **fiska skal**, f. *muselskal, snäckskal.* ostracites är en sten liter them fiska skalommen som kallas ostree PMSkr 485. — ***fiska spordher** (fyska spol), m. *fena.* (Jfr Sv. dial. spord, spol, fena. Rietz 659 a, Vendell 919 b). pinna fyska spol GU C 20 s. 449. — ***fiska sumper**, m. *fisksump.* STb 3: 86 (1492). — ***fiska torgh**, n. fiskkorg. ss egennamn. engin kasta orenligheet pa stoora torget eller fiska torget vidh sina xij (12) marck STb 1: 432 (1459, burspr.). SJ 62 (1480) etc. — ***fiska trä**, n. *fisktunna.* them (_ fiskarna) skal fogten . . . besöria mz . . . fiscaträä Arnell Brask Bil 29. — ***fiska tunna** (pl. med art. tunnnaner STb 1: 424 (1459, burspr)), f. *fisktunna.* STb 1: 434 (1459, burspr), 474 (1478, burspr), 492 (1492, burspr.). — ***fiska vräkare** (fiske-. -vrakare), m. »fiskvräkare», person som det är ålagt att kontrollera den i en sal saluförda fisken. fiske wrakare ST Ämb 49 (1450). ib 51 (1451). fiska vrakara oc at wraka näffwer, bräde och kalk ib 58 (1454). Jfr fiskvräkare.

Finally, we have the entry of Schlyter, which is, in this case, rather sparse.



Fisker (Schlyter) (fisk, Sk. acc. pl. -ka l. -kia), m. fisk. ÖG.* U.* SM.* VM.* H.* G.* Sk.* ME. B. 20: 2; _j. 28; St. Kp. 15: 1; Chr. Kp. 6: 2. at -kum faræ, Sk.* Jfr. Sma-, Stokfisker.

3 Paradigms

The linguistic model we are using is *word and paradigm*, a concept coined by Hockett [4]. A paradigm is a collection of words inflected in the same manner, and is typically represented with an inflection table. The way we describe our morphology is to assign paradigm identifiers to the citation forms of our

lexicon, which are translated into full inflection table by the *inflection engine* implemented in FM.

As an example, consider the citation form *fisker*, which we here assign the paradigm identifier *nn_m_fisker*. The paradigm identifier does not 'mean' anything, it could just as well be a number, but here we chose a mnemonic encoding, which can be read as: it is a masculine noun inflected in the same way as 'fisker' (which is trivially true in this case). If we feed the paradigm name and the citation form into the inflection engine, it generates the information below. To keep the presentation compact, we have contracted some word forms, i.e. the parenthesised letters are optional.

nn_m_fisker fisker ⇒

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

The starting point of the paradigmatic specification, besides the dictionaries themselves, are A. Noreen [6], E. Wessén [13, 14, 12], and G. Pettersson [7]. The paradigm description has been done by R. Johnson. Here is a table showing the number of paradigms in the current description, in their respective part of speech.

Word Class	Paradigm Number
Noun (indefinite and definite)	30
Adjective (strong and weak declension, comparison)	6
Numeral	7
Pronoun	15
Adverb	3
Verb	6

4 Ideas on dealing with the variation

As the reader may already have noticed in the presentation of the *fisker* paradigm, some variation occurring in the suffixes has already been added. The stem, how-

ever, must be treated in a different manner, since we do not have the same direct access to it as to the suffixes.

The first approach will be the use of edit distance, i.e. if an input word form is supplied, then the word forms in the lexicon (and their corresponding dictionary entries) that are close to the input word form (in terms of editing operations) should be returned. This can be done efficiently with an *universal Levenshtein automata*, see, for example, Mihov [5].

However, this approach may turn out to be too naive, i.e. generate to many false positives. Another approach would be to use a set of rules, which translates the input word form into a more idealized form.

Our initial feeling is that the problem requires a hybrid solution, where we investigate what kind of false positives the edit distance gives, and try to write rules that remedy the situation.

5 Implementation

We will now give an example on how a paradigm is defined in FM, here with some verb paradigms of Old Swedish. The presentation will be brief, and many details left out. The main objective is to provide a sense of what is involved in the paradigm definition of FM, and the interested reader is referred to one of the papers on FM.

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.

An inflection table is represented as a finite function. The intuition is that since the function is finite, then we can enumerate all its arguments and by that, create an inflection table.

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

The type system defines the inflectional parameters of Old Swedish verbs. By the `Param` instance, we ensure that the parameters is enumerable, and by `Dict` instance, we accomplish a connection between the inflection functions and the generic dictionary.

```
... definition of Modus, Number, Vox, Person, Person12

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

```
instance Param VerbForm where
  values = ... all forms in VerbForm

instance Dict VerbForm where
  category _ = "vb"
```

The next step is to define an inflection function. We start with the paradigm exemplified with the word *ælskar*.

`aelskar` in the function is a variable. If we supply `aelskar_rule` with a string, e.g. `"ælskar"`, then a `Verb` is created. Since we know how to enumerate the arguments, we can translate the function to an inflection table.

Note that the function is built up from a set of helper functions, e.g. `passivum`.

```
aelskar_rule :: String -> Verb
aelskar_rule aelskar p =
  case p of
   (PresSg Ind Act)   -> strings [aelskar,aelsk++"a"]
   (PresSg Ind Pass)  -> strings [aelska ++"s"]
   (ImperSg)          -> strings [aelsk++"a"]
   (ImperPl per)      -> imperative_pl      per       aelsk
   (PresPl per m v)   -> indicative_pl      (per,m,v) aelsk
   (PretInd Pl per v) -> preteritum_ind_pl  (per,v)   aelsk
   (PretConjPl per v) -> preteritum_conj_pl (per,v)  (aelsk++"a")
   (PresSg Conj v)    -> passivum v [aelsk++"i",aelsk++"e"]
   (PretInd Sg _ v)   -> passivum v [aelska++thi]
   (PretConjSg v)     -> passivum v [aelska++thi, aelska++the]
 where aelsk  = tk 2 aelskar
       aelska = tk 1 aelskar
```

We continue by defining two additional verb paradigms. What is worth noting here is that `foerir_rule` is defined in terms of `aelskar_rule`, and `liver_rule` in terms of `foerir_rule`.

```
foerir_rule :: String -> Verb
foerir_rule foerir p =
 case p of
   (PresSg Ind Act)     -> strings [foerir, foer++"i"]
   (PresSg Ind Pass)    -> strings [foer++"s"]
   (ImperSg)            -> strings [foer]
   (PretInd Pl per v)   -> preteritum_ind_pl  (per,v) foer
   (PretConjPl per v)   -> preteritum_conj_pl (per,v) foer
   (PretInd Sg _ v)     -> passivum v [foer++thi]
   (PretConjSg v)       -> passivum v [foer++thi, foer++the]
   _                    -> aelskar_rule foerir p
 where foer = tk 2 foerir

liver_rule :: String -> Verb
liver_rule liver p =
 case p of
```

```
   (PresSg Ind Act)  -> strings [liver, liv++"ir", liv++"i"]
   (PresSg Ind Pass) -> strings [lif++"s"]
   (ImperSg)         -> strings [lif]
   _                 -> foerir_rule (lif++"er") p
 where liv = tk 2 liver
       lif = v_to_f liv
```

After we defined our inflection functions, we need to create interface functions that translate dictionary forms into entries in the generic dictionary. Since we already defined an instance of `Dict`, it is done by the following, homogeneous, definition. If the current part of speech has any inherent parameters, e.g. gender, it would show up here.

```
vb_aelskar :: String -> Entry
vb_aelskar = entry . aelskar_rule

vb_foerir :: String -> Entry
vb_foerir  = entry . foerir_rule

vb_liver  :: String -> Entry
vb_liver   = entry . liver_rule
```

Now we are almost done. The last step involves assigning paradigm names to the interface functions, and to provide an example word form for each paradigm.

```
  ("vb_aelskar", ["ælskar"] , app1 vb_aelskar),
  ("vb_foerir",  ["førir"]  , app1 vb_foerir),
  ("vb_liver",   ["liver"]  , app1 vb_liver),
```

Finally, we can start developing our lexicon. Here in the file `fornsvenska.lexicon` we show one entry: `vb_liver liver`.

```
.. fil: fornsvenska.lexicon
...
 vb_liver liver
...
```

6 Final comment

The paradigm system has been defined together with a lexicon of 3k entries. The next step, besides finding an appropriate method to deal with the variation, is to create a prototype system that connects the morphological component with the dictionaries, and to create a facility to input real text.

We are considering a solution where the real text is a HTML page that is analyzed by the system, and the result is a new HTML page, where the word forms becomes hyperlinked to a list of lemma candidates. These candidates are, in turn, hyperlinked to the dictionaries.

References

[1] K. R. Beesley and L. Karttunen. *Finite State Morphology*. CSLI Publications, Stanford University, United States,, 2003.

[2] M. Forsberg and A. Ranta. Functional Morphology. *Proceedings of the Ninth ACM SIGPLAN International Conference of Functional Programming, Snowbird, Utah*, pages 213–223, 2004.

[3] M. Forsberg and A. Ranta. Functional morphology. `http://www.cs.chalmers.se/~markus/FM`, 2007.

[4] C. F. Hockett. Two models of grammatical description. *Word*, 10:210–234, 1954.

[5] S. Mihov and K. Schulz. Fast approximate search in large dictionaries. *Comput. Linguist.*, 30(4):451–477, 2004.

[6] A. Noreen. *Altschwedische Grammatik*. Halle: Max Niemeyer, 1904.

[7] G. Pettersson. *Svenska språket under sjuhundra år*. Lund, Sweden, 2005.

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

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

[10] K. Söderwall. *Ordbok Öfver svenska medeltids-språket. Vol I-III*. Lund, Sweden, 1884-1918.

[11] K. Söderwall. *Ordbok Öfver svenska medeltids-språket. Supplement. Vol IV—V*. Lund, Sweden, 1953—1973.

[12] E. Wessén. *Svensk språkhistoria: Grundlinjer till en historisk syntax*. Stockholm, Sweden, 1965.

[13] E. Wessén. *Svensk språkhistoria: Ljudlära och ordböjningslöra*. Stockholm, Sweden, 1969.

[14] E. Wessén. *Svensk språkhistoria: Ordböjningslöra*. Stockholm, Sweden, 1971.