Using Finite State Transducers for Making Efficient Reading Comprehension Dictionaries

Ryan Johnson, Lene Antonsen, Trond Trosterud

University of Tromsø, Norway

ryan.txanson@gmail.com, lene.antonsen@uit.no, trond.trosterud@uit.no

ABSTRACT

This article presents a novel way of combining finite-state transducers (FSTs) with electronic dictionaries, thereby creating efficient reading comprehension dictionaries. We compare a North Saami - Norwegian and a South Saami - Norwegian dictionary, both enriched with an FST, with existing, available dictionaries containing pre-generated paradigms, and show the advantages of our approach. Being more flexible, the FSTs may also adjust the dictionary to different contexts. The finite state transducer analyses the word to be looked up, and the dictionary itself conducts the actual lookup. The FST part is crucial for morphology-rich languages, where as little as 10 % of the wordforms in running text actually consists of lemma forms. If a compound or derived word, or a word with an enclitic particle is not found in the dictionary, the FST will give the stems and derivation affixes of the wordform, and each of the stems will be given a separate translation. In this way, the coverage of the FST-dictionary will be far larger than an ordinary dictionary of the same size.

KEYWORDS: Lexicography, Computational Morphology, Orthographic Variation, Finite-state Transducers, Electronic Dictionaries.

1 Introduction

The article presents work on enriching bilingual dictionaries with existing finite state transducers (FST). There is a need for reading comprehension dictionaries of morphology-rich languages like the Saami languages. They are also members of the North European Sprachbund, and contain dynamic compounds.

All of the Saami languages are minority languages. An effective digital dictionary is a necessity for language learners, but it is also important for Saami speakers: when they write in Saami, they are often met with complaints from those who do not understand the language well enough, and are asked to write in the majority language instead. This repeatedly appears in discussions in the Facebook group of one of the Saami organisations in Norway (Facebook-group, 2012).

The paper is structured as follows: in Section 2, we look at why we need dictionaries with morphology and dynamic compounding for highly inflected languages. Section 3 describes how we produce dictionaries by combining lexical resources, finite-state transducers, and open-source software. In Section 4, we evaluate the North Saami and South Saami FST-dictionaries and compare their performance with wordform dictionaries. In Section 5, we look at more possibilities that this gives us, by adjusting the FST for special needs. Our conclusion is presented in Section 6.

2 Saami Lexicography and its Challenges

Saami lexicography boasts a history spanning over 250 years, and contains several multi-volume works of high quality (for overview and discussion, see (Larsson, 1997) and (Magga, 2012)). During the first 200 years, the goal in Saami lexicography was to present the Saami languages to scientific audiences. In recent decades, an increasing use of Saami in writing has posed new challenges for Saami lexicography. For a discussion on how to present Saami in dictionaries with Saami as a target language, see (Trosterud, 2000); and for a discussion on the source language vocabulary behind Norwegian - Saami dictionaries, see (Trosterud and Eskonsipo, 2012).

Electronic North Saami dictionaries are a fairly recent phenomenon. Apart from the work presented here, there are two such dictionaries available offering lemma pairs: a small (5500 lemma) North Saami <-> English dictionary, by Renato B. Figueiredo[1]; and in 2013 a larger dictionary between Norwegian and North Saami was released from the publisher Davvi girji[2], presenting the lemma pairs from their large paper dictionaries. Giellatekno, at the University of Tromsø, has published wordform dictionaries for North and South Saami[3]. These are discussed in Section 4 below.

The idea of using FSTs for making passive dictionaries is not new. One version similar to ours is (Maxwell and Poser, 2004). It elaborates upon the idea of unifying dictionaries with FSTs, but does not cite any actual implemetation. A simpler non-FST approach would be to create a static list of every inflected form, combined with their respective lemmas, or articles in the dictionary which these forms relate to. Commercial products rarely present their methodology,

[1] `http://www.freelang.net/online/sami.php`

[2] `http://533.davvi.no/ordbok_samnor.php?lang=sam`

[3] Vuosttaš Digisánit for North Saami in 2008 (`http://giellatekno.uit.no/words/dicts/index.eng.html`) and Voestes Digibaakoeh (`http://giellatekno.uit.no/words/dicts/index-sma.eng.html` for South Saami in 2010.

but this is probably the method underlying dictionaries for morphology-poor languages like English and German.

The Saami languages are morphologically complex, suffixing languages with much non-concatenative morphology. Most lemmas have large inflection paradigms: North Saami verbs have about 55 different wordforms, with more than 40 of these being finite forms. Nouns have about 80 different wordforms, 70 of which include possessive suffixes. Adjectives have about 30 wordforms. In addition, variants of these forms exist within the normative orthography. Many of the word forms are transparent, but others are not. For example, the North Saami wordforms *gullái* and *bođii* are the first person singular past tense form of the verbs *gullát* 'to start to wake up' and *boahtit* 'to come', the latter of which has a less obvious basic form, due to diphthong simplification and consonant gradation, frequent phonological processes in the language.

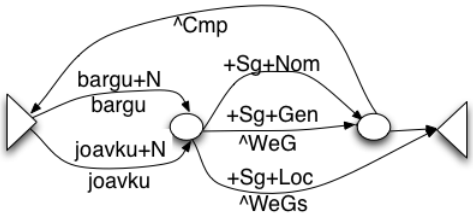
The Saami languages also rely heavily on derivation. Lexical aspects like continuative and inchoative, the causative, and even the passive voice are the result of derivation. North Saami compounding causes vowel changes in the stem vowel, so it is not possible to split a compound mechanically into two parts. For example, *bargojoavku* 'work team' is a compound of the nouns *bargu* 'work' and *joavku* 'team', but a noun like **bargo* does not exist. This is also very common: in a corpus of 1.1 million words, 26.7 % are noun compounds, comprising 7.0 % of all of the words (Antonsen et al., 2009). In addition, the Saami languages have clitic particles. In North Saami for instance, there is a question enclitic and 11 focus enclitics which can be written as a part of almost all words.

The conclusion is that we need comprehension dictionaries with inflectional word forms, as well as dynamic derivation, compounding and enclitisation. Used as an electronic dictionary, a dictionary containing only lemmas is simply not sufficient, particularly because only 7.9 % of the word forms in North Saami running text are identical to the lemma form, as in Table 2. Another possibility would be to use *stemming*, but as shown in (Antonsen and Trosterud, 2010), a stemmer containing all North Saami inflectional suffixes still assigns the wrong stem to 31 % of a large test corpus. The reason for this poor result is wide range of non-concatenative morphological processes in North Saami: consonant gradation, and root and stem vowel changes require an approach beyond concatenative affixation.

	North Saami	Finnish	Norwegian
Wordforms in test material	252,461	45,144	64,994
Lemmata in automaton	99,071	94,111	38,983
Coverage	7.9 %	10.0 %	30.5 %

Table 1: Coverage of dictionary without a morphological component, (Antonsen et al., 2009).

Our previous approach for a morphological dictionary was presented in (Antonsen et al., 2009) and consists of lemmas and wordforms for each lemma. The last version for North Saami was compiled in 2012 and contains 252,787 wordforms referring to 9,999 lemma articles, and the South Saami dictionary contains 180,352 wordforms referring to 10,657 lemma articles. The wordforms are generated ahead of time with the same FST as is used in the implementation in this article. In Section 4, we will compare the performance of this wordform dictionary to the FST dictionary described in 3.




```
rg -> rgg || _ Vow* ^WeG ;
vk -> vkk || _ Vow* ^WeG ;
u  -> o   || _   ^Cmp ;
```

triggers:
^WeG = Weak grade
^Cmp = Compound

Figure 1: The finite-state transducer produces the word forms *bargu, barggu, barggus* ('work.N') and *joavku, joavkku, joavkkus* ('team.N') and it maps between the wordform and the grammatical word. There is a compound path for both the genitive and the nominative form, so the FST can produce e.g. *boargojoavku, joavkobargu* ('work team, team work') (cfr. Section 2). Changes in the stem are carried out simultaneously in another transducer `twolc` by means of triggers.

3 Morphologically Sensitive Dictionaries

3.1 Building an FST-dictionary on Existing Resources

Our FST-dictionary, which we call *Neahttadigisánit* (`http://sanit.oahpa.no`) , takes existing lexical resources and morphological resources and combines them to produce a dictionary which is able to look up a larger set of possible inputs, by analysing wordforms and finding lemmas, but also by splitting up compound words to provide either lemmas that combine to make the compound, or either the whole compound word if a translation is available. The application also runs without need for relational databases, as linguistic resources are all contained within static files and external command-line tools.

In this system, we use bilingual lexicons for different language pairs, which we will evaluate later in this article. Currently, the dictionaries are built on a North Saami dictlexicon containing 9,999 lemmas, and a South Saami dictlexicon containing 10,657 lemmas. The lexicons are stored in XML and combined with existing FSTs for these two languages.

Both the North Saami and South Saami FSTs consist of lexical transducers written in `lexc` with respectively 110,000 and 85,000 entries, and phonological transducers implemented in `twolc` for the suprasegmental processes (Koskenniemi, 1983) and (Moshagen et al., 2004). They may be compiled with both the Xerox (Beesley and Karttunen, 2003) and HFST (Lindén et al., 2009) compilers, and are available as open-source [4] under the terms of the GNU General Public License.

Instead of compiling what is essentially a list of all word forms in a language, the FST approach involves listing the stems and affixes separately, and combining them to individual word forms by means of finite-state automata. A finite-state transducer is a finite-state automaton that maps between two strings of characters: the word form itself and the grammatical word, as in Figure 1: *girjin* (the lower level) and *girji+N+Ess* (the upper level). An FST can run both ways: giving the grammatical word from a wordform, or generating a wordform from the grammatical word.

Using an FST to generate paradigms for the dictionary demands some adaptations to make it

[4] `https://victorio.uit.no/langtech/trunk/gt/`

possible to generate the correct paradigm for each lemma, because there are certain considerations that are not present from the perspective of a morphological analyser, which may simply accept any and all input in a descriptive manner. Instead, generation requires attention to lexicalisation of certain word forms as their own individual words. For example, some lemmas have different meanings in singular or plural, such as *gaskabeaivi* 'midday (sg.)', and *gaskabeaivvit* 'dinner (pl.)'. Some lemmas also have a homonymous basic form, but have different paradigms and different translations. Some extra tags in the FST are used to keep these apart.

The FSTs already build the backbone to pedagogical systems for new beginners and heritage speakers, spellchecking and grammar checking systems for text proofing, and machine translation systems. As such, some modifications are also required in order to mark certain words or word forms for inclusion or exclusion from individual systems, to control for normative outputs, or to compensate for specific kinds of input from second-language learners.

3.2 Software

3.2.1 User Interface

Thanks to the open-source community, there are numerous resources available which make it easy to produce designs with good cross-browser compatibility. Previously, troubleshooting these issues for each individual browser would take time, when one would rather focus on implementation and basically, producing usable software.

In this case, we used Twitter Bootstrap[5] to get the most for less, and it has resulted in an easy to use and very minimal layout, see Figure 2. The layout works simultaneously on all the major browsers for desktop operating systems, as well as the most popular mobile browsers, see Figure 2. Thus, there is no real need to produce code specific to Apple's iOS or the Android operating system, or pay for the licensing setup involved with iOS development, and we get all of these things for free.

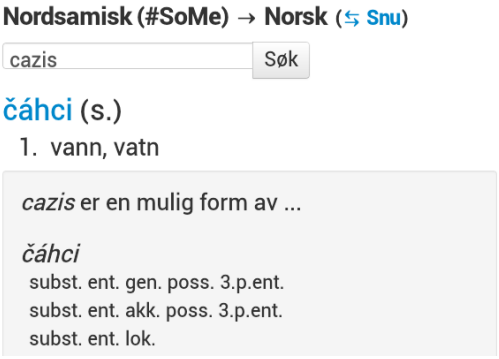


Figure 2: The FST-dictionary on the mobile.

3.2.2 Server Architecture

Having to not worry about the design meant that there was more time left for developing functionality. Our dictionary is based on Flask[6], a light, and flexible web framework for Python.

[5] `http://twitter.github.com/bootstrap/`
[6] `http://flask.pocoo.org/`

As mentioned above, the lexical data used in this application is stored in an XML format, with one file per language pair, per direction (thus making separate files for language 1 to language 2, and language 2 to language 1). These files range in size from 2MB to 5MB, and are used in the live site, without the need for a relational database to store the data. On our server, queries end up being quite fast, but to ensure that this continues to be true for larger dictionaries, we have also used one of the quickest XML libraries for Python currently available, *lxml*[7] benchmarks[8] This allows us to simply update the files, and restart the service, and any new lexical entries are immediately available to users.

Our previous wordform dictionaries demanded installation in two steps: installing a separate dictionary program (StarDict for Windows and Linux; and the preinstalled Dictionary.app for Mac OS X), and then downloading and installing the linguistic files in the dictionary program. New and updated dictionary versions demanded new downloading and installing. Our new, web-based approach naturally avoids all of this, as users only require the URL. The web dictionary may also be updated by the providers at any time, without the need for users to be aware of and perform the updates themselves.

Compared to our web dictionary, the wordform dictionaries had one major advantage: they could be used to click on words in any application running within the operating system in order to get an analysis and definition, whereas the similar functionality provided in the web dictionary only works on web pages within the browser. However, newer versions of Mac OSX have lost a user-friendly means of installing additional dictionaries to the preinstalled dictionary application, as such, this has become a point in favor of web-based solutions.

There is also an advantage for the providers of the dictionary, programmers and linguists alike. With the previous wordform dictionaries, new versions of the software (such as with StarDict), required adjustments in the format of the dictionary files, and we would often find ourselves concerned over whether we should add more linguistic content, or aim for smaller file sizes. As such, running the dictionary on a server with already existing lexicons is a big step forward.

Having a server-based system also allows us to pay attention to actual usage of the systems. As such, we log all incoming queries along with their results, in order to detect areas where the dictionary needs expansion, and these updates are then available to users as they are made.

3.3 Dictionary API

In addition to being searchable via a form in the web interface, we provide detailed lexical entries in an easily linkable HTML format, and in a more bare-bones format, JSON (JavaScript Object Notation). JSON is a widely adopted, and open standard for communication between applications, specifically with a focus on web applications. The intent here is that data is provided not just for our web-based dictionary via the interface that we provide, but that it may also be used within external applications, on other websites, and even potentially in mobile services.

The data is exposed in a couple of public-facing API endpoints or URL paths, more or less following REST (Representational State Transfer) architecture. One of the endpoints, which provides detailed word entries with inflectional paradigms has already been included in MultiDict's

[7] `http://lxml.de/`
[8] `http://lxml.de/performance.html`

Wordlink[9], a reading comprehension tool that includes many other languages and dictionaries. WordLink is quite nice, but naturally, we had some of our own designs for how to use this API.

3.4 Example Applications

3.4.1 Wordpress Plugin and Cross-browser Bookmarklet

Two of the learning tools already constructed for North Saami are *Kursa* and *Oahpa*. *Kursa* is a free, multimedia-rich set of online course materials in North Saami, containing lessons with text, and audio recordings, which are implemented in WordPress[10], a free and open-source blogging tool. There is another version on the way for South Saami.

To go with these learning materials, we have created a plugin for WordPress written in JavaScript, jQuery, and Twitter Bootstrap, which provides access to lemmatisation, compound analysis and lexicon lookup. Users simply `Alt/Opt+Double Click` a word, and it is highlighted with a text-bubble appearing below that contains word translations and wordform analysis 2. Users can quickly and easily look up as many words as they need to comprehend a text, which erases one of the barriers to reading in a new language, namely: the need to frequently look up words in a dictionary, while being unacquainted with potential "dictionary" word forms.

The modular nature of the core library within the plugin allows it to be inserted into several other potential situations with ease. For example, it could be included on a specific page or website, or inserted via a web browser plugin in any page. We have ensured that the library works in the most commonly used, and current web browsers, as such, this functionality is available on Windows, Mac OS X and Linux; in Internet Explorer, Firefox, Chrome, Opera, and Safari.

In addition to plugin for *Kursa*, we have produced a cross-browser solution which is similar to a browser extension, but instead, is accessible via a *bookmarklet*, which is a bookmark providing functionality, instead of a link to a website. As it turns out, this option has been much more preferable to developing (and also convincing users to install) browser specific plugins, and "installation" is simply a drag-and-drop affair. Thus, when on a page they wish to read, users may simply click the bookmarklet, which downloads and includes the plugin source in the HTML document structure facing the user. Now the world of news, blogs, or even Facebook, is accessible in all of the language pairs that we support.

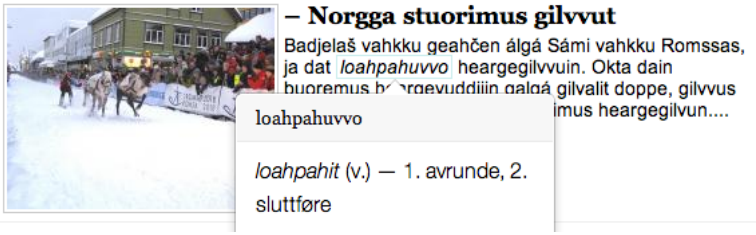


Figure 3: Reading a text with the FST-dictionary on the Saami news website site *Avvir.no*.

[9] `http://multidict.net/multidict/`
[10] `http://www.wordpress.org/`

4 Evaluation

The evaluation was run against two separate corpora, one North Saami corpus, containing 44,256 words, and one South Saami corpus, containing 60,037 words (excluding Arabic numerals, punctuation, web-addresses, and proper nouns). Since the task was to evaluate web dictionaries, the test corpora contain a balanced selection of the main text genres found online: web sites of official institutions, news texts, and political blogs. Note that Saami text on the Internet is poorly proofread, and 4.4% of the North Saami corpus (Antonsen, 2013) are words not written according to normative orthography. A part of them are misspellings, a part are subforms which are included in the FST. This should be taken into account when evaluating the data in Table 4.

The corpora were tested with both the FST-dictionary and the wordform dictionary in order to compare performance. The wordform dictionary can only recognise whole words based on sets of pre-generated word forms; while the FST-dictionary also recognises individual parts of compound words, so these translations are categorised as partial translations in Table 4.

	Translation		Partly transl.		No transl.		100%
FST-dict: all wds	39,920	90.2%	549	1.2%	3787	8.6%	44,256
wf-dict: all wds	36,231	81.9%	–	–	8025	18.1%	44,256
FST-dict: unique wds	10,862	79.7%	448	3.3%	2322	17.0%	13,632
wf-dict: unique wds	7874	57.8%	–	–	5758	42.2%	13,632

Table 2: Coverage of an North Saami FST-dictionary compared to a wordform-dictionary.

In Table 4, the coverage of the FST-dictionary is as high as 91.4% (90.2% + 1.2%) for all words in the running text. The numbers for unique words in the corpus give a more realistic picture of the usefulness of the dictionary for a language learner. The FST-dictionary leaves 17% of the unique words without any translation, while the word-form dictionary is unable to translate three times as many. If we look closer at the words that are not translated by the wordform dictionary, there are derivations such as *bivnnutvuođa*, which receives the analysis *bivnnut+A+Der/vuohta+N+Sg+Acc*. The FST-dictionary finds the lemma *bivnnut* with 'popular', while the whole word together means 'popularity'. There are also compounds which get all parts translated by the FST-dictionary, such as *sáhpánjagiid sáhpán+N+SgNomCmp+Cmp#jahki+N+Pl+Gen*, translations: *sáhpán* 'mouse', *jahki* 'year'. And, there are compounds with partial translations, such as *divamávssu diva+N+SgNomCmp+Cmp#máksu+N+Sg+Acc*. Here, only one of the lemmas is translated: *máksu* 'payment, signification'.

The FST-dictionary also manages to translate words with an enclitic, like *oidnoge oidnot+V+Ind+Prs+ConNeg+Foc/ge*. The lemma is translated: *oidnot* 'to show', *ge* is an enclitic. And the FST-dictionary is more tolerant to non-normative forms, such as *eandalit*, which is recognised by the FST as *eandalii+Adv* 'absolutely'. Among the unique 2,322 words for which the FST-dictionary does not give any translation to at all, 47.6% are unknown for the FST, mostly misspellings, or even Norwegian or Finnish quotes. 52.4% are missing in the dictionary.

In Table 4, a similar evaluation was performed with a South Saami corpus and dictionaries. The South Saami wordform dictionary leaves more words without translation than the North Saami wordform dictionary, which correlates to the fact that the South Saami dictionary has a smaller

	Translation		Partly transl.		No transl.		100 %
FST-dict: all wds	53,295	88.8 %	475	0.8 %	6266	10.4 %	60,037
wf-dict: all wds	44,989	74.7 %	–	–	15,268	25.3 %	60,257
FST-dict: unique wds	8039	67.0 %	308	2.6 %	3660	30.5 %	12,008
wf-dict: unique wds	4945	41.1 %	–	–	7085	58.9 %	12,030

Table 3: Coverage of a South Saami FST-dictionary compared to a wordform-dictionary. Data for partly translation is relevant only for the FST dictionary, the wordform dictionary is not able to handle compounds.

amount of wordforms (180,352 vs. 252,787). The FST-dictionary leaves as much as 30.5 % of the unique words without translation. Of these 3,660 words, 78.4 % are lacking in the FST. This is due to the fact that the South Saami FST is not as good as the North Saami ones, and also that there are even more misspellings in South Saami than in North Saami. Note that even though the amount of lemmas in the South Saami dictionary is slightly higher than for North Saami, they are not as relevant for the corpus as the lemmas in the North Saami dictionary are for the corpus.

South Saami also has some additional orthographic challenges not present in North Saami: namely, the orthography lacks a single norm that is used in the whole of the South Saami speaking region. Generally, one can expect certain sounds to be represented with characters that are found in each region's majority language, such as Swedish *ä* and *ö*, compared to Norwegian *æ* and *ø*, however these are often found mixed in single texts; thus the FSTs take into account certain "spelling relaxations" to handle this. In addition, there is a character, *ï*, which does not exist in either Norwegian or Swedish, and is not obligatory to write since its distribution is phonologically predictable.

5 More FST possibilites

5.1 Modularising the North Saami FST-dictionary

Using finite state transducers makes it possible to make more flexible dictionaries. Many text genres are written outside the written norm, or using input devices with keyboard restrictions. A case in point is the North Saami Facebook group *Ártegis ságat*[11]. The group is open, and has over 1800 registered members, a number equivalent to approximately 10 % of the whole speaker community. Many Facebook users read and write using smartphones. Smartphones come without preinstalled Saami keyboards, and for iPhone and Nokia smartphones there also are no such keyboards available at all. An investigation of the discussion during a three months' period, measuring the 30 most frequent words containing North Saami characters, revealed that almost 20 % of the text was written without a Saami keyboard. In a North Saami running text, 38 % of the words contain North Saami characters, and are thus the remainder are outside of the reach of electronic dictionaries.

Table 5.1 shows the performance of the dictionary combined with the ordinary FST and with a Facebook-FST with spell relax, thus allowing it to accept letters without diacritics as North Saami letters (e.g. the characters *acdsz* are accepted as representatives of *áčđšž*). The corpus

[11] `https://www.facebook.com/groups/336300756303/?fref=ts`

is taken from *Ártegis ságat* during the period 24.10.2012 – 21.01.2013 and contains 67,265 words (excluding Arabic numerals, punctuation, web-addresses and proper nouns).

	Translation		Partly transl.		No transl.		100 %
Fb-FST: all wds	54,197	80.6 %	315	0.5 %	12,753	19.0 %	67,265
ord. FST: all wds	50,263	76.3 %	250	0.4 %	15,364	23.3 %	65,877
Fb-FST: unique wds	10,596	59.8 %	286	1.6 %	6825	38.5 %	17,707
ord. FST: unique wds	8813	50.8 %	224	1.3 %	8326	48.0 %	17,363

Table 4: Coverage of a North Saami FST-dictionary. The corpus is analysed with the ordinary FST and an FST adapted to the Facebook orthography.

The overall results shown in Table 5.1 are worse than the results given in Table 4 above, but this is due to the nature of the Facebook corpus, containing a higher amount of orthographic errors and non-Saami text than is found in published text. The important point when reading Table 5.1 is the difference between the ordinary and the adjusted dictionaries. For unique words, the adapted FST recognises 59.8 % of the words, as opposed to 50.8 % with the ordinary FST.

Finite state transducers are flexible tools, not only for analysing wordforms, but also for changing their shapes. In this way, we are able to adjust the dictionary to different types of input. In an ordinary dictionary, the preference will still be to show the lexicalised words, while filtering out other analyses, however in a student dictionary it would be useful to show all the potential analyses. A Facebook-oriented dictionary will be like the ordinary dictionary, but in addition be able to understand letters lacking the proper diacritics.

5.2 Flexible FSTs: Nynorsk and Bokmål in one

Most bilingual Norwegian-L1 dictionaries are made with Bokmål as the source language. There are lexical differences between Bokmål and Nynorsk (*ikke/ikkje* 'not', *forskjell/skilnad* 'difference'), but the main difference between Nynorsk and Bokmål is morphological. In the nominal morphology, Nynorsk masculines have *-ar, -ane*, whereas Bokmål has *-er, -ene*. Similar differences are to be found in other parts of speech. In order to deal with this, we have made a special dictionary transducer. At the outset, it was an ordinary Bokmål transducer, but we added the main lexical differences between Bokmål and Nynorsk, as well as the Nynorsk morphology. Thus, rather than adding *ikkje* to the dictionary, we made *ikke* the lemma form of *ikkje*, which is then used in lexicon lookups. Nynorsk forms like *handlingar*, *diskusjonar*, were recognised as plural forms of *handling* 'action', *diskusjon* 'discussion' on par with Bokmål *handlinger*, *diskusjoner*. For a paper dictionary user, such plural forms do not pose problems, but the electronic comprehension dictionary needs a mechanism for coping with it.

In Table 5.2, we analyse two corpora of 10,606,263 words from the Nynorsk and Bokmål Wikipedia (excluding words containing capital letters or symbols outside of the Norwegian alphabet). The reference Bokmål transducer is not of outstanding quality: it recognised only 93.36 % of the Bokmål test corpus. Analysing the Nynorsk corpus with the same transducer resulted in coverage of 79.34 % of the wordforms, but minor changes to the analyser (referred to above) to include Nynorsk forms improves the coverage to 89.62 %, cf. 5.2[12].

[12]The dictionary coverage is poorer than the FST coverage, this reflects the size of the Norwegian - North Saami dictionary, and is irrelevant to the discussion on FST flexibility.

Text		FST coverage		dictionary coverage	
Nynorsk text	Conservative Bokmål	2,191,428	79.34 %	3,504,733	66.96 %
	All Bokmål varieties	1,849,654	82.56 %	3,206,796	69.77 %
	Bokmål with Nynorsk	1,101,116	89.62 %	2,530,995	76.14 %
Bokmål text	All Bokmål varieties	703,950	93.36 %	1,644,286	84.50 %

Table 5: Coverage of an Bokmål - North Saami FST-dictionary on Nynorsk text.

5.3 Reuse of FSTs for Reading Comprehension Dictionaries

FSTs are in use in many areas apart from lexicography, such as in parsing and spellchecking. The demands on a FST when used for electronic dictionaries are very different from these other usages, however. In order to be efficient for spellchecking and parsing, the FST must have an accuracy rate very close to 100 %. If it drops to, say 95 %, there will, on average, be one error in almost every sentence. For sentence analysis and syntactic disambiguation, every error has the potential of destroying larger parts of the sentence analysis, and for a spellchecker, one false alarm in each sentence will make the spellchecker useless. For an e-dictionary, on the other hand, 95 % correct implies that 19 out of 20 words will have a relevant analysis. Since the individual errors will not destroy the overall result, the dictionary is much more tolerant to errors, and even an FST recognising two thirds of the wordforms would result in a great improvement over an e-dictionary without any FST.

There are openly available finite state transducers for all Nordic languages, for all languages taught as foreign languages in Nordic schools, for all official EU languages except for Greek, and for most of the minority languages in Europe. With an increasing amount of parallel texts online, the number of bilingual dictionary resources is increasing rapidly. Given the dictionary setup described here, the missing link is a finite state transducer, which will turn a bilingual dictionary into a useful comprehension dictionary for reading of online texts.

Although our example evaluations have shown high quality FSTs in use in comprehension dictionaries, even a low quality FST would result in a drastic improvement for morphologically complex languages, where a running text contains a relatively small percentage of word forms that are also the lemma. Adding one more noun to such an FST system where each noun may have upwards of 40 word forms would thus result in coverage of 40 additional tokens in a text, as opposed to just one word form as with a non-morphologically sensitive dictionary. Gains to text coverage by adding more frequent words to such a system would result in dramatic gains to coverage. Thus, one does not need a high quality morphology or a high quality lexicon to gain the benefits that come from connecting these two things.

5.4 FSTs for Vacillating Norms

Several literary languages have vacillating norms, for one reason or another. The reason may be purely technical, as is the case for Romanian, where the Turkic *ş* (*s* with cedilla) is used more often than the correct Romanian *ș* (*s* with comma below); or when users of minority languages in Russia use Latin letters instead of modified Cyrillic ones. Writers may also lack access to keyboard layouts for their language, as demonstrated with North Saami above, or alternatively writers may just write without need or attention to detail.

In cases where there are several dialects competing for the status of being the written standard, or where the standard is new or otherwise not yet firmly rooted in the language community, the variation within written text may be considerable. To the extent that one may predict the variation, it is a trivial matter to use an FST to reduce the varying forms to a common lemma, thus increasing the coverage of a dictionary drastically.

6 Conclusion

The present article has focused upon the role that FSTs may play in coping with linguistic variation of one type or another with the goal of building reading comprehension dictionaries. Written language occurs in many varieties and in many contexts, and FSTs provide dictionaries that are flexible enough to cope with this variation, while at the same time maintaining the integrity of the dictionary itself. This article gave one example using a North Saami Facebook corpus, but many other settings may be envisaged, and coped with in a similar way.

Here, we presented bilingual dictionaries combined with finite-state transducers for morphology-rich languages. In our evaluation, we compared our existing wordform dictionaries with this finite-state solution, and found that use of an FST improves the coverage of the dictionary tremendously, from 57.8 % to 83.3 % for North Saami, and from 41.4 % to 69.6 % for South Saami.

In addition, there are other ways of analysing the linguistic inputs. For instance, words do not occur in isolation, but in sentences, so with morphological analysis as a starting point, one might also disambiguate the word grammatically, syntactically, and lexically, in order to pick a verbal reading rather than a nominal one, and to give word-sense disambiguation based upon context. These perspectives are left to future research, but it is also worth note that given the availability of FSTs in many other highly inflected languages, these possibilities are also available for many more languages than just those presented here.

Running the dictionary on a server is also a far better user experience to providing users with a dictionary to install on their computers, and has the added benefit of providing a mobile-friendly dictionary for users on the go. The implementation has also resulted in a means for reading running texts on any website without having to look away from the text itself, with a high success rate for text coverage. Together, these solutions strike down one of the larger barriers to text comprehension for language learners, who will no longer need to spend their time looking up words in dictionaries, and can instead use it for reading more texts.

Acknowledgments

Our thanks go to Norway Open Universities for funding the project "Interactive Saami Instruction on the Internet", which this work is a part of, to Berit Merete N. Eskonsipo and Inger Ellen Márjá Eira for working on the North Saami dictionary content, and to our other collegues at Giellatekno, Divvun and Aajege for participating in the project as a whole.

References

Antonsen, L. (2013). Čállinmeattáhusaid guorran. [English summary: Tracking misspellings.]. University of Tromsø.

Antonsen, L. and Trosterud, T. (2010). Manne dihtor galgá máhttit grammatihka? [English summary: Why the computer should know its Sami grammar.]. *Sámi dieđalaš áigečála*, 1:3–28.

Antonsen, L., Trosterud, T., Gerstenberger, C.-V., and Moshagen, S. N. (2009). Ei intelligent ordbok for samisk. *LexicoNordica*, 16:271–283.

Beesley, K. R. and Karttunen, L. (2003). *Finite State Morphology*. CSLI publications in Computational Linguistics, USA.

Facebook-group (2012). Discussions in NSR – a Norwegian Saami Organisation's facebook group. `https://www.facebook.com/groups/norskesamersriksforbund/?fref=ts`. [last visited on 25/01/2013].

Koskenniemi, K. (1983). *Two-level morphology : a general computational model for word-form recognition and production*. Helsingin yliopisto, Helsinki.

Larsson, L.-G. (1997). Prästen och ordet. Ur den samiska lexikografins historia. *LexicoNordica*, 4:101–117.

Lindén, K., Silfverberg, M., and Pirinen, T. (2009). HFST tools for morphology – An Efficient Open-Source Package for Construction of Morphological Analyzers. In *Proceedings of the Workshop on Systems and Frameworks for Computational Morphology*, Zürich, Switzerland.

Magga, O. H. (2012). Lexicography and indigenous languages. In Fjeld, R. V. and Torjusen, J. M., editors, *Proceedings of the 15th EURALEX International Congress*, pages 3–18, Oslo, Norway. Department of Linguistics and Scandinavian Studies, University of Oslo.

Maxwell, M. and Poser, W. (2004). Morphological interfaces to dictionaries. In Zock, M., editor, *COLING 2004 Enhancing and using electronic dictionaries*, pages 65–68, Geneva, Switzerland. COLING.

Moshagen, S., Sammallahti, P., and Trosterud, T. (2004). Twol at work. In Arppe, A., Carlson, L., Lindén, K., Piitulainen, J., Suominen, M., Vainio, M., Westerlund, H., and Yli-Jyrä, A., editors, *Inquiries into Words, Constraints and Contexts*, pages 94–105, Stanford, CA. CSLI.

Trosterud, T. (2000). Kåven, Brita E. (red) 2000: Stor norsk-samisk ordbok [book review]. *LexicoNordica*, 8:283–306.

Trosterud, T. and Eskonsipo, B. N. (2012). A North Sami translator's mailing list seen as a key to minority language lexicography. In Fjeld, R. V. and Torjusen, J. M., editors, *Proceedings of the 15th EURALEX International Congress*, pages 250–256, Oslo, Norway. Department of Linguistics and Scandinavian Studies, University of Oslo.