Extracting inflectional class assignment in Pite Saami
Nouns, verbs and those pesky adjectives

Joshua Wilbur
Albert-Ludwigs-Universität Freiburg
Department of Scandinavian Studies
Freiburg Research Group in Saami Studies
joshua.wilbur@skandinavistik.uni-freiburg.de

2017-12-24

Abstract

The main goal of this paper is to describe to what extent the three main open word classes in Pite Saami (nouns, verbs and adjectives) can be automatically assigned to inflectional classes in language technology, specifically for a Finite State Transducer. For each of these word classes, the relevant structural features necessary for determining inflectional class membership are described. In this, a clear difference between the behavior of nouns and verbs, on the one hand, and that of adjectives, on the other hand, is ascertained. While morphophonology, as seen in the paradigmatic behavior of all three word classes, is complex and features a number of types of stem alternations, nouns and verbs are predictable, while adjectives are not. With this in mind, a basic algorithm for extracting inflectional class assignment for nouns and verbs is presented for use in a LEXC framework. In contrast to this, adjectives must be assigned to inflectional classes manually. The main TWOLC rules used to trigger morphophonological alternations are also outlined. The Pite Saami lexicographic database that forms the backbone for the LEXC stem files is managed using FileMaker Pro database software, and the workflow used to extract and update LEXC files from that database is described, focussing on the differences between nouns and verbs, and adjectives. In this, light is shed on how, on the one hand, nominal and verbal inflectional patters are highly complex yet reliably systematic, while adjective morphophonology is complex and unpredictable.

Kokkuvõte

Selle artikli peamine eesmärk on kirjeldada, mil määral saab kolme põhilist avatud sõnaklassi (substantiive, verbe ja adjektiive) pite saami keeles automaatselt flekteerida kasutades keeletehnoloogia FST-d. Artiklis kirjeldatakse iga sõnaliigi muuttüübi määramiseks vajalikke strukturaalseid omadusi ning näidatakse, et adjektiivid on substantiividest ja verbidest selgelt erinevad. Samal ajal kui kõigi kolme sõnaklassi paradigmaatilist käitumist iseloomustab kompleksne paljusid tüvevahelduse tüüpe hõlmav morfofonoloogia, saab substantiivide ja verbide muutumist ennustada, kuid adjektiivide oma mitte. Seega kirjeldatakse artiklis

Proceedings of the 4th International Workshop for Computational Linguistics for Uralic Languages (IWCLUL 2018), pages 156–170,
Helsinki, Finland, January 8–9, 2018.

substantiivide ja verbide muuttüübi määramiseks kasutatavat algoritmi, mille väljund on LEXC formaadis. Adjektiivide fleksiooniklass tuleb aga määrata käsitsi. Tuuakse välja ka peamised TWOLC reeglid, mida kasutatakse morfofonoloogilise vahelduse tekitamiseks. LEXC tüvefailide põhialuseks on pite saami keele leksikograafiline andmebaas, mida hallatakse FileMaker tarkvaraga; artiklis kirjeldatakse sellest andmebaasist LEXC failide väljavõtmise ja nende uuendamise töövoogu, keskendudes erinevustele nimisõnade ja verbide, ning adjektiivide vahel. Näidatakse, et substantiivide ja verbide fleksioonimustrid on küll komplekssed, kuid väga süstemaatilised, samas kui adjektiivide morfofonoloogia on keeruline ning raskesti ennustatav.

1 Introduction

Pite Saami is a critically endangered Saami language spoken in Swedish Lapland, mainly in and around the municipality of Arjeplog.[1] No pedagogical materials have been published for Pite Saami, although a few speakers have put together some basic teaching resources for their own use. No digital resources exist from before the turn of the millennium, and, until recently, no standard orthography existed either. However, starting in 2008, a local lexicographic project and language documentation project created a foundation for digital language technology resources for Pite Saami, as outlined below. Specifically, a group of Pite Saami language activists carried out a wordlist project titled *Insamling av pitesamiska ord* 'Collection of Pite Saami words' (Bengtsson et al. 2008-2012), which ultimately resulted in the publication of the first Pite Saami dictionary (Bengtsson et al. 2016) and orthographic standard (Wilbur, 2016b), together in a single volume (Wilbur, 2016a). As a further result of this collaboration, additional digital linguistic resources are now available for Pite Saami. A searchable version of the ever-growing wordlist is available on-line at http://saami.uni-freiburg.de/psdp/pite-lex/. Language technology tools are under development for Pite Saami in cooperation with Giellatekno at the University of Tromsø; specifically, these consist of a Finite State Transducer (FST) and a Constraint Grammar.

In mid-2016, the project *Syntactic patterns in Pite Saami: A corpus-based exploration of 130 years of variation and change*[2] began compiling a thorough corpus of both spoken and written Pite Saami texts, the oldest of which were published in the late 19th century in Halász (1893). Using the combination of the electronic lexicography resources and language technology tools mentioned above, corpus creation for this syntax project is completed as automatically as possible. This idea is presented in detail in Gerstenberger et al. (2017), but a brief overview is provided here. The corpus consists of Pite Saami texts (in both spoken and written mode) transcribed in current orthographic standard and collected in the ELAN format[3] and following the common structure stipulated by projects carried out by the Freiburg Research Group in Saami Studies.[4] A python script runs each token through the FST processor, and then automatically creates annotations for lemma, morphological categories and part of speech based on this. Simultaneously, an implementation of constraint grammar is used in order to reduce the number of and, ideally, completely rule out, ambiguities

[1] Cf. Valijärvi and Wilbur (2011) and Wilbur (2014, 1-7) for more on the sociolinguistic situation of Pite Saami.

[2] Cf. saami.uni-freiburg.de/psdp/syntax/.

[3] ELAN is free software used to annotate multimedia recordings; cf. https://tla.mpi.nl/tools/tla-tools/elan/.

[4] This common ELAN structure can be found at https://github.com/langdoc/FRechdoc/wiki/ELAN-tiers.

that occur during FST processing; however, as this project is still at an early stage in describing syntactic structures in Pite Saami, the constraint grammar implementation for Pite Saami is also only in an initial stage.

The LEXC files that provide the lexical input to FST are a sort of lexicon themselves, since they present a collection of stems assigned to inflectional classes (based on their morphophonological behavior), as well as a TWOLC file providing orthographic (phonological) rules for generating and analyzing wordforms. However, the information contained in the LEXC files for stems is mainly limited to the base form, the underlying representation and an inflectional class assignment; on the other hand, the current Pite Saami lexical database contains significantly more information for each entry, including e.g. translations, gradation patterns and stem extension patterns, just to name a few categories. Because this extensive database exists and is continually being corrected and supplemented, it is clearly preferable to extract the relevant LEXC stem files from it, rather than adding these by hand on an individual basis. The Pite Saami lexical database is managed using the database software FileMaker Pro.[5] While FileMaker Pro is hardly ideal from the point of view of likely all computational linguists,[6] FileMaker Pro databases can be highly complex, and this is currently the structure the Pite Saami data is in. Although a better, open-source solution is desired in the medium-term, for the time being, this is the tool used in the project.

2 Pite Saami morphophonology and FST

While Pite Saami language structures may be represented in the corpus using various transcription methods, every text in the corpus is at least transcribed using the current standard orthography (as presented in Wilbur (2016b)).[7] Entries in the lexicographic database also use this standard. For this reason, this is also the transcription standard which provides the character strings used for processing with FST.

As is true for all the Saami languages, paradigmatic stem alternations can be used to define inflectional classes, and these are prevalent throughout the language's inflected words.[8] In addition to outlining the LEXC rules for the main open word classes of nouns, verbs and adjectives, the main TWOLC rules intended for dealing with the Pite Saami stem alternations in both consonants and vowels (including effects of vowel harmony) will be presented briefly.

Aside from being stored and run locally, the resources presented in the following sections are hosted by Giellatekno[9] at the University of Tromsø. The source documents (LEXC and TWOLC, among others) can be accessed at https://gtsvn.uit.no/langtech/trunk/langs/sje/. Language technology tools for analyzing and generating Pite Saami wordforms can be found online at http://giellatekno.uit.no/cgi/index.sje.eng.html.

[5] http://www.filemaker.com.

[6] Cf. Wilbur (2017) for both a description and a critique of FileMaker Pro as a lexical database.

[7] Note that there is no officially recognized orthography for Pite Saami. By the fall of 2017, both the Norwegian and Swedish Saami parliaments indicated they intend to officially recognize a Pite Saami orthography in the near future. It remains to be seen to what extent any official Pite Saami orthography will adhere to the orthography presented in Wilbur (2016b) (and the accompanying website at saami.uni-freiburg.de/psdp/stavningsregler/).

[8] An initial approach like this is presented for nouns, verbs and adjectives in the relevant chapters in Wilbur (2014), but the present analysis goes beyond this, and is informed by continuing research that has taken place since that publication.

[9] http://giellatekno.uit.no.

Noun class				Example		
		Stem-final-C				
Name	Syll.	NOM.SG	all others	Lemma	NOM.PL	Gloss
N_EVEN	2			juällge	juolge	'leg/foot'
N_CONTR	2	✓		båtsoj	buhtsu	'reindeer'
N_ODD	2	✓	✓	árran	árrana	'fireplace'
N_ODD_OPEN	2		✓	biena	biednaga	'dog'
N_EVEN4	3	✓	✓	mánnodak	mánnodaga	'Monday'

Table 1: The main criteria for determining FST-Inflectional classes for Pite Saami nouns, including representative examples

	Number	
Case	SINGULAR	PLURAL
NOM	juällge	juolge
GEN	juolge	julgij
ACC	juolgev	julgijd
ILL	juallgáj	julgijd
INESS	juolgen	julgijn
ELAT	juolgest	julgijst
COM	julgijn	julgij
ABESS	juolgedak	juolgedaga
ESS	juällgen	

Table 2: An example noun paradigm, showing *juällge* 'leg/foot'.

2.1 LEXC inflectional classes for nouns

Noun lemmata are named using the nominative singular wordform, and thus entered in the database using this form. Example nouns and the information required to calculate the inflectional class for each class are presented in table 1, and an entire nominal inflectional paradigm for *juällge* 'leg/foot' is provided in table 2. The naming system for inflectional classes for nouns is based roughly on Saami linguistic tradition, and uses the terms *even* and *odd* (referring to the syllable count of the nominative plural form[10]) and *contr* (for 'contracted' stems). Subdividing nouns into inflectional classes is done based on the syllable count of the base form (under 'Syll.' in table 2), and the behavior of a stem final consonant, if there is one at all (under 'stem-final-C'). For lemmata with a stem final consonant, noun classes are further determined by whether the stem final consonant occurs in the nominal singular form and/or in the other paradigm slots. These features are sufficient to unambiguously assign any given noun lemma to the correct general inflectional class, and this is done using pattern recognition in the FileMaker Pro database (see § 3). In fact, further subclasses arising due to morphophonemic alternations (such as consonant gradation and umlaut) and variation within Pite Saami also exist, but this somewhat simplified version fully illustrates the main functionality of the algorithm.

Some recent loan words with three or more syllables and with principle stress on the penultimate syllable (such as *antänna* 'antenna' or *universitähtta* 'university') initially may appear to deviate from this system. However, by simply treating the

[10]In Pite Saami, the nominative plural form and the genitive singular forms are syncretic.

Verb class			Example		
Name	Syll.	-j-ext	Lemma	3SG.PRS	Gloss
V_EVEN	2		båhtet	båhtá	'come'
V_CONTR	2	✓	gullit	gullija	'fish'
V_ODD	3		ságastit	ságasta	'speak'

Table 3: The main criteria for determining FST-Inflectional classes for Pite Saami verbs, including representative examples

Tense/ Mood	Person	Number		
		SINGULAR	DUAL	PLURAL
IND-PRS	1st	buoldáv	bulldin	buälldep
	2nd	buoldá	buälldebehtin	buälldebehtit
	3rd	bualldá	buälldeba	bulldi
IND-PST	1st	bulldiv	buldijme	buldijme
	2nd	bulldi	buldijden	buldijde
	3rd	buldij	buldijga	bulldin
IMP	2nd	buolde	buällden	bulldit

Table 4: An example verb paradigm, showing *buälldet* 'ignite'.

penultimate and final syllables as the stem, these behave in the same fully predictable manner as native lemmata.

2.2 LEXC inflectional classes for verbs

Verb lemmata are named using the infinitive wordform, and thus entered in the database using this form. Example verbs and the information required to calculate the inflectional class for each class are presented in table 3, and an entire verbal inflectional paradigm for *buälldet* 'ignite' is provided in table 4. As with the noun classes, the naming system for inflectional classes for verbs is based roughly on Saami linguistic tradition, and uses the terms *even* and *odd* (referring to the syllable count of the infinitive form) and *contr* (for 'contracted' stems). Subdividing verbs into inflectional classes is done based on the syllable count of the infinitive form (under 'Syll.' in table 3), and whether a stem extension is present in finite forms.[11] These features are sufficient to unambiguously assign any given verb lemma to the correct general inflectional class, and this is done using pattern recognition in the FileMaker Pro database (see § 3).

Some recent loan words with three or more syllables and with principle stress on the penultimate syllable (such as *adopterit* 'adopt') initially may appear to deviate from this system. However, by simply treating the penultimate and final syllables as the stem, these behave in the same fully predictable manner as native lemmata. With this in mind, the only actual exceptions to the above classes are the copula verb (*årrot* in the infinitive, but with most inflectional forms based on an *l*- stem) and the negation verb (which lacks non-finite forms). As a result, the latter two verb lemmata are each in an inflectional class of their own.

[11]Lemmata in the V_EVEN class are further subdivided based on the vowel of the second syllable, but these supplementary details are not shown here for reasons of simplicity.

Criterion	Attributive	Predicative	
		SINGULAR	PLURAL
stem gradation	str/wk/ø	str/wk/ø	str/wk/ø
suffix	-s/ø	-s/ø	-s/ø
extension	-X/ø	-X/ø	-X/ø

Table 5: Morphophonological criteria used to determine assignment to adjective inflectional classes.

Adjective class Name	Attrributive	Predicative		Gloss
		SINGULAR	PLURAL	
A_BIVVAL	-is bivvalis	ø bivval	-a bivvala	'warm'
A_AMAS	wk amas	wk amas	str/-a abmasa	'foreign'
A_AAJDNA	ø ájdna	ø ájdna	ø ájdna	'only'
A_UNNE	-a unna	-e unne	-e unne	'little'
A_FIEROK	ø fierok	ø fierok	-a fieroga	'finished'
A_GALMAS	str galbma	wk/-s galmas	str/-sa galbmasa	'cold'

Table 6: Selection of six inflectional classes for adjectives and the morphophonological features that distinguish them, including the representative examples.

2.3 LEXC inflectional classes for adjectives

The descriptions for noun and verb inflectional classes in the previous two sections show that assignment to inflectional classes in those cases are quite straightforward. While the morphophonology is rather complex, especially for nouns, the system is very consistent, and thus predictable. Indeed, the total number of classes is reasonable and clearly limited.

On the other hand, the morphophonological behavior of Pite Saami adjectives is quite the opposite, despite the fact that the basic adjective inflectional paradigm implemented in the current Pite Saami FST only consists of three slots: attributive, predicative singular and predicative plural (as opposed to 17 slots for nouns and at least 21 for verbs).[12] As is the case with assigning inflectional classes to nouns and verbs, the morphophonemic behavior (as reflected in the orthographic representation) of a given adjective lemma throughout a paradigm is extracted; this includes stem alternations in the initial vowel slot and the consonant center (stem gradation), stem extensions, as well as the form of any discernible suffixes present. The possible values of the relevant morphophonological criteria are presented in table 5.

Table 6 provides a few examples of differing adjective paradigms in order to provide an impression for the variation in patterning. Wilbur (2014) sets forth nine adjec-

[12]Whether comparative and superlative forms are types of morphological derivation or inflection, and whether case-marked adjective forms which occur in elliptical constructions should be included in the core adjectival paradigm, are theoretical discussions well beyond the scope of the current paper.

tive inflectional classes,[13] but the current analyses is well beyond that, and currently has 28 inflectional classes, with new classes added on a fairly regular basis as new adjectives are added to the database.[14]

To provide still another impression of how inconsistent adjectives are, the following figures are provided anecdotally. At the time of writing, 99 adjective paradigms have been deduced as part of ongoing analyses of the Pite Saami lexical database,[15] and, based on this set, 28 different inflectional classes have been found.[16] This number is likely to increase since more adjective entries in the database have yet to be scrutinized, but even if it does not, the ratio of adjective inflectional classes to lemmata is significantly higher than for nouns or verbs.

While historically, a *-s*-suffix may be posited as a marker for attributive forms (Rießler, 2016, 215-228), this is clearly not a productive rule in Pite Saami.[17] Not only are there plenty of instances for attributive forms without such a suffix, but there are examples for predicative forms with such a suffix. Any attempts to link stem mutation patterns such as umlaut or consonant gradation to marking attributive forms is also fruitless, since there are numerous counter examples.

In addition, variation in adjective wordforms is also more common and seemingly random compared to noun and verb lemmata, and this makes it even more difficult to assign adjectives to a specific inflectional class, as they often can belong to more than one class. For instance, *guhkke* 'long' is the predicative form, while two different attributive forms are found, in seemingly free variation: *guhka* and *guhkes*. Currently, not enough is known about which adjectives are subject to such variation, so it is too early to decide whether such lemmata should be considered belonging to a single inflectional class with variation in its forms, or to two different lemmata, each in its own inflectional class. For the time being, the latter analysis is preferred.

Ultimately, due to the lack of correlation between the 'basic' form of any given adjective lemma (whether this is considered the predicative singular form, as is typically the case in Saami lexicography, or the attributive form) and the inflectional class it belongs to, each adjective lemma has to be assigned to an inflectional class manually. Or, at a very minimum, the attributive form and the predicative form must be paired manually, and the inflectional class extracted from these forms. This is a significant

[13]Wilbur (2014) uses the phrase "correspondence patterns" (131) in a seeming attempt to avoid the term *inflectional class* for adjectives altogether. Indeed, he claims that "there is no clear or consistent morphological relationship synchronically between attributive adjectives and the corresponding predicative adjectives" (134).

[14]Note that this seeming lack of any consistency and higher frequency in variation for adjectives compared to other open word classes is nothing specific to Pite Saami, but true for other Saami languages as well; cf. e.g., the 25 inflectional classes posited for North Saami attributive adjective forms alone in Sammallahti (1998, 71-73), the 12 inflectional classes posited for North Saami attributive adjective forms in Svonni (2009, 75-76) (who also points out that "[D]en attributiva formen … bildas (delvis) oregelmässigt" (74)), or the claim by Rießler (2016, 201) that "the system of attributive and predicative marking is highly irregular in the Saamic languages".

[15]This dataset formed the basis for the Pite Saami dictionary published as Bengtsson et al. (2016), but is continually being revised, improved and expanded. It can be accessed at http://saami.uni-freiburg.de/psdp/pite-lex/.

[16]Just as a comparison, the database contains 2437 noun lemmata in five inflectional classes, and 1642 verb lemmata in three inflectional classes.

[17]It is common in Saami linguistics to posit the predicative singular form as the base form; however, I choose not to follow this tradition because of a complete lack of any consistent evidence in the synchronic system to indicate that the attributive form can be predicted based on the predicative form. Indeed, both predicative and attributive forms should be included as lexical entries in any lexicographic data collection, as Svonni (2009, 74) points out for North Saami: "Den attributiva formen böjs alltså inte och bildas (delvis) oregelmässigt och brukar därför anges i ordböcker".

Base V		Resulting V
á	➜	ä
a	➜	i
ie, ä	➜	e
å, ua/uä, uo	➜	u

Figure 1: Sets of vowel alternations triggered by vowel harmony, with the base vowel on the left, and the resulting raised vowel on the right.

difference to the noun and verb lemmata described above, which can be reliably assigned to the correct inflectional class based on the basic form and a few supplemental pieces of information, as indicated in table 1 and table 3 above. Currently, FST inflectional categories for Pite Saami adjectives are named after the predicative singular form of one of the adjectives in each class, unlike the more generic names used for noun and verb classes.

2.4 Morphophonological processes in TWOLC

As evidenced by the inflectional classes for nouns, verbs and adjectives described above, Pite Saami features complex morphophonology (as is typical for all Saami languages). In addition to marking certain morphological categories using suffixes, almost every wordform also features non-concatenative morphology in the form of stem alternations, both in the word-initial vowel slot and in the "consonant center" (the consonant slot between the first and second vowels of a Pite Saami foot[18]). These two main morphophonological processes occur together, but operate on the initial vowel and the consonant center independently, and are referred to here as *umlaut* and *consonant gradation*; these are described here briefly. Umlaut is seen in the paradigmatic alternation of two vowel sets. In the one set, *ua* (and its allophone *uä*) alternate with *uo*, and in the other, *ä* alternates with *ie*, with the former set of each pair limited to wordforms in grade III. Consonant gradation is a similar alternation in principle, but concerns paradigmatic alternations in the segments in the consonant center. Here, a number of patterns are evident, all of which alternate at least in quantity, and sometimes in quality as well. For instance, a geminate can alternate with a singleton (*rr~r*), a preaspirated segment can alternate with its unaspirated equivalent (*hp~b*), or a tripartite consonant cluster can alternate with a bipartite cluster, thereby losing its middle member (*jbm~jm*).[19] For more details on Pite Saami morphophonology, see Wilbur (2014, 74-79).[20]

In addition, there is one phonological feature significant enough to mention here, partly because it is reflected in orthographic forms: regressive phonological assimilation. It is referred to as *vowel harmony*, and only applies within a prosodic foot. In this, a closed back vowel (*i* or *u*) in the second vowel slot triggers raising of the initial vowel. The choice of the resulting vowel depends on the base vowel affected by the harmony; the possibilities are presented in figure 1.

To analyze and generate these morphophonological alternations and the phonological rule as represented by orthographic character strings in FST, a **Two-L**evel

[18]Cf. Wilbur (2014, 25-30) for a description of word-level prosodic structures in Pite Saami.

[19]In the actual phonetic realization of this third type, the second member is always pronounced as an unreleased plosive homorganic with the third consonant in the cluster, e.g. [jp̚m] for <jbm>.

[20]Note however that the description of vowel harmony on pages 79–81 in Wilbur (2014) is correct in its essence, but does not accurately reflect the status of this phenomenon as being truly phonological.

Tag	Function
^WG	require weak grade
^G3	require grade III
^UAUML	trigger *ua* diphthong
^IJ	V2-*e* becomes *i*
^V2O2U	V2-*o* becomes *u*
^V2E2AA	V2-*e* becomes *á*
^CDEL	delete stem final consonant

Table 7: Pite Saami TWOLC morphophonological triggers and their functions.

Compiler (TWOLC) is used (Beesley and Karttunen, 2003). Triggers are defined which cause these alternations, and are assigned to morphological slots in the definition files of the various inflectional classes (the LEXC affix files). For instance, *weak grade* is represented in the input by the tag ^WG, and since specific choices for an umlaut vowel or a consonant set align with weak grade, this tag is introduced in the relevant slots in the inflectional class definitions. For example, for the nominal inflectional class N_EVEN, this means that NOM.PL, GEN, ACC, ILL.PL, INESS, ELAT, COM and ABESS slots include the tag ^WG. A list of the triggers and their functions is found in table 7. A more thorough example is provided below in § 2.5.

In summary, the Pite Saami TWOLC file contains 13 rules for implementing the various consonant gradation patters, 5 rules for umlaut, and one for vowel harmony. In addition, there are supplementary rules for deleting a final consonant, selecting the voiceless variant of a final plosive, and triggering slot-specific vowel alternations in the second vowel slot.

2.5 Implementation example

In this section, a selection of TWOLC and LEXC code examples are provided to illustrate how morphophonology is implemented in the Pite Saami FST. Specifically, the generation of an accusative plural form jävrijd for the noun lemma *jávvre* 'lake' is presented.

To begin with, the code presented in figure 2 is the entry in the LEXC noun stem file. This provides the upper and lower forms for the lemma,[21] the assignment to the nominal inflectional class N_EVEN, and an English translation (just as a reference).

```
jávvre:jávvre N_EVEN "lake" ;
```

Figure 2: Code snippet for an example noun stem entry in the LEXC lexicon

From here, the LEXC nominal affix file adds inflectional tags and suffixes to the upper and lower forms, respectively, as shown in the code presented in figure 3. Here, the continuation classes mark the form to be generated as being a noun (+N) in plural accusative (+Pl+Acc) with a suffix –jd (added in two steps). Furthermore, it is morphophonologically characterized by weak grade using the tag ^WG, and as subject to a raising of the second vowel slot's vowel using the tag ^IJ.

[21]In this example, the entry is redundant because both the upper and lower sides are identical. However, since many Pite Saami lemmata do not have identical upper and lower representations (cf. e.g. the examples in 1, 2 and 3 below), all entries contain both sides explicitly, even when redundant.

```
LEXICON N_EVEN
+N:^WG N_EVEN_WK ;

LEXICON N_EVEN_WK
:^IJ%>j N_EVEN_J ;

LEXICON N_EVEN_J
+Pl+Acc:d   ENDLEX ;
```

Figure 3: Code snippet for continuation classes adding accusative plural morphology to `N_EVEN` nouns in the LEXC lexicon

For this example, three morphophonological rules stored in the TWOLC file are relevant; these are presented here. The rule named "Consonant Gradation for xxy:xy" in figure 4 deals with the alternation in the stem's consonant center. Here, when

```
"Consonant Gradation for xxy:xy"
Cx:0 <=> Vow:+ Cx _ Cy Vow:+ Cns:* %^WG: ;
     where Cx in ( ŋ ŋ ŋ v v v v v v v )
           Cy in ( g k n d j k l r g s )
          matched ;
```

Figure 4: Code snippet for a consonant gradation rule for the pattern xxy:xy

matched pairs of characters for which the first character is doubled,[22] the second instance of the doubled character is deleted when preceded by a vowel as well as followed by a vowel, an optional consonant, and, crucially, a `^WG` tag. In the example, the antepenultimate set is present: `v` and `r` in the lemma `já`**`vvr`**`e`, so the second `r` is deleted, resulting in a consonant center `rv`.

The treatment of vowel characters requires two steps in this example. Initially, the vowel in the second vowel slot is altered from `e` to `i`[23] by the rule named "V2 E to I before j-suffixes" and presented in figure 5. This occurs when preceded by at least

```
"V2 E to I before j-suffixes"
e:i <=> Vow:+ Cns:+ _ %^IJ:0 ;
```

Figure 5: Code snippet for the raising of the second vowel triggered by certain suffixes featuring a -j-segment

one consonant and at least one vowel character, and, crucially, when followed by a `^IJ` tag. As a result, the `e` in the example *jávvre* becomes `i`.

The resulting `i` in the second vowel slot then provides the input for the general vowel harmony rule, which is called "Default VH" and is presented in figure 6. For this rule to take effect, a vowel character from a subset represented by the tag `VHtrig` and defined in the TWOLC file to consist of `[e:i | i | o:u | u ]` in a word's second vowel slot must be present.[24] This then triggers an alternation in the initial vowel

[22]Note that the set of character pairs in this rule is in fact much longer, but for reasons of space, it has been shortened significantly in figure 4.

[23]Phonologically speaking, the vowel is raised.

[24]In the rule in figure 6, the tags **`#`** and **`.#.`** require a word or compound boundary at the left edge, thus restricting the affected slot to the initial vowel position.

```
"Default VH"
Vx:Vy <=> [#|.#.] Cns:* _ Cns:+ VHtrig Cns:* Vow:* %>:0 ;
    where Vx in ( á a ä å )
          Vy in ( ä i e u )
          matched ;
```

Figure 6: Code snippet for the vowel harmony rule (regressive vowel height assimilation in the initial vowel slot with i/u in the second vowel slot)

slot as set forth in the set of matched characters in this rule. In this example, á then becomes ä.

In summary, the LEXC stem and affix files together with the TWOLC rules can be implemented in FST to correctly generate and anaylze the form jävrijd as being a noun in accusative plural for the input lemma string jávvre. Partly with the help of LEXC tags, the consonant rule outputs rv from an initial rrv, and the two vowel rules alter the vowels in the first and second vowel slot from initially á and e to ä and i, respectively, in the inflected form jävrijd.

3 Determining inflectional classes with FileMaker Pro

The Pite Saami lexical database, which is the source of the LEXC files, is currently a FileMaker Pro database. While this GUI-based software is far from the ideal choice for programmers or coders, in it essence, it successfully allows one to keep complex relational data sets.[25] The program has its own powerful but cumbersome GUI-based scripting methodology, and this is used to extrapolate inflectional class assignments for nouns, verbs and adjectives. The ability to export the database into XML format makes it possible to then use XSLT to transform the data into the desired plain-text output structure, which, in this case, is a LEXC stem file.[26]

3.1 Automatic inflectional classes for nouns and verbs

While stem alternations in word forms within inflectional classes for nouns and verbs are complex, they are also surprisingly systematic, as indicated in § 2.1 and § 2.2 above. In almost every case, membership in a specific inflectional class can be derived exclusively from the shape of the citation form. As a result, it is possible to set up algorithms which automatically assign lemmata to the correct inflectional class.

Here, the basic process is explained. While the lexical database contains more than just citation forms,[27] only entries corresponding to the citation form of a lemma (in other words, nominative singular for nouns and infinitive for verbs) are subject to evaluation to begin with. In an initial script, nouns and verbs are identified based on the value in the part-of-speech field. The syllable count of each entry is determined automatically with a script that counts vowel grapheme and grapheme combinations.

[25] Cf. Wilbur (2017, 305-307) for a discussion of advantages and, crucially, disadvantages to using FileMaker Pro as a lexical database, or indeed for any data set.

[26] Note that, while an XLT stylesheet can be applied automatically during export from FileMaker Pro, it is only possible using the outdated version 1.0 of XSLT.

[27] Because the source of the majority of the entries in the database is a group of native speakers without any training in linguistics or lexicography, a not insignificant number of entries consist of inflected forms or sometimes entire phrases.

Then, lemmata are further divided into groups based on the syllable count (groups of bisyllabic and trisyllabic lemmata).

Then, for nouns, the existence of a stem-final consonant is determined, and whether this is present in the nominative singular form (the entry itself) and/or in oblique case-number paradigm slots. This is sufficient to allot the main nominal inflectional classes (as portrayed in table 1), but sub-classes for inflectional patterns reflecting more specific morphophonological alternations (such as stem final -e- alternating with stem final -á-) are assigned based on unicode string values of stem-final segments.

For verbs the process is even more straightforward. Trisyllabic verbs are assigned to class `V_ODD`, bisyllabic verbs with a stem extension are in class `V_CONTR` and all others are in `V_EVEN`, as illustrated in table 3. Membership in the further sub-classes can be determined unambiguously based on the vowel immediately preceding the *-t* infinitive suffix.

Compounds are marked by hand in the database, and the resulting inflectional classes are determined based only on the final compound element. A compound boundary is inserted which prevents phonological rules (TWOLC) from applying before it.

With the above process in mind, a noun or verb lexical entry can clearly be assigned to the correct inflectional class. Then, a few other smaller FileMaker Pro scrips extract the appropriate form for the LEXC database files in preparation for exporting. Thus a noun entry such as *biena* 'dog', which has a bisyllabic stem and lacks a stem-final consonant, is correctly identified as belonging to `N_ODD`. The right and left sides of the LEXC entry are thus calculated in FileMaker Pro as `biena` and `biednag`. Similarly, the verb *gullit* 'fish', which has a bisyllabic stem in the infinitive but a -j-extention for the stem in a number of paradigm slots, is assigned to the `V_CONTR` class. These data and the English translation are then exported from FileMaker Pro into XML. In the export process, an XSLT style sheet is applied so that the lines in (1) and (2) are included in the appropriate LEXC stem files:

(1) `biena:biednag N_ODD "dog" ;`

(2) `gullit:gul'li V_CONTR "fish" ;`

3.2 Semi-automatic inflectional classes for adjectives

As described in detail in § 2.3, the morphophonology of Pite Saami adjectives is, on a lexeme-by-lexeme basis, equally complex with that of nouns and verbs, but the assignment of adjectives to inflectional classes is significantly less transparent. This is due not only to common variation among speakers, but also to the vast number of classes, despite the fact that adjective paradigms only have three slots in their most basic form. Partly because of this variation and mainly due the lack of any consistent correspondence of morphophonological behavior across paradigms, it is not possible to automatically assign inflectional classes to adjective lemmata based on the form of the lexeme itself.[28]

Note that it is possible to separate adjectives into a (seemingly) limited set of inflectional classes (currently 28), but, crucially, membership in a specific class cannot

[28] As mentioned above, the question as to whether the attributive form or the predicative form should be considered representative for a given adjective lemma is in fact impossible to answer in a satisfactory way (at least synchronically) due to the lack of any consistent correspondence between the two forms; this will therefore not be further addressed here.

be derived from the representative lemma form, syllable count and another data category. While the citation form of nouns and verbs is sufficient for inflectional class assignment in those cases, both the predicative singular and the attributive forms of a given adjective have their own entry in the Pite Saami lexical database (although in cases where these forms are the same, there is only one entry, and this is marked as being valid for both paradigmatic slots). Once the assignment to a specific class is set, then other wordforms (such as the predicative plural) can be derived (in a computational sense, not morphologically) using the LEXC adjective affix file. For purely practical purposes, the predicative singular form is used as the base for calculating the entry in the LEXC adjective stem file.

To deal with the lack of derivable assignability, the current solution in the FileMaker Pro lexical database consists of two steps. First, inflectional classes are defined in a related database table. Second, attributive and predicative singular forms are assigned to the correct inflectional class on an individual basis. Once inflectional classes are assigned, then the LEXC stem file can be updated automatically by exporting to XML and applying an XSLT style sheet, just as for noun and verb lemmata. Due to the unpredictable nature of adjectives, processing them requires more manual work. But ultimately, an adjective entry such as *årås* 'new', which has an attributive form *årrå* (strong grade, without a stem-final consonant), a predicative singular form *årås* (weak grade, with a stem-final consonant) and a predicative plural form *årråsa* (strong grade, with a stem-final consonant and plural-marking *-a* suffix), is classified as belonging to `A_GALMAS`, even though this is done manually. The right and left sides of the LEXC entry are thus calculated in FileMaker Pro as `galmas` and `galbma`, and the result is a line in the LEXC adjective stem file, as shown in (3).

(3) `galmas:galbma A_GALMAS "cold" ;`

While the LEXC file is created automatically, the actual assignment to the correct inflectional class is a manual process, and thus the process is semi-automatic.

4 Summary and implications

In this paper, I have outlined how inflectional classes can be determined for the main open word classes in Pite Saami (nouns, verbs and adjectives), and how class assignment can be computed using a FileMaker Pro lexical database. Noun and verb morphophonology is quite complex, but is easily predictable based on the representative lemma form, syllable count and stem-final consonant behavior. Adjectives, on the other hand, are equally complex, but not reliably assignable to a specific adjective inflectional class, and thus require manual assignment. Furthermore, I have covered the basic phonological rules as implemented in TWOLC for Pite Saami. Regardless of how inflectional class assignment occurs, LEXC files can be extracted on an ongoing basis from the lexical database in expanding and supplementing the potency of the Pite Saami FST generator and analyzer.

While the FST backbone is certainly nothing new, its successful implementation for Pite Saami is an important step towards not only recognition of Pite Saami as an official Saami language in Norway and Sweden, but also for other revitalization efforts that are based on language technology (such as spell checkers and pedagogical tools). The language technology infrastructure is not entirely complete at this stage; for instance, closed word class LEXC files need updating and expanding, and a

constraint grammar implementation for reliably avoiding unnecessary wordform ambiguities in Pite Saami texts is only in its embryonic stages. Nonetheless, the current set of tools is already being used successfully to automatically annotate tokens in the Pite Saami corpus of both spoken and written texts by adding lemma, part of speech, morphophonological categories and English glosses. While using FileMaker Pro is hardly an ideal solution, it is clearly an effective one in this particular case. Continued improvement and refinement of the Pite Saami language technology infrastructure should prove to be useful for both the language community and linguistics research in the future.

Abbreviations

ABESS	abessive case
ACC	accusative case
COM	comitative case
ELAT	elative case
ESS	essive case
FST	Finite State Transducer
GEN	genitive case
ILL	illative case
IMP	imperative mood
IND	indicative mood
INESS	inessive case
-j-ext	*-j-* stem extension
NOM	nominative case
ø	no overt marker (zero morpheme)
PL	plural
PRS	present tense
PRT	past tense (preteritum)
SG	singular
str	strong grade
syll.	syllable count
V	vowel segment
wk	weak grade

References

Kenneth R. Beesley and Lauri Karttunen. 2003. Two-level rule compiler http://web.stanford.edu/ laurik/.book2software/twolc.pdf.

Nils-Henrik Bengtsson, Marianne Eriksson, Inger Fjällås, Eva-Karin Rosenberg, Gry Helen Sivertsen, Valborg Sjaggo, Dagny Skaile, and Peter Steggo. 2008-2012. Insamling av pitesamiska ord. (Pite Saami wordlist project, unpublished).

Nils-Henrik Bengtsson, Marianne Eriksson, Inger Fjällås, Eva-Karin Rosenberg, Gry Helen Sivertsen, Valborg Sjaggo, Dagny Skaile, Peter Steggo, and Joshua Wilbur. 2016. Pitesamisk ordbok. In Wilbur (2016a), pages 13–121.

Ciprian Gerstenberger, Niko Partanen, Michael Rießler, and Joshua Wilbur. 2017. In-

stant annotations. Association for Computational Linguistics, ACL Anthology, pages 25–36.

Ignácz Halász. 1893. *Népköltési gyűjtemény*, volume 5 of *Svéd-Lapp Nyelv*. Magyar tudományos akadémia. Arjeplog.

Michael Rießler. 2016. *Adjective attribution*. Number 2 in Studies in Diversity Linguistics. Language Science Press.

Pekka Sammallahti. 1998. *The Saami languages*. Davvi girji.

Mikael Svonni. 2009. *Samisk grammatik*. Universitetet i Tromsø.

Riitta-Liisa Valijärvi and Joshua Wilbur. 2011. The past, present and future of the Pite Saami language. *Nordic Journal of Linguistics* 34:295–329.

Joshua Wilbur. 2014. *A grammar of Pite Saami*. Number 5 in Studies in Diversity Linguistics. Language Science Press.

Joshua Wilbur, editor. 2016a. *Pitesamisk ordbok samt stavningsregler*. Number 2 in Samica. Albert-Ludwigs-Universität Freiburg.

Joshua Wilbur. 2016b. Stavningsregler. In Wilbur (2016a), pages 123–197.

Joshua Wilbur. 2017. The Pite Saami lexicographic backbone. ГОУ ВО КРАГСиУ, pages 299–308.