

The Principal Parts of Finnish Nominals

Mans Hulden¹ and Miikka Silfverberg²

¹ University of Colorado

² University of British Columbia

`first.last@colorado.edu`

`msilfver@mail.ubc.ca`

Abstract. We design an FST-driven computational method to calculate the minimal number of nominal forms—the *principal parts*—one must know to be able to fully inflect a lexeme in standard Finnish. To do this, we model the nominal inflection pattern as an FST according to the KOTUS declension types. Our results show that knowing five forms always suffices to uniquely determine a nominal’s inflectional behavior, and to subsequently correctly inflect all the remaining forms. This contrasts with most sources in the literature that tend to assume seven forms are needed.

Keywords: Finnish morphology · Principal parts · Finite-State grammars.

1 Principal parts

The principal parts of a verb or noun inflection pattern has a long tradition in language learning pedagogy, particularly as regards the classical languages Latin and Greek. The notion of principal parts commonly refers to the minimal set of inflected forms of a lexeme that must be known to reliably deduce all the other possible inflections. Some authors, e.g. Finkel and Stump [3], also tie the number of principal parts in a language’s nouns or verbs to the inflectional complexity of that language’s morphology.

A typical illustration of the implicational structure of a language’s inflectional paradigms is shown in Table 1a.³ Here the morphological exponents present in each form are abstracted through letters a-o. As the example shows, knowing the inflections of W, X, and Y suffices to determine the exponent present in Z, indicating that the system consists of three principal parts. This kind of a structure is typical for languages where the morphological exponent is clearly distinguishable—for example, where it is represented by an unambiguous suffix. Given such a table of exponents, determining the minimal set of principal parts can be readily solved through various computational methods.⁴

Interestingly, principal parts in Finnish presents a more complex problem than many other languages, which is why the above approach to determining

³ This kind of a structure is referred to as a ‘plat’ in [8].

⁴ As is done in e.g. <https://www.cs.uky.edu/~raphael/linguistics/analyze.html>

principal parts will not work as is. The scheme shown in Table 1a implicitly assumes that if one witnesses an inflected form of some type, the morphological exponent can be detected from that particular form. In Latin, for example, the nouns **ager** (‘field’ nom sg.) and **pater** (‘father’ nom sg.) belong to different inflectional classes as their gen sg. forms are **agrī** and **patris**, respectively. Here, witnessing only the gen. sg. forms, the morphological exponent is obvious (**-ī** vs. **-is**). But consider witnessing a nonce Finnish word form such as **kavun** (gen. sg.) and trying to reason about the morphological exponents present. Obviously the suffix is **-n**, but in order to deduce which inflectional class the lexeme belongs to, we also need to know about another morphological exponent: has the lexeme undergone consonant gradation? That information is not always available to us by examining a single form. In the above example, we would have to contrast the given form with some non-gradating form, such as the nom. sg. (which could be **kapu** or **kavu**). In other words, a naive notion of morphological exponents is not sufficient to operationalize any automatic deduction of principal parts since some lexemes have a more subtle realization of exponence than is implicated by e.g. Table 1a.

	W	X	Y	Z
I	a	e	i	m
II	b	e	i	m
III	c	f	j	n
IV	c	g	j	n
V	d	h	k	o
VI	d	h	l	o

(a) Minimal set of principal parts: Form W is required to distinguish classes I and II, form X to distinguish classes III and IV and form Y to distinguish classes V and VI. Finally, the morphological exponents in columns W, X, and Y uniquely determine the remaining exponents in column Z and Z is, therefore, not a principal part. After Stump and Finkel [8].

	NOM	GEN	PART
I	valo (- \emptyset)	valon (-n)	valoa (-a)
IV	laatikko (- \emptyset)	laatikon (-n + CG)	laatikko (-a)
VII	ovi (- \emptyset)	oven (i~e -n)	ovea (i~e -a)

(b) Truncated ‘toy’ Finnish principal parts: the NOM and GEN suffice to predict the PART form in all cases. Crucially, NOM and PART do not suffice since we need to see both a consonant gradation (CG) form and a non-CG form to identify morphological exponents.

Table 1: Principal parts in the abstract (a) and in a toy grammar of Finnish (b).

1.1 Principal parts in Finnish

In the literature, it is commonly assumed that, for nominals, seven principal parts (‘teemamuoto’ in Finnish) are needed: the singular forms of the nominative, genitive, partitive and illative as well as the plural forms of the genitive, partitive, and illative [9].⁵ There is some variation in generalizations: Nykysuomen Sanakirja [7] lists eight forms, while the Perussanakirja lists the aforementioned seven forms [2].

1.2 Parallel forms

Complicating the exact calculation of the principal parts is that several inflectional classes allow for a number of parallel forms. For example, the KOTUS⁶ inflectional class⁷ listing features five genitive plurals in class 11, exemplified by **omena**: “**omenien omenoiden omenoitten (omenojen) (omenain)**”, with the forms given in parentheses considered archaic or rare.

In this investigation, we calculate the principal parts in two different ways. In our first calculation, we assume there is only one form in each inflectional class which we infer from the frequency of each form through Google searches. For example, in the case of **omena**, we find that by far the most commonly attested form in the genitive plural is **omenoiden**; hence we assume that only that genitive plural form is possible for the **omena**-class. In our second calculation, we include all the parallel forms given by KOTUS except the ones marked in parentheses, i.e. for **omena** we include only **omenien**, **omenoiden**, and **omenoitten** as acceptable genitive plurals in that inflectional class. As we shall see, the two methods yield different sets of principal parts.

2 A finite-state grammar for calculating Principal Parts

Reasoning about principal parts from a computational point of view requires stricter definitions and assumptions than are usually provided in the linguistic literature. A commonly seen definition such as “a set of PRINCIPAL PARTS for a paradigm P is a minimal subset of P’s members from which all of P’s other members can be deduced” [3, p. 40] is computationally inadequate insofar as it does not specify an explicit mechanism for how “other members can be deduced”. To make the mechanism explicit, we observe that the transformation between two inflected forms like the singular genitive and singular nominative in a KOTUS class does not depend on the specific word. For example, transforming **ovi** (nom. sg.) to **oven** (gen. sg.) entails deleting the stem vowel **-i** and appending another stem vowel (**-e**) and the genitive suffix **-n**. The exact same transformation applies to e.g. **koski** and all other instances of KOTUS inflection class 7. These transformations between inflected forms uniquely identify

⁵ <https://kaino.kotus.fi/visk/sisallys.php?p=63>

⁶ Kotimaisten kielten keskus / Institute for the Languages of Finland

⁷ <https://kaino.kotus.fi/sanat/nykysuomi/taivutustyypit.php>

each KOTUS class (which follows from the fact that all lexemes in a KOTUS class inflect identically). Such transformations are also very straightforward to implement with FST technology.

Our general approach is to design a finite-state transducer (FST) grammar that accepts as input any number of actual inflected forms of a lexeme together with tags (such as `NOMSG`, `PARTPL`, etc.). This FST returns (1) all the compatible inflectional classes, and (2) the logically corresponding nominative singular form of each given inflected form.⁸ Implementing all the KOTUS inflectional classes as such FST-transformations allows us to investigate which combinations of inflected forms provide sufficient information to uniquely determine the inflectional class for any lexeme. Once we can uniquely identify an inflectional class from some number of given forms, completing the missing inflected forms can be done with the same FST grammar.

Further, after calculating all the combinations of inflected forms that are sufficient for each inflectional class, we calculate the minimal set of inflected forms that can uniquely determine the inflectional class for all classes—i.e. the smallest number of combined forms that can always uniquely identify a class. This last part corresponds to the *principal parts* for all nominals.

2.1 A toy grammar

To illustrate our approach, we provide the following ‘toy grammar’ of Finnish (see also Table 1b), which only contains three inflectional classes—KOTUS’ 1 (**valo**), 4 (**laatikko**), and 7 (**ovi**)—and furthermore only inflects nominals in three case forms: the singular nominative (`NOM`), genitive (`GEN`), and partitive (`PART`). For our experiments we use the *foma* toolkit [4], which employs the Xerox formalism for regular expressions [1].⁹

```

1 def IDS    [1|4|7] ;
2 def TAGS   [GEN|PART|NOM] ;
3 def SEP    "|" :0 ;
4 def NT     [? - TAGS - IDS - "|"] ;
5 def P      SEP NT+ ;
6
7 def Gr     1:0 (P 0:NOM) (P 0:n 0:GEN) (P 0:a 0:PART) |
8           4:0 (P 0:NOM) (P k:0 o 0:n 0:GEN) (P 0:a 0:PART) |
9           7:0 (P i 0:NOM) (P i:e 0:n 0:GEN) (P i:e 0:a 0:PART);

```

With such a grammar, we can provide one or multiple inflected forms as input, and receive as output all the valid paradigm numbers together with the base forms. For example:

⁸ This choice of which form to treat as the citation form is arbitrary—we could have chosen any of the case forms. To ease grammar design we used the standard nominative singular.

⁹ Note that parentheses in the formalism denote optionality, not grouping.

```
foma[1]: up valonGEN
1|valo
4|valko
```

Here, we see that the genitive form **valon** is compatible with inflectional classes 1 and 4, and that in the case of 1, the citation form is **valo**, whereas it is **valko** if the class is 4. Class 7 is ruled out since that class defines nouns such as **ovi** which have an **e~i** alternation between the nominative singular and other singular forms; the provided example **valon** does not contain an **e** required for membership in class 7.

When several forms are given, we may produce a situation where the provided forms correspond to different citation forms:

```
foma[1]: up valonGENvaloaPART
1|valo|valo
4|valko|valo
```

Here, the forms (**valon/valoa**) are mapped to both classes 1 and 4 as above, but the mapping to class 4 corresponds to two different citation forms **valko** and **valo**, an inconsistency that we rule out in our principal parts calculation method by non-FST means. This example illustrates that the genitive and partitive singular forms are sufficient to uniquely determine class 1 nominals in our toy grammar.

Note that there is a subtlety in how we have interpreted the KOTUS class 7 and the behavior of KOTUS classes in general. We make the assumption that the alternation given through the example noun **ovi** which defines class 7 is indeed **e~i** (line 9 of code snippet):

```
7:0 (P i 0:NOM) (P i:e 0:n 0:GEN) (P i:e 0:a 0:PART)
```

We could have defined the alternation in another way, perhaps saying that the last letter—whatever it is—in the nominative singular alternates with **e**, or, conversely, that the letter preceding suffixes in the **e**-containing forms (**oven**, **ovea**) etc. alternates with **i** in the nominative singular, in both cases without specifying the relevant letter by using the **?**-wildcard in regex notation:

```
7:0 (P 0:NOM) (P ?:e 0:n 0:GEN) (P ?:e 0:a 0:PART)
```

or

```
7:0 (P ?:i 0:NOM) (P 0:n 0:GEN) (P 0:a 0:PART)
```

These are different generalizations but would both work equally well to model all forms of **ovi** which defines class 7. However, such a modification would change the calculation of principal parts since the class in question would not then be restricted to having a stem that ends exclusively in either **e** or **i**. These design

decisions, which are ultimately linguistic in nature, are of course unavoidable whenever inflectional behavior needs to be strictly formalized with an FST.¹⁰

2.2 Full Grammar

With our grammar, we can query the FST by giving various combinations of inflected forms from each class, using the KOTUS-specified example forms. Figure 1 illustrates a query against our full grammar using two forms of the lexeme **laatikko**, showing that those two forms are sufficient to rule out all inflectional classes except class 4. Following this, we can establish all combinations of forms that are sufficient for each inflectional class.

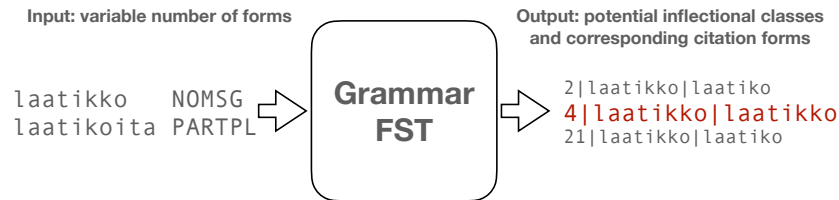


Fig. 1: Our grammar takes as input any combination of inflected forms together with tags and outputs all the compatible inflectional classes. Only those compatible classes where all given forms map to the same citation forms is an actual candidate class. In the example given, class 4 (shown in red) is uniquely determined by the NOMSG and PARTPL forms of **laatikko**. The other classes returned all presume inconsistent stem mappings and are ruled out by non-FST means.

In our full grammars,¹¹ we model the inflection of 48 of the 51 KOTUS nominal inflectional classes (classes 1 through 48). We exclude classes 50 and 51 which pertain to two-stem compounds where one or both stems inflect and also class 49 which models rare alternate stems (**askel/askele**). The grammar includes the eight forms commonly given as example forms: the nominative, genitive, partitive, and illative singulars and plurals, e.g. “ovi oven ovia oven ovet ovien ovia oviin”. As the remaining case forms¹² can readily be derived if these forms are known, we focus on finding the set of principal parts among these eight forms. The set of principal parts of Finnish is then the minimal set of inflectional forms which is sufficient to uniquely identify any word in any KOTUS inflectional class. Our grammar, being bidirectional, can of course also be used to fill in the missing forms once the inflectional class of a lexeme is known.

¹⁰ We have attempted to make reasonable judgments in these cases, consulting the Iso Suomen Kielioppi [9] as well as the more computational and FST-oriented discussion about inflectional classes in [6].

¹¹ <https://github.com/mhulden/finnishprincipalparts>

¹² The inessive, elative, illative, adessive, ablative, allative, essive, translative, instructive, abessive, and comitative.

As mentioned above, we construct both a grammar that allows for alternate forms and a single (most frequent) form grammar and calculate the principal parts separately for each. A few classes collapse when only allowing single allomorphs (e.g. only allowing **omenoiden** in the genitive plural as opposed to accepting all three listed by KOTUS: **omenien**, **omenoiden**, and **omenoitten**. These are classes 11 and 12 (**omena** and **kulkija**) and classes 23 and 24 (**tiili** and **uni**).

The FST model for each inflectional class is an elaboration upon the toy grammar code given above. For example, the class 7 (**ovi**) mapping is defined as follows in our full grammar:

```

1 def WeakCG [{kk}:k|{pp}:p|{tt}:t|k:0|p:v|t:d|{nk}:{ng}|
2           {mp}:{mm}|{lt}:{ll}|{nt}:{nn}|{rt}:{rr}|k:j|k:v};
3           # Weakening CG
4           ...
5 7:0 (P Cons i 0:NOMSG) ([P|P WeakCG] i:{en} 0:GENSG)
6     (P Cons i:{eA} 0:PARTSG) (P Cons i:{een} 0:ILLSG)
7     ([P|P WeakCG] i:{et} 0:NOMPL)
8     ([P|P WeakCG] 0:{en} 0:GENPL) (P Cons i 0:{A} 0:PARTPL)
9     (P Cons i 0:{in} 0:ILLPL)

```

The above illustrates a class that allows for members that undergo consonant gradation. Capital letters in suffixes—archiphonemes—such as A in suffixes eA and A are later rewritten to their actual harmonic realizations (a or ä), as is standard when modeling Finnish morphophonology computationally [5].

2.3 Additional restrictions

Many of the KOTUS inflectional classes have structural membership restrictions. These include (1) whether the class allows consonant gradation, as in the example above, (2) some restrictions on number of syllables that stems have in the class, and (3) various restrictions on stem endings. For example, class 2 (**palvelu**) only pertains to nominals that end in **-o**, **-u**, **-ö**, or **-y**, where the stem has minimally three syllables, and where the lexeme does not undergo consonant gradation. We encode all such structural restrictions into the FST grammar to avoid associating inflected word forms with classes where they would not fit.

The above code for inflecting **ovi**-type nominals reflects such restrictions—we require the stem to have the structure **-Ci** in certain forms (e.g. **oviin**) and **-Ce** in others (as in **oveen**).

3 Results

When including alternate forms of lexemes like the singular genitives **omenien**, **omenoiden** and **omenoitten** for the lexeme **omena**, we discover two equivalent minimal sets of principal parts each containing four forms:

PARTSG ILLSG GENSG/
NOMPL GENPL

Cl. NOMSG	GENSG	PARTSG	ILLSG	NOMPL	GENPL	PARTPL	ILLPL
1	valo	valon	valoa	valoon	valot	valojen	valoihin
2	palvelu	palvelun	palvelua	palveluun	palvelut	palveluiden	palveluihin
3	valtio	valtion	valtiota	valtioon	valtiot	valtioiden	valtioihin
4	laatikko	laatikon	laatikkoa	laatikkoon	laatikot	laatikoiden	laatikoihin
5	risti	ristin	ristiä	ristiin	ristit	ristien	ristejä
6	paperi	paperin	paperia	paperiin	paperit	papereiden	papereita
7	ovi	oven	ovea	oveen	ovet	ovien	oviin
8	nalle	nallen	nallea	nalleen	nallet	nallejen	nalleihin
9	kala	kalan	kalaa	kalaan	kalat	kalojen	kaloja
10	koira	koiran	koiraa	koiraan	koirat	koirien	koiriin
11	omena	omenan	omenaa	omenaan	omenat	omenoiden	omenoihin
12	kulkija	kulkijan	kulkijaa	kulkijaan	kulkijat	kulkijoiden	kulkijoihin
13	katiska	katiskan	katiskaa	katiskaan	katiskat	katiskojen	katiskoita
14	solakka	solakan	solakkaa	solakkaan	solakat	solakoiden	solakoihin
15	korkea	korkean	korkeaa	korkeaan	korkeat	korkeiden	korkeita
16	vanhempi	vanhemman	vanhempaa	vanhempaan	vanhemmat	vanhempien	vanhempia
17	vapaa	vapaan	vapaata	vapaaseen	vapaat	vapaiden	vapaita
18	maa	maan	maata	maahan	maat	maiden	maita
19	suo	suon	suota	suohon	suot	soiden	soita
20	filee	fileen	fileetä	fileeseen	fileet	fileiden	fileitä
21	rosé	rosén	roséta	roséhen	rosét	roséiden	roséita
22	parfait	parfait'n	parfait'ta	parfait'hen	parfait't	parfait'iden	parfait'ita
23	tiili	tiilen	tiiltä	tiileen	tiilet	tiilien	tiiliä
24	uni	unen	unta	uneen	unet	unien	unia
25	toimi	toimen	tointa	toimeen	toimet	toimien	toimia
26	pieni	pienen	pientä	pieneen	pienet	pienten	pieniä
27	käsi	käden	kättä	käteen	kädet	käsien	käsiä
28	kynsi	kynnen	kynttä	kynteen	kynnet	kynsien	kynsiä
29	lapsi	lapsen	lasta	lapseen	lapset	lasten	lapsia
30	veitsi	veitsen	veistä	veitseen	veitset	veitsien	veitsiä
31	kaksi	kahden	kahta	kahteen	kahdet	kaksien	kaksia
32	sisar	sisaren	sisarta	sisareen	sisaret	sisarten	sisaria
33	kytkin	kytkimen	kytkintä	kytkimeen	kytkimet	kytkinten	kytkimiä
34	onneton	onnettoman	onnetonta	onnettomaan	onnettomat	onnettomien	onnettomia
35	lämmin	lämpimän	lämmintä	lämpimään	lämpimät	lämpimien	lämpimiä
36	sisin	sisimmän	sisintä	sisimpään	sisimmät	sisimpien	sisimpiä
37	vasen	vasemman	vasenta	vasempaan	vasemmat	vasempien	vasempia
38	nainen	naisen	naista	naiseen	naiset	naisten	naisia
39	vastaus	vastauksen	vastausta	vastaukseen	vastaukset	vastausten	vastauksia
40	kalleus	kalleuden	kalleutta	kalleuteen	kalleudet	kalleuksien	kalleuksia
41	vieras	vieraan	vierasta	vieraaseen	vieraat	vieraiden	vieraita
42	mies	miehen	miestä	mieheen	miehet	miesten	miehiä
43	ohut	ohuen	ohutta	ohueen	ohuet	ohuiden	ohuita
44	kevät	kevään	kevättä	kevääseen	kevät	keväiden	keväitä
45	kahdeksas	kahdeksannen	kahdeksatta	kahdeksanteen	kahdeksannet	kahdeksansien	kahdeksansia
46	tuhat	tuhannen	tuhatta	tuhanteen	tuhannet	tuhansien	tuhansia
47	kuollut	kuolleen	kuollutta	kuolleeseen	kuolleet	kuolleiden	kuolleita
48	hame	hameen	hameetta	hameeseen	hameet	hameiden	hameita

Table 2: The choice of forms in our model where a single most frequent allomorph is chosen. In boldface is shown an example set of a minimal number of forms per class to uniquely inflect the class. Sometimes only a single inflected form of a lexeme suffices to completely infer its class, as in e.g. **valoa**. Note that, because of our selection of single allomorphs, classes 11&12, and 23&24 are identical.

GENPL	GENSG	ILLPL	ILLSG	NOMPL	NOMSG	PARTPL	PARTSG
1.5	8.4	1.8	4.6	8.4	6.4	1.6	2.5

Table 3: Average class ambiguity after giving a single form. These numbers pertain to the case where we can have parallel forms in the same slot.

	GENPL	GENSG	ILLPL	ILLSG	NOMPL	NOMSG	PARTPL	PARTSG
GENPL	-	1.2	1.1	1.1	1.2	1.4	1.2	1.1
GENSG	-	-	1.5	3.9	8.3	2.4	1.3	1.8
ILLPL	-	-	-	1.4	1.5	1.8	1.4	1.3
ILLSG	-	-	-	-	3.9	2.0	1.2	1.7
NOMPL	-	-	-	-	-	2.4	1.3	1.8
NOMSG	-	-	-	-	-	-	1.5	2.5
PARTPL	-	-	-	-	-	-	-	1.2
PARTSG	-	-	-	-	-	-	-	-

Table 4: Average class ambiguity remaining after giving two forms. These numbers pertain to the case where we can have parallel forms in the same slot.

GENPL	GENSG	ILLPL	ILLSG	NOMPL	NOMSG	PARTPL	PARTSG
10.9	8.1	8.0	4.6	8.0	6.4	6.7	4.6

Table 5: Average class ambiguity after giving a single form. These numbers pertain to the case where we choose a single exemplar of parallel forms in the same slot using frequency.

	GENPL	GENSG	ILLPL	ILLSG	NOMPL	NOMSG	PARTPL	PARTSG
GENPL	-	2.3	4.7	2.0	2.3	1.8	5.4	1.6
GENSG	-	-	3.0	3.8	7.9	2.3	2.6	2.2
ILLPL	-	-	-	2.7	3.0	1.9	5.1	1.7
ILLSG	-	-	-	-	3.8	2.0	2.3	2.2
NOMPL	-	-	-	-	-	2.3	2.6	2.2
NOMSG	-	-	-	-	-	-	1.6	2.6
PARTPL	-	-	-	-	-	-	-	1.5
PARTSG	-	-	-	-	-	-	-	-

Table 6: Average class ambiguity remaining after giving two forms. These numbers pertain to the case where we choose a single exemplar of parallel forms in the same slot using frequency.

The singular partitive, singular illative and plural genitive are always required to distinguish classes. In addition, either the singular genitive or the plural nominative form is required. When choosing one exemplar from each set of parallel forms based on maximal corpus frequency, we find the following minimal sets of principal parts each containing five forms:

NOMSG PARTSG ILLSG PARTPL GENPL

As mentioned above, there is no substantive distinction between classes 11 (**omena**) and 12 (**kulkija**), or 23 (**tiili**) and 24 (**uni**) when disallowing parallel forms since they define exactly the same inflectional behavior when parallel forms are ruled out. In actuality, those classes could in principle still be distinguished for *some* nouns that have a special structure. Class 11 is defined as consisting of 3-syllable stems that end in **-Ca** or **-Cä** while class 12 consists likewise of 3-syllable stems, but ones that end in **-a** or **-ä**. Both the KOTUS example lexemes **omena** and **kulkija** thus fit both classes based on their shapes. However, a word such as **anarkia** would only fit class 12 as it doesn't have a consonant before the final vowel. Similarly, the difference between structural constraints for the **uni** and **tiili** classes is that **uni**-class stems end in **-(h,n,l,n,r,s)i** while **tiili**-class stems end in **-(h,l,n)i**. Therefore, a word that only fits one of the classes (such as **kusi**) could in principle be used to separate the two classes. These are subtle corner cases that have little bearing on the actual principal parts since the classes themselves have the same inflected forms when parallel forms are not included.

Table 2 shows all the example forms in the 48 KOTUS classes where a single allomorph is allowed in each inflectional slot; it also includes an example of form combinations that suffice to deduce that a lexeme is a member of that particular inflectional class.

We conducted an additional experiment to determine how well individual forms and pairs of forms distinguish between classes. We display results separately for the case where one slot can be occupied by several parallel inflected forms, when allomorphy is present, and the case where a single example is chosen based on corpus frequency. Table 3 shows the average number of remaining possible classes after observing a single form and Table 4 gives this average ambiguity after observing two forms, when we allow for parallel forms in the same slot. Tables 5 and 6 show the same information when a single representative is selected from the set of parallel forms. In general, we can see that the plural genitive is the most informative individual form in the parallel form setting leading to only 1.5 possible classes remaining on average for each of our lexemes. However, results change drastically when filtering by frequency where the genitive plural results in 10.9 possible classes for each lexeme on average. This is understandable because the most frequent plural genitive is the form taking a suffix *-den* or *-jen* in most classes (as in **omenoiden** gen.pl. and **valojen** gen.pl.) which effectively makes this case form very uninformative overall. In general, we can see that ambiguity increases when filtering by form frequency if parallel forms are possible in the slot but remains the same, otherwise. The most restrictive pair of forms are the plural genitive and singular partitive (when we allow for parallel forms). Only 1.1 classes on average are possible after observing this pair. It is, therefore, that both of these forms are included in our set of principal parts. Again, pairs of forms are far less restrictive when we do not allow for parallel forms. Here the best disambiguation performance is achieved by the plural and singular partitive. After observing these forms 1.5, classes on average are possible.

4 Discussion

Our core finding—that only four or five principal parts are needed for identifying the inflectional class of any Finnish nominal—is somewhat surprising given that standard sources usually list 7–8 forms.¹³ We suspect that the number of principal parts could perhaps be reduced further—at least in the case of parallel forms—by tightening the phonological structural restrictions on membership in classes.

The discrepancy between the current results and earlier results in the literature may simply be a result of different assumptions about the implication structure between inflected forms; our principal parts model relies on a specific model of inflection (encoded as an FST), while the underlying assumption behind various comprehensive grammars may be different. A different choice of class membership restrictions could also affect the result—we have encoded substantial structural restrictions on allowable word stems in most of the classes, which may play a part in reducing the number of principal parts.

The fundamental advantage of calculating the principal parts using an FST is that all the assumptions about affixes and participating morphophonological processes are made explicit in the grammar, and the results are verifiable.

References

1. Beesley, K.R., Karttunen, L.: *Finite State Morphology*. CSLI Publications, Stanford, CA (2003)
2. Eronen, R.: *Taivutusmuotoja ikkunassa. Kielikello (1/1997)* (1997)
3. Finkel, R., Stump, G.: *Principal parts and morphological typology*. *Morphology* **17**(1), 39–75 (2007)
4. Hulden, M.: *Foma: a finite-state compiler and library*. In: *Proceedings of the Demonstrations Session at EACL 2009*, pp. 29–32. Association for Computational Linguistics, Athens, Greece (Apr 2009), <https://www.aclweb.org/anthology/E09-2008>
5. Koskeniemi, K.: *Two-level morphology: A general computational model for word-form recognition and production*. Publication 11, University of Helsinki, Department of General Linguistics, Helsinki (1983)
6. Pirinen, T.: *Suomen kielen äärellistilainen automaattinen morfologinen jäsenin avoimen lähdekoodin resurssein*. University of Helsinki, MA Thesis (2008)
7. Sadeniemi, M.: *Nyky-suomen sanakirja: lyhentämätön kansanpainos*. W. Söderström (1980)
8. Stump, G., Finkel, R.A.: *Morphological typology: From word to paradigm*. Cambridge University Press (2013)
9. Vilkuna, M., Korhonen, R., Kovisto, V., Heinonen, T.R., Alho, I.: *Iso Suomen Kieliooppi*. Suomalaisen Kirjallisuuden Seura (2004)

¹³ The smallest set we have seen described is Jukka Korpela’s speculation that six forms should suffice: <http://jkorpele.fi/suomi/sijataivutus.html>