Språk-
BANKEN

GÖTEBORGS UNIVERSITET

What can we learn from inflection tables?

Markus Forsberg
in collaboration with Måns Huldén and Malin Ahlberg
BAULT 2016

Today's question:
What can we (machine) learn from a set of inflection tables?

Declinatio [+/-]

f.	sing.	plur.	
nom.	**rŏsa**	rŏsae	I
gen.	rŏsae	rŏsārum	II
dat.	rŏsae	rŏsīs	III
acc.	rŏsam	rŏsās	IV
abl.	**rŏsā**	rŏsīs	VI
voc.	**rŏsa**	rŏsae	V

f.	sing.	plur.	
nom.	**mensa**	mensae	I
gen.	mensae	mensārum	II
dat.	mensae	mensīs	III
acc.	mensam	mensās	IV
abl.	**mensā**	mensīs	VI
voc.	**mensa**	mensae	V

Verbum finitum

Thema	Vox activa						
vīv-	Tempus praesens			Imperfectum		futurum	
Persona	indicativ.	coniunct.	imperat.	indicativ.	coniunct.	indicativ.	imperat.
I. sing.	vīvō	vīvam		vīvēbam	vīverem	vīvam	
II. sing.	vīvis	vīvās	vīve!	vīvēbās	vīverēs	vīvēs	vīvitō!
III. sing.	vīvit	vīvat		vīvēbat	vīveret	vīvet	vīvitō!
I. plur.	vīvimus	vīvāmus		vīvēbāmus	vīverēmus	vīvēmus	
II. plur.	vīvitis	vīvātis	vīvite!	vīvēbātis	vīverētis	vīvētis	vīvitōte!
III. plur.	vīvunt	vīvant		vīvēbant	vīverent	vīvent	vīvuntō!

Thema	Vox activa				
vīx-	Tempus perfectum		plusquam perfectum		futurum exactum
Persona	indicativ.	coniunct.	indicativ.	coniunct.	
I. sing.	vīxī	vīxerim	vīxeram	vīxissem	vīxerō
II. sing.	vīxistī	vīxeris	vīxerās	vīxissēs	vīxeris
III. sing.	vīxit	vīxerit	vīxerat	vīxisset	vīxerit
I. plur.	vīximus	vīxerimus	vīxerāmus	vīxissēmus	vīxerimus
II. plur.	vīxistis	vīxeritis	vīxerātis	vīxissētis	vīxeritis
III. plur.	vīxērunt	vīxerint	vīxerant	vīxissent	vīxerint

Verbum infinitum

Modus	Infinitivus			participium		
Tempus	praesens	perfectum	futurum	praesens	perfectum	futurum
Vox activa	vīvere	vīxisse	vīctūrum, -am, -um esse	vīvēns		vīctūrus, -a, -um

Gerundium	Gerundivum	Supinum	
vīvendī	vīvendus, -a, -um	—	—

Why this interest in inflection tables?

There is a lot of inflection tables out there:

Wiktionary

quote logo

Wiktionary
The free dictionary

Multilingual portal
Full list of languages

Wiktionary is a project to create a multilingual free content dictionary in every language. This means each project seeks to use a particular language to define all words in *all* languages. It actually aims to be much more extensive than a typical dictionary, including thesauri, rhymes, translations, audio pronunciations, etymologies, and quotations. The project started in December 2002, and as of June 2016 is available in over 170 languages with over 25,000,000 entries in all. The largest language edition is English, with 4,733,000 entries. Then Malagasy, French, Serbo-Croatian, Spanish, Chinese, Russian and Lithuanian follow. All seven of them have more than 600,000 entries each, while 41 other languages have more than 100,000 entries each. In total, 116 languages have at least 1,000 entries.

Wiktionary works in collaboration with the Wikimedia Commons. Many sound files have been uploaded to Commons to provide Wiktionary and other projects with examples of pronunciation.

Some learning possibilites we will look into

1. Derivation of inflection engines
   => ***paradigm induction***

2. Learn how to inflect unseen words
   => ***paradigm prediction***

3. Derivation of **morphological analyzers**

1. Paradigm induction

What does it mean to say that a word is inflected as another word?

- **Statement**: The German word '*Anfang'* is inflected in the same way as the word '*Frack'*.

And here you have the inflection table of Frack:

	Singular	**Plural**
Nominative	Frack	Fräcke
Genitive	Frackes, Fracks	Fräcke
Dative	Frack, Fracke	Fräcken
Accusative	Frack	Fräcke

So how do we inflect '*Anfang*', given this information?

Like this:

	Singular	**Plural**
Nominative	Anfang	Anfänge
Genitive	Anfanges, Anfangs	Anfänge
Dative	Anfang, Anfange	Anfängen
Accusative	Anfang	Anfänge

Did you guess right? Can you explain why?

If you know German, pretend that you don’t.

First some terminology

- **Paradigm function**: a function that given one (typically the baseform) or more word forms, produces the full inflection table.

f(Anfang) =

	Singular	Plural
Nominative	Anfang	Anfänge
Genitive	Anfanges, Anfangs	Anfänge
Dative	Anfang, Anfange	Anfängen
Accusative	Anfang	Anfänge

- Words inflect in the same way = they share the same paradigm function.
- **Inflection engine**: a set of paradigm functions.
- **Paradigm induction**: derivation of paradigm functions.

Paradigm Induction

	Singular	Plural
Nominative	**Frack**	**Fräck**e
Genitive	**Frack**es, **Frack**s	**Fräck**e
Dative	**Frack**, **Frack**e	**Fräck**en
Accusative	**Frack**	**Fräck**e

	Singular	Plural
Nominative	**Anfang**	**Anfäng**e
Genitive	**Anfang**es, **Anfang**s	**Anfäng**e
Dative	**Anfang**, **Anfang**e	**Anfäng**en
Accusative	**Anfang**	**Anfäng**e

Induction

$f(x_1, x_2) =$

	Singular	Plural
Nominative	x_1+a+x_2	x_1+ä+x_2+e
Genitive	x_1+a+x_2+es, x_1+a+x_2+s	x_1+ä+x_2+e
Dative	x_1+a+x_2, x_1+a+x_2+e	x_1+ä+x_2+en
Accusative	x_1+a+x_2	x_1+ä+x_2+e

The method

- **LCS** = Longest common subsequence
- **subsequence** = a string that can be obtained from another string by deleting zero or more characters from that string.
- **substrings** in the subsequence becomes **variables**. I.e, What is common in all words are the variable parts.
- The method: LCS + heuristics to resolve LCS ambiguity.

	Singular	**Plural**
Nominative	Frack	Fräcke
Genitive	Frackes, Fracks	Fräcke
Dative	Frack, Fracke	Fräcken
Accusative	Frack	Fräcke

LCS: Frck

LCS ambiguity

Competing alignments

compr*ar*, ***compr****a*, ***compr****o*

comp*ra****r***, ***compr****a*, ***compr****o*

Competing LCS

seg*e****l,*** ***segl****et****,*** ***segl****en* LCS: *segl*

sege*l*, ***seg****l****e****t*, ***seg****l****e****n* LCS: *sege*

LCS ambiguity resolution through heuristics

- **Heuristic 1**: minimize the number of variables

compr*ar*, ***compr****a*, ***compr****o*

compr*a****r***, ***compr****a*, ***compr****o*

- **Heuristic 2**: minimize the number of infix segments

seg*el*, ***segl****et*, ***segl****en* LCS: *segl*

sege*l*, ***seg****l****e****t*, ***seg****l****e****n* LCS: *sege*

- and some additional heuristics, but above is the major ones.

The paradigm function

- From a function accepting variable instantiation to word form(s)?

$$f(x_1, x_1, .., x_n) \Rightarrow f(w_1, w_1, ..., w_n)$$

- We **match** the input word(s) with **any word pattern(s)** in the paradigm function (often just the lemma with the lemma pattern). This gives us the **variable instantiations** we need to compute the forms.

- The matching may be **ambiguous**, so we need a **matching strategy. Longest match** seems to work best for suffixing languages.

match(x_1+a+x_2, ”Frack”) = {x_1=Fr, x_1=ck}

Regular expression with groups

Ambiguity

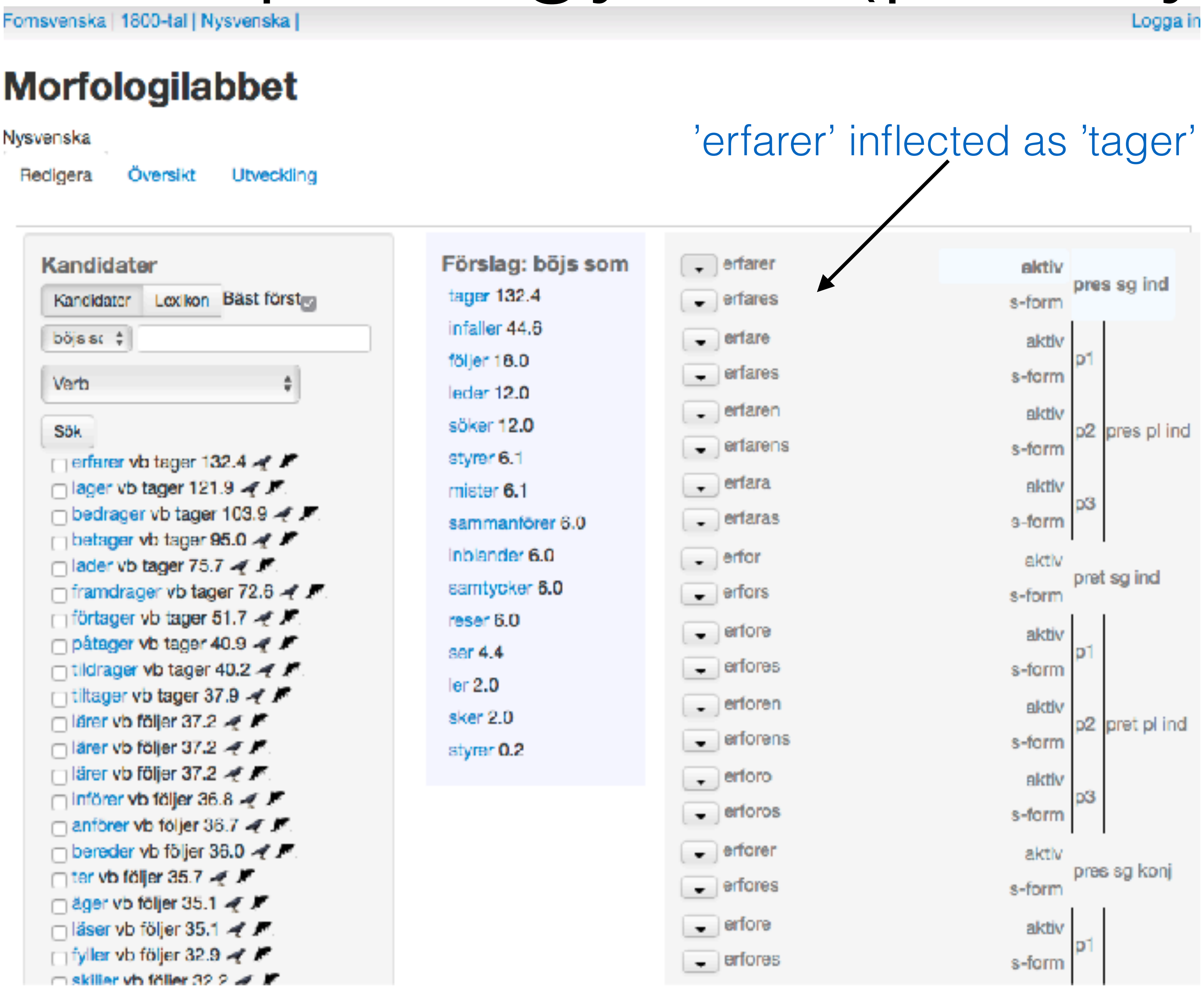
match(x_1+a+x_2, ”Ananas”) = {x_1=An, x_2=nas},
{x_1=Anan, x_2=s}

What have we achieved?

- We can actually **hide away the paradigm functions** and describe inflections by statements such as: *word X is inflected as word Y (or equivalent, this set of words S).*

- Might this be more natural way for a linguist to **define a computational morphology**?

The morphology lab (prototype)



Built-in paradigm induction and prediction

2. Paradigm prediction

Prediction task

- Given a **word form** (typically the lemma), **predict** its **paradigm function**/inflection table.
- The paradigm induction **gives us set of words for each paradigm function**, sharing that function.
- **Idea**: predict the appropriate paradigm function for an **input lemma** by comparing it to the words of the paradigms, and **chose the set of words it is most similar to**.

The classifier

- We first defined a **hand-crafted classifier** for the task (in AFH14).
- We then improved on it using a **linear SVM** (one-vs-the-rest multi-class) with **edge-anchored features** (i.e., prefixes and suffixes).
- We also tried other substring variants, but with worse results.

Evaluation data

- **Evaluation set 1**
  Inflection tables for three languages from Wiktionary tables (Durrett & DeNero, 2013). Languages: **Finnish** (nouns/adjectives, verbs), **Spanish** (verbs), **German** (nouns, verbs).
  Clean data with no defective or variant forms.

- **Evaluation set 2**
  Additional inflection tables gathered from various resources for: **Catalan** (nouns, verbs), **English** (verbs), **French** (nouns, verbs), **Galician** (nouns, verbs), **Italian** (nouns, verbs), **Portuguese** (nouns, verbs), **Russian** (nouns), **Maltese** (verbs).
  More messy data with defective tables, variants forms (e.g., cactuses - cacti), et cetera.

Eval 1: paradigm induction

Data	Input: inflection tables	Output: abstract paradigms
DE-VERBS	1827	140
DE-NOUNS	2564	70
ES-VERBS	3855	97
FI-VERBS	7049	282
FI-NOUNS-ADJS	6200	258

(dev: 200 tables)
(test: 200 tables)

Eval 1: Results comparison with D&DN13

Data	Per table accuracy			Per form accuracy			Oracle acc. per form (table)
	SVM	AFH14	D&DN13	SVM	AFH14	D&DN13	
DE-VERBS	**91.5**	68.0	85.0	**98.11**	97.04	96.19	99.70 (198/200)
DE-NOUNS	**80.5**	76.5	79.5	**89.88**	87.81	88.94	100.00 (200/200)
ES-VERBS	**99.0**	96.0	95.0	**99.92**	99.52	99.67	100.00 (200/200)
FI-VERBS	**94.0**	92.5	87.5	**97.14**	96.36	96.43	99.00 (195/200)
FI-NOUNS-ADJS	**85.5**	85.0	83.5	**93.68**	91.91	93.41	100.00 (200/200)

Eval 2: Table accuracy

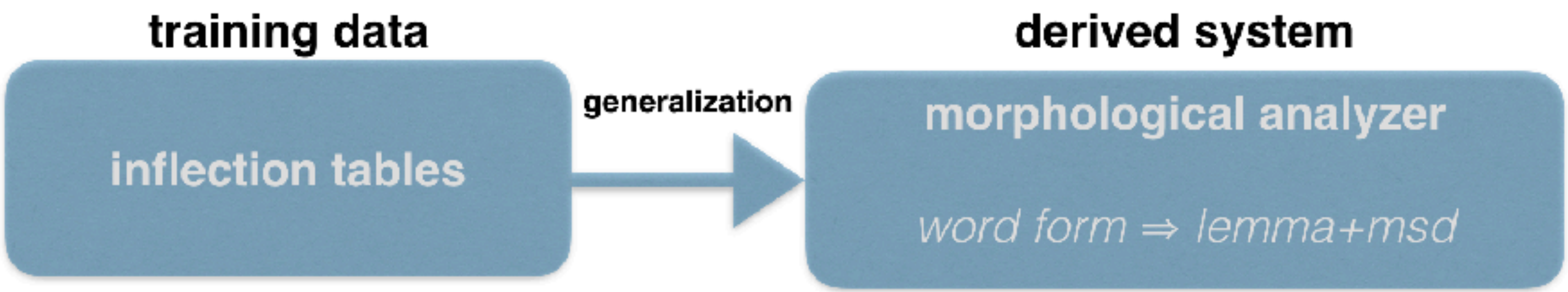
Data	#tbl	#par	mfreq	AFH14	SVM	Oracle
DE-N	2,210	66	18.99	76.09	**77.68**	98.99
DE-V	1,621	125	52.77	65.02	**83.59**	95.45
ES-V	3,243	90	70.42	92.25	**93.48**	96.59
FI-N&A	4,000	233	26.52	**83.20**	82.84	98.12
FI-V	4,000	204	43.04	**91.88**	91.64	94.76
MT-V	826	200	10.68	18.83	**38.64**	85.63
CA-N	4,000	49	44.12	94.00	**94.92**	99.44
CA-V	4,000	164	60.44	90.76	**93.40**	98.48
EN-V	4,000	161	77.12	89.40	**90.00**	97.40
FR-N	4,000	57	92.16	91.60	**93.96**	98.72
FR-V	4,000	95	81.52	93.72	**96.48**	98.80
GL-N	4,000	24	88.36	90.48	**95.08**	99.80
GL-V	3,212	101	45.21	58.92	**60.87**	98.95
IT-N	4,000	39	83.84	92.32	**93.76**	99.40
IT-V	4,000	115	63.96	89.68	**91.56**	98.68
PT-N	4,000	68	74.52	88.12	**90.88**	99.04
PT-V	4,000	92	62.00	76.96	**80.20**	99.20
RU-N	4,000	260	15.76	64.12	**66.36**	96.80

Eval 2: Form accuracy

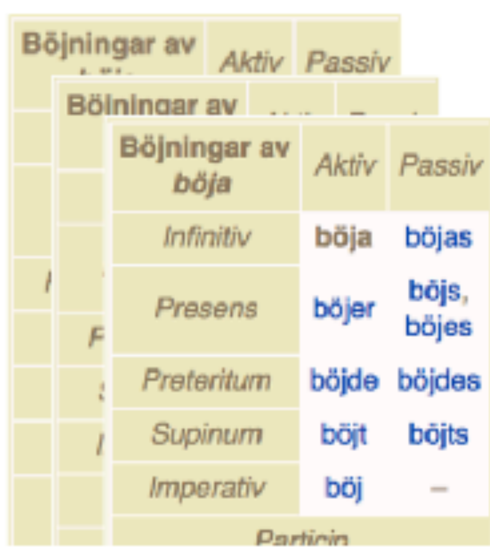
Data	#forms	mfreq	AFH14	SVM	Oracle
DE-N	8	57.36	89.72	**90.25**	99.69
DE-V	27	87.35	**96.12**	95.28	99.20
ES-V	57	93.80	98.72	**98.83**	99.47
FI-N&A	233	52.15	91.03	**91.06**	98.95
FI-V	54	70.38	**95.27**	95.22	96.76
MT-V	16	39.75	54.66	**61.15**	95.49
CA-N	2	71.30	96.89	**97.33**	97.93
CA-V	53	86.89	98.18	**98.89**	99.77
EN-V	6	91.43	95.93	**96.16**	99.28
FR-N	2	93.24	92.48	**94.68**	99.08
FR-V	51	91.47	97.09	**98.33**	99.02
GL-N	2	91.92	92.82	**95.38**	99.78
GL-V	70	94.89	**98.48**	98.32	99.67
IT-N	3	89.36	93.38	**94.59**	97.44
IT-V	51	89.51	97.76	**98.21**	99.64
PT-N	4	83.35	89.78	**91.97**	98.60
PT-V	65	92.62	96.81	**97.20**	99.68
RU-N	12	25.16	88.19	**89.35**	99.15

3. Deriving morphological analyzers

Morphological analyzers



example (swe)



Böjningar av *böja*	*Aktiv*	*Passiv*
Infinitiv	böja	böjas
Presens	böjer	böjs, böjes
Preteritum	böjde	böjdes
Supinum	böjt	böjts
Imperativ	böj	–

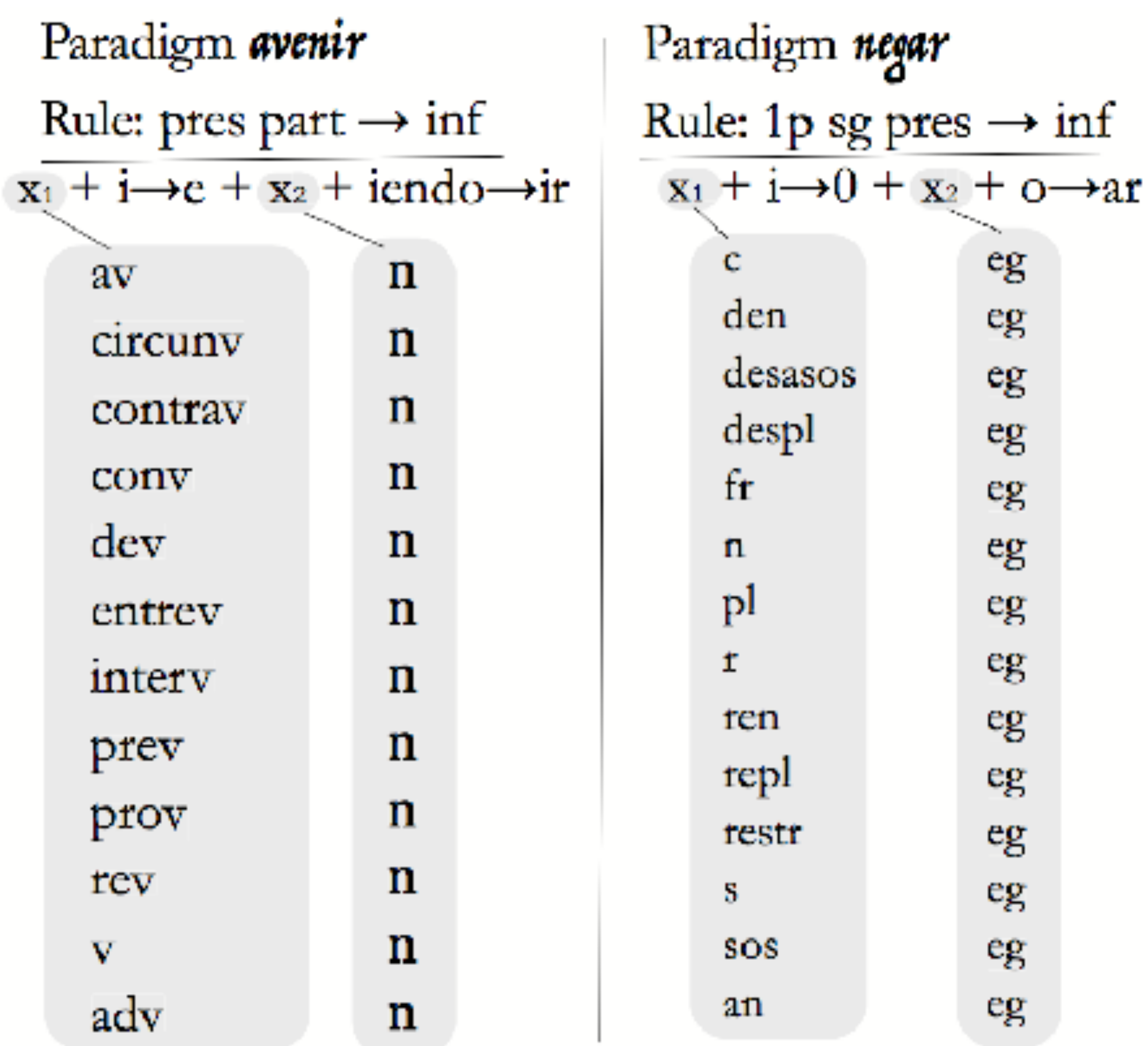
example run (swe)

uppvärmde ⇒ uppvärma:verb+aktiv+preteritum [+other analyses]

From inflection table to FST

- An inflection table may be interpreted as a set of string relations. In particular:
  wordform => **lemma** +wordform **msd**.
- And we can build a **FST** over these relations.
- **Problem**: allowing variables to match any substring may **overgenerate** a lot.
- So we need to **constrain the variables**.

Learning variable constraints



$$p_{\text{unseen}} = (1 - \frac{1}{t+1})^n$$

$$p_{\text{unseen}} < 0.05 \Rightarrow \textit{set is closed}$$

Constraining the variables of the **avenir** paradigm:

$$x_1 = (\Sigma^* v) \quad x_2 = n$$

Hierarchical analyses

We generate three separate analyzers: **Original**, where variables only matches previously seen instantiations; **Constrained,** where variables are constrained; **Unconstrained**, where all variables are completely unconstrained. These analyzers can be combined into one large transducer by, e.g., an operation commonly called *priority union*:

$$\textbf{Original} \cup_P \textbf{Constrained} \cup_P \textbf{Unconstrained}$$

Evaluation: D&D-data any analysis

Language		L-recall	L+M-recall	L/W	L+M/W
German	nouns	95.30	95.06	2.08	9.52
	verbs	91.18	92.44	4.16	9.57
	nouns+verbs	92.11	93.04	4.91	14.10
Spanish	verbs	98.06	97.98	1.93	2.20
Finnish	nounadj	88.69	88.48	4.10	5.30
	verbs	94.52	94.47	3.77	4.60
	nounadj+verbs	92.63	92.43	12.56	16.40

L-recall: correct lemma constructed
L+M-recall: correct lemma+MSD constructed
L/W: candidate lemma/word form
L+MSD/W: candidate lemma+msd/word form

Evaluation: D&D-data
Selecting the top ranked

Language		Lemma	L+MSD	MSD
German	nouns	77.06	69.44	79.50
	verbs	90.02	89.76	92.78
Spanish	verbs	96.92	96.92	97.43
Finnish	nounadj	70.29	69.68	91.59
	verbs	90.44	90.44	98.02

Thanks for listening! and some references

1. Forsberg, M., Hulden, M. (2016). **Learning Transducer Models for Morphological Analysis from Example Inflections**. *In Proceedings of StatFSM.* Association for Computational Linguistics.

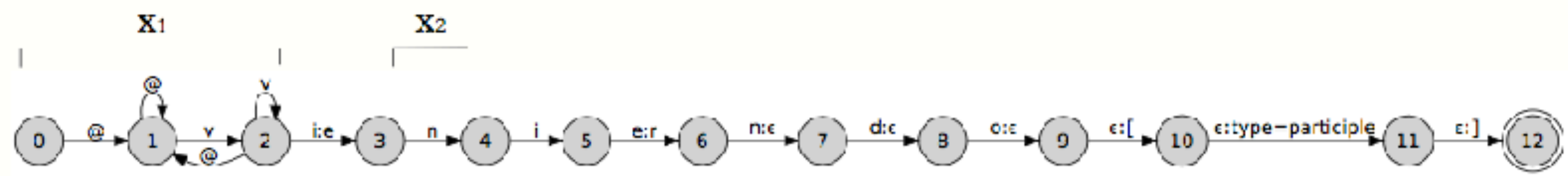
2. Forsberg, M., Hulden, M. (2016). **Deriving Morphological Analyzers from Example Inflections**. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC-2016).

3. Ahlberg, M., Forsberg, M., Hulden, M. (2015). **Paradigm classification in supervised learning of morphology**. In *Proceedings of NAACL-HLT 2015*.

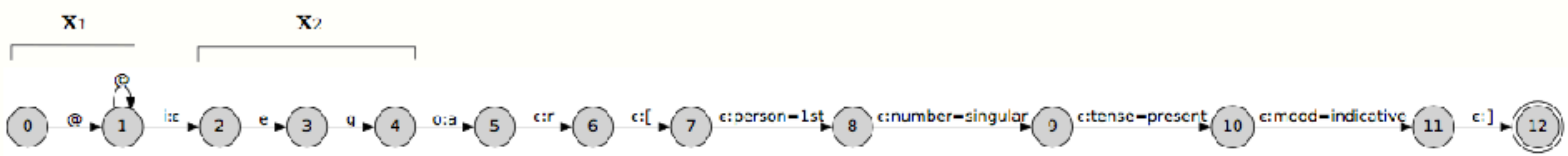
4. Adesam, Y., Ahlberg, M., Andersson, P., Bouma, G., Forsberg, M., Hulden, M. (2014). **Computer-aided morphology expansion for Old Swedish**. In *Proceedings of LREC 2014*.

5. Hulden, M.; Forsberg, M., Ahlberg, M. (2014). **Semi-supervised learning of morphological paradigms and lexicons**. In *EACL 2014*.

Deriving morphological analyzers



aviniendo → avenir[type=participle]



ciego → cegar[person=1st number=singular mood=indicative]

Two single generalized word forms mapped to lemma+msd. The variables x_1 and x_2 are marked.

Prediction and NN

- *SIGMORPHON 2016 Shared Task on Morphological Reinflection*: **Kann et al. 2016**

Language	Task 1	Task 2	Task 3
Arabic	95.47%	97.38%	96.52%
Finnish	96.80%	97.40%	96.56%
Georgian	98.50%	99.14%	98.87%
German	95.80%	97.45%	95.60%
Hungarian	99.30%	99.67%	99.50%
Maltese	88.99%	88.17%	87.83%
Navajo	91.48%	96.64%	96.20%
Russian	91.46%	91.00%	89.91%
Spanish	98.84%	98.74%	97.96%
Turkish	98.93%	97.94%	99.31%

Table 2: Exact-match accuracy per language for the standard track of the SIGMORPHON 2016 Shared Task.

- So we are interested in the **combination of NN and our paradigmatic representations**.