Lexicon building

Markus Forsberg

GF summer school in Riga 2017

Today's talk

- **Part I: computational morphology**
  - What can we learn from inflection tables?

- **Part II: Word senses in GF**
  - a few slides; if there is time

Part II:
Computational morphology

What can we learn from inflection tables?

work done together with Måns Huldén and Malin Ahlberg

Think about this question for a minute: 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

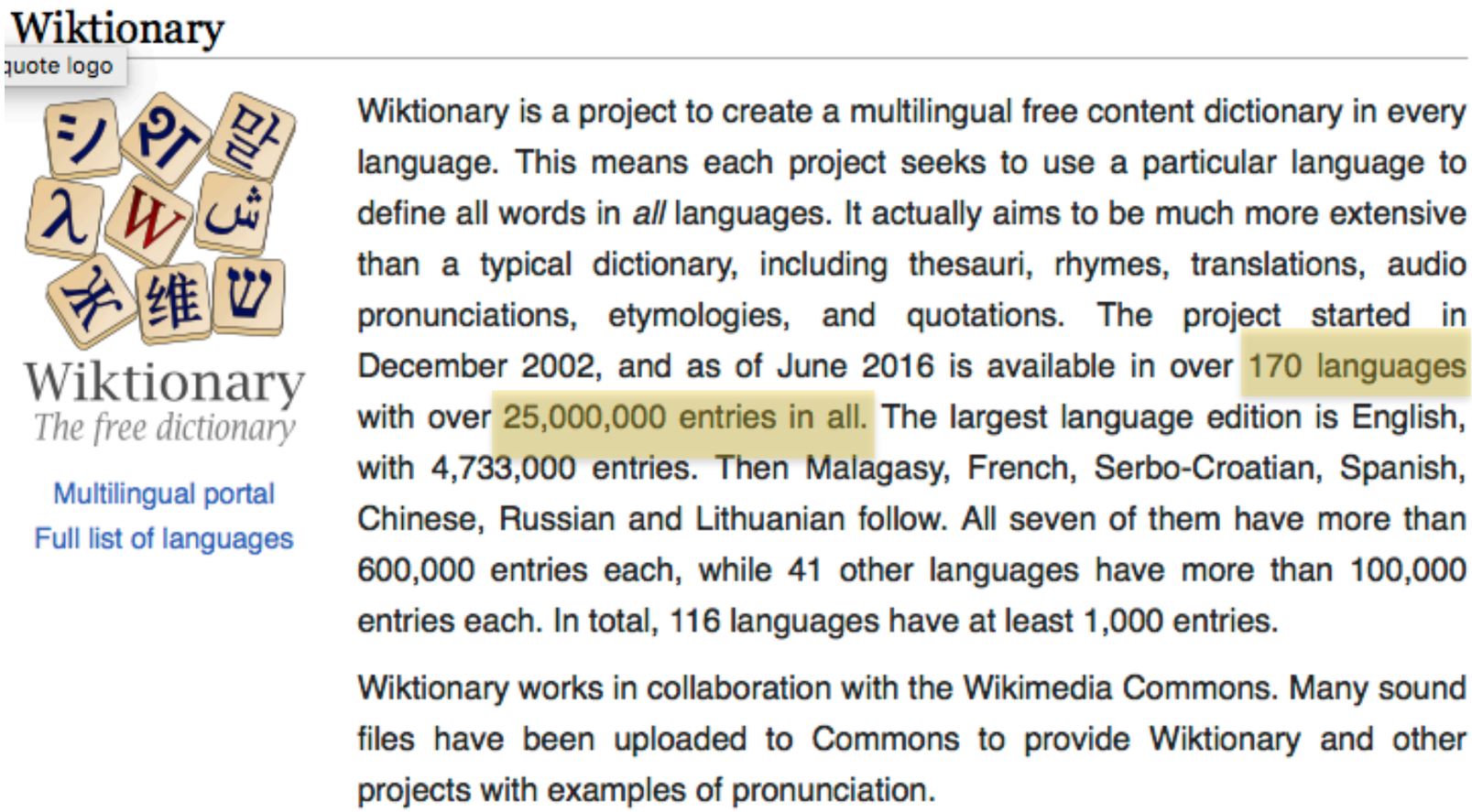
Verbum infinitum

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:



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.

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	**Fr**a**ck**	**Fr**ä**ck**e
Genitive	**Fr**a**ck**es, **Fr**a**ck**s	**Fr**ä**ck**e
Dative	**Fr**a**ck**, **Fr**a**ck**e	**Fr**ä**ck**en
Accusative	**Fr**a**ck**	**Fr**ä**ck**e

	Singular	Plural
Nominative	**Anf**a**ng**	**Anf**ä**ng**e
Genitive	**Anf**a**ng**es, **Anf**a**ng**s	**Anf**ä**ng**e
Dative	**Anf**a**ng**, **Anf**a**ng**e	**Anf**ä**ng**en
Accusative	**Anf**a**ng**	**Anf**ä**ng**e

Induction

$f(x_1,x_2) =$

	Singular	Plural
Nominative	$\mathbf{x_1}$+a+$\mathbf{x_2}$	$\mathbf{x_1}$+ä+$\mathbf{x_2}$+e
Genitive	$\mathbf{x_1}$+a+$\mathbf{x_2}$+es, $\mathbf{x_1}$+a+$\mathbf{x_2}$+s	$\mathbf{x_1}$+ä+$\mathbf{x_2}$+e
Dative	$\mathbf{x_1}$+a+$\mathbf{x_2}$, $\mathbf{x_1}$+a+$\mathbf{x_2}$+e	$\mathbf{x_1}$+ä+$\mathbf{x_2}$+en
Accusative	$\mathbf{x_1}$+a+$\mathbf{x_2}$	$\mathbf{x_1}$+ä+$\mathbf{x_2}$+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**rar*, ***compr**a*, ***compr**o*

Competing LCS

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

***sege**l*, ***seg**let*, ***seg**len* LCS: *sege*

LCS ambiguity resolution through heuristics

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

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

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

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

segel*, *seglet*, *seglen LCS: *segl*

segel*, *seglet*, *seglen 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)?

$$f(x_1, x_1, .., x_n) => f(w_{1,} w_{1,} ..., w_{n)}$$

- 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.

$$match(x_1+a+x_2, \text{"Frack"}) = \{x_1=Fr, x_1=ck\}$$

Regular expression with groups

Ambiguity

$$match(x_1+a+x_2, \text{"Ananas"}) = \{x_1=An, x_2=nas\}, \{x_1=Anan, x_2=s\}$$

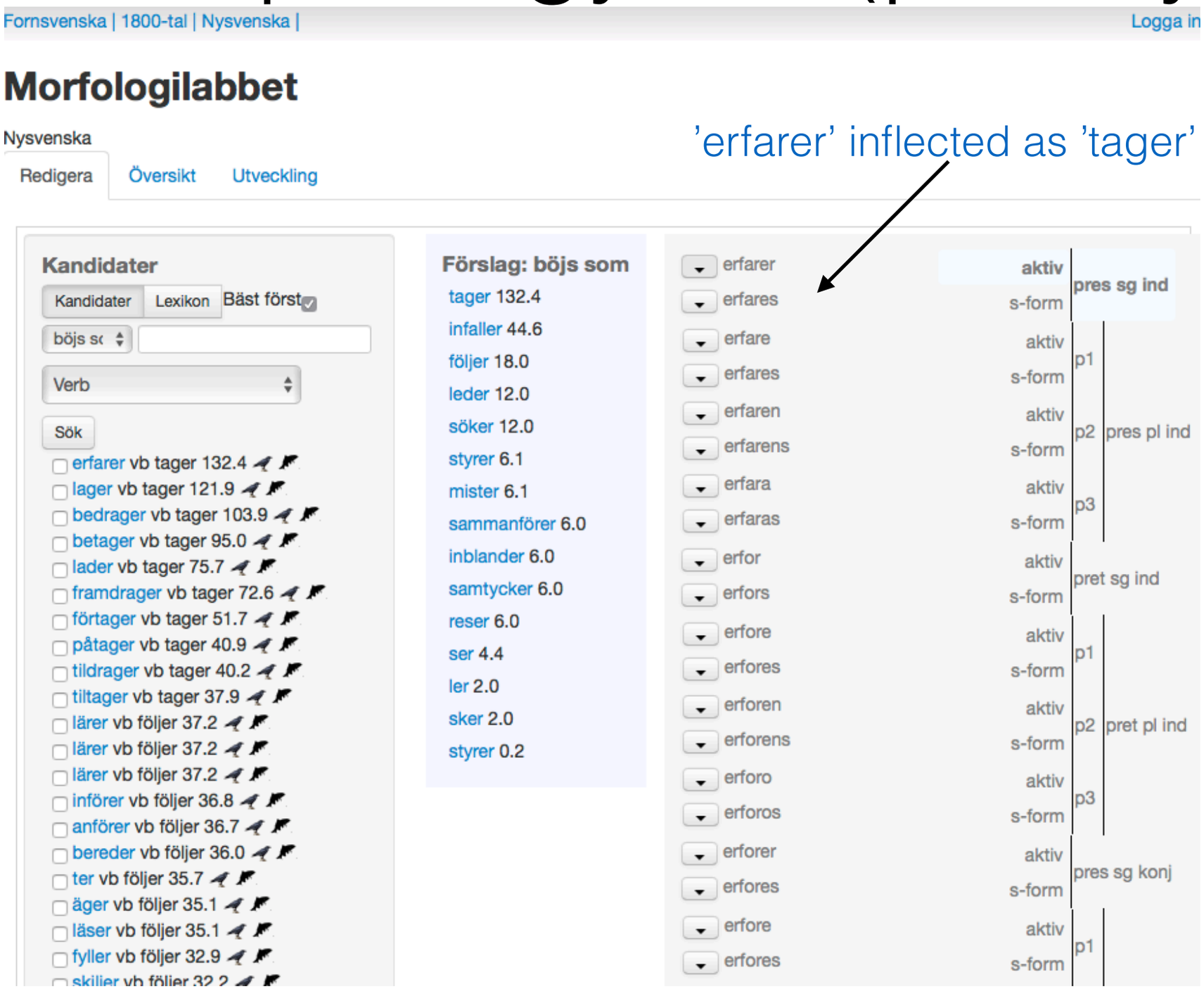
What have we achieved?

- We can actually keep the **the paradigm functions** hidden in the background.

- Specifying inflection becomes: w*ord X is inflected as some other word Y (with an already known inflection table).*

- Might this be more natural way for a non-computational 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 (described 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

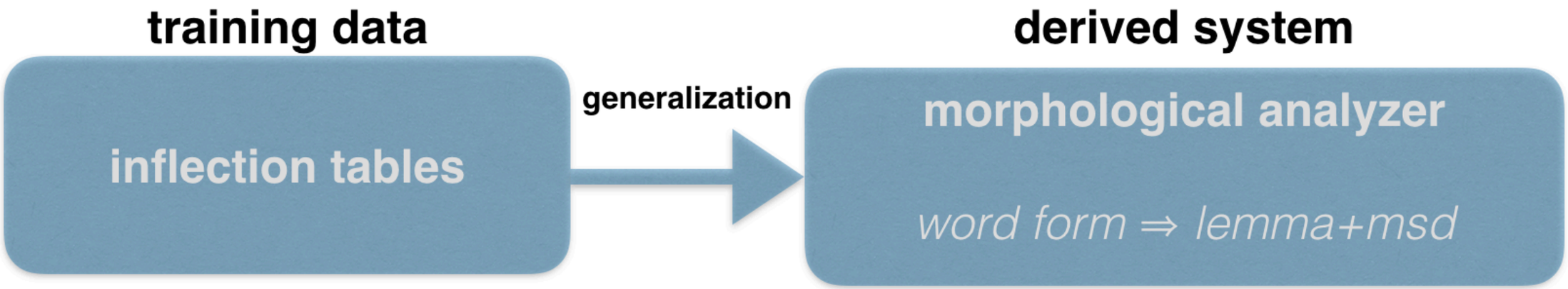
Paradigm prediction in GF: smart paradigms

- A smart paradigm in GF is a gateway function that selects the approriate inflection function based on the input form(s). E.g. (from Detréz and Ranta 2012):

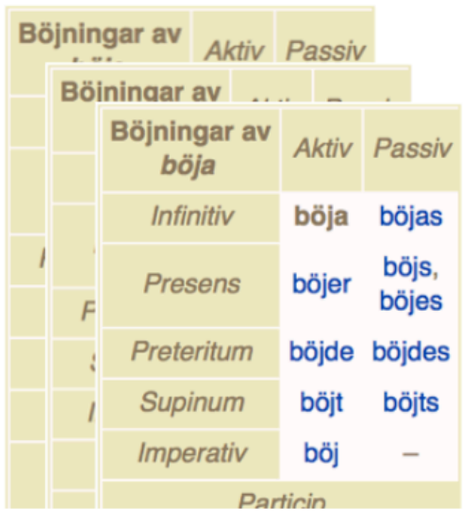
```
mkV : Str -> V
mkV s = case s of {
_ + "ir" -> conj19finir s ;
_ + ("eler"|"eter") -> conjlljeter s ;
_ + "er" -> conj06parler s ;
}
```

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	–
Particip		

example run (swe)

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

A similar task to paradigm prediction, but here the input is any word form.

From inflection table to FST

- An inflection table may be interpreted as a set of string relations. In particular:
  wordform => **lemma** +**wordform's msd**.

- 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

Paradigm ***avenir***

Rule: pres part → inf

x_1 + i→e + x_2 + iendo→ir

x_1	x_2
av	n
circunv	n
contrav	n
conv	n
dev	n
entrev	n
interv	n
prev	n
prov	n
rev	n
v	n
adv	n

Paradigm ***negar***

Rule: 1p sg pres → inf

x_1 + i→0 + x_2 + o→ar

x_1	x_2
c	eg
den	eg
desasos	eg
despl	eg
fr	eg
n	eg
pl	eg
r	eg
ren	eg
repl	eg
restr	eg
s	eg
sos	eg
an	eg

Learning variable constraints

- Assume uniform distribution (just a heuristic!)
- Calculate the probability that there is an unseen string in a variable.
- If the probability is low, assume that we seen everything already.
- If the probability is high, do the same thing for prefixes and suffixes (with smaller and smaller strings).

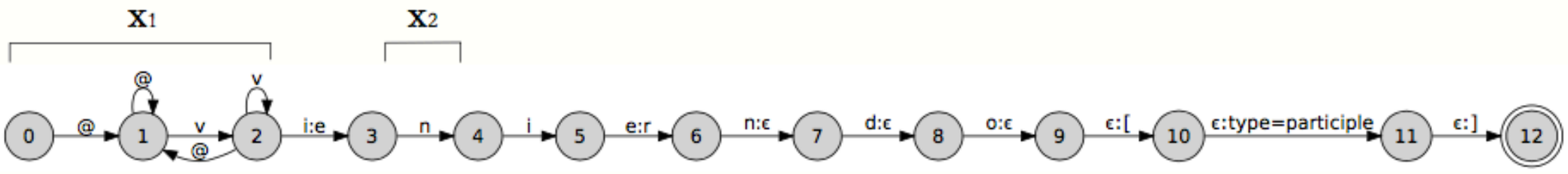
$$p_{\text{unseen}} = (1 - \frac{1}{t+1})^n$$

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

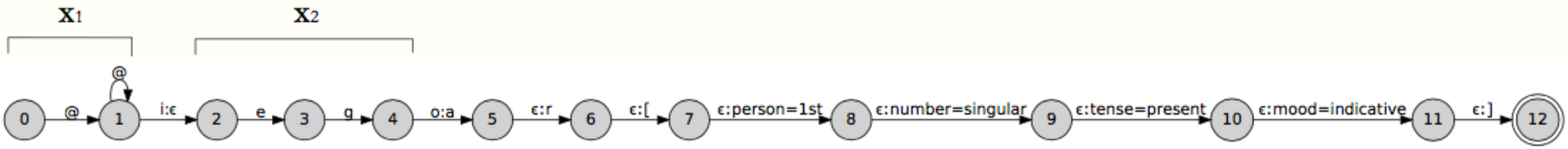
Constraining the variables of the **avenir** paradigm:

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

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.

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*:

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

Ranking

- The analyser has until now been **unweighted**, i.e., its goal is to **give all plausible analyses** while curbing the unwanted ones.

- But for practical use, we want the plausible analyses to be ranked, to get at the **most plausible analysis**.

- We do that by creating a **language model for each variable**.

- The ranking depends on how well a plausible analysis **fits its variables' language models**.

Evaluation: D&D-data unweighted (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 weighted (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

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*.

Part II:

Word senses in GF

The lexicon in GF

- No clear (theoretical) distinction between the lexicon and anything else.

- Probably the best distinction: The lexicon is the set of zero-place functions.

  ```
  fun word: PoS;
  ```

- These zero-place functions correspond to word senses.

  ```
  fun word : PoS ; -- a particular sense of the word 'word'.
  ```

Making sense out of words

xkcd.com

- Just to get it out of the way: word senses are just abstractions — a word has no god-given number of senses.

- As Kilgariff sensibly put it in ”*I don’t believe in word senses*” (1997):
”*[…] word senses exist only relative to a task.*”

How many senses?
The two extremes

(1) ***A word in an unique context constitutes a word sense****.*
(= we all produce (slightly) new word senses continously, since no two words are in exactly in the same context)

(2) ***A lemma has exactly one sense****.*
(= we don't need to care about word senses, just forms)

The middle: splitters vs lumpers

1. paper – (a material made of cellulose pulp derived mainly from wood or rags or certain grasses)
2. composition, paper, report, theme – (an essay (especially one written as an assignment); "he got an A on his composition")
3. newspaper, paper – (a daily or weekly publication on folded sheets; contains news and articles and advertisements; "he read his newspaper at breakfast")
4. paper – (a scholarly article describing the results of observations or stating hypotheses; "he has written many scientific papers")
5. paper – (medium for written communication; "the notion of an office running without paper is absurd")
6. newspaper, paper, newspaper publisher – (a business firm that publishes newspapers; "Murdoch owns many newspapers")
7. newspaper, paper – (a newspaper as a physical object; "when it began to rain he covered his head with a newspaper"

⇒

1. paper – material (1 and 5)
2. paper – composition, article (2 and 4)
3. paper – newspaper, publication, publisher (3,6,7)

Homonymy and polysemy

- **homonymy**: same form, unrelated meaning
  - (a baseball) **bat** vs **bat** (a furry flying object)
    - probably realized with different words in other languages (e.g., Swedish: ’basebollträ’ vs ’fladdermus’)
- **Polysemy**: same form, related meaning
  - **university** (the institution) vs **university** (the building)
  - Often with the same word in other languages as well.

Regular polysemy

- **animal** ~ **food** (I saw a duck; I ate duck yesterday)
- **kind** ~ **portion** (two beer = two servings of beer/two kinds of beer)
- **causative** ~ **inchoative** (John broke the window/The window broke)
- **container** ~ **content** (He drank a bottle/He dropped the bottle)
- **object** ~ **person** (The cello is playing great tonight)
- **location** ~ **government** ~ **representative** (He visited China; China signs the trade agreement; China attended the peace conference)
- …

Examples collected from http://www.cs.upc.edu/~gboleda/pubs/talks/WSD_regularpolysemyIMS.pdf

So, where does this leave us?

- How should we think about word senses in GF?
- Well, if the GF task is **to create as good multilingual translations as possible.**
- then we should only make a **sense distinction** if it actually **improves the translation quality**.
- not because some monolingual dictionary makes a particular sense distinction.

Nothing more.

Thanks for listening!