University of Colorado **Boulder**

UNIVERSITY OF GOTHENBURG

Paradigm classification in supervised learning of morphology

Malin Ahlberg
Markus Forsberg
Mans Hulden

University of Colorado **Boulder**

UNIVERSITY OF GOTHENBURG

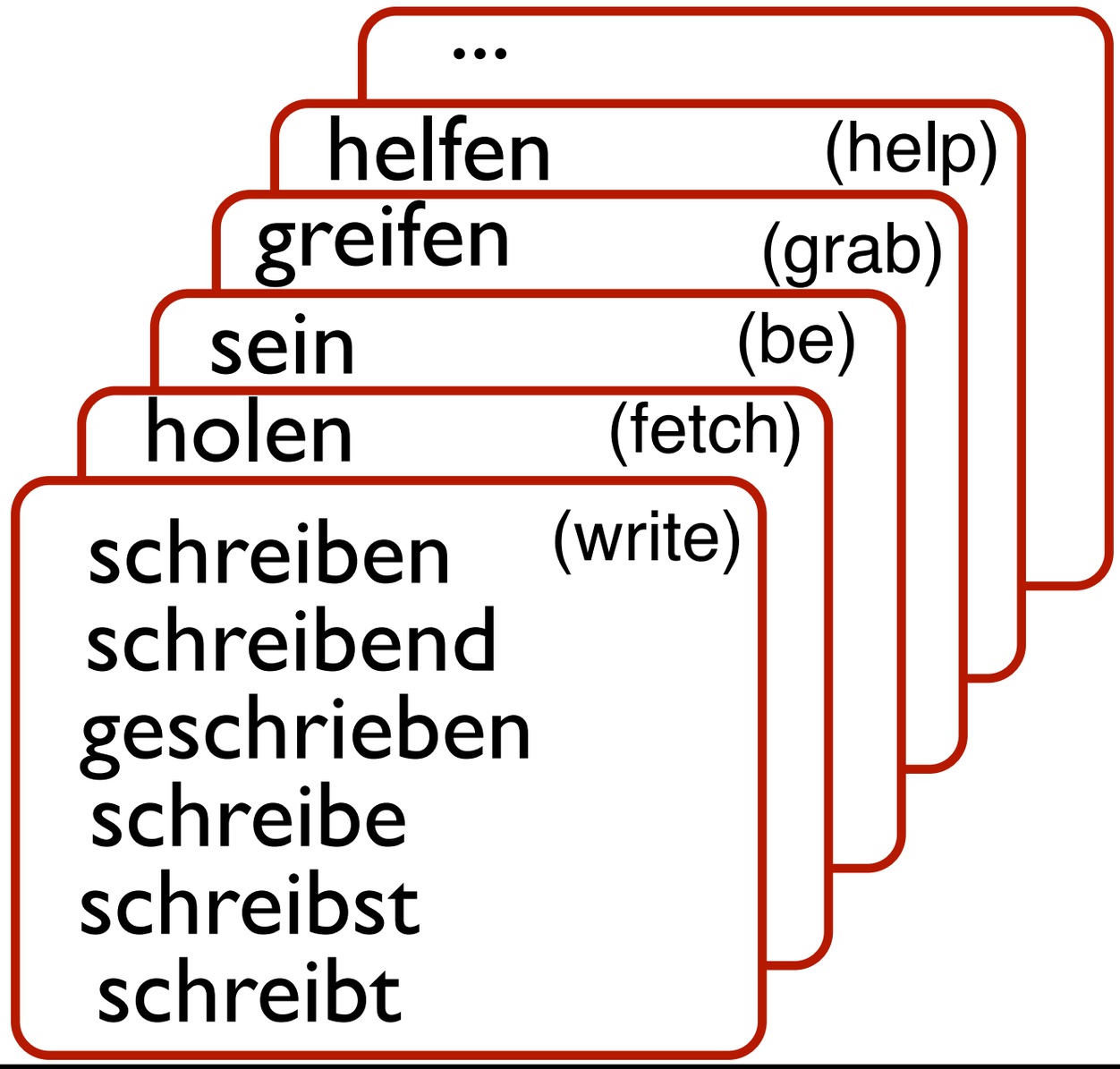
Overview

Goal: learn to inflect (unseen) words from annotated data in a language-independent way

Overview

Goal: learn to inflect (unseen) words from annotated data in a language-independent way

Data (inflection tables)
German verbs example:

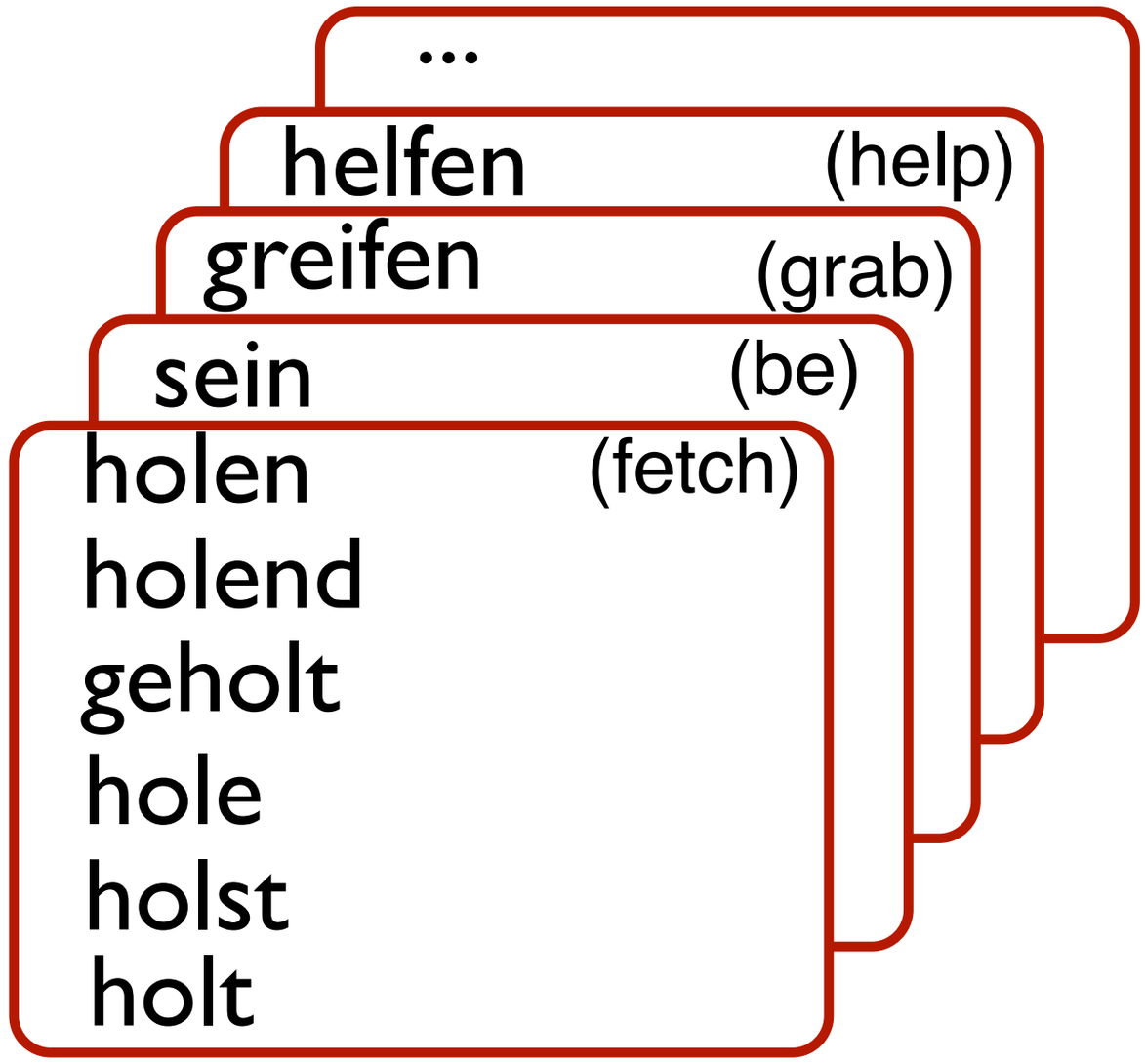


Overview

Goal: learn to inflect (unseen) words from annotated data in a language-independent way

Data (inflection tables)

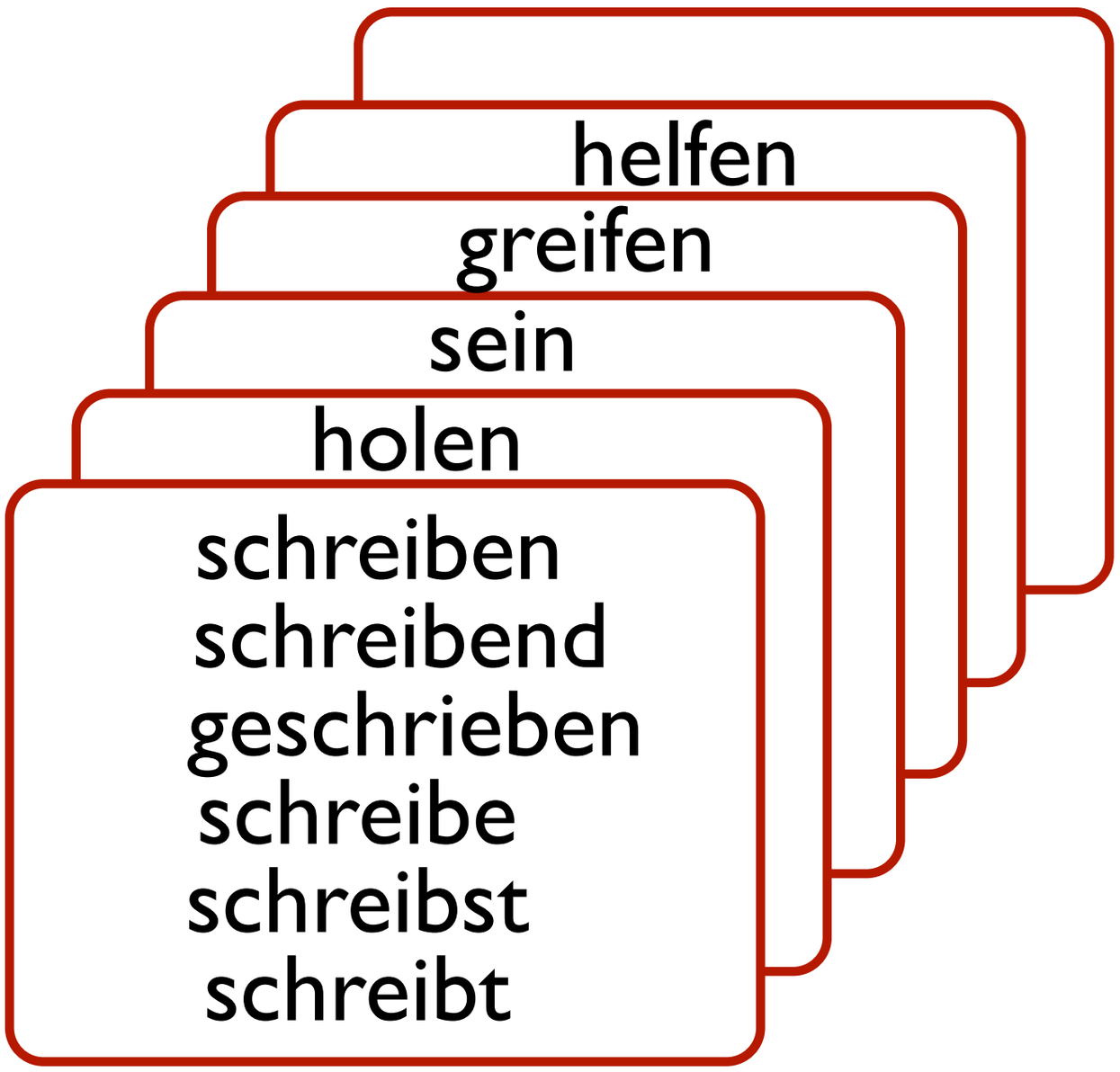
German example:



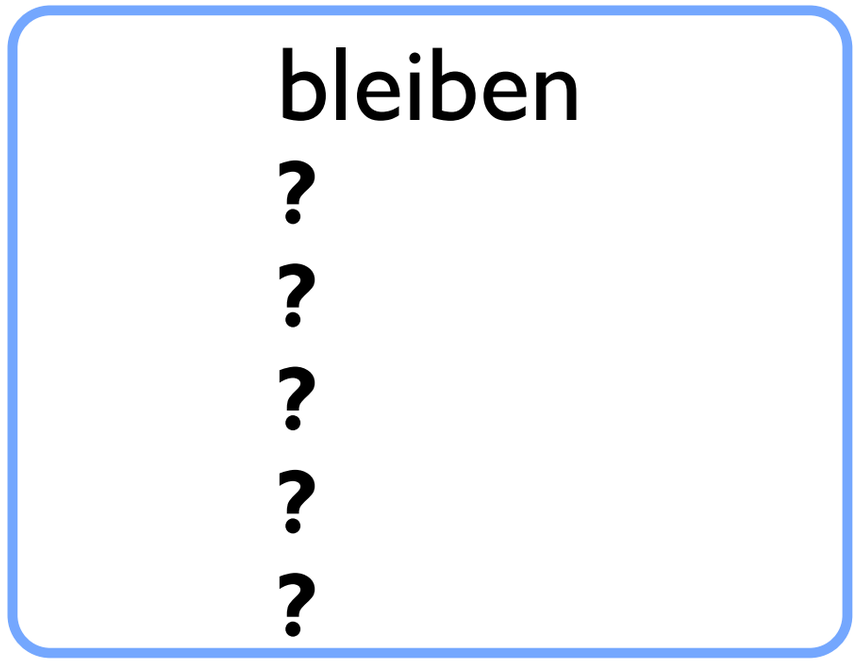
Overview

Training data

...



Reconstruct unseen forms from lemma/base form



'stay'

Previous work (recent)

Dreyer and Eisner (2011)

- Semi-supervised Bayesian model that learns from a small amount of "seed paradigms"

Durrett and DeNero (2013)

- Supervised discriminative model that learns rule transformations to reconstruct paradigms from many examples

Ahlberg et al. (2014)

- Symbolic model; adaptable to supervised/semi-supervised settings

University of Colorado **Boulder**

UNIVERSITY OF GOTHENBURG

Method

Produce an abstract representation of inflection paradigms by extracting the longest common subsequence (LCS) from each inflection table, and assigning piecewise discontinuous subsequences to variables Ahlberg et al. (2014), Hulden (2014)

infl. table	"paradigm"
schreiben	x_1+**e**+x_2+x_3+**en**
schreibend	x_1+**e**+x_2+x_3+**end**
geschrieben	**ge**+x_1+x_2+**e**+x_3+**en**
schreibe	x_1+**e**+x_2+x_3+**e**
schreibst	x_1+**e**+x_2+x_3+**st**
schreibt	x_1+**e**+x_2+x_3+**t**

LCS = **schrib**

x_1 = **schr**

x_2 = **i**

x_3 = **b**

University of Colorado **Boulder**

UNIVERSITY OF GOTHENBURG

Method

Toy example (English verbs)

Input: inflection tables

① Extract LCS

```
ring  }
rang  }  rng
rung  }

swim  }
swam  }  swm
swum  }
```

University of Colorado **Boulder**

UNIVERSITY OF GOTHENBURG

Method

Toy example (English verbs)

Input: inflection tables

① Extract LCS

② Fit LCS to table

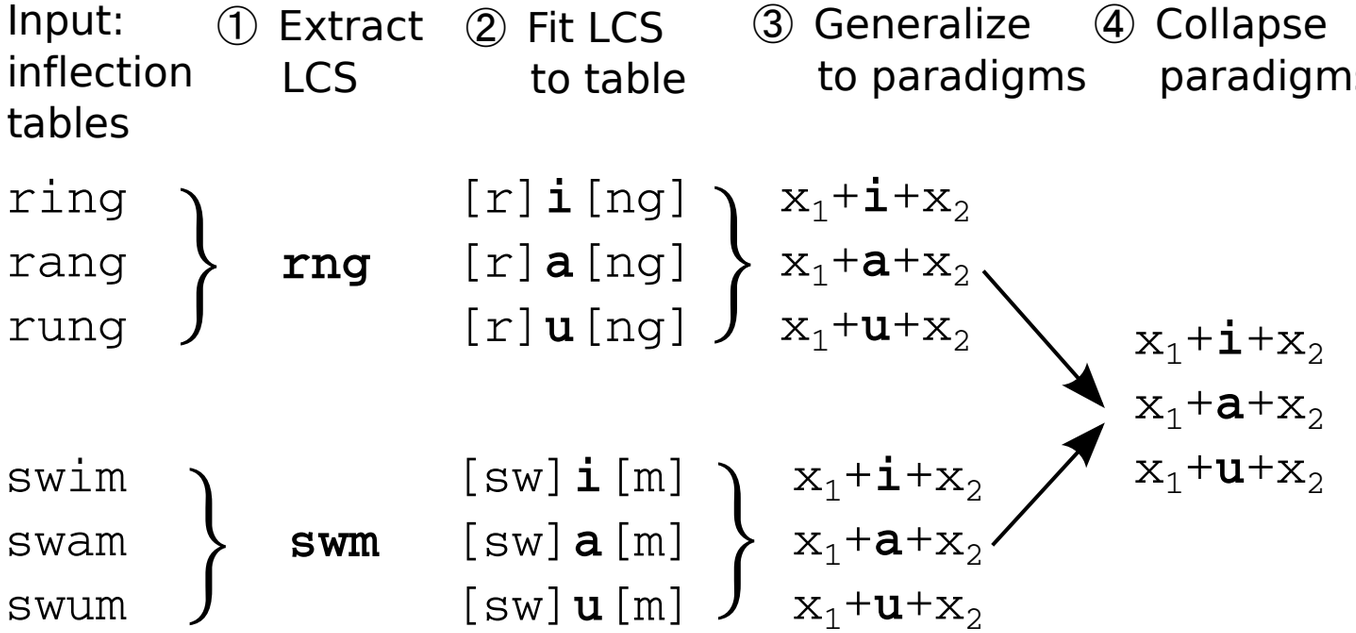
ring	**rng**	[r]**i**[ng]
rang		[r]**a**[ng]
rung		[r]**u**[ng]
swim	**swm**	[sw]**i**[m]
swam		[sw]**a**[m]
swum		[sw]**u**[m]

University of Colorado **Boulder**

UNIVERSITY OF GOTHENBURG

Method

Toy example (English verbs)



ring and *swim* belong to the same class

ring: $x_1 = r, \quad x_2 = g$

swim: $x_1 = sw, x_2 = m$

University of Colorado **Boulder**

UNIVERSITY OF GOTHENBURG

Method

Collapsing paradigms

WIKTIONARY
the free dictionary

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

Comparison:
Thompson (1998) lists 79 "classes" of Spanish verbs
Kotus (2007) Finnish grammar uses 51 noun (& adj) paradigms

*Wiktionary data from Durrett and DeNero (2013)

University of Colorado **Boulder**

UNIVERSITY OF GOTHENBURG

Reconstruction

An inflection table can be reconstructed from a lemma according to an abstract paradigm:

lemma	abstract paradigm guess
show	x_1
	x_1 +ed
	x_1 +n
	x_1 +ing
panic	x_1
	x_1 +ked
	x_1 +ked
	x_1 +king

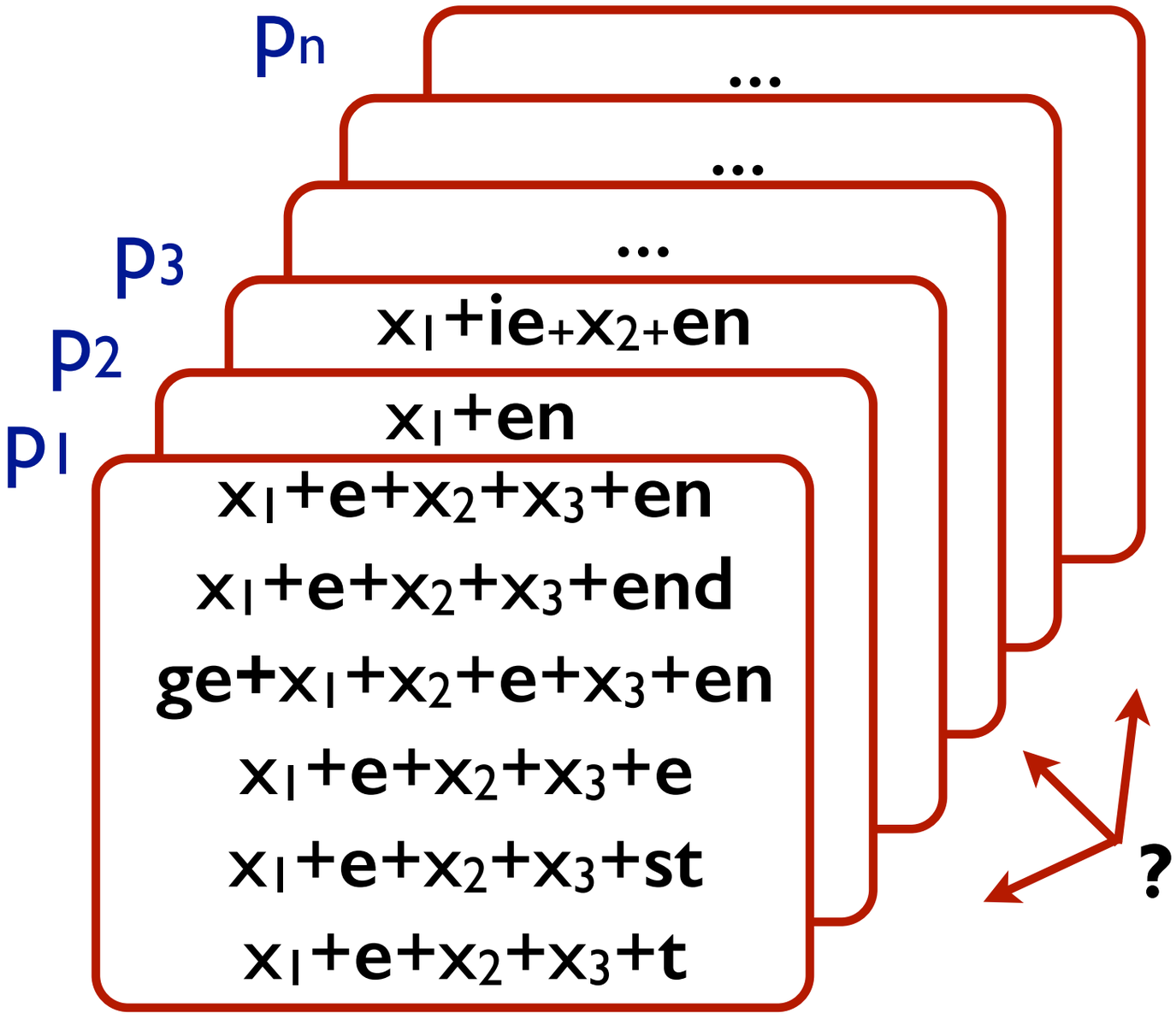
Reconstruction

An inflection table can be reconstructed from a lemma according to an abstract paradigm:

lemma	abstract paradigm guess
show	x_1
showed	x_1 +ed
shown	x_1 +n
showing	x_1 +ing
panic	x_1
panicked	x_1 +ked
panicked	x_1 +ked
panicking	x_1 +king

Reconstruction from lemma

Reduces to choosing the appropriate paradigm for the unknown lemma



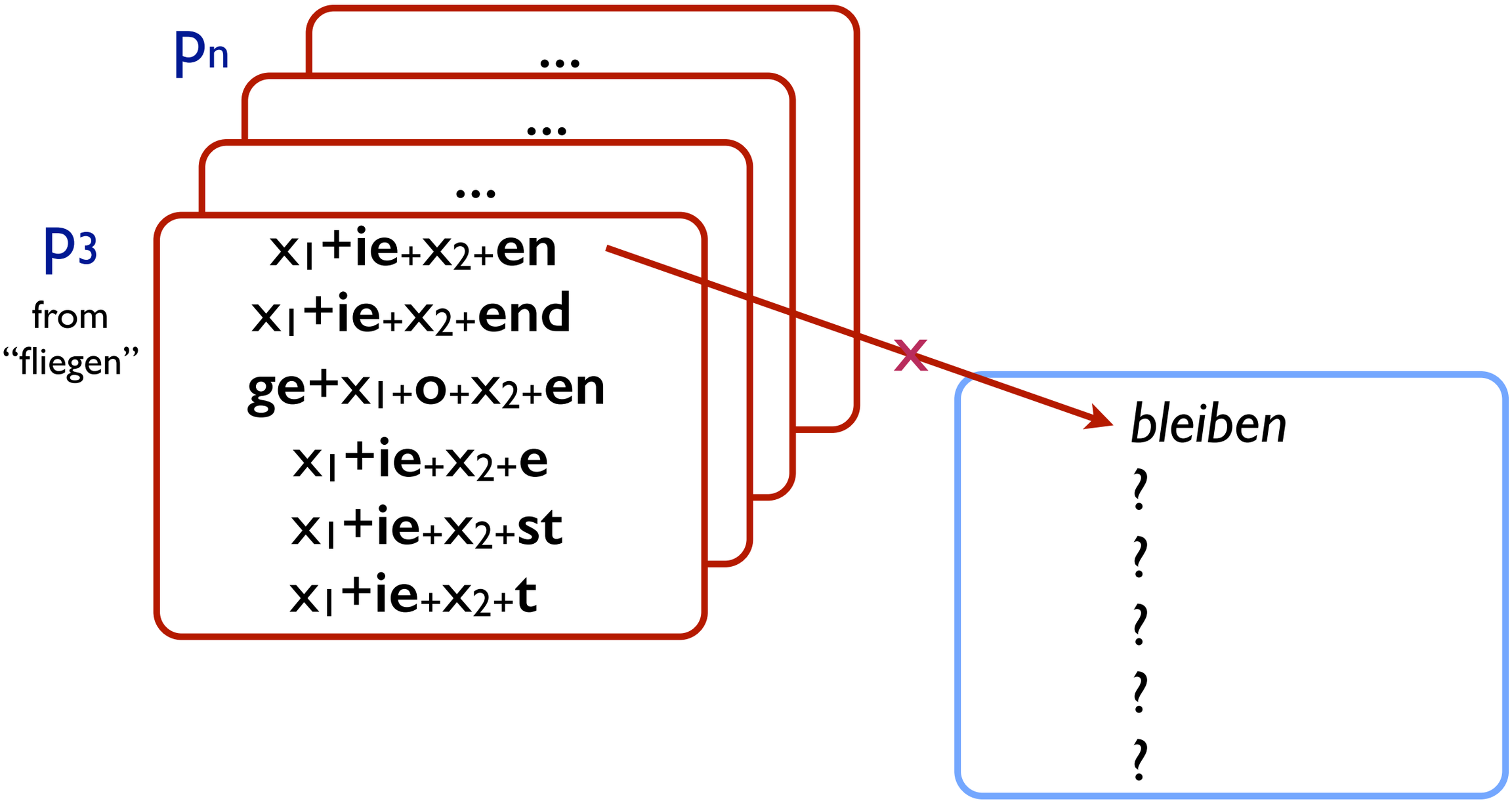
bleiben
?
?
?
?
?

University of Colorado **Boulder**

UNIVERSITY OF GOTHENBURG

Reconstruction from lemma

(many competing paradigms can be ruled out by simple inspection)

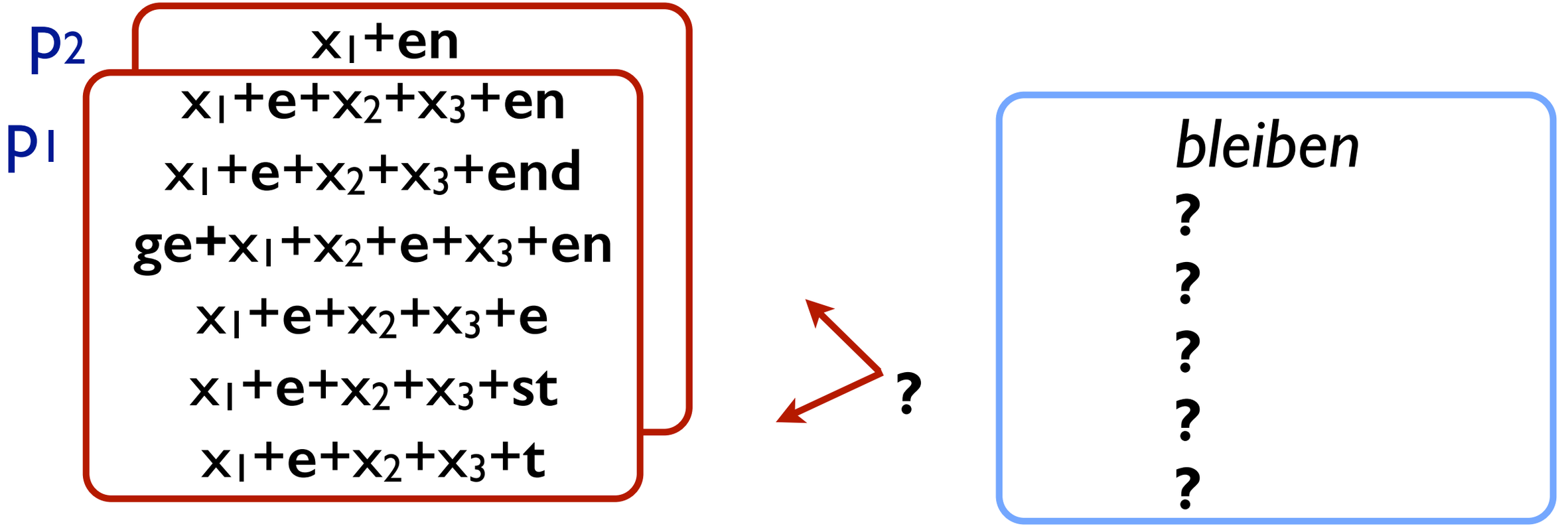


University of Colorado **Boulder**

UNIVERSITY OF GOTHENBURG

Reconstruction from lemma

Train a classifier for the remaining choices

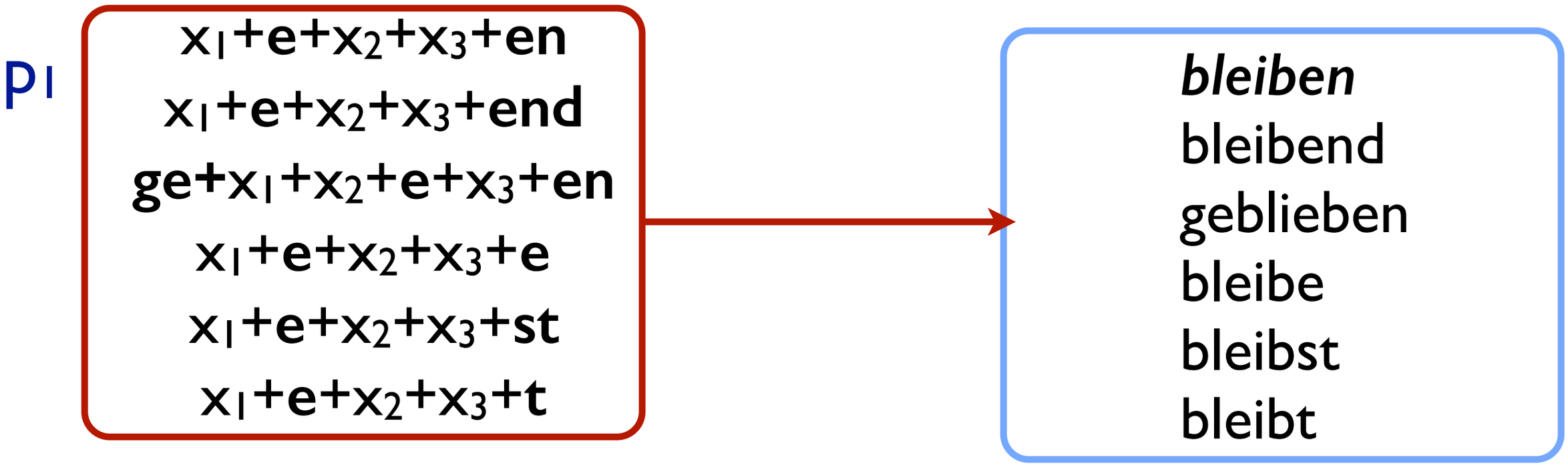


University of Colorado **Boulder**

UNIVERSITY OF GOTHENBURG

Reconstruction from lemma

After classification, complete table can be reconstructed from lemma



SVM classifier

- Only use lemmas seen in paradigms as training data
- Use edge-anchored substrings as binary features, e.g.

  f("lesen") = {#l, #le, #les, #lese, #lesen, lesen#, esen#, sen#, en#, n#}

- Linear SVM (one-vs-the-rest multi-class)
- Feature selection using dev set on maximum length of prefix/suffix to use (3-9 symbols), and whether to include prefix/suffix at all
- (Other types of substring-features were explored, with worse results)

Evaluation

(1)

Inflection tables for three languages from Wiktionary tables (Durrett & DeNero, 2013): *Finnish* (nouns/adjectives, verbs), *Spanish* (verbs), *German* (nouns, verbs)

(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)

(1) tables very clean, no defective forms/parallel forms
(2) contains defective tables, parallel forms (cactuses ~ cacti), etc.

Evaluation

(1)

Inflection tables for three languages from Wiktionary tables (Durrett & DeNero, 2013): Finnish (nouns/adjectives, verbs), Spanish (verbs), German (nouns, verbs)

WIKTIONARY
the free dictionary

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)

Results (1)

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)

Oracle = always picks the best paradigm

SVM = current method
AFH14 = Ahlberg, Forsberg, Hulden (2014) [LCS + suffix-based classifier]
D&DN13 = Durrett & DeNero (2013) [discriminative string transformation]

Results (2)

mfreq=pick most "popular" paradigm

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

mean (SVM) = 84.18

accuracy per table (entire inflection table correctly reconstructed)

Results (2)

mfreq=pick most "popular" paradigm

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

mean (SVM) = 93.46

accuracy per form (entire inflection table correctly reconstructed)

Discussion

Best:

Data	#forms	mfreq	AFH14	SVM	Oracle
CA-V	53	86.89	98.18	**98.89**	99.77

Worst:

Data	#forms	mfreq	AFH14	SVM	Oracle
MT-V	16	39.75	54.66	**61.15**	95.49

Maltese has ‘mixed’ lexicon of Semitic, Italian & Sicilian, English

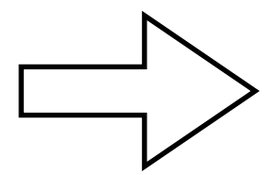
Maltese exhibits Semitic interdigitation (root-and-pattern paradigms) in verbs

Future work

Learn morphophonology

Finnish (consonant gradation)

ma.to	x_1+t+x_2
ma.dot ma.don	x_1+d+x_2+t x_1+d+x_2+n
ma.to.jen	x_1+t+x_2+jen
ma.to.a	x_1+t+x_2+a
…	…



t in open syllable onsets, d in closed syllable onsets

Learn “subparadigms"

Verb [edit]

naalnish

1. he/she is working
2. he/she is employed

Conjugation [edit]

IMPERFECTIVE	singular	duoplural	plural
1st person	naashnish	neiilnish	nideiilnish
2nd person	nanilnish	naołnish	nidaałnish
3rd person	naalnish		nidaalnish
4th person	nijilnish		nidajilnish

PERFECTIVE	singular	duoplural	plural
1st person	nishishnish	nishiilnish	nidashiilnish
2nd person	nishínílnish	nishoołnish	nidashoołnish
3rd person	naashnish		nidaashnish
4th person	nijishnish		nidajishnish

Related terms [edit]

- IMPERFECTIVE: -nish
- PERFECTIVE: -nish
- FUTURE: -nish
- ITERATIVE: -nish
- OPTATIVE: -nish

Navajo

x_1=n
x_2=nish

naashnish > x_1+aash+x_2
neiilnish > x_1+eiil+x_2
…

Ungeneralizable paradigm in Navajo

Verb [edit]

hałeeh

1. he/she is causing it, generating it

LCS = h

Conjugation [edit]

IMPERFECTIVE	singular	duoplural	plural
1st person	hashłeeh	hwiidleeh	dahwiidleeh
2nd person	hółeeh	hohłeeh	dahohłeeh
3rd person	hałeeh		dahałeeh
4th person	hojiłeeh		dahojiłeeh

PERFECTIVE	singular	duoplural	plural
1st person	hosétɬʔ	hosiidlʔ	dahosiidlʔ
2nd person	hosíníɬʔ	hosoołʔ	dahosoołʔ
3rd person	hasɬʔ		dahasɬʔ
4th person	hojisɬʔ		dahojisɬʔ

University of Colorado **Boulder**

UNIVERSITY OF GOTHENBURG

Summary

An LCS-based method for inferring paradigmatic behavior yields competitive generalizations when coupled with a discriminative classifier

Relatively easy to implement - model is human readable

Fairly language-independent approach (gives paradigms that capture infixation, templatic processes, etc.)

Thank you

Code and language data at:

https://svn.spraakbanken.gu.se/clt/naacl/2015/extract

Stand-alone paradigm extractor tool:

http://pextract.googlecode.com