



Learning Transducer Models for Morphological Analysis from Example Inflections

Markus Forsberg
Mans Hulden

StatFSM
Aug 12, 2016



Morphological Analyzers



Morphological Analyzers

runs ————— word form

Morphological Analyzers

runs ————— word form



Morph.Analyzer



Morphological Analyzers

runs ————— word form

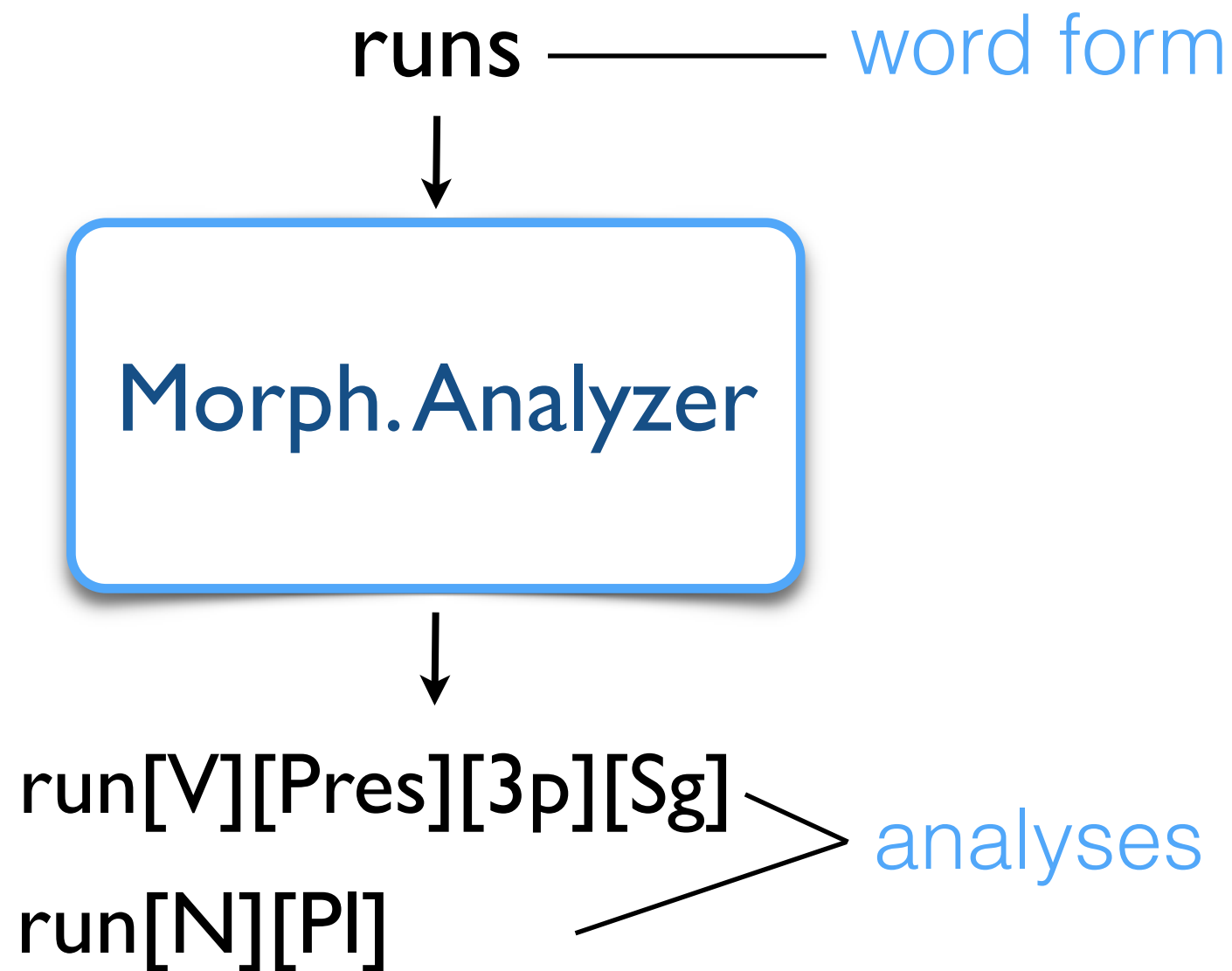


Morph.Analyzer



run[V][Pres][3p][Sg]
run[N][PI] > analyses

Morphological Analyzers



- Useful for NLP
- Necessary for many tasks
- Difficult to construct



Morphological Guessers



Morphological Guessers

blargashes — word form

Morphological Guessers

blargashes — word form



Morph.Analyzer



Morphological Guessers

blargashes — word form



Morph.Analyzer



blargashe[V][3p][Sg]

blargash[V][3p][Sg]

blargash[N][PI]

...

analyses



FSTs from grammars

FSTs from grammars

Grammatical Description

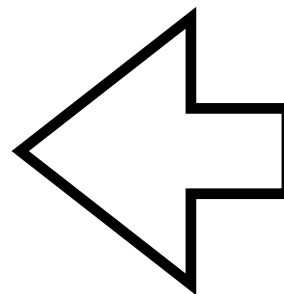
```
def GlideFormation i -> j, ĩ -> j̃, ɪ -> j, ã -> j̃ || V _ V ;
def GlideEpenthesis [..] -> j || V V _ V ;
def Nasalization a -> ã, ɛ -> ě, ɪ -> ĩ, j -> j̃, r -> n // [Nas-C] _ ;
def Lateralization r -> l || \l l _ ;
def Degemination ll -> l || V C* V _ ;
def Hardening r -> d || Nas _ ;
def rMerger r r -> t ;
def VLowering i -> e, ɪ -> ɛ, ʊ -> ɔ, u -> o || _ [e|ɛ] ;
def eDeletion [ɛ | ě] -> 0 || [ɛ | ě] _ .# . ;
def VowelEpenthesis [..] -> X || C _ .# . ;
def LabialHarmony X -> ʊ || LabRnd C* _ .o. X -> ɪ ;
```

...

FSTs from grammars

Grammatical Description

compile



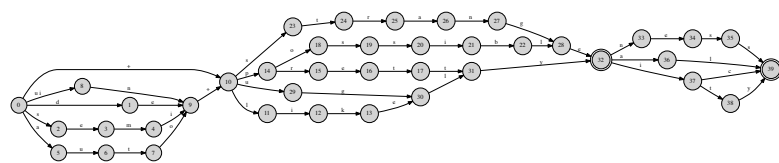
```
def GlideFormation i -> j, ĩ -> j̃, ɪ -> j, ï -> j̃ || V _ V ;
def GlideEpenthesis [..] -> j || V V _ V ;
def Nasalization a -> ã, ɛ -> ě, ɪ -> ĩ, j -> j̃, r -> n // [Nas-C] _ ;
def Lateralization r -> l || \l l _ ;
def Degemination ll -> l || V C* V _ ;
def Hardening r -> d || Nas _ ;
def rMerger r r -> t ;
def VLowering i -> e, ɪ -> ɛ, ʊ -> ɔ, u -> o || _ [e|ɛ] ;
def eDeletion [ɛ | ě] -> 0 || [ɛ | ě] _ .# . ;
def VowelEpenthesis [..] -> X || C _ .# . ;
def LabialHarmony X -> ʊ || LabRnd C* _ .o. X -> ɪ ;
```

...

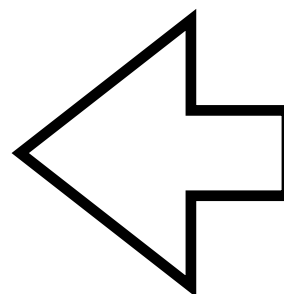
FSTs from grammars

Grammatical Description

FST



compile



```
def GlideFormation i -> j, ĩ -> ĵ, ɪ -> j, ȩ -> ĵ || V _ V ;
def GlideEpenthesis [..] -> j || V V _ V ;
def Nasalization a -> ā, ɛ -> ĕ, ɪ -> ĩ, j -> ĵ, r -> n // [Nas-C] _ ;
def Lateralization r -> l || \l l _ ;
def Degemination ll -> l || V C* V _ ;
def Hardening r -> d || Nas _ ;
def rMerger rr -> t ;
def VLowering i -> e, ɪ -> ɛ, ʊ -> ɔ, u -> o || _ [e|ɛ] ;
def eDeletion [ɛ | ĕ] -> 0 || [ɛ | ĕ] _ .# . ;
def VowelEpenthesis [..] -> X || C _ .# . ;
def LabialHarmony X -> ʊ || LabRnd C* _ .o. X -> ɪ ;
```

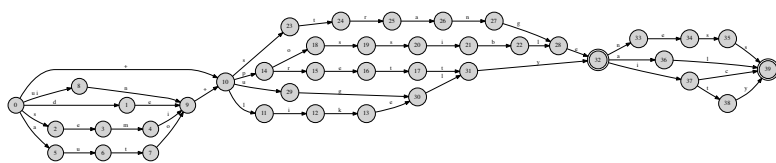
...

FSTs from grammars

runs



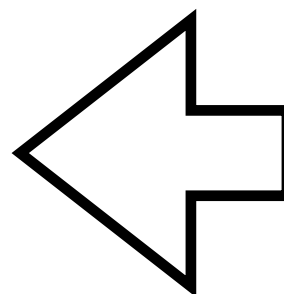
FST



run[V][3p][Sg]

run[N][PI]

compile



Grammatical Description

```
def GlideFormation i -> j, ĩ -> ĵ, ɪ -> j, ȩ -> ĵ || V _ V ;
def GlideEpenthesis [..] -> j || V V _ V ;
def Nasalization a -> ā, ɛ -> ĕ, ɪ -> ĩ, j -> ĵ, r -> n // [Nas-C] _ ;
def Lateralization r -> l || \l l _ ;
def Degemination ll -> l || V C* V _ ;
def Hardening r -> d || Nas _ ;
def rMerger rr -> t ;
def VLowering i -> e, ɪ -> ɛ, ʊ -> ɔ, u -> o || _ [e|ɛ] ;
def eDeletion [ɛ | ĕ] -> 0 || [ɛ | ĕ] _ .# . ;
def VowelEpenthesis [..] -> X || C _ .# . ;
def LabialHarmony X -> ʊ || LabRnd C* _ .o. X -> ɪ ;
```

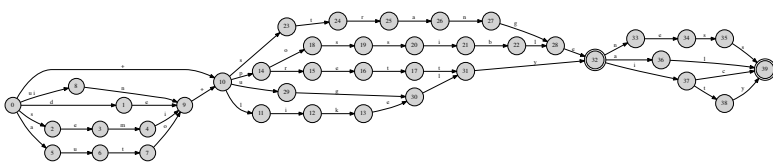
...

FSTs from grammars

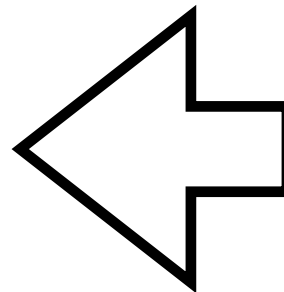
runs



FST



compile



Grammatical Description

```
def GlideFormation i->j, ĩ->ĩ, I->j, Ñ->Ñ || V _ V ;
def GlideEpenthesis [..]->j || V V _ V ;
def Nasalization    a->ã, ε->ẽ, I->Ĩ, i->ĩ as-C] _ ;
def Lateralization  r->l || C _ # . ;
def Desonorantize  n->[] || C _ # . ;
de
def
def
def
def
def
def
def
```

Complex!

Complex!

...

run[V][3p][Sg]

run[N][PI]

FSTs from examples

FSTs from examples

Inflection examples

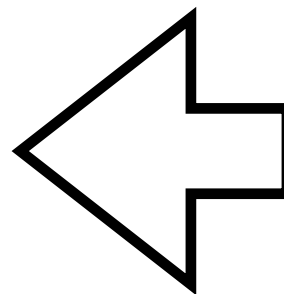
	Indicative	
present	ich schreibe	wir schreiben
	du schreibst	ihr schreibt
	er schreibt	sie schreiben
preterite	ich schrieb	wir schrieben
	du schriebst	ihr schriebt
	er schrieb	sie schrieben
imperative	schreib (du) schreibe (du)	schreibt (ihr)



FSTs from examples

Inflection examples

compile

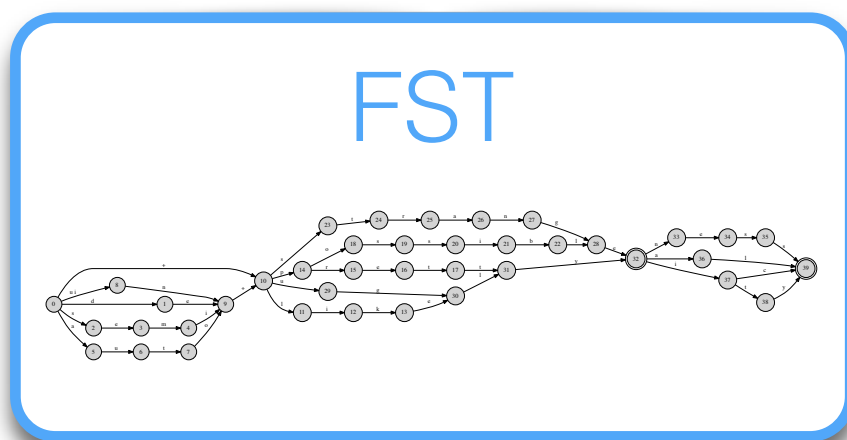


	indicative	
present	ich schreibe	wir schreiben
	du schreibst	ihr schreibt
	er schreibt	sie schreiben
preterite	ich schrieb	wir schrieben
	du schriebst	ihr schriebt
	er schrieb	sie schrieben
imperative	schreib (du) schreibe (du)	schreibt (ihr)

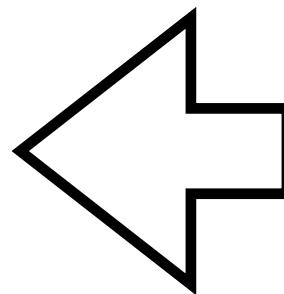


FSTs from examples

Inflection examples



compile



	Indicative	
present	ich schreibe	wir schreiben
	du schreibst	ihr schreibt
	er schreibt	sie schreiben
preterite	ich schrieb	wir schrieben
	du schriebst	ihr schriebt
	er schrieb	sie schrieben
imperative	schreib (du) schreibe (du)	schreibt (ihr)



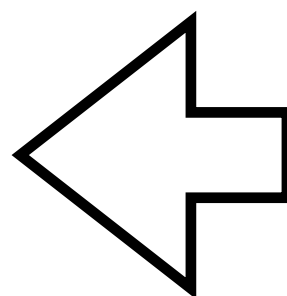
FSTs from examples

schreibe



FST

compile



Inflection examples

	Indicative	
present	ich schreibe	wir schreiben
	du schreibst	ihr schreibt
	er schreibt	sie schreiben
preterite	ich schrieb	wir schrieben
	du schriebst	ihr schriebt
	er schrieb	sie schrieben
imperative	schreib (du)	schreibt (ihr)
	schreibe (du)	



schreiben[V][Pres][Ind][Ip][Sg]



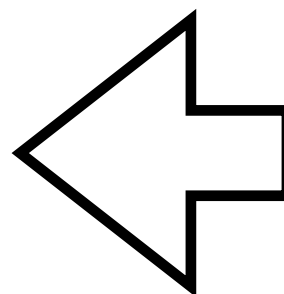
Guessers from examples

blarge



FST

convert



Inflection examples

	Indicative	
present	ich schreibe	wir schreiben
	du schreibst	ihr schreibt
	er schreibt	sie schreiben
preterite	ich schrieb	wir schrieben
	du schriebst	ihr schriebt
	er schrieb	sie schrieben
imperative	schreib (du)	schreibt (ihr)
	schreibe (du)	

blargen[V][Pres][Ind][Ip][Sg]



Outline

(1) Paradigmatic model

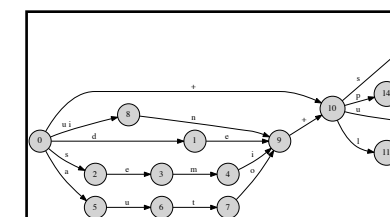
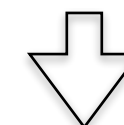
$x_1 \mathbf{i} x_2$
 $x_1 \mathbf{a} x_2$
 $x_1 \mathbf{u} x_2$

Outline

(1) Paradigmatic model

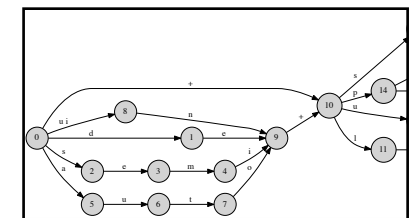
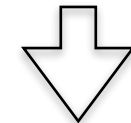
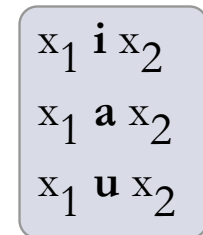
(2) Model \Rightarrow FST analyzer/guesser

$x_1 \text{ i } x_2$
 $x_1 \text{ a } x_2$
 $x_1 \text{ u } x_2$



Outline

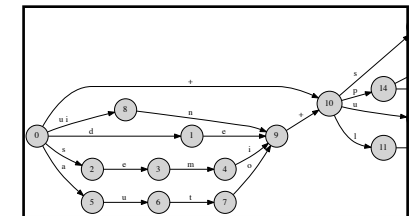
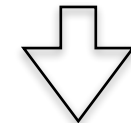
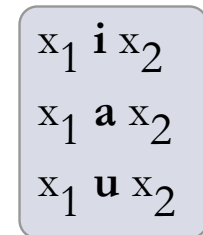
- (1) Paradigmatic model
- (2) Model \Rightarrow FST analyzer/guesser
- (3) Ranking analyses (implementable as WFSTs)



1. blargashe[V][3p][Sg]
2. blargash[V][3p][Sg]
3. blargash[N][PI]

Outline

- (1) Paradigmatic model
- (2) Model \Rightarrow FST analyzer/guesser
- (3) Ranking analyses (implementable as WFSTs)
- (4) Experiments & Results



1. blargashe[V][3p][Sg]
2. blargash[V][3p][Sg]
3. blargash[N][PI]



Generalizing from inflection tables

‘write’

DE

schreiben
schreibend
geschrieben
schreibe
schreibst
schreibt
schreiben
schreibt

...

‘buy’

ES

comprar
comprando
comprado
compro
compras
compra
compramos
compráis

...

‘store’[N]

FI

kauppa
kaupat
kaupan
kauppojen
kauppaa
kauppoja
kaupassa
kaupoissa

...

Generalizing from inflection tables

‘write’

DE

schreiben

schreibend

geschrieben

schreibe

schreibst

schreibt

schreiben

schreibt

...

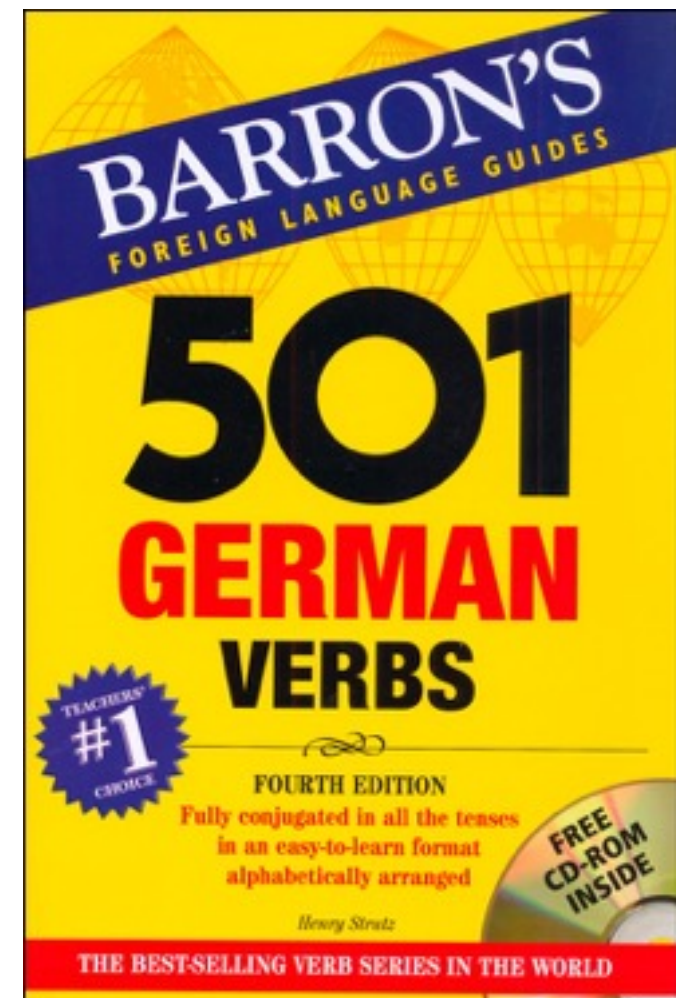
Generalizing from inflection tables

‘write’

DE

schreiben
schreibend
geschrieben
schreibe
schreibst
schreibt
schreiben
schreibt

...



Generalizing from inflection tables

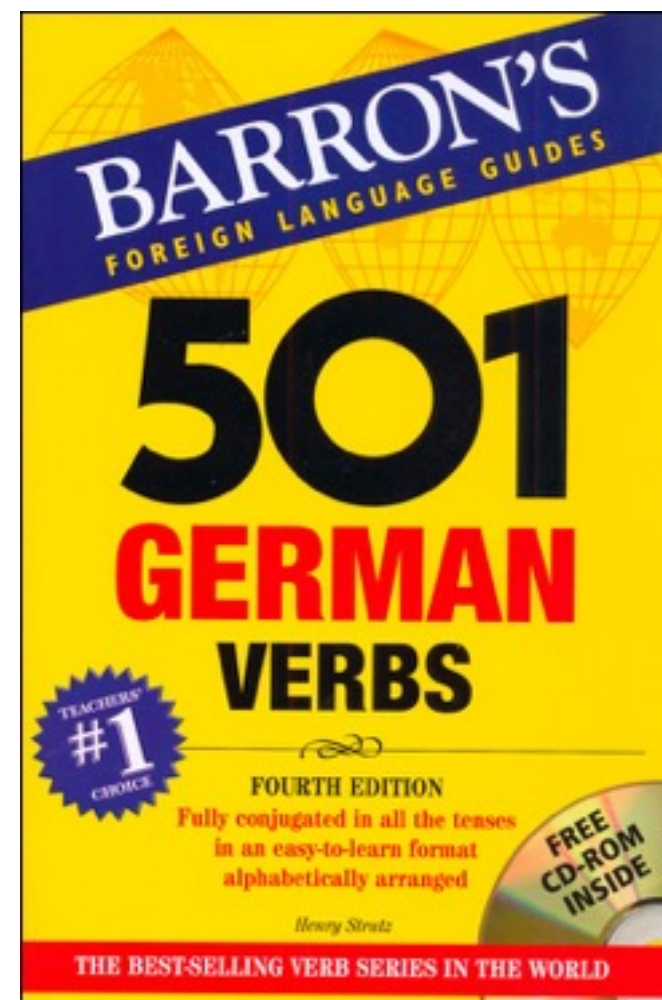
“leihen inflects like schreiben”

‘write’

DE

schreiben
schreibend
geschrieben
schreibe
schreibst
schreibt
schreiben
schreibt

...



Generalizing from inflection tables

“leihen inflects like schreiben”

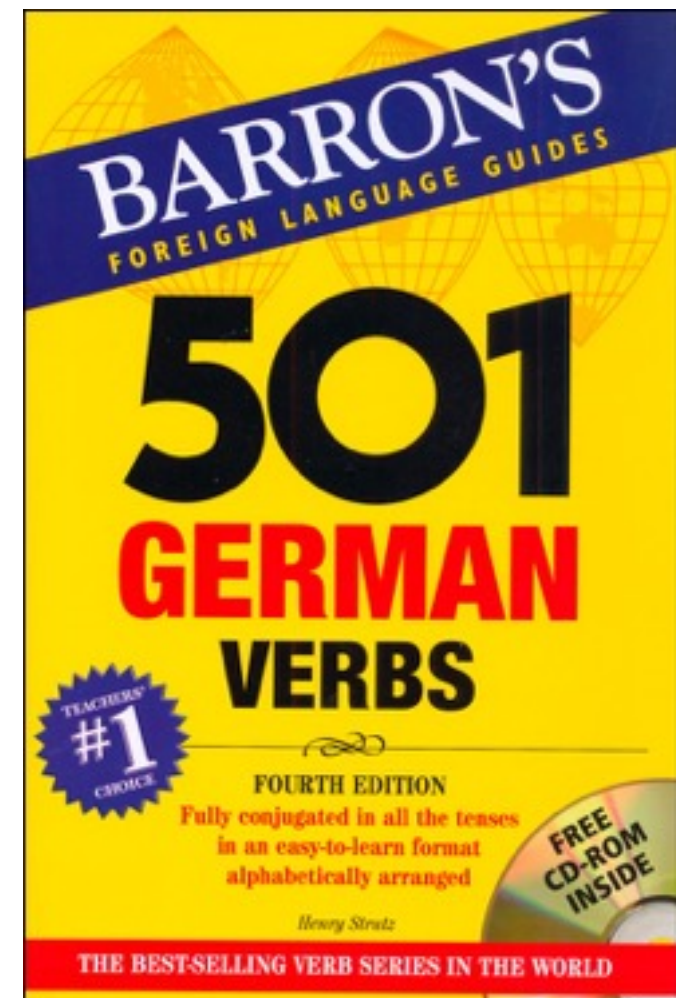
‘write’

DE

schreiben
schreibend
geschrieben
schreibe
schreibst
schreibt
schreiben
schreibt

...

If **leihen** (to lend) is conjugated like **schreiben** (to write), what are the forms?



Linguistic intuition

"Car la notion du thème est postérieure aux formes concrètes composant le paradigme: on trouve le thème en dégageant les éléments communs à toutes les formes casuelles du paradigme (quand il s'agit de la déclinaison)"

[For the notion of the stem is dependent on the concrete forms composing the paradigm: one **finds the stem** in **disengaging the elements that are common to all the case forms** of a paradigm (when dealing with declension)].

Kuryłowicz (1949 p. 159)

Generalization

Formal claim: the common parts (stem) are calculated by extracting the **Longest Common Subsequence** from related forms*

inflection table

schreiben
schreibend
geschrieben
schreibe
schreibst
schreibt

“to write” (German)

*Ahlberg, Forsberg, Hulden (2014, 2015)

Generalization

Formal claim: the common parts (stem) are calculated by extracting the **Longest Common Subsequence** from related forms*

inflection table

schrei	ben	
schrei	bend	
geschr	ieben	
schrei	be	
schrei	b	st
schrei	b	t

*Ahlberg, Forsberg, Hulden (2014, 2015)

Generalization

Formal claim: the common parts (stem) are calculated by extracting the **Longest Common Subsequence** from related forms*

inflection table

LCS = **schrib**

schrei	ben	
schrei	bend	
geschr	ieben	
schrei	be	
schrei	b	st
schrei	b	t

*Ahlberg, Forsberg, Hulden (2014, 2015)

Generalization

Formal claim: the common parts (stem) are calculated by extracting the **Longest Common Subsequence** from related forms*

inflection table	<u>LCS = schrib</u>
schrei ben	$x_1 = \text{schr}$
schrei bend	$x_2 = \text{i}$
geschr ieben	$x_3 = \text{b}$
schrei be	
schrei b	st
schrei b	t
<hr/> x_1 x_2 x_3	

*Ahlberg, Forsberg, Hulden (2014, 2015)

Generalization

Formal claim: the common parts (stem) are calculated by extracting the **Longest Common Subsequence** from related forms*

inflection table	<u>LCS = schrib</u>	“paradigm”
schrei ben	$x_1 = \text{schr}$	$x_1 + \mathbf{e} + x_2 + x_3 + \mathbf{en}$
schrei bend	$x_2 = \mathbf{i}$	$x_1 + \mathbf{e} + x_2 + x_3 + \mathbf{end}$
geschr ieben	$x_3 = \mathbf{b}$	$\mathbf{ge} + x_1 + x_2 + \mathbf{e} + x_3 + \mathbf{en}$
schrei be		$x_1 + \mathbf{e} + x_2 + x_3 + \mathbf{e}$
schrei b	st	$x_1 + \mathbf{e} + x_2 + x_3 + \mathbf{st}$
schrei b	t	$x_1 + \mathbf{e} + x_2 + x_3 + \mathbf{t}$
<u> </u> <u> </u> <u> </u>		
x_1 x_2 x_3		

*Ahlberg, Forsberg, Hulden (2014, 2015)

Generalization

Formal claim: the common parts (stem) are calculated by extracting the **Longest Common Subsequence** from related forms*

inflection table	<u>LCS = schrib</u>	“paradigm”
schrei ben	$x_1 = \text{schr}$	$x_1 + \mathbf{e} + x_2 + x_3 + \mathbf{en}$
schrei bend	$x_2 = \mathbf{i}$	$x_1 + \mathbf{e} + x_2 + x_3 + \mathbf{end}$
geschr ieben	$x_3 = \mathbf{b}$	$\mathbf{ge} + x_1 + x_2 + \mathbf{e} + x_3 + \mathbf{en}$
schrei be		$x_1 + \mathbf{e} + x_2 + x_3 + \mathbf{e}$
schrei b	st	$x_1 + \mathbf{e} + x_2 + x_3 + \mathbf{st}$
schrei b	t	$x_1 + \mathbf{e} + x_2 + x_3 + \mathbf{t}$
<u> </u> <u> </u> <u> </u>		
x_1 x_2 x_3		

*Ahlberg, Forsberg, Hulden (2014, 2015)

Generalization

“paradigm”

$x_1 + \mathbf{e} + x_2 + x_3 + \mathbf{en}$

$x_1 + \mathbf{e} + x_2 + x_3 + \mathbf{end}$

$\mathbf{ge} + x_1 + x_2 + \mathbf{e} + x_3 + \mathbf{en}$

$x_1 + \mathbf{e} + x_2 + x_3 + \mathbf{e}$

$x_1 + \mathbf{e} + x_2 + x_3 + \mathbf{st}$

$x_1 + \mathbf{e} + x_2 + x_3 + \mathbf{t}$

$x_1 = \mathbf{l}$

$x_2 = \mathbf{i}$

$x_3 = \mathbf{h}$



leihen

leihend

geliehen

leihe

leihst

leiht

Generalization

Generalization

Inflection tables

Generalization

Inflection tables

swim

swam

swum

jump

jumped

jumped

ring

rang

rung

Generalization

Inflection tables

swim
swam
swum



jump
jumped
jumped



ring
rang
rung



Generalization

Inflection tables

Paradigms

swim
swam
swum



$x_1 + i + x_2$
 $x_1 + a + x_2$
 $x_1 + u + x_2$

jump
jumped
jumped



x_1
 $x_1 + ed$
 $x_1 + ed$

ring
rang
rung



$x_1 + i + x_2$
 $x_1 + a + x_2$
 $x_1 + u + x_2$

Generalization

Inflection tables

Paradigms

swim
swam
swum



$x_1 + i + x_2$
 $x_1 + a + x_2$
 $x_1 + u + x_2$

jump
jumped
jumped



x_1
 $x_1 + ed$
 $x_1 + ed$

ring
rang
rung



$x_1 + i + x_2$
 $x_1 + a + x_2$
 $x_1 + u + x_2$

Generalization

Inflection tables

Paradigms

Collapsed

swim
swam
swum



jump
jumped
jumped



ring
rang
rung



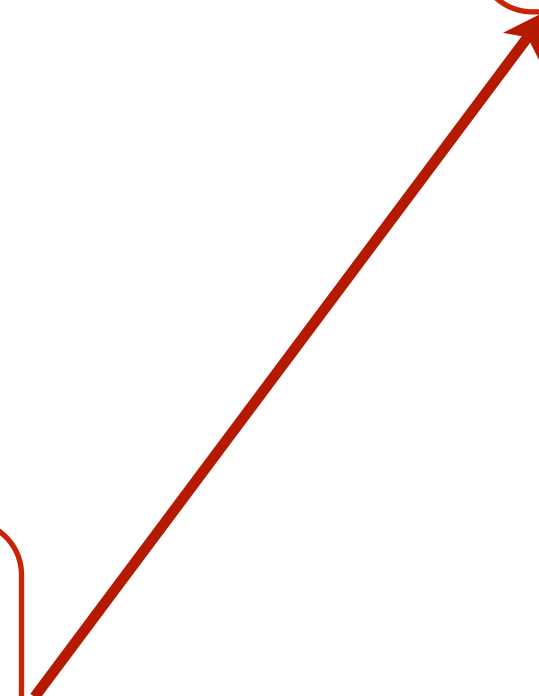
$x_1 + i + x_2$
 $x_1 + a + x_2$
 $x_1 + u + x_2$



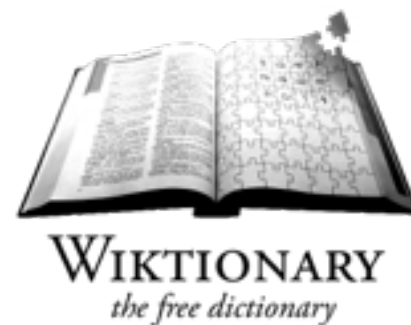
$x_1 + i + x_2$
 $x_1 + a + x_2$
 $x_1 + u + x_2$

x_1
 $x_1 + ed$
 $x_1 + ed$

$x_1 + i + x_2$
 $x_1 + a + x_2$
 $x_1 + u + x_2$



Generalization



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

From paradigm to FST

Sp. 'reconcile'

Lemmatization

From paradigm to FST

Sp. 'reconcile'

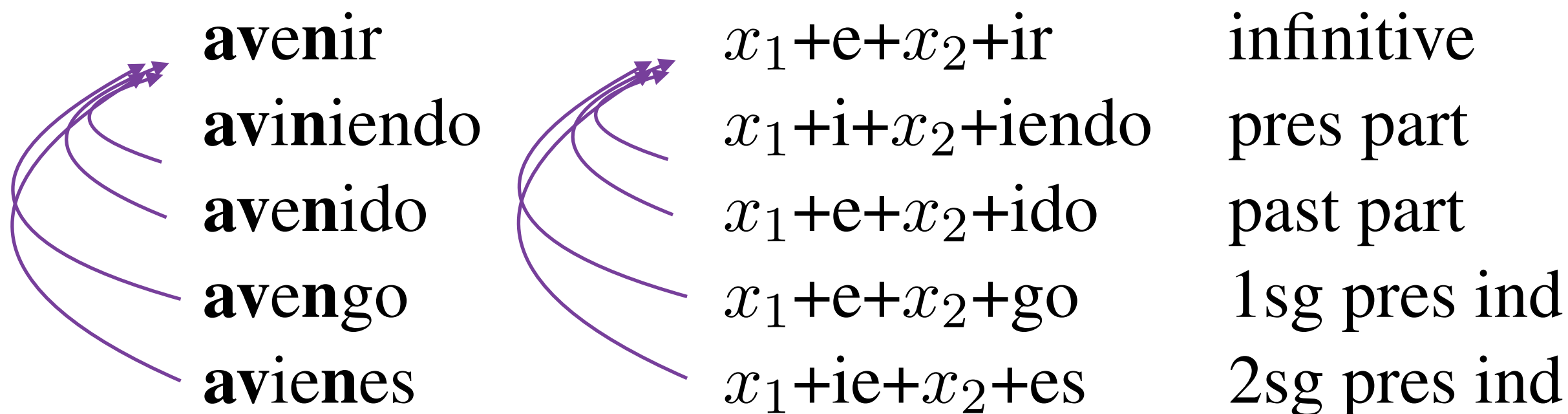
Lemmatization

avenir	$x_1 + e + x_2 + ir$	infinitive
aviniendo	$x_1 + i + x_2 + iendo$	pres part
avenido	$x_1 + e + x_2 + ido$	past part
avengo	$x_1 + e + x_2 + go$	1sg pres ind
avienes	$x_1 + ie + x_2 + es$	2sg pres ind

From paradigm to FST

Sp. 'reconcile'



Lemmatization




From paradigm to FST


Sp. 'reconcile'

Lemmatization

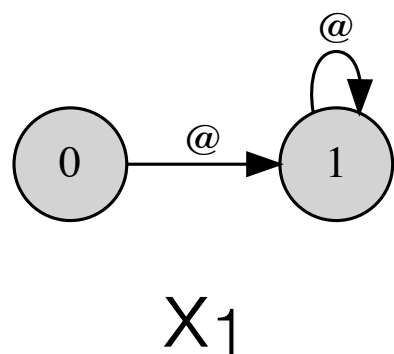
	avenir		$x_1 + e + x_2 + ir$	infinitive
	aviniendo		$x_1 + i + x_2 + iendo$	pres part
	avenido		$x_1 + e + x_2 + ido$	past part
	avengo		$x_1 + e + x_2 + go$	1sg pres ind
	avienes		$x_1 + ie + x_2 + es$	2sg pres ind

From paradigm to FST



avenir
aviniendo
avenido
avengo
avienes



 $x_1 + e + x_2 + ir$
 $x_1 + i + x_2 + iendo$
 $x_1 + e + x_2 + ido$
 $x_1 + e + x_2 + go$
 $x_1 + ie + x_2 + es$

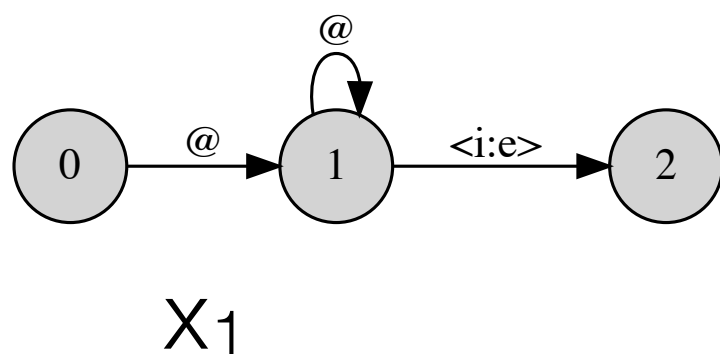
infinitive
 pres part
 past part
 1sg pres ind
 2sg pres ind




From paradigm to FST



avenir
aviniendo
avenido
avengo
avienes

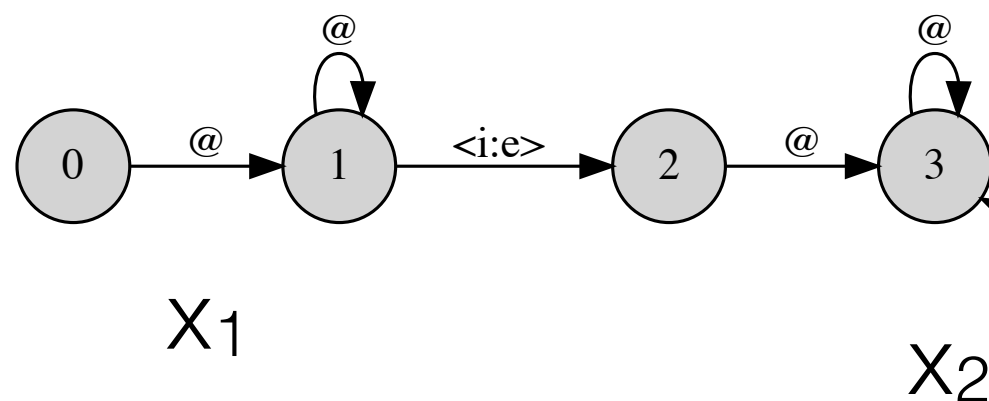

 $x_1 + e + x_2 + ir$ infinitive
 $x_1 + i + x_2 + iendo$ pres part
 $x_1 + e + x_2 + ido$ past part
 $x_1 + e + x_2 + go$ 1sg pres ind
 $x_1 + ie + x_2 + es$ 2sg pres ind



From paradigm to FST

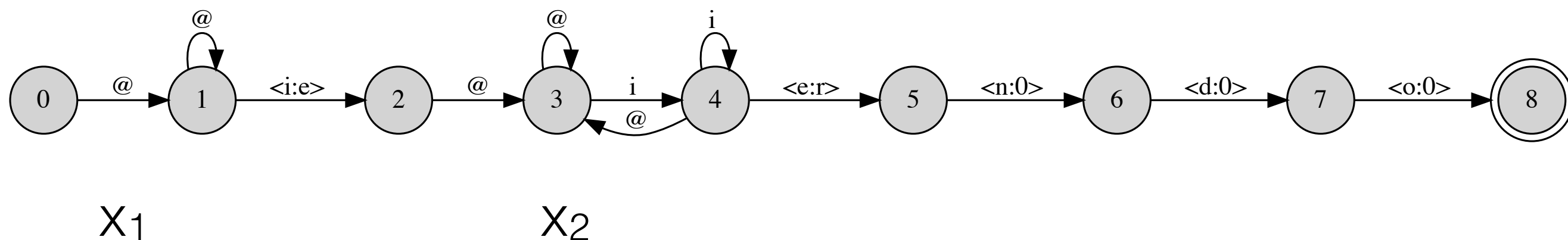

avenir
aviniendo
avenido
avengo
avienes


 $x_1 + e + x_2 + ir$ infinitive
 $x_1 + i + x_2 + iendo$ pres part
 $x_1 + e + x_2 + ido$ past part
 $x_1 + e + x_2 + go$ 1sg pres ind
 $x_1 + ie + x_2 + es$ 2sg pres ind



From paradigm to FST

avenir	$x_1 + e + x_2 + ir$	infinitive
aviniendo	$x_1 + i + x_2 + iendo$	pres part
avenido	$x_1 + e + x_2 + ido$	past part
avengo	$x_1 + e + x_2 + go$	1sg pres ind
avienes	$x_1 + ie + x_2 + es$	2sg pres ind



From paradigm to FST

↪ **avenir**
 ↪ **aviniendo**
avenido
avengo
avienes

$x_1 + e + x_2 + ir$

$x_1 + i + x_2 + iendo$

$x_1 + e + x_2 + ido$

$x_1 + e + x_2 + go$

$x_1 + ie + x_2 + es$

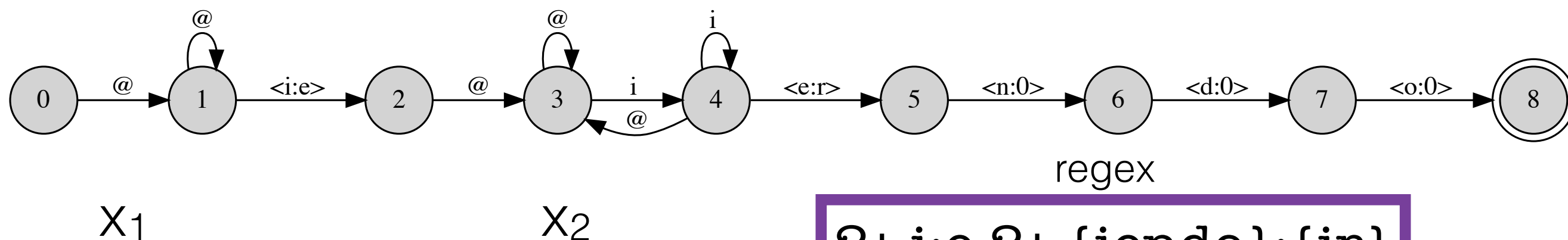
infinitive

pres part

past part

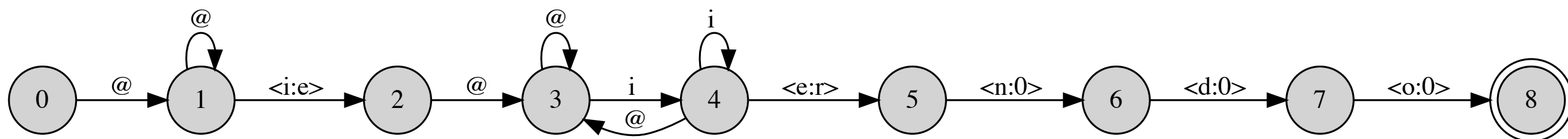
1sg pres ind

2sg pres ind

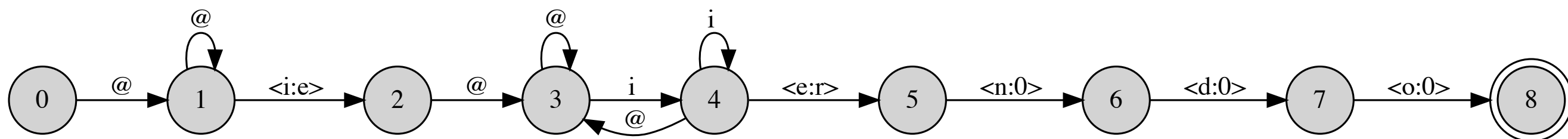


?+ i:e ?+ {iendo}:{ir}

From paradigm to FST

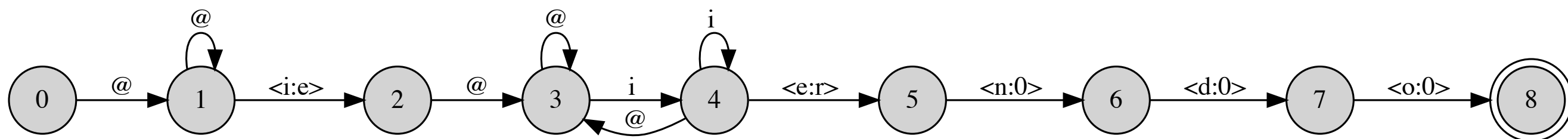


From paradigm to FST



aviniendo > avenir

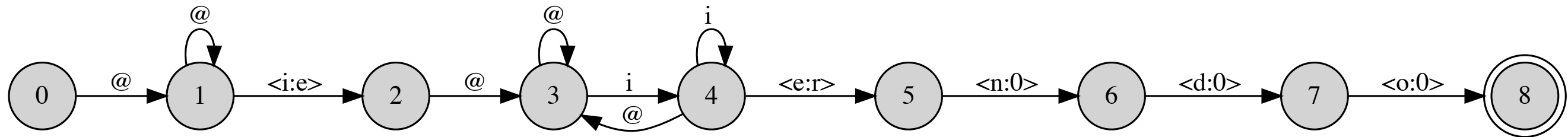
From paradigm to FST



aviniendo > avenir

viniendo > venir

From paradigm to FST

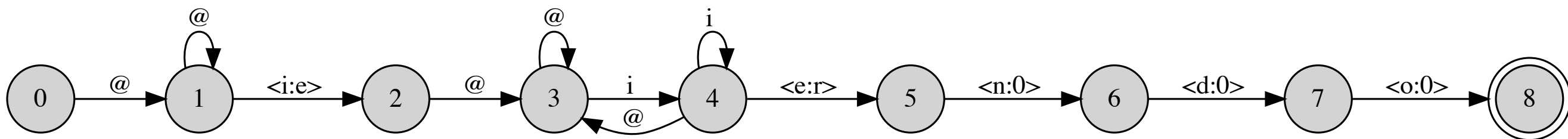


aviniendo > avenir

viniendo > venir

proviniendo > provenir

From paradigm to FST



aviniendo > avenir

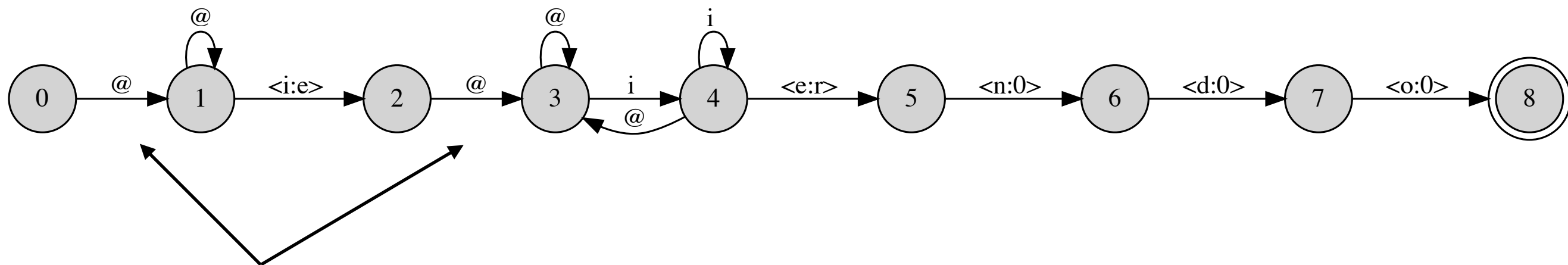
viniendo > venir

proviniendo > provenir

viviendo > *vevir

should be vivir

From paradigm to FST



too general!

aviniendo > avenir

viniendo > venir

proviniendo > provenir

viviendo > *vevir

should be vivir

From paradigm to FST

avenir	$x_1 + e + x_2 + ir$	infinitive
aviniendo	$x_1 + i + x_2 + iendo$	pres part
avenido	$x_1 + e + x_2 + ido$	past part
avengo	$x_1 + e + x_2 + go$	1sg pres ind
avienes	$x_1 + ie + x_2 + es$	2sg pres ind

From paradigm to FST

avenir	$x_1 + e + x_2 + ir$	infinitive
aviniendo	$x_1 + i + x_2 + iendo$	pres part
avenido	$x_1 + e + x_2 + ido$	past part
avengo	$x_1 + e + x_2 + go$	1sg pres ind
avienes	$x_1 + ie + x_2 + es$	2sg pres ind

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

From paradigm to FST

avenir	$x_1 + e + x_2 + ir$	infinitive
aviniendo	$x_1 + i + x_2 + iendo$	pres part
avenido	$x_1 + e + x_2 + ido$	past part
avengo	$x_1 + e + x_2 + go$	1sg pres ind
avienes	$x_1 + ie + x_2 + es$	2sg pres ind

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

From paradigm to FST

av nir	$x_1 + e + x_2 + ir$	infinitive
av iniendo	$x_1 + i + x_2 + iendo$	pres part
av nido	$x_1 + e + x_2 + ido$	past part
av engo	$x_1 + e + x_2 + go$	1sg pres ind
av ienes	$x_1 + ie + x_2 + es$	2sg pres ind

“ends in **v**”

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

From paradigm to FST

av nir	$x_1 + e + x_2 + ir$	infinitive
av iniendo	$x_1 + i + x_2 + iendo$	pres part
av nido	$x_1 + e + x_2 + ido$	past part
av engo	$x_1 + e + x_2 + go$	1sg pres ind
av ienes	$x_1 + ie + x_2 + es$	2sg pres ind

“ends in **v**”

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

n
n
n
n
n
n
n
n
n
n
n
n

“is always **n**”

From paradigm to FST

avenir	$x_1 + e + x_2 + ir$	infinitive
aviniendo	$x_1 + i + x_2 + iendo$	pres part
avenido	$x_1 + e + x_2 + ido$	past part
avengo	$x_1 + e + x_2 + go$	1sg pres ind
avienes	$x_1 + ie + x_2 + es$	2sg pres ind

“ends in **v**”

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

n
n
n
n
n
n
n
n
n
n
n
n

“is always **n**”

Add restriction:

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

From paradigm to FST

From paradigm to FST

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

From paradigm to FST

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



regex

n
n
n
n
n
n
n
n
n
n
n
n



regex

From paradigm to FST

av

circunv

contrav

conv

dev

entrev

interv

prev

prov

rev

v

adv



regex

n

n

n

n

n

n

n

n

n

n

n

n



regex



From paradigm to FST

av

circunv

contrav

conv

dev

entrev

interv

prev

prov

rev

v

adv

n

n

n

n

n

n

n

n

n

n

n

n

n



given m tokens of n types, what is the probability of encountering new unseen type?

↓
regex

↓
regex

From paradigm to FST

av

circunv

contrav

conv

dev

entrev

interv

prev

prov

rev

v

adv



regex

n

n

n

n

n

n

n

n

n

n

n

n



regex



given m tokens of n types, what is the probability of encountering new unseen type?

Good (1953), Ogino (1999)
Kaegura and Sekine (1999)

From paradigm to FST

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

n
n
n
n
n
n
n
n
n
n
n
n



given morphological
probabilities

the
seen type?

Details in paper!

Good (1953), Ogino (1999)
Kaegura and Sekine (1999)

↓
regex

↓
regex

From paradigm to FST

Add inflection information

avenir

aviniendo

avenido

avengo

avienes

$x_1 + e + x_2 + ir$

$x_1 + i + x_2 + iendo$

$x_1 + e + x_2 + ido$

$x_1 + e + x_2 + go$

$x_1 + ie + x_2 + es$

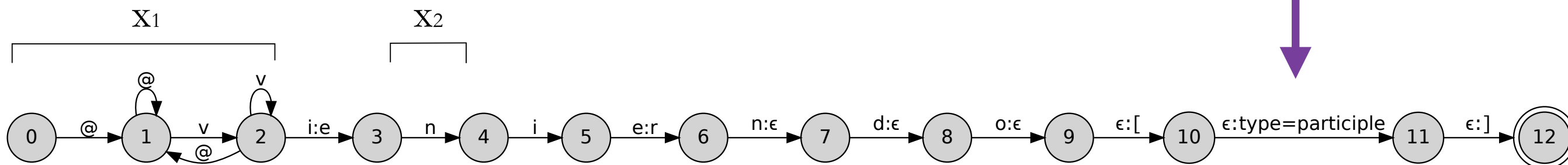
infinitive

pres part

past part

1sg pres ind

2sg pres ind



From paradigm to FST

Add inflection information

avenir

aviniendo

avenido

avengo

avienes

$x_1 + e + x_2 + ir$

$x_1 + i + x_2 + iendo$

$x_1 + e + x_2 + ido$

$x_1 + e + x_2 + go$

$x_1 + ie + x_2 + es$

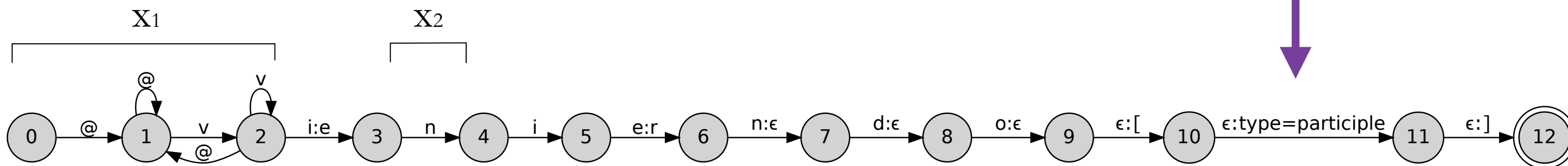
infinitive

pres part

past part

1sg pres ind

2sg pres ind



proviniendo > provenir[type=participle]

From paradigm to FST

Add inflection information

avenir

aviniendo

avenido

avengo

avienes

$x_1 + e + x_2 + ir$

$x_1 + i + x_2 + iendo$

$x_1 + e + x_2 + ido$

$x_1 + e + x_2 + go$

$x_1 + ie + x_2 + es$

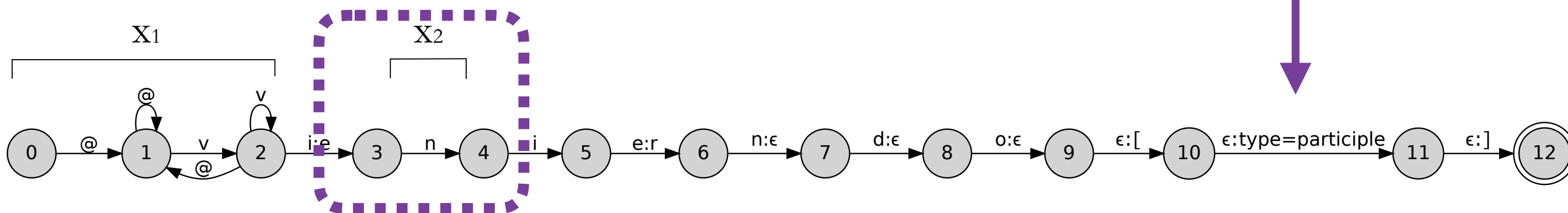
infinitive

pres part

past part

1sg pres ind

2sg pres ind



proviniendo > provenir[type=participle]

viviendo > ∅

Building the analyzer

Building the analyzer

Paradigm

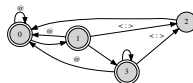
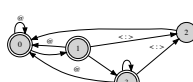
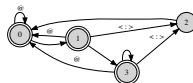
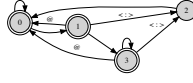
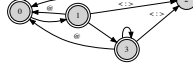
$x_1 + e + x_2 + ir$
 $x_1 + i + x_2 + iendo$
 $x_1 + e + x_2 + ido$
 $x_1 + e + x_2 + go$
 $x_1 + ie + x_2 + es$

Paradigm

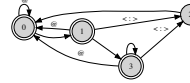
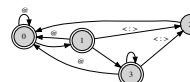
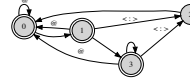
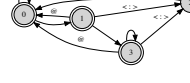

$x_1 + x_2 + ar$
 $x_1 + x_2 + ando$
 $x_1 + x_2 + ado$
 $x_1 + i + x_2 + o$
 $x_1 + i + x_2 + as$

Building the analyzer

Paradigm

$x_1 + e + x_2 + ir$	
$x_1 + i + x_2 + iendo$	
$x_1 + e + x_2 + ido$	
$x_1 + e + x_2 + go$	
$x_1 + ie + x_2 + es$	

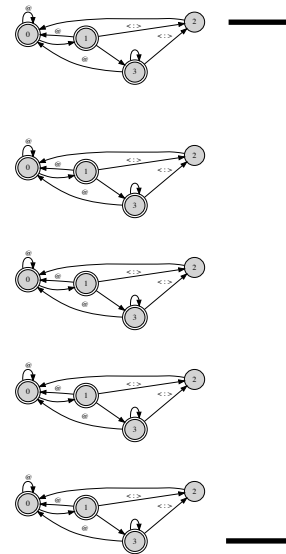
Paradigm

$x_1 + x_2 + ar$	
$x_1 + x_2 + ando$	
$x_1 + x_2 + ado$	
$x_1 + i + x_2 + o$	
$x_1 + i + x_2 + as$	

Building the analyzer

Paradigm

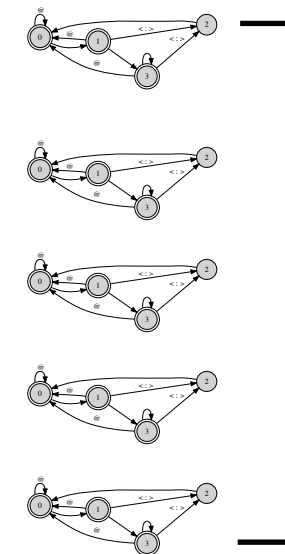
$x_1 + e + x_2 + ir$
 $x_1 + i + x_2 + iendo$
 $x_1 + e + x_2 + ido$
 $x_1 + e + x_2 + go$
 $x_1 + ie + x_2 + es$



analyzers

Paradigm

$x_1 + x_2 + ar$
 $x_1 + x_2 + ando$
 $x_1 + x_2 + ado$
 $x_1 + i + x_2 + o$
 $x_1 + i + x_2 + as$

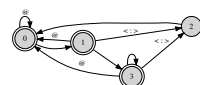
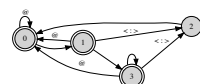
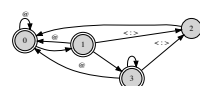
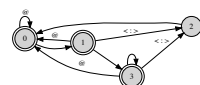
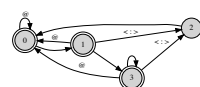


analyzers

Building the analyzer

Paradigm

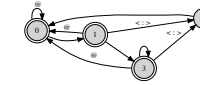
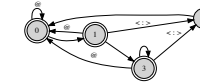
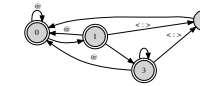
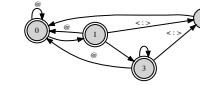
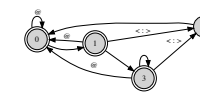
$x_1 + e + x_2 + ir$
 $x_1 + i + x_2 + iendo$
 $x_1 + e + x_2 + ido$
 $x_1 + e + x_2 + go$
 $x_1 + ie + x_2 + es$



analyzers

Paradigm

$x_1 + x_2 + ar$
 $x_1 + x_2 + ando$
 $x_1 + x_2 + ado$
 $x_1 + i + x_2 + o$
 $x_1 + i + x_2 + as$



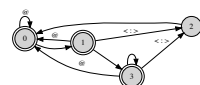
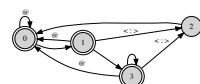
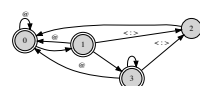
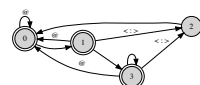
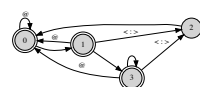
analyzers

m transducers

Building the analyzer

Paradigm

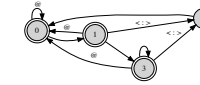
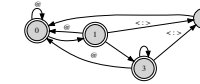
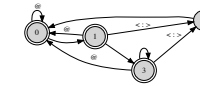
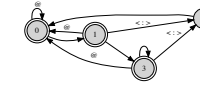
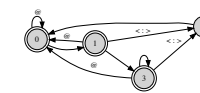
$x_1 + e + x_2 + ir$
 $x_1 + i + x_2 + iendo$
 $x_1 + e + x_2 + ido$
 $x_1 + e + x_2 + go$
 $x_1 + ie + x_2 + es$



analyzers

Paradigm

$x_1 + x_2 + ar$
 $x_1 + x_2 + ando$
 $x_1 + x_2 + ado$
 $x_1 + i + x_2 + o$
 $x_1 + i + x_2 + as$



analyzers

m transducers

$$\text{Analyzer} = f_1 \cup f_2 \cup \dots \cup f_1 \cup \dots \cup f_m$$



Prioritizing analyses

Prioritizing analyses

$x_1 + e + x_2 + \dot{i}r$

$x_1 + \dot{i} + x_2 + \dot{i}endo$

$x_1 + e + x_2 + \dot{i}do$

$x_1 + e + x_2 + go$

$x_1 + \dot{i}e + x_2 + es$

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

Prioritizing analyses

unconstrained

U

$$X_1 = \Sigma +$$

$$X_2 = \Sigma +$$

$x_1 + e + x_2 + ir$

$x_1 + i + x_2 + iendo$

$x_1 + e + x_2 + ido$

$x_1 + e + x_2 + go$

$x_1 + ie + x_2 + es$

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

Prioritizing analyses

unconstrained

constrained

U

C

$$x_1 = \Sigma^+$$

$$x_2 = \Sigma^+$$

$$x_1 = \Sigma^* v$$

$$x_2 = n$$

$x_1 + e + x_2 + ir$

$x_1 + i + x_2 + iendo$

$x_1 + e + x_2 + ido$

$x_1 + e + x_2 + go$

$x_1 + ie + x_2 + es$

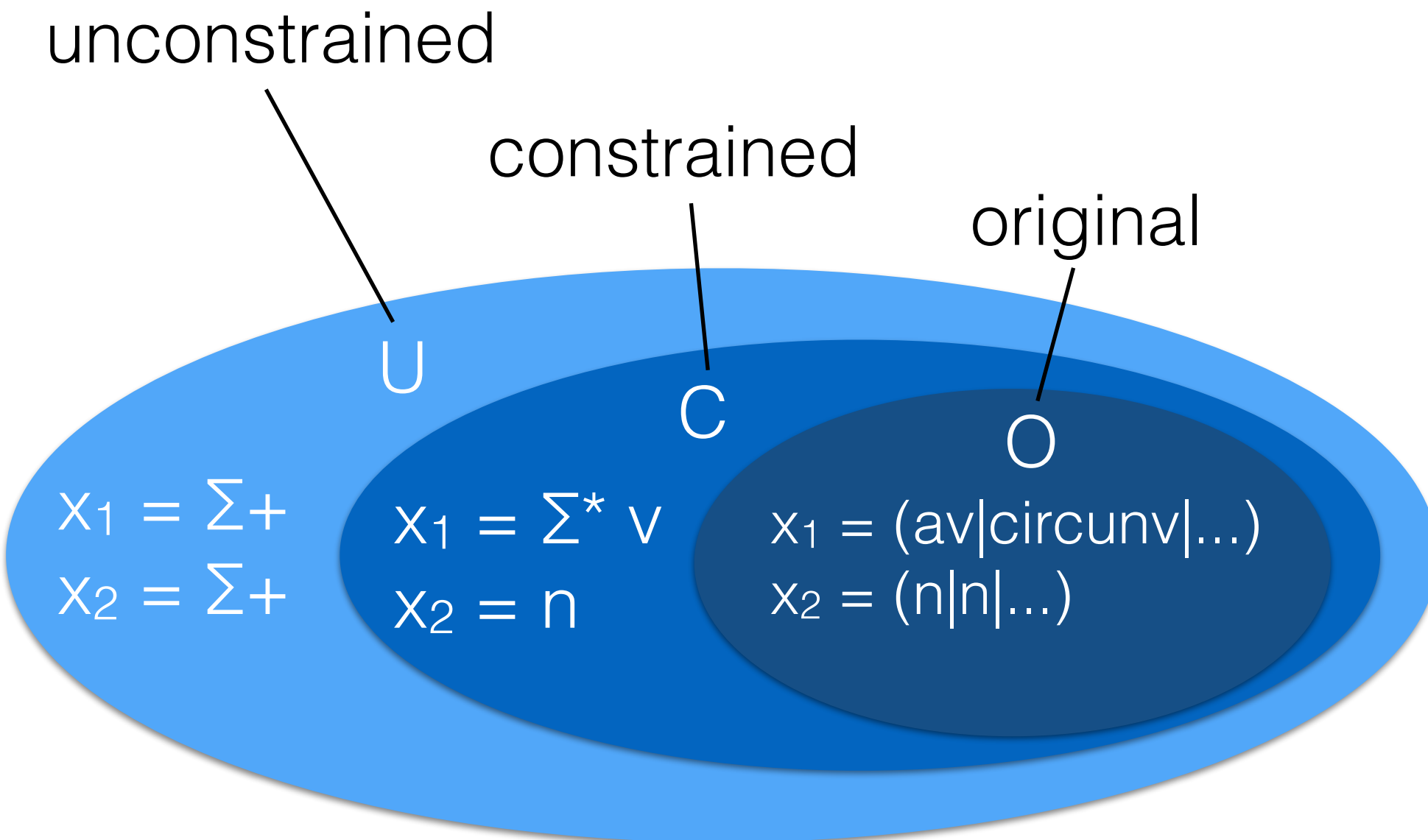
av	n
circunv	n
contrav	n
conv	n
dev	n
entrev	n
interv	n
prev	n
prov	n
rev	n
v	n
adv	n

Prioritizing analyses

unconstrained

constrained

original



$x_1 + e + x_2 + ir$
 $x_1 + i + x_2 + iendo$
 $x_1 + e + x_2 + ido$
 $x_1 + e + x_2 + go$
 $x_1 + ie + x_2 + es$

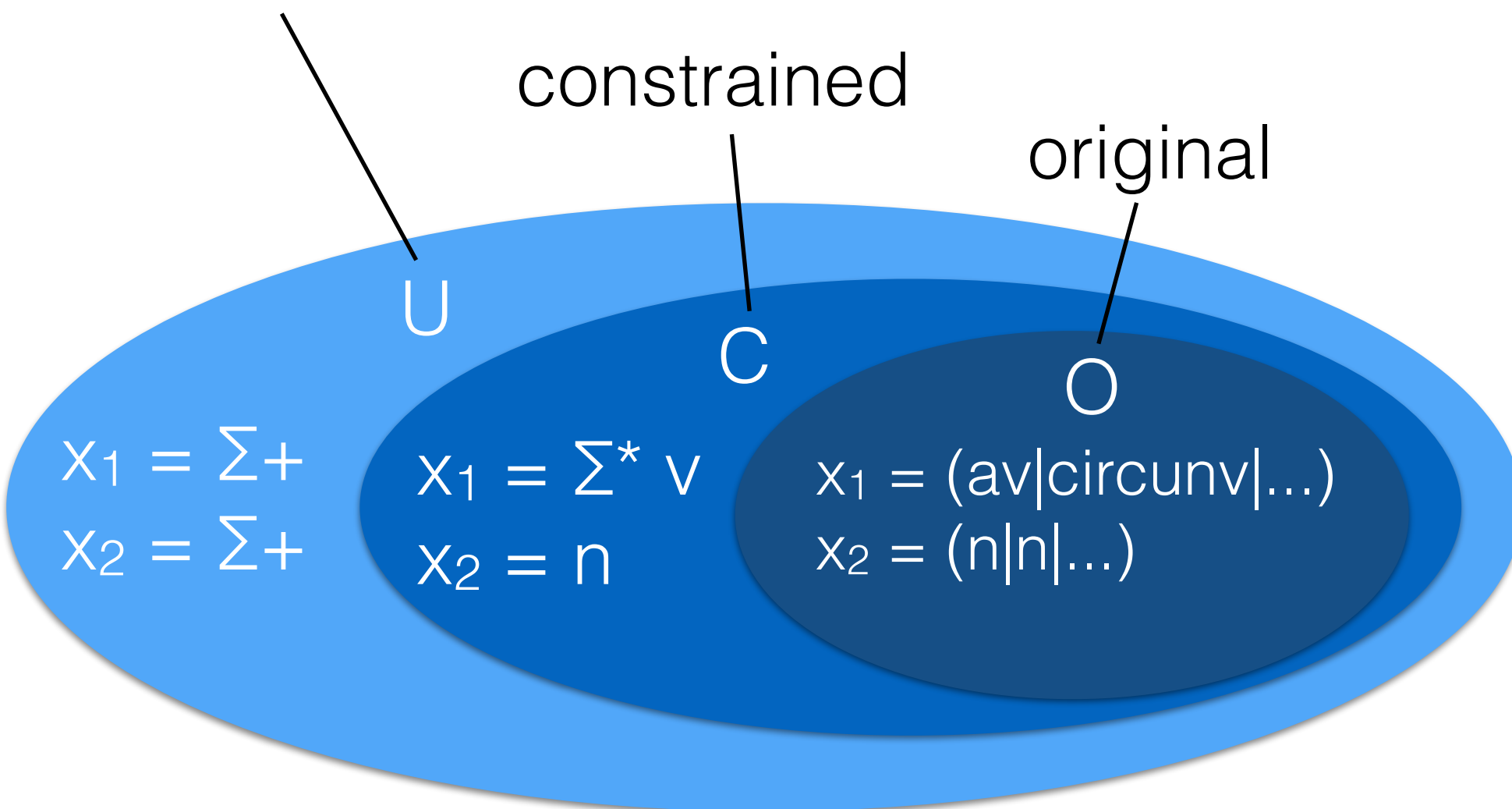
av	n
circunv	n
contrav	n
conv	n
dev	n
entrev	n
interv	n
prev	n
prov	n
rev	n
v	n
adv	n

Prioritizing analyses

unconstrained

constrained

original



Combine with “priority union”:

$x_1 + e + x_2 + ir$
 $x_1 + i + x_2 + iendo$
 $x_1 + e + x_2 + ido$
 $x_1 + e + x_2 + go$
 $x_1 + ie + x_2 + es$

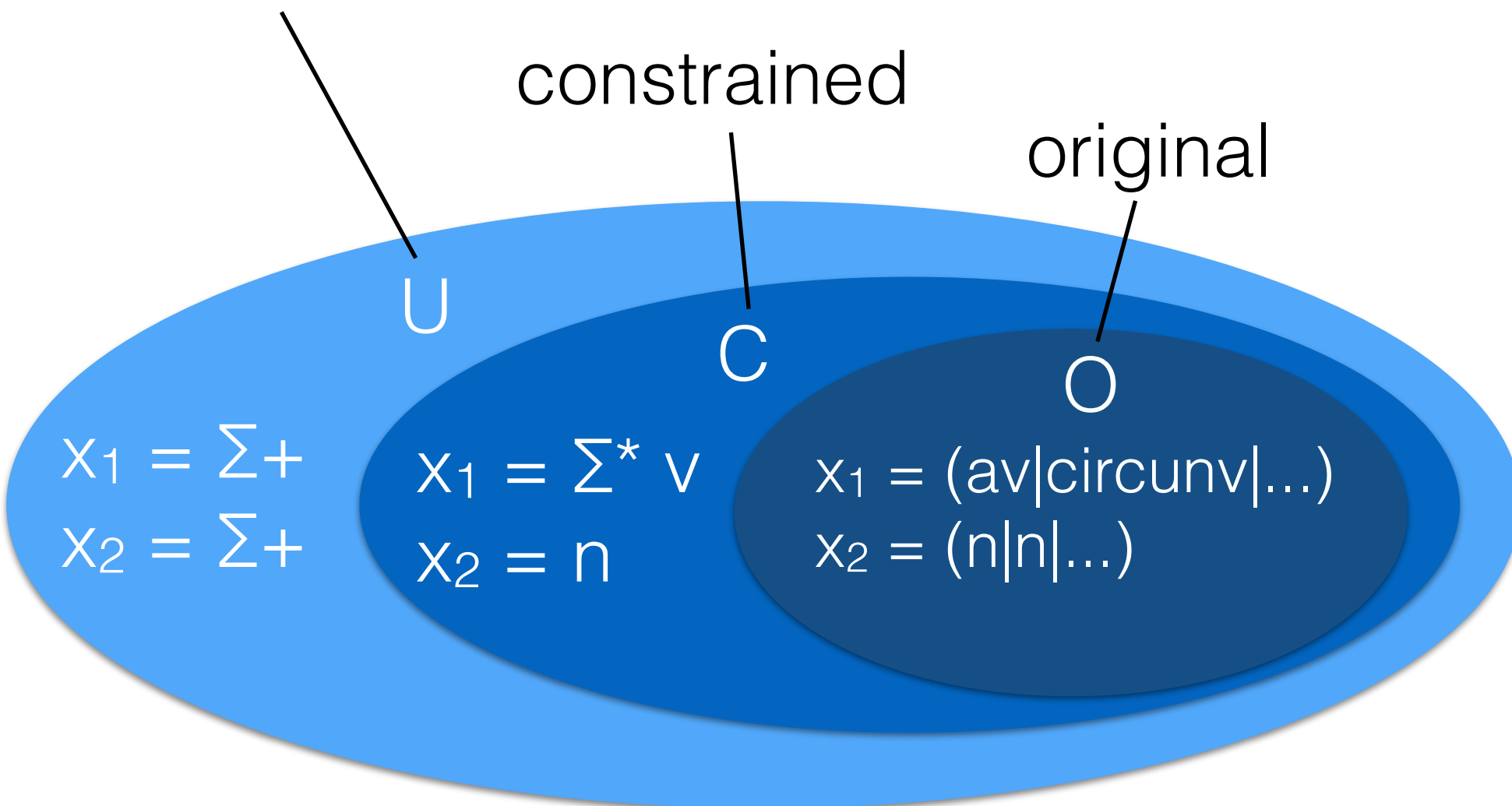
av	n
circunv	n
contrav	n
conv	n
dev	n
entrev	n
interv	n
prev	n
prov	n
rev	n
v	n
adv	n

Prioritizing analyses

unconstrained

constrained

original



Combine with “priority union”:

$$\text{Analyzer} = O \cup_P C \cup_P U$$

$x_1 + e + x_2 + ir$
 $x_1 + i + x_2 + iendo$
 $x_1 + e + x_2 + ido$
 $x_1 + e + x_2 + go$
 $x_1 + ie + x_2 + es$

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

Language models over variables (WFSTs)

avenir	$x_1 + e + x_2 + ir$	infinitive
aviniendo	$x_1 + i + x_2 + iendo$	pres part
avenido	$x_1 + e + x_2 + ido$	past part
avengo	$x_1 + e + x_2 + go$	1sg pres ind
avienes	$x_1 + ie + x_2 + es$	2sg pres ind

“ends in **v**”

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

“is always **n**”

Add restriction:

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

Language models over variables (WFSTs)

avenir
 $x_1 + e + x_2 + ir$

infinitive

aviniendo
 $x_1 + i + x_2 + iendo$

pres part

avenido
 $x_1 + e + x_2 + ido$

past part

avengo
 $x_1 + e + x_2 + go$

1sg pres ind

avienes
 $x_1 + ie + x_2 + es$

2sg pres ind

“ends in **v**”



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

n
n
n
n
n
n
n
n
n
n
n
n

“is always **n**”

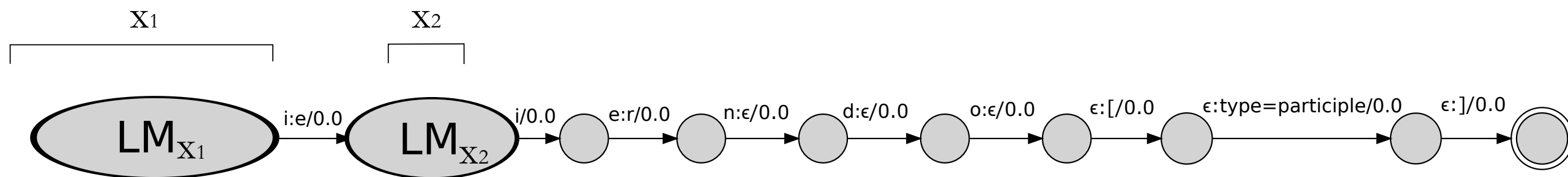
Add
Infer a language model!

From paradigm to FST

	avenir		$x_1 + e + x_2 + ir$	infinitive
	aviniendo		$x_1 + i + x_2 + iendo$	pres part
	avenido		$x_1 + e + x_2 + ido$	past part
	avengo		$x_1 + e + x_2 + go$	1sg pres ind
	avienes		$x_1 + ie + x_2 + es$	2sg pres ind

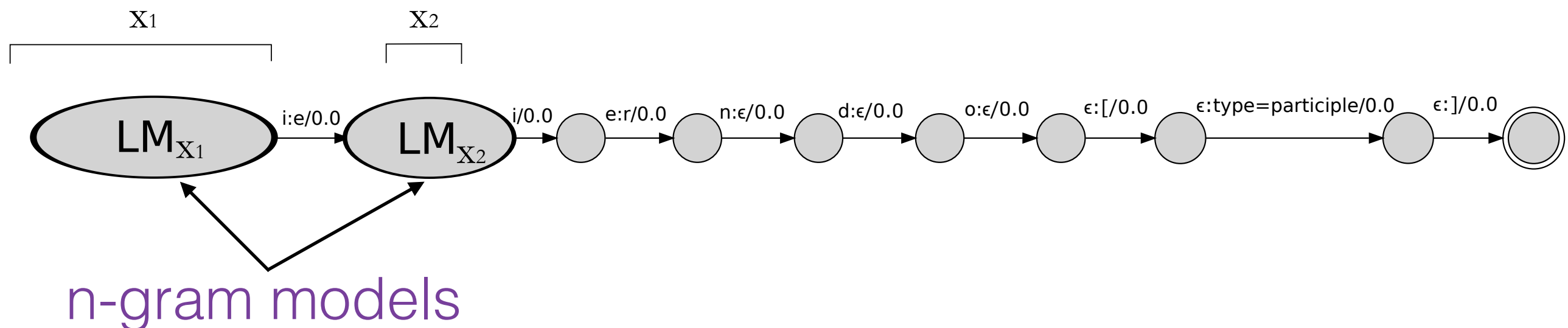
From paradigm to FST

avénir	$x_1 + e + x_2 + ir$	infinitive
aviniendo	$x_1 + i + x_2 + iendo$	pres part
avenido	$x_1 + e + x_2 + ido$	past part
avengo	$x_1 + e + x_2 + go$	1sg pres ind
avienes	$x_1 + ie + x_2 + es$	2sg pres ind



From paradigm to FST

avénir	$x_1 + e + x_2 + ir$	infinitive
aviniendo	$x_1 + i + x_2 + iendo$	pres part
avenido	$x_1 + e + x_2 + ido$	past part
avengo	$x_1 + e + x_2 + go$	1sg pres ind
avienes	$x_1 + ie + x_2 + es$	2sg pres ind



Example analysis (unweighted)

Example analysis (unweighted)

Analysis: **peleaste**

O	pelear	[pers=2 number=sg tense=past mood=ind]
----------	---------------	---

C	pelestar	[pers=1 num=sg tense=pres mood=subj]
	pelestar	[pers=3 num=sg tense=pres mood=subj]
	pelestar	[pers=3 num=sg tense=pres mood=ind]
	pelear	[pers=2 num=sg tense=past mood=ind]

U	pelestar	[pers=3 num=sg tense=pres mood=ind]
	pelestar	[pers=1 num=sg tense=pres mood=subj]
	pelestar	[pers=3 num=sg tense=pres mood=subj]
	pelestar	[pers=3 num=sg tense=pres mood=ind]
	pelear	[pers=2 num=sg tense=past mood=ind]
	pelestar	[pers=3 num=sg tense=pres mood=ind]
	pelestar	[pers=3 num=sg tense=pres mood=ind]

Example analysis (unweighted)

Analysis: **peleaste**

O	pelear	[pers=2 number=sg tense=past mood=ind]
C	pelestar	[pers=1 num=sg tense=pres mood=subj]
	pelestar	[pers=3 num=sg tense=pres mood=subj]
	pelestar	[pers=3 num=sg tense=pres mood=ind]
	pelear	[pers=2 num=sg tense=past mood=ind]
U	pelestar	[pers=3 num=sg tense=pres mood=ind]
	pelestar	[pers=1 num=sg tense=pres mood=subj]
	pelestar	[pers=3 num=sg tense=pres mood=subj]
	pelestar	[pers=3 num=sg tense=pres mood=ind]
	pelear	[pers=2 num=sg tense=past mood=ind]
	pelestar	[pers=3 num=sg tense=pres mood=ind]
	pelestar	[pers=3 num=sg tense=pres mood=ind]

correct

Example analysis (unweighted)

Analysis: **peleaste**

O	pelear	[pers=2 number=sg tense=past mood=ind]
C	pelestar	[pers=1 num=sg tense=pres mood=subj]
	pelestar	[pers=3 num=sg tense=pres mood=subj]
	pelestar	[pers=3 num=sg tense=pres mood=ind]
	pelear	[pers=2 num=sg tense=past mood=ind]
U	pelestar	[pers=3 num=sg tense=pres mood=ind]
	pelestar	[pers=1 num=sg tense=pres mood=subj]
	pelestar	[pers=3 num=sg tense=pres mood=subj]
	pelestar	[pers=3 num=sg tense=pres mood=ind]
	pelear	[pers=2 num=sg tense=past mood=ind]
	pelestar	[pers=3 num=sg tense=pres mood=ind]
	pelestar	[pers=3 num=sg tense=pres mood=ind]

correct

plausible

Example analysis (unweighted)

Analysis: **peleaste**

O	pelear	[pers=2 number=sg tense=past mood=ind]
C	pelestar	[pers=1 num=sg tense=pres mood=subj]
	pelestar	[pers=3 num=sg tense=pres mood=subj]
	pelestar	[pers=3 num=sg tense=pres mood=ind]
	pelear	[pers=2 num=sg tense=past mood=ind]
U	pelestar	[pers=3 num=sg tense=pres mood=ind]
	pelestar	[pers=1 num=sg tense=pres mood=subj]
	pelestar	[pers=3 num=sg tense=pres mood=subj]
	pelestar	[pers=3 num=sg tense=pres mood=ind]
	pelear	[pers=2 num=sg tense=past mood=ind]
	pelestar	[pers=3 num=sg tense=pres mood=ind]
	pelestar	[pers=3 num=sg tense=pres mood=ind]

correct

plausible

some noise

Example analysis (weighted)

rank	w	paradigm	vars	lemma	analyses
1	14.10	p1_abadernar	(1=compr)	comprar	[pers=2 num=sg tense=past mood=ind]
2	18.22	p1_abadernar	(1=comprast)	comprastar	[pers=1 num=sg tense=pres mood=sub]
				comprastar	[pers=3 num=sg tense=pres mood=sub]
3	23.57	p5_abogar	(1=compr)	comprar	[pers=2 num=sg tense=past mood=ind]
4	24.58	p4_abolir	(1=comprast)	comprastir	[pers=3 num=sg tense=pres mood=ind]
5	24.58	p8_acrecentar	(1=com,2=pr)	comprar	[pers=2 num=sg tense=past mood=ind]
6	25.51	p37_colgar	(1=c,2=mpr)	comprar	[pers=2 num=sg tense=past mood=ind]
7	26.20	p10_acostar	(1=c,2=mpr)	comprar	[pers=2 num=sg tense=past mood=ind]
8	26.61	p7_acceder	(1=comprast)	compraster	[pers=3 num=sg tense=pres mood=ind]
9	26.87	p8_acrecentar	(1=comp,2=r)	comprar	[pers=2 num=sg tense=past mood=ind]
10	29.98	p20_cegar	(1=c,2=ompr)	comprar	[pers=2 num=sg tense=past mood=ind]

compraste

Boldface = correct

Experiments

- Complete inflection tables for various languages:

German verbs (1,827 tables)

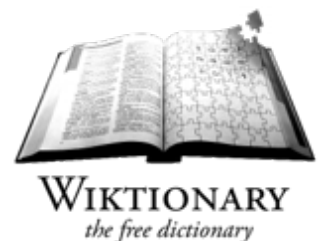
German nouns (2,564 tables)

Spanish verbs (3,855 tables)

Finnish verbs (7,049 tables)

Finnish nouns/adjectives (6,200 tables)

Durrett & DeNero (2013)



Conjugation [\[edit\]](#)

conjugation of schreiben [hide ▲]					
infinitive		schreiben			
present participle		schreibend			
past participle		geschrieben			
auxiliary		haben			
	indicative			subjunctive	
present	ich	schreibe	wir	schreiben	
	du	schreibst	ihr	schreibt	
	er	schreibt	sie	schreiben	
preterite	ich	schrieb	wir	schrieben	
	du	schriebst	ihr	schriebt	
	er	schrieb	sie	schrieben	
imperative		schreib (du)		schreibt (ihr)	
		schreibe (du)			
composed forms of schreiben [show ▼]					

Experiments

Unweighted (return multiple analyses)

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

Train/test split from D&DN13, all test words unseen

Experiments

Weighted (return single top-scoring lemma + MSD)

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

Train/test split from D&DN13, all test words unseen

Wrap-up

Wrap-up

- ▶ Simple method to construct FST analyzers & guessers from labeled data

Wrap-up

- ▶ Simple method to construct FST analyzers & guessers from labeled data
- ▶ Yields weighted/unweighted FSTs


Wrap-up

- ▶ Simple method to construct FST analyzers & guessers from labeled data
- ▶ Yields weighted/unweighted FSTs
- ▶ Robust performance for inflectional morphology

Wrap-up

- ▶ Simple method to construct FST analyzers & guessers from labeled data
- ▶ Yields weighted/unweighted FSTs
- ▶ Robust performance for inflectional morphology
- ▶ Can also use as generator

Thank You

Code and language data at: 

<https://github.com/marfors/paradigmextract/>