Unsupervised Code-Switching for Multilingual Historical Document Transcription

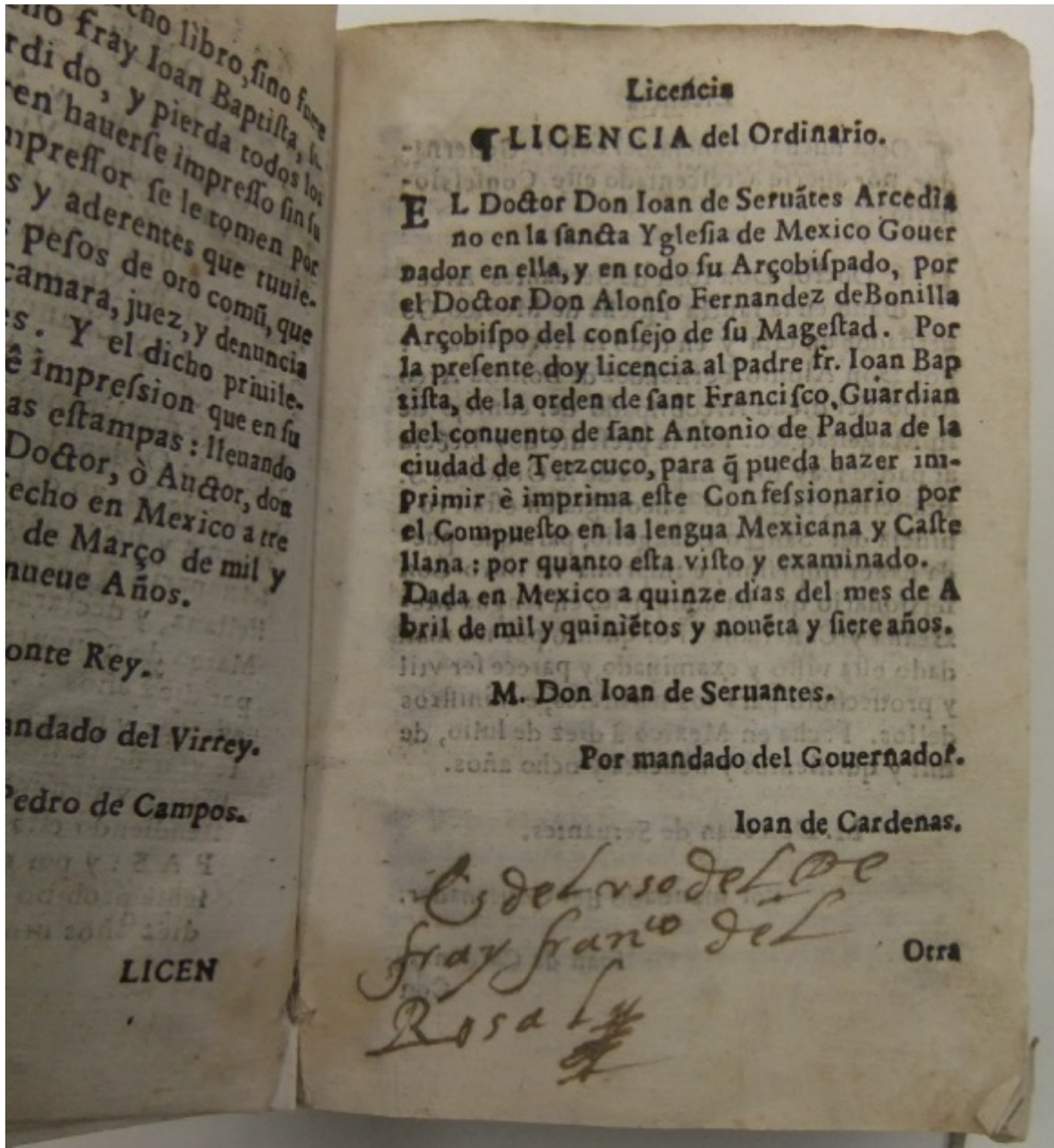
Dan Garrette UT-Austin Computer Science
Hannah Alpert-Abrams UT-Austin Comparative Literature

Taylor Berg-Kirkpatrick UC Berkeley Computer Science
Dan Klein UC Berkeley Computer Science

Historical Document Transcription

Working with scholars in humanities who want to study texts from the 1500s.

Standard OCR systems don't work well on printing-press books.



cho libro, ſino fuere
o fray Ioan Baptiſta, lo
rdi do, y pierda todos los
en hauerſe impreſſo ſin ſu
mpreſſor ſe le tomen por
s y aderentes que tuuie-
pesos de oro comũ, que
camara, juez, y denuncia
s. Y el dicho priuile-
ê impreſsion que en ſu
as eſtampas: llenando
Doctor, ò Auctor, don
echo en Mexico a tre
de Março de mil y
nueue Años.

onte Rey.

ndado del Virrey.

edro de Campos.

LICEN

Licencia

¶ LICENCIA del Ordinario.

EL Doctor Don Ioan de Seruãtes Arcedia
no en la ſancta Yglesia de Mexico Gouer
nador en ella, y en todo ſu Arçobiſpado, por
el Doctor Don Alonſo Fernandez deBonilla
Arçobiſpo del conſejo de ſu Mageſtad. Por
la preſente doy licencia al padre fr. Ioan Bap
tiſta, de la orden de ſant Franciſco, Guardian
del conuento de ſant Antonio de Padua de la
ciudad de Terzcuco, para q̃ pueda hazer im-
primir è imprima eſte Confeſsionario por
el Compueſto en la lengua Mexicana y Caſte
llana: por quanto eſta viſto y examinado.
Dada en Mexico a quinze dias del mes de A
bril de mil y quiniẽtos y nouẽta y ſiete años.

M. Don Ioan de Seruantes.

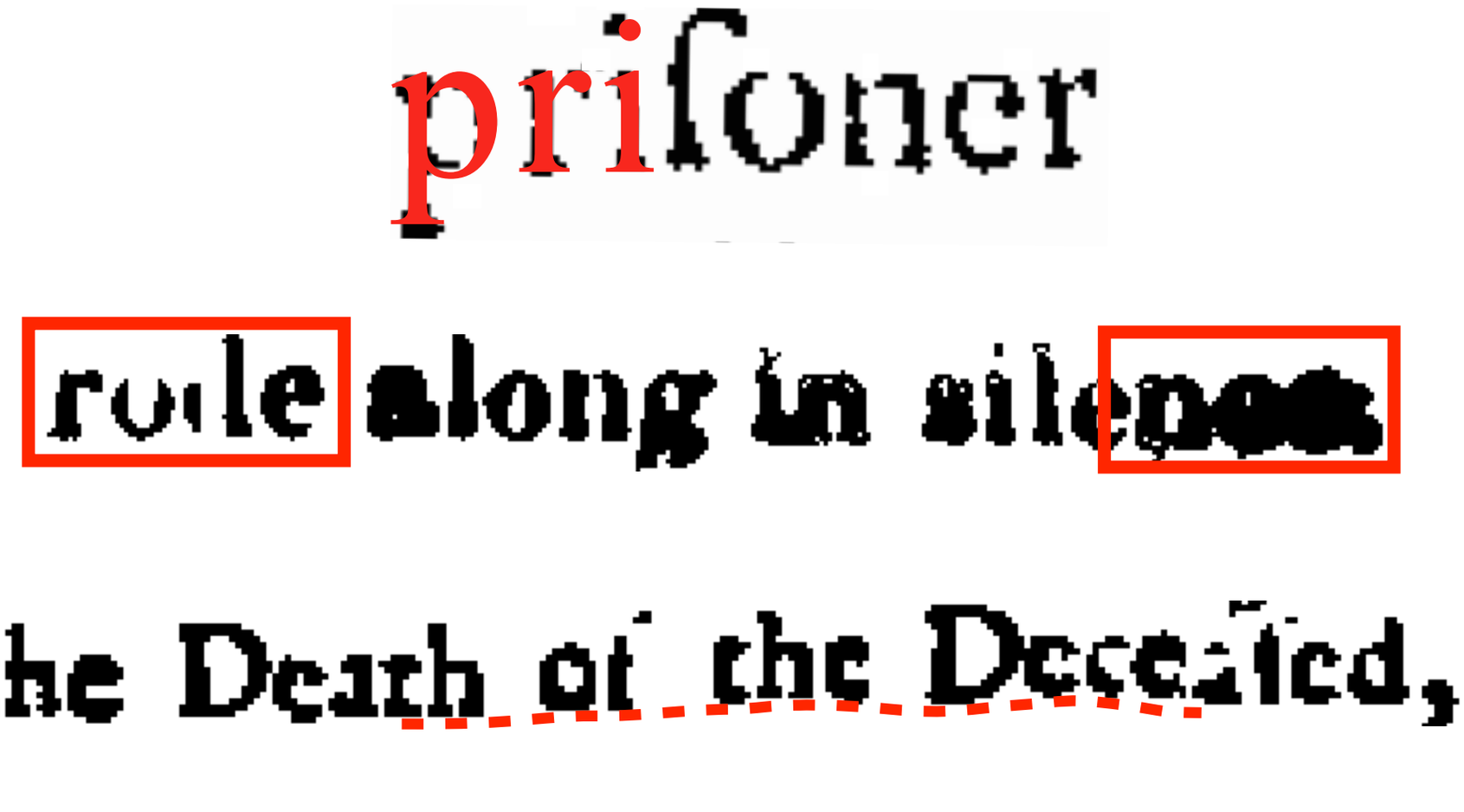
Por mandado del Gouernador.

Ioan de Cardenas.

Otra

State-of-the-Art: *Ocular*

- Berg-Kirkpatrick, Durrett, and Klein 2013



Slide courtesy of Taylor Berg-Kirkpatrick [Berg-Kirkpatrick et al. 2013]

Multilingual Texts

- But many historical documents are written in, and switch readily between, multiple languages.

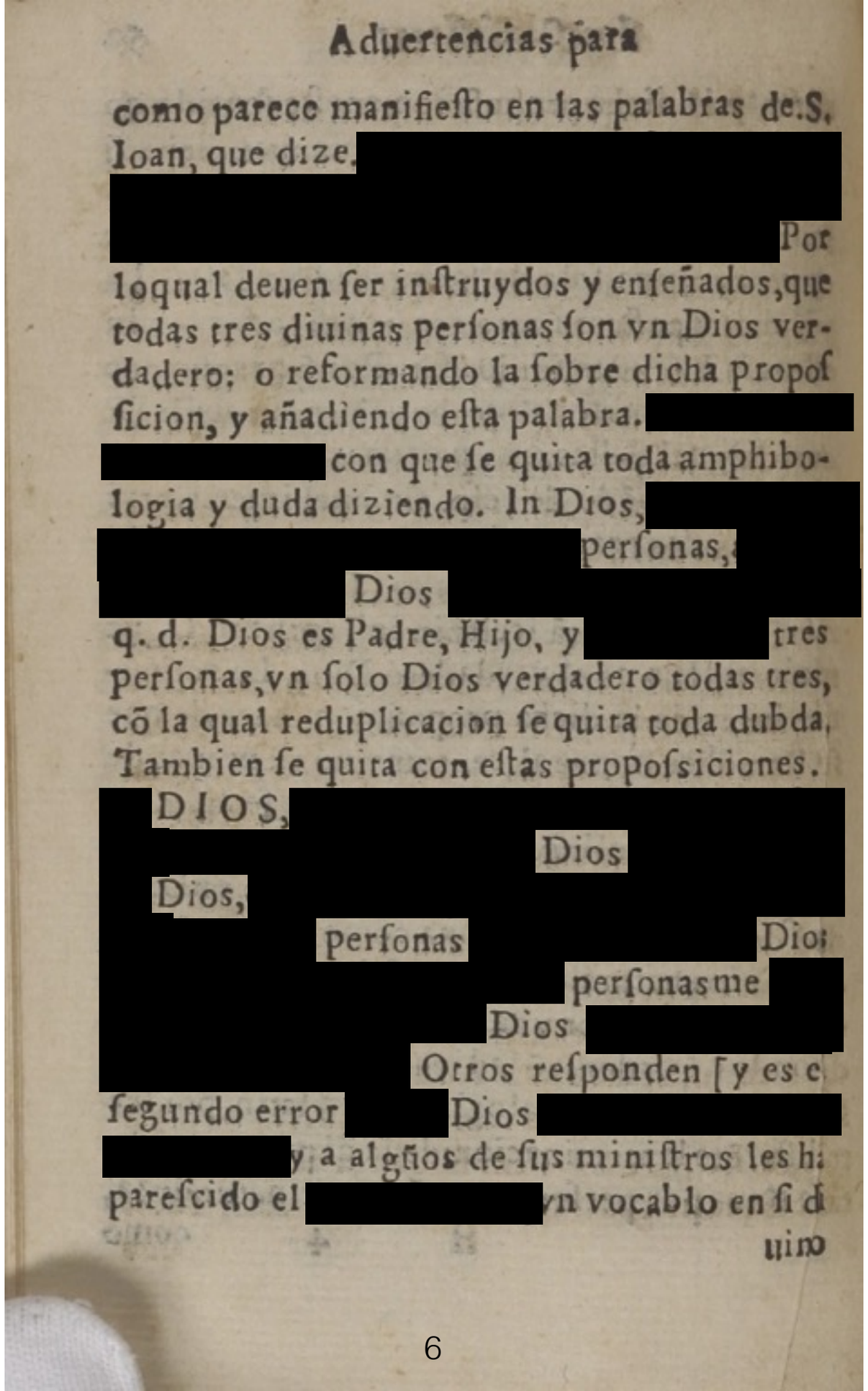
como parece manifiesto en las palabras de S. Ioan, que dize. Tres sunt qui testimoniũ dãt in cœlo Pater, Verbum, & Spiritussanctus: & hi tres vnum sunt. 1. Ioann. vltimo. Por loqual deuen ser instruydos y enseñados, que todas tres diuinas personas son vn Dios verdadero: o reformando la sobre dicha proposicion, y añadiendo esta palabra. In huel ime ixtintzitzin, con que se quita toda amphibologia y duda diziendo. In Dios, ca Tettatzin Tepiltzin, Spiritu sancto, ei personas, çan ce huelnelli teutl Dios in huel imeixtintzitzin. q. d. Dios es Padre, Hijo, y Spũ sancto tres personas, vn solo Dios verdadero todas tres, cõ la qual reduplicacion se quita toda dubda. Tambien se quita con estas proposiciones. In DIOS, ca Tettatzin, Tepiltzin, Spiritusanc̃to, çan huel iceltzin teutl Dios tlahtohuani. In Dios, ca Tettatzin, Tepiltzin, Spiritusanc̃to, imeixtin personas çan huel iceltzin Dios tlahtohuani. Ca inimeixtin personas me ca-çan huel iceltzin teutl Dios tlahtohuani in-huel imeixtin. ¶ Otros responden [y es c segundo error] ça ce Dios tlahtohuani, imme teihttotica, y a algũos de sus ministros les h parescido el meteihttotica, vn vocablo en si d

uino

Spanish

Latin

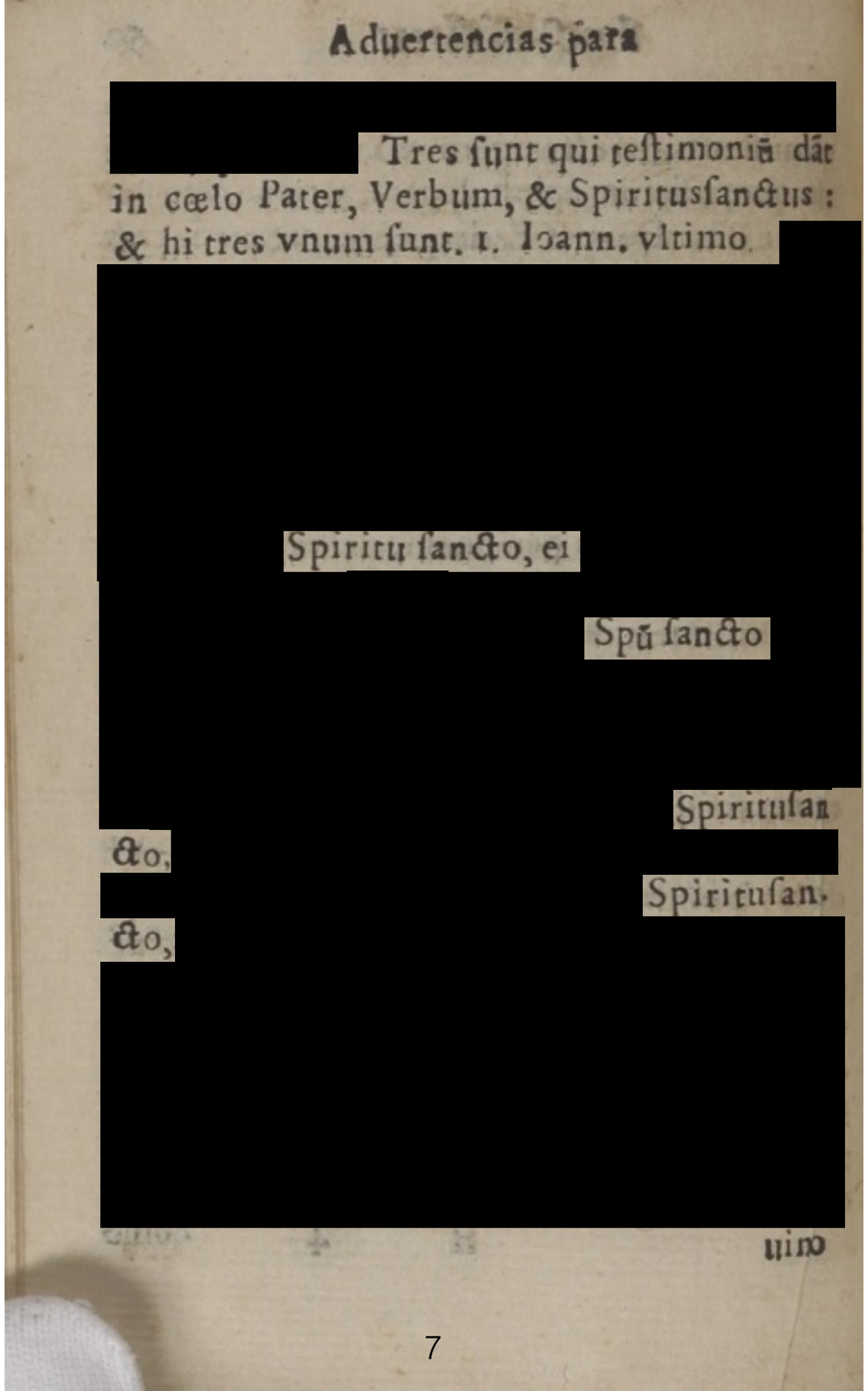
Nahuatl



Spanish

Latin

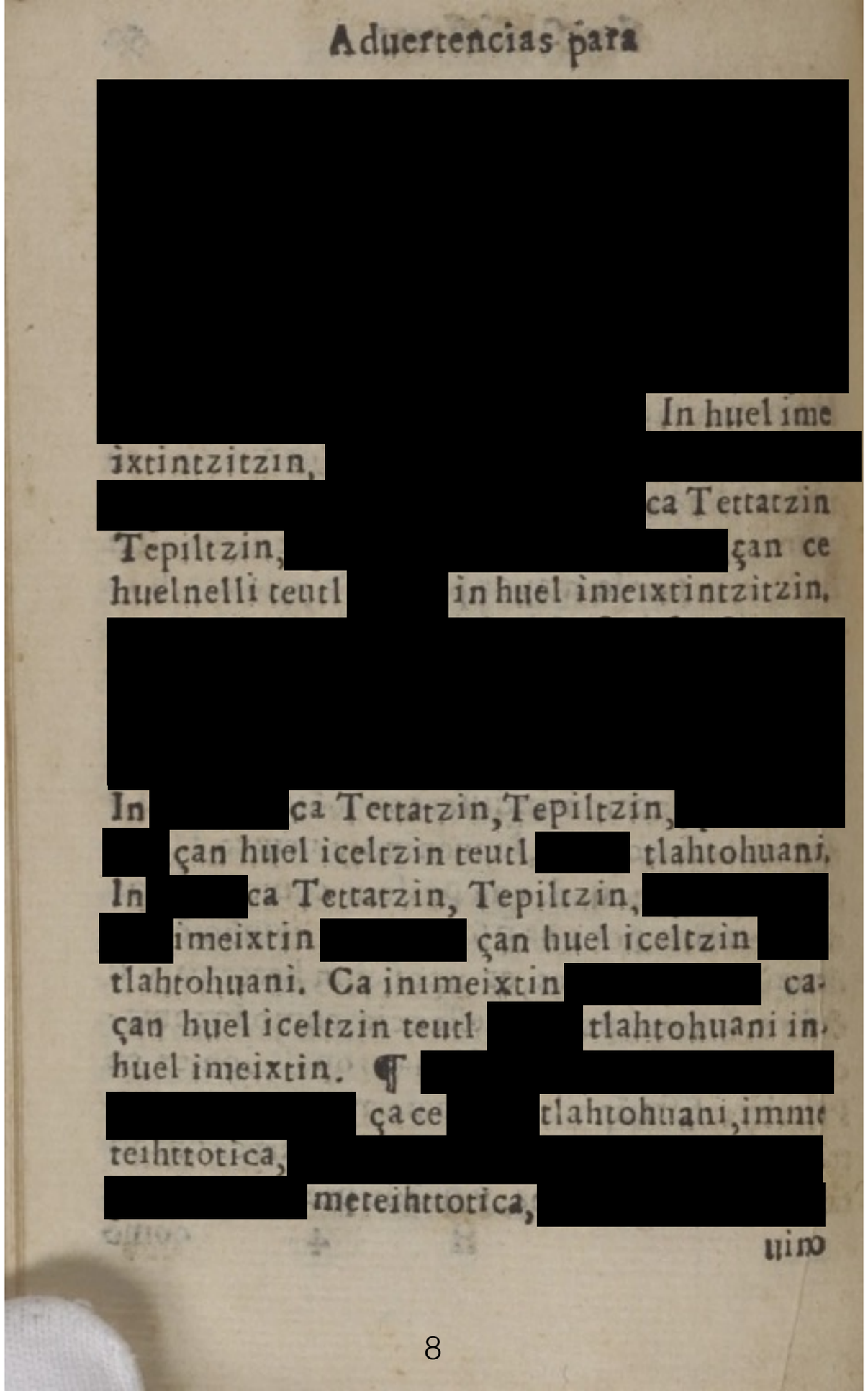
Nahuatl



Spanish

Latin

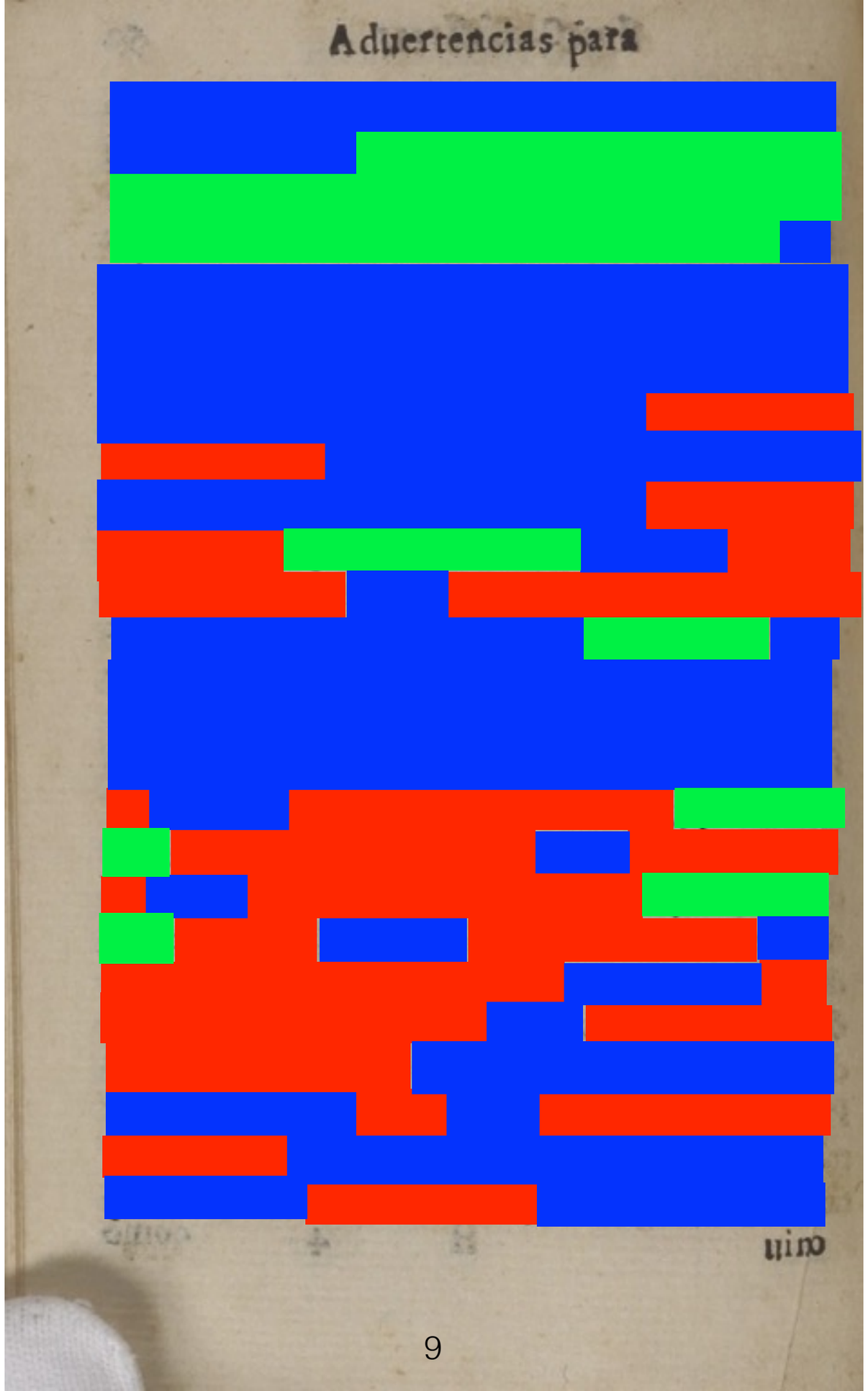
Nahuatl



Spanish

Latin

Nahuatl



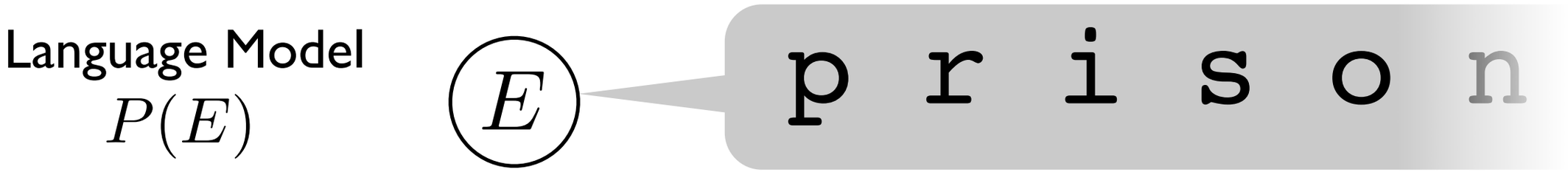
Starting Point: *Ocular*

Generative Model in 3 parts:

1. Language model
2. Typesetting model
3. Rendering model

[Berg-Kirkpatrick et al. 2013]

Ocular's Generative Model

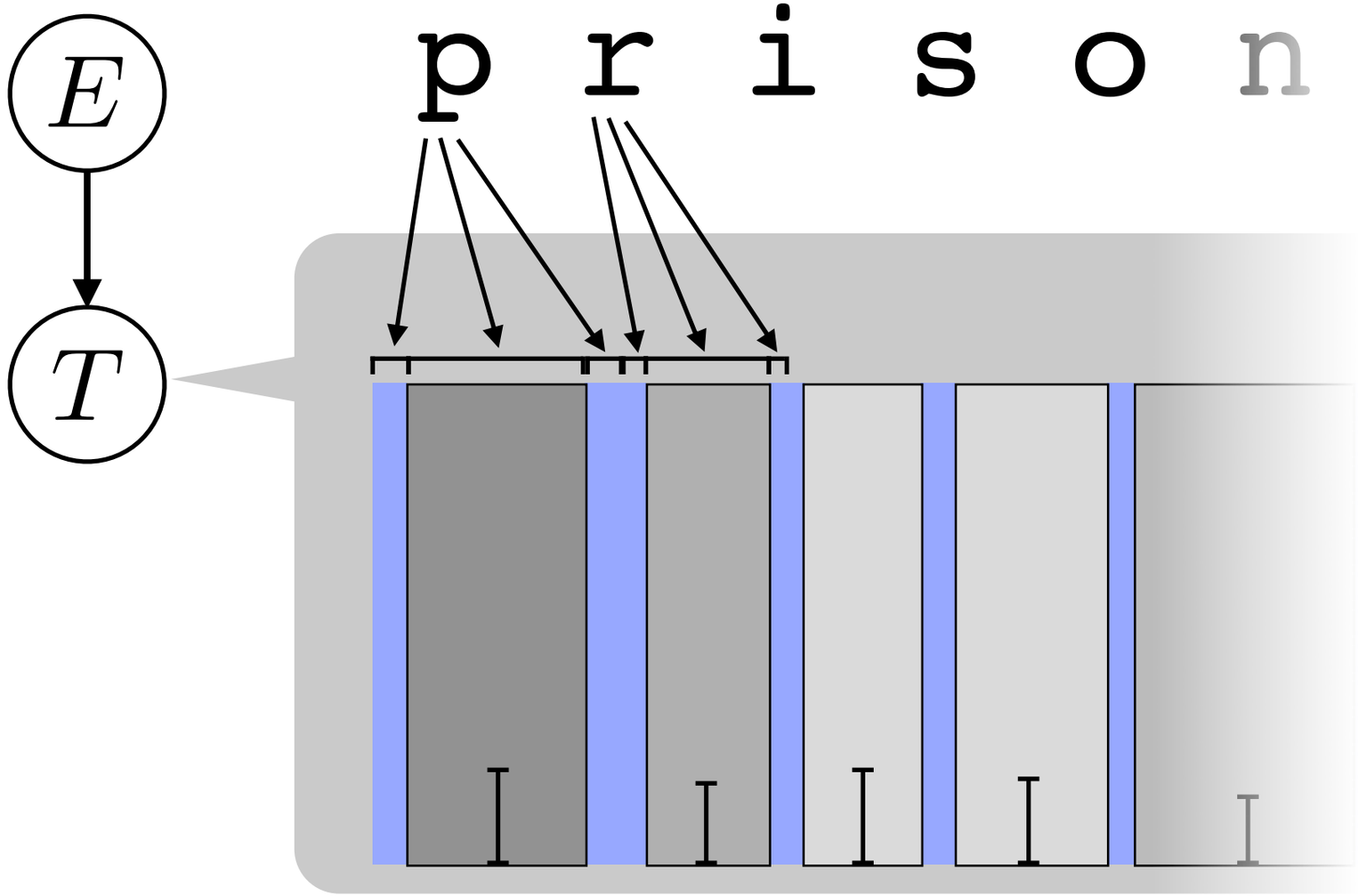


Slide courtesy of Taylor Berg-Kirkpatrick [Berg-Kirkpatrick et al. 2013]

Ocular's Generative Model

Language Model
$P(E)$

Typesetting Model
$P(T|E)$



Slide courtesy of Taylor Berg-Kirkpatrick

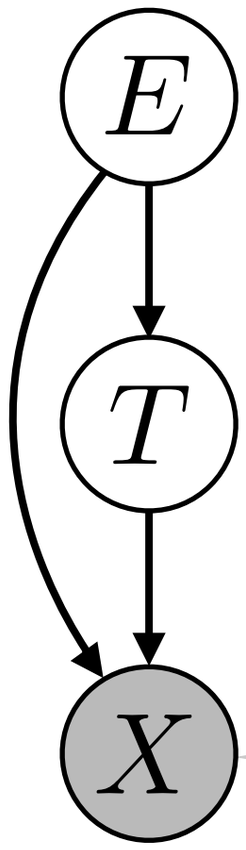
[Berg-Kirkpatrick et al. 2013]

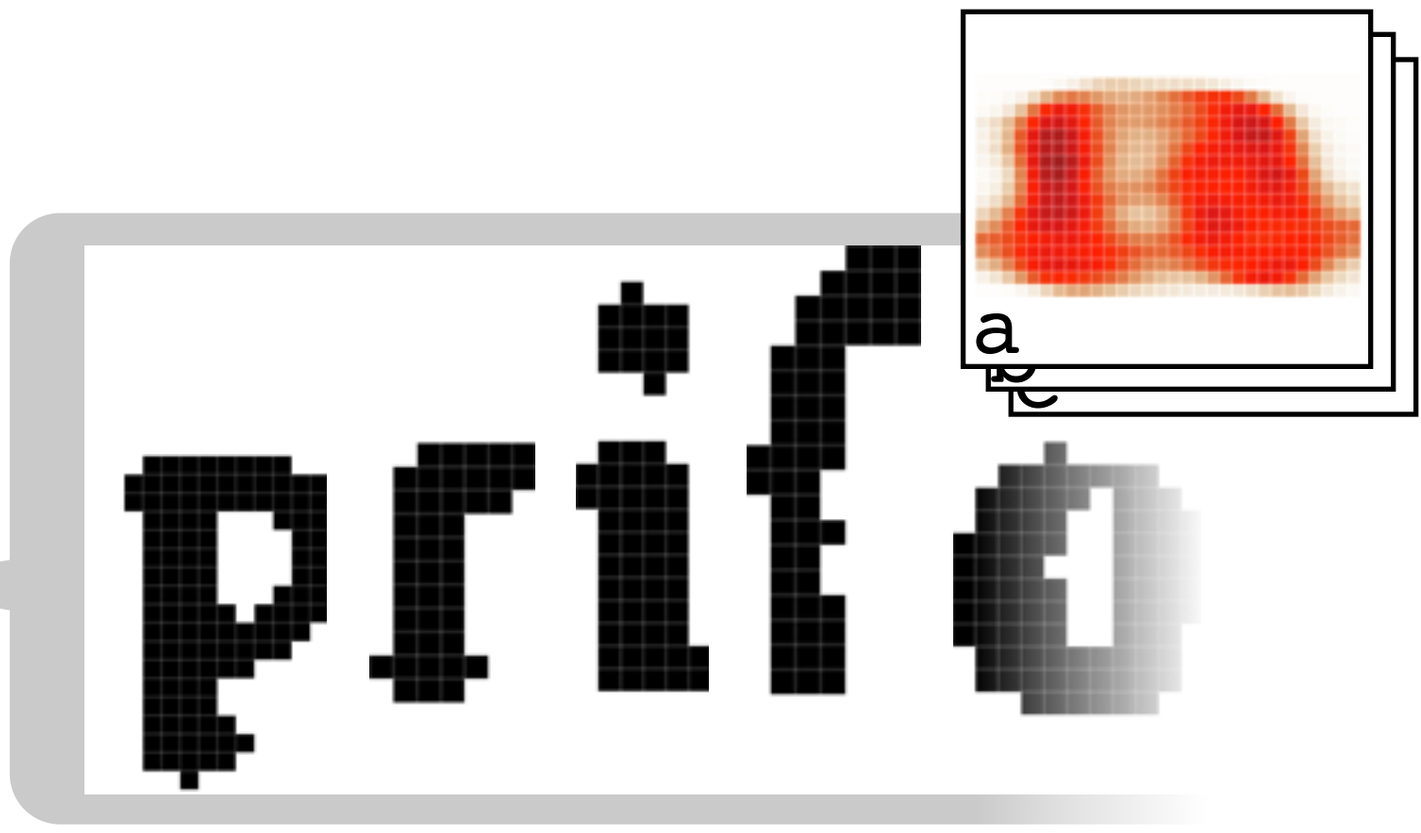
Ocular's Generative Model

Language Model
$P(E)$

Typesetting Model
$P(T|E)$

Rendering Model
$P(X|E,T)$



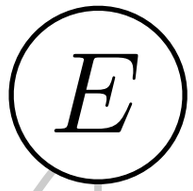


Slide courtesy of Taylor Berg-Kirkpatrick

[Berg-Kirkpatrick et al. 2013]

Our Focus

Language Model
$P(E)$



p r i s o n

Typesetting Model
$P(T|E)$

Rendering Model
$P(X|E,T)$

Slide courtesy of Taylor Berg-Kirkpatrick

[Berg-Kirkpatrick et al. 2013]

Starting Point: *Ocular*

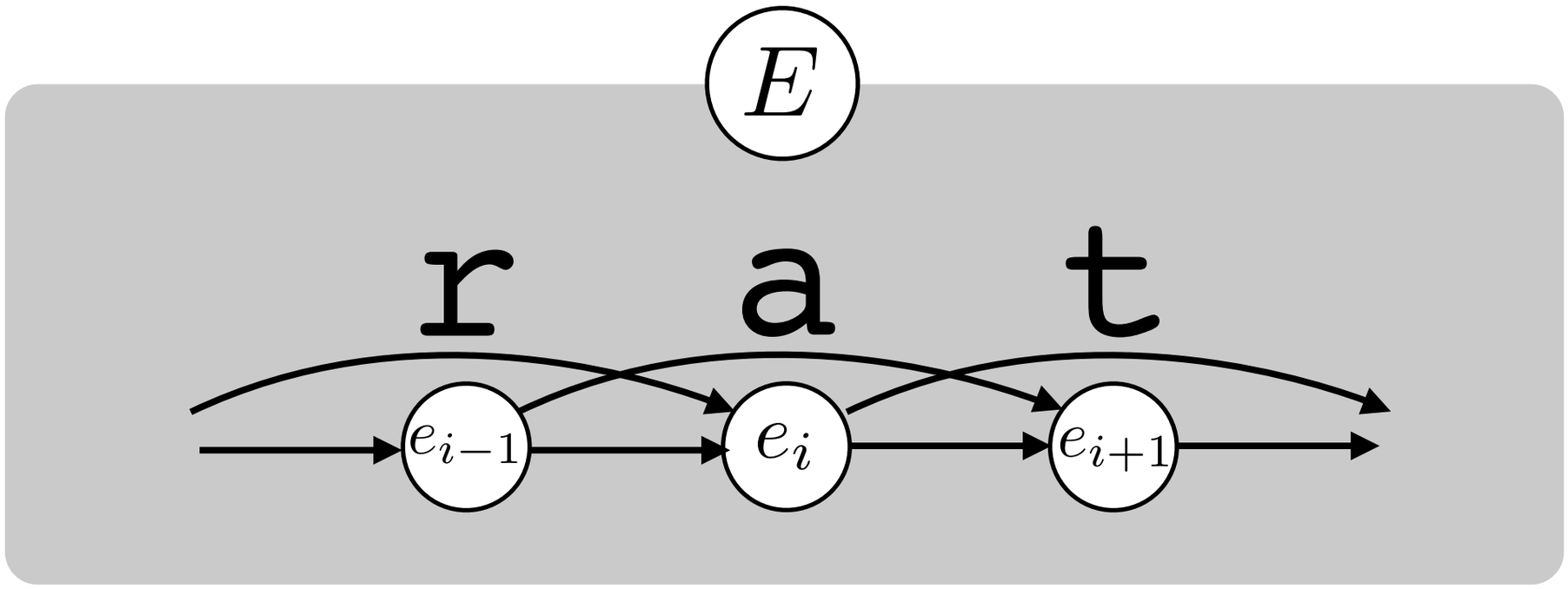
- The language model helps Ocular work well, but creates additional challenges for many documents.

- Our work helps to overcome those challenges.

Our Focus

1. Multilingual code-switching
2. Inconsistent/outdated orthography

Ocular's Language Model

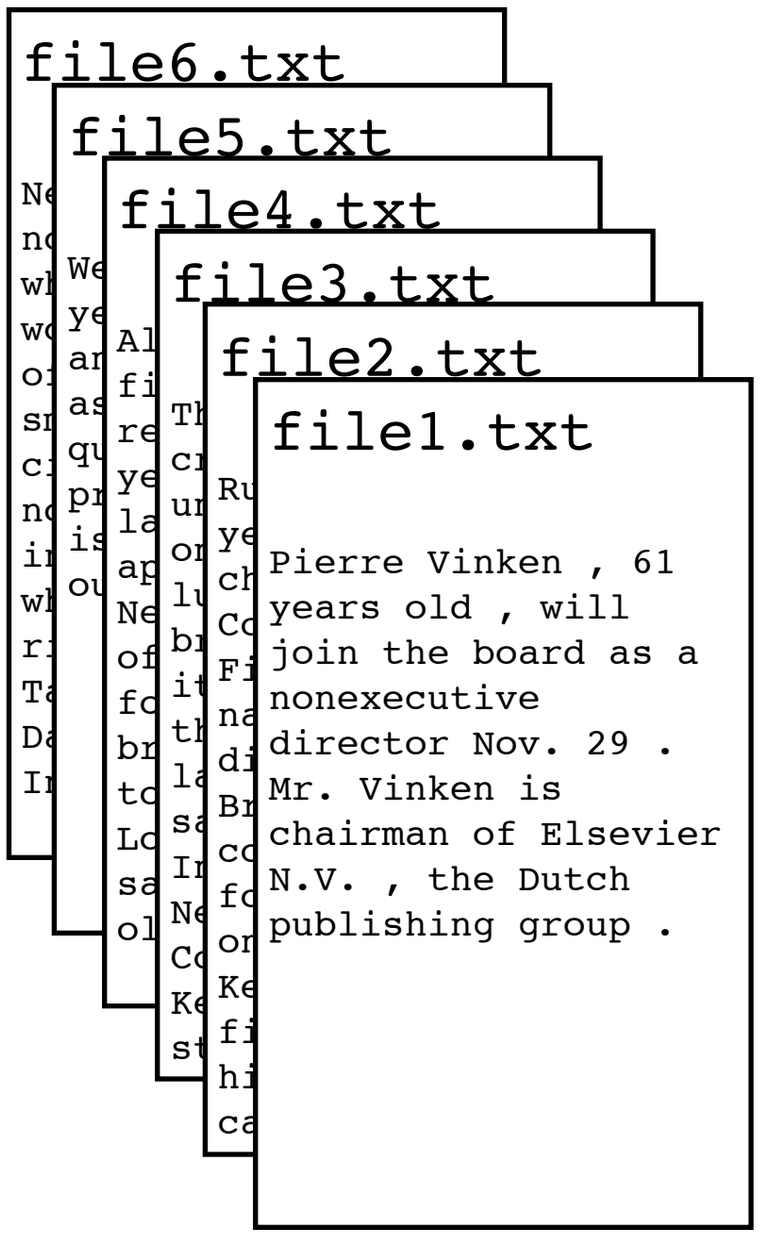


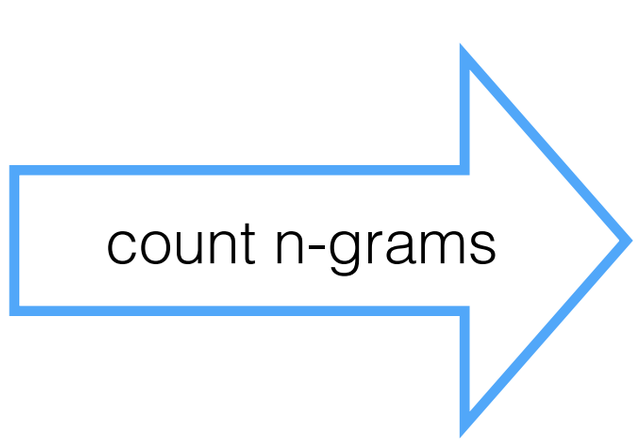
Kneser-Ney smoothed **character** 6-gram

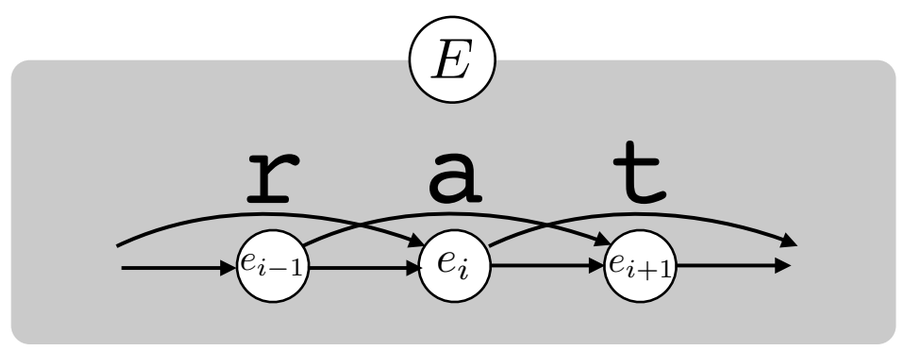
Slide courtesy of Taylor Berg-Kirkpatrick

[Berg-Kirkpatrick et al. 2013]

Ocular's Language Model

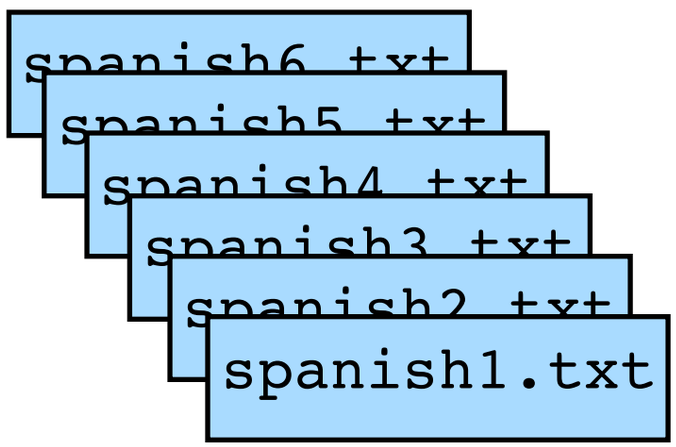


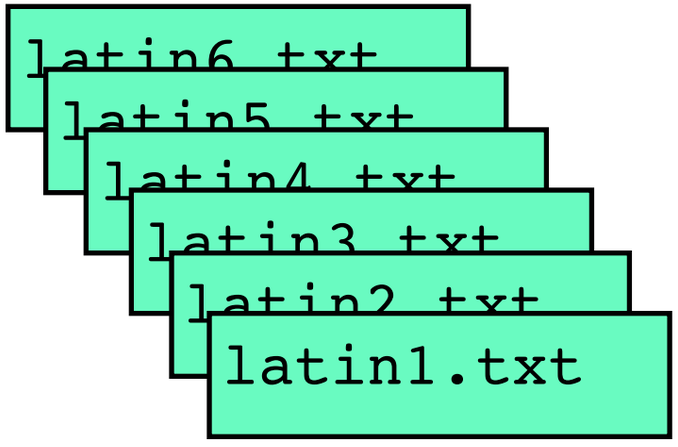


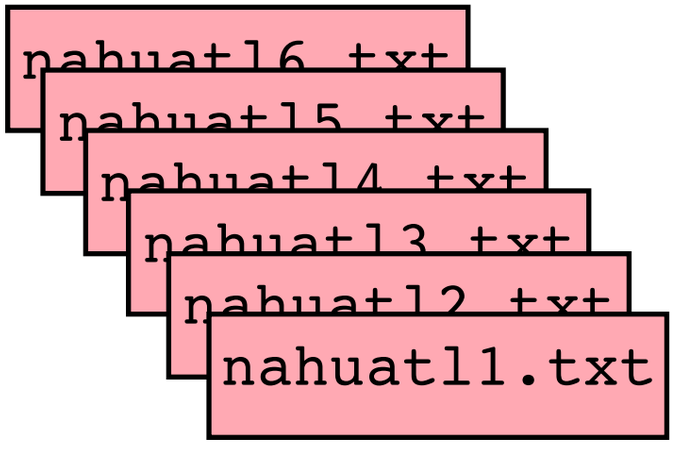


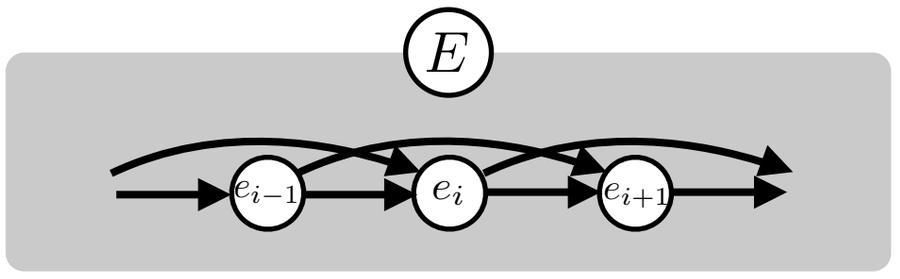
[Berg-Kirkpatrick et al. 2013]

Baseline Multilingual Model

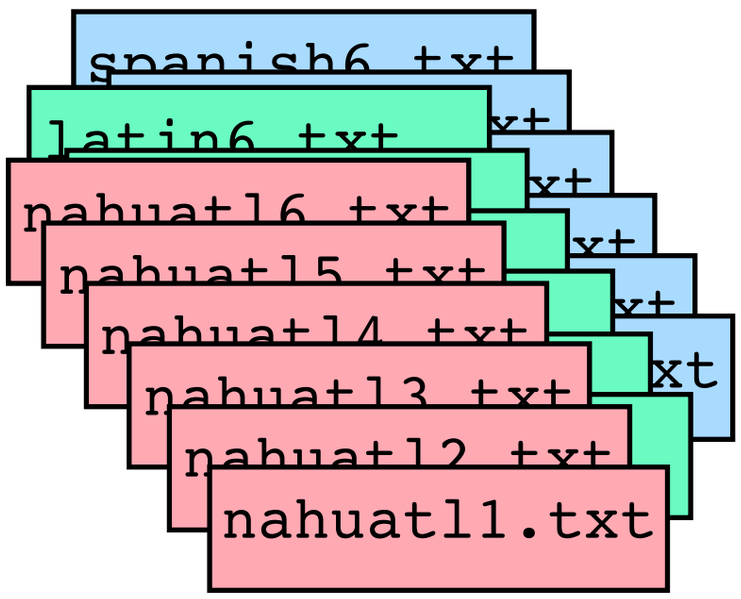




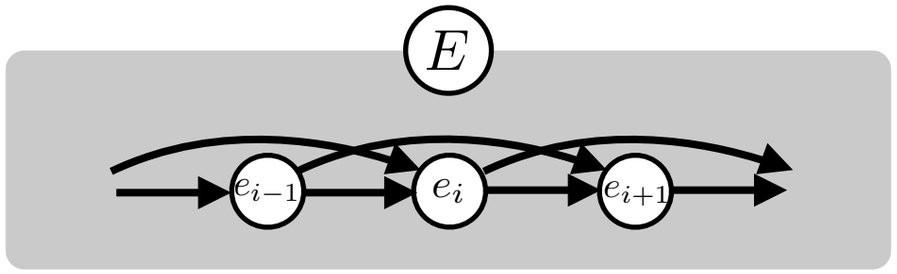




Baseline Multilingual Model



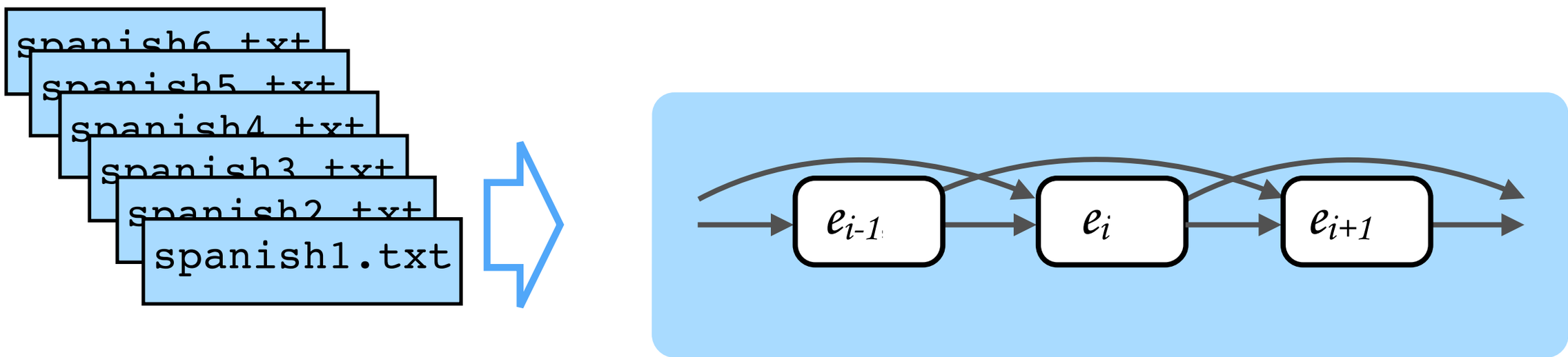
count n-grams

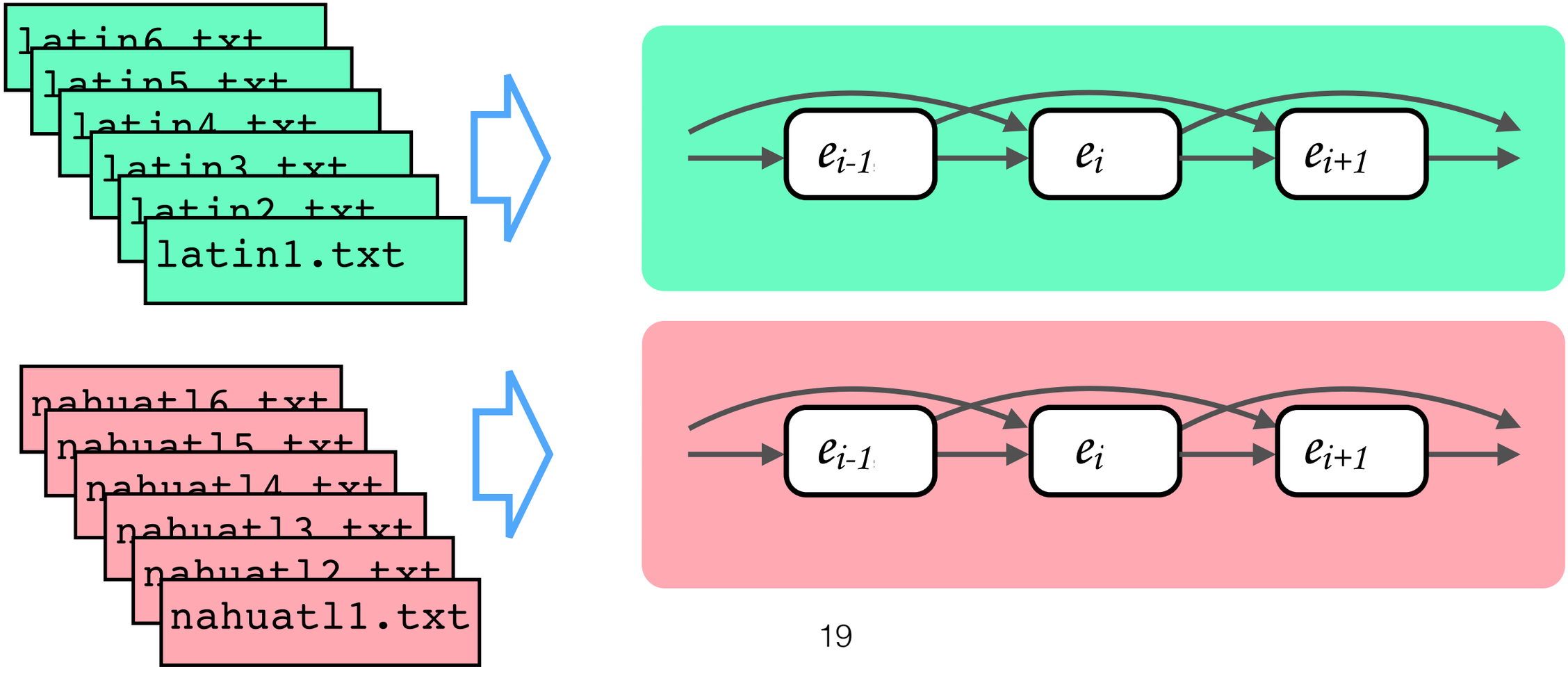


Baseline Multilingual Model

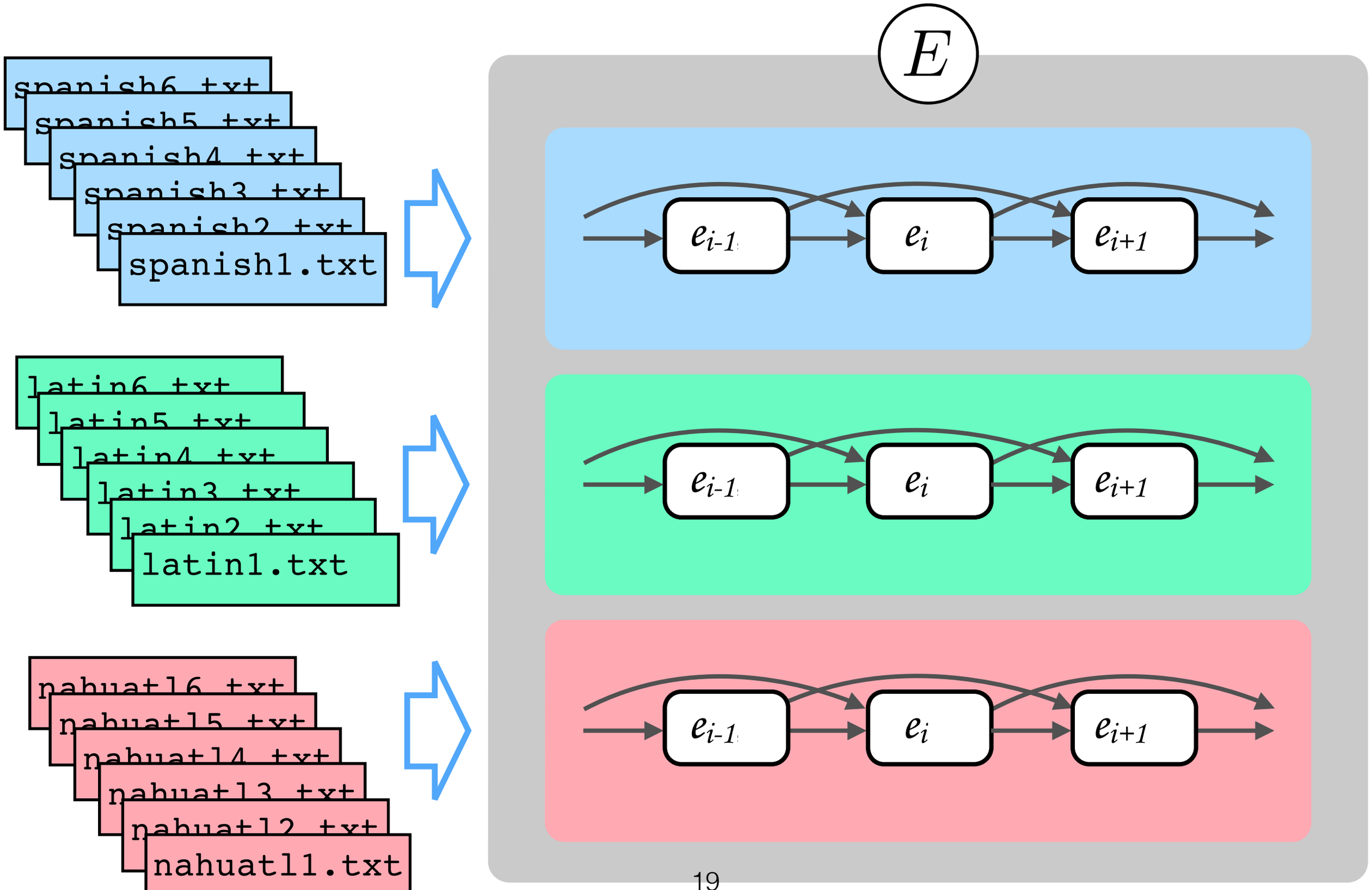
- Poor results
- “Multilingual blur”

Code-Switching Language Model

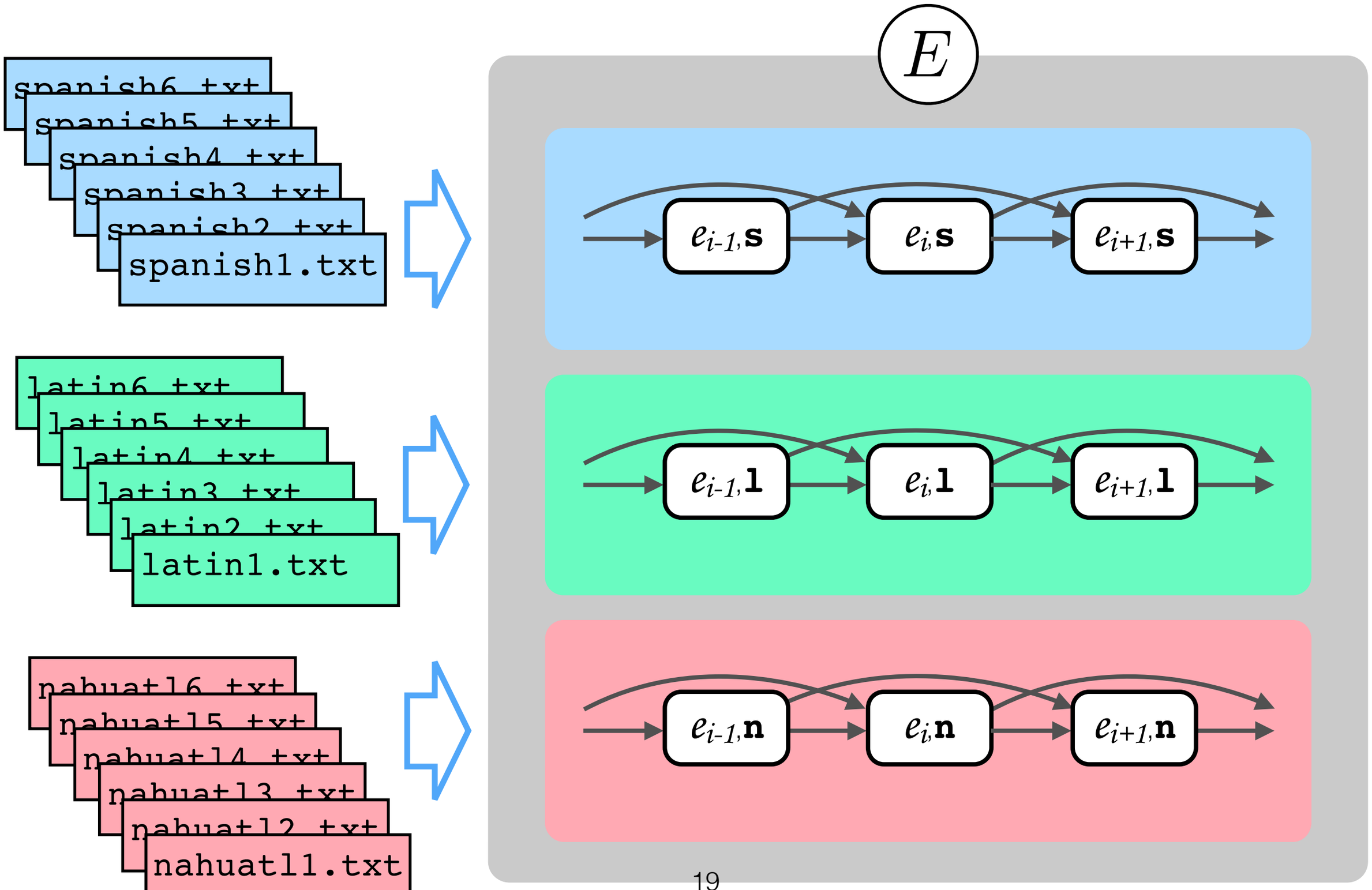




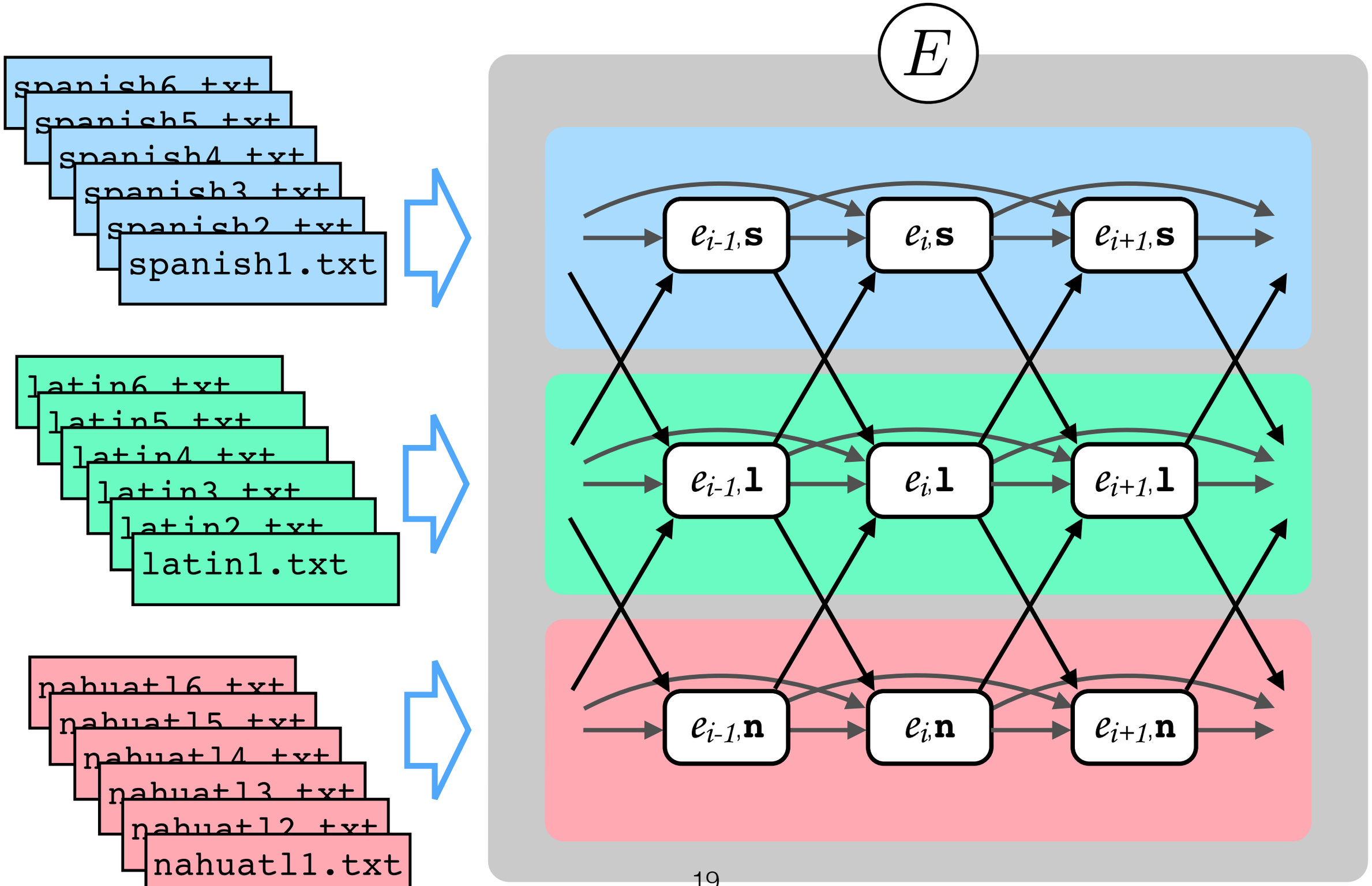
Code-Switching Language Model



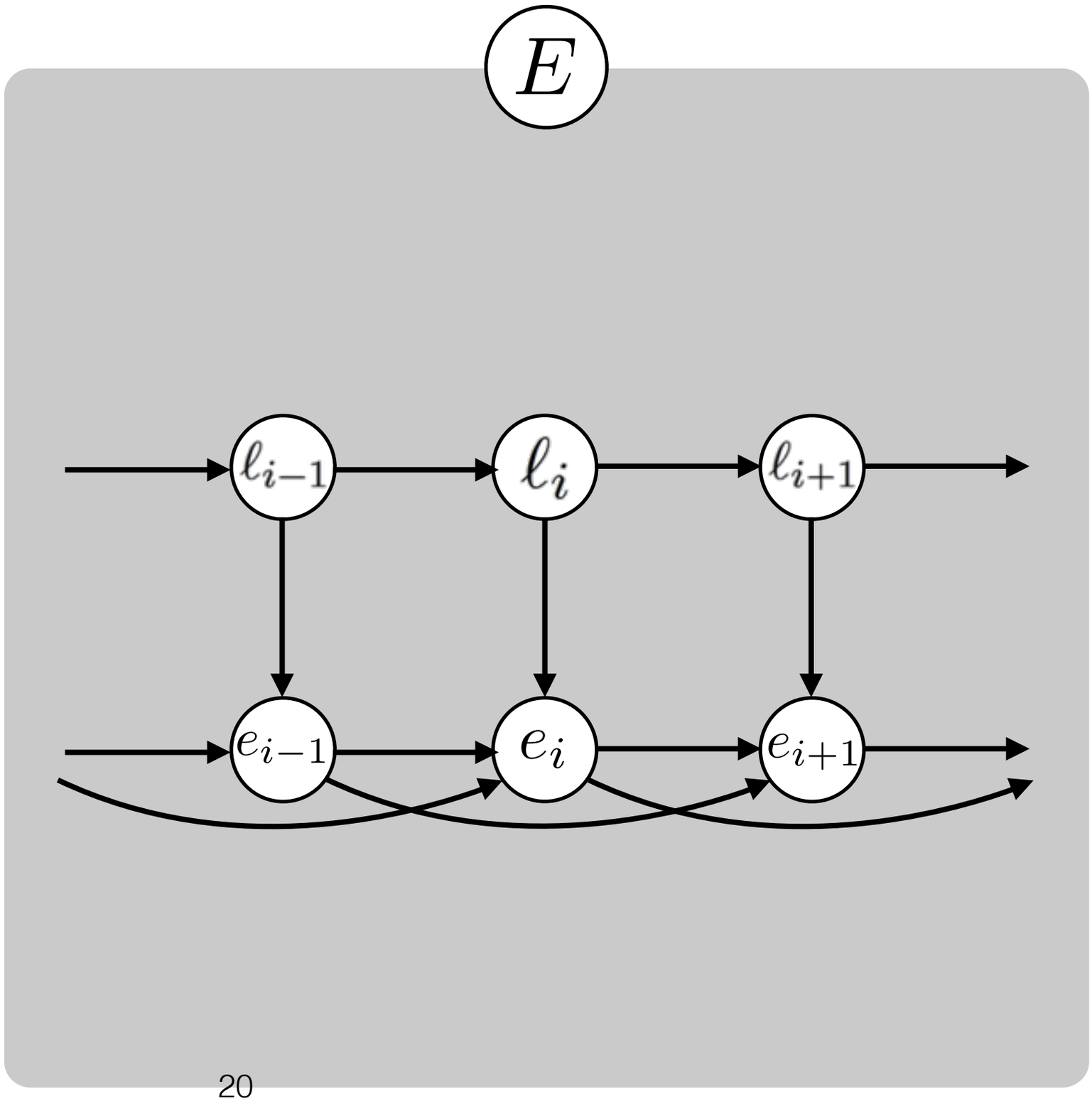
Code-Switching Language Model



Code-Switching Language Model



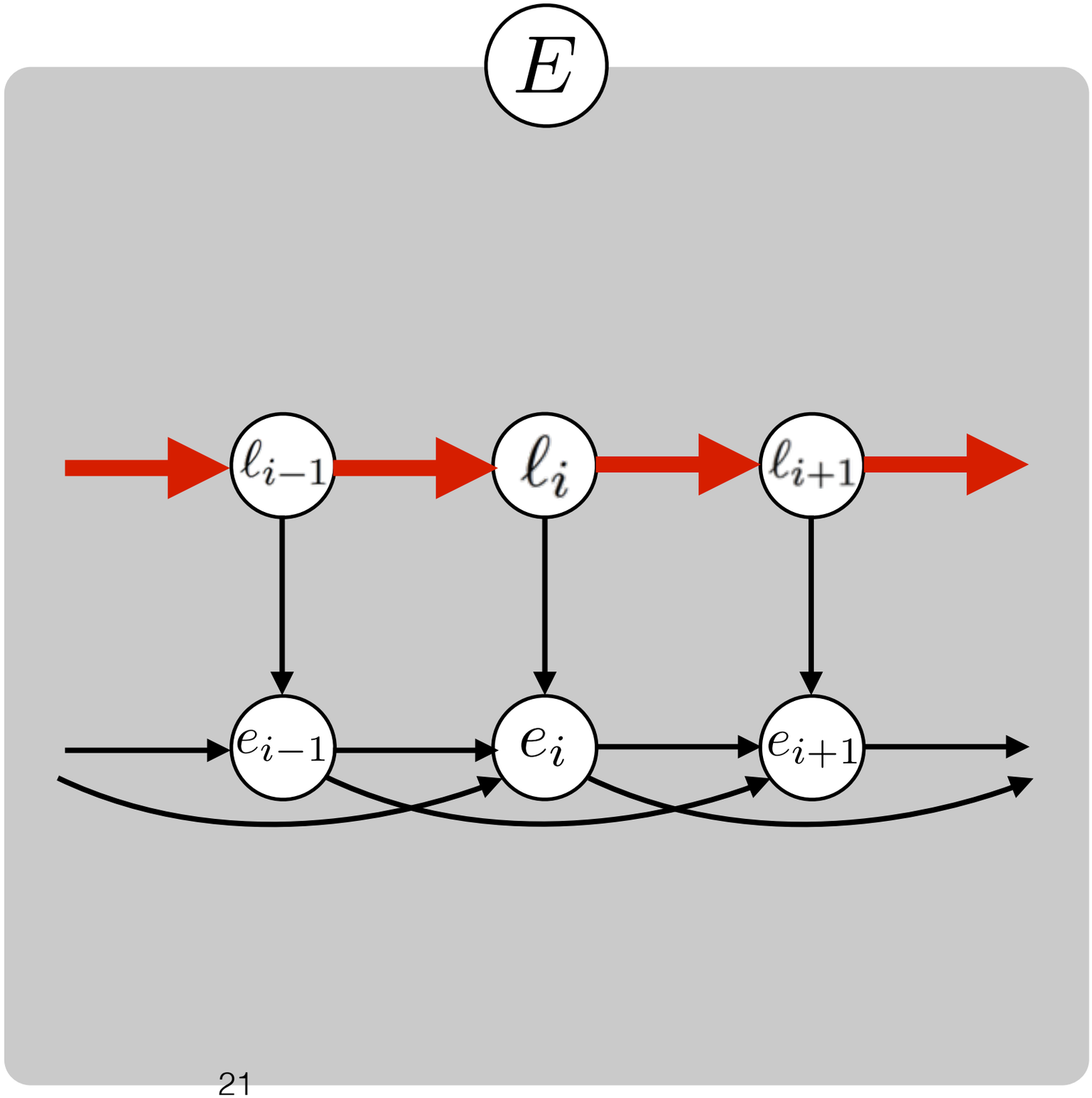
Code-Switching Language Model



Code-Switching Language Model

$P(\ell_i|\ell_{i-1})$ is learned unsupervised via EM,

with a hyperparameter biasing the model toward **not** switching (long language spans)



como parece manifiesto en las palabras de S. Ioan, que dize. Tres sunt qui testimoniũ dãt in cœlo Pater, Verbum, & Spiritussanctus: & hi tres vnum sunt. 1. Ioann. vltimo. Por loqual deuen ser instruydos y enseñados, que todas tres diuinas personas son vn Dios verdadero: o reformando la sobre dicha proposicion, y añadiendo esta palabra. In huel ime ixtintzitzin, con que se quita toda amphibologia y duda diziendo. In Dios, ca Tettatzin Tepiltzin, Spiritu sancto, ei personas, çan ce huelnelli teutl Dios in huel imeixtintzitzin. q. d. Dios es Padre, Hijo, y Spũ sancto tres personas, vn solo Dios verdadero todas tres, cõ la qual reduplicacion se quita toda dubda. Tambien se quita con estas proposiciones. In DIOS, ca Tettatzin, Tepiltzin, Spiritusanc̃to, çan huel iceltzin teutl Dios tlahtohuani. In Dios, ca Tettatzin, Tepiltzin, Spiritusanc̃to, imeixtin personas çan huel iceltzin Dios tlahtohuani. Ca inimeixtin personas me ca-çan huel iceltzin teutl Dios tlahtohuani in-huel imeixtin. ¶ Otros responden [y es e segundo error] ça ce Dios tlahtohuani, imme teihttotica, y a algũos de sus ministros les h parescido el meteihttotica, vn vocablo en si d

uino

ſicion, y añadiendo eſta palabra. In huel ime
ixtintzitzin, con que ſe quita toda amphibo-
logia y duda diziendo. In Dios, ca Tettatzin

añadiendo eſta palabra. *In huel ime ixtintzitzin*, con que ſe quita toda amphibologia y duda diziendo. In Dios,

ixtintzitzin, con que

Spanish

AÁBCDÉFGHIÍJKLMÑOÓPQRSTUÚVWXYZ
aábcdéfghiíjklmñoópqrstuúvwxyz
01234567890.,/\()?!”’:;-

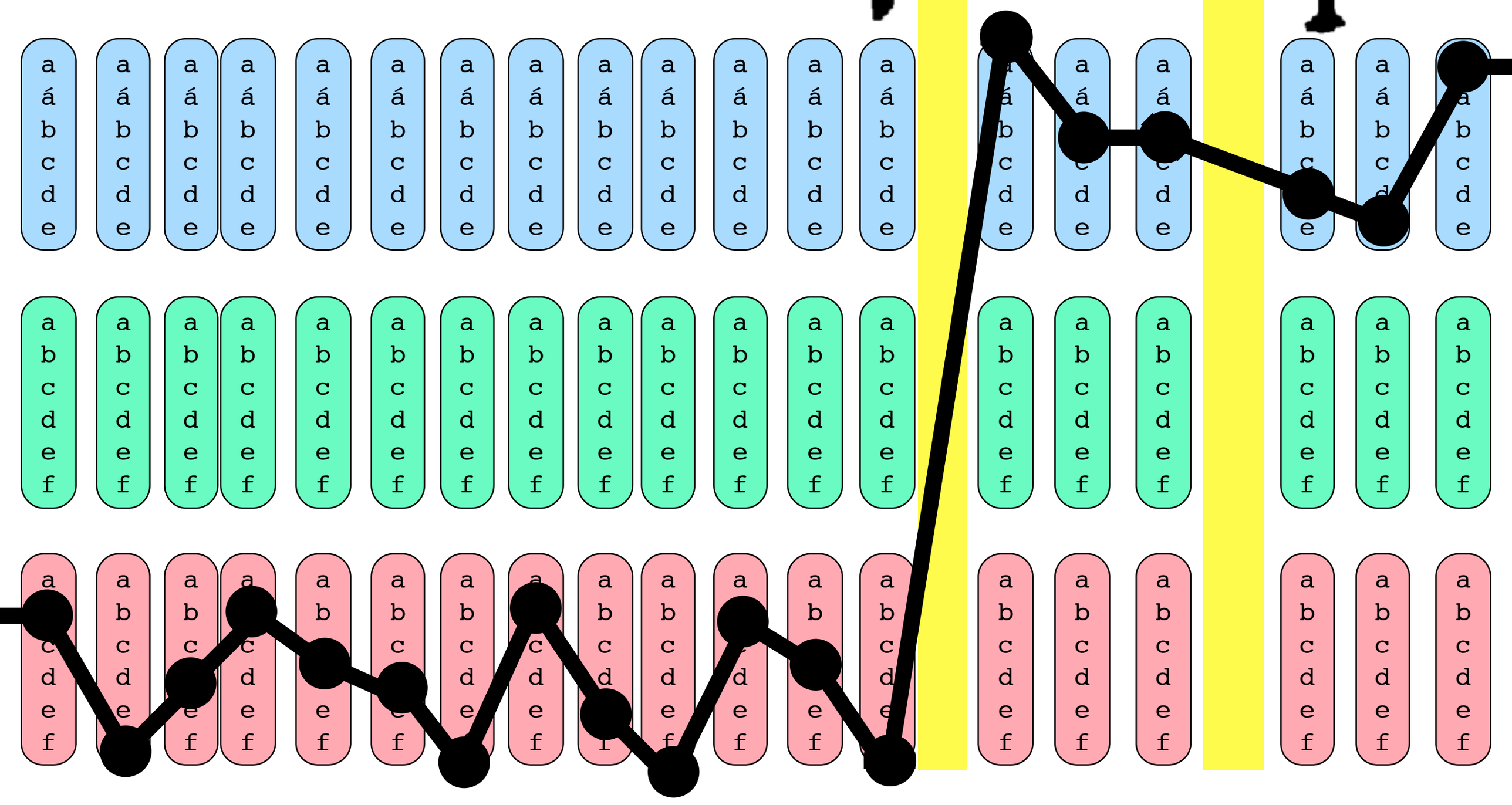
Latin

ABCDFGHIJKLMOPQRSTUVWXYZ
abcdfghijklmopqrstuvwxyz
01234567890.,/\()?!”’:;-

Nahuatl

ABCDFGHIJKLMOPQRSTUVWXYZ
abcdfghijklmopqrstuvwxyz
01234567890.,/\()?!”’:;-

ixtintzitzin, conque



Code-Switching Language Model

- Improves transcription quality, *and*

- Implicitly identifies language spans in text (metadata of the transcription)

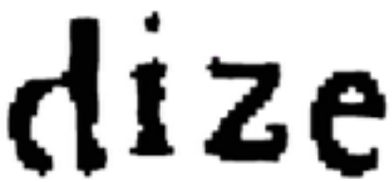
Orthographic Variability

Orthographic Variability

- We train our language models from available text (e.g. Project Gutenberg)

- Modern transcribers use modern spellings, which often do not match the printed documents

Orthographic Variability

	Transcription	Modern Form
	di**z**e	di**c**e
	n**u**mero	n**ú**mero
	D**õ**de	D**on**de

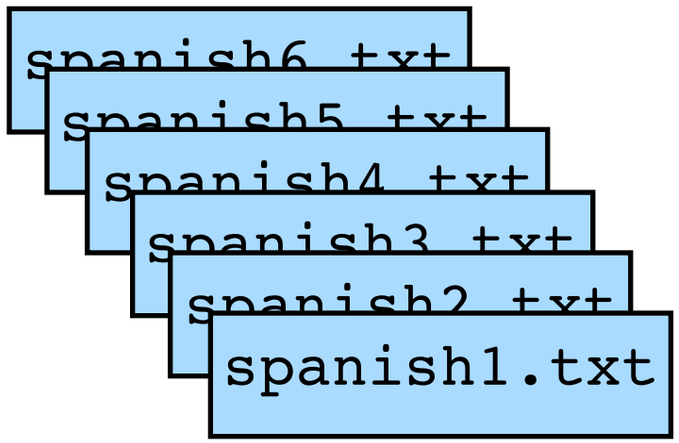
Orthographic Variability

Simple solution:

Modify the modern corpora to use old conventions.

Orthographic Variability

Modern Spanish



Replacement Rules

u → v

c → z

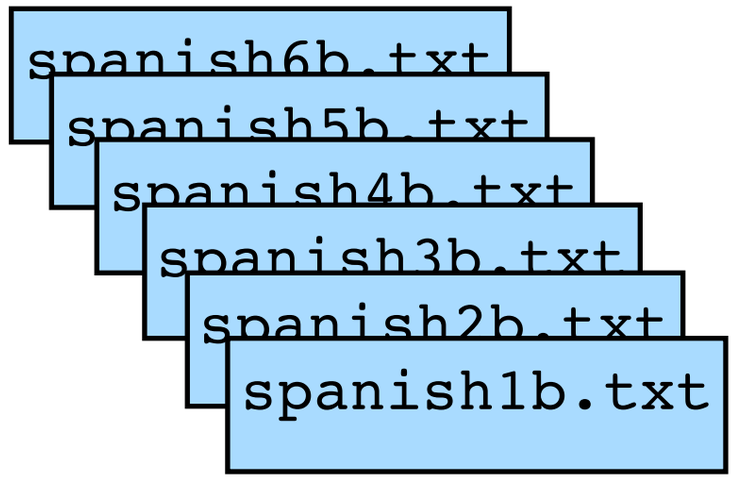
ú → u

on → õ

que → q̃

…

Old Spanish



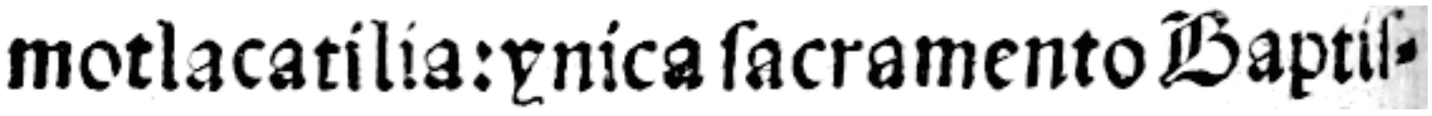
Experiments

Experiments

- Evaluated on five different books *Primeros Libros*
- Years 1553 to 1600
- Differing fonts, language proportions, clarity

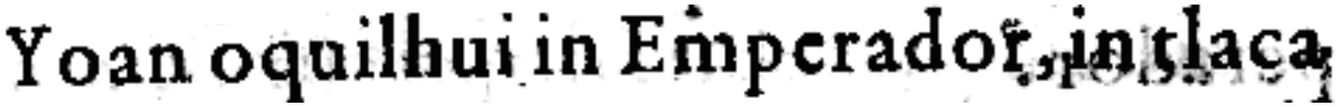
Unknown Fonts

Gante (1553)

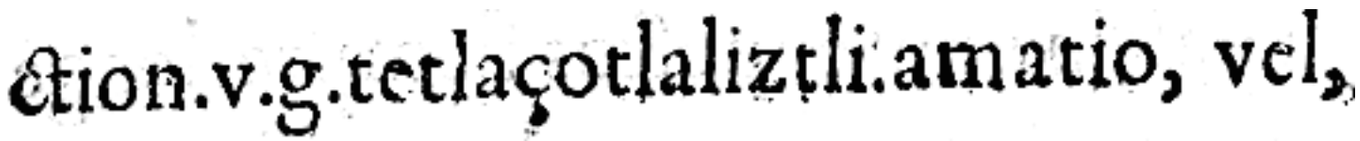


Anunciación (1565)

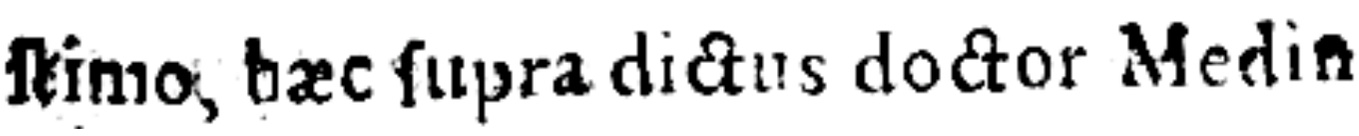
Sahagún (1583)



Rincón (1595)



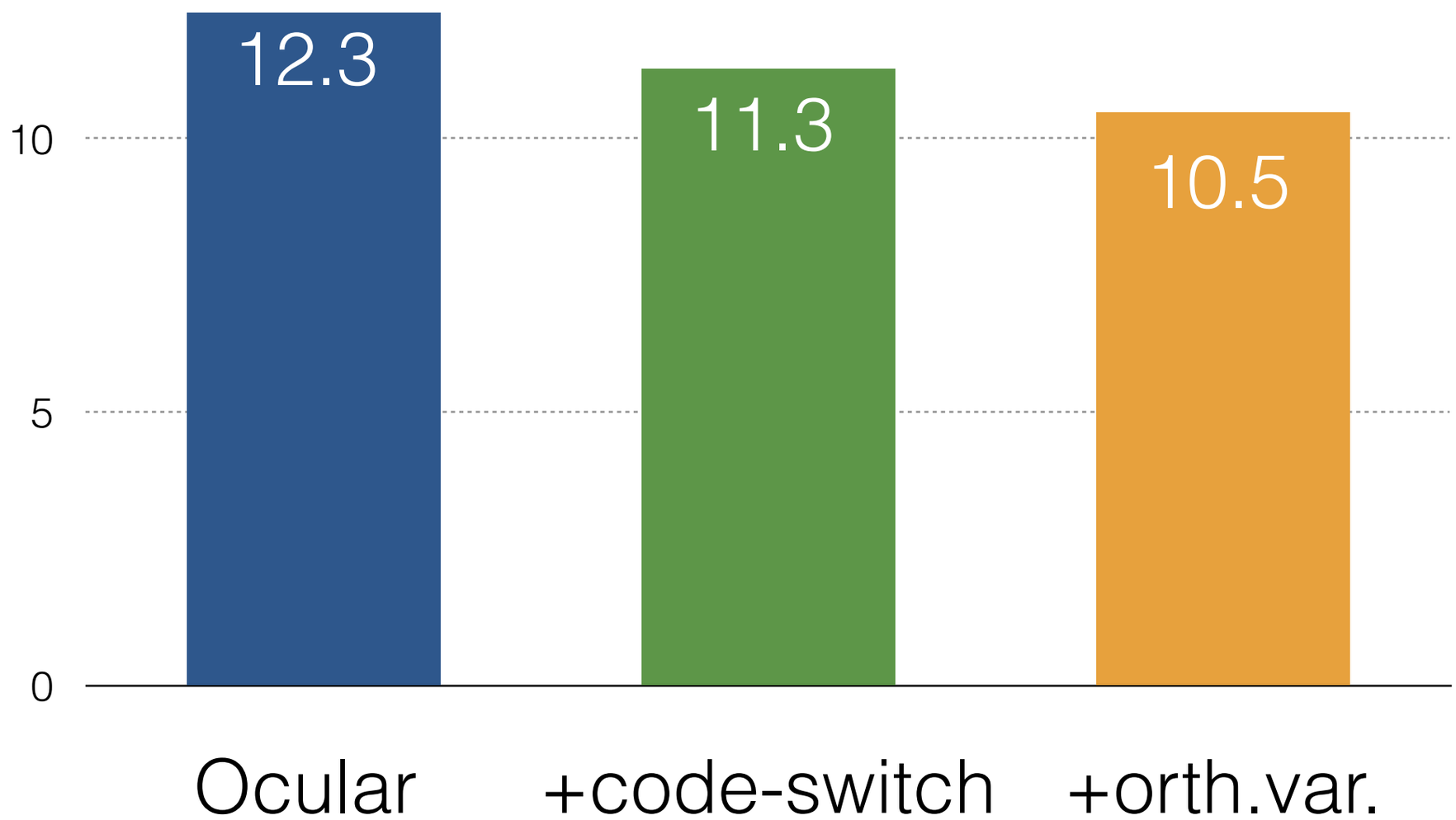
Bautista (1600)



Experimental Results

(lower is better. ~90% characters are correct)

Character
Error Rate



A thing we do well

Without handling orth. variation: merita

With handling orth. variation: mẽtira

Modern form: mentira

A thing we do wrong

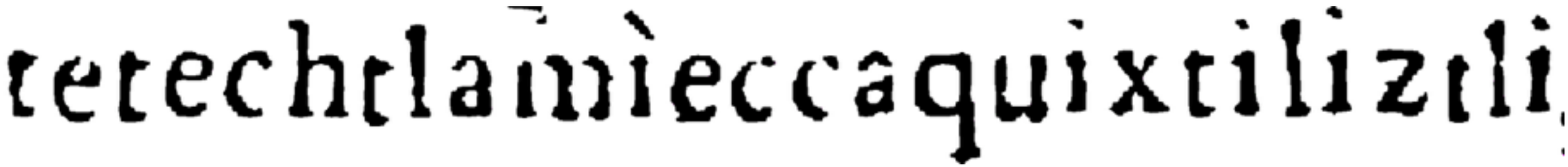
Los verbales en,li,o en tli, antepo-

A thing we do wrong

Model output:	**sí**	**tĩi**	← Spanish
Gold transcription:	**li**	**tli**	← Nahuatl

Model avoids switching languages, but this is actually from a description of Nahuatl grammar.

A thing that's hard



tetechtla miec caquixtiliztli

- All letters are correct, but the model adds spaces
- No agreed-upon standards for Nahuatl spacing
- Hard to evaluate what is "correct"

Conclusion

Conclusion

- By accounting for **multilingual text** and **obsolete orthography**, we can improve the state-of-the-art for historical OCR.
- These are common characteristics of texts from all over the world and from all eras.
- Expansion of OCR abilities means a wider range of texts may be available for study.