Carnegie Mellon University

Unlocking Resources for Under-resourced Languages

Graham Neubig

Shruti Rijhwani, Xinyi Wang

Antonios Anastasopoulos, Daisy Rosenblum, Sebastian Ruder

Language technologies in a multilingual world

Language technologies in a multilingual world

Considerable recent progress in expanding NLP to many languages!

Language technologies in a multilingual world

Considerable recent progress in expanding NLP to many languages!

Multilingual benchmark datasets for NLP tasks

WMT MasakhaNER
TyDiQA XNLI FLORES-101
UD Treebank XTREME

Language technologies in a multilingual world

Considerable recent progress in expanding NLP to many languages!

Multilingual benchmark datasets for NLP tasks

WMT MasakhaNER
TyDiQA XNLI FLORES-101
UD Treebank XTREME

Pretrained multilingual language models

XLM-R mBERT
mT5 mBART ERNIE-M
Turing ULR

Language technologies in a multilingual world

Considerable recent progress in expanding NLP to many languages!

Multilingual benchmark datasets for NLP tasks

WMT MasakhaNER
TyDiQA XNLI FLORES-101
UD Treebank XTREME

Pretrained multilingual language models

XLM-R mBERT
mT5 mBART ERNIE-M
Turing ULR

Commercial models that support many languages

Voice assistants
Predictive keyboards Translation
Web search Text analytics

Language technologies in a multilingual world

Pretrained multilingual language models

XLM-R mBERT

mT5 mBART ERNIE-M

Turing ULR

Language technologies in a multilingual world

Pretrained multilingual language models

XLM-R mBERT
mT5 mBART ERNIE-M
Turing ULR

Trained on unlabeled text corpora
(e.g., Wikipedia and Common Crawl)

Language technologies in a multilingual world

Pretrained multilingual language models

XLM-R mBERT
mT5 mBART ERNIE-M
Turing ULR

Support 100 – 200 languages

Language technologies in a multilingual world

Pretrained multilingual language models

XLM-R mBERT
mT5 mBART ERNIE-M
Turing ULR

Support 100 – 200 languages

Enables NLP applications through cross-lingual transfer

- Named entity recognition
- Entity linking
- Web search
- Machine translation
- Question answering

…

Language technologies in a multilingual world

Pretrained multilingual language models

XLM-R mBERT
mT5 mBART ERNIE-M
Turing ULR

Support 100 – 200 languages

Language technologies in a multilingual world

Pretrained multilingual language models

XLM-R mBERT
mT5 mBART ERNIE-M
Turing ULR

Support 100 – 200 languages

Multiple societal benefits of NLP that includes many languages!

Language technologies in a multilingual world

Pretrained multilingual language models

XLM-R mBERT
mT5 mBART ERNIE-M
Turing ULR

Support 100 – 200 languages

Automatic translation of information

Multiple societal benefits of NLP that includes many languages!

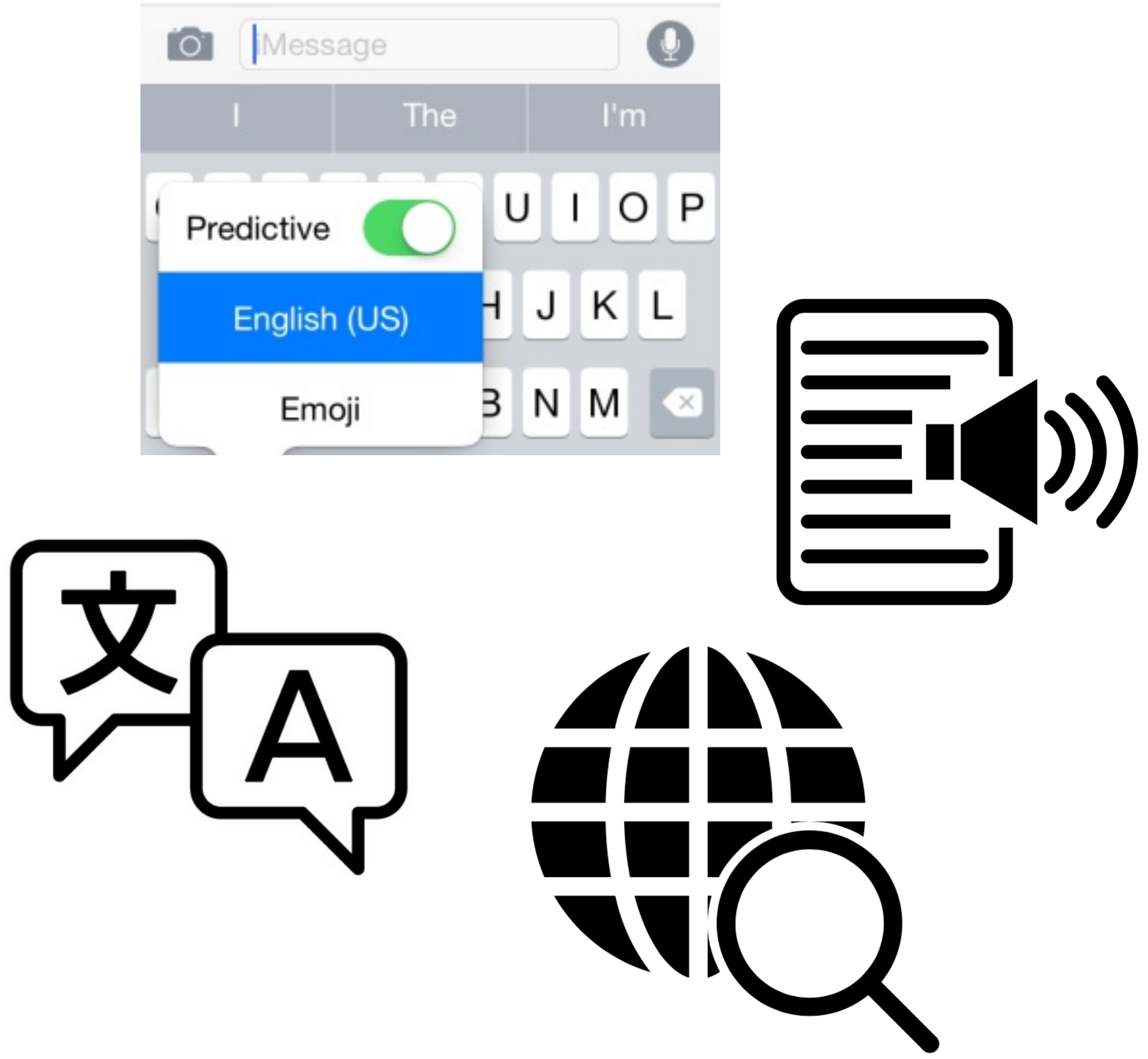
Access to information and education from other languages

Language technologies in a multilingual world

Pretrained multilingual language models

XLM-R mBERT
mT5 mBART ERNIE-M
Turing ULR

Support 100 – 200 languages



Multiple societal benefits of NLP that includes many languages!

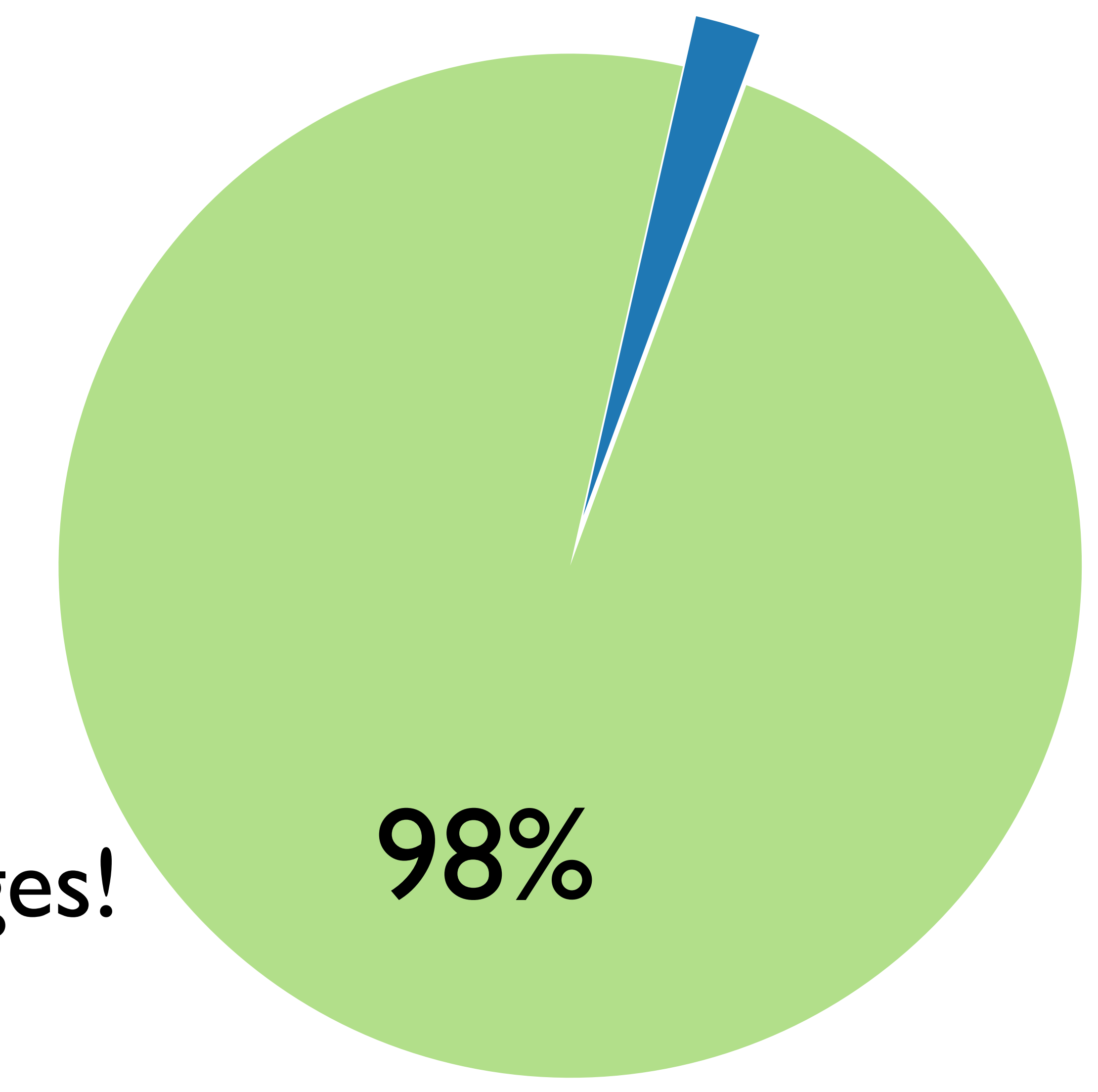
Access to information and education from other languages

Language technologies that serve many more people!

Language technologies in a multilingual world

Pretrained multilingual language models

XLM-R mBERT
mT5 mBART ERNIE-M
Turing ULR

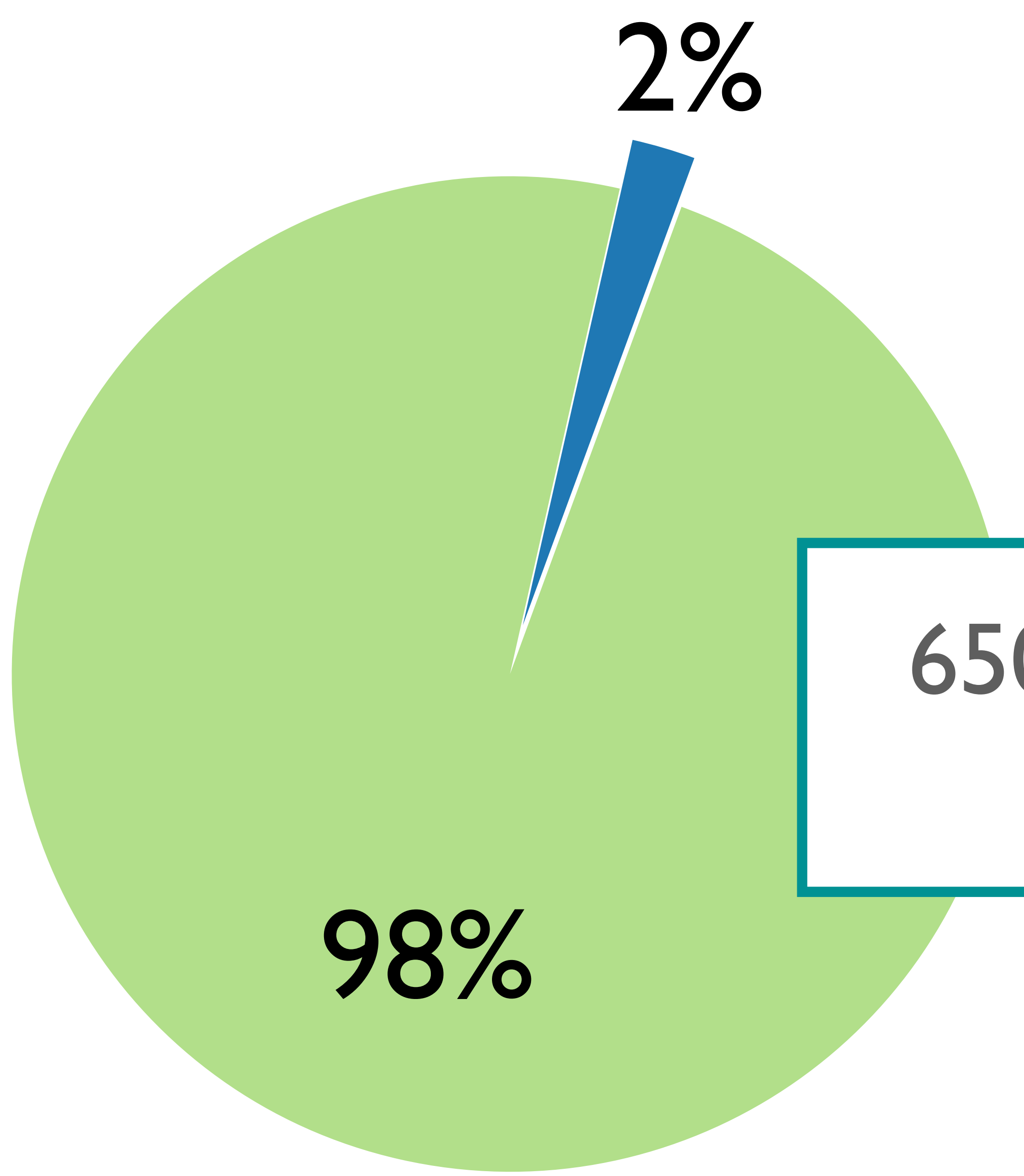


Language technologies in a multilingual world

Pretrained multilingual language models

XLM-R mBERT
mT5 mBART ERNIE-M
Turing ULR

Support 100 – 200 languages



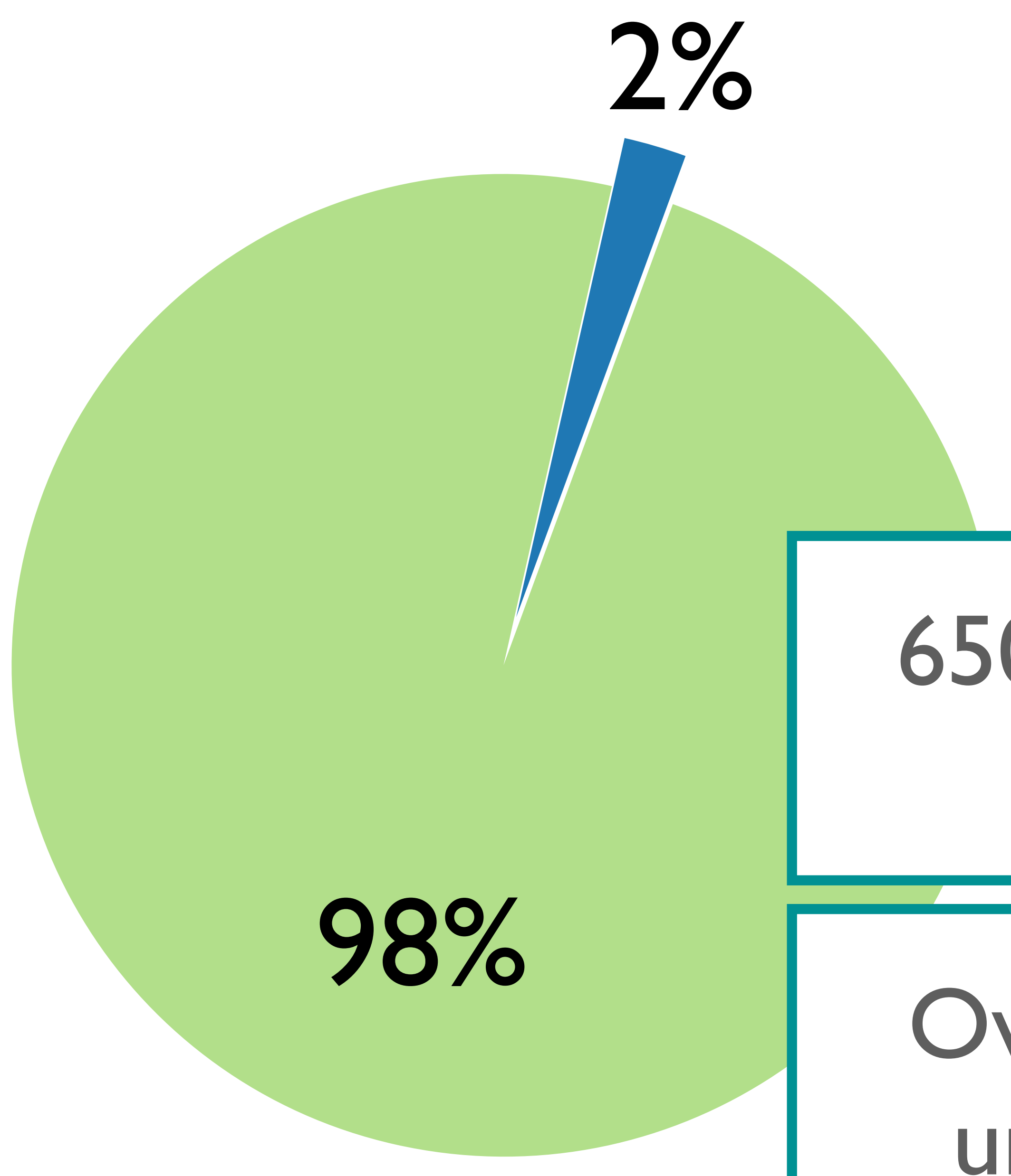
6500+ languages are under-represented in NLP

Language technologies in a multilingual world

Pretrained multilingual language models

XLM-R mBERT
mT5 mBART ERNIE-M
Turing ULR

Support 100 – 200 languages



6500+ languages are under-represented in NLP

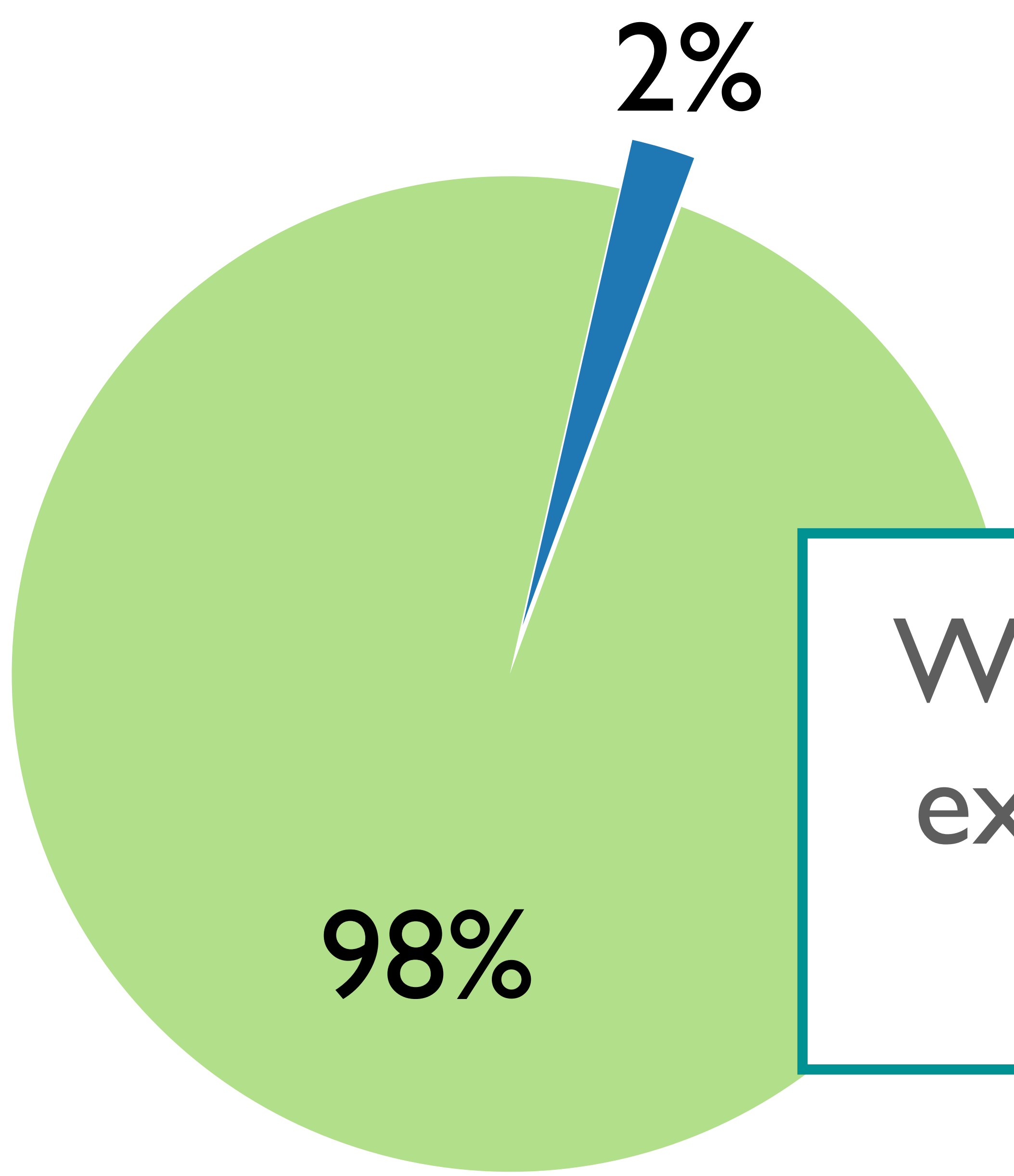
Over 2.2 billion people are under-served by modern language technologies

Language technologies in a multilingual world

Pretrained multilingual language models

XLM-R mBERT
mT5 mBART ERNIE-M
Turing ULR

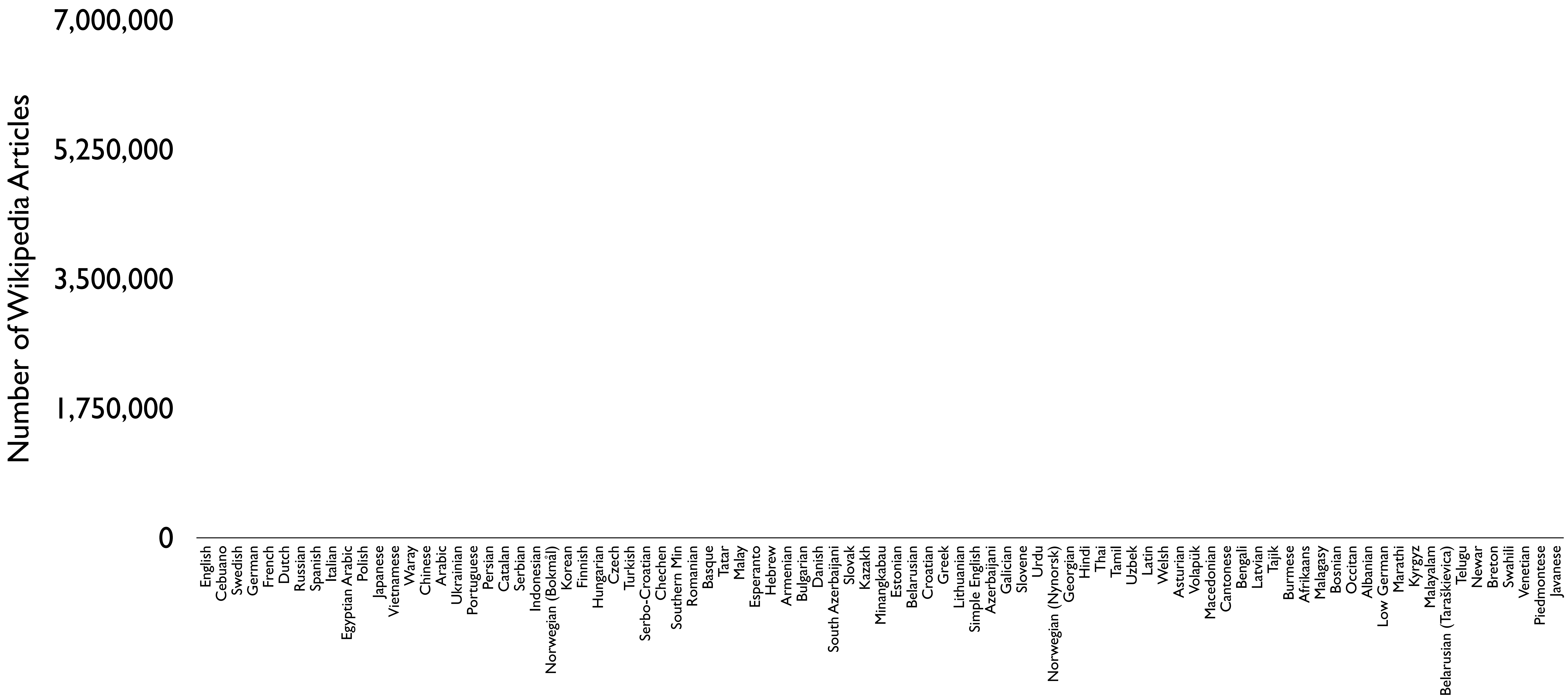
Support 100 – 200 languages



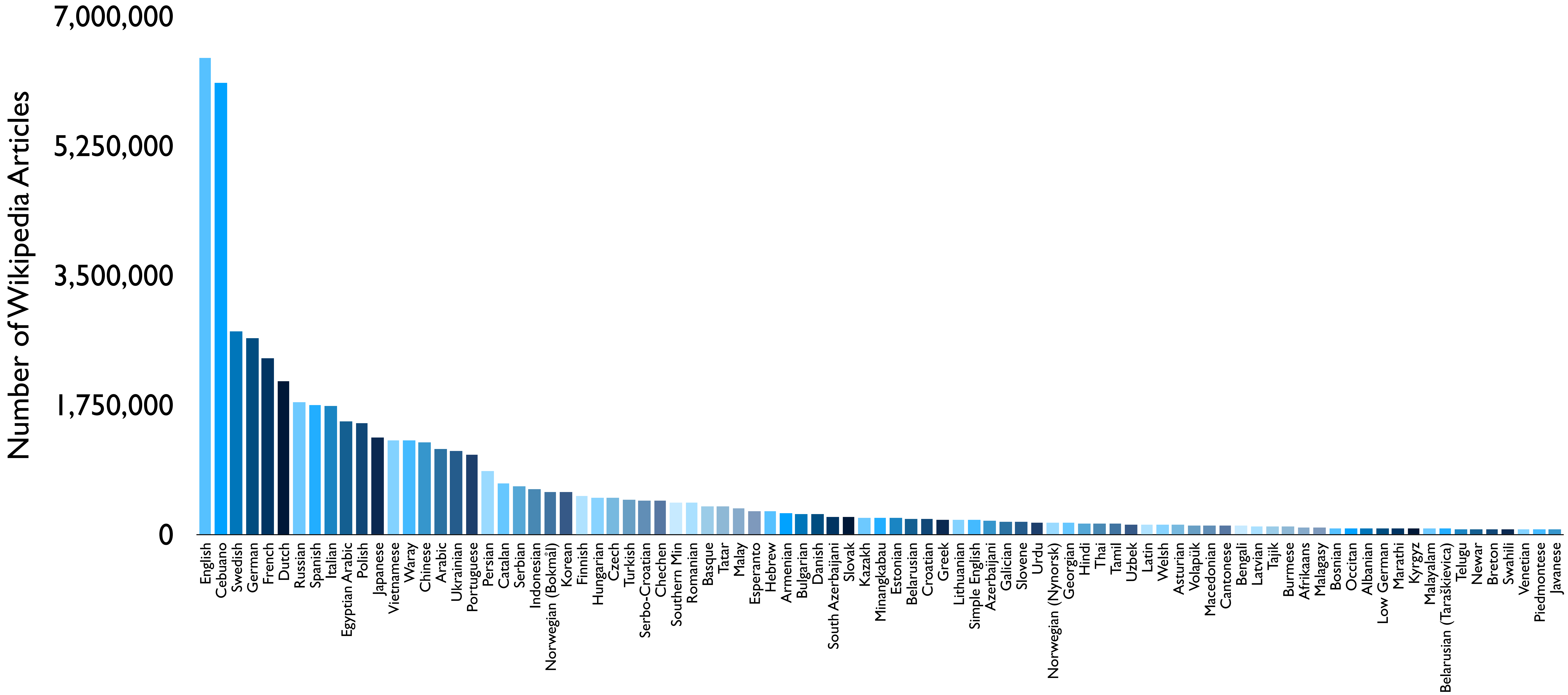
What's stopping us from expanding NLP systems to more languages?

The unlabeled text bottleneck

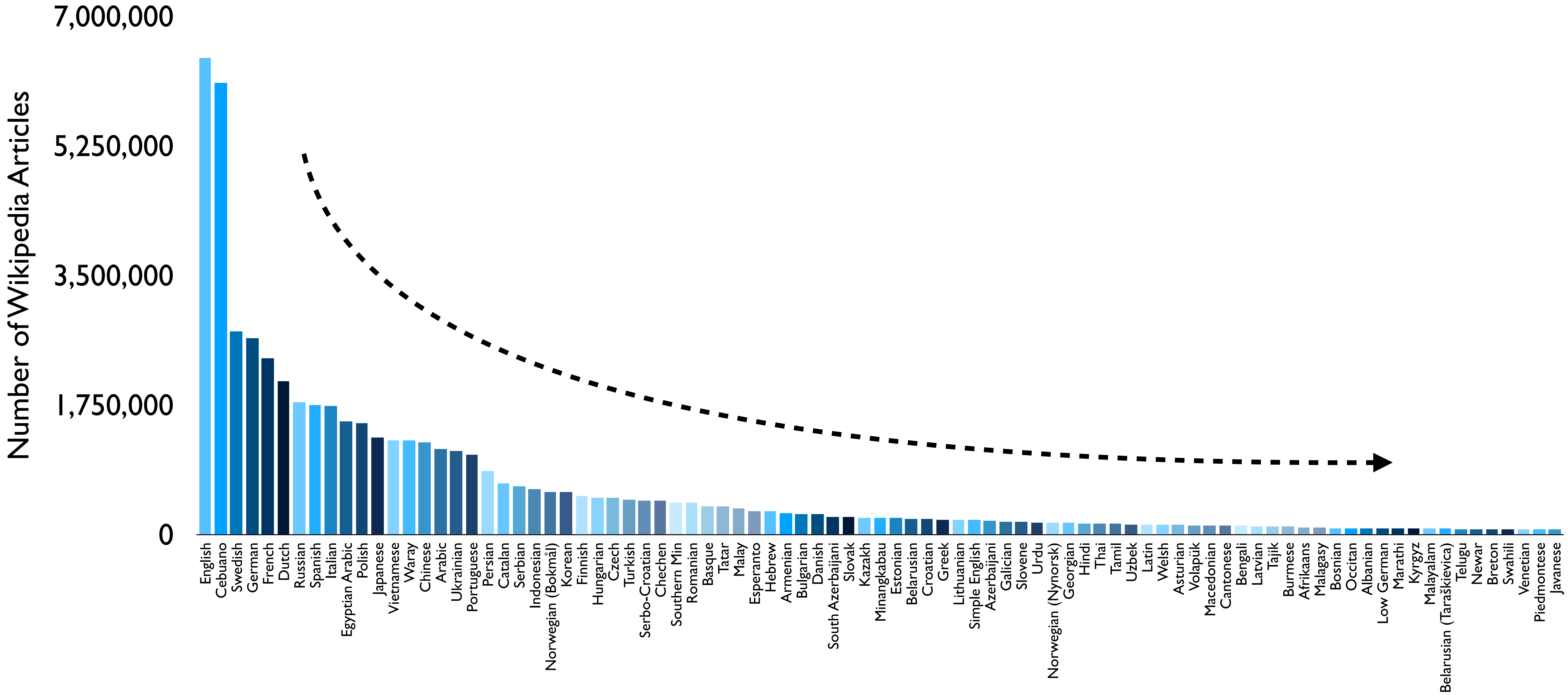
The unlabeled text bottleneck



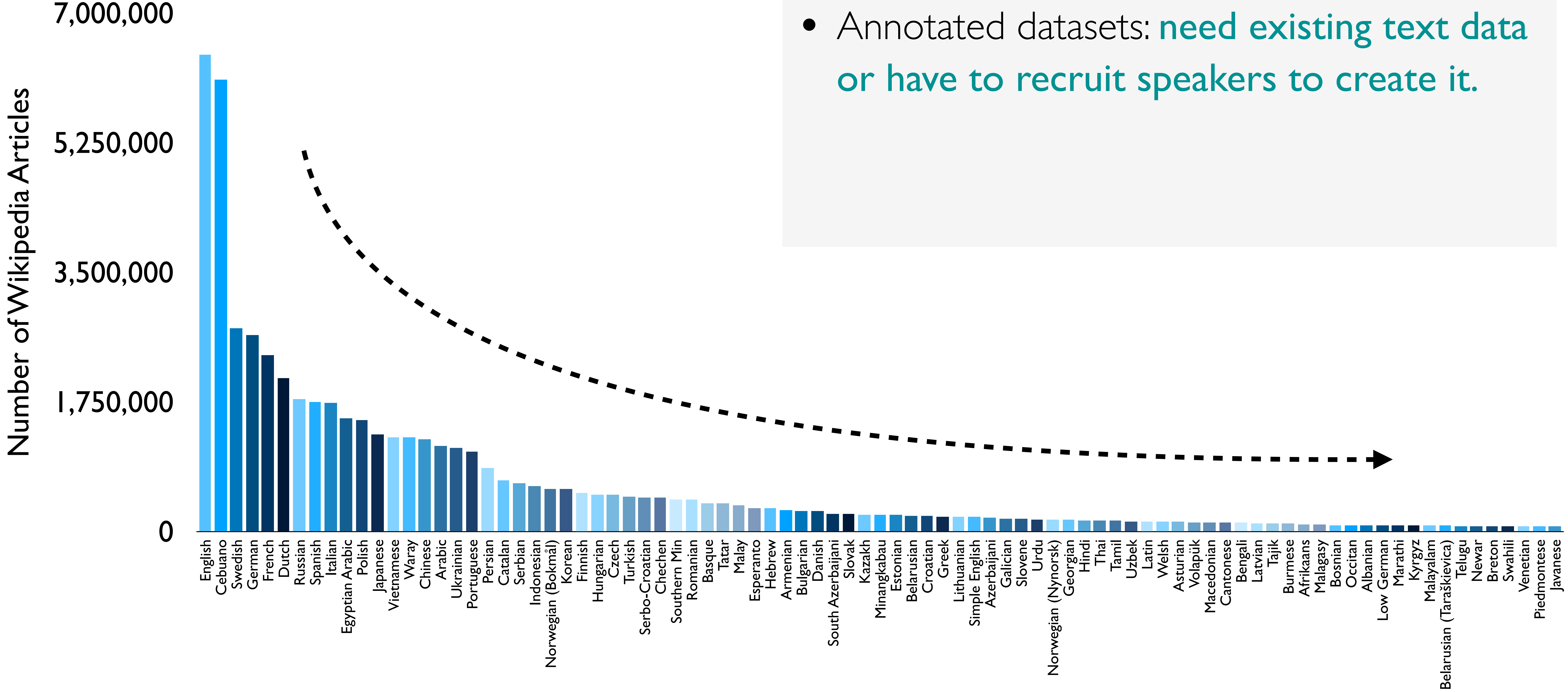
The unlabeled text bottleneck



The unlabeled text bottleneck

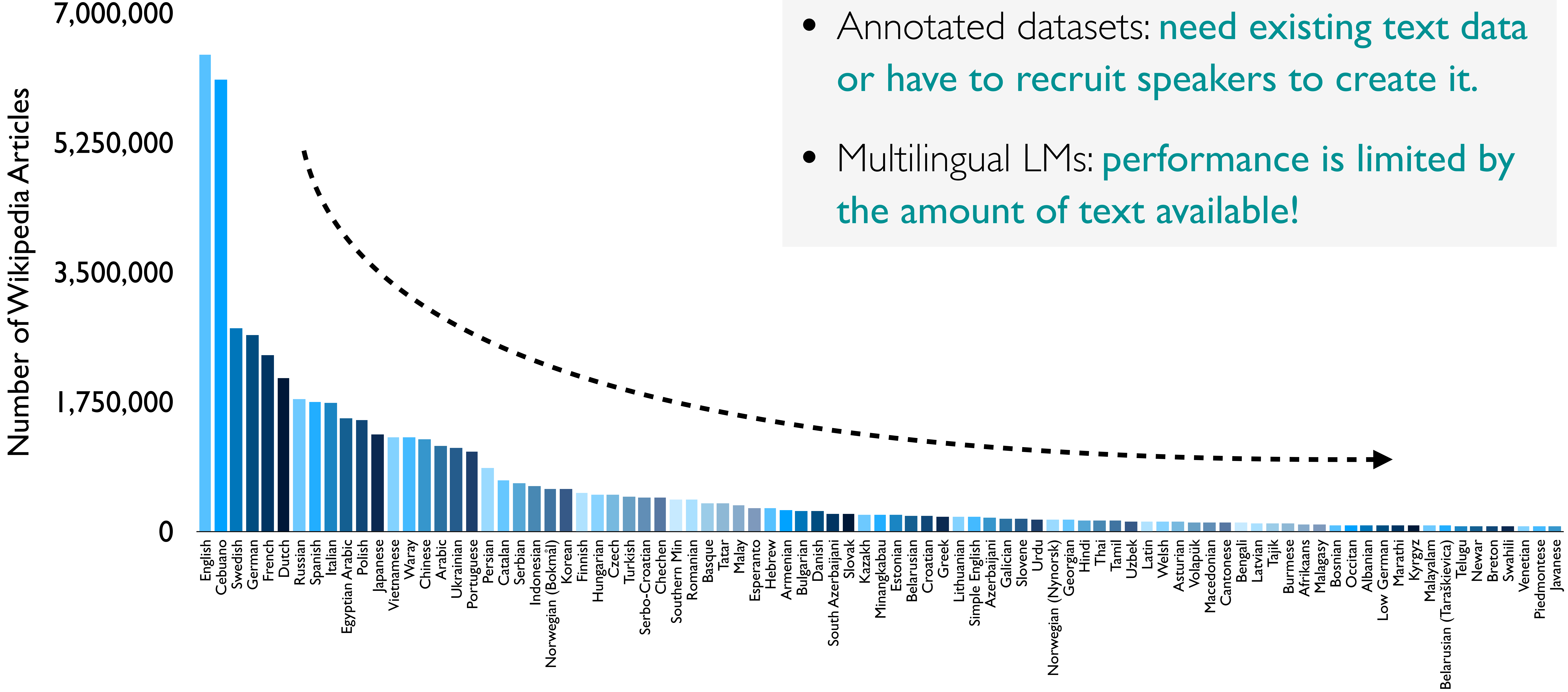


The unlabeled text bottleneck



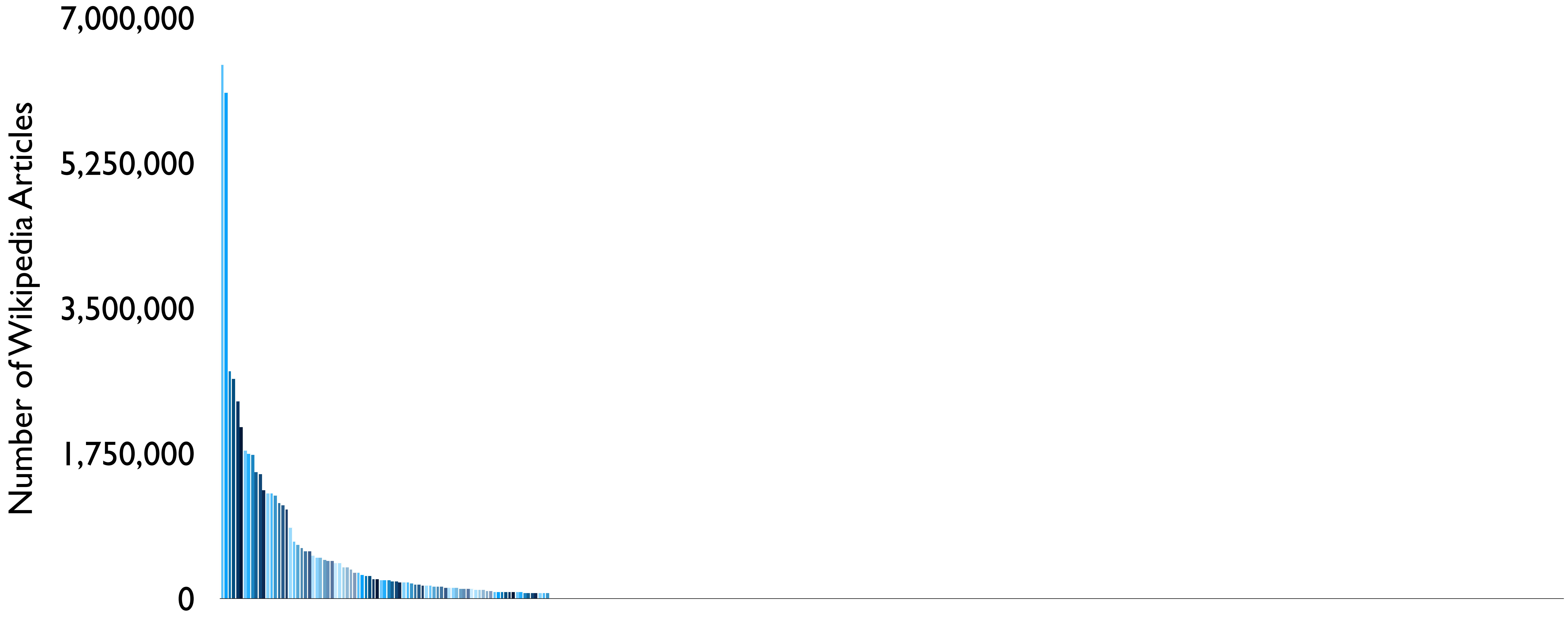
- Annotated datasets: need existing text data or have to recruit speakers to create it.

The unlabeled text bottleneck

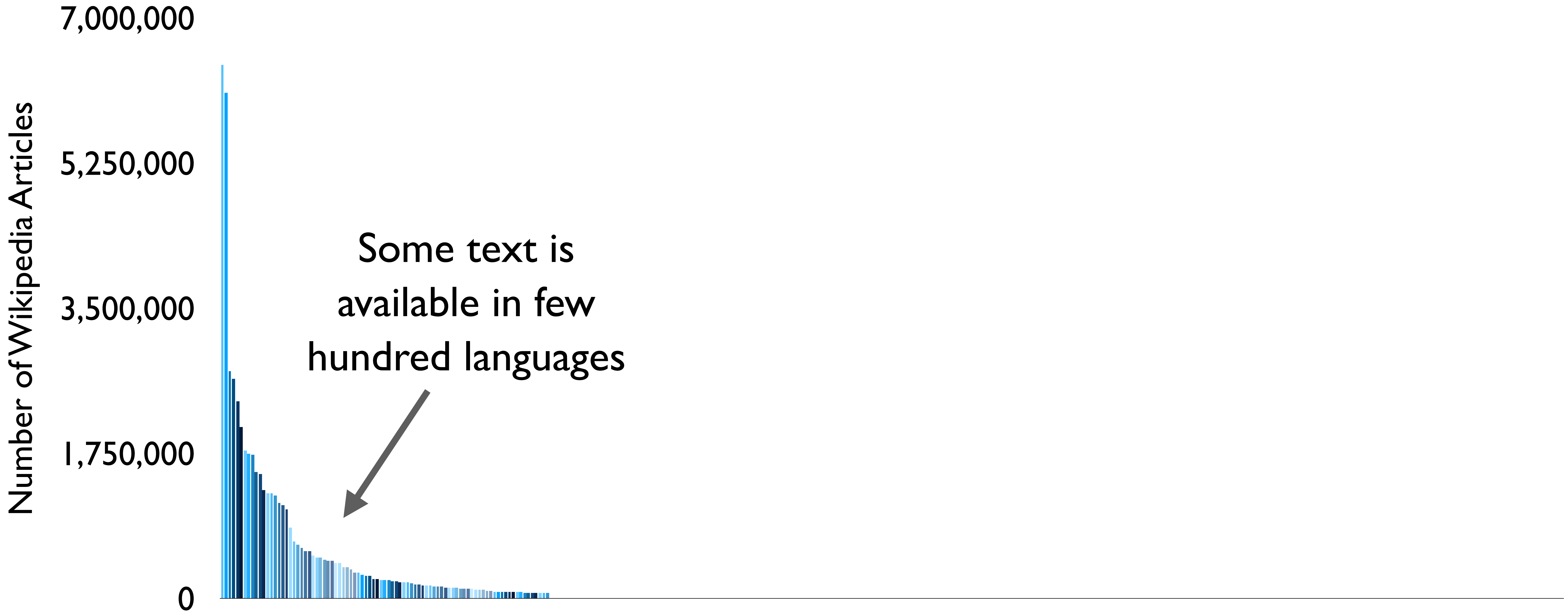


- Annotated datasets: need existing text data or have to recruit speakers to create it.
- Multilingual LMs: performance is limited by the amount of text available!

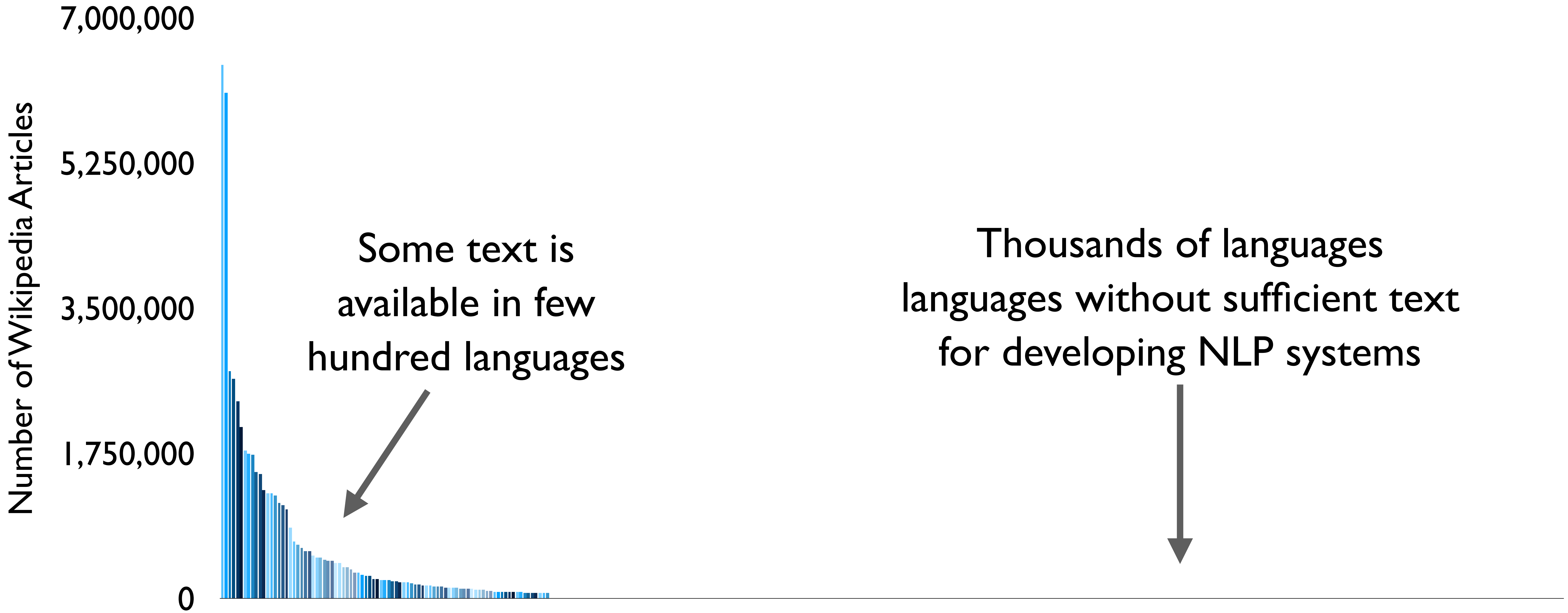
The unlabeled text bottleneck



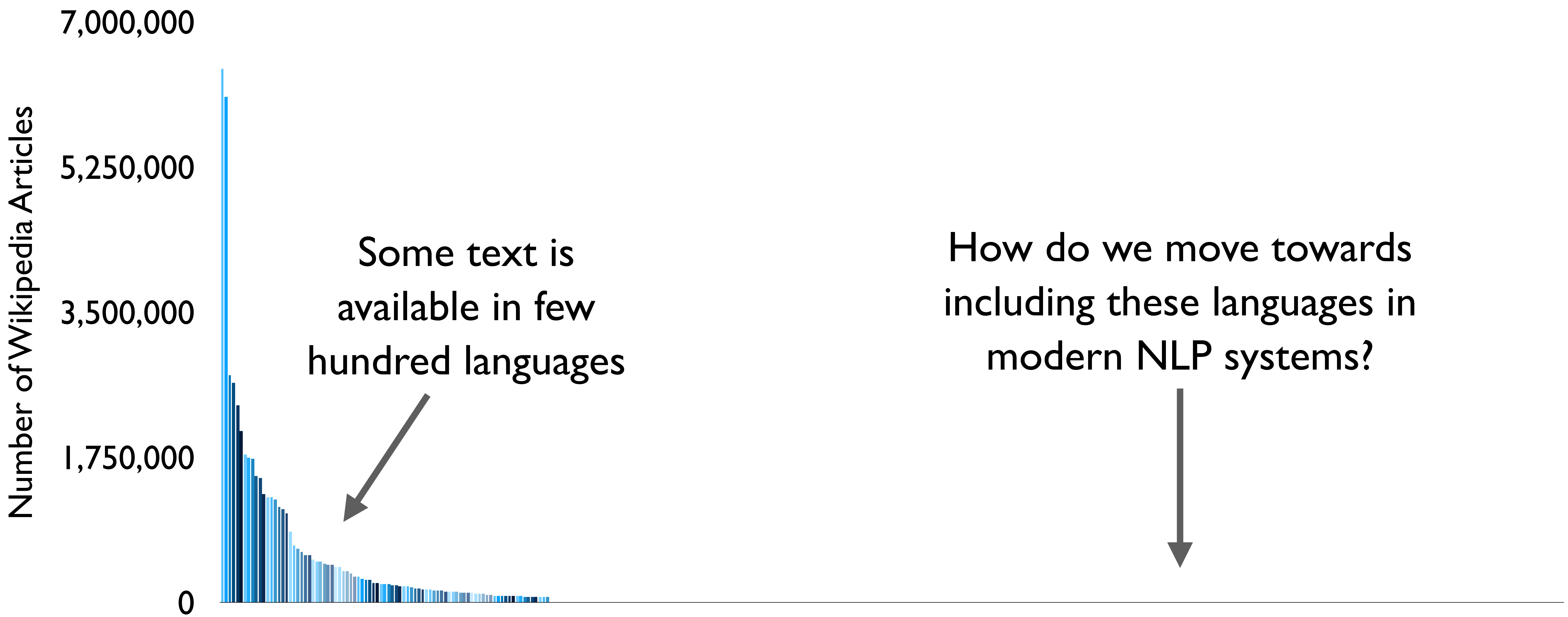
The unlabeled text bottleneck



The unlabeled text bottleneck



The unlabeled text bottleneck



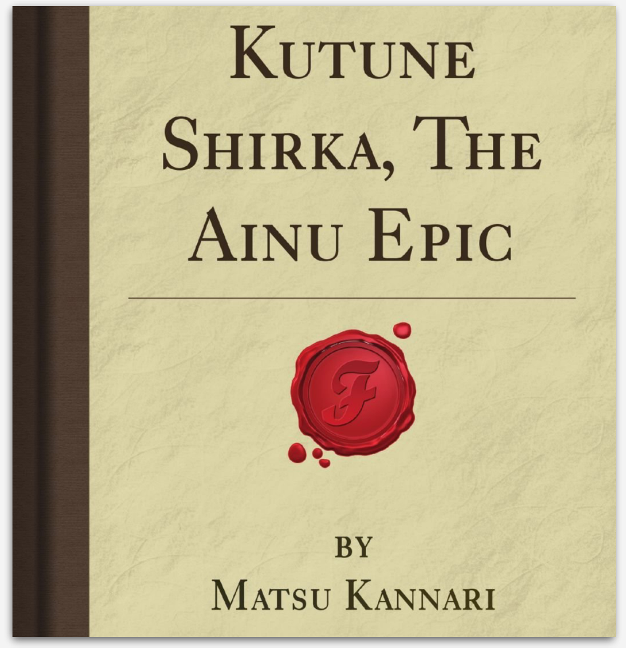
Text resources do exist in many more languages!

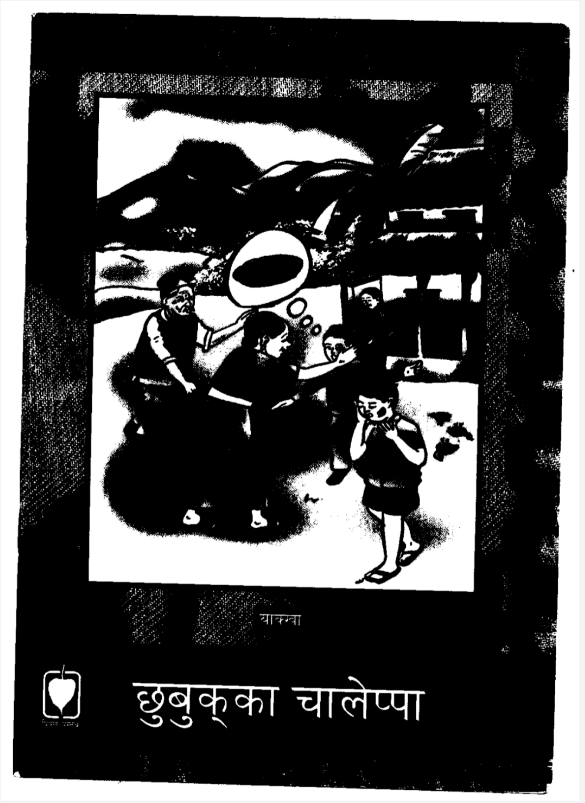
Text resources do exist in many more languages!

But locked away in formats that are not machine-readable

Text resources do exist in many more languages!

But locked away in formats that are not machine-readable

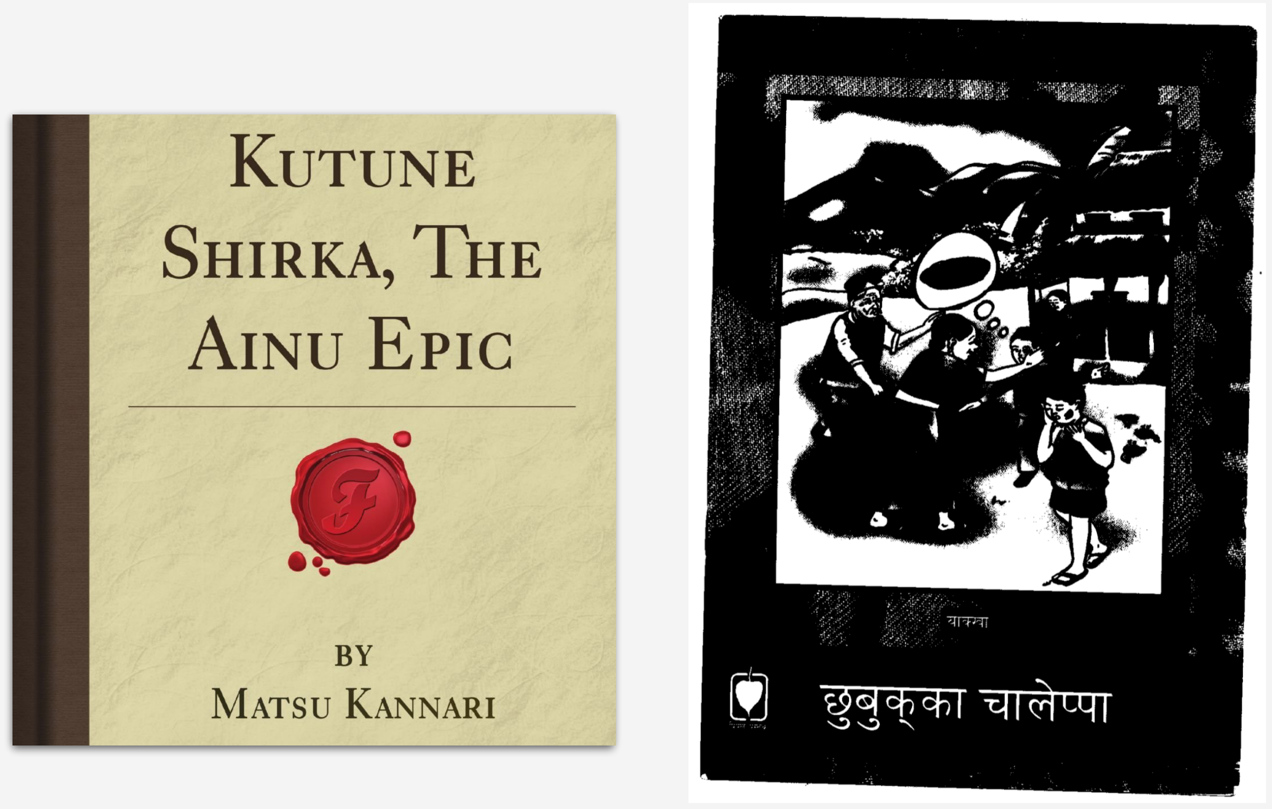




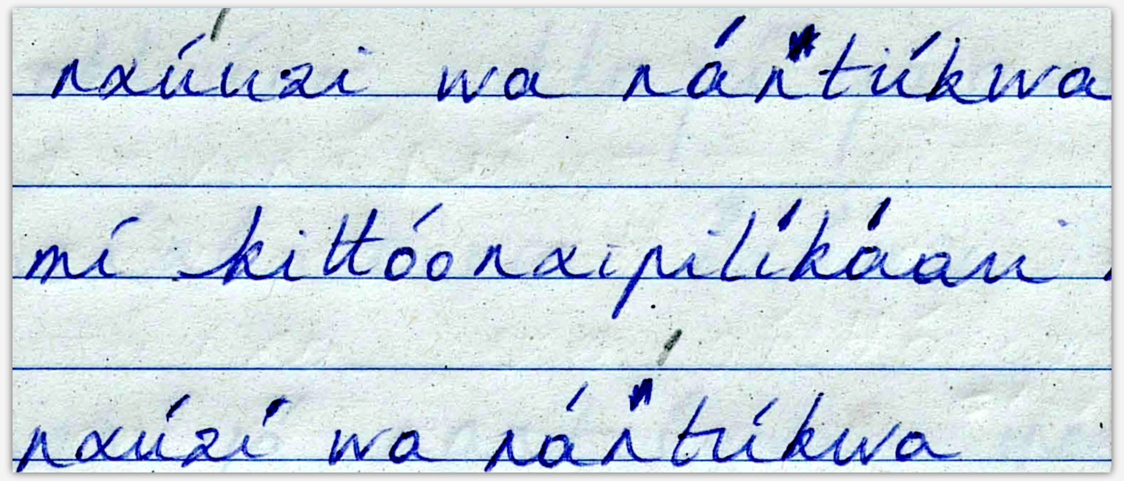
Printed books

Text resources do exist in many more languages!

But locked away in formats that are not machine-readable



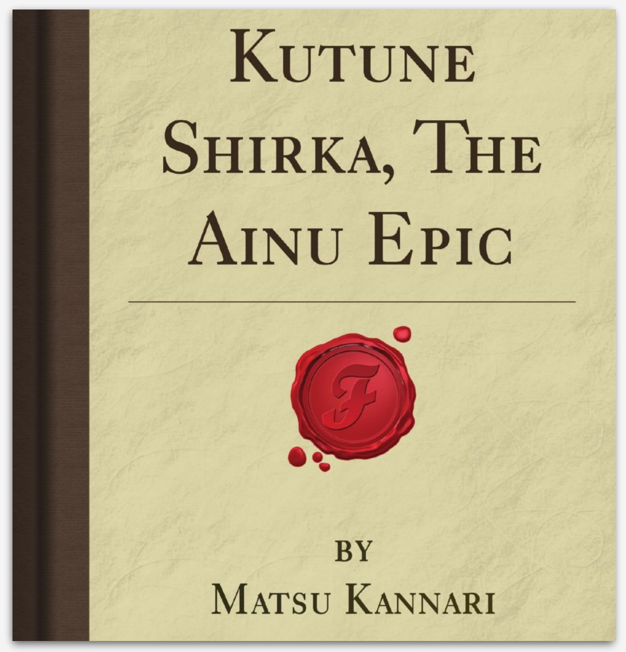
Printed books

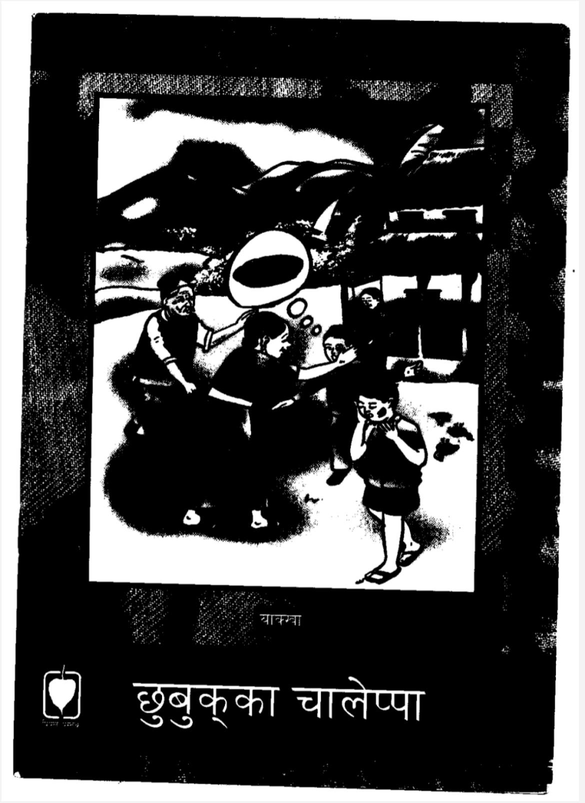


Handwritten notes

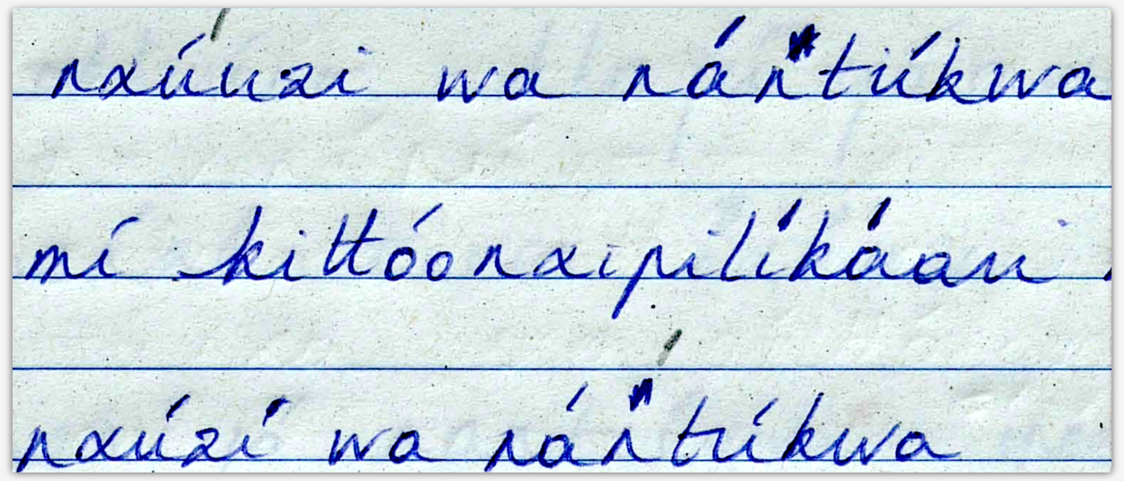
Text resources do exist in many more languages!

But locked away in formats that are not machine-readable

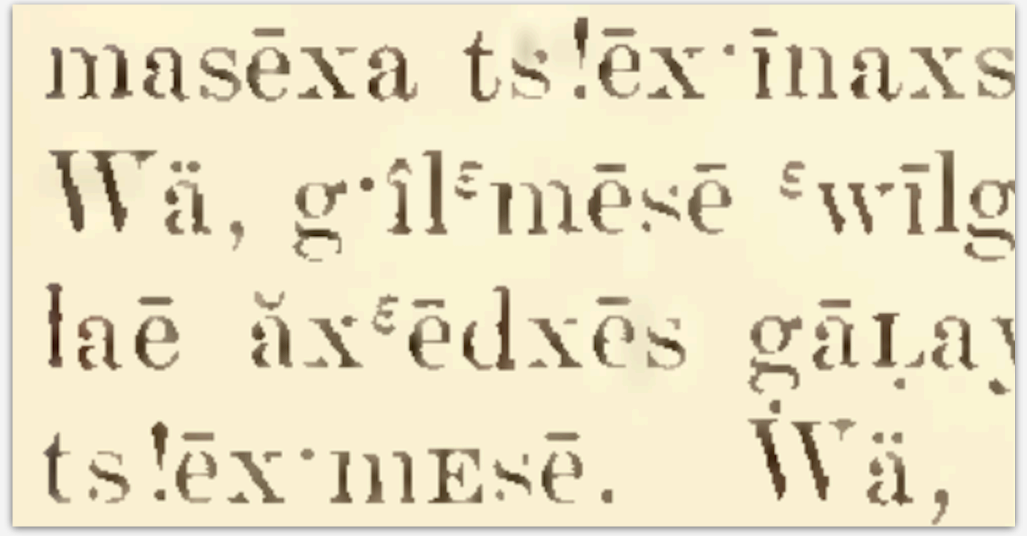




Printed books


naúuzi wa nántťúkwa
mí kittóonaipilikáan
naúzi wa nántťúkwa

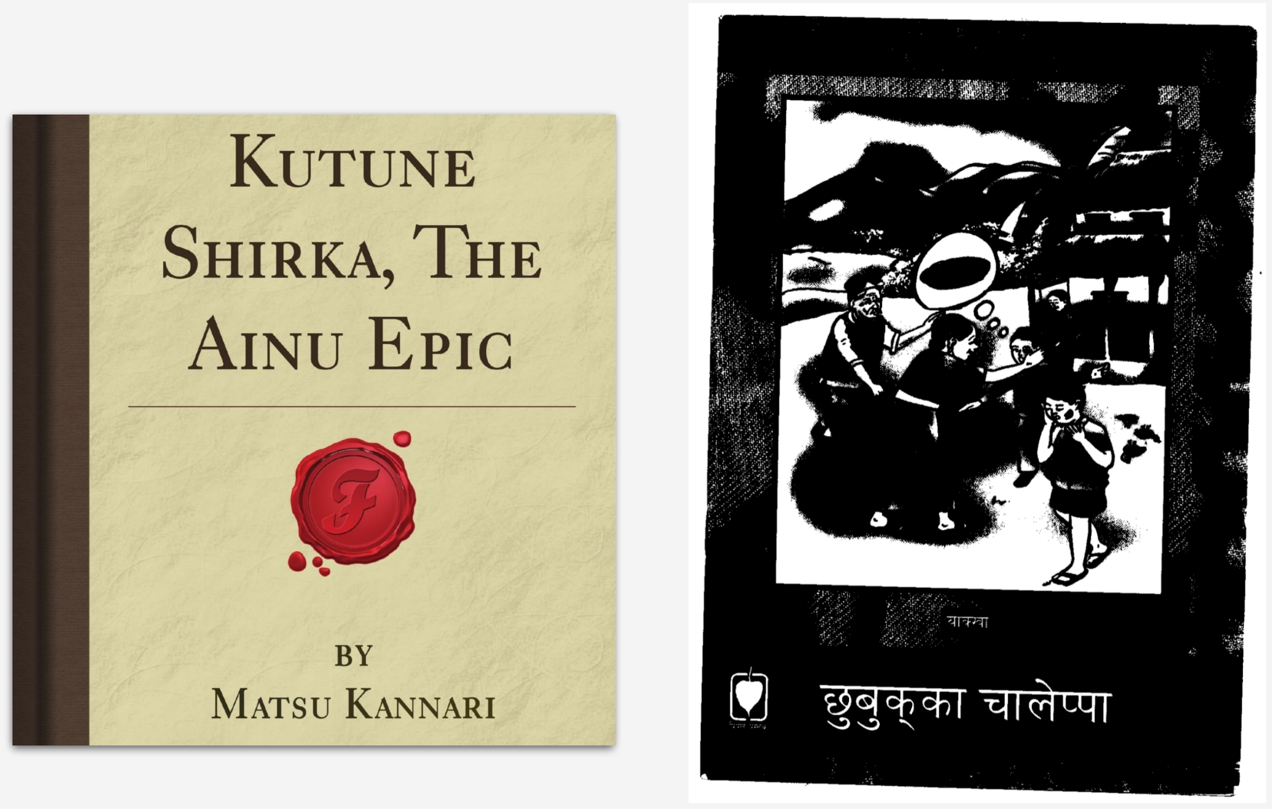
Handwritten notes


masēxa ts!ēx·inaxs
Wä, g·îlᵋmēsē ᵋwīlg
laē ăxᵋēdxēs gāLay
ts!ēx·mEsē. Wä,

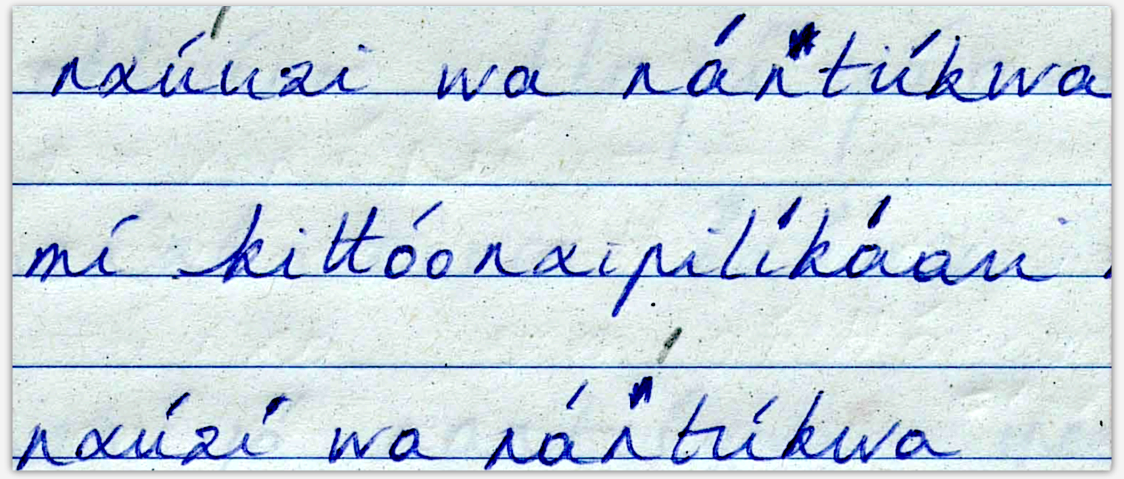
Typewritten documents

Text resources do exist in many more languages!

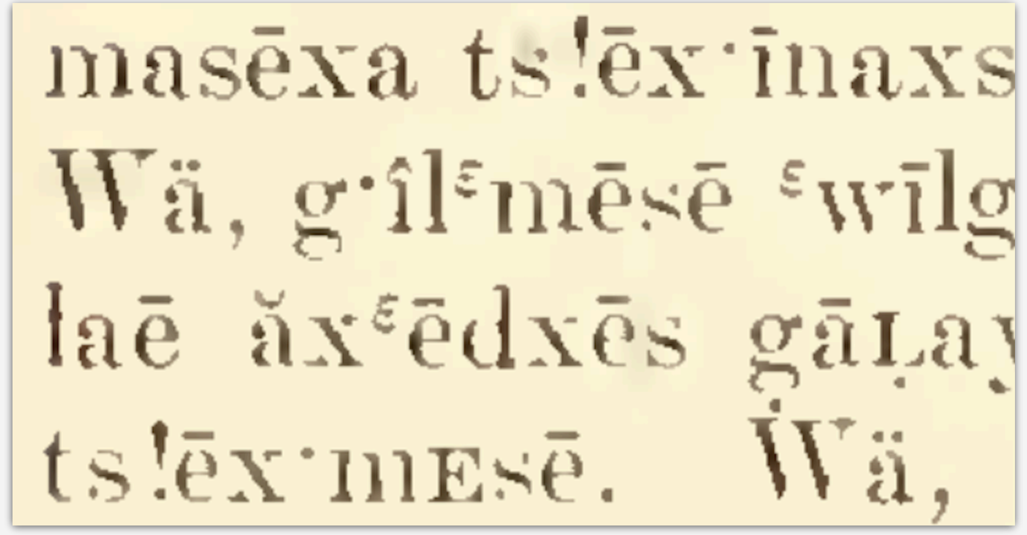
But locked away in formats that are not machine-readable



Printed books



Handwritten notes

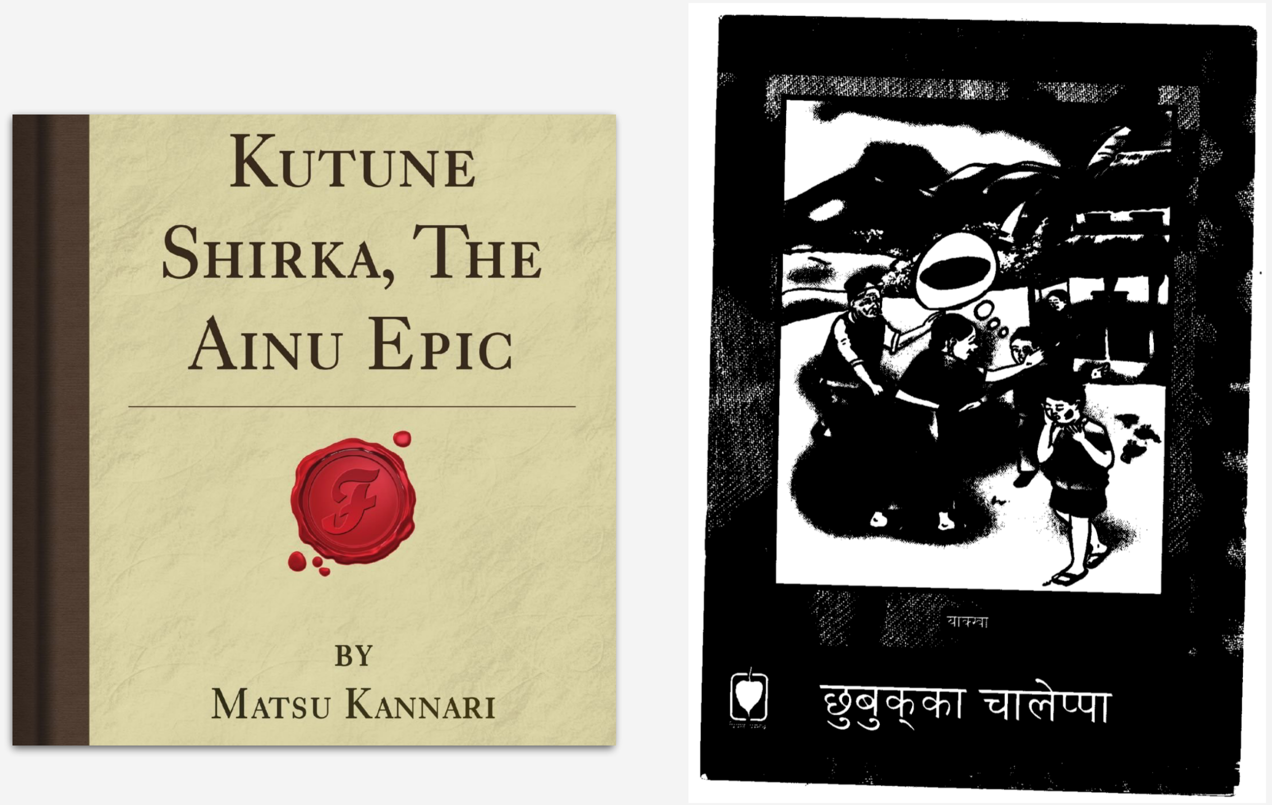


Typewritten documents

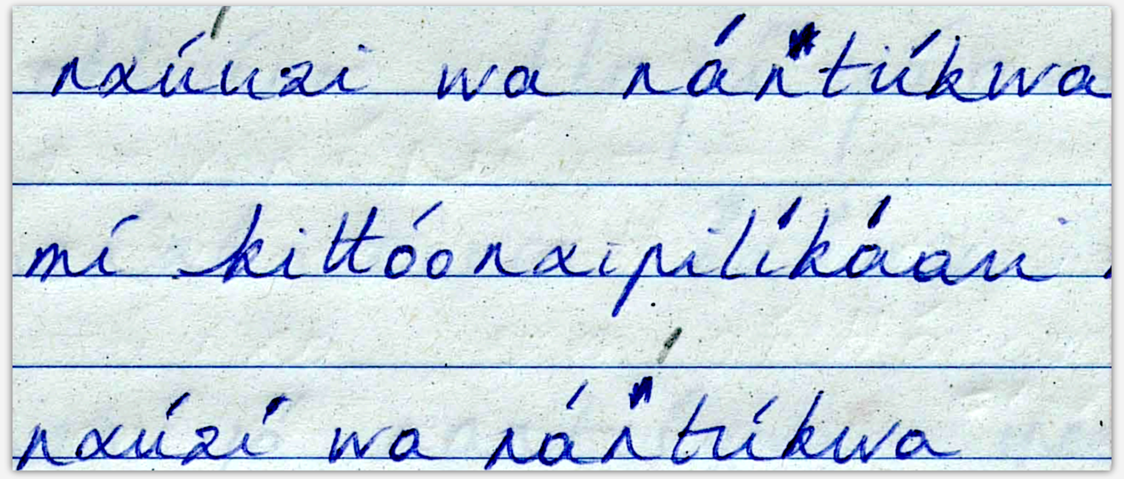
or in other formats such as bilingual lexicons

Text resources do exist in many more languages!

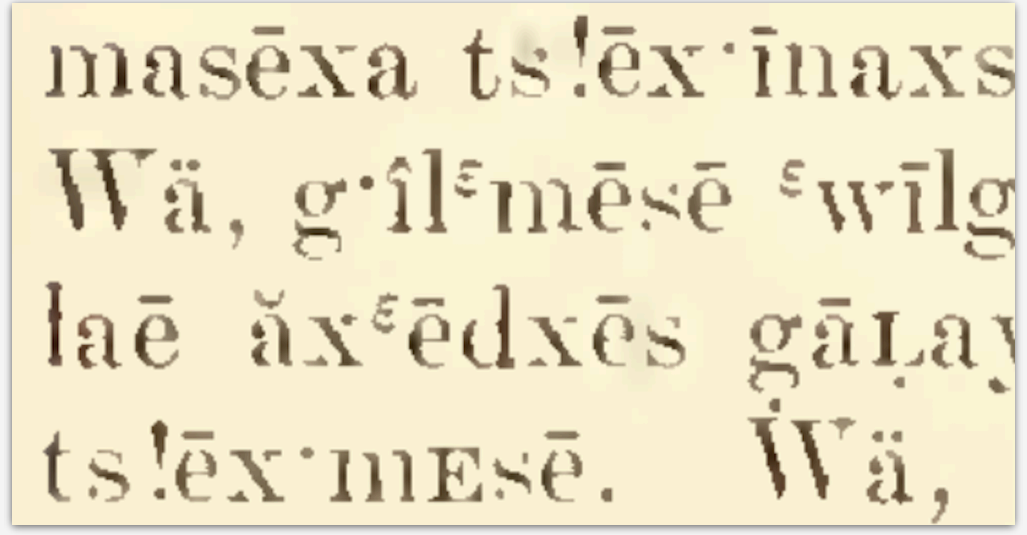
But locked away in formats that are not machine-readable



Printed books



Handwritten notes



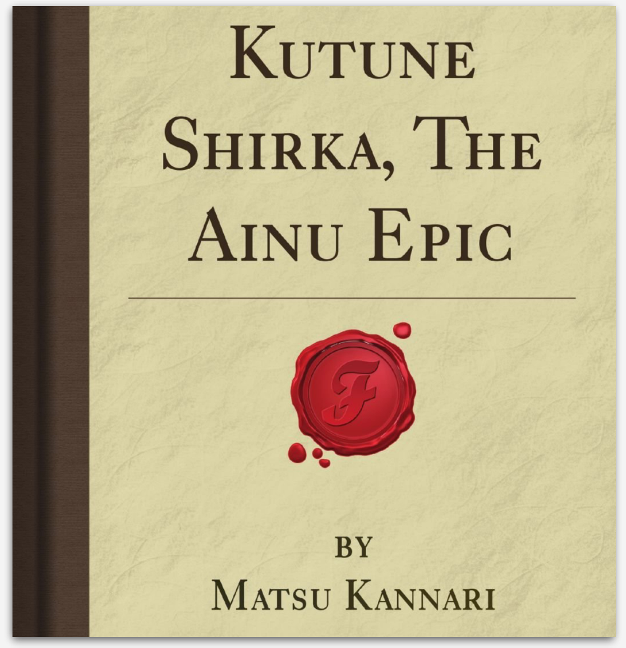
Typewritten documents

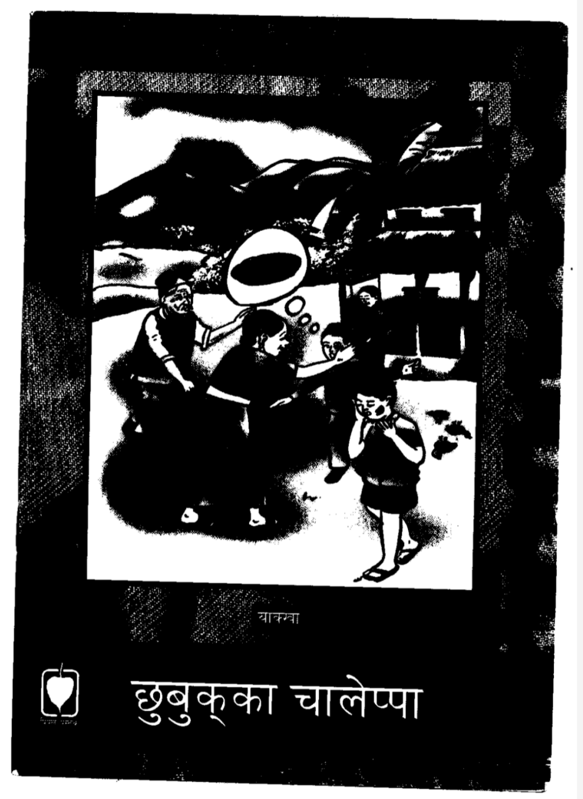
or in other formats such as bilingual lexicons

PanLex

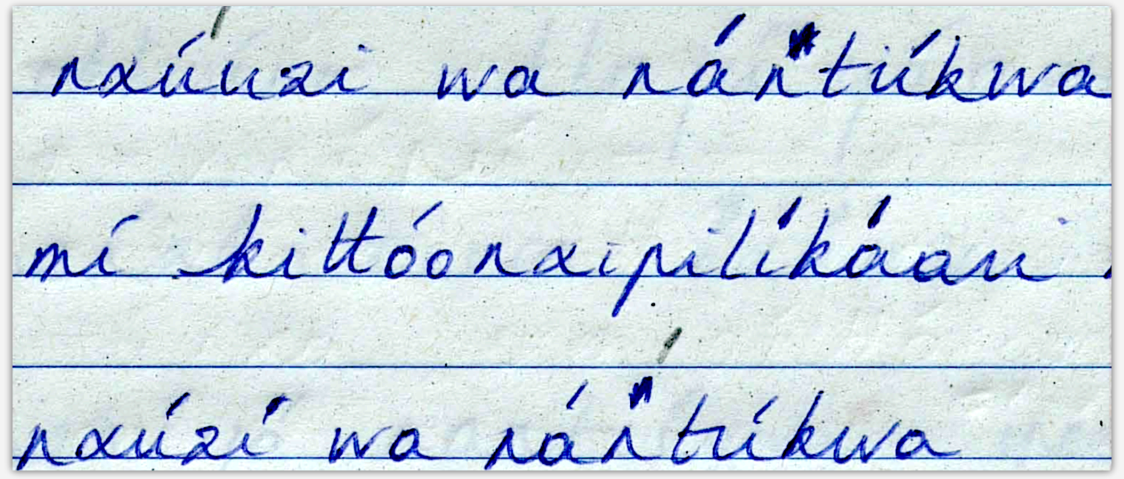
Text resources do exist in many more languages!

But locked away in formats that are not machine-readable

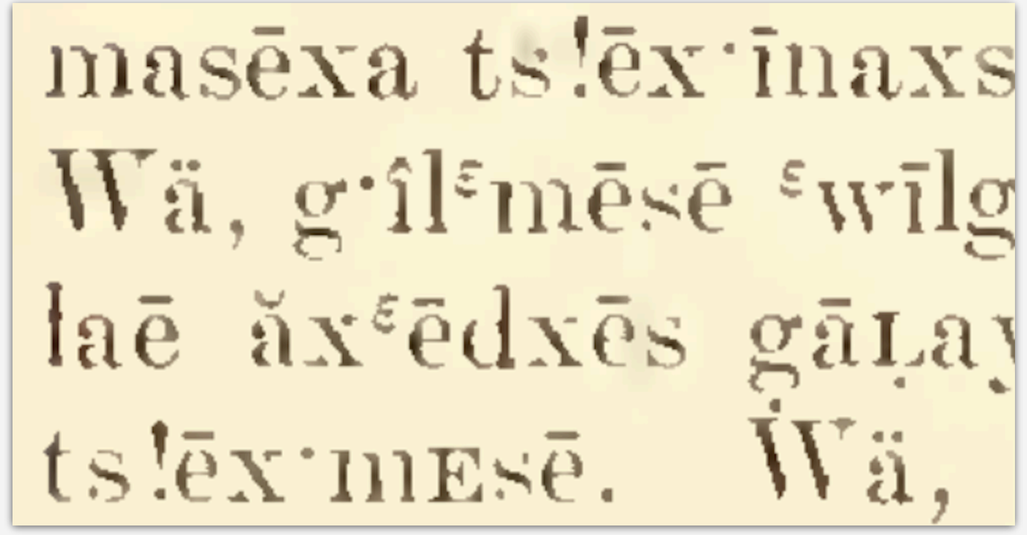




Printed books



Handwritten notes



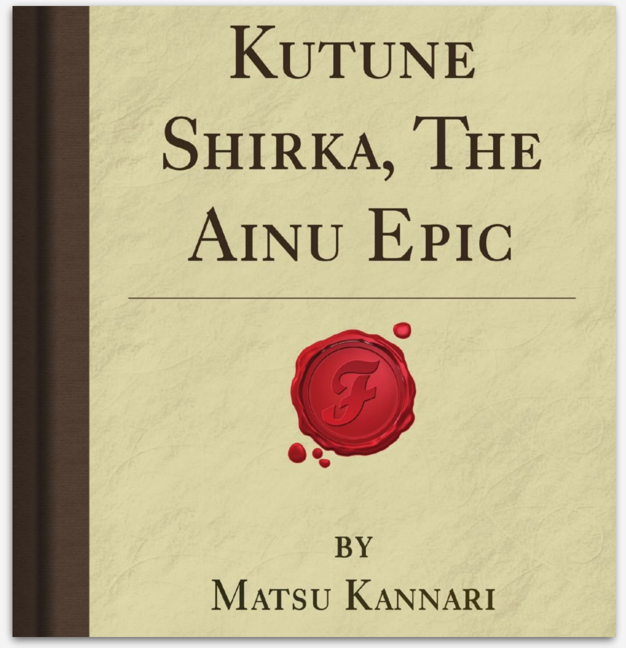
Typewritten documents

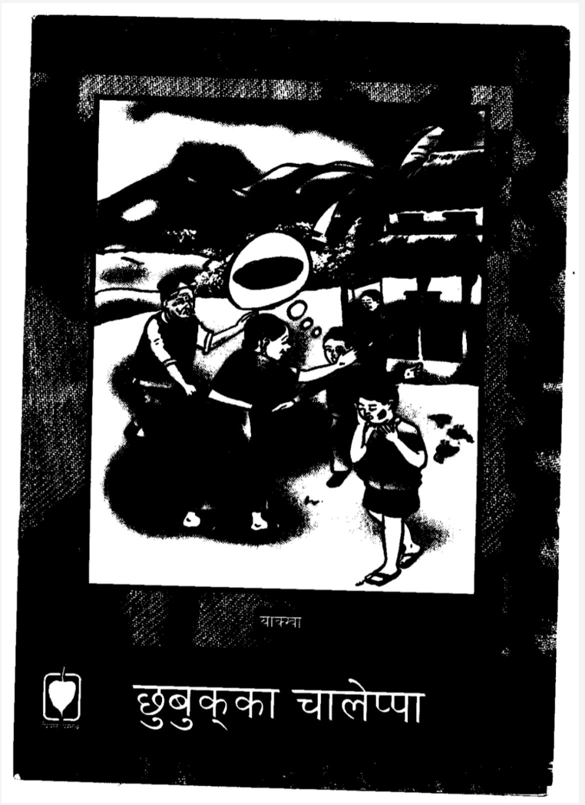
or in other formats such as bilingual lexicons

PanLex

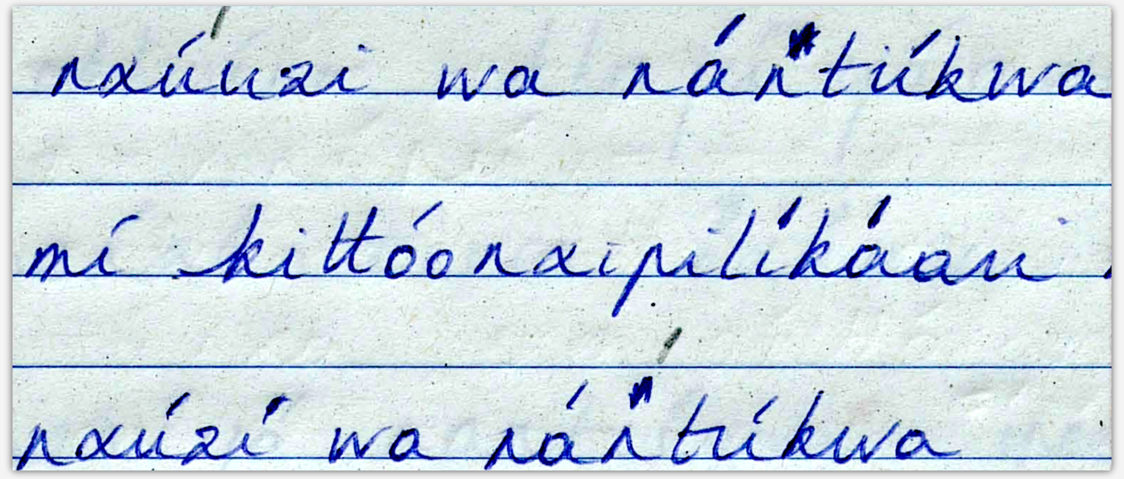
Text resources do exist in many more languages!

But locked away in formats that are not machine-readable

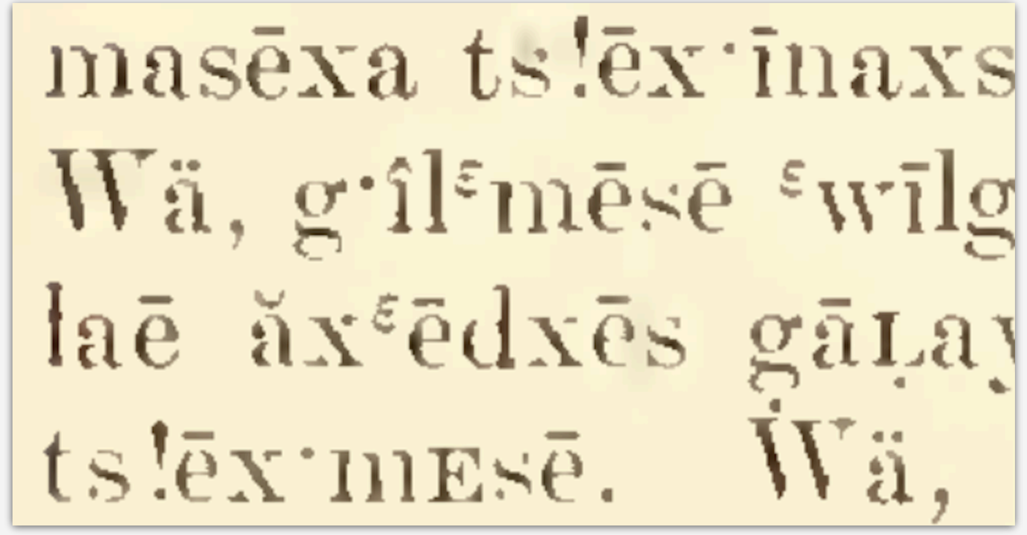




Printed books



Handwritten notes



Typewritten documents

or in other formats such as bilingual lexicons

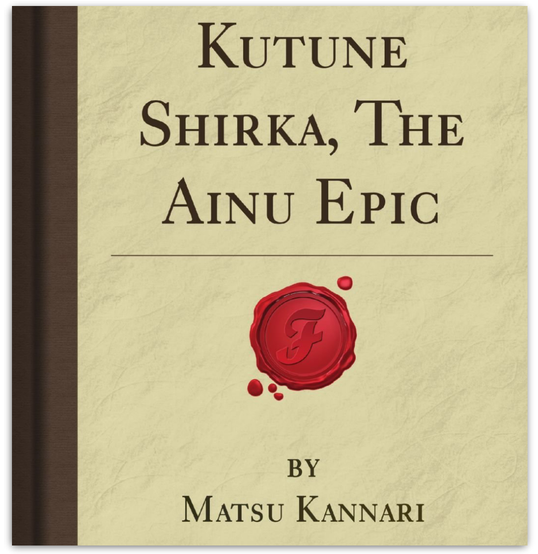
PanLex

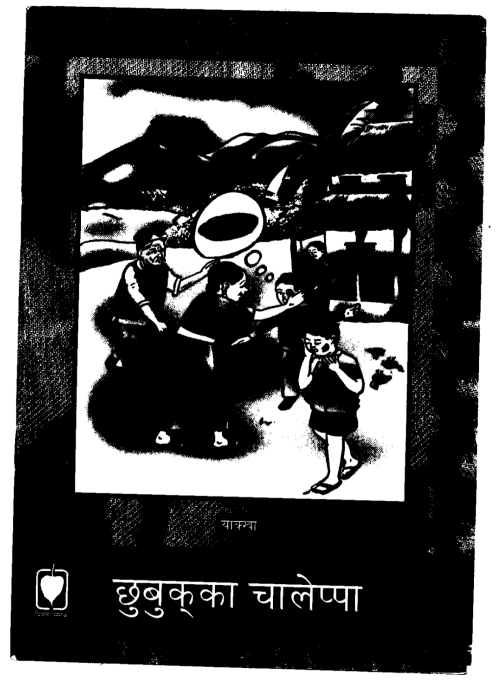
SIL

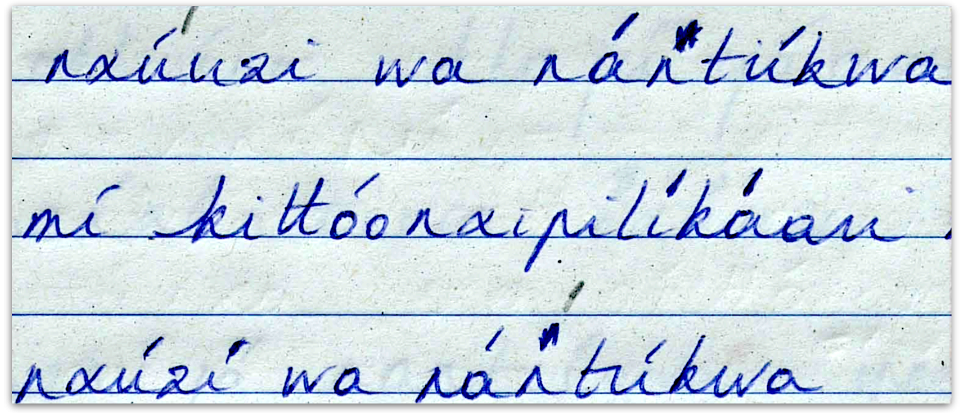
what can we do?!

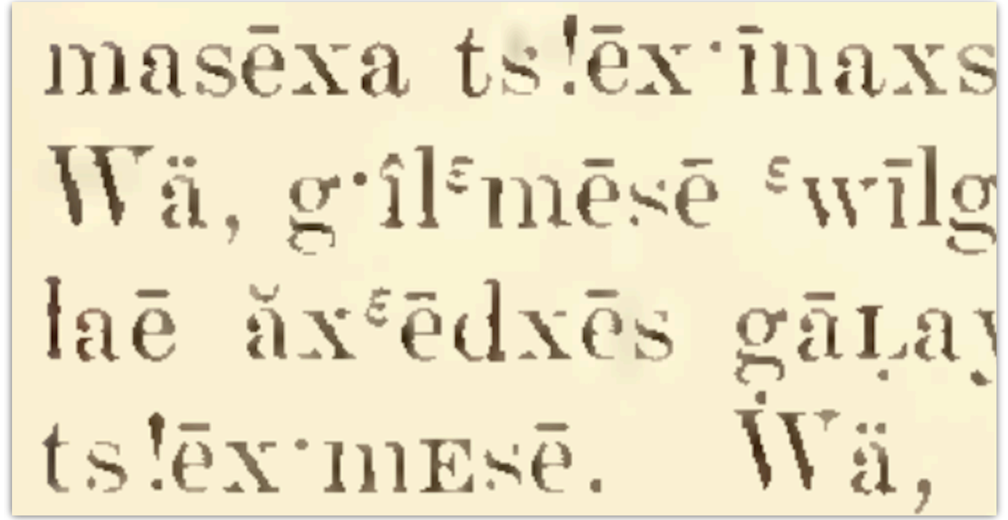
Text resources do exist in many more languages!

Text resources do exist in many more languages!



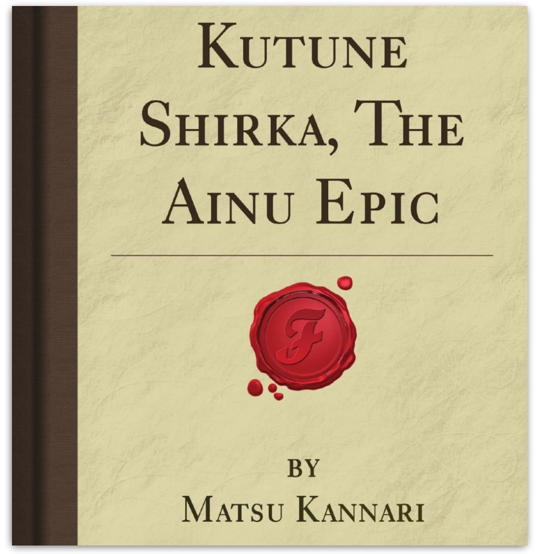


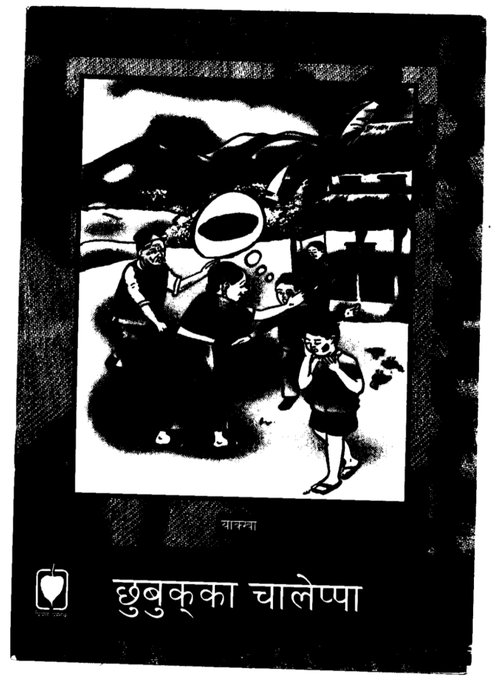


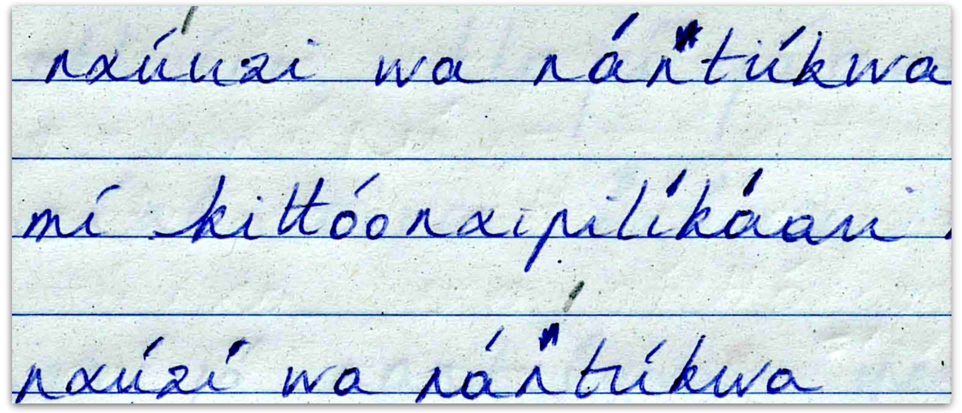


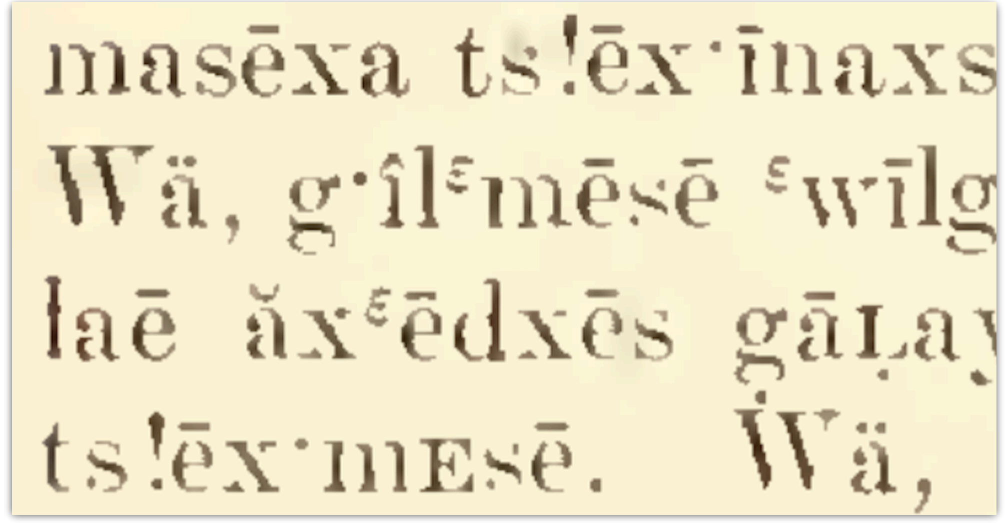
PanLex

Text resources do exist in many more languages!





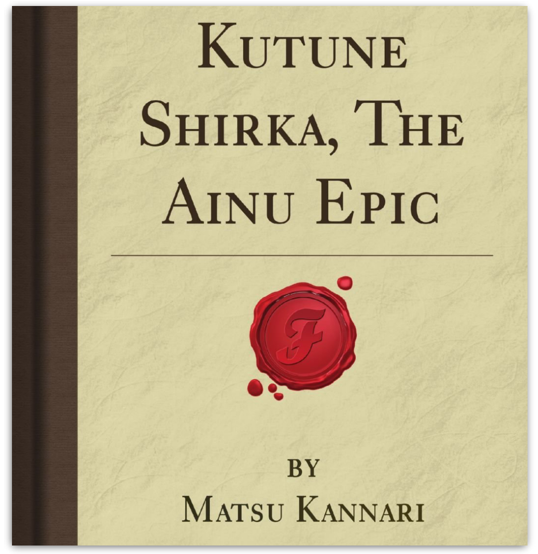


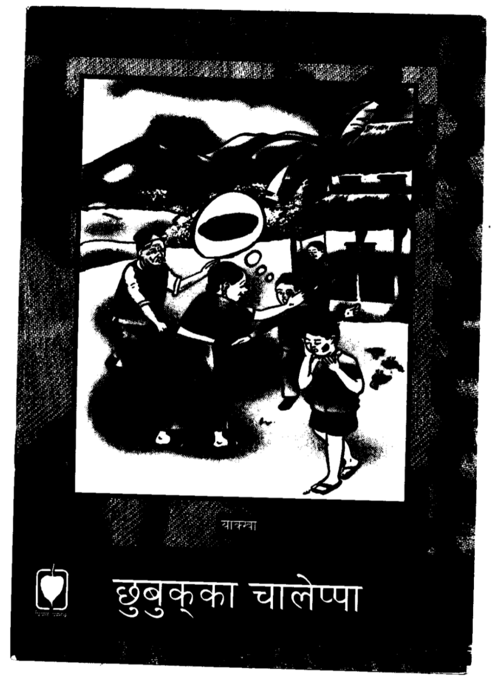


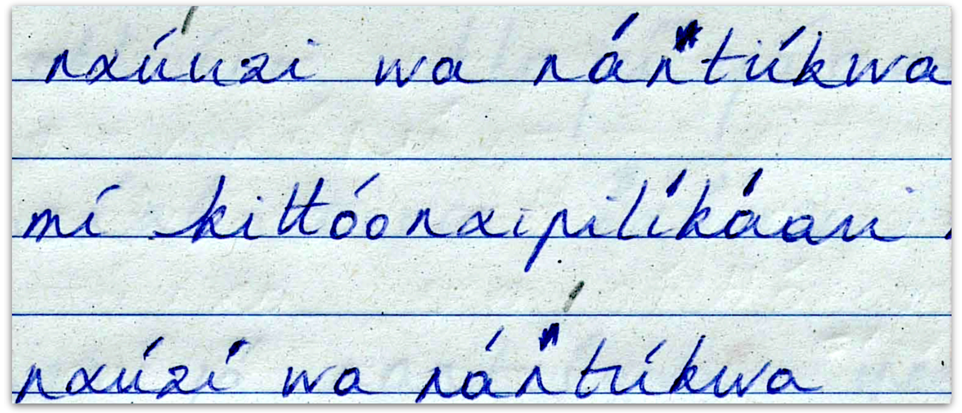
PanLex

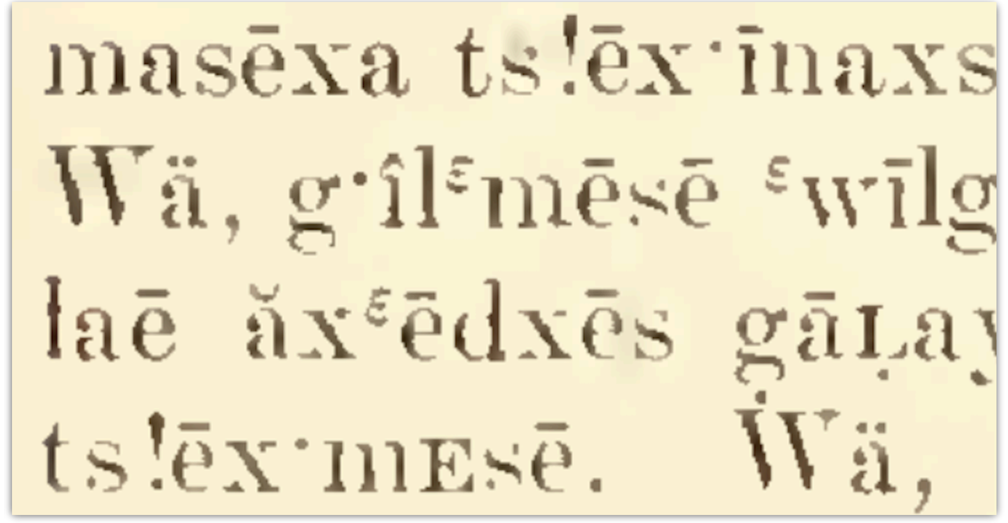
Unlocking non-traditional resources

Text resources do exist in many more languages!







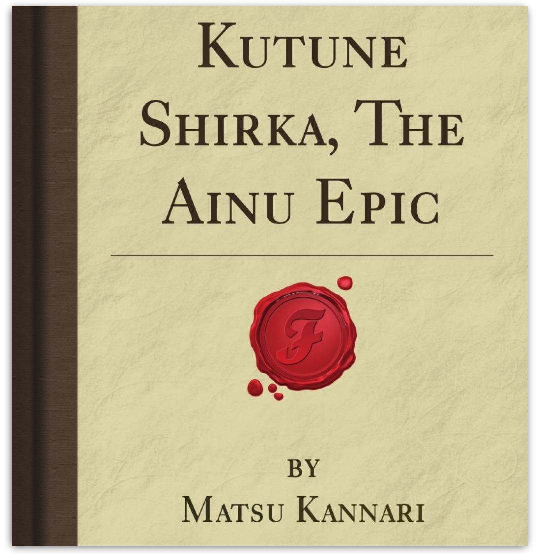


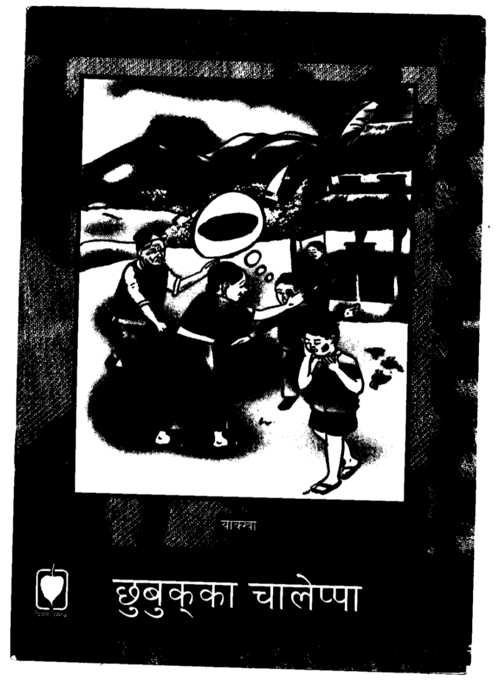
PanLex

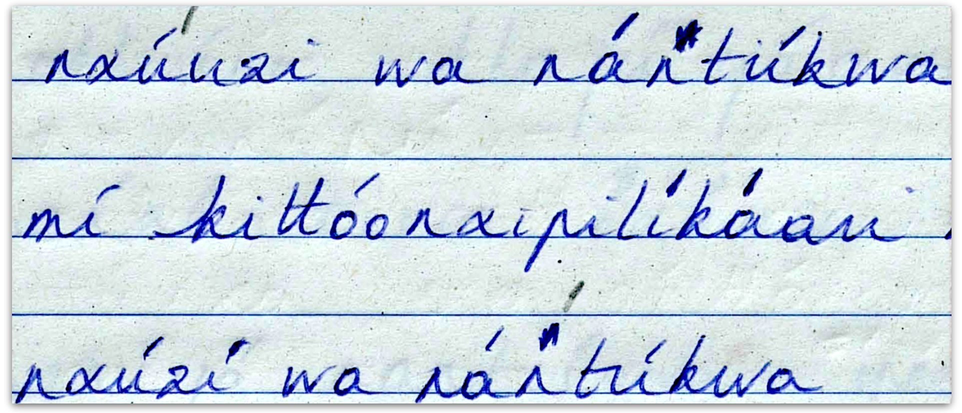
Unlocking non-traditional resources

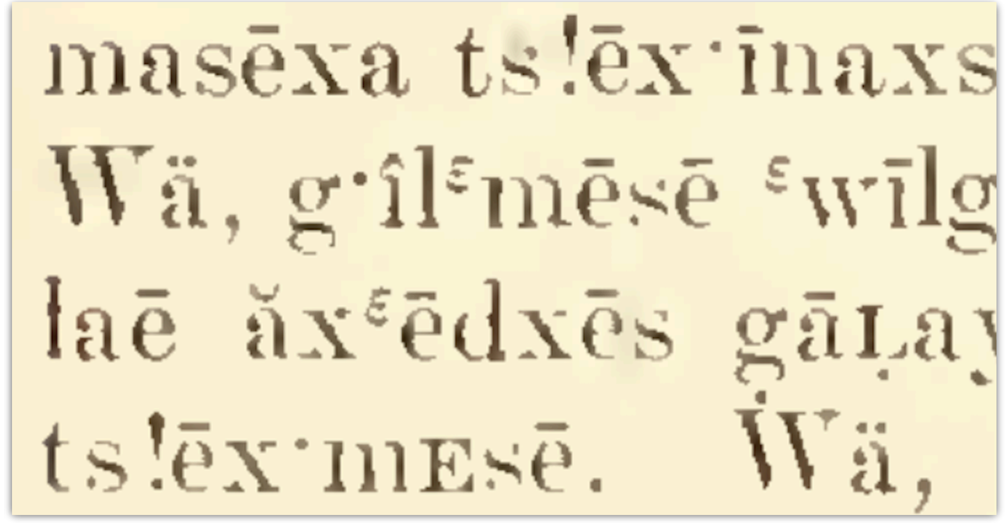
Enable NLP for under-resourced languages

Text resources do exist in many more languages!









PanLex

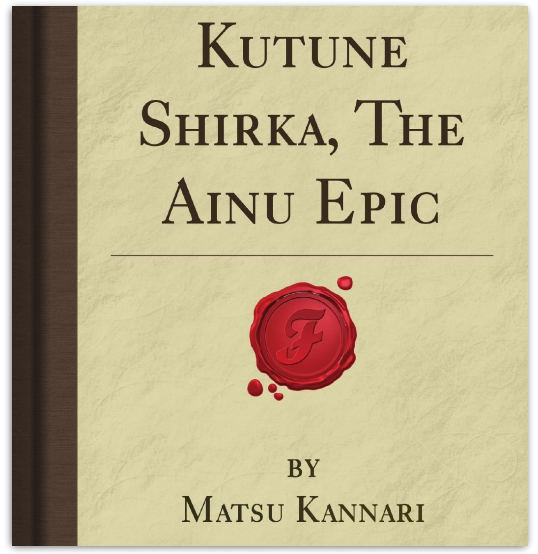
Unlocking non-traditional resources

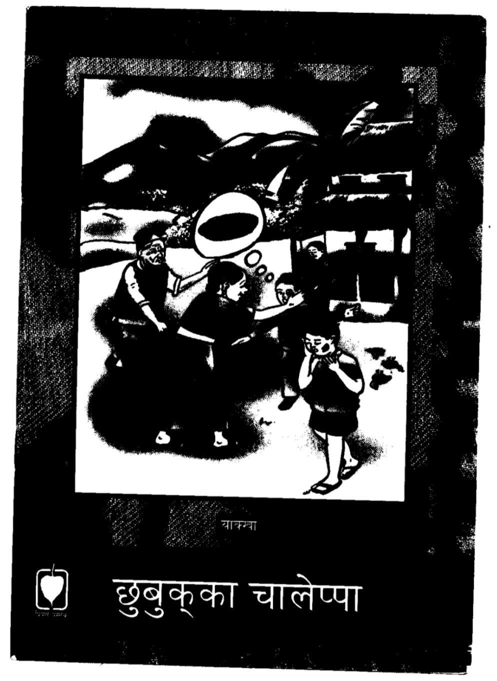
Enable NLP for under-resourced languages

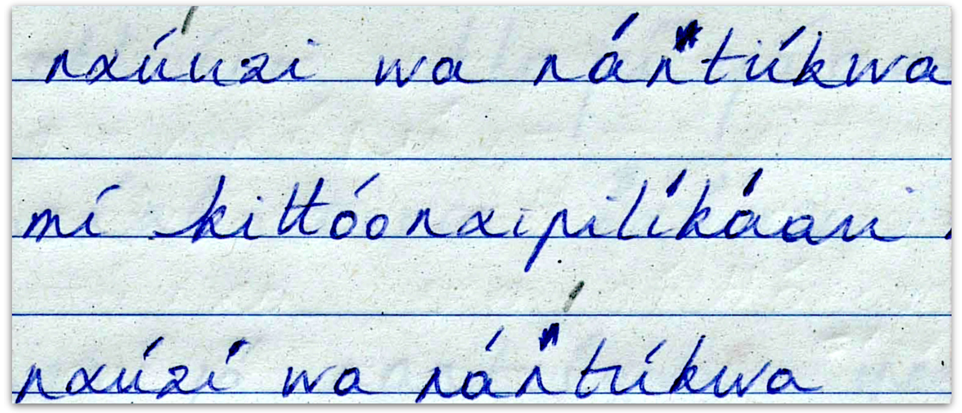
Expand multilingual LMs to more languages

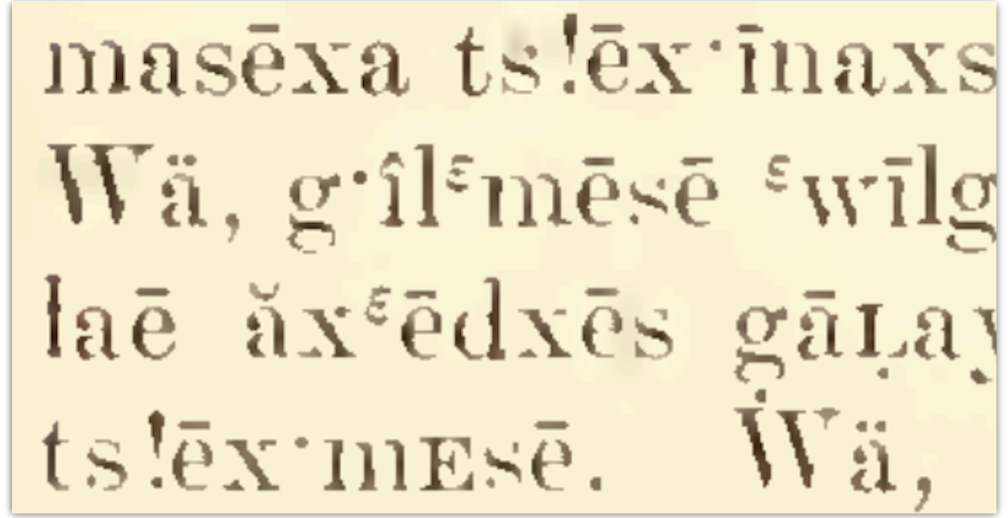
XLM-R mBERT
mT5 mBART ERNIE-M
Turing ULR

Text resources do exist in many more languages!








masēxa ts!ēx·înaxs
Wä, g·îlᵋmēsē ᵋwīlg
laē ăxᵋēdxēs gāɫay
ts!ēx·mᴇsē. Wä,

PanLex

Unlocking non-traditional resources

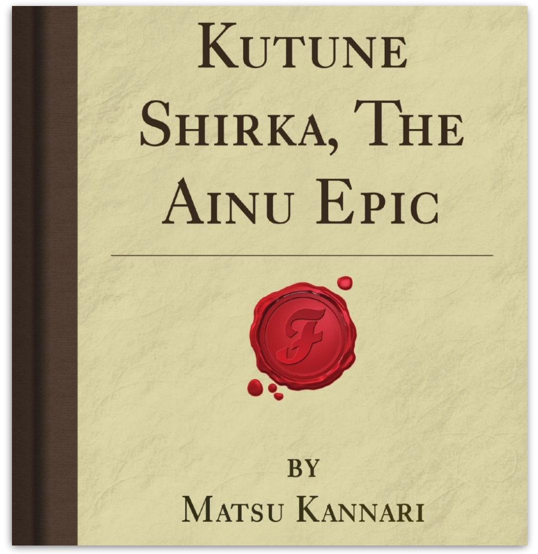
Enable NLP for under-resourced languages

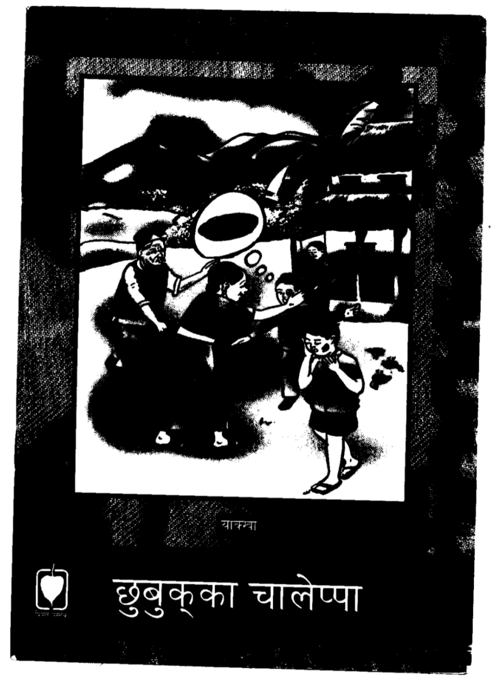
Expand multilingual LMs to more languages

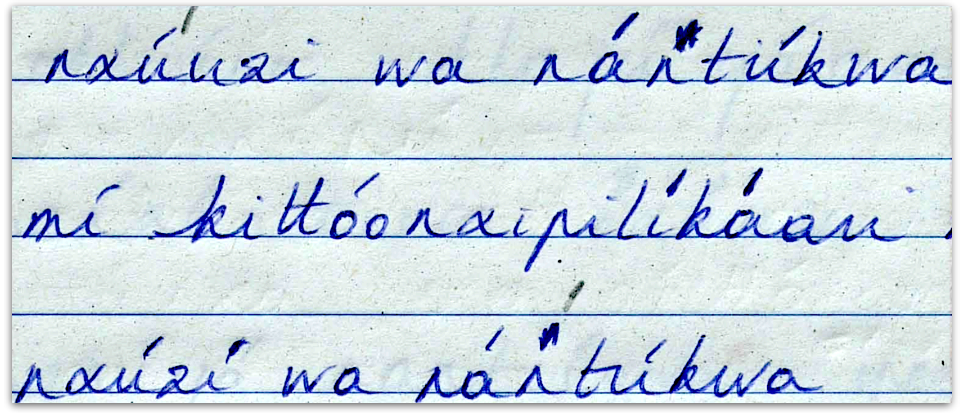
XLM-R mBERT
mT5 mBART ERNIE-M
Turing ULR

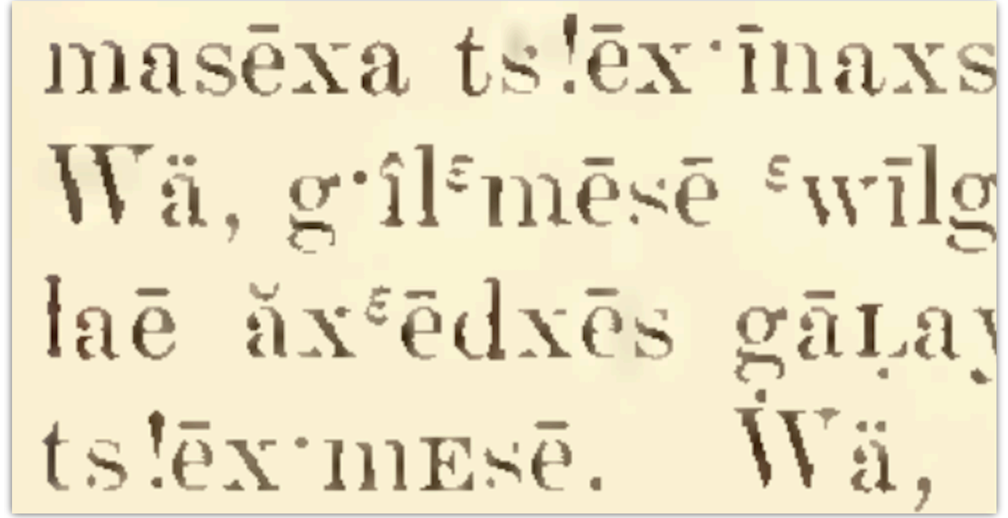
Annotate datasets for downstream NLP tasks

Text resources do exist in many more languages!









PanLex

Unlocking non-traditional resources

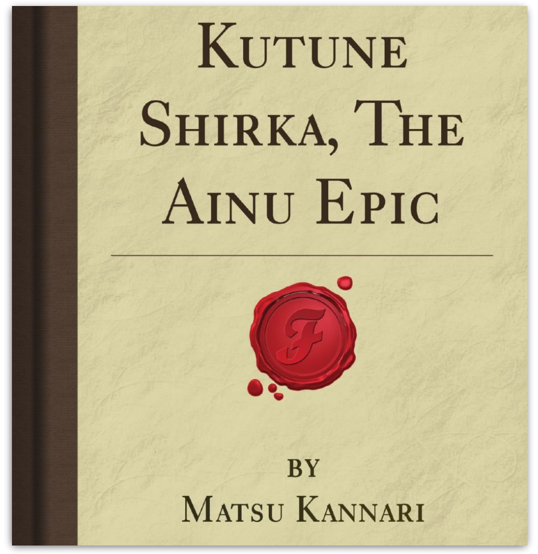
Enable NLP for under-resourced languages

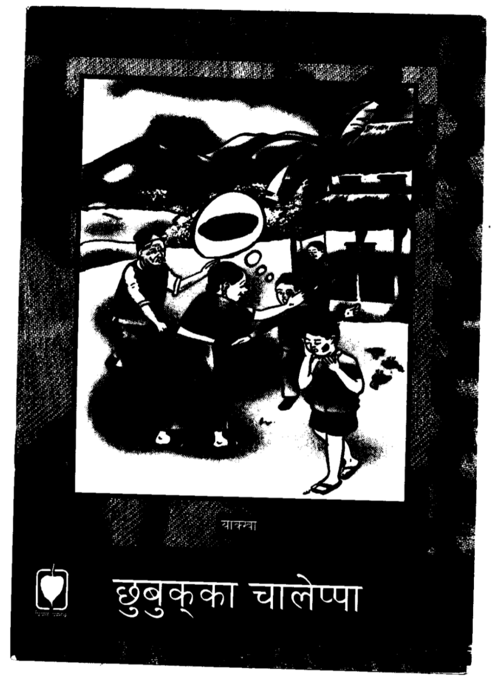
Expand multilingual LMs to more languages — XLM-R, mBERT, mT5, mBART, ERNIE-M, Turing ULR

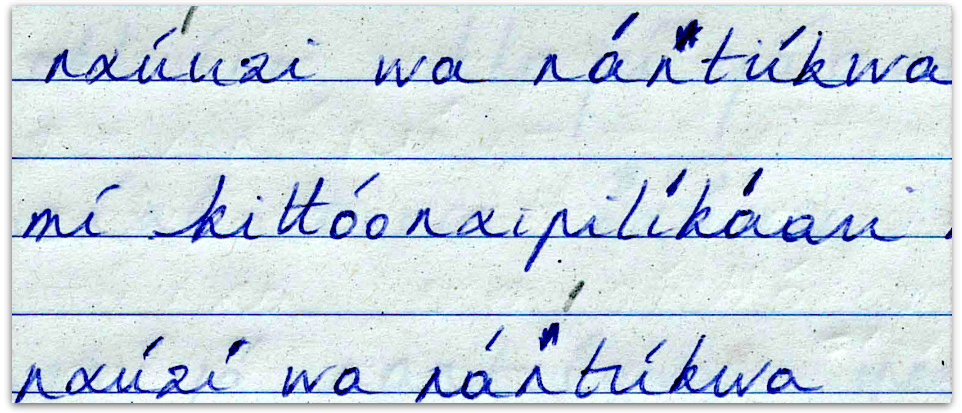
Annotate datasets for downstream NLP tasks

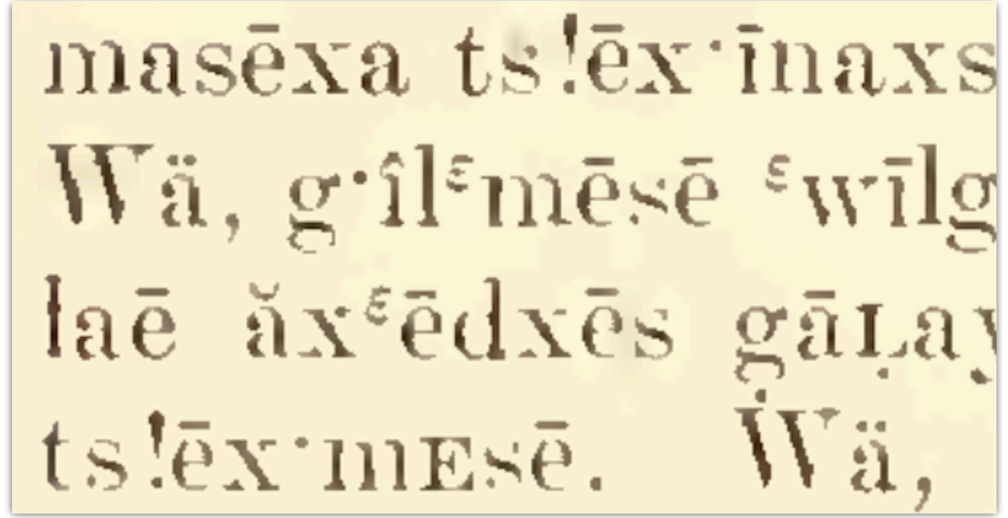
Support communities that speak these languages

Text resources do exist in many more languages!









PANLEX

Unlocking non-traditional resources

Enable NLP for under-resourced languages

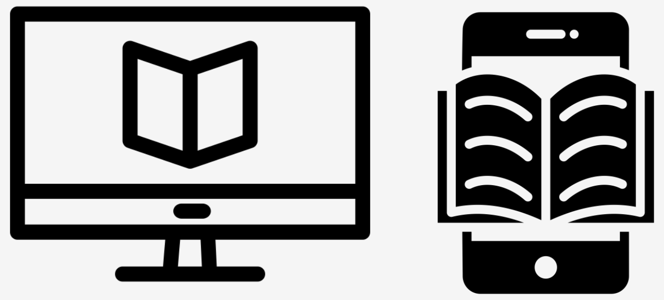
Expand multilingual LMs to more languages

XLM-R mBERT
mT5 mBART ERNIE-M
Turing ULR

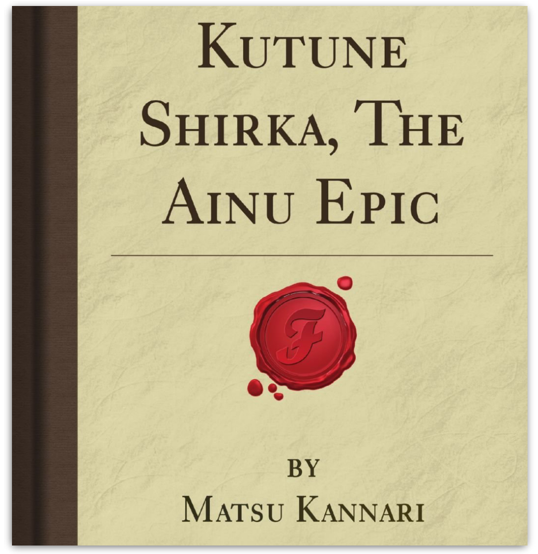
Annotate datasets for downstream NLP tasks

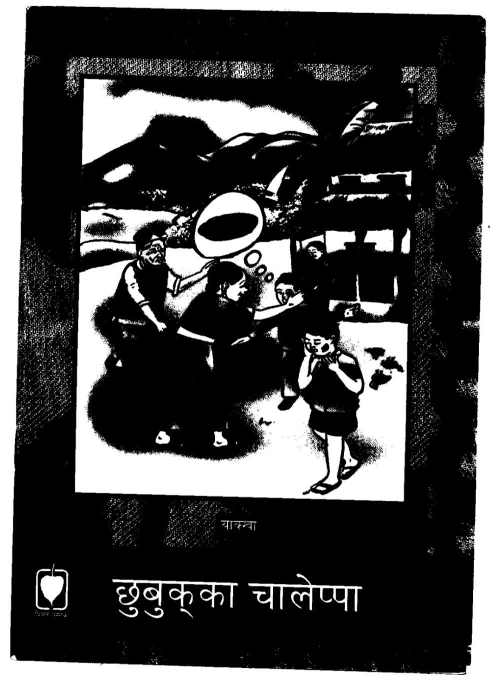
Support communities that speak these languages

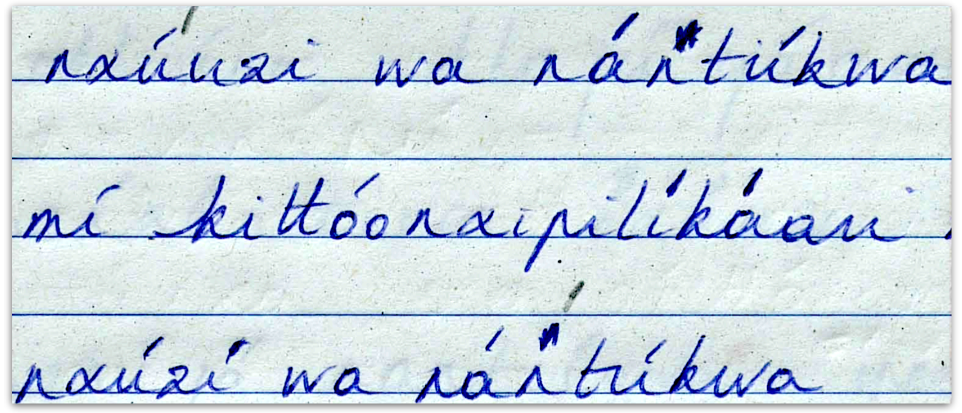
Make native texts digitally accessible and searchable

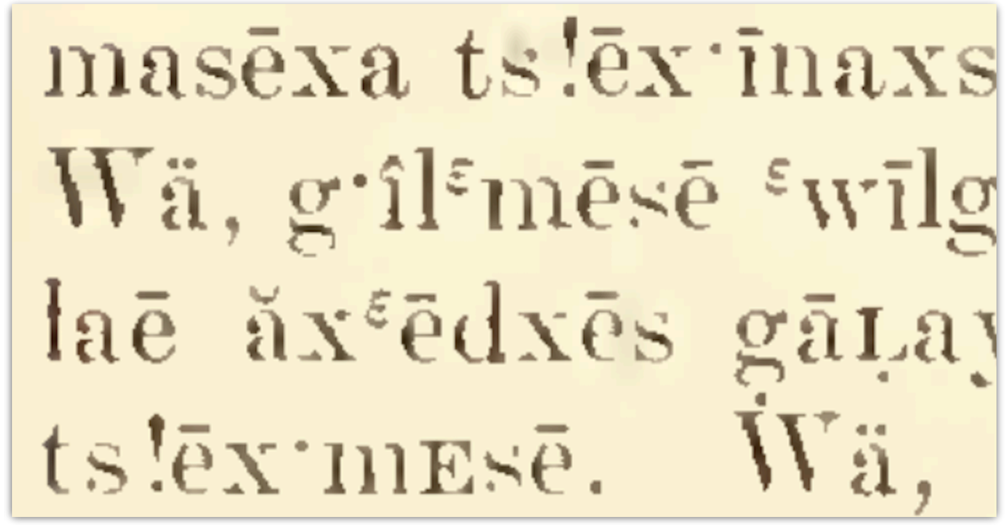


Text resources do exist in many more languages!








masēxa ts!ēx·inaxs
Wä, g·îlᵋmēsē ᵋwīlg
laē ăxᵋēdxēs gāɫay
ts!ēx·mɛsē. Wä,

PanLex

Unlocking non-traditional resources

Enable NLP for under-resourced languages

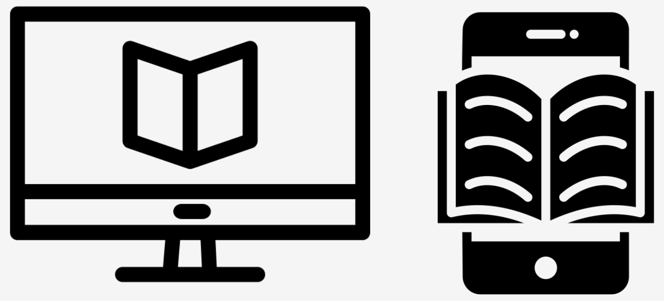
Expand multilingual LMs to more languages

XLM-R mBERT mT5 mBART ERNIE-M Turing ULR

Annotate datasets for downstream NLP tasks

Support communities that speak these languages

Make native texts digitally accessible and searchable



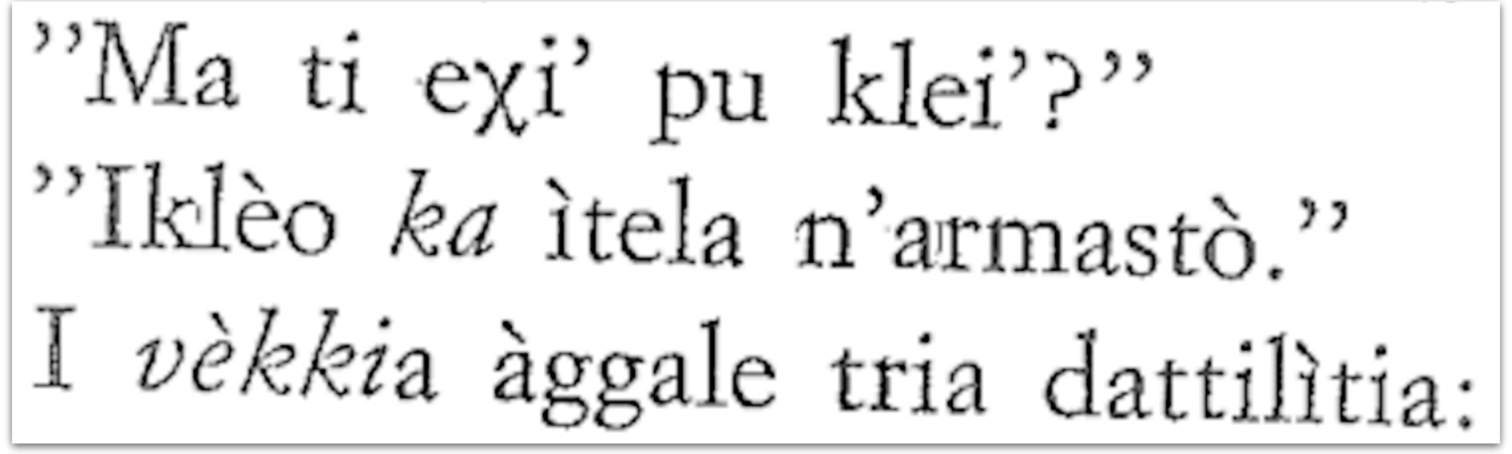
Aid language researchers, educatators, libraries…

Unlocking Un-digitized Text

Shruti Rijhwani, Antonios Anastasopoulos, Graham Neubig.
OCR Post-Correction for Endangered Language Texts.
EMNLP 2020.

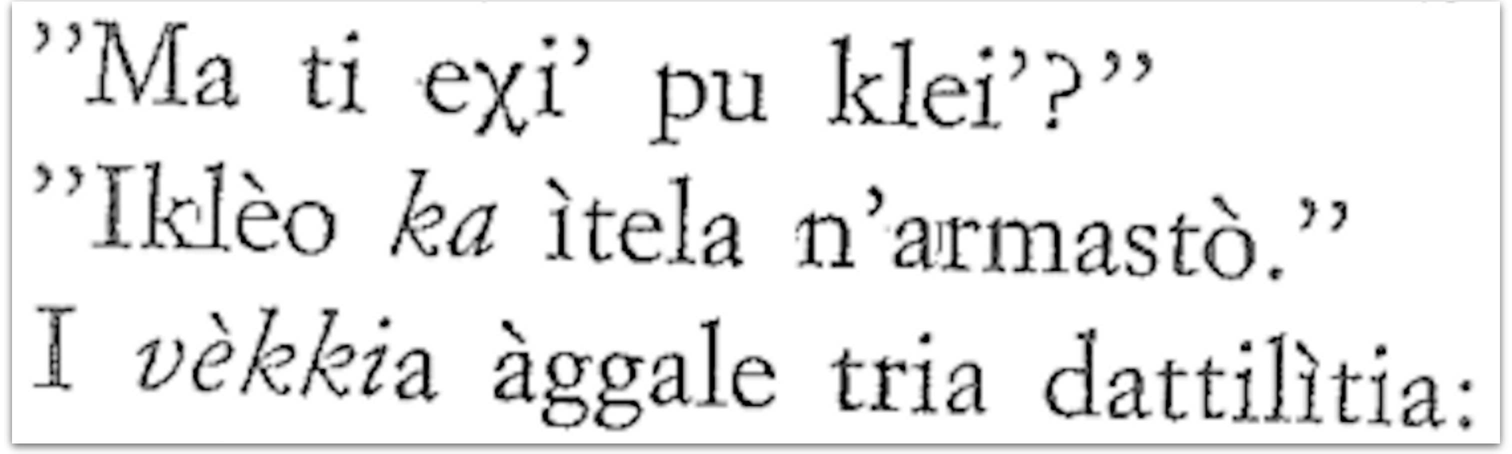
Shruti Rijhwani, Daisy Rosenblum, Antonios Anastasopoulos, Graham Neubig.
Lexically-Aware Semi-Supervised Learning for OCR Post-Correction.
TACL 2021.

Extracting text from scanned documents


"Ma ti eχi' pu klei'?"
"Iklèo *ka* ìtela n'armastò."
I *vèkki*a àggale tria dattilìtia:

Scanned document

Extracting text from scanned documents

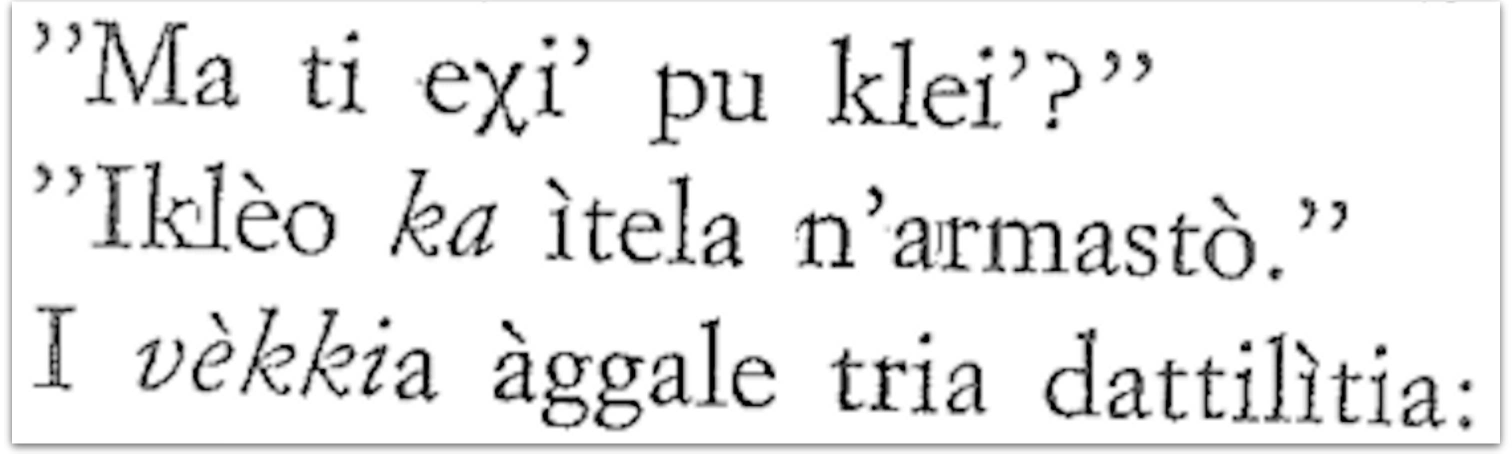


"Ma ti eχi' pu klei'?"
"Iklèo *ka* ìtela n'armastò."
I *vèkki*a àggale tria dattilìtia:

Scanned document

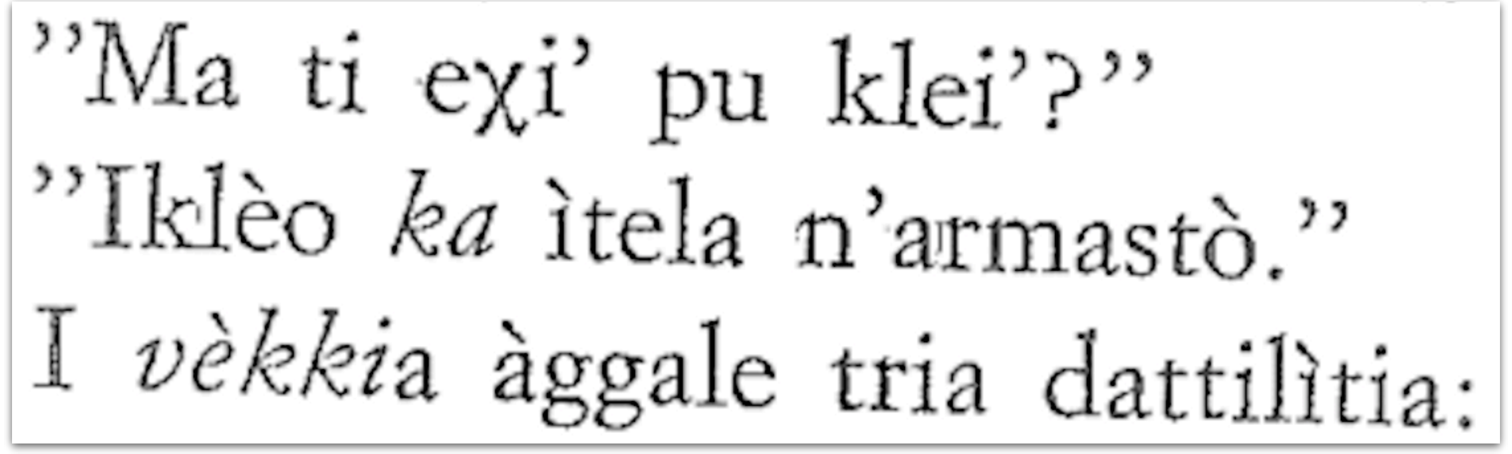
Scan from a book of folk tales in Griko

Extracting text from scanned documents


’’Ma ti eχi’ pu klei’?’’
’’Iklèo *ka* ìtela n’armastò.’’
I *vèkki*a àggale tria dattilìtia:

Scanned document

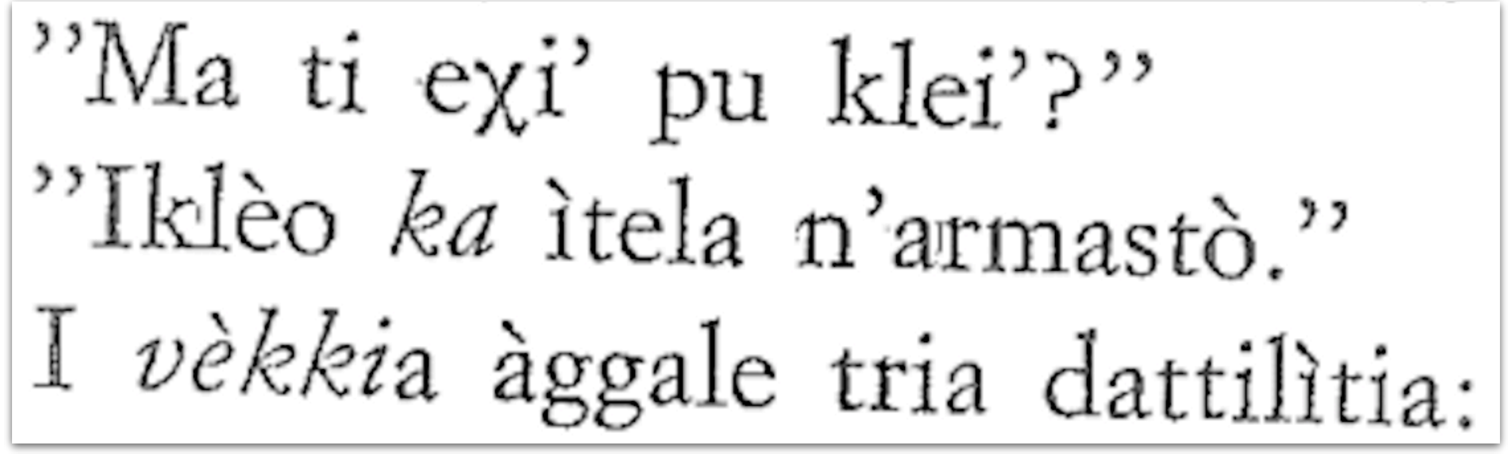
Extracting text from scanned documents


”Ma ti eχi’ pu klei’?”
”Iklèo *ka* ìtela n’armastò.”
I *vèkki*a àggale tria dattilìtia:

Scanned document

Optical Character Recognition (OCR)

Extracting text from scanned documents


''Ma ti eχi' pu klei'?''
''Iklèo *ka* ìtela n'armastò.''
I *vèkkia* àggale tria dattilìtia:

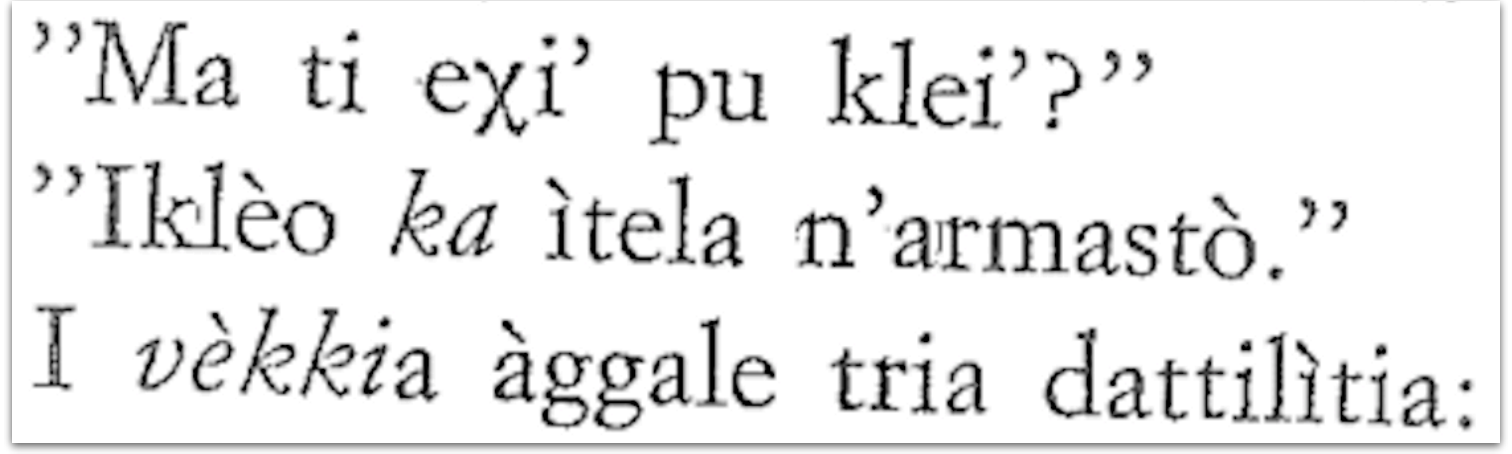
Scanned document

Optical Character Recognition (OCR)

''Ma ti eχi' pu klei'?''
''Iklèo ka ìtela n'armastò.''
I vèkkia àggale tria dattilìtia:

Machine readable text

Extracting text from scanned documents


”Ma ti eχi’ pu klei’?”
”Iklèo *ka* ìtela n’armastò.”
I *vèkkia* àggale tria dattilìtia:

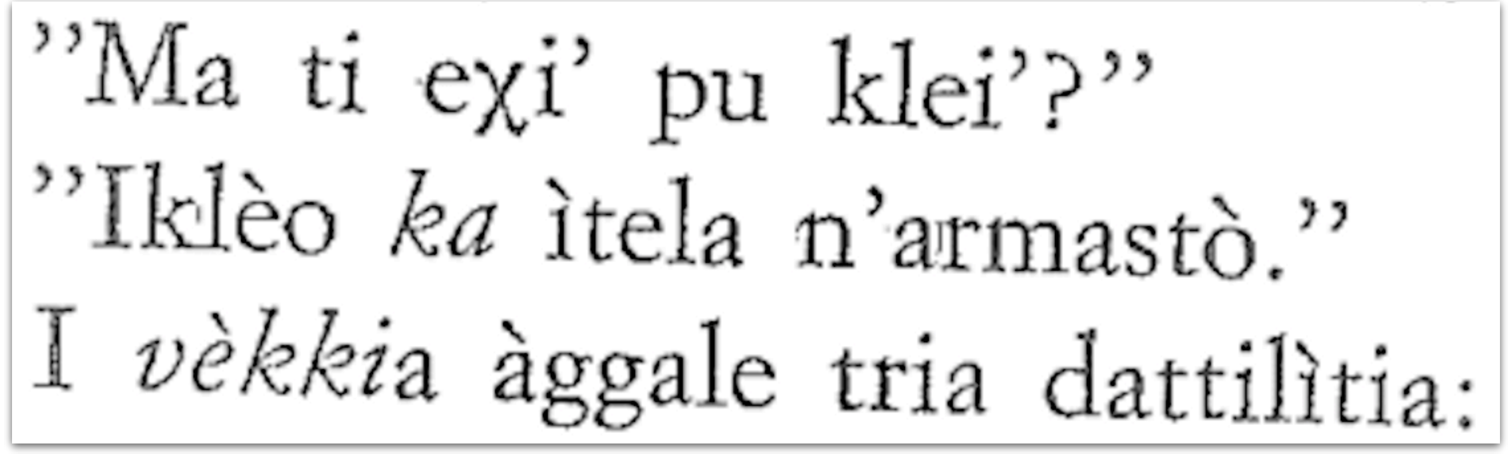
Scanned document

Optical Character Recognition (OCR)

”Ma ti eχi’ pu klei’?”
”Iklèo ka ìtela n’armastò.”
I vèkkia àggale tria dattilìtia:

Machine readable text

Extracting text from scanned documents


"Ma ti eχi' pu klei'?"
"Iklèo *ka* ìtela n'armastò."
I *vèkkia* àggale tria dattilìtia:

Scanned document

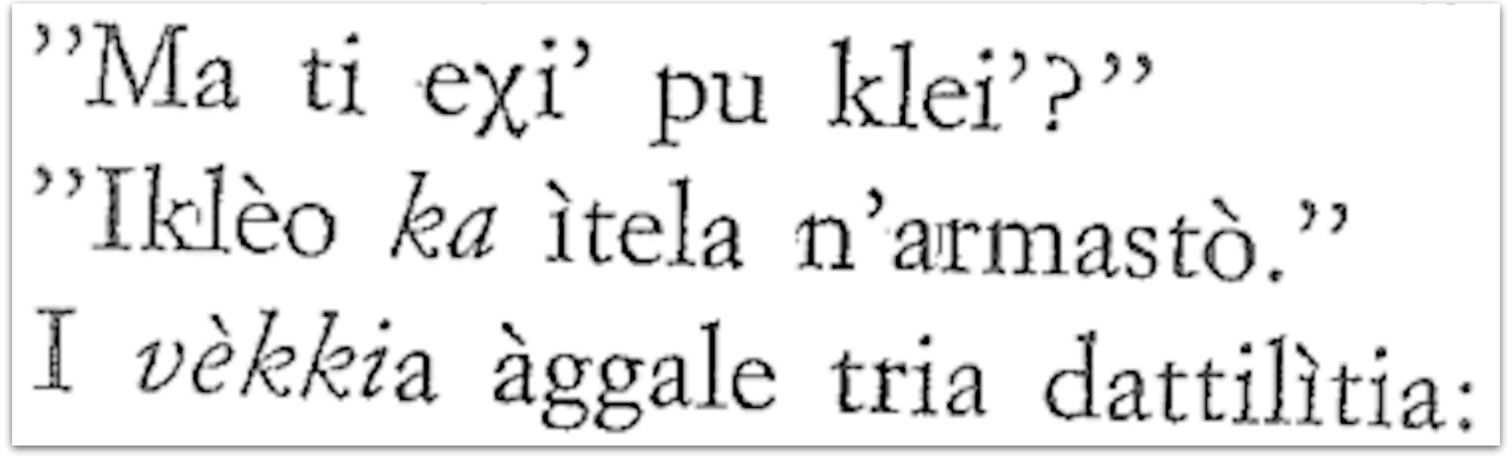
Optical Character Recognition (OCR)

"Ma ti eχi' pu klei'?"
"Iklèo ka ìtela n'armastò."
I vèkkia àggale tria dattilìtia:

Machine readable text

- High accuracy on languages that have easily available resources!

Extracting text from scanned documents


”Ma ti eχi’ pu klei’?”
”Iklèo *ka* ìtela n’armastò.”
I *vèkkia* àggale tria dattilìtia:

Scanned document

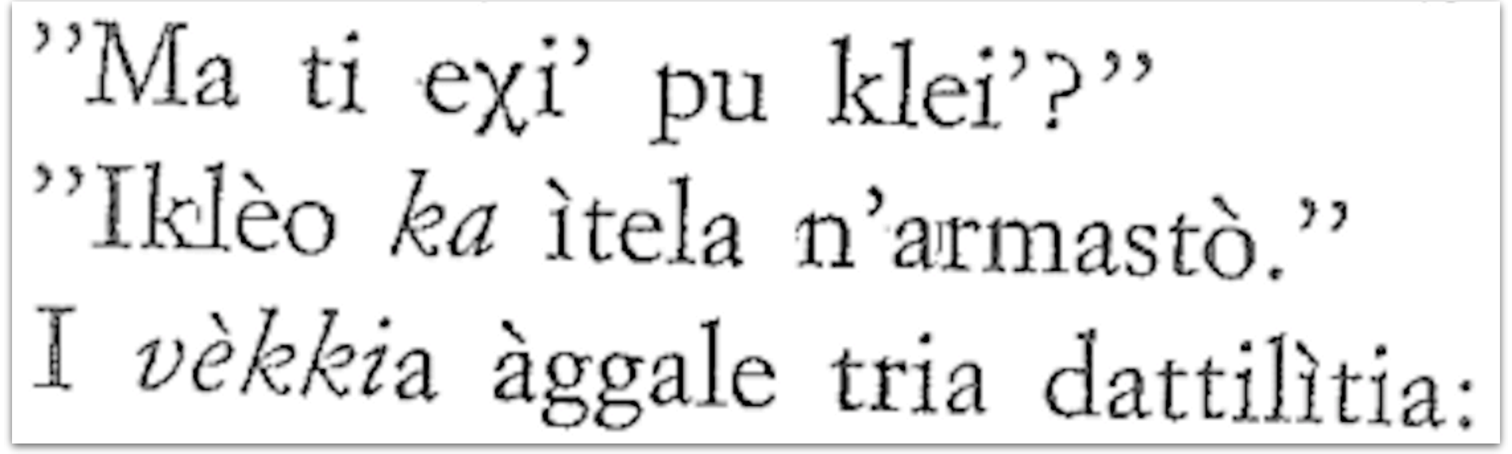
Optical Character Recognition (OCR)

”Ma ti eχi’ pu klei’?”
”Iklèo ka ìtela n’armastò.”
I vèkkia àggale tria dattilìtia:

Machine readable text

- High accuracy on languages that have easily available resources!
- Off-the-shelf tools support many scripts and languages

Extracting text from scanned documents


''Ma ti eχi' pu klei'?''
''Iklèo *ka* ìtela n'armastò.''
I *vèkkia* àggale tria dattilìtia:

Scanned document

Optical Character Recognition (OCR)

''Ma ti eχi' pu klei'?''
''Iklèo ka ìtela n'armastò.''
I vèkkia àggale tria dattilìtia:

Machine readable text

- High accuracy on languages that have easily available resources!
- Off-the-shelf tools support many scripts and languages

Google Vision
Tesseract
EasyOCR
…

Extracting text from scanned documents

"Ma ti eχi' pu klei'?"
"Iklèo *ka* ìtela n'armastò."
I *vèkkia* àggale tria dattilìtia:

Scanned document

Optical Character Recognition (OCR)

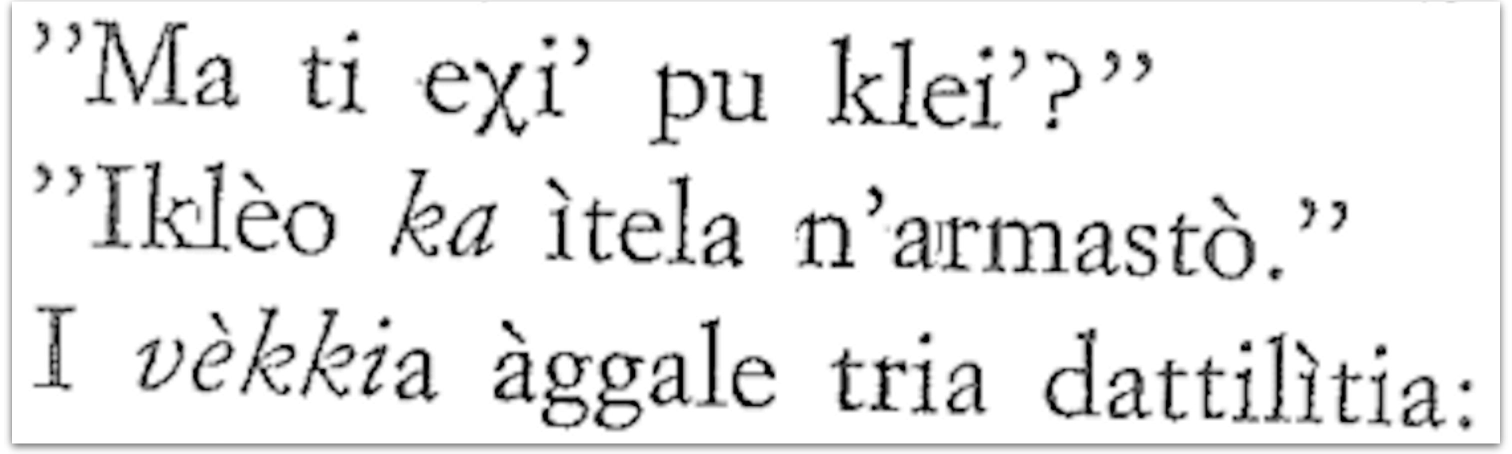
"Ma ti eχi' pu klei'?"
"Iklèo ka ìtela n'armastò."
I vèkkia àggale tria dattilìtia:

Machine readable text

- High accuracy on languages that have easily available resources!
- Off-the-shelf tools support many scripts and languages

Support 80-100 languages
- Google Vision
- Tesseract
- EasyOCR
- …

Extracting text from scanned documents


"Ma ti eχi' pu klei'?"
"Iklèo *ka* ìtela n'armastò."
I *vèkkia* àggale tria dattilìtia:

Scanned document

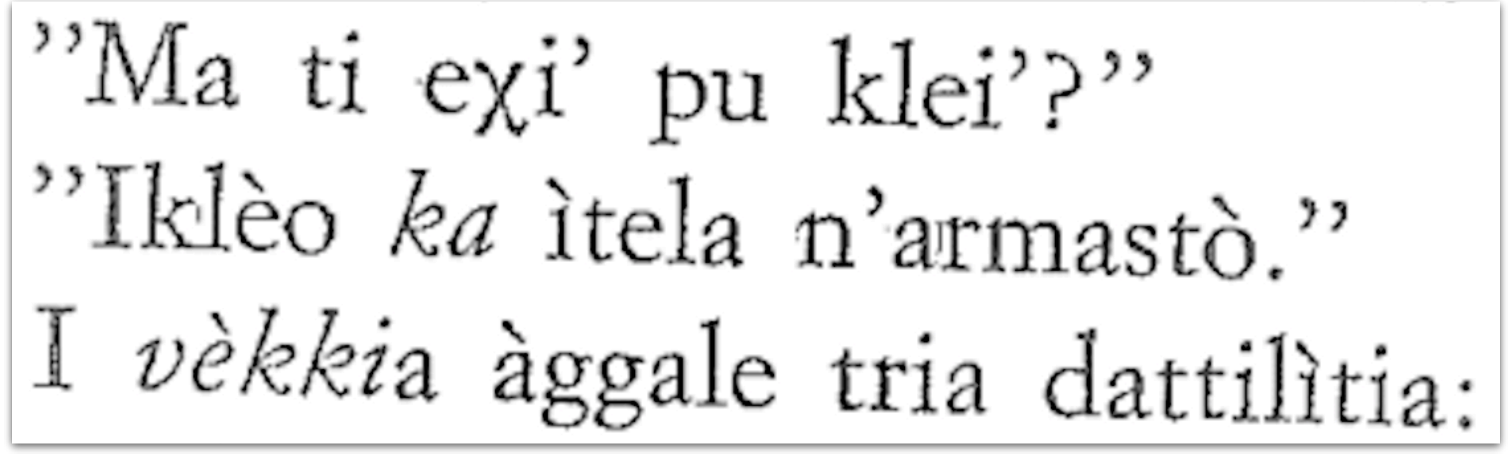
Optical Character Recognition (OCR)

"Ma ti eχi' pu klei'?"
"Iklèo ka ìtela n'armastò."
I vèkkia àggale tria dattilìtia:

Machine readable text

- High accuracy on languages that have easily available resources!
- Off-the-shelf tools support many scripts and languages
- Little to no prior work on very low-resourced settings

Extracting text from scanned documents


''Ma ti eχi' pu klei'?''
''Iklèo *ka* ìtela n'armastò.''
I *vèkkia* àggale tria dattilìtia:

Scanned document

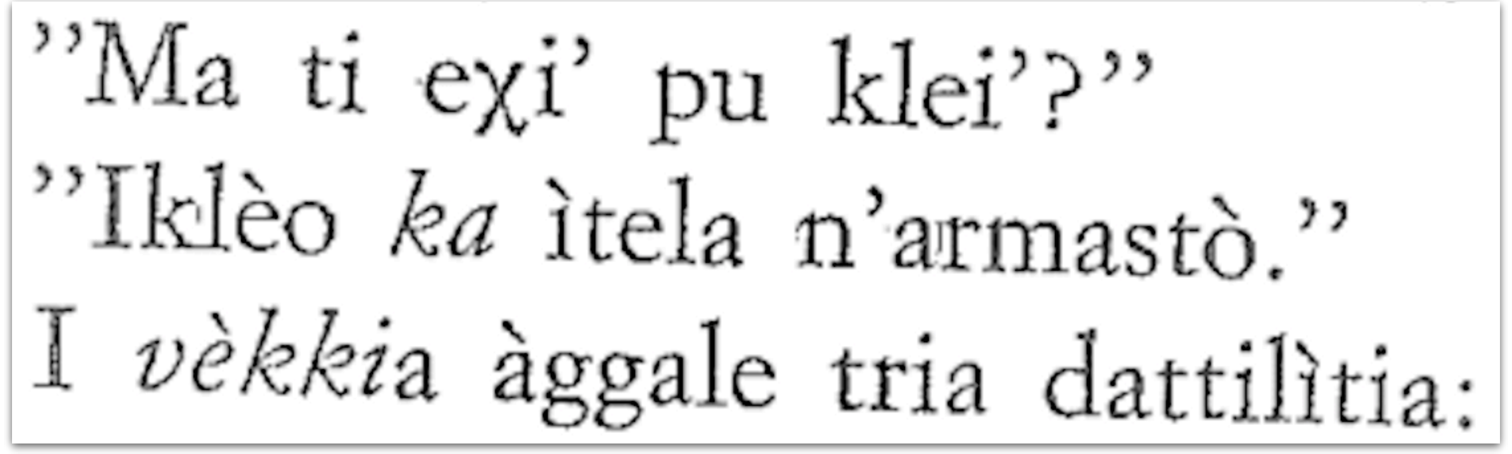
Optical Character Recognition (OCR)

''Ma ti eχi' pu klei'?''
''Iklèo ka ìtela n'armastò.''
I vèkkia àggale tria dattilìtia:

Machine readable text

- Little to no prior work on very low-resourced settings

Extracting text from scanned documents


"Ma ti eχi' pu klei'?"
"Iklèo *ka* ìtela n'armastò."
I *vèkkia* àggale tria dattilìtia:

Scanned document

Optical Character Recognition (OCR)

"Ma ti eχi' pu klei'?"
"Iklèo ka ìtela n'armastò."
I vèkkia àggale tria dattilìtia:

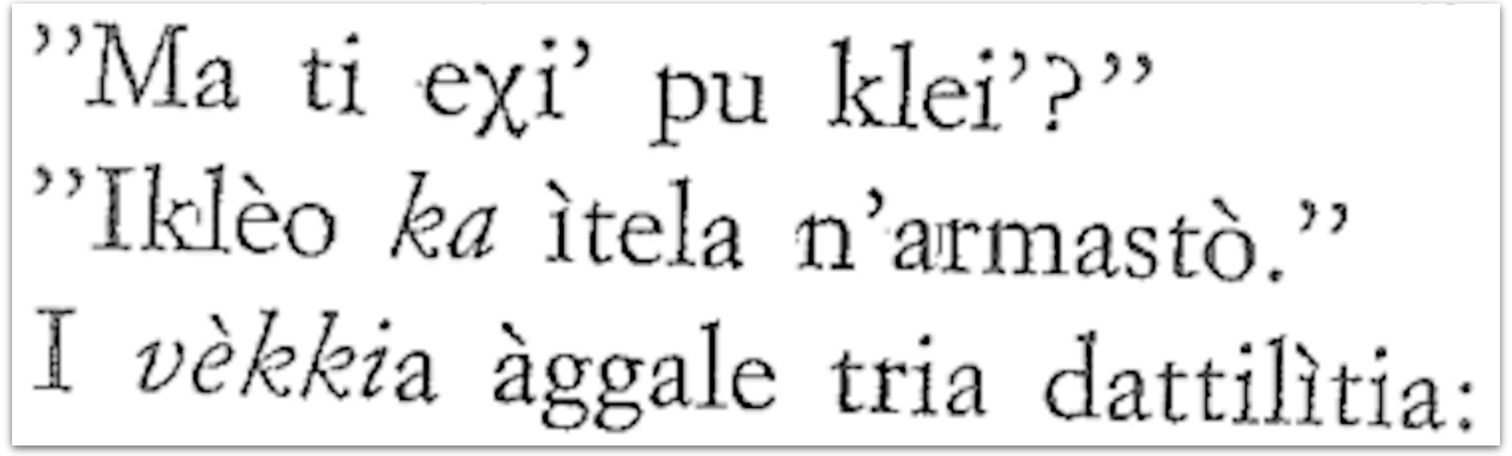
Machine readable text

- Little to no prior work on very low-resourced settings

Evaluation dataset
Promises and pitfalls of existing methods
Neural models for improving OCR performance in low-resource settings

Rijhwani, Anastasopoulos, Neubig. EMNLP 2020.

Extracting text from scanned documents


"Ma ti eχi' pu klei'?"
"Iklèo *ka* ìtela n'armastò."
I *vèkkia* àggale tria dattilìtia:

Scanned document

Optical Character Recognition (OCR)

"Ma ti eχi' pu klei'?"
"Iklèo ka ìtela n'armastò."
I vèkkia àggale tria dattilìtia:

Machine readable text

- Little to no prior work on very low-resourced settings

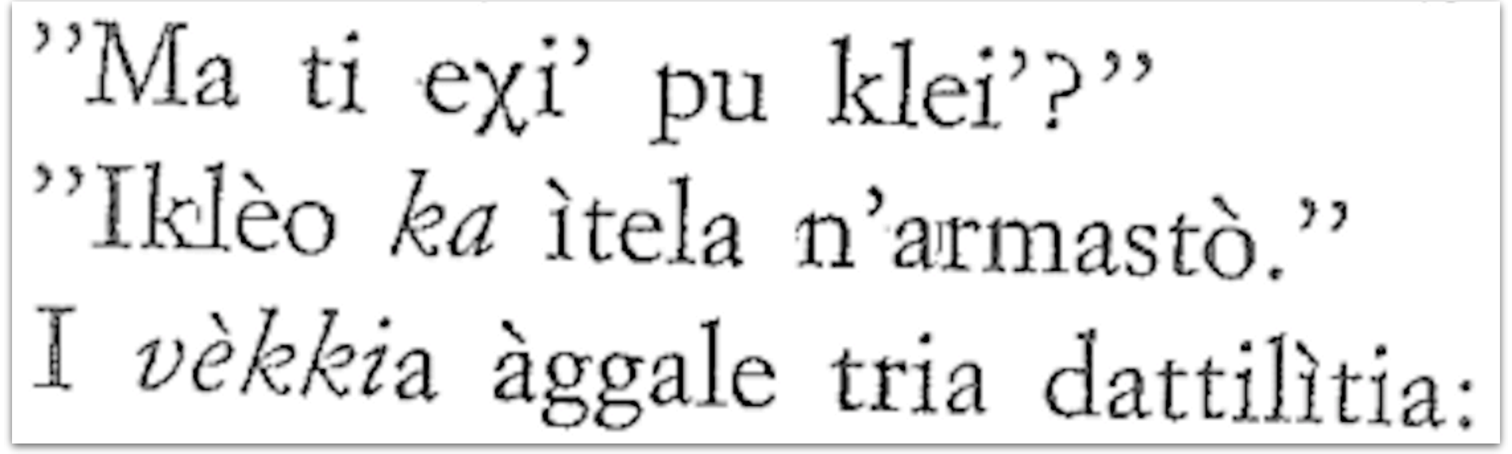
Evaluation dataset

Promises and pitfalls of existing methods

Neural models for improving OCR performance in low-resource settings

Rijhwani, Anastasopoulos, Neubig. EMNLP 2020.

Extracting text from scanned documents


"Ma ti eχi' pu klei'?"
"Iklèo *ka* ìtela n'armastò."
I *vèkkia* àggale tria dattilìtia:

Scanned document

Optical Character Recognition (OCR)

"Ma ti eχi' pu klei'?"
"Iklèo ka ìtela n'armastò."
I vèkkia àggale tria dattilìtia:

Machine readable text

- Little to no prior work on very low-resourced settings

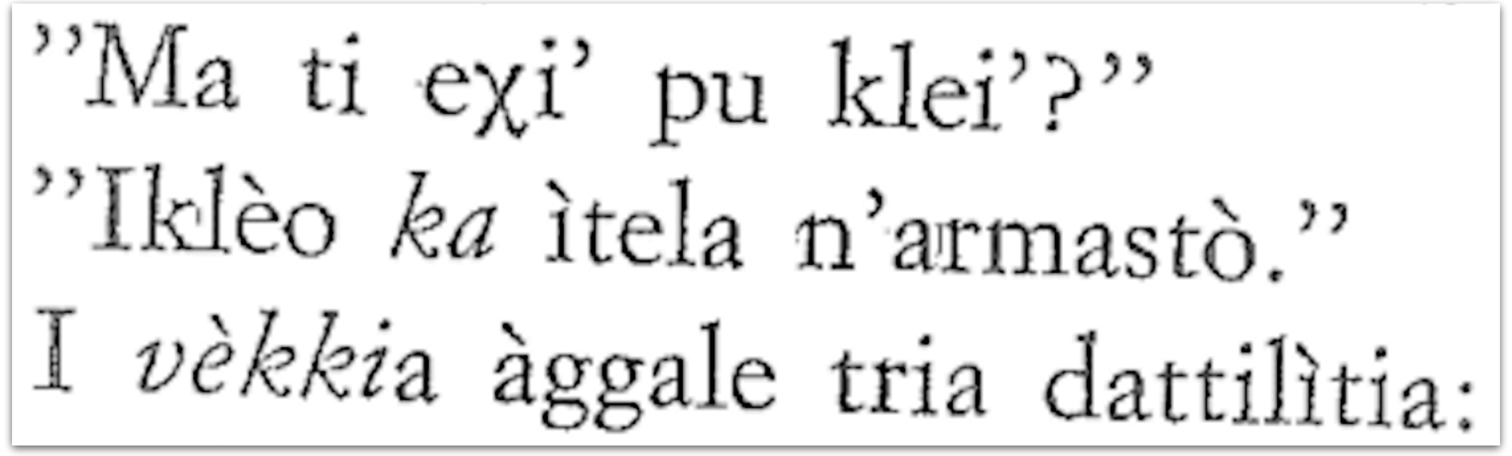
Evaluation dataset

Promises and pitfalls of existing methods

Neural models for improving OCR performance in low-resource settings

Rijhwani, Anastasopoulos, Neubig. EMNLP 2020.

Extracting text from scanned documents


"Ma ti eχi' pu klei'?"
"Iklèo *ka* ìtela n'armastò."
I *vèkkia* àggale tria dattilìtia:

Scanned document

Optical Character Recognition (OCR)

"Ma ti eχi' pu klei'?"
"Iklèo ka ìtela n'armastò."
I vèkkia àggale tria dattilìtia:

Machine readable text

- Little to no prior work on very low-resourced settings

Evaluation dataset
Promises and pitfalls of existing methods
Neural models for improving OCR performance in low-resource settings

Rijhwani, Anastasopoulos, Neubig. EMNLP 2020.

Semi-supervised learning to improve performance with unlabeled images

Rijhwani, Rosenblum, Anastasopoulos, Neubig. TACL 2021.

Evaluation dataset for low-resource OCR

Evaluation dataset for low-resource OCR

Ainu
(Japan)

kira-an patek
aeyairamshitne[1]
5760 hushkotoi wano[2]
iki-an aine

Evaluation dataset for low-resource OCR

Ainu
(Japan)

kira-an patek
aeyairamshitne[1]
5760 hushkotoi wano[2]
iki-an aine

Griko
(Italy)

"Ma ti eχi' pu klei'?"
"Iklèo *ka* ìtela n'armastò."
I *vèkki*a àggale tria dattilìtia:

Evaluation dataset for low-resource OCR

Ainu
(Japan)

kira-an patek
aeyairamshitne[1]
5760 hushkotoi wano[2]
iki-an aine

Griko
(Italy)

''Ma ti eχi' pu klei'?''
''Iklèo *ka* ìtela n'armastò.''
I *vèkkia* àggale tria dattilìtia:

Yakkha
(Nepal)

मा, ना चिगा निङ्वामाङ् ओम,
हाखोक्ङागो लेम्साङ् खाʔला लुया,
"पिछानाछा लेङ्माहोङ प्याक छो छो
लाप्लाप मेन्जोक्माहा।"

Rijhwani, Anastasopoulos, Neubig. EMNLP 2020.

Evaluation dataset for low-resource OCR

Ainu (Japan)

kira-an patek
aeyairamshitne[1]
5760 hushkotoi wano[2]
iki-an aine

Griko (Italy)

''Ma ti eχi' pu klei'?''
''Iklèo *ka* ìtela n'armastò.''
I *vèkki*a àggale tria dattilìtia:

Yakkha (Nepal)

मा, ना चिगा निङ्वामाङ् ओम,
हाखोक्ङागो लेम्साङ् खा?ला लुया,
"पिछानाछा लेङ्माहोङ प्याक छो छो
लाप्लाप मेन्जोक्माहा।"

Kwak'wala (Canada)

q!âLElax gwēg·ilasasa lExēläxa lExaᵋyē
lExEläsa nEkwäxa nEk!ŭlē. Wä, hëᵋn
wä, lä hëlēda ᵋnEmsgEmē; wä, hëᵋmi
lExEläs. Wä hëEm ḷēgEmsa ᵋwālēgaᵋ

Rijhwani, Anastasopoulos, Neubig. EMNLP 2020.

Evaluation dataset for low-resource OCR

Ainu
(Japan)

kira-an patek
aeyairamshitne[1]
5760 hushkotoi wano[2]
iki-an aine

Griko
(Italy)

"Ma ti eχi' pu klei'?"
"Iklèo *ka* ìtela n'armastò."
I *vèkkia* àggale tria dattilìtia:

Yakkha
(Nepal)

मा, ना चिगा निङ्वामाङ् ओम,
हाखोक्डागो लेम्साङ् खाʔला लुया,
"पिछानाछा लेङ्माहोङ प्याक छो छो
लाप्लाप मेन्जोक्माहा।"

Kwak'wala
(Canada)

q!âʟᴇlax gwēg·ilasasa lᴇxēläxa lᴇxaᵋyē
lᴇxᴇläsa nᴇkwäxa nᴇk!ŭlē. Wä, hēᵋn
wä, lä hëlēda ᵋnᴇmsgᴇmē; wä, hëᵋmi
lᴇxᴇläs. Wä hëᴇm ḷēgᴇmsa ᵋwālēgaᵋ

- Orthographically, typologically, geographically diverse

Evaluation dataset for low-resource OCR

Ainu (Japan)

kira-an patek
aeyairamshitne(1)
5760 hushkotoi wano(2)
iki-an aine

Latin

Griko (Italy)

”Ma ti eχi’ pu klei’?”
”Iklèo *ka* ìtela n’armastò.”
I *vèkkia* àggale tria dattilìtia:

Yakkha (Nepal)

मा, ना चिगा निङ्वामाङ् ओम,
हाखोक्ङागो लेम्साङ् खाʔला लुया,
“पिछानाछा लेङ्माहोङ प्याक छो छो
लाप्लाप मेन्जोक्माहा।”

Kwak’wala (Canada)

q!âLElax gwēg·ilasasa lExēläxa lExaᵋyē
lExEläsa nEkwäxa nEk!ŭlē. Wä, hēᵋn
wä, lä hëlēda ᵋnEmsgEmē; wä, hëᵋmi
lExEläs. Wä hëEm ḷēgEmsa ᵋwālēgaᵋ

- Orthographically, typologically, geographically diverse

Evaluation dataset for low-resource OCR

Ainu (Japan)

kira-an patek
aeyairamshitne(1)
5760 hushkotoi wano(2)
iki-an aine

Latin

Griko (Italy)

"Ma ti eχi' pu klei'?"
"Iklèo *ka* ìtela n'armastò."
I *vèkkia* àggale tria dattilìtia:

Latin+Greek

Yakkha (Nepal)

मा, ना चिगा निङ्वामाङ् ओम,
हाखोक्ङागो लेम्साङ् खाʔला लुया,
“पिछानाछा लेङ्माहोङ प्याक छो छो
लाप्लाप मेन्जोक्माहा।”

Kwak'wala (Canada)

q!âLElax gwēg·ilasasa lExēläxa lExaᵋyē
lExEläsa nEkwäxa nEk!ŭlē. Wä, hēᵋn
wä, lä hēlēda ᵋnEmsgEmē; wä, hēᵋmi
lExEläs. Wä hëEm ļēgEmsa ᵋwālēgaᵋ

- Orthographically, typologically, geographically diverse

Evaluation dataset for low-resource OCR

Ainu (Japan)	kira-an patek aeyairamshitne(1) 5760 hushkotoi wano(2) iki-an aine	Latin
Griko (Italy)	”Ma ti eχi’ pu klei’?” ”Iklèo *ka* ìtela n’armastò.” I *vèkkia* àggale tria dattilìtia:	Latin+Greek
Yakkha (Nepal)	मा, ना चिगा निङ्वामाङ् ओम, हाखोक्ङागो लेम्साङ् खा?ला लुया, “पिछानाछा लेङ्माहोङ प्याक छो छो लाप्लाप मेन्जोक्माहा।”	Devanagari
Kwak’wala (Canada)	q!âʟElax gwēg·ilasasa lExēläxa lExaᵋyē lExEläsa nEkwäxa nEk!ŭlē. Wä, hēᵋn wä, lä hëlēda ᵋnEmsgEmē; wä, hëᵋmi lExEläs. Wä hëEm ḷēgEmsa ᵋwālēgaᵋ	

- Orthographically, typologically, geographically diverse

Evaluation dataset for low-resource OCR

Ainu (Japan)

kira-an patek
aeyairamshitne(1)
5760 hushkotoi wano(2)
iki-an aine

Latin

Griko (Italy)

”Ma ti eχi’ pu klei’?”
”Iklèo *ka* ìtela n’armastò.”
I *vèkkia* àggale tria dattilìtia:

Latin+Greek

Yakkha (Nepal)

मा, ना चिगा निङ्वामाङ् ओम,
हाखोक्ङागो लेम्साङ् खाʔला लुया,
“पिछानाछा लेङ्माहोङ पयाक छो छो
लाप्लाप मेन्जोक्माहा।”

Devanagari

Kwak’wala (Canada)

q!âʟᴇlax gwēg·ilasasa lᴇxēläxa łᴇxaᵋyē
łᴇxᴇläsa nᴇkwäxa nᴇk!ŭłē. Wä, hēᵋn
wä, lä hēłēda ᵋnᴇmsgᴇmē; wä, hēᵋmi
łᴇxᴇläs. Wä hēᴇm ḻēgᴇmsa ᵋwālēgaᵋ

Boas

- Orthographically, typologically, geographically diverse

Rijhwani, Anastasopoulos, Neubig. EMNLP 2020.

Evaluation dataset for low-resource OCR

Ainu (Japan)

kira-an patek
aeyairamshitne(1)
5760 hushkotoi wano(2)
iki-an aine

Griko (Italy)

"Ma ti eχi' pu klei'?"
"Iklèo *ka* ìtela n'armastò."
I *vèkkia* àggale tria dattilìtia:

Yakkha (Nepal)

मा, ना चिगा निङ्वामाङ् ओम,
हाखोक्डागो लेम्साङ् खाʔला लुया,
"पिछानाछा लेङ्माहोङ प्याक छो छो
लापलाप मेन्जोक्माहा।"

Kwak'wala (Canada)

q!âʟᴇlax gwēg·ilasasa lᴇxēläxa lᴇxaᵋyē
lᴇxᴇläsa nᴇkwäxa nᴇk!ŭlē. Wä, hēᵋn
wä, lä hēlēda ᵋnᴇmsgᴇmē; wä, hēᵋmi
lᴇxᴇläs. Wä hēᴇm ʟ̣ēgᴇmsa ᵋwālēgaᵋ

- Orthographically, typologically, geographically diverse

- The languages currently have:
  - No Wikipedia/Common Crawl text
  - Not supported by multilingual LMs
  - No easily accessible bilingual lexica

Rijhwani, Anastasopoulos, Neubig. EMNLP 2020.

Evaluation dataset for low-resource OCR

Ainu (Japan)

kira-an patek
aeyairamshitne[1]
5760 hushkotoi wano[2]
iki-an aine

Griko (Italy)

"Ma ti eχi' pu klei'?"
"Iklèo *ka* ìtela n'armastò."
I *vèkkia* àggale tria dattilìtia:

Yakkha (Nepal)

मा, ना चिगा निङ्वामाङ् ओम,
हाखोक्ङागो लेम्साङ् खाʔला लुया,
"पिछानाछा लेङ्माहोङ प्याक छो छो
लाप्लाप मेन्जोक्माहा।"

Kwak'wala (Canada)

q!âLElax gwēg·ilasasa lExēläxa lExaᵋyē
lExEläsa nEkwäxa nEk!ŭlē. Wä, hēᵋn
wä, lä hēlēda ᵋnEmsgEmē; wä, hēᵋmi
lExEläs. Wä hëEm ḷēgEmsa ᵋwālēgaᵋ

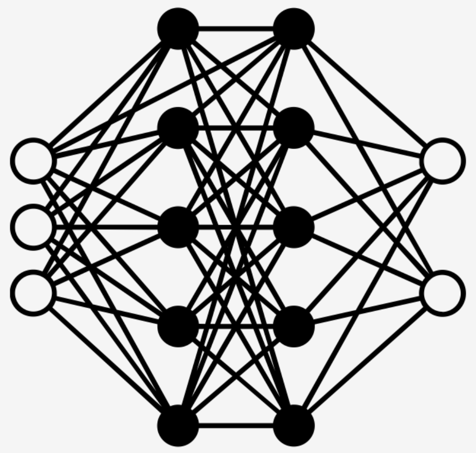
- Orthographically, typologically, geographically diverse
- The languages currently have:
  - No Wikipedia/Common Crawl text
  - Not supported by multilingual LMs
  - No easily accessible bilingual lexica
- <1000 transcribed lines per language

Rijhwani, Anastasopoulos, Neubig. EMNLP 2020.

Existing OCR methods

Existing OCR methods

Supervised

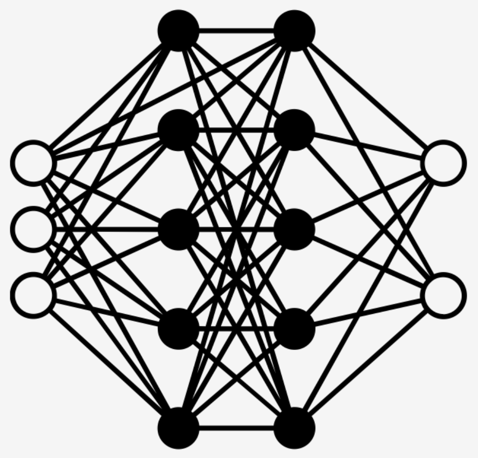


Large neural networks

Requires: 10000s of transcribed images

Existing OCR methods

Supervised



Large neural networks

Requires: 10000s of transcribed images

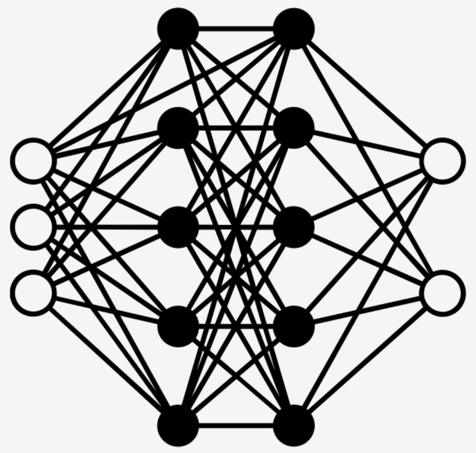
Unsupervised

- Unlabeled images
- Language model

Requires: text corpus or lexicon in the target language

Existing OCR methods

Supervised



Large neural networks

Requires: 10000s of transcribed images

Unsupervised

- Unlabeled images
- Language model

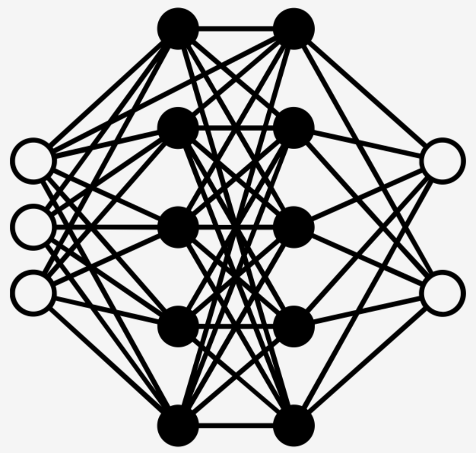
Requires: text corpus or lexicon in the target language

Off-the-shelf

- Support ~100 languages
- Not trained on our target languages
- Can act as a general character recognizer for many scripts

Existing OCR methods

Supervised



Large neural networks

Requires: 10000s of transcribed images

Unsupervised

- Unlabeled images
- Language model

Requires: text corpus or lexicon in the target language

Off-the-shelf

- Support ~100 languages
- Not trained on our target languages
- Can act as a general character recognizer for many scripts

Existing OCR methods

Supervised

Large neural networks

Requires: 10000s of transcribed images

Unsupervised

- Unlabeled images
- Language model

Requires: **text corpus or lexicon in the target language**

Off-the-shelf

- Support ~100 languages
- Not trained on our target languages
- Can act as a **general character recognizer for many scripts**

Existing OCR methods: promises and pitfalls

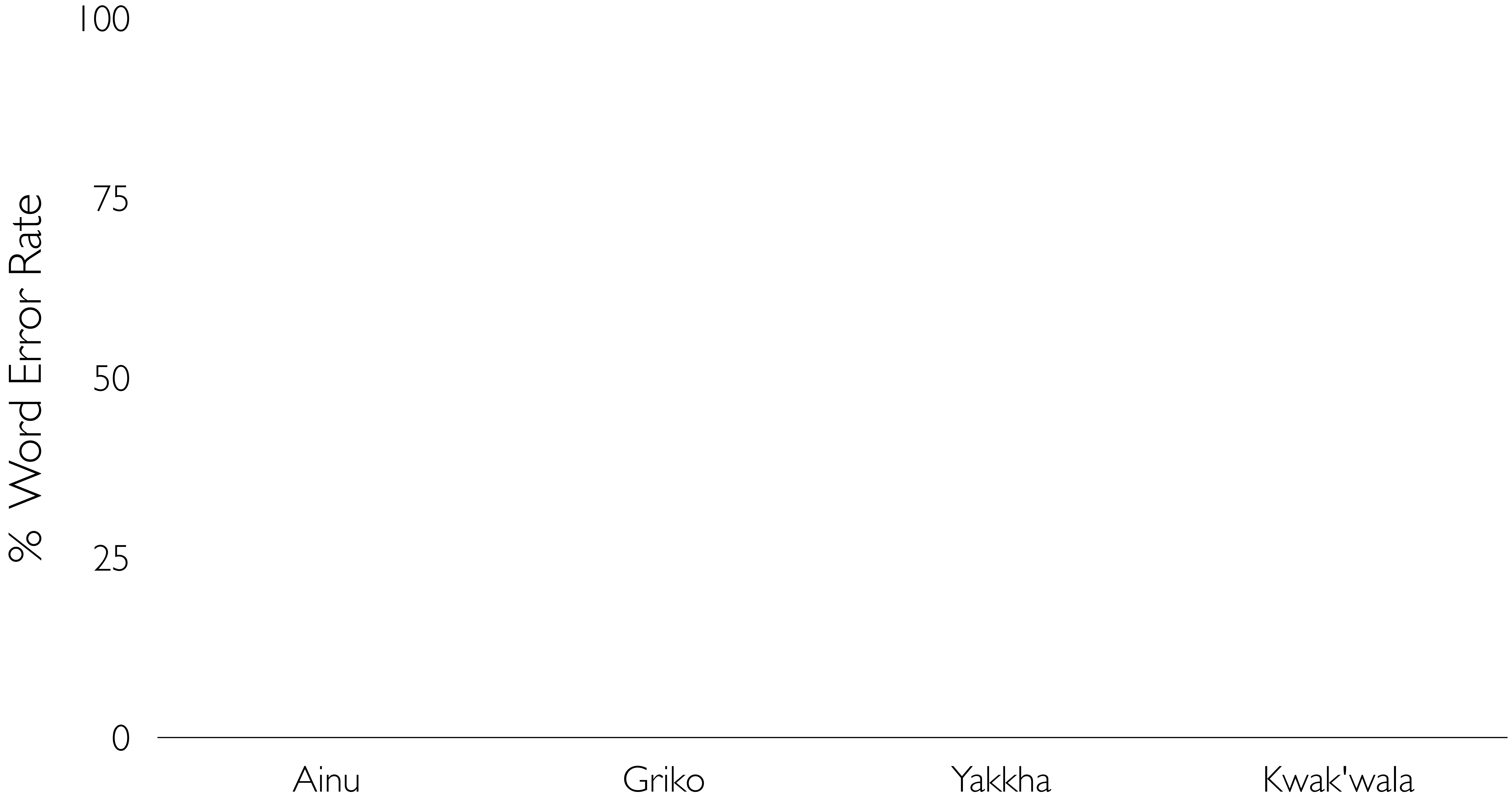
Existing OCR methods: promises and pitfalls

Existing OCR methods: promises and pitfalls

$$\frac{\text{word edit distance between prediction and reference}}{\text{number of words in reference}}$$

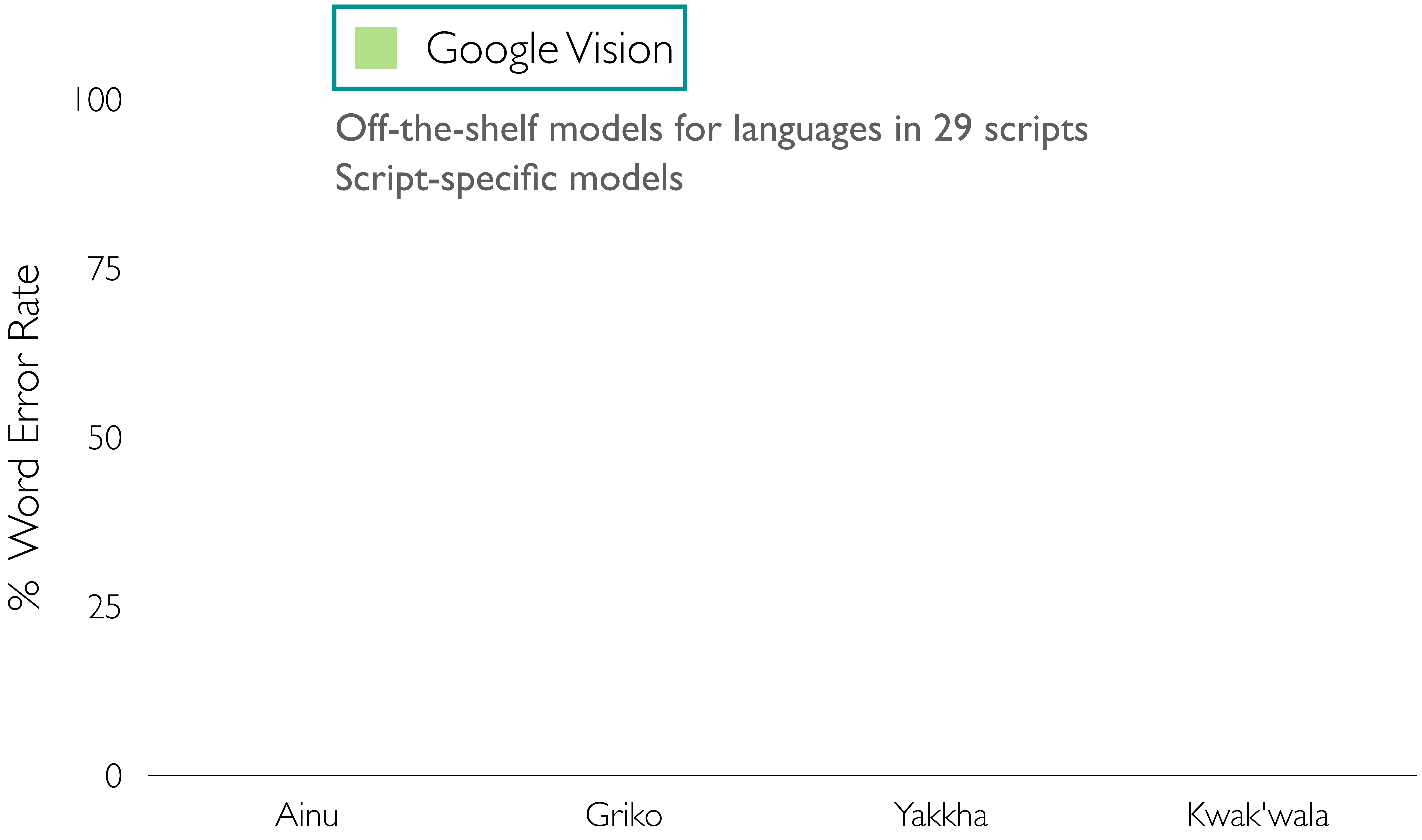
Existing OCR methods: promises and pitfalls

Lower is better!



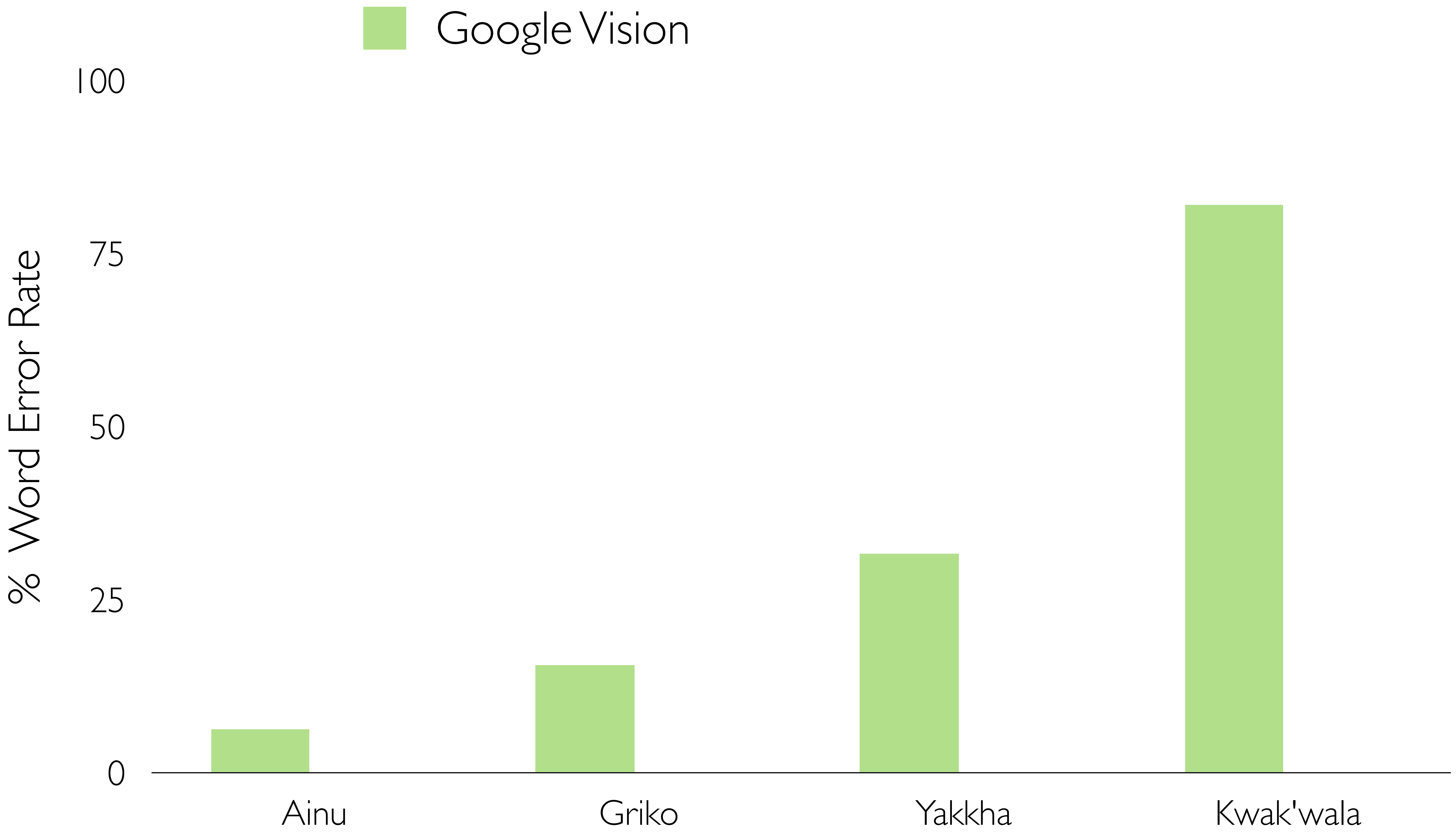
Existing OCR methods: promises and pitfalls

Lower is better!

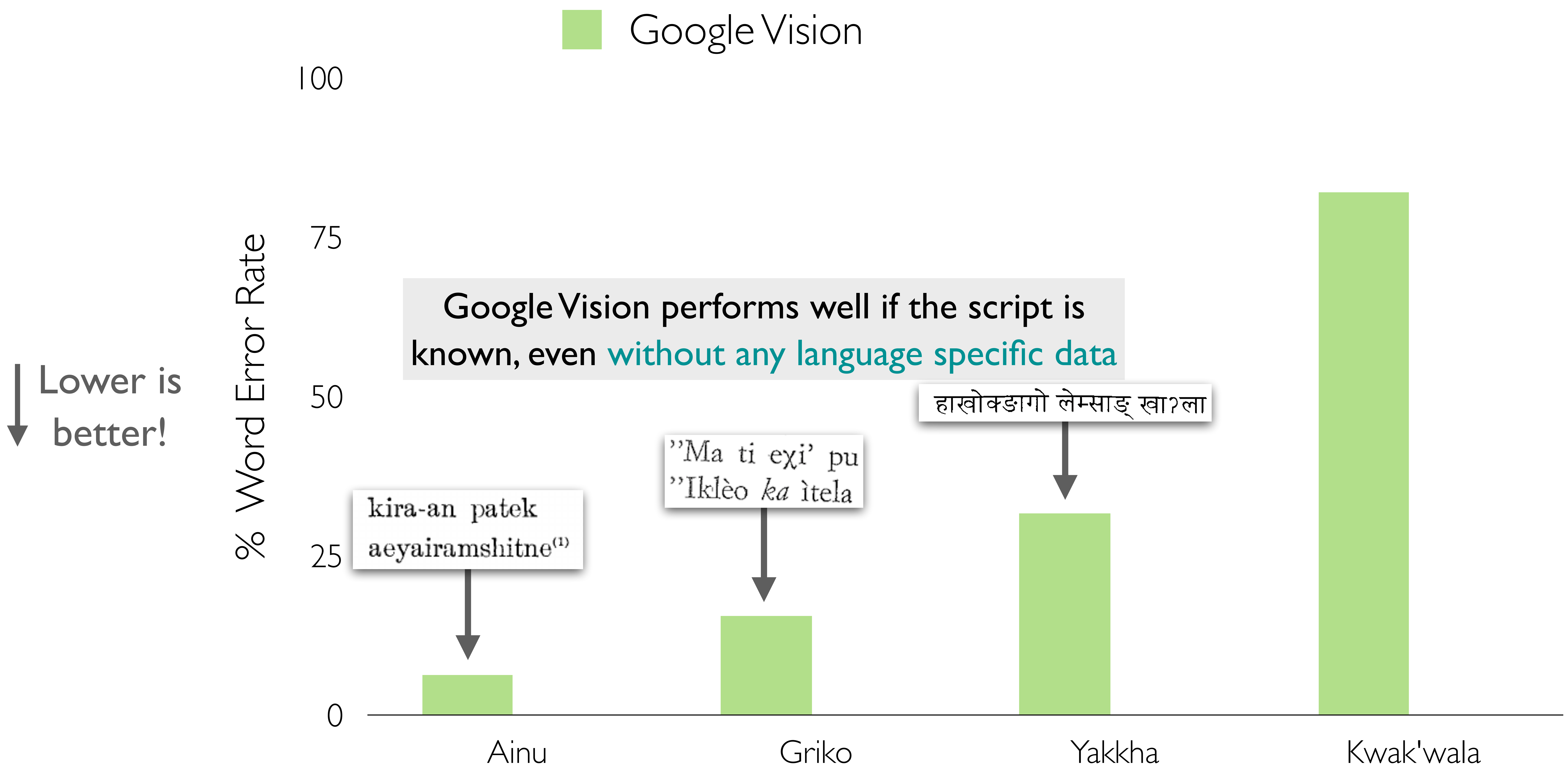


Existing OCR methods: promises and pitfalls

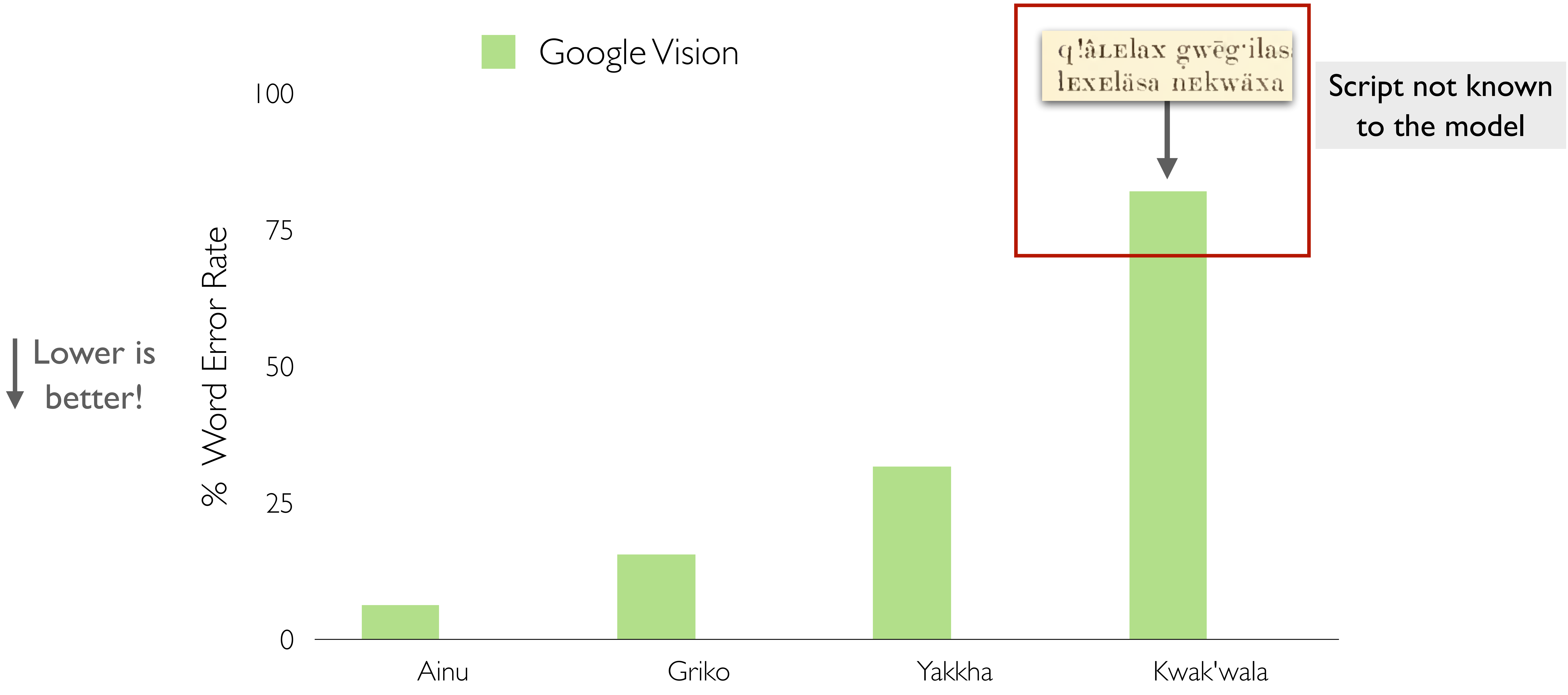
Lower is better!



Existing OCR methods: promises and pitfalls



Existing OCR methods: promises and pitfalls



Existing OCR methods: promises and pitfalls

Existing OCR methods: promises and pitfalls

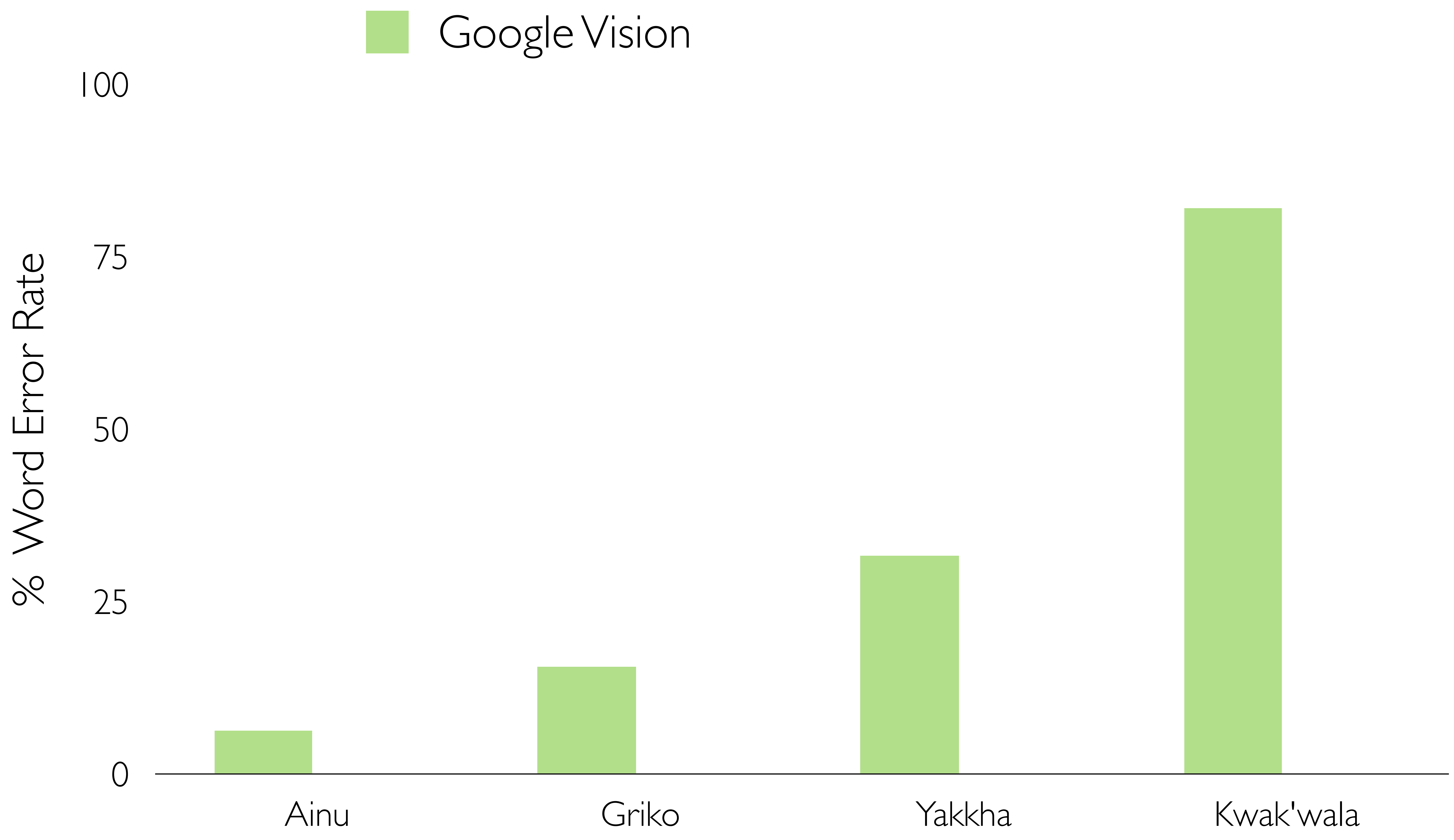
Existing OCR methods: promises and pitfalls

Existing OCR methods: promises and pitfalls

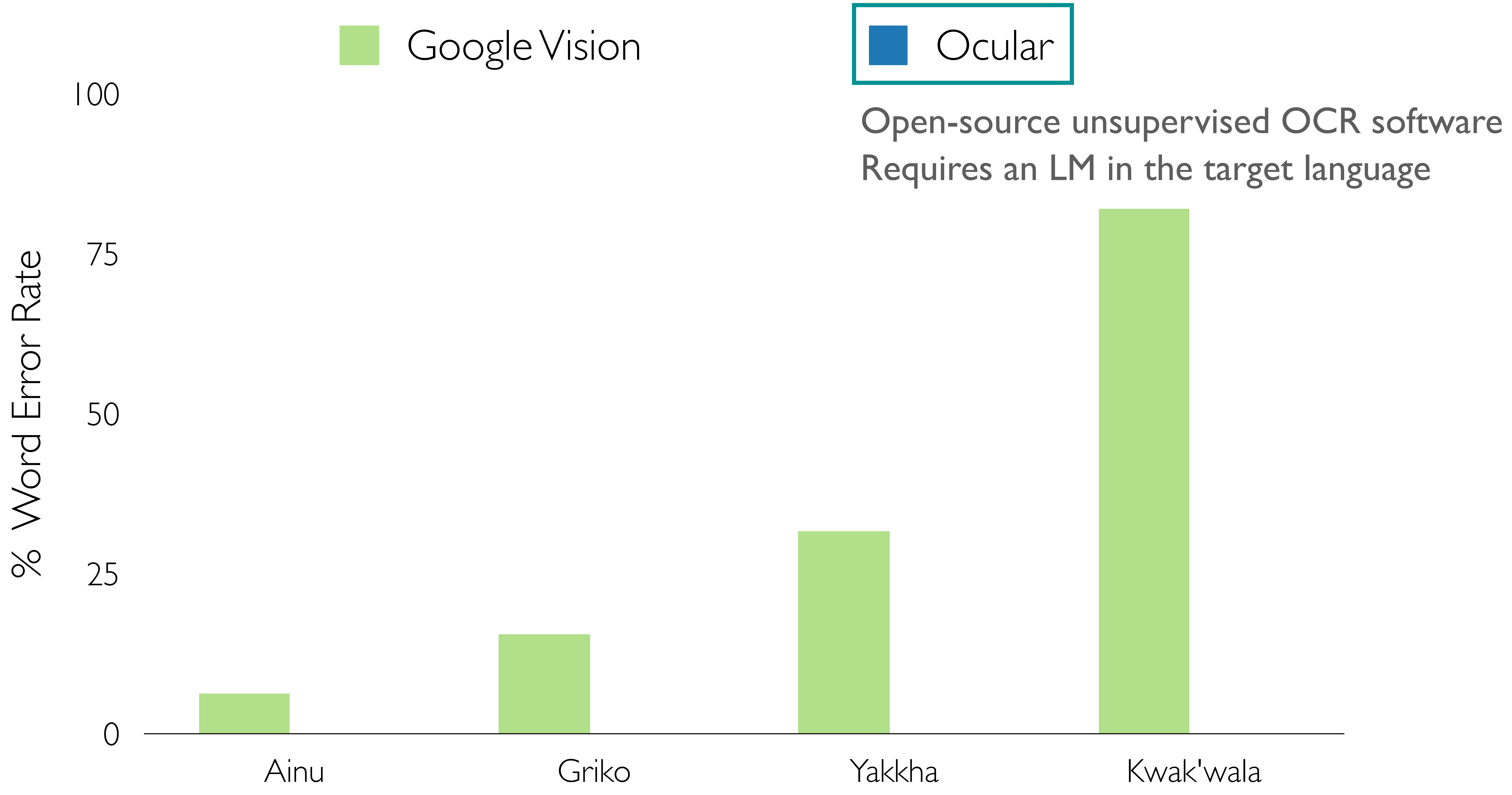
Existing OCR methods: promises and pitfalls

Existing OCR methods: promises and pitfalls

Lower is better!

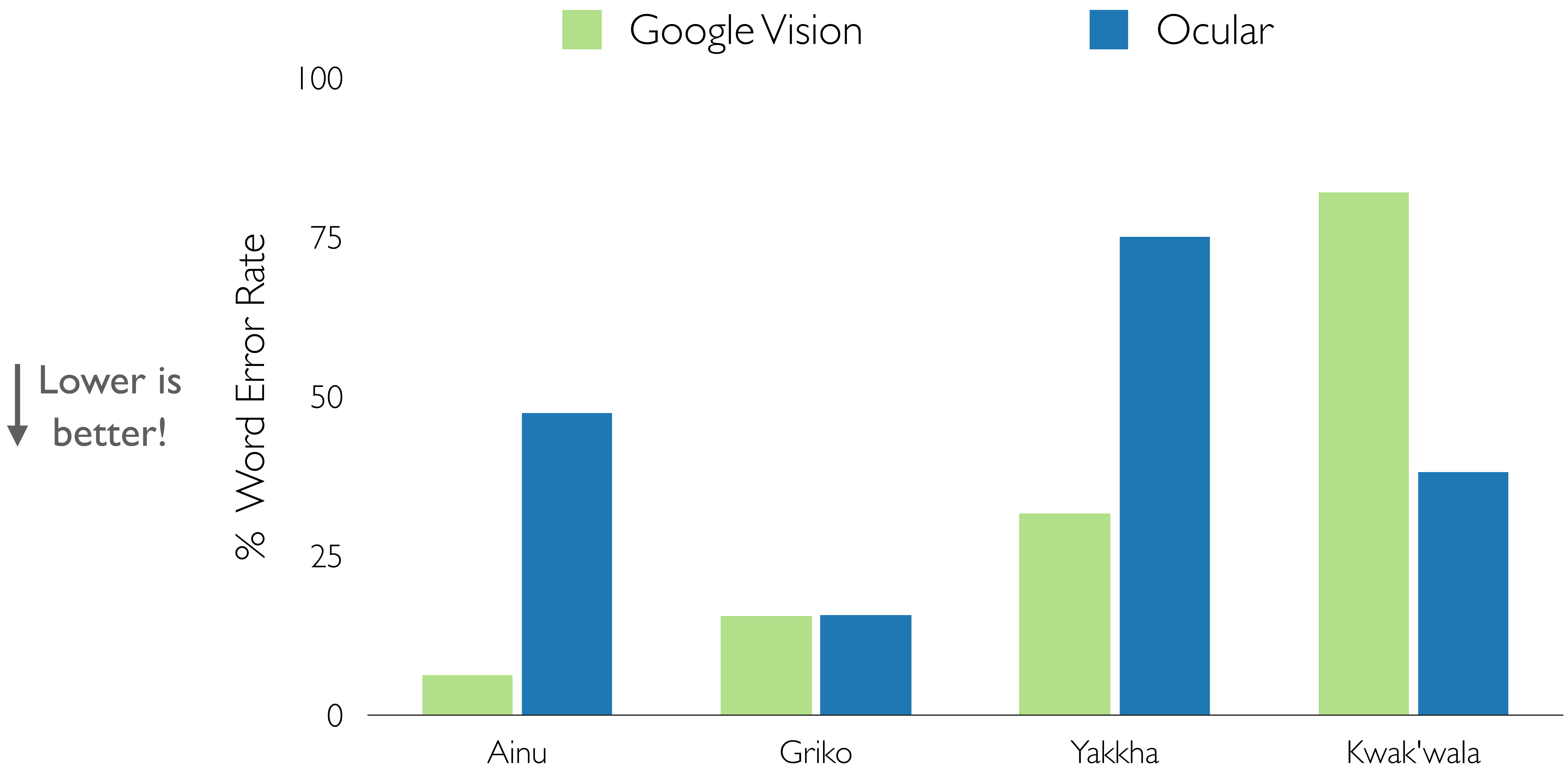


Existing OCR methods: promises and pitfalls

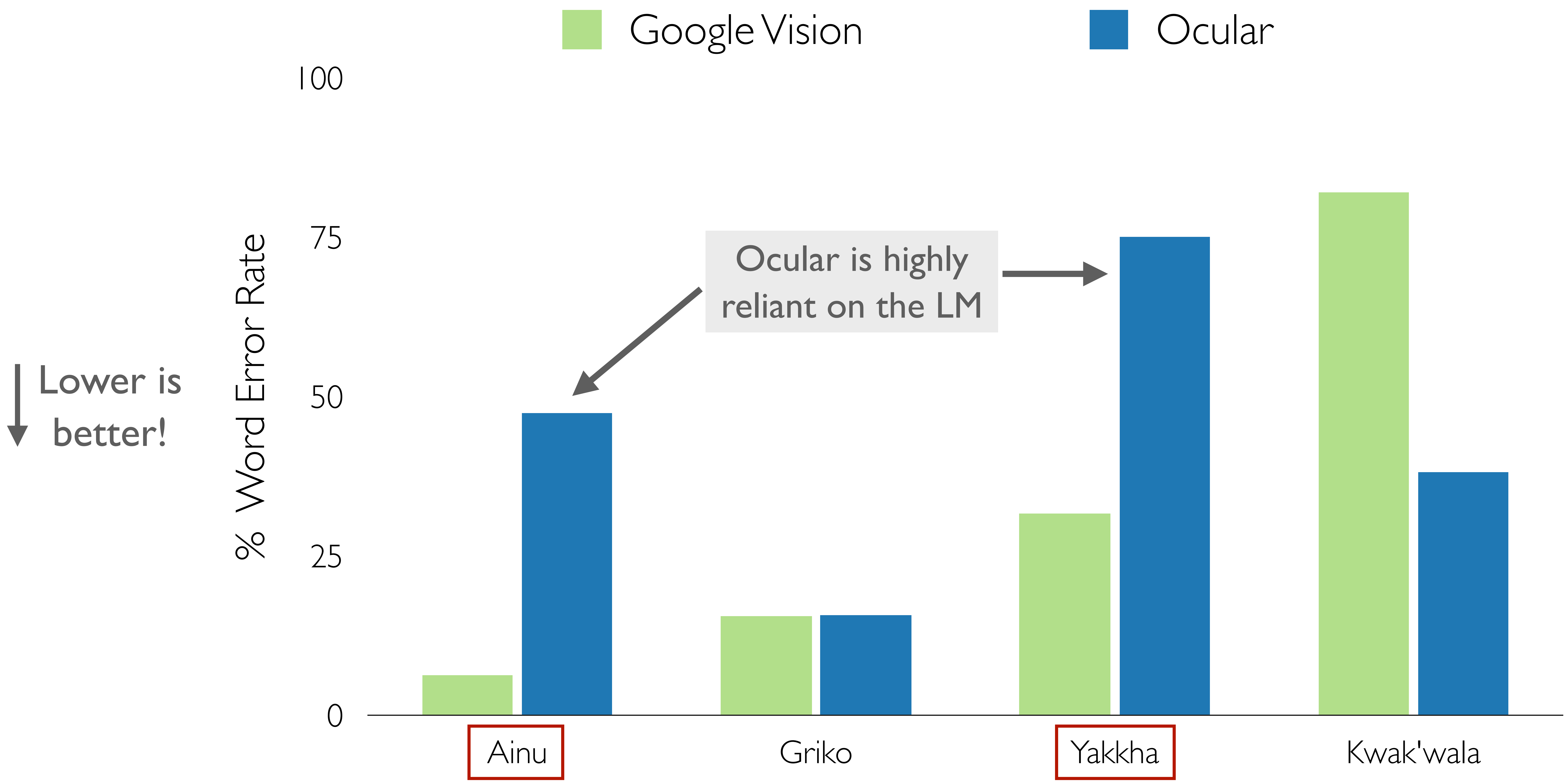


Lower is better!

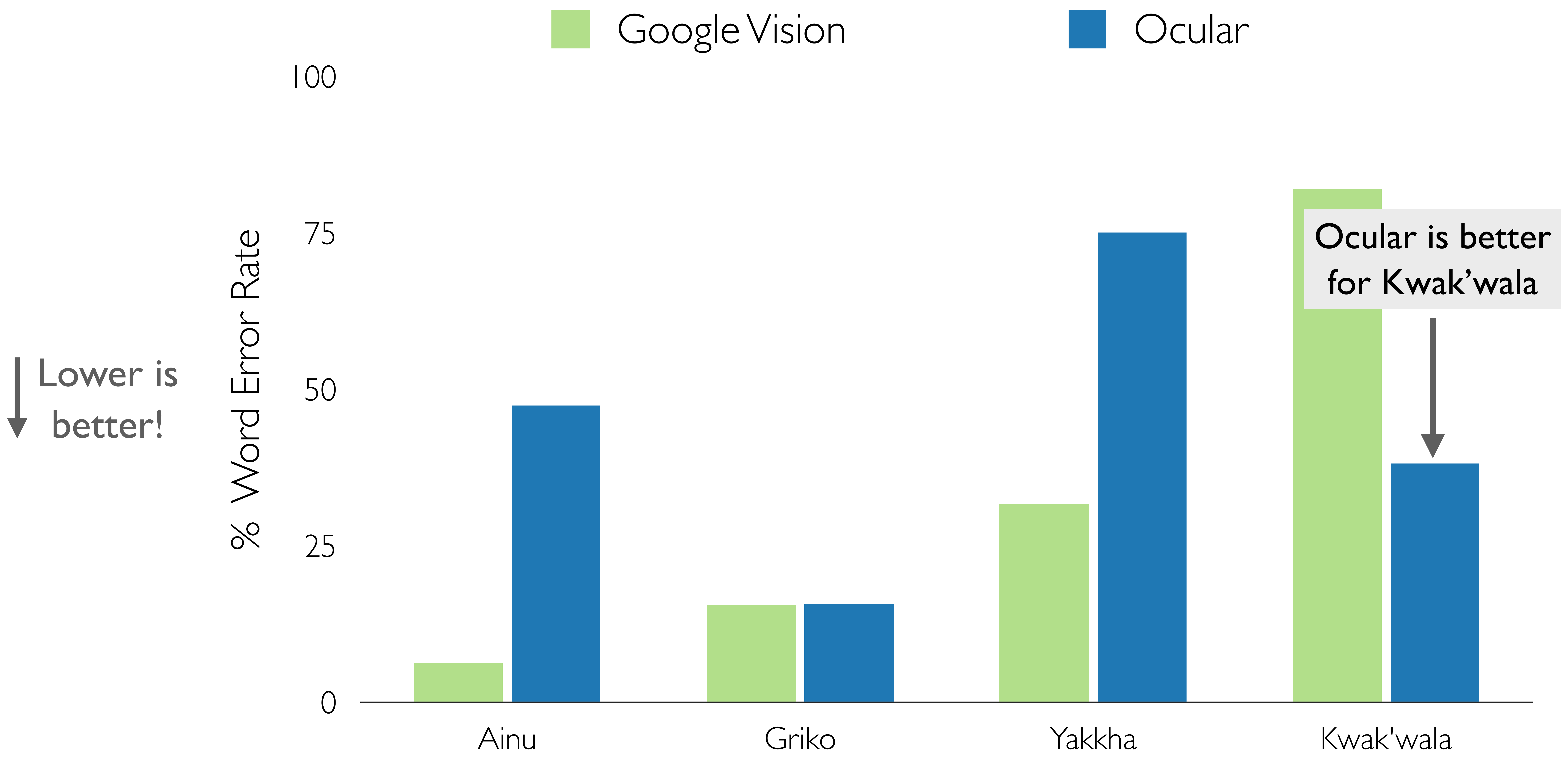
Existing OCR methods: promises and pitfalls



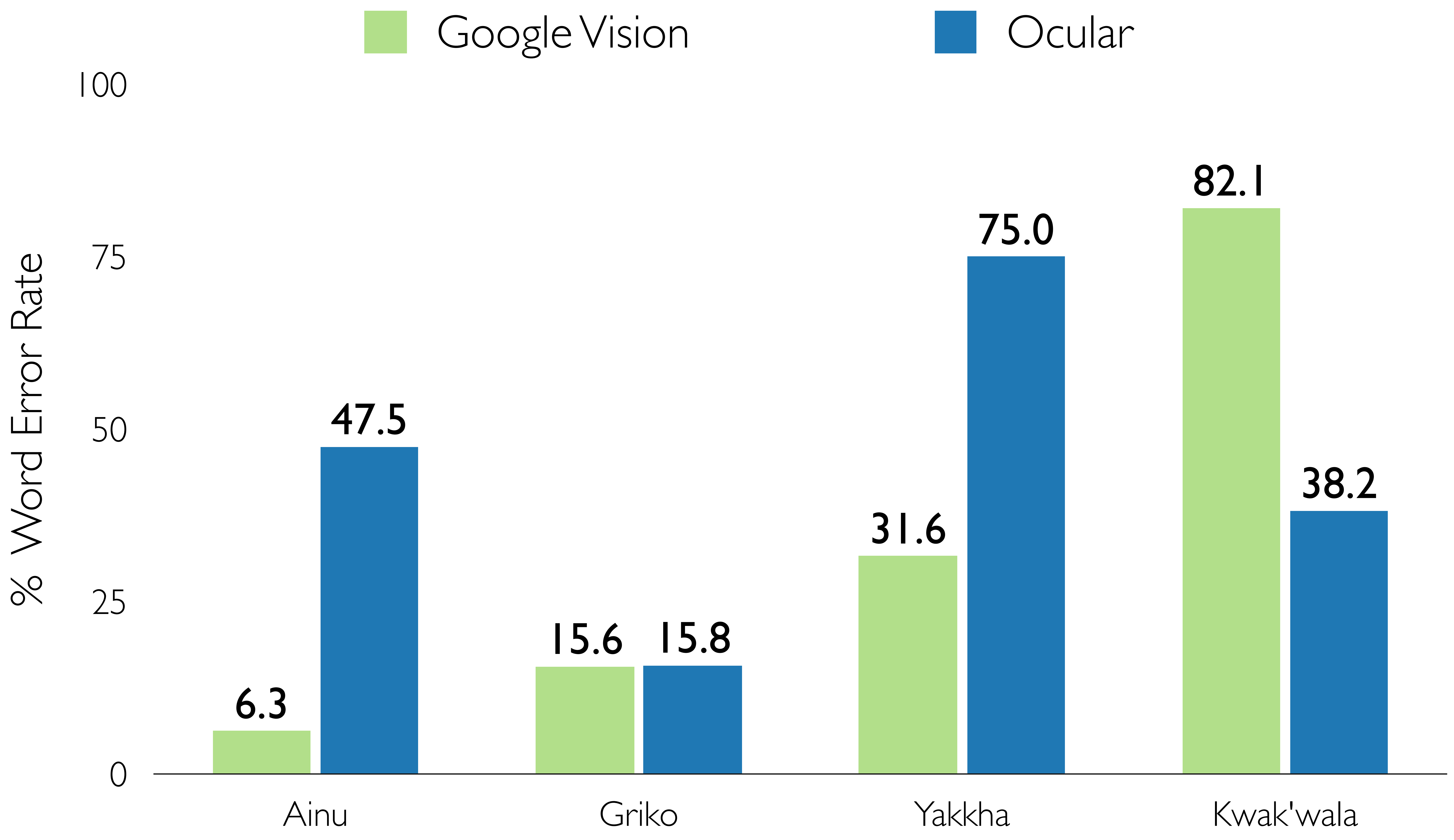
Existing OCR methods: promises and pitfalls



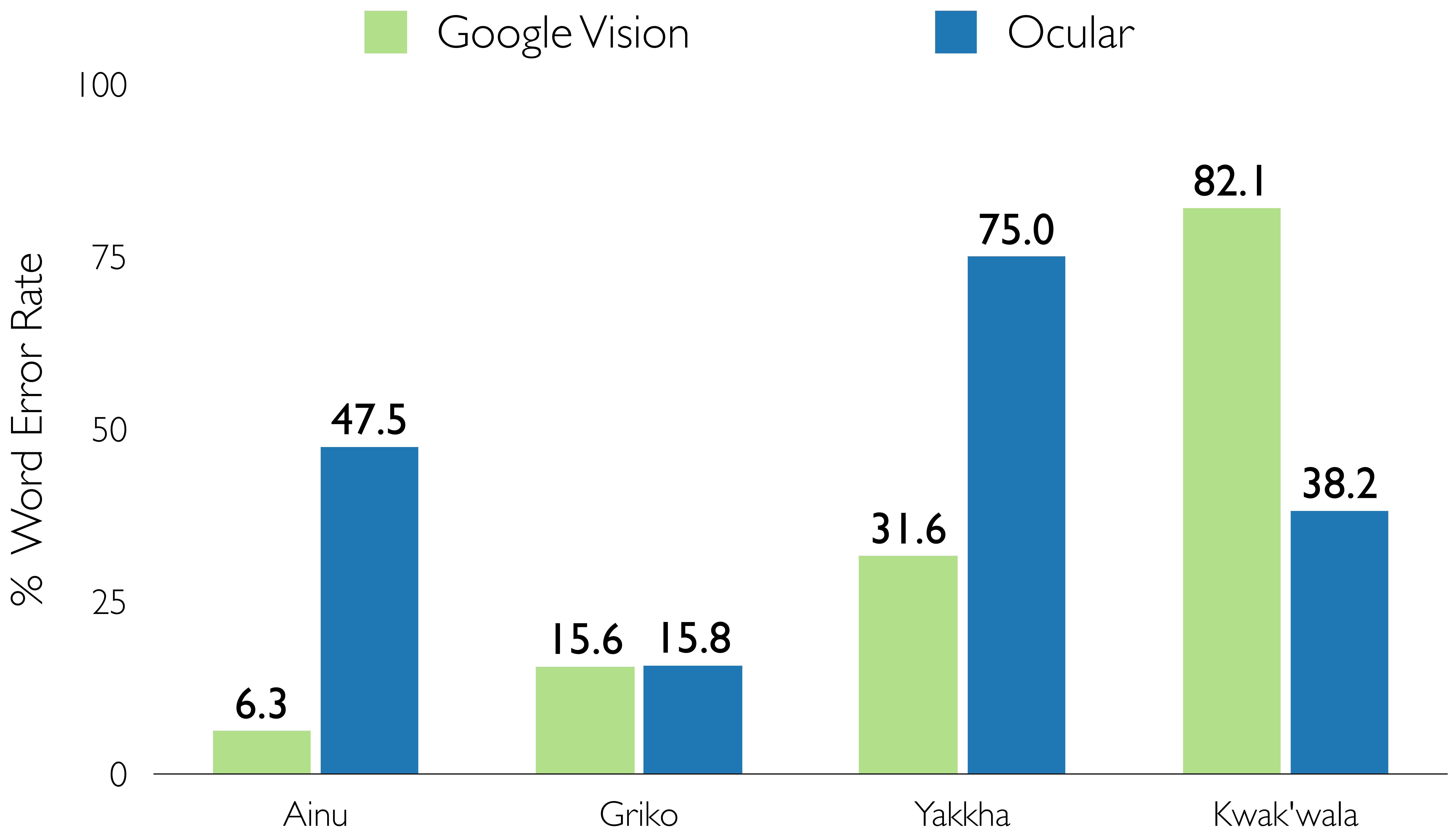
Existing OCR methods: promises and pitfalls



Existing OCR methods: promises and pitfalls

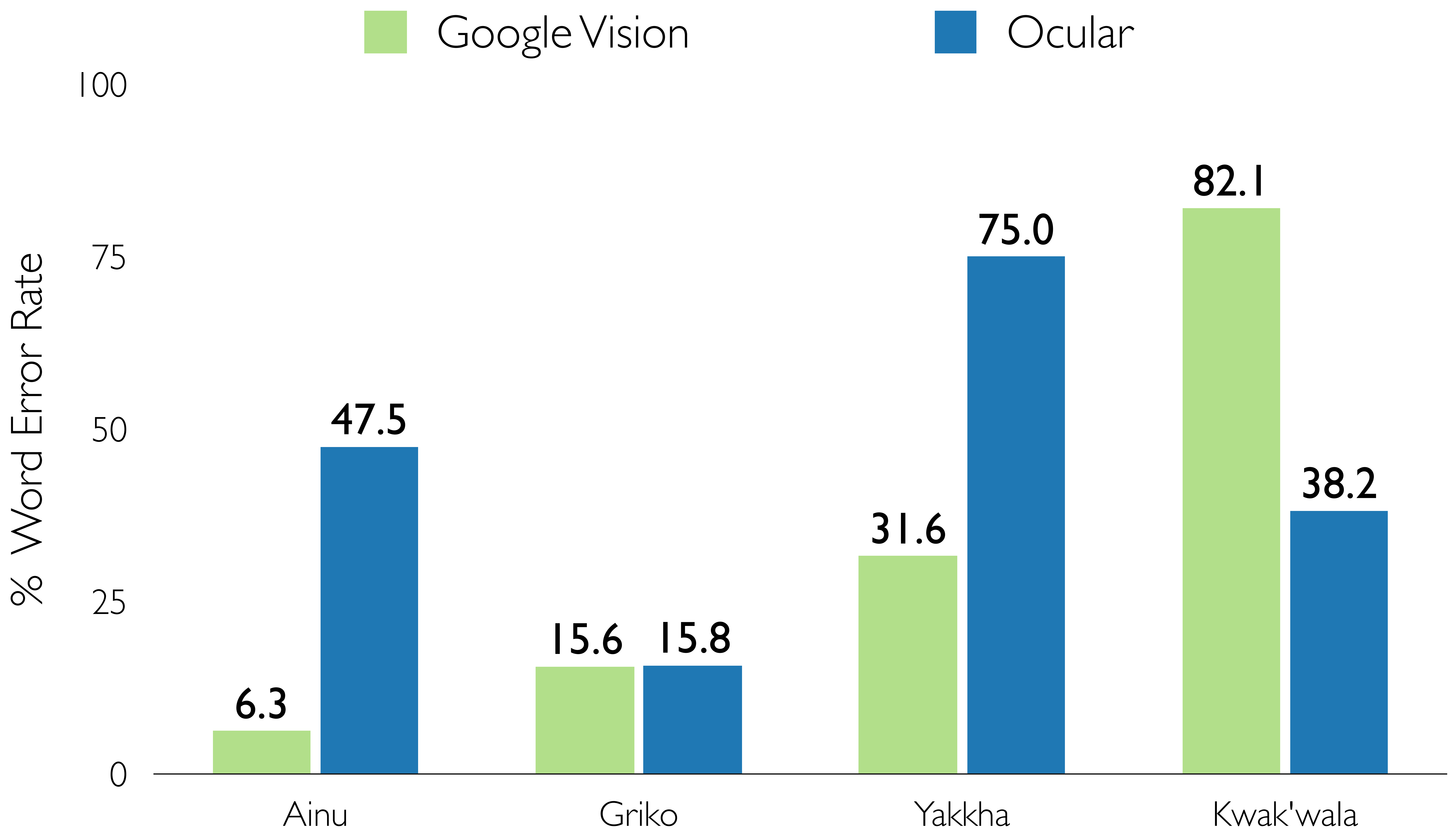


Existing OCR methods: promises and pitfalls



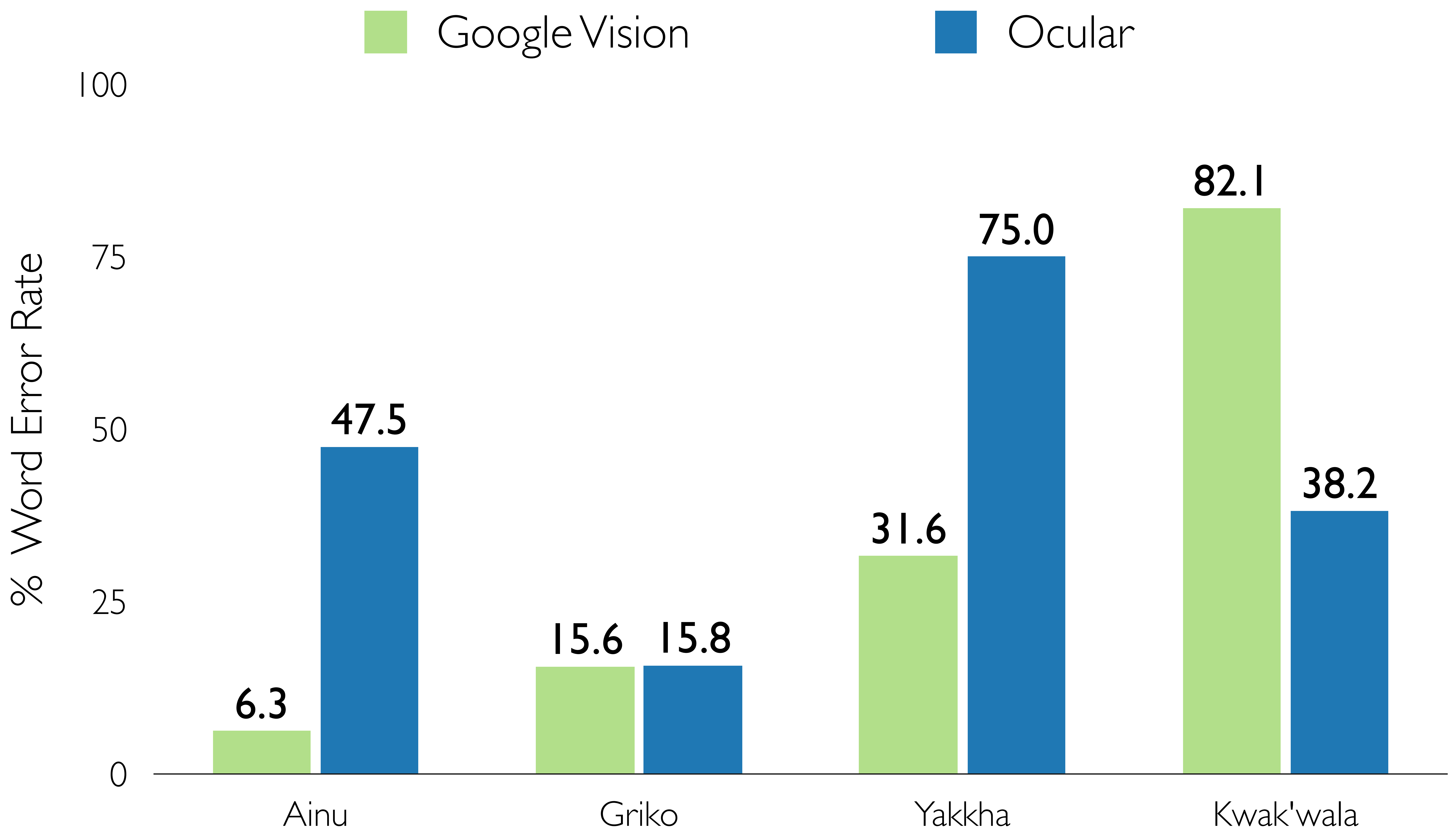
- Considerable room for improvement compared to high-resource languages

Existing OCR methods: promises and pitfalls



- Considerable room for improvement compared to high-resource languages
- Recognizes the majority of words correctly

Existing OCR methods: promises and pitfalls



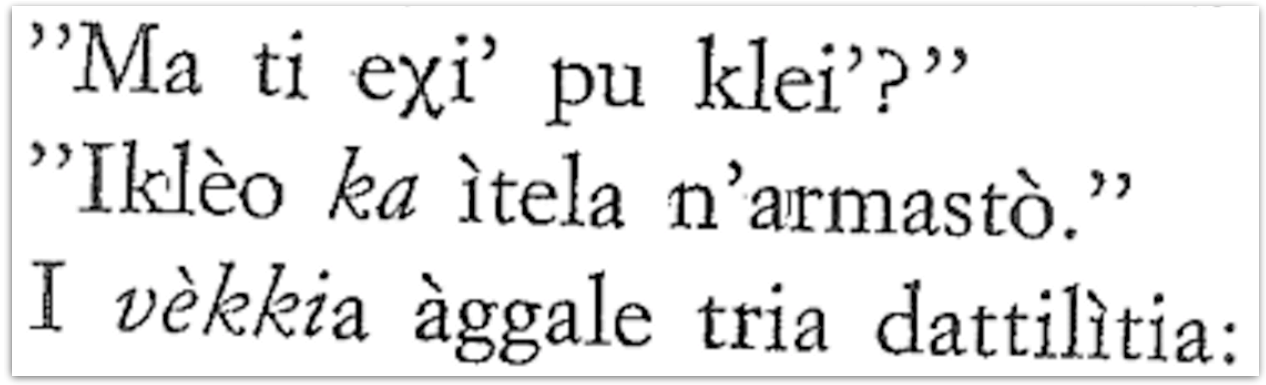
- **Considerable room for improvement** compared to high-resource languages
- Recognizes the majority of words correctly
- **Reliable starting point** for further improvements

Improving the results of existing OCR systems

Improving the results of existing OCR systems

"Ma ti eχi' pu klei'?"
"Iklèo *ka* ìtela n'armastò."
I *vèkki*a àggale tria dattilìtia:

Improving the results of existing OCR systems


”Ma ti eχi’ pu klei’?”
”Iklèo *ka* ìtela n’armastò.”
I *vèkkia* àggale tria dattilìtia:

”Ma ti e**x**i’ pu klei’?”
”Ikl**e**o ka ìtela _armastò.”
I v**e**kkia **a**ggale tria dattil**i**tia:

OCR output (“first pass”)

Improving the results of existing OCR systems

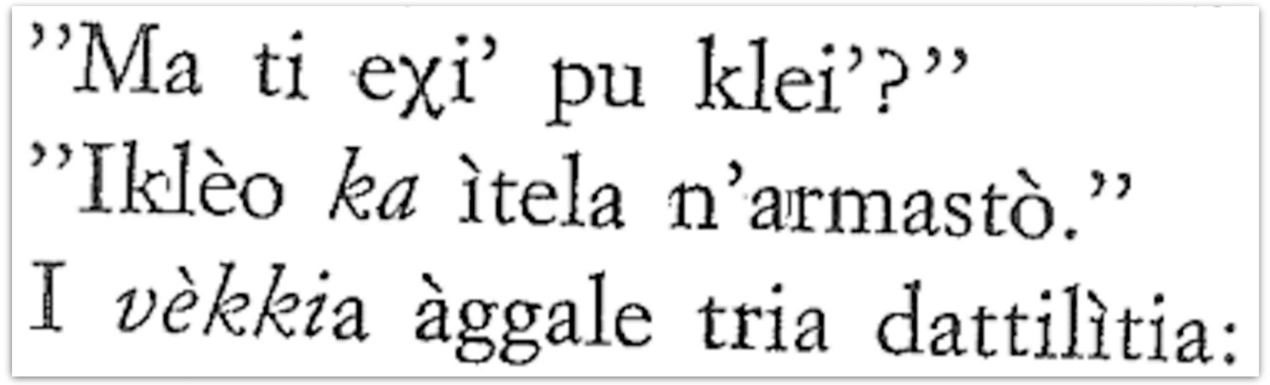
"Ma ti eχi' pu klei'?"
"Iklèo *ka* ìtela n'armastò."
I *vèkkia* àggale tria dattilìtia:

"Ma ti e**x**i' pu klei'?"
"Ikl**e**o ka ìtela **_**armastò."
I v**e**kkia **a**ggale tria dattil**i**tia:

OCR output ("first pass")

OCR output has
some errors

Improving the results of existing OCR systems

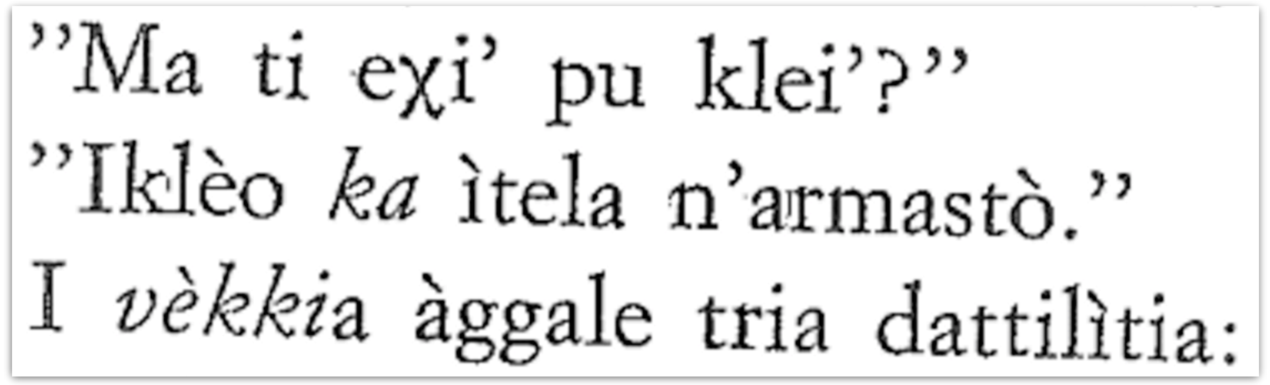

”Ma ti eχi’ pu klei’?”
”Iklèo *ka* ìtela n’armastò.”
I *vèkkia* àggale tria dattilìtia:

”Ma ti e**x**i’ pu klei’?”
”Ikl**e**o ka ìtela _armastò.”
I v**e**kkia **a**ggale tria dattil**i**tia:

OCR output (“first pass”)

Automatic OCR
Post-Correction

Improving the results of existing OCR systems


”Ma ti eχi’ pu klei’?”
”Iklèo *ka* ìtela n’armastò.”
I *vèkkia* àggale tria dattilìtia:

”Ma ti e**x**i’ pu klei’?”
”Ikl**e**o ka ìtela **_**armastò.”
I v**e**kkia **a**ggale tria dattil**i**tia:

OCR output (“first pass”)

Automatic OCR Post-Correction

”Ma ti e**χ**i’ pu klei’?”
”Ikl**è**o ka ìtela **n’**armastò.”
I v**è**kkia **à**ggale tria dattil**ì**tia:

Corrected transcription

Improving the results of existing OCR systems

”Ma ti eχi’ pu klei’?”
”Iklèo *ka* ìtela n’armastò.”
I *vèkkia* àggale tria dattilìtia:

”Ma ti exi’ pu klei’?”
”Ikleo ka ìtela _armastò.”
I vekkia aggale tria dattilitia:

OCR output (“first pass”)

Automatic OCR
Post-Correction

”Ma ti eχi’ pu klei’?”
”Iklèo ka ìtela n’armastò.”
I vèkkia àggale tria dattilìtia:

Corrected transcription

Previous work: improve results for unseen fonts, layouts, domains.
This talk: low-resourced languages.

Improving the results of existing OCR systems

”Ma ti e**x**i’ pu klei’?”
”Ikl**e**o ka ìtela **_**armastò.”
I v**e**kkia **a**ggale tria dattil**i**tia:

OCR output (“first pass”)

Automatic OCR
Post-Correction

Text-based sequence-to-sequence task

”Ma ti e**χ**i’ pu klei’?”
”Ikl**è**o ka ìtela **n’**armastò.”
I v**è**kkia **à**ggale tria dattil**ì**tia:

Corrected transcription

Improving the results of existing OCR systems

”Ma ti e**x**i’ pu klei’?”
”Ikl**e**o ka ìtela _armastò.”
I v**e**kkia **a**ggale tria dattil**i**tia:

OCR output (“first pass”)

Text input

Automatic OCR
Post-Correction

Text-based sequence-
to-sequence task

”Ma ti e**χ**i’ pu klei’?”
”Ikl**è**o ka ìtela **n’**armastò.”
I v**è**kkia **à**ggale tria dattil**ì**tia:

Corrected transcription

Improving the results of existing OCR systems

"Ma ti exi' pu klei'?"
"Ikleo ka ìtela _armastò."
I vekkia aggale tria dattilitia:

OCR output ("first pass")

Text input

Automatic OCR
Post-Correction

Text-based sequence-to-sequence task

"Ma ti eχi' pu klei'?"
"Iklèo ka ìtela n'armastò."
I vèkkia àggale tria dattilìtia:

Corrected transcription

Text output

Adapting to low-resource settings

Prior work: character-level encoder-decoder with attention

- Add structural biases to the model
  - Diagonal attention loss, copy mechanism, coverage mechanism

Adapting to low-resource settings

Prior work: character-level encoder-decoder with attention

- Add structural biases to the model
  - Diagonal attention loss, copy mechanism, coverage mechanism

- Leverage additional information from the source document

What additional information is available?

Matiaxh	Xhunik	jos.om	marímpa
Mathías	John	work wood (tv).agent	marimba
Matiaxh	Xhunik	wood-worker	(of) marimbas

ruwe-ne noine	もの如く
poro ape are wa	澤山に火を焚きて
hekota rok wa	そこに向ひて坐して
uweneusar[1]	昔噺やお伽などをし
kor okai.	つゝ暮らし居たりき。
Inkar ne wa	只それを見るのみ
akip ne korka	にはあれど
ine-ap-kusu	いかばかり
arushka wa	わが腹立たしくて

ཐག་ཉེ་ (thag-ngea) adv. near, close or at a shorter distance.

ཐག་ཀོ་ (thag-ko) n. rope.

ཐག་ཆོད་ (thag-choth) v. to be dedicated/settled/resolved.

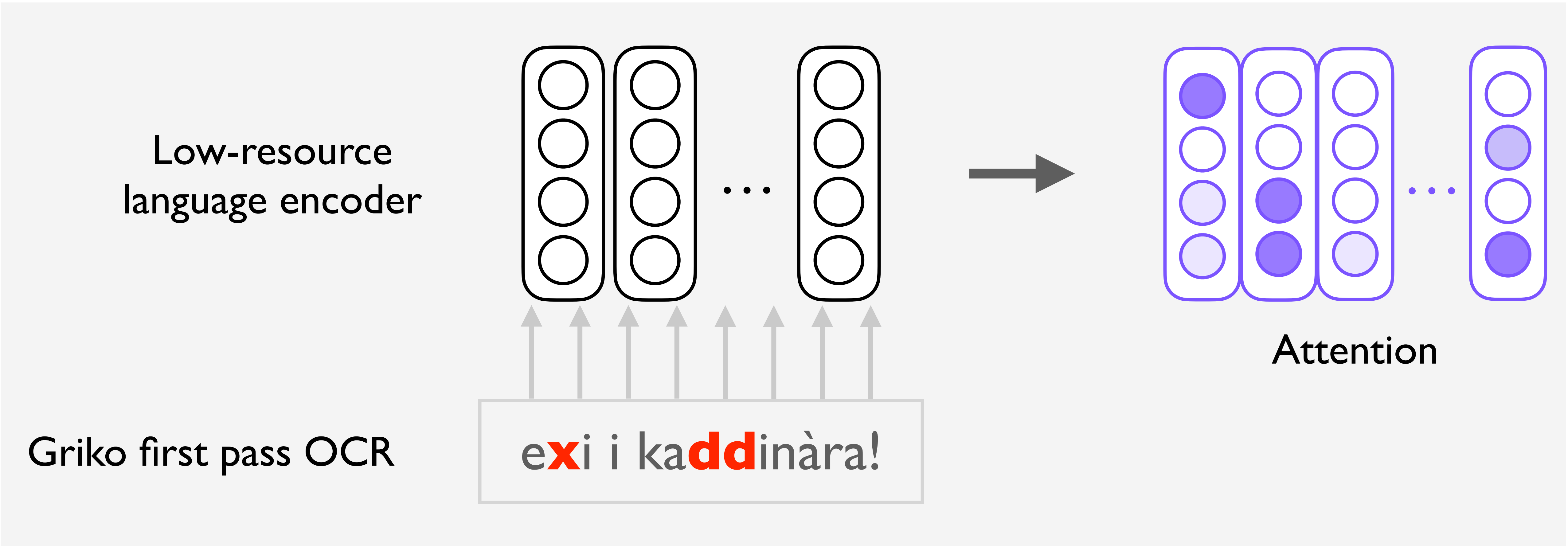
ཐག་མཆོང་ (thag-chong) n. Rope skipping, jumping. v. to skip, to jump.

Seall thall thar an aiseig am fasgadh nan craobh, Am bothan beag glan ud, 's e gealaicht' le aol ; Sud agaibh mo dhachaidh : 's i dachaidh mo ghaoil, Gun chaisteal 'san t-saoghal a 's feàrr leam.	Ayont by the ferry, whaur woodlands are green, My cantie cot housie stan's tidy an' clean ; I envy nae laird in his castle, I ween, I 'm happy an' bien in my ain house.

- Many documents containing text in low-resource languages also **contains a translation of the text**
- Interlinear glosses, dictionaries, linguistic documentation, language learning material…

Multi-source model for post-correction

Multi-source model for post-correction

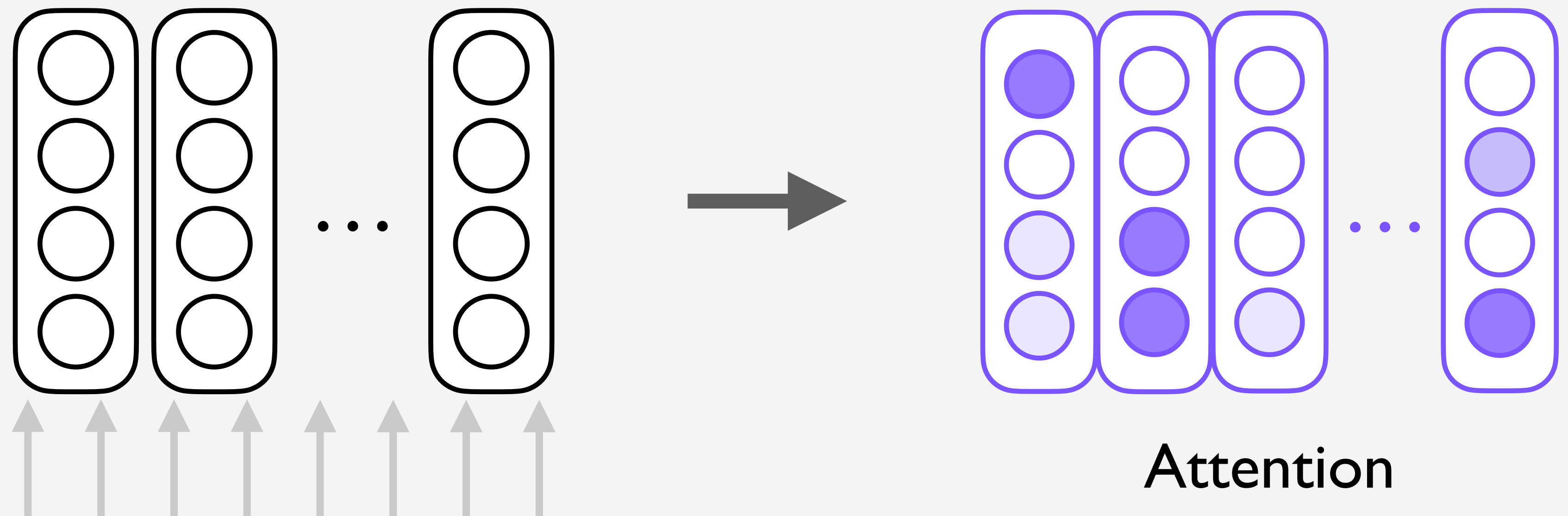


Multi-source model for post-correction

Italian first pass OCR

possiede la gallinaia

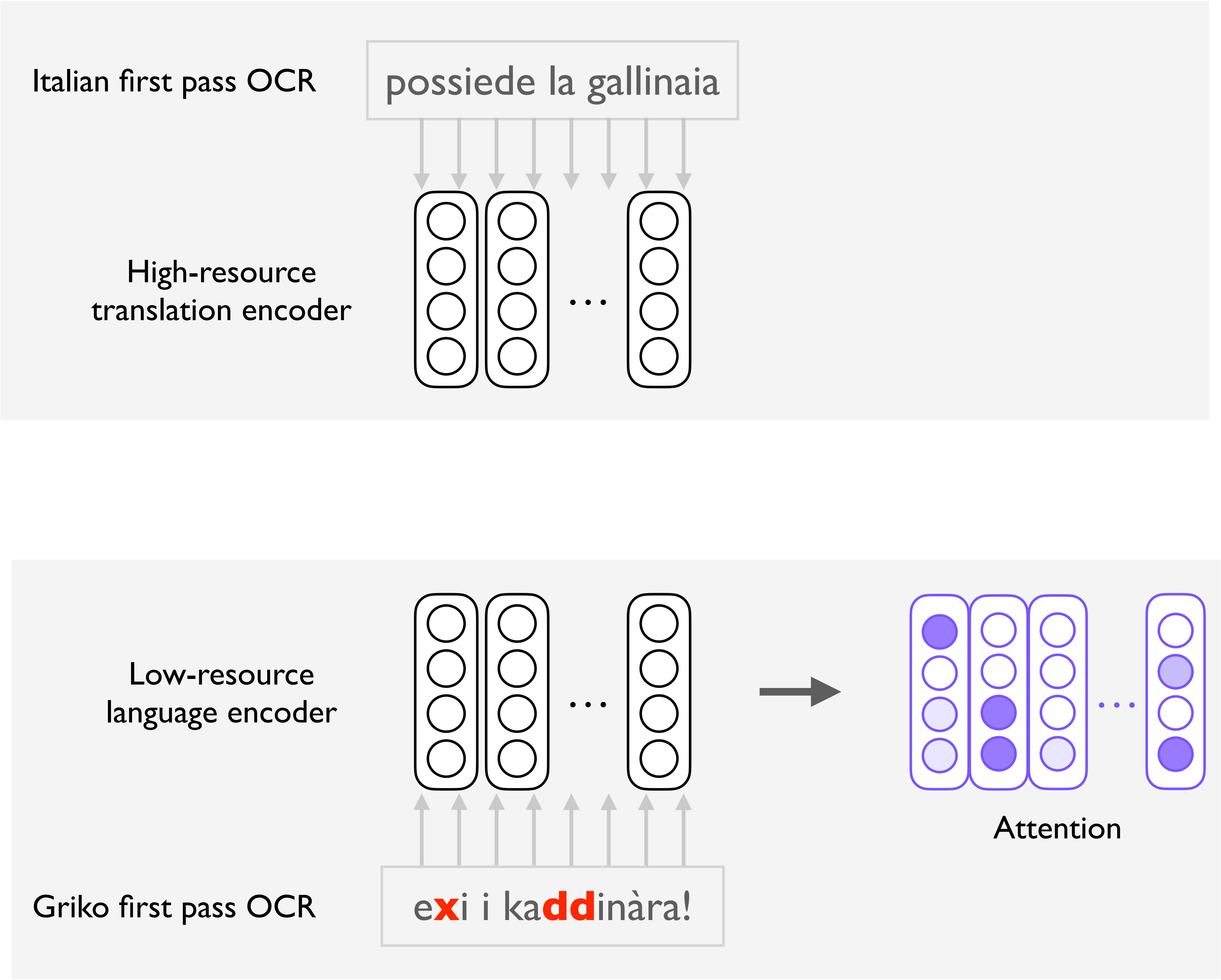
Low-resource language encoder



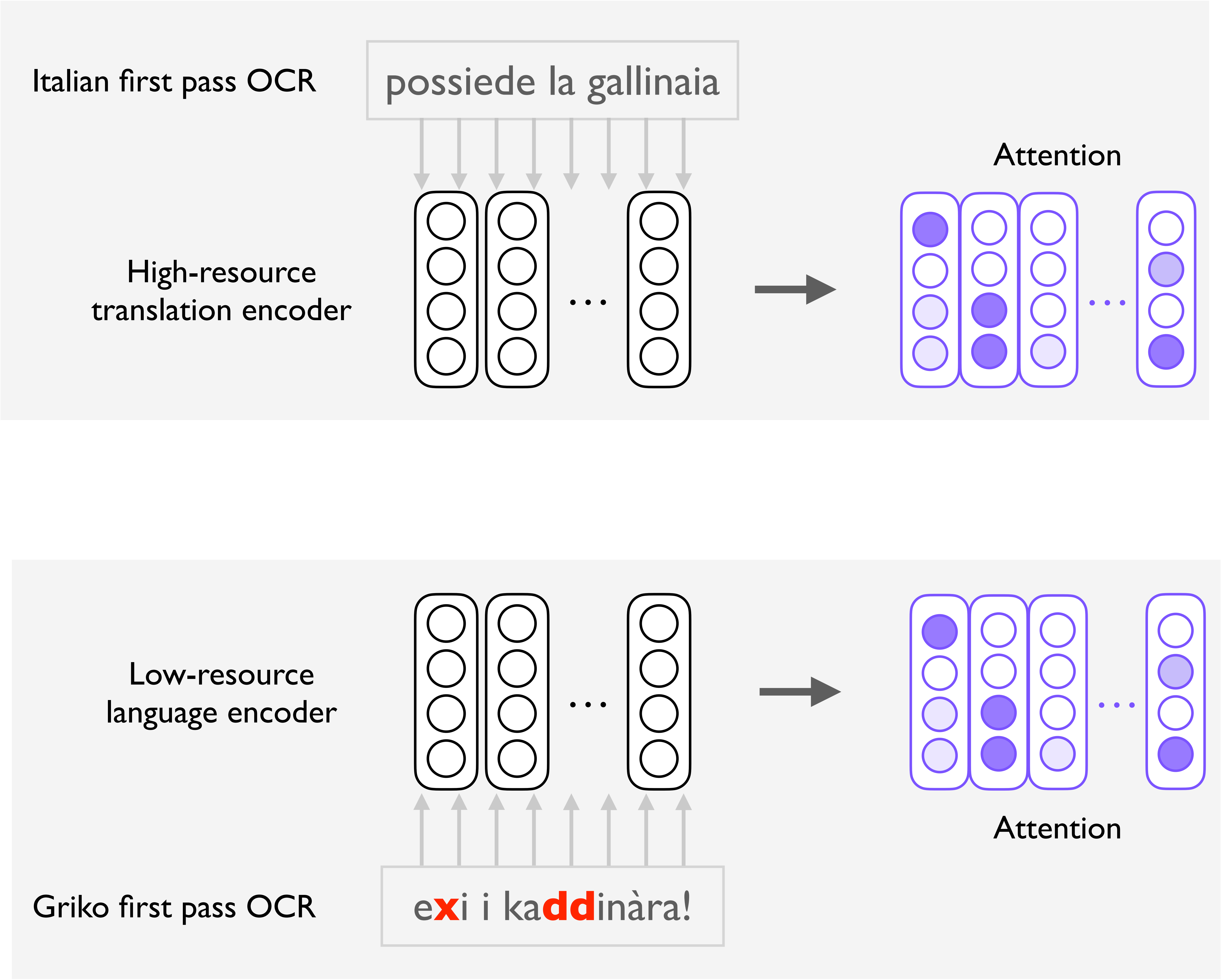
Griko first pass OCR

e**x**i i ka**dd**inàra!

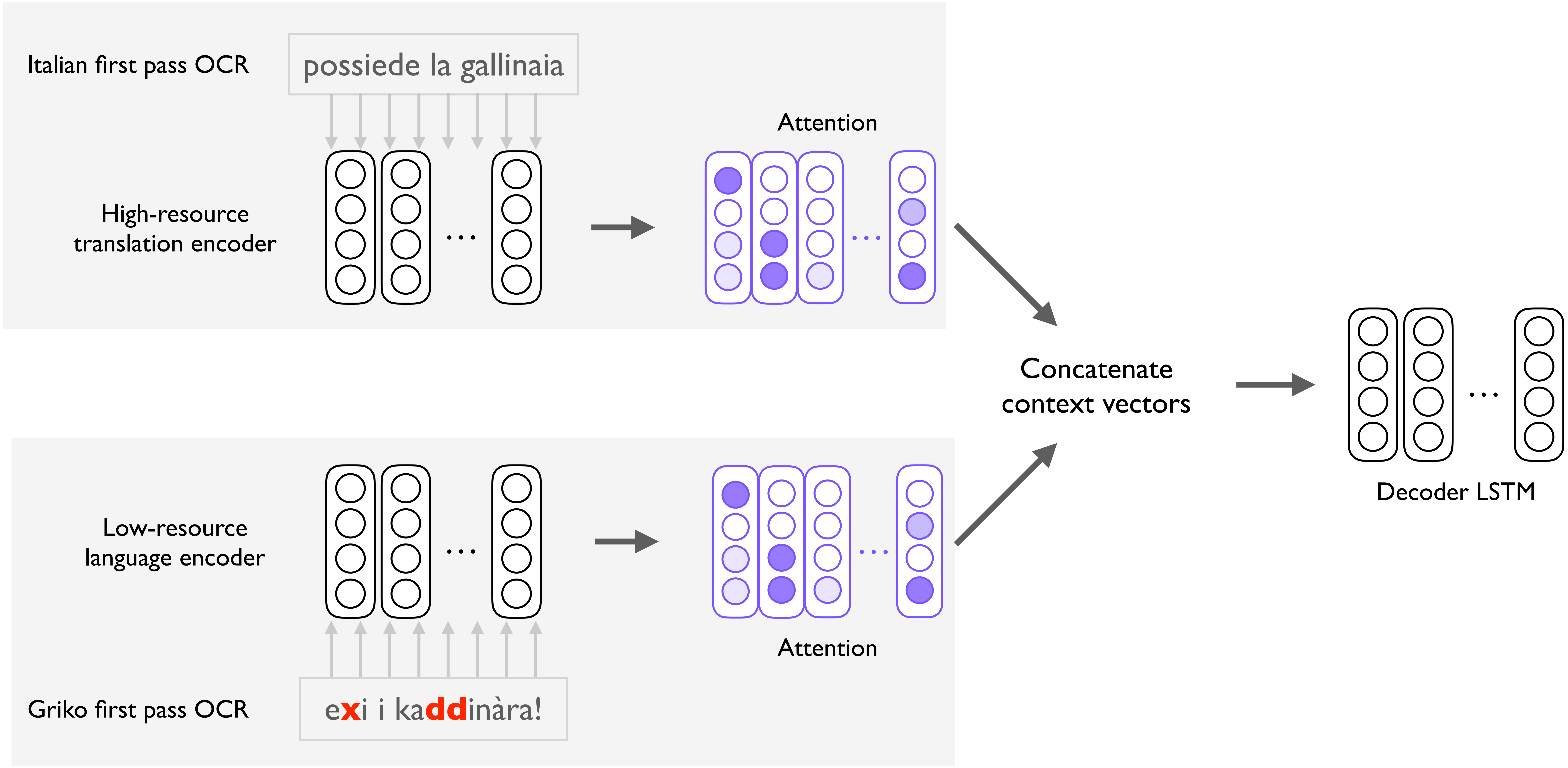
Multi-source model for post-correction



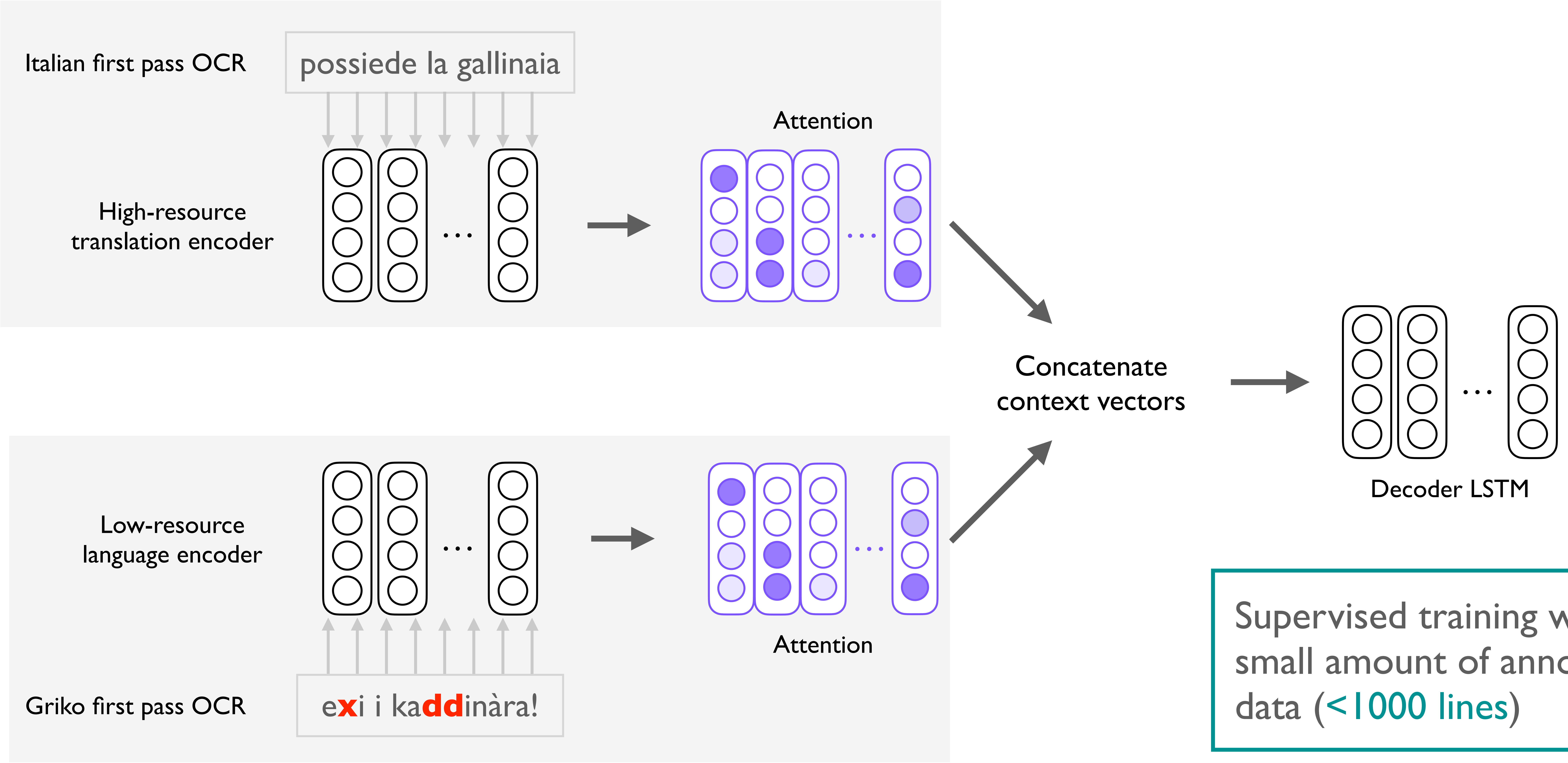
Multi-source model for post-correction



Multi-source model for post-correction



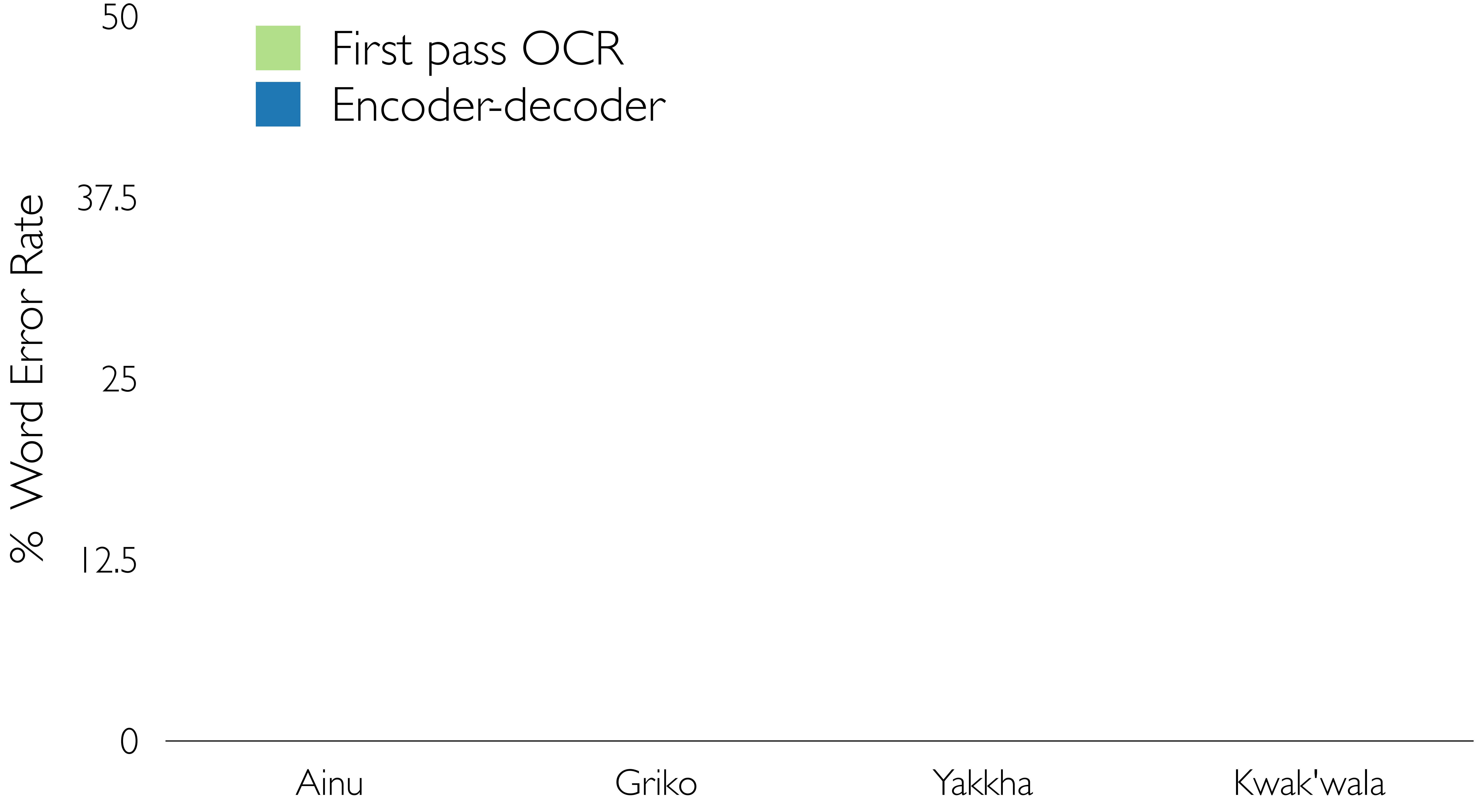
Multi-source model for post-correction



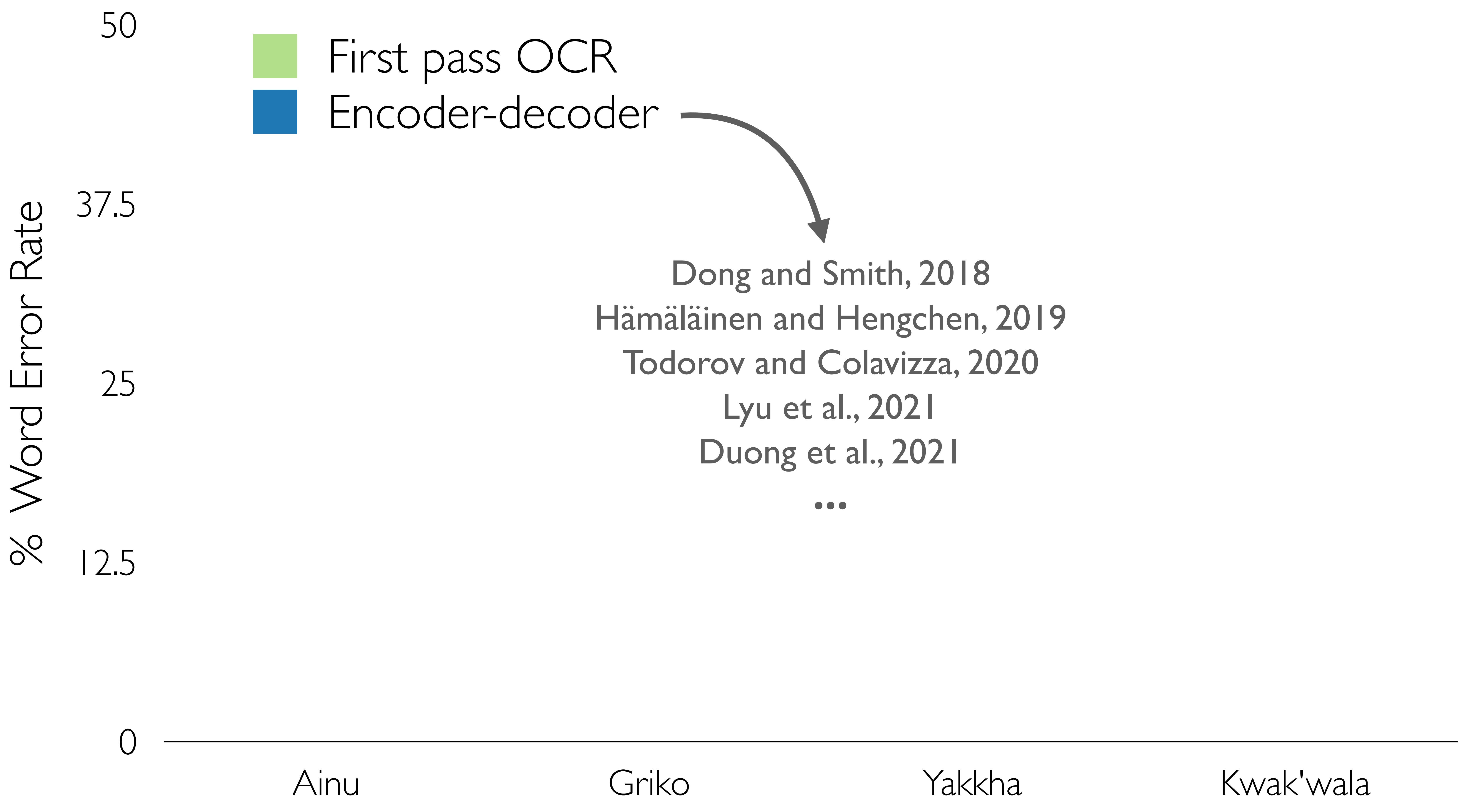
Supervised training with a small amount of annotated data (<1000 lines)

Experiments: how do existing post-correction methods perform?

Experiments: how do existing post-correction methods perform?

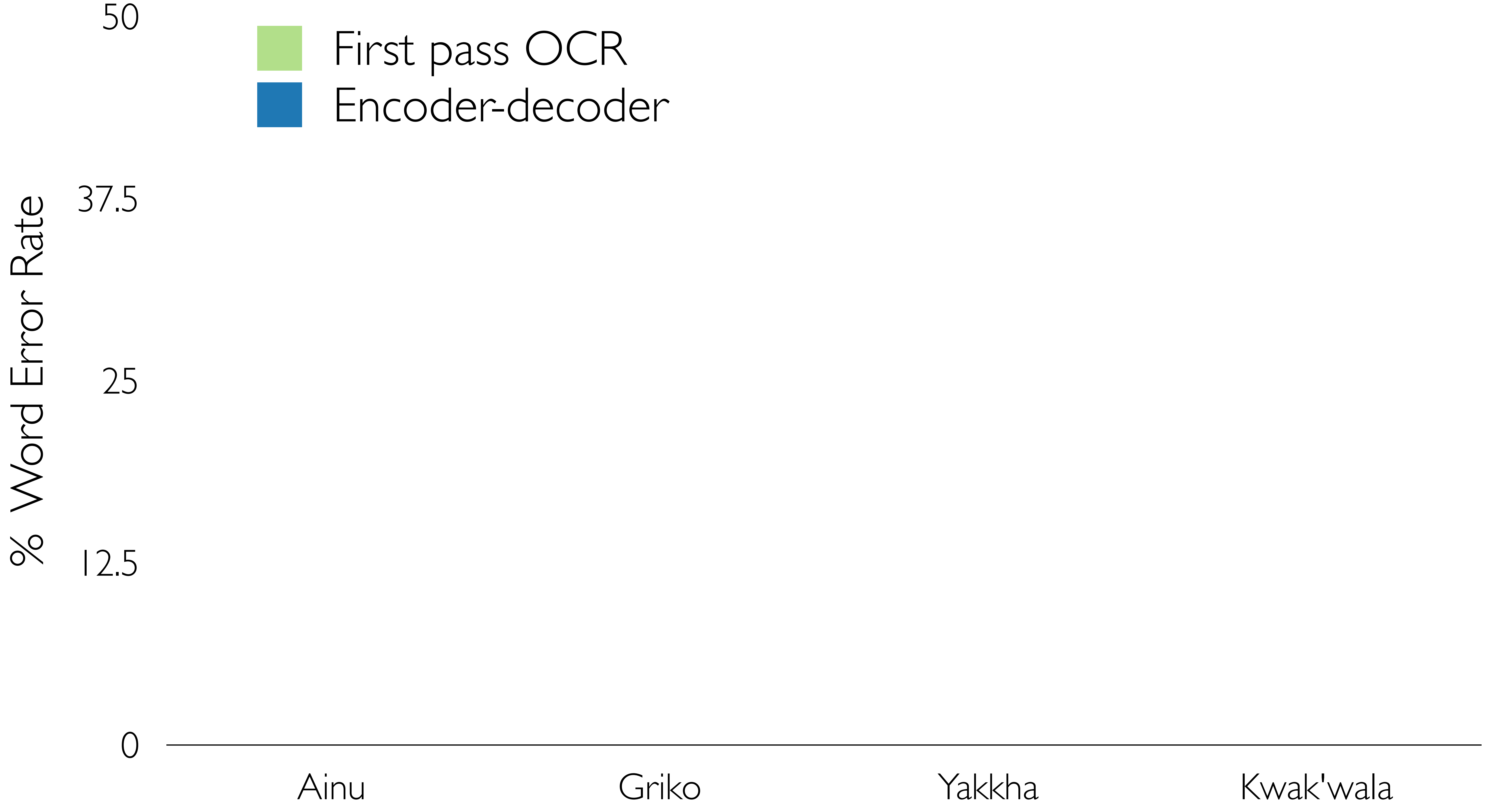


Experiments: how do existing post-correction methods perform?



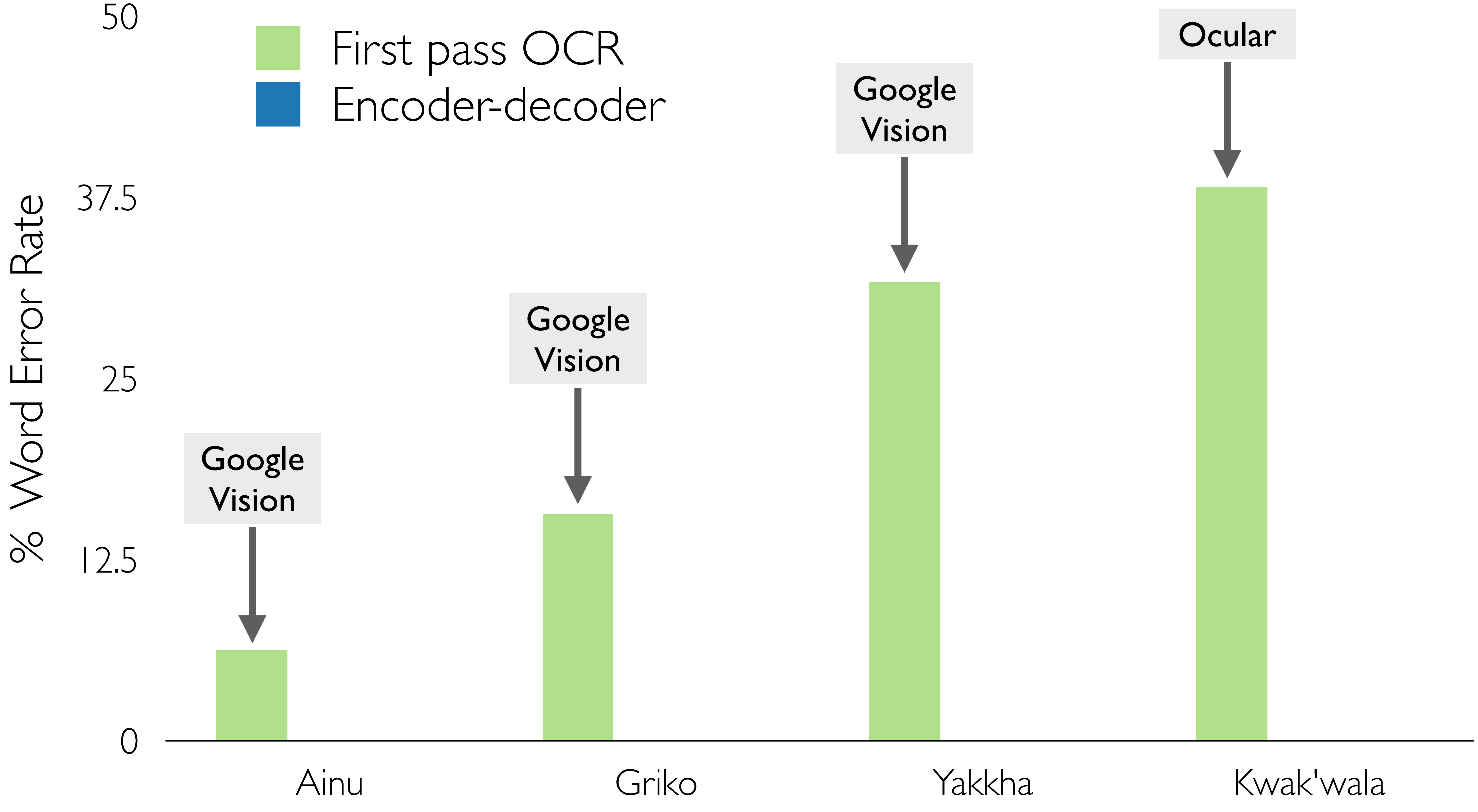
Experiments: how do existing post-correction methods perform?

Cross validation



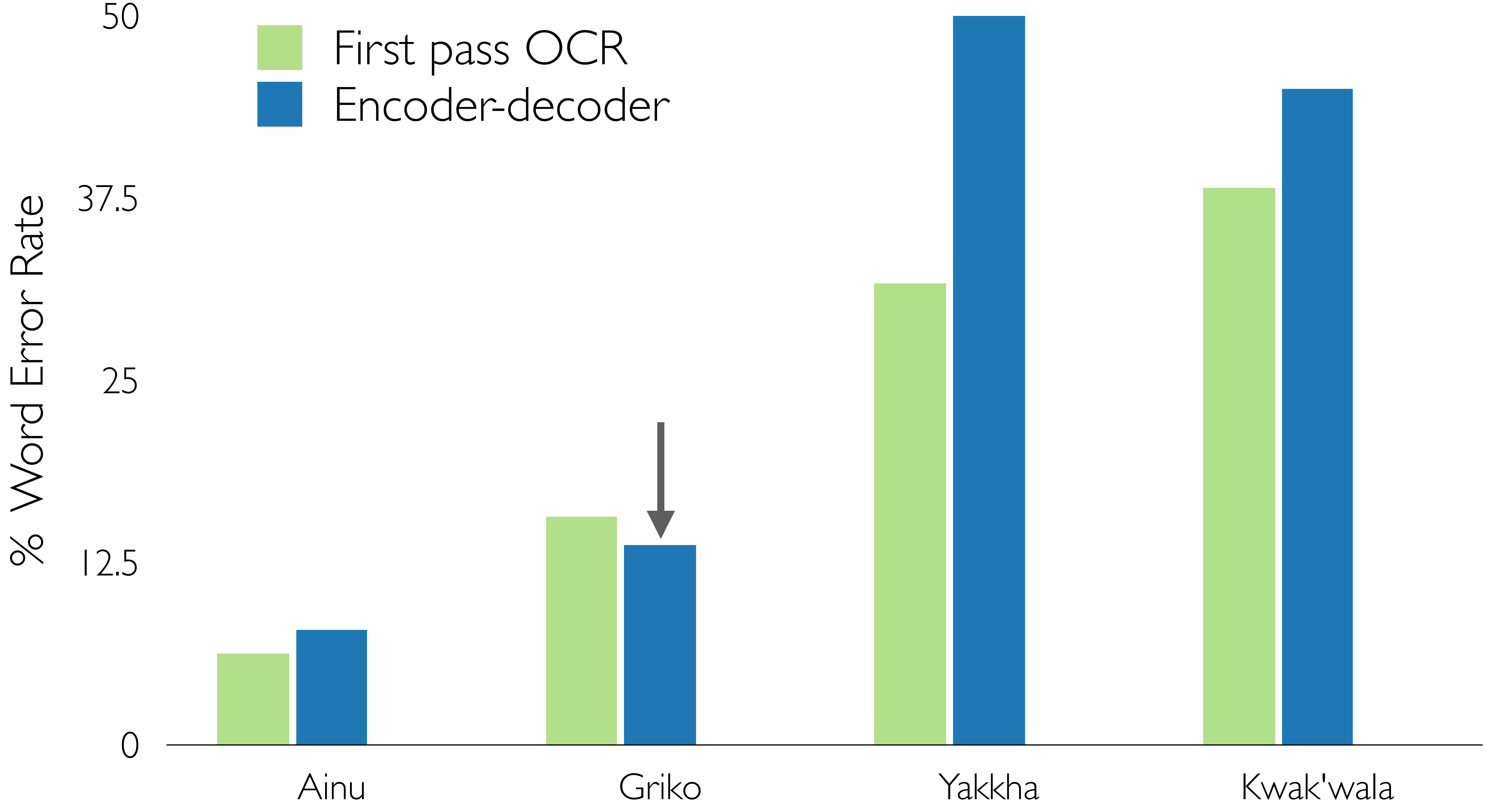
Experiments: how do existing post-correction methods perform?

Lower is better!

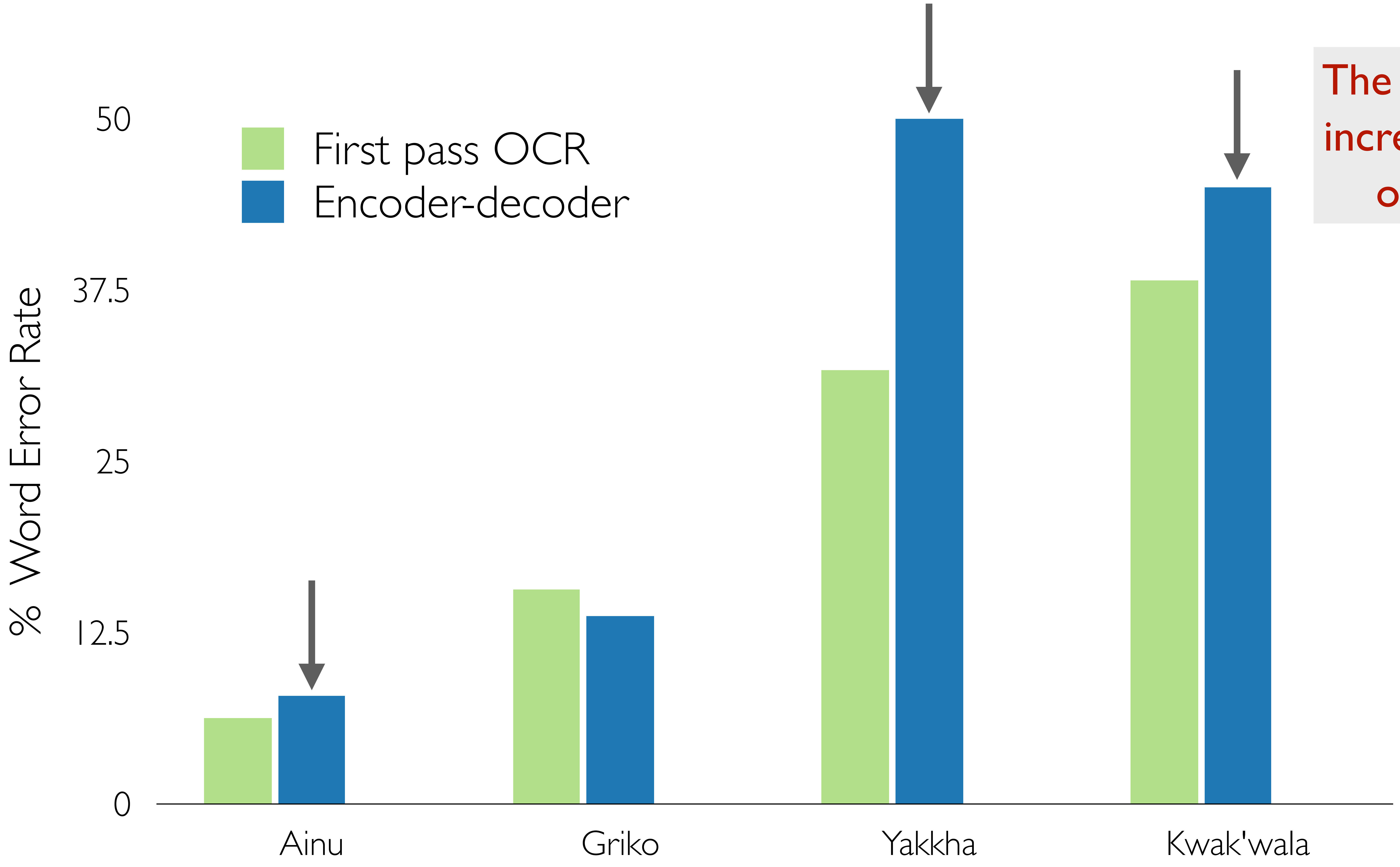


Experiments: how do existing post-correction methods perform?

Lower is better!



Experiments: how do existing post-correction methods perform?

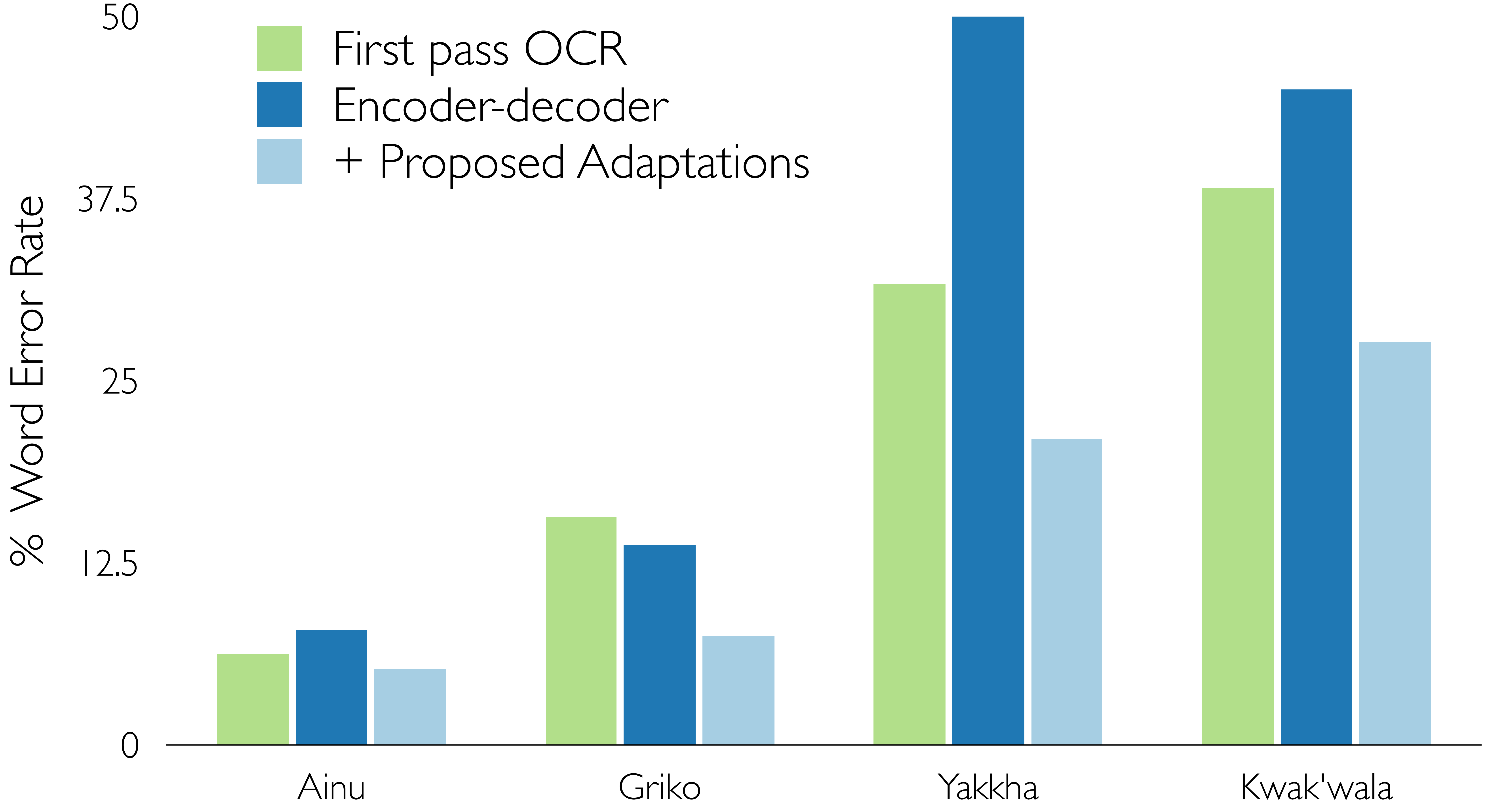


Lower is better!

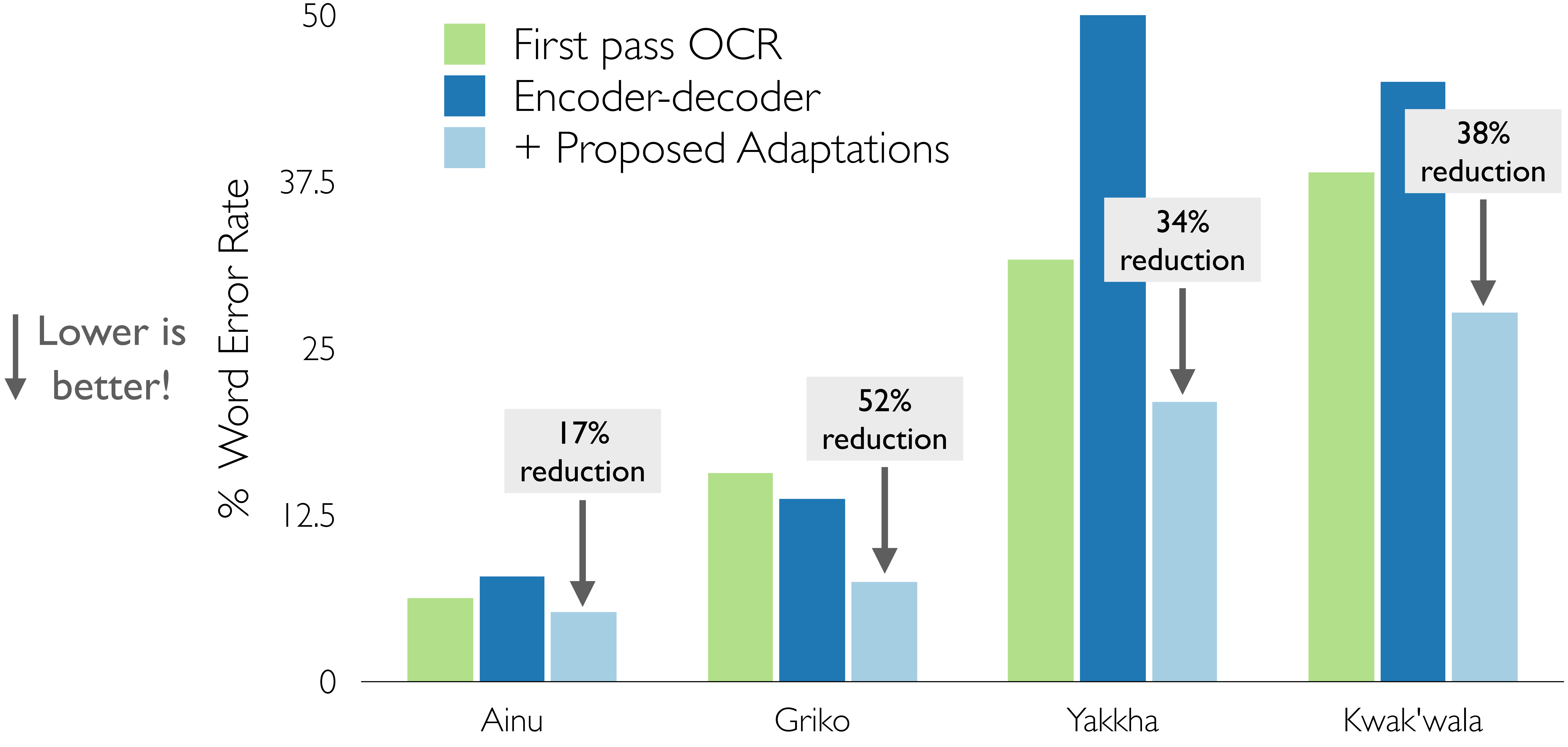
The baseline sometimes increases the error rate over the first pass

Experiments: do the adaptions help low-resource learning?

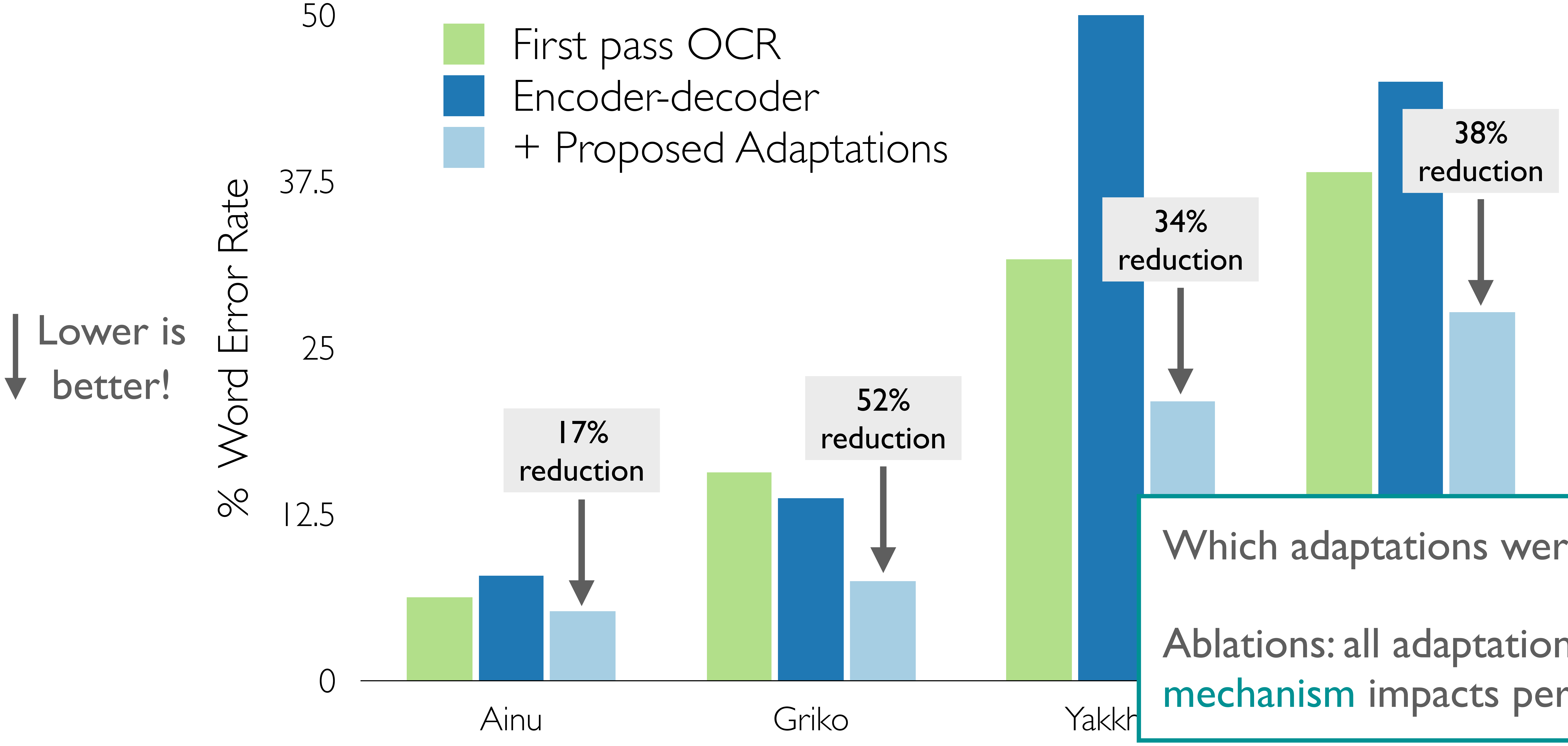
Lower is better!



Experiments: do the adaptions help low-resource learning?



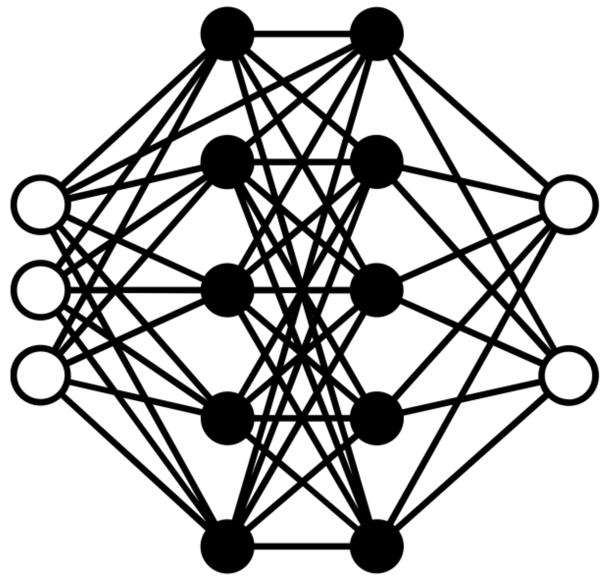
Experiments: do the adaptions help low-resource learning?



Which adaptations were most helpful?

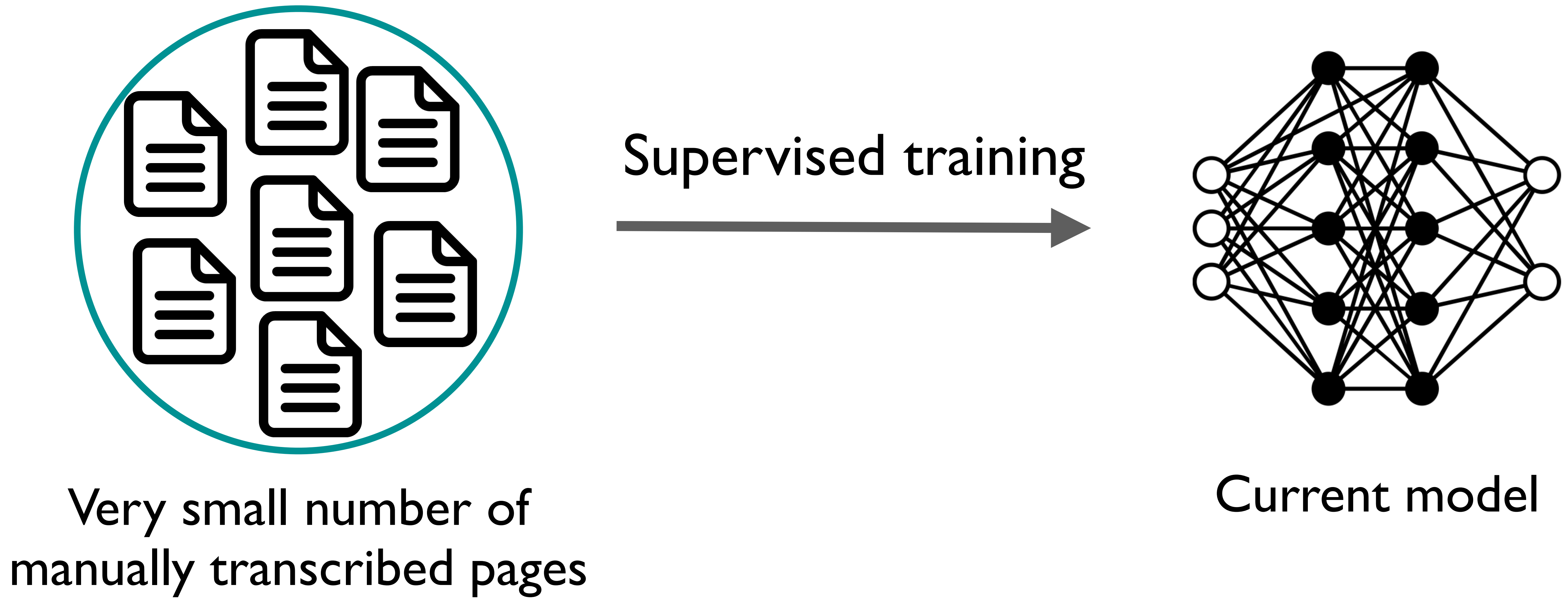
Ablations: all adaptations are useful, copy mechanism impacts performance the most

Improving performance without additional annotation



Current model

Improving performance without additional annotation

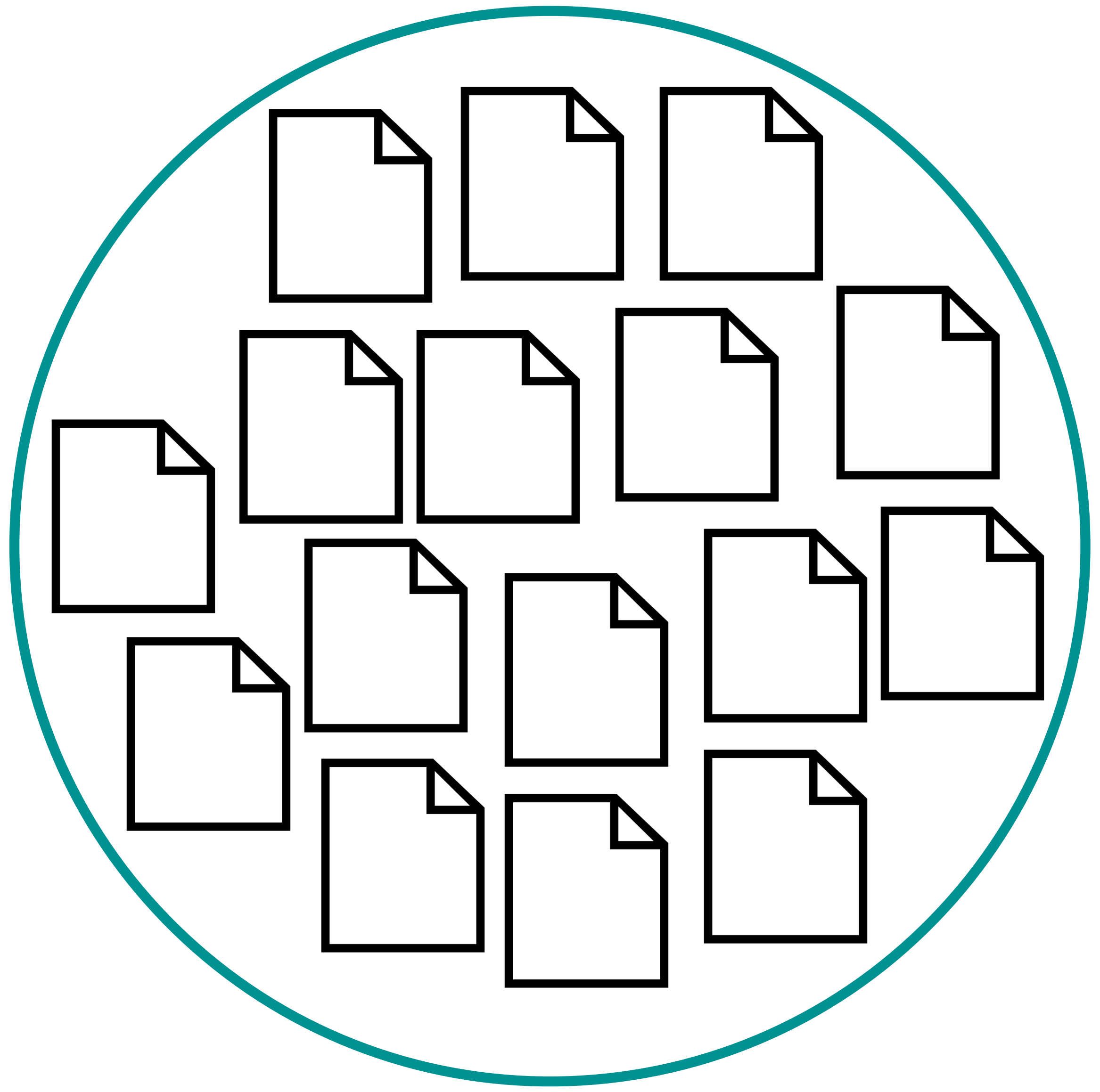


Improving performance without additional annotation

Very small number of
manually transcribed pages

Improving performance without additional annotation

Very small number of
manually transcribed pages

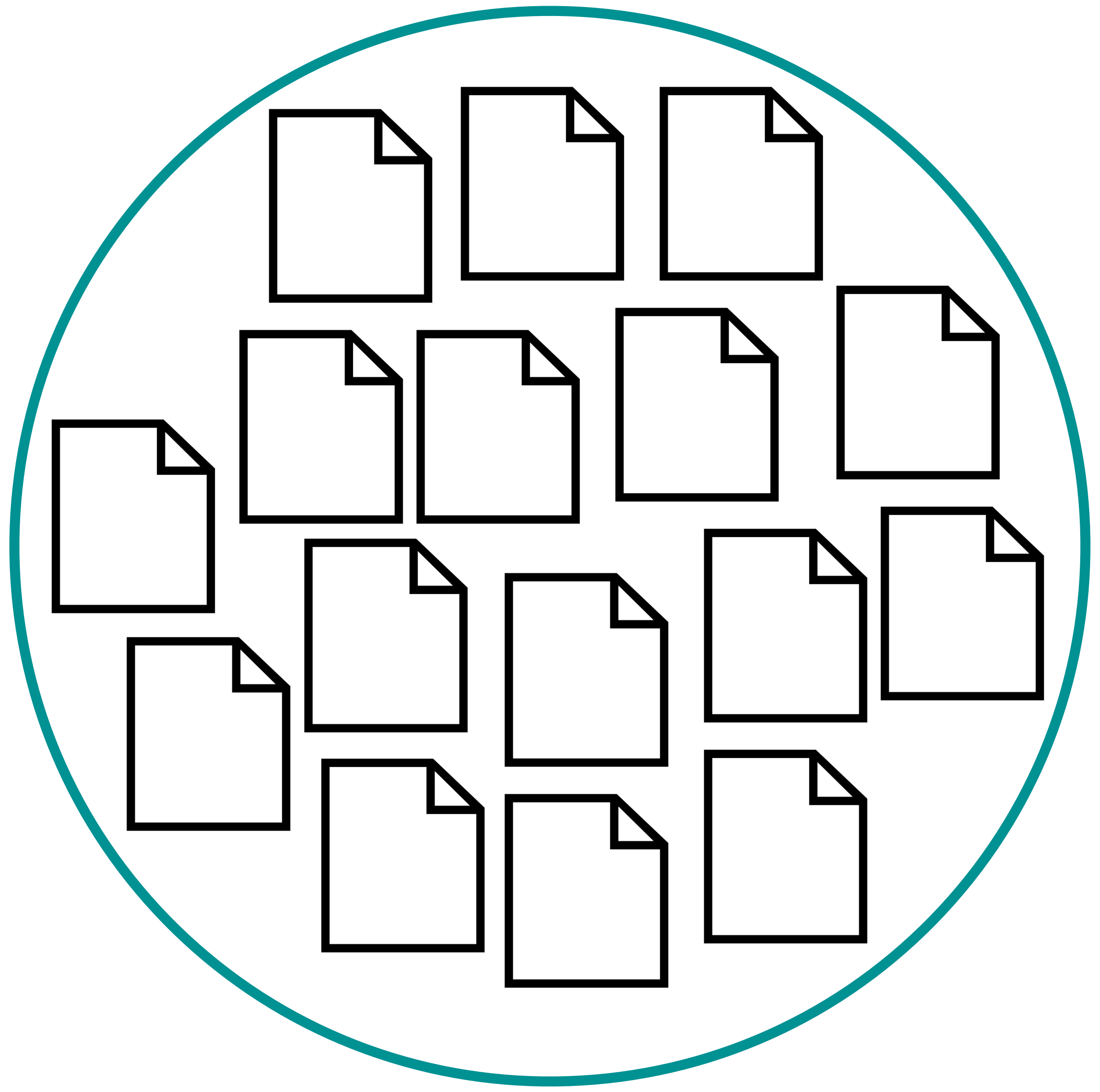


Relatively larger number
of raw images that need
to be digitized

Rijhwani, Rosenblum, Anastasopoulos, Neubig. TACL 2021.

Improving performance without additional annotation

Very small number of
manually transcribed pages



Relatively larger number
of raw images that need
to be digitized

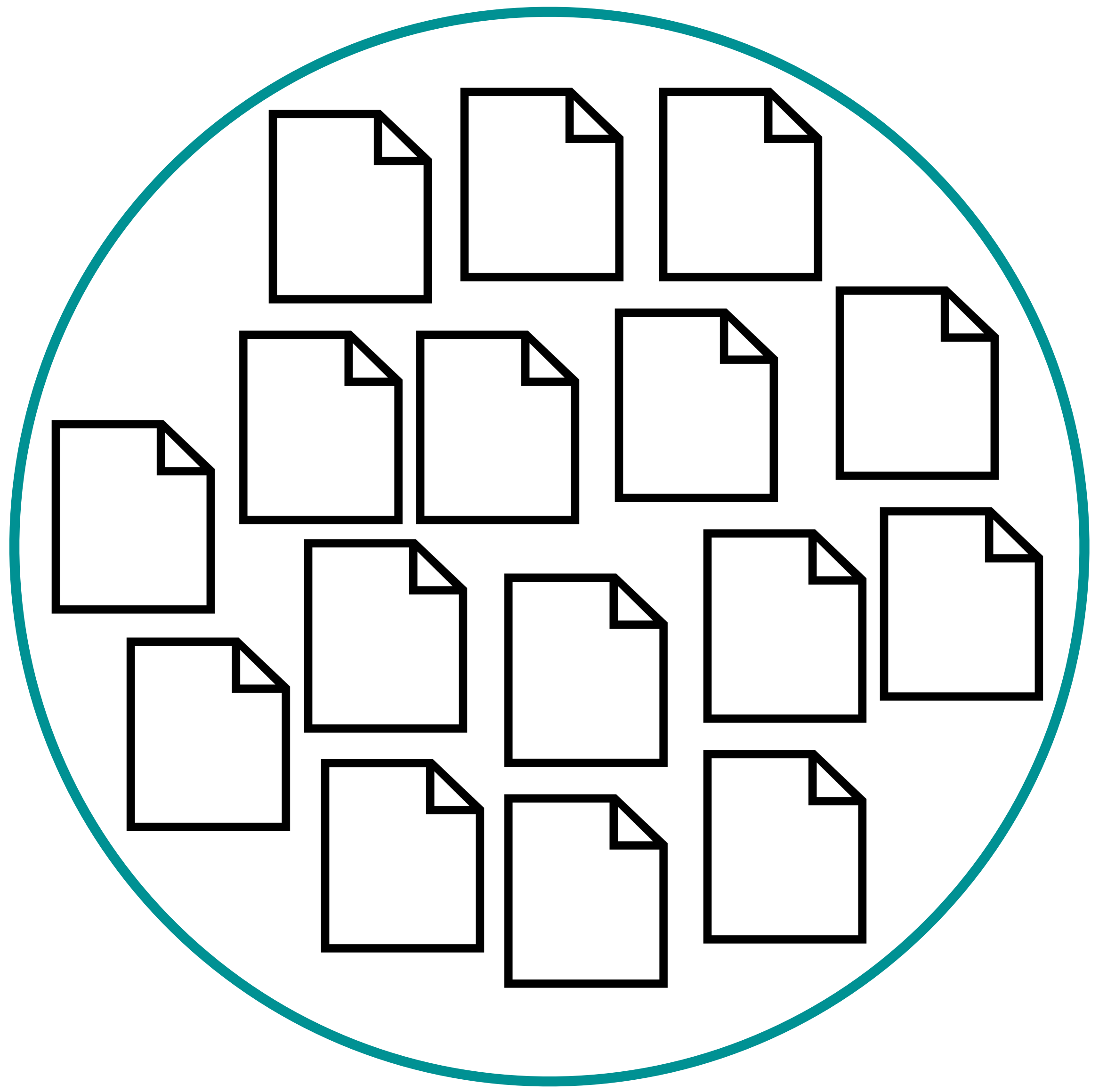
Our dataset:

Documents contain
300 – 800 pages

Only ~30 are
manually transcribed

Improving performance without additional annotation

Very small number of
manually transcribed pages



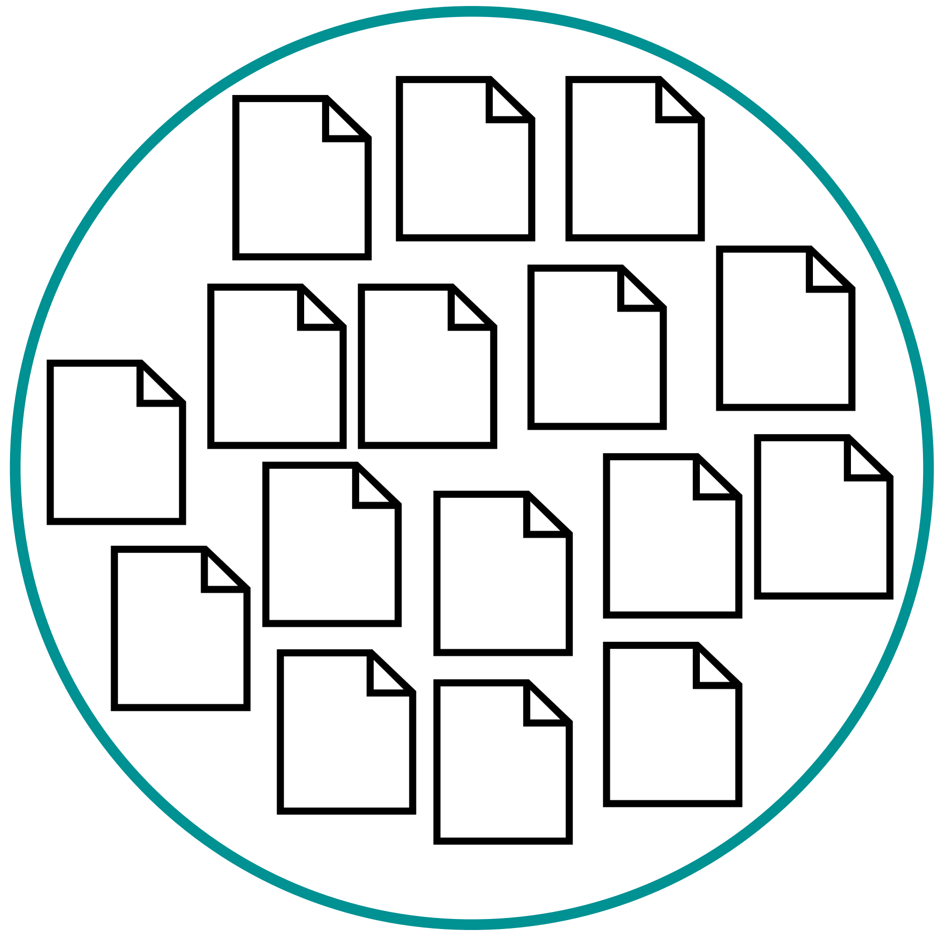
Relatively larger number
of raw images that need
to be digitized

Semi-supervised learning
for efficient use of the
unlabeled images

Rijhwani, Rosenblum, Anastasopoulos, Neubig. TACL 2021.

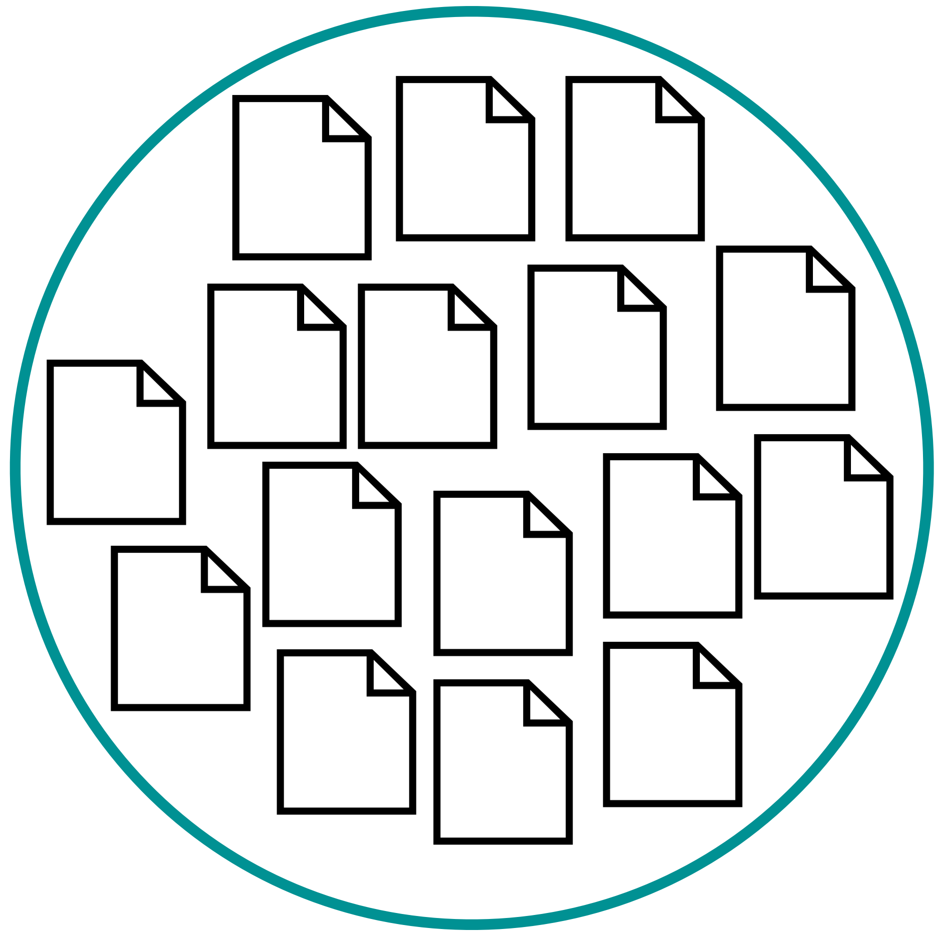
Self-training for OCR post-correction

Self-training for OCR post-correction

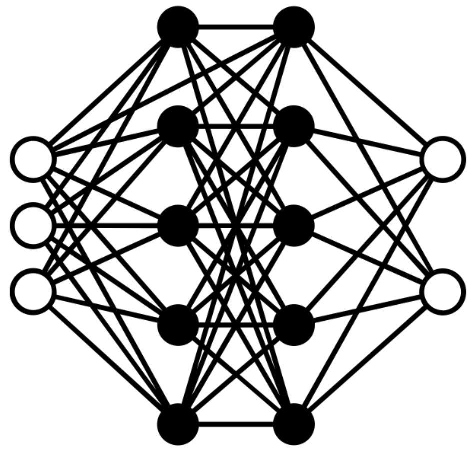


First pass OCR on unlabeled images

Self-training for OCR post-correction

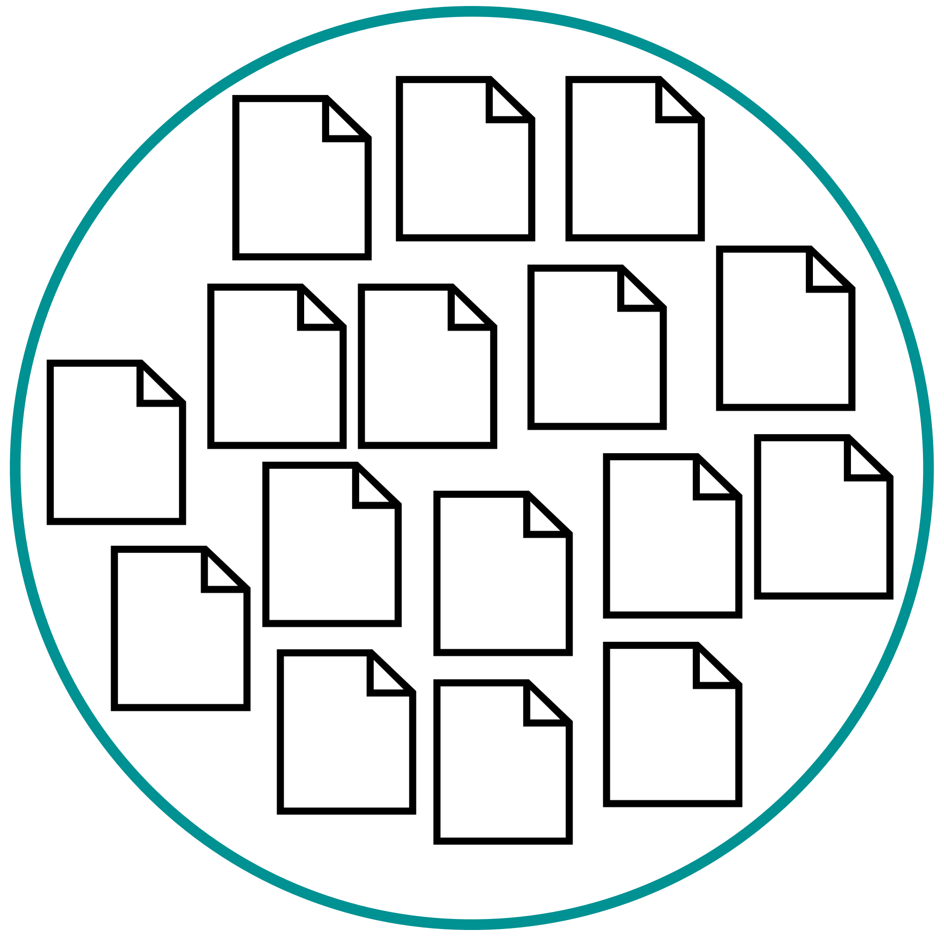


First pass OCR on
unlabeled images

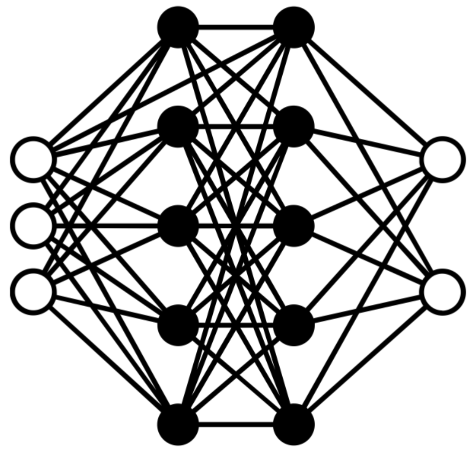


Trained post-
correction model

Self-training for OCR post-correction



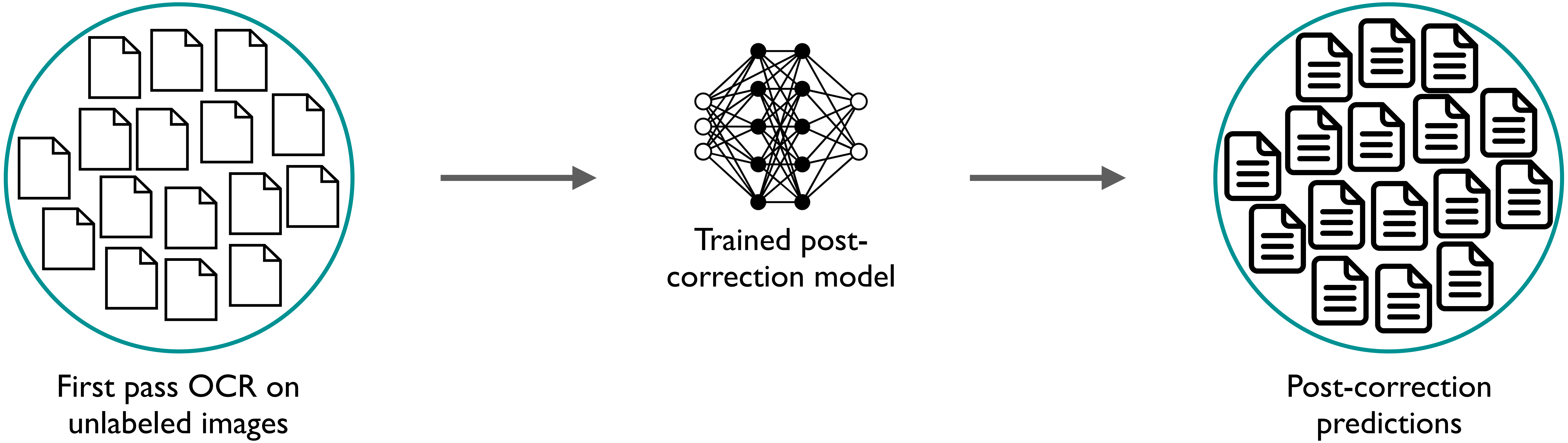
First pass OCR on
unlabeled images



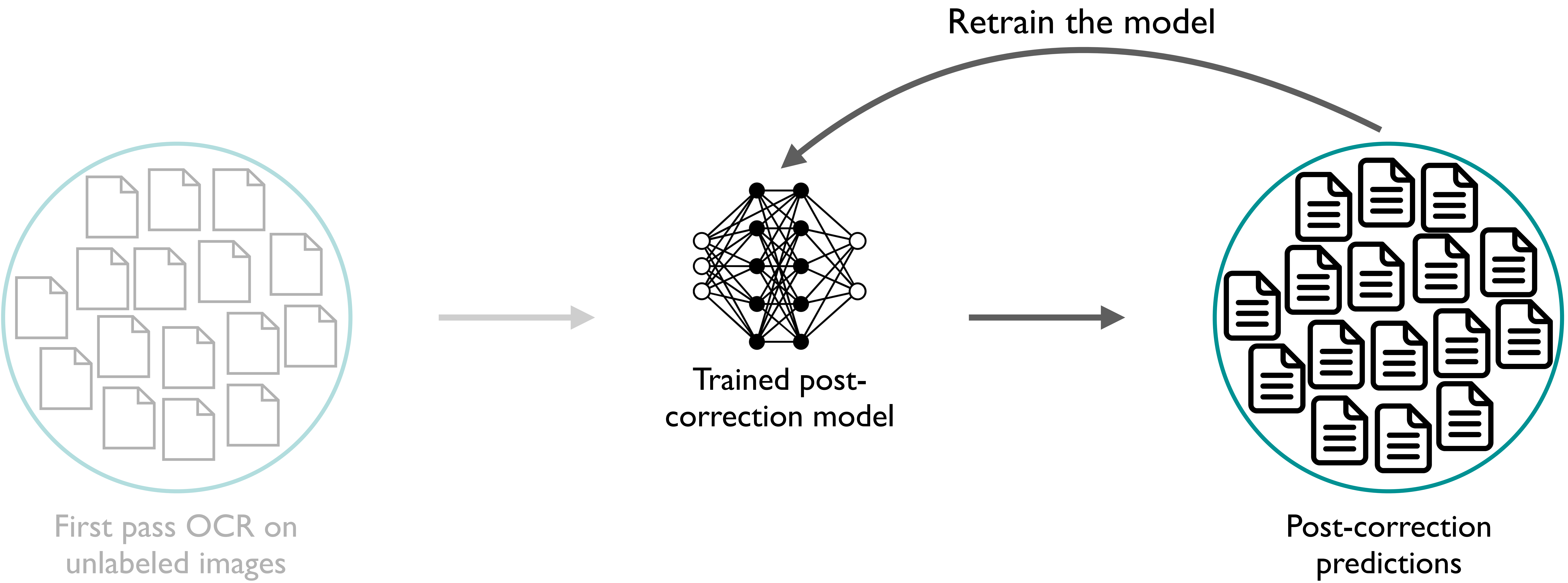
Trained post-
correction model

Previous best
supervised model

Self-training for OCR post-correction



Self-training for OCR post-correction



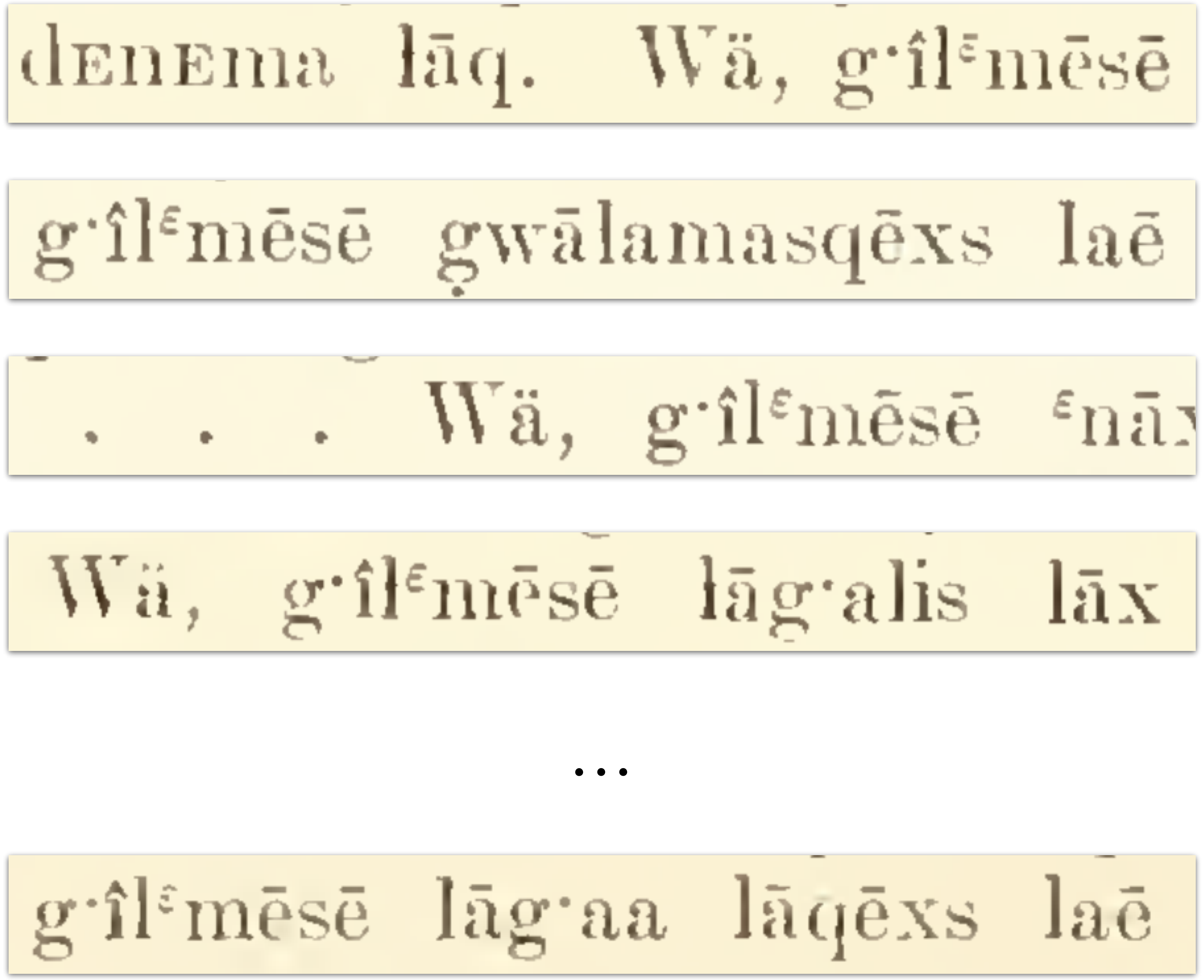
Self-training may introduce noise

Self-training may introduce noise

Can we bias post-correction towards generating correct words?

Self-training may introduce noise

Can we bias post-correction towards generating correct words?


dᴇnᴇma łāq. Wä, g·îlᵋmēsē

g·îlᵋmēsē gwālamasqēxs laē

. . . Wä, g·îlᵋmēsē ᵋnā

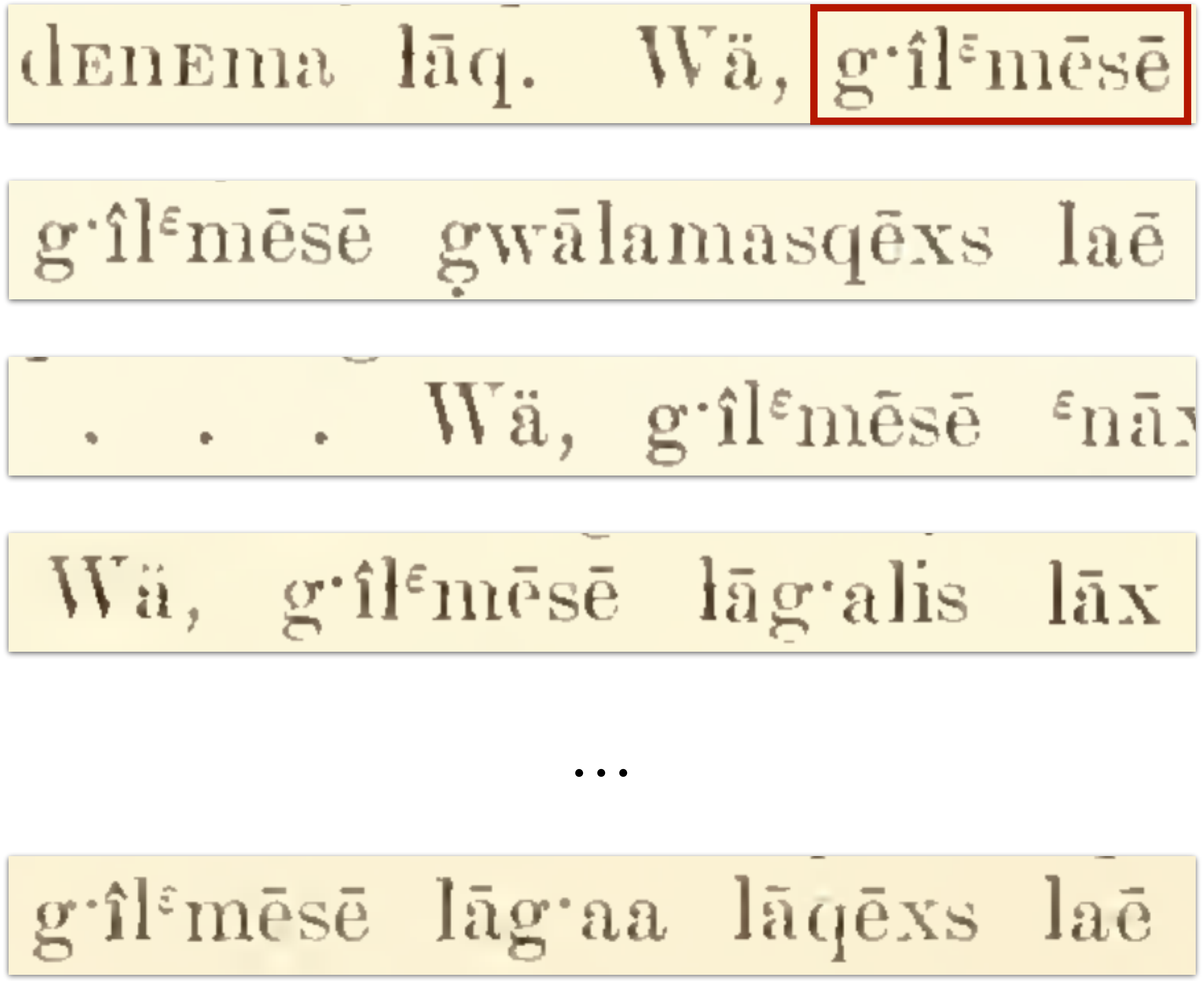
Wä, g·îlᵋmēsē łāg·alis lāx

…

g·îlᵋmēsē lāg·aa lāqēxs laē

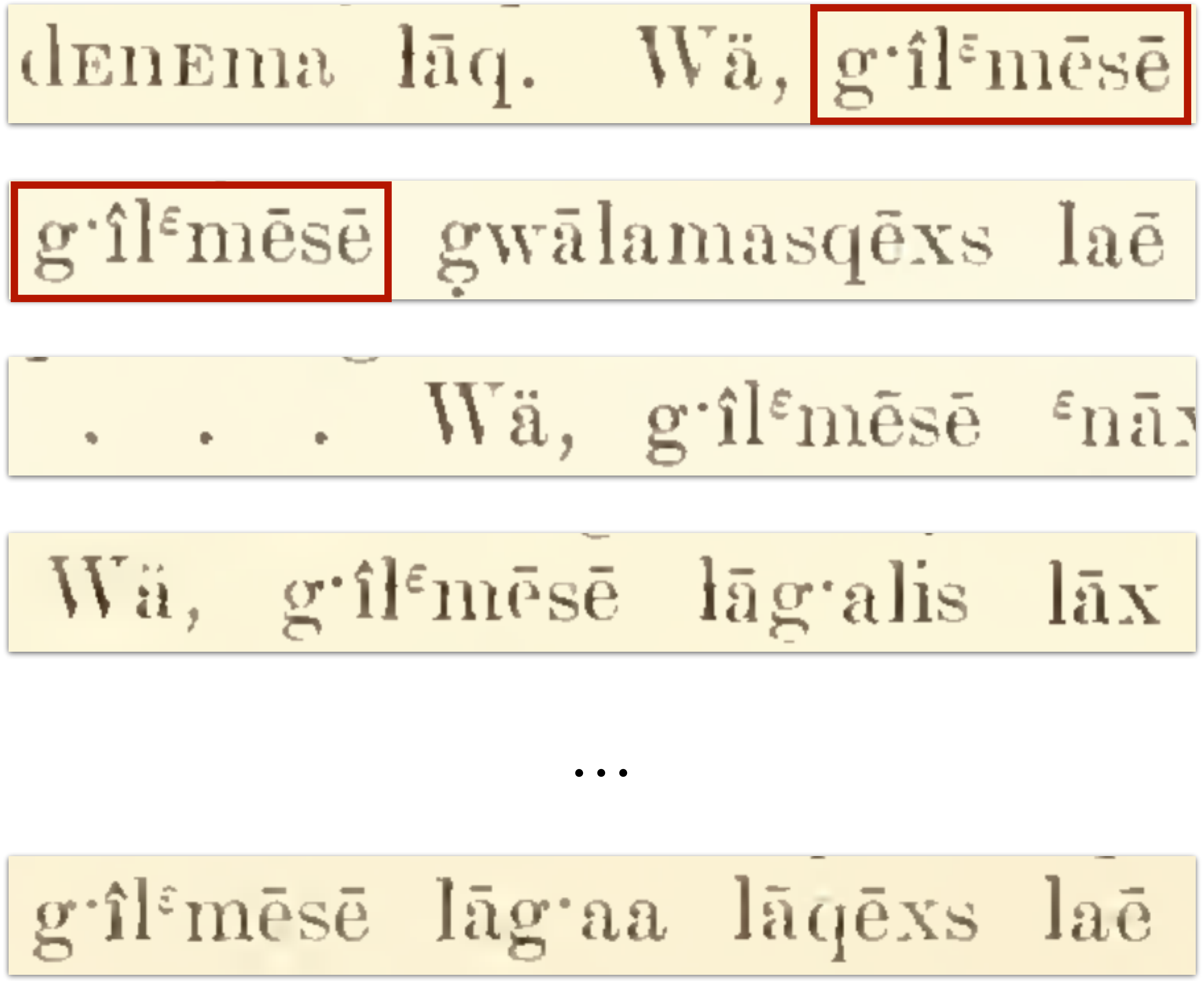
Self-training may introduce noise

Can we bias post-correction towards generating correct words?


dEnEma lāq. Wä, g·îlᵋmēsē
g·îlᵋmēsē gwālamasqēxs laē
. . . Wä, g·îlᵋmēsē ᵋnā
Wä, g·îlᵋmēsē lāg·alis lāx
…
g·îlᵋmēsē lāg·aa lāqēxs laē

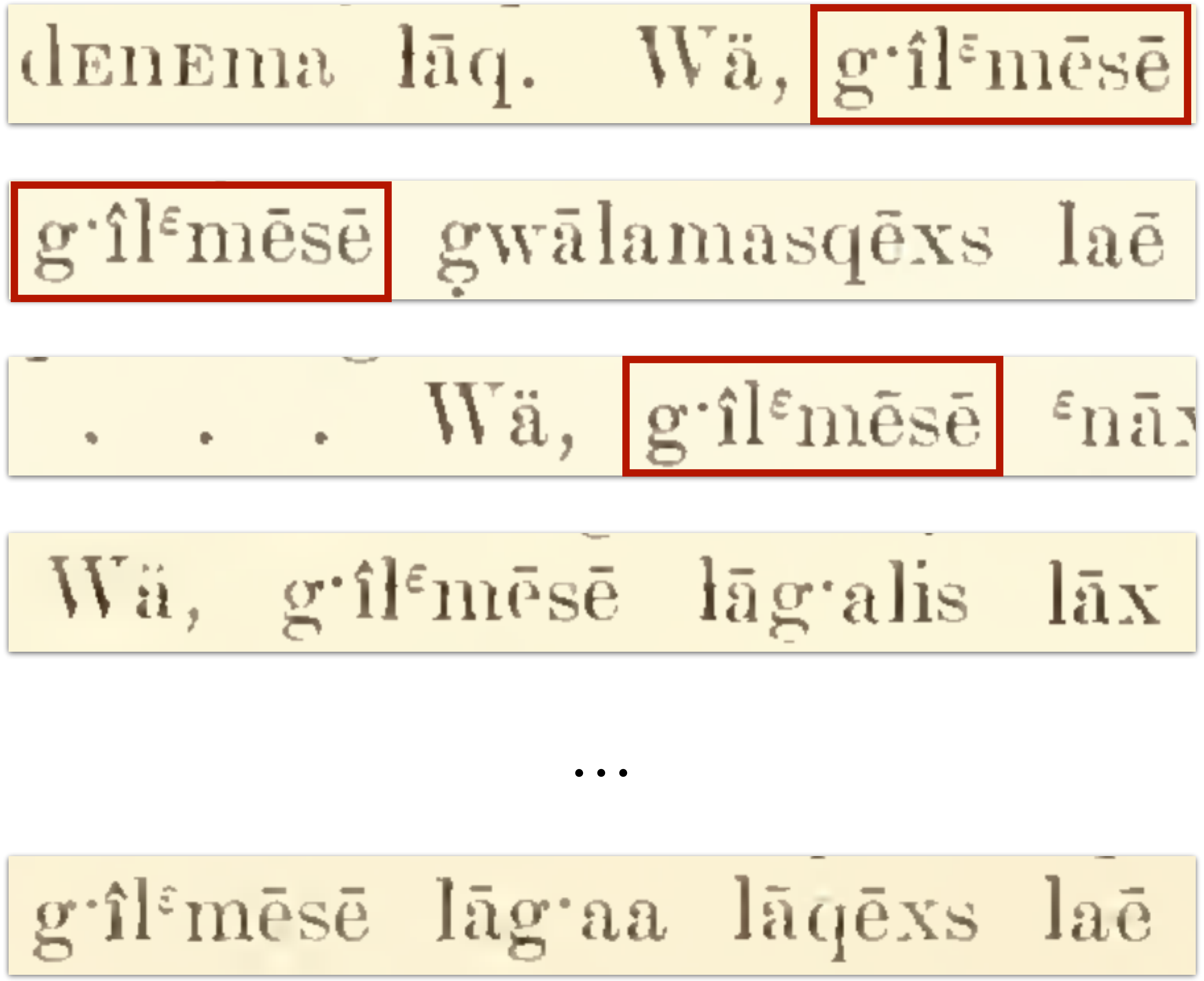
Self-training may introduce noise

Can we bias post-correction towards generating correct words?



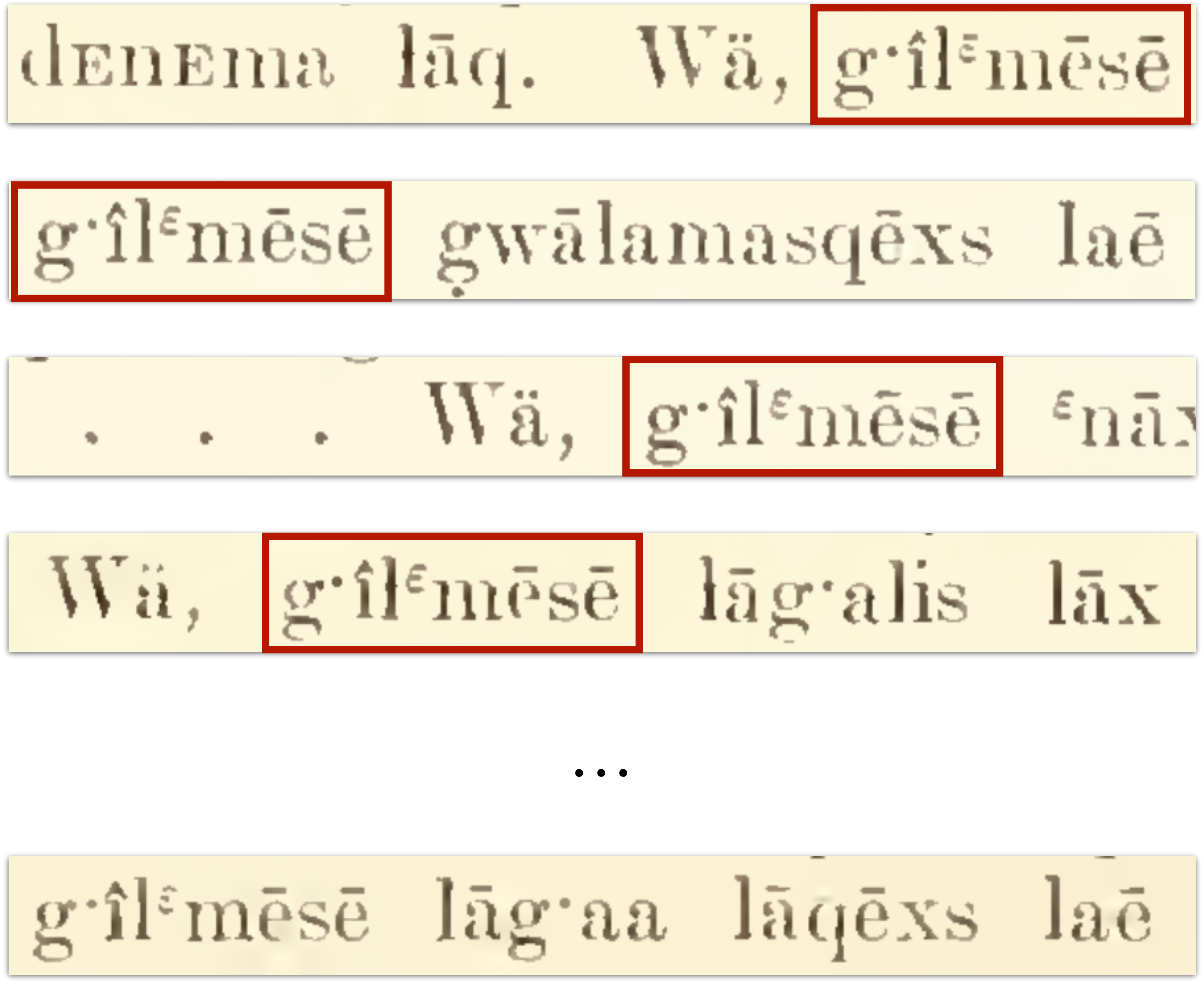
Self-training may introduce noise

Can we bias post-correction towards generating correct words?



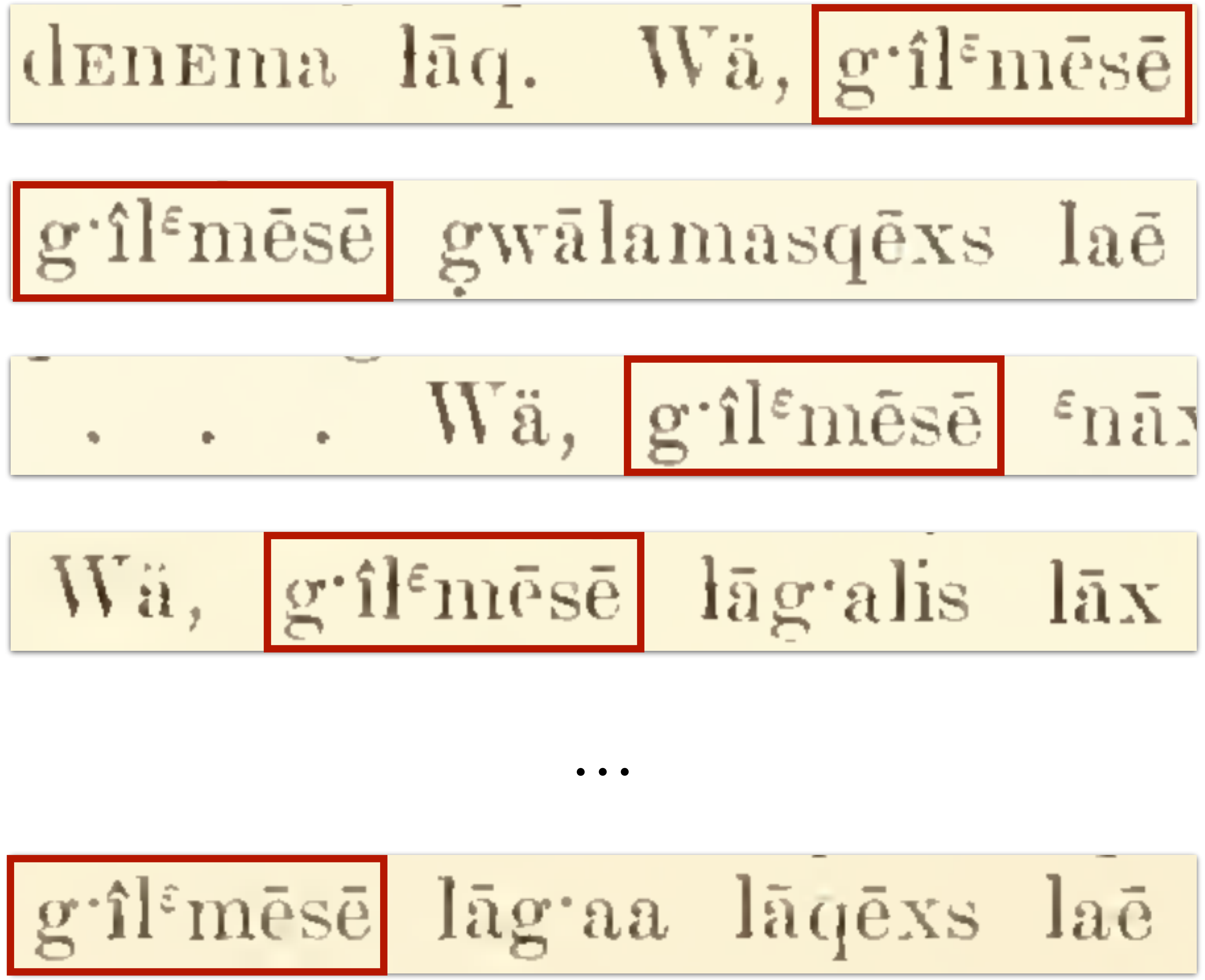
Self-training may introduce noise

Can we bias post-correction towards generating correct words?



Self-training may introduce noise

Can we bias post-correction towards generating correct words?


dᴇnᴇma lāq. Wä, g·îlᵋmēsē

g·îlᵋmēsē gwālamasqēxs laē

. . . Wä, g·îlᵋmēsē ᵋnā

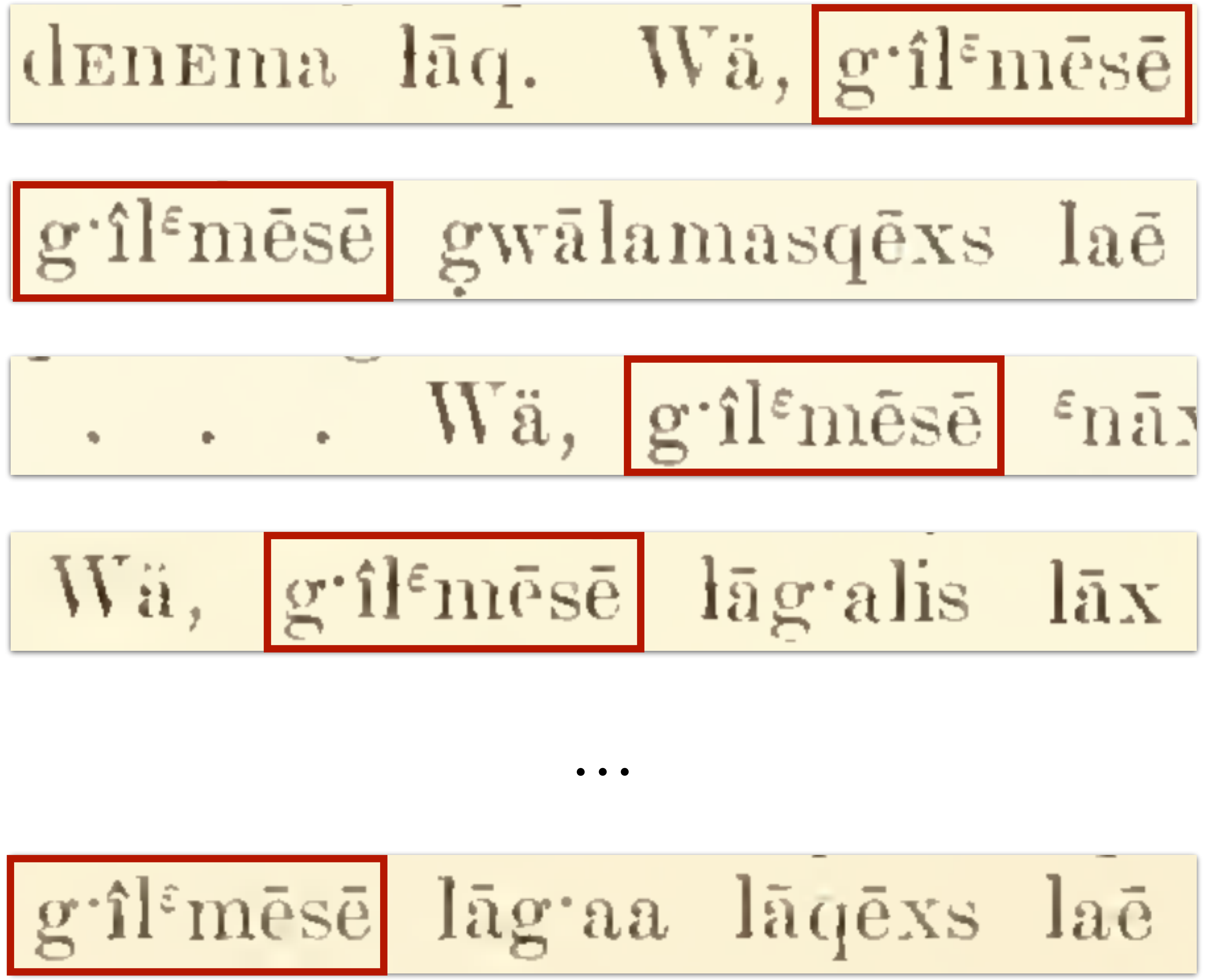
Wä, g·îlᵋmēsē lāg·alis lāx

…

g·îlᵋmēsē lāg·aa lāqēxs laē

Self-training may introduce noise

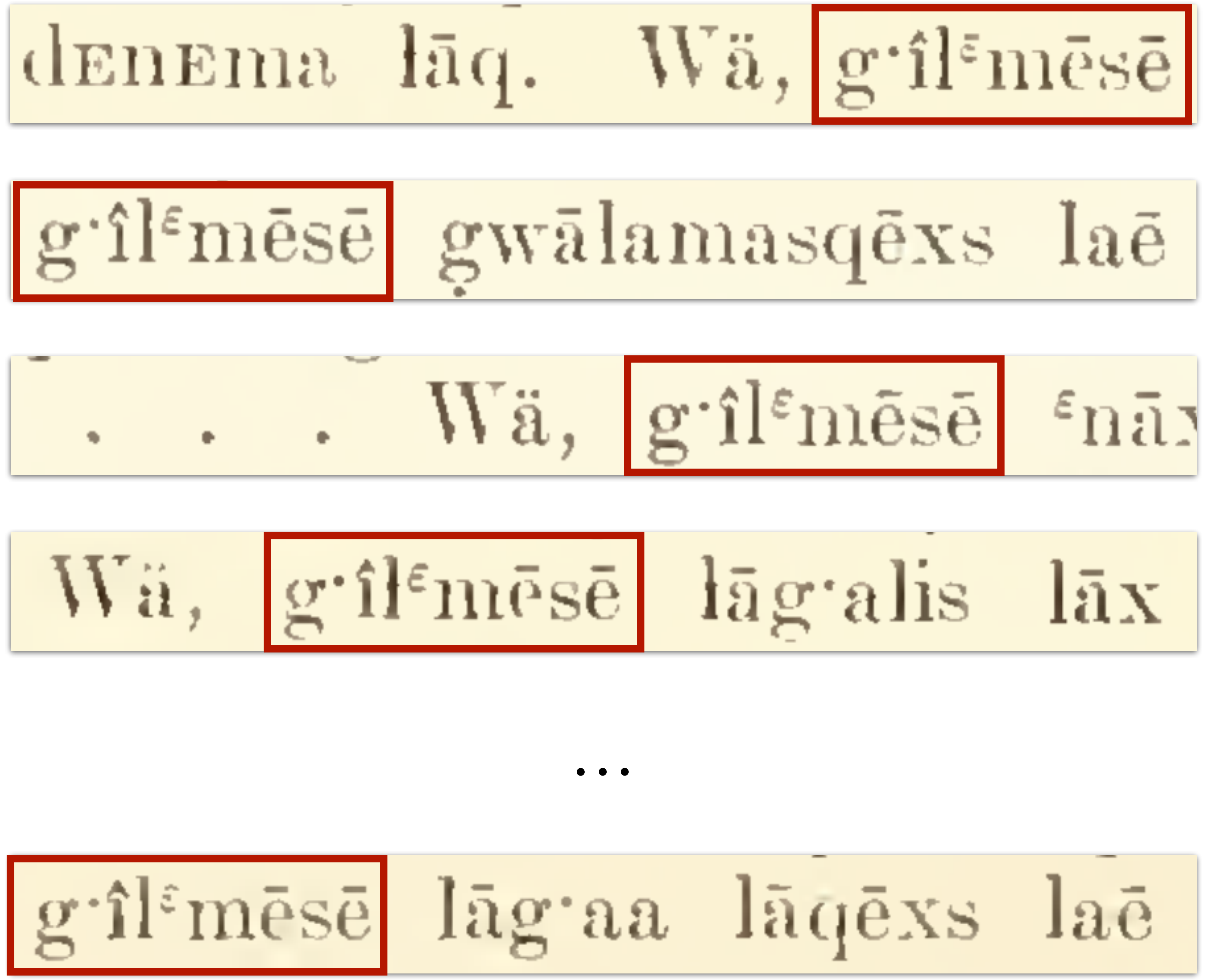
Can we bias post-correction towards generating correct words?



g·îl$^{\varepsilon}$mēsē (7)

Self-training may introduce noise

Can we bias post-correction towards generating correct words?

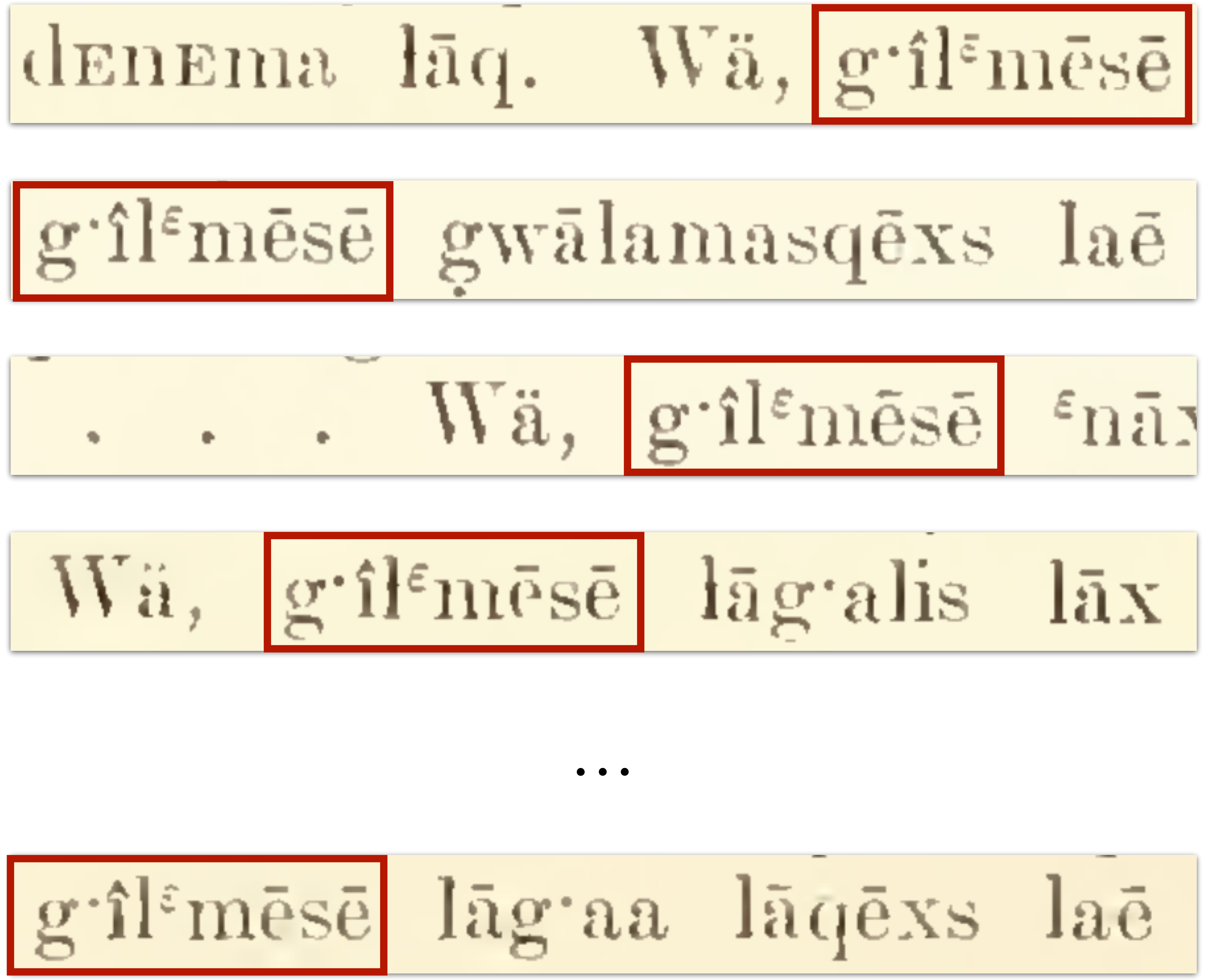


g·îl$^{\varepsilon}$mēsē (7)

g·il$^{\varepsilon}$mēsē (5)

Self-training may introduce noise

Can we bias post-correction towards generating correct words?



…

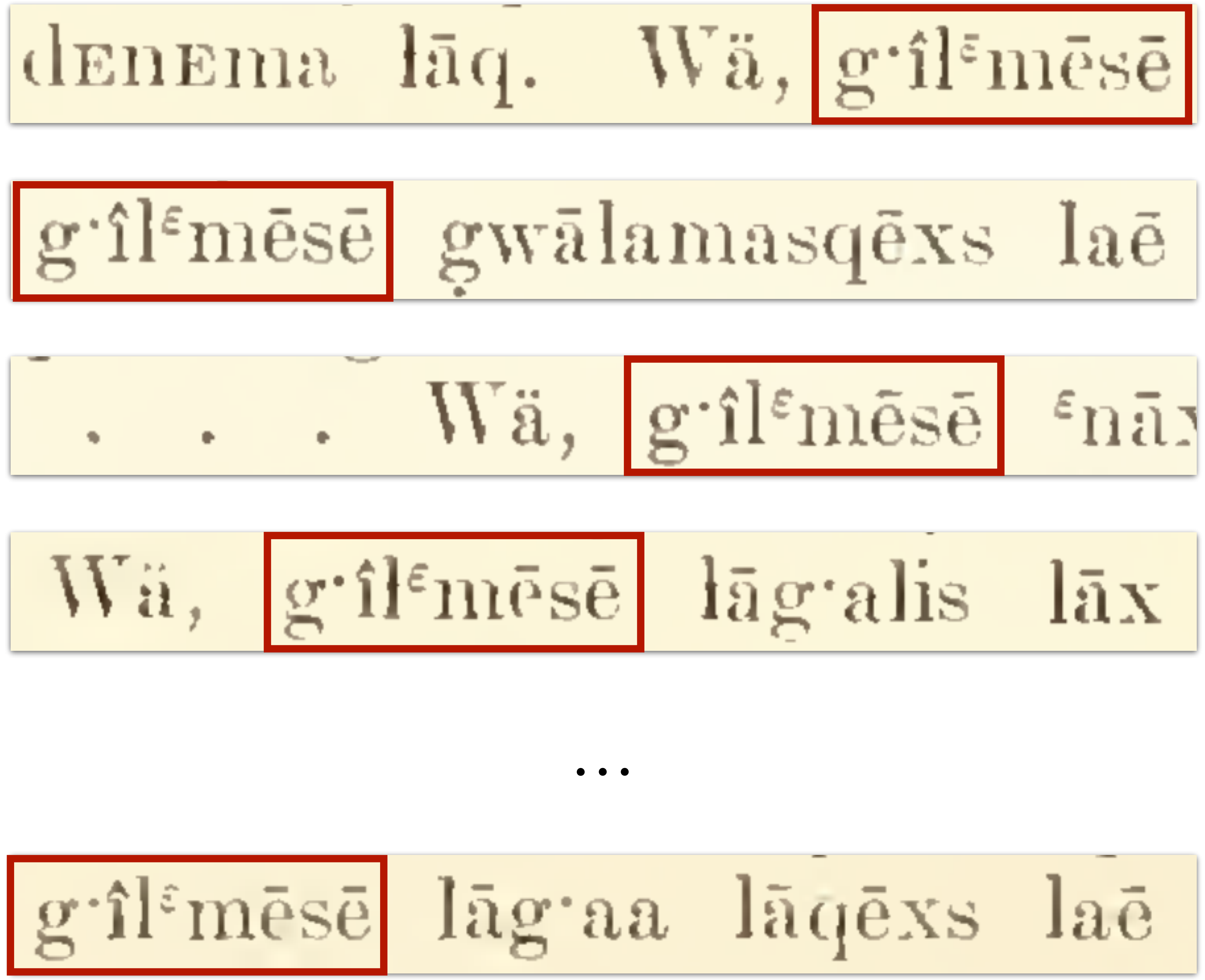
g·îlᵋmēsē (7)

g·ilᵋmēsē (5)

g·îlᵋmᵋsē (2)

Self-training may introduce noise

Can we bias post-correction towards generating correct words?



g·îl$^{\varepsilon}$mēsē (7)

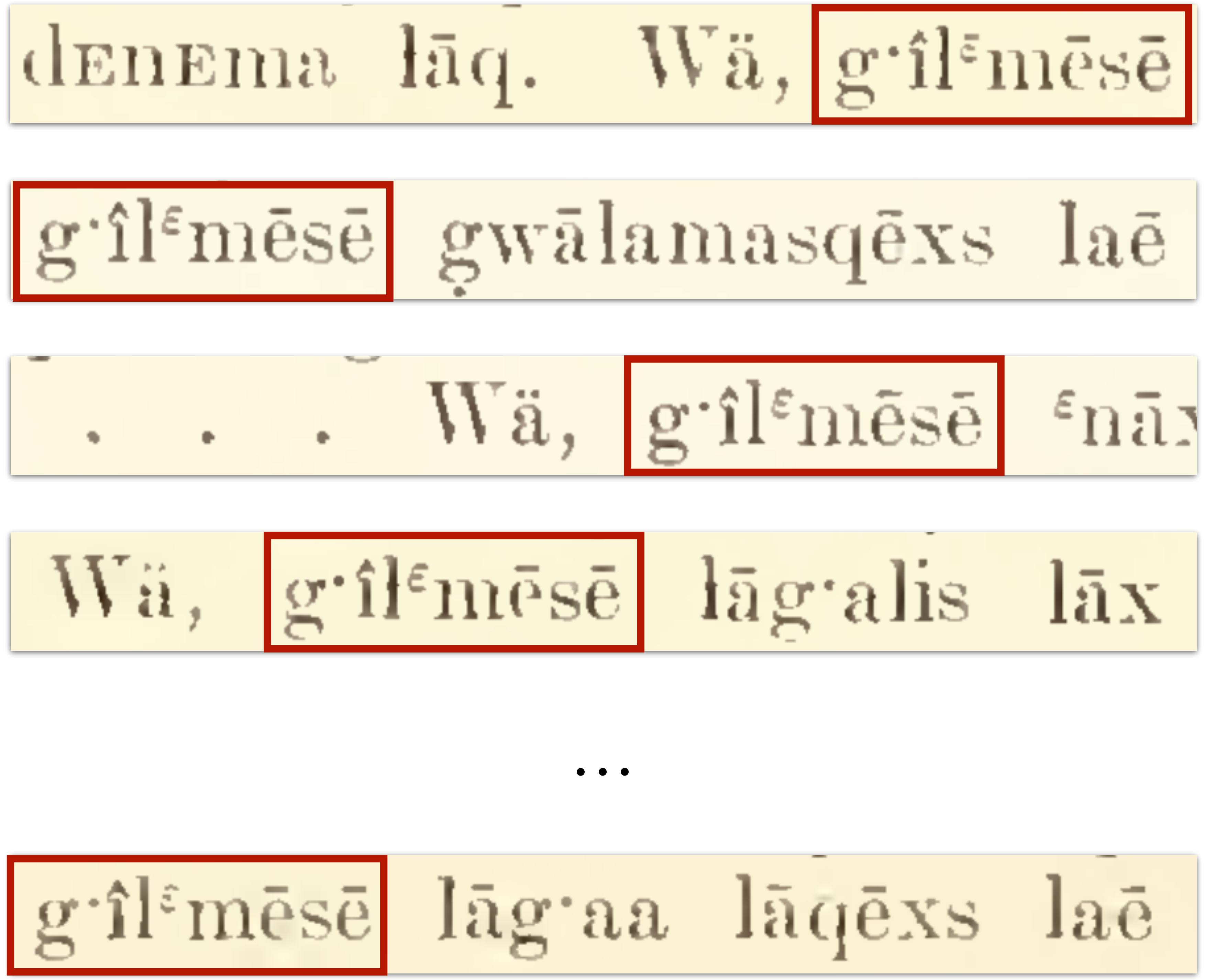
g·il$^{\varepsilon}$mēsē (5)

g·îl$^{\varepsilon}$m$^{\varepsilon}$sē (2)

g·îl$^{\varepsilon}$mi$^{\varepsilon}$sē (2)

Self-training may introduce noise

Can we bias post-correction towards generating correct words?



g·îl$^{\varepsilon}$mēsē (7)

g·il$^{\varepsilon}$mēsē (5)

g·îl$^{\varepsilon}$m$^{\varepsilon}$sē (2)

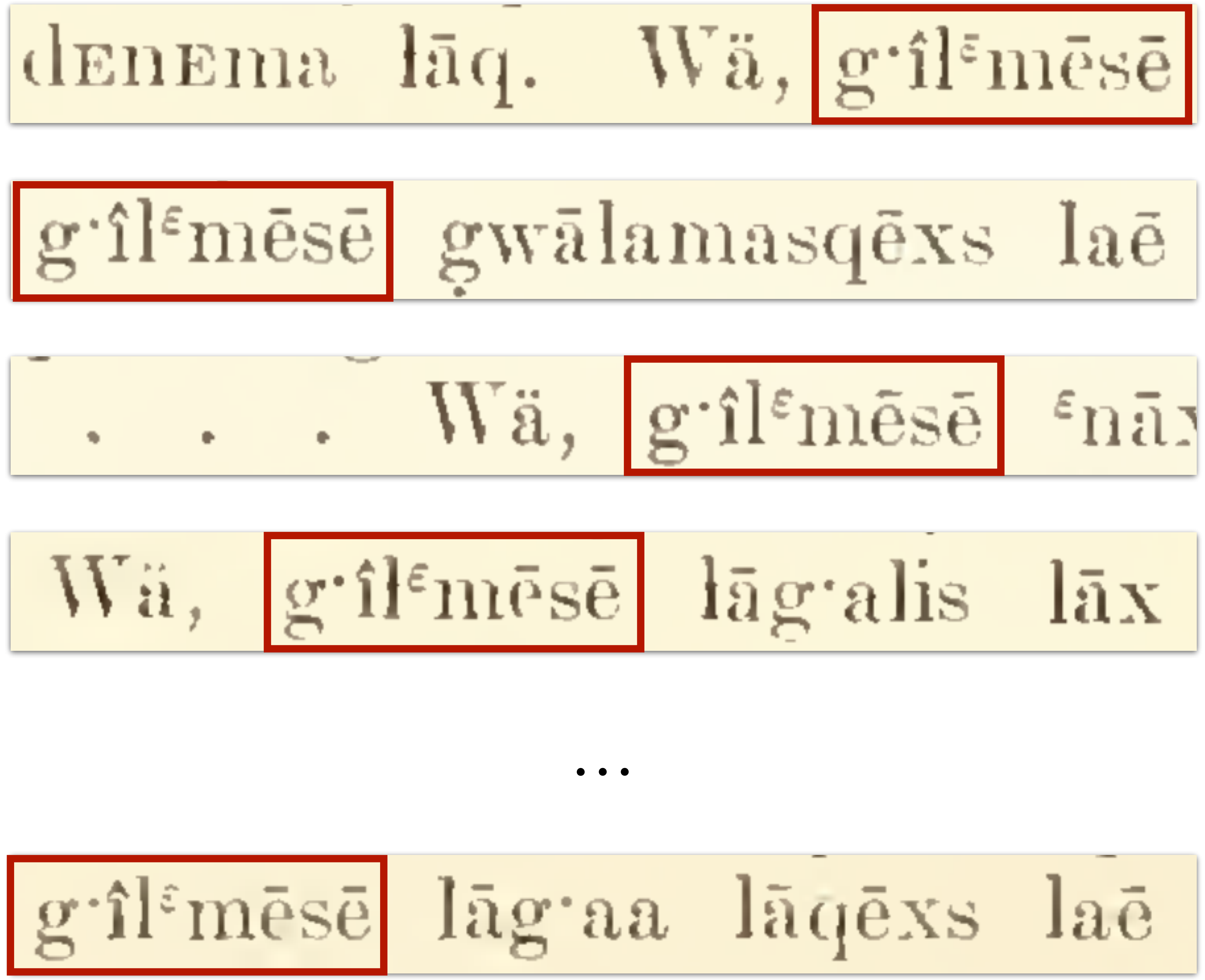
g·îl$^{\varepsilon}$mi$^{\varepsilon}$sē (2)

Different subsets of characters are incorrect

Empirical observations

Self-training may introduce noise

Can we bias post-correction towards generating correct words?



✔ g·îl$^{\varepsilon}$mēsē (7)

✘ g·il$^{\varepsilon}$mēsē (5)

✘ g·îl$^{\varepsilon}$m$^{\varepsilon}$sē (2)

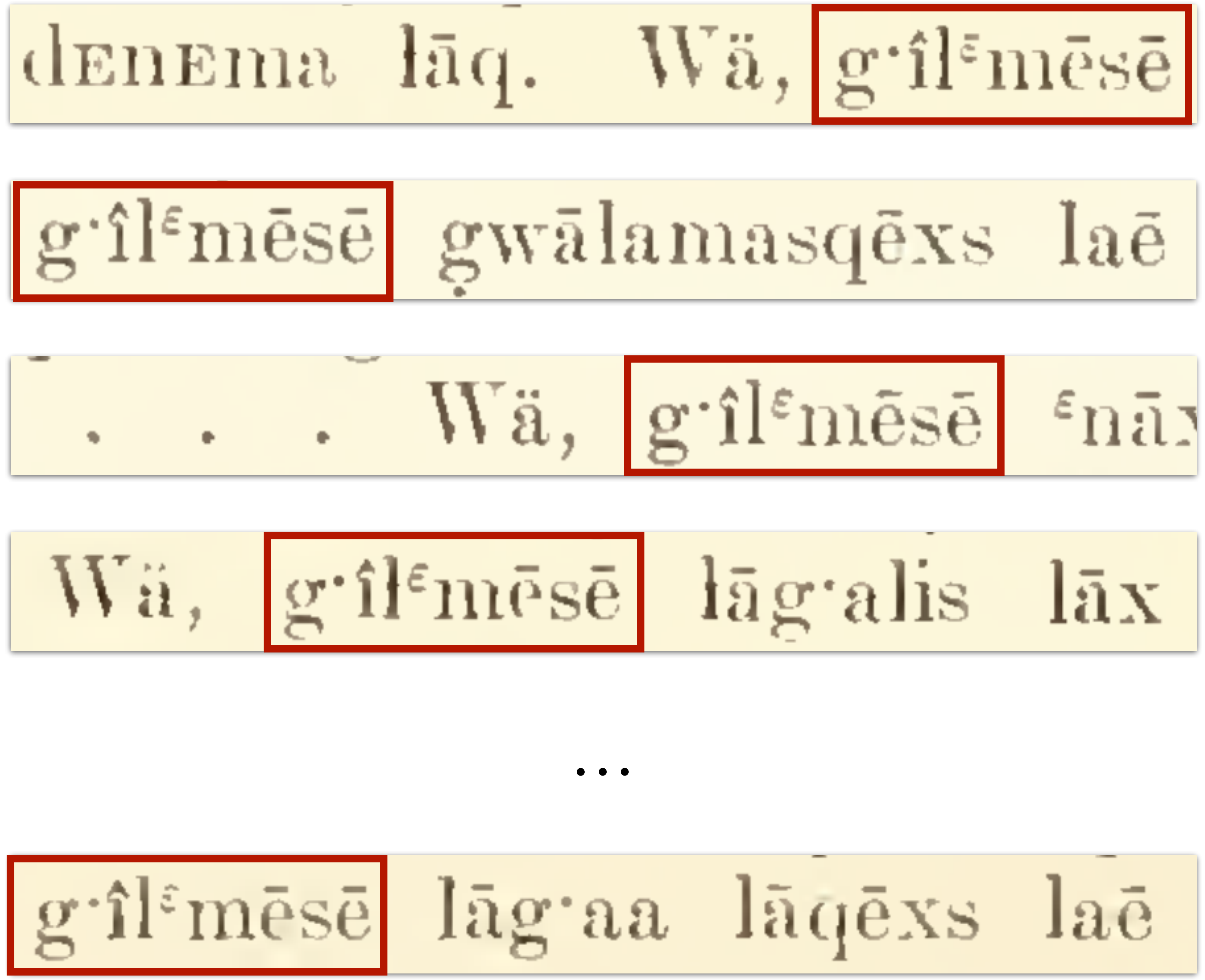
✘ g·îl$^{\varepsilon}$mi$^{\varepsilon}$sē (2)

Noise from self-training is typically inconsistent at the word-level

Empirical observations

Self-training may introduce noise

Can we bias post-correction towards generating correct words?


dEnEma lāq. Wä, g·îl$^{\varepsilon}$mēsē

g·îl$^{\varepsilon}$mēsē gwālamasqēxs laē

. . . Wä, g·îl$^{\varepsilon}$mēsē $^{\varepsilon}$nā

Wä, g·îl$^{\varepsilon}$mēsē lāg·alis lāx

…

g·îl$^{\varepsilon}$mēsē lāg·aa lāqēxs laē

g·îl$^{\varepsilon}$mēsē (7) ← Correct form of the word ends up being more frequent

g·il$^{\varepsilon}$mēsē (5)

g·îl$^{\varepsilon}$m$^{\varepsilon}$sē (2)

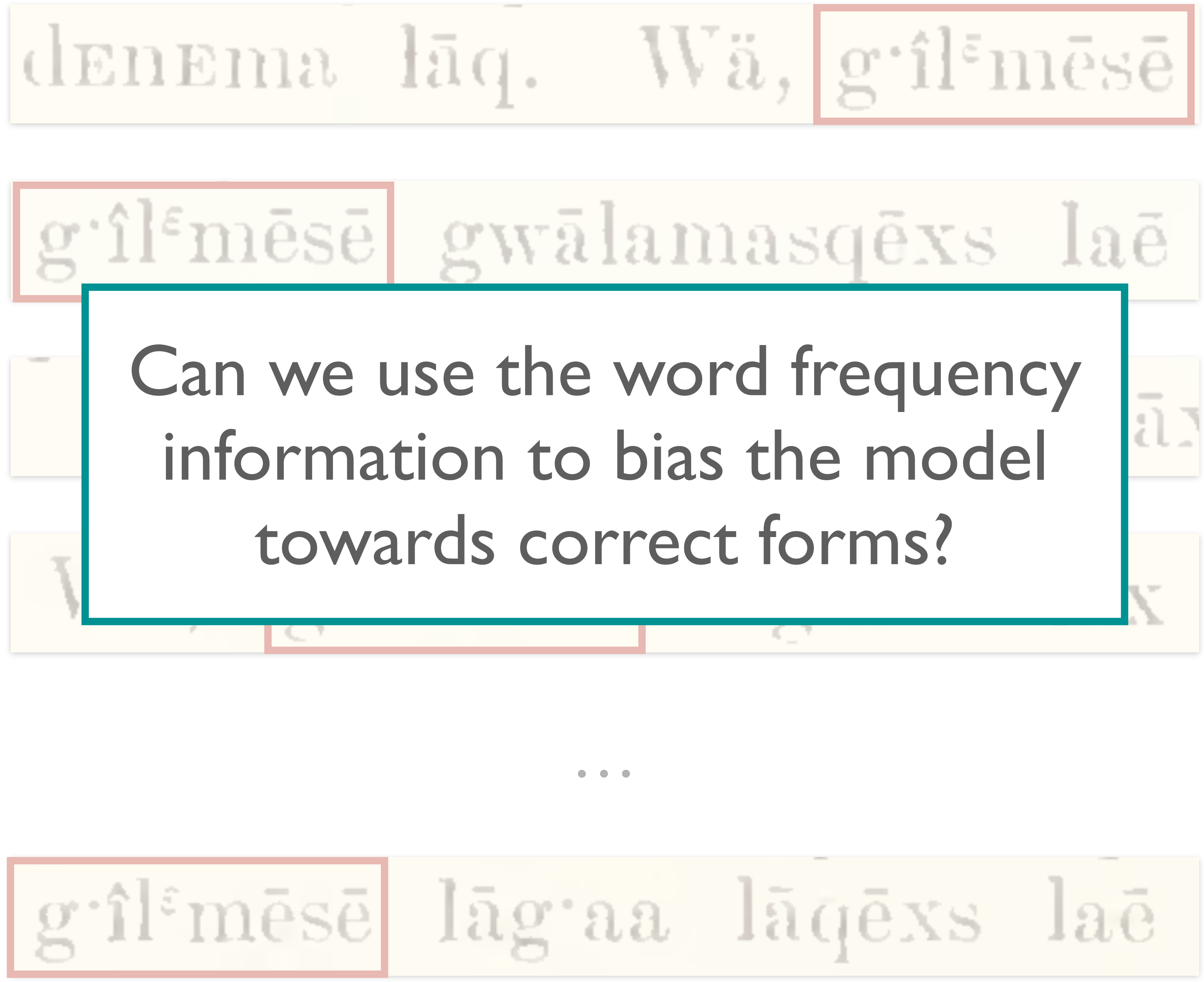
g·îl$^{\varepsilon}$mi$^{\varepsilon}$sē (2)

Noise from self-training is typically inconsistent at the word-level

Empirical observations

Self-training may introduce noise

Can we bias post-correction towards generating correct words?



Can we use the word frequency information to bias the model towards correct forms?

g·îl$^{\varepsilon}$mēsē (7)

g·il$^{\varepsilon}$mēsē (5)

g·îl$^{\varepsilon}$m$^{\varepsilon}$sē (2)

g·îl$^{\varepsilon}$mi$^{\varepsilon}$sē (2)

Correct form of the word ends up being more frequent

Incorporating word frequency information

Incorporating word frequency information

$$P(y) = p_{\mathrm{lstm}}(y)$$

Incorporating word frequency information

$$P(y) = p_{\text{lstm}}(y)$$

Next character
probability

Incorporating word frequency information

$$P(y) = p_{\mathrm{lstm}}(y)$$

Next character probability

Decoder probability

Incorporating word frequency information

$$P(y) = p_{\text{lstm}}(y) \qquad\qquad p_{\text{freq}}$$

Frequency-based probability to explicitly bias the model

Incorporating word frequency information

$$P(y) = p_{\text{lstm}}(y) \qquad\qquad p_{\text{freq}}$$

How do we get probabilities based on word frequency?

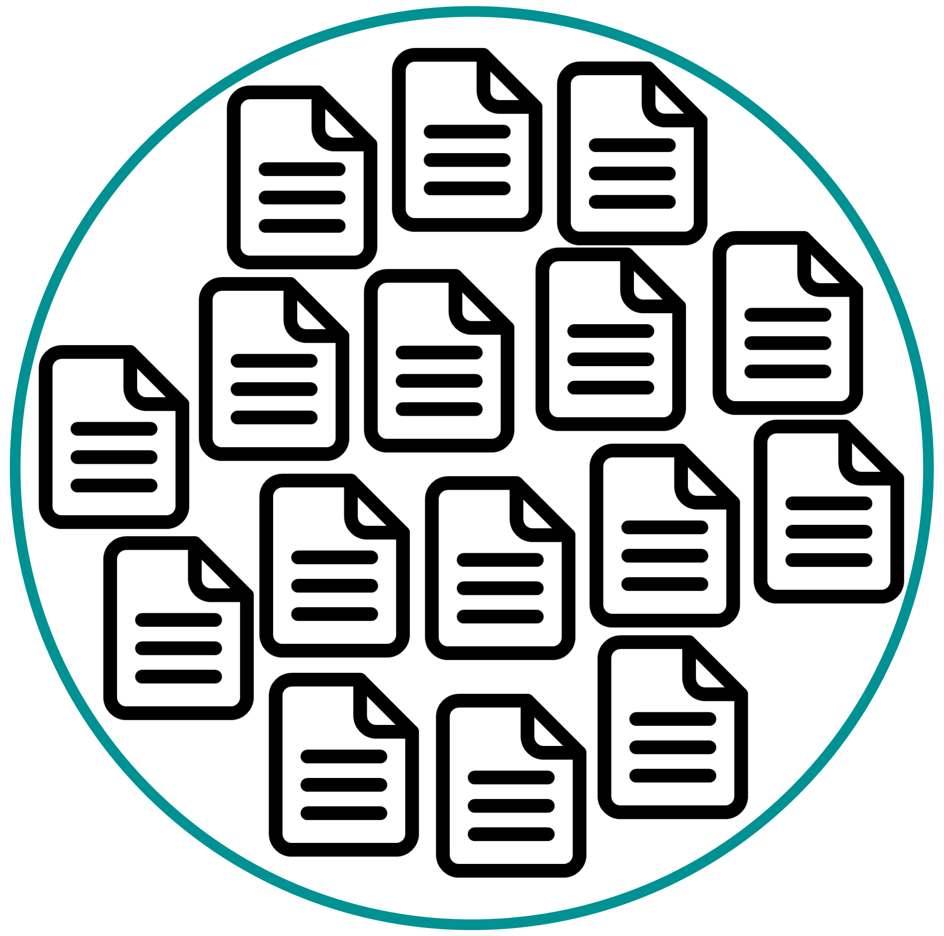
Modeling word frequency

Modeling word frequency

Simple model for word frequency: count-based language model

Modeling word frequency

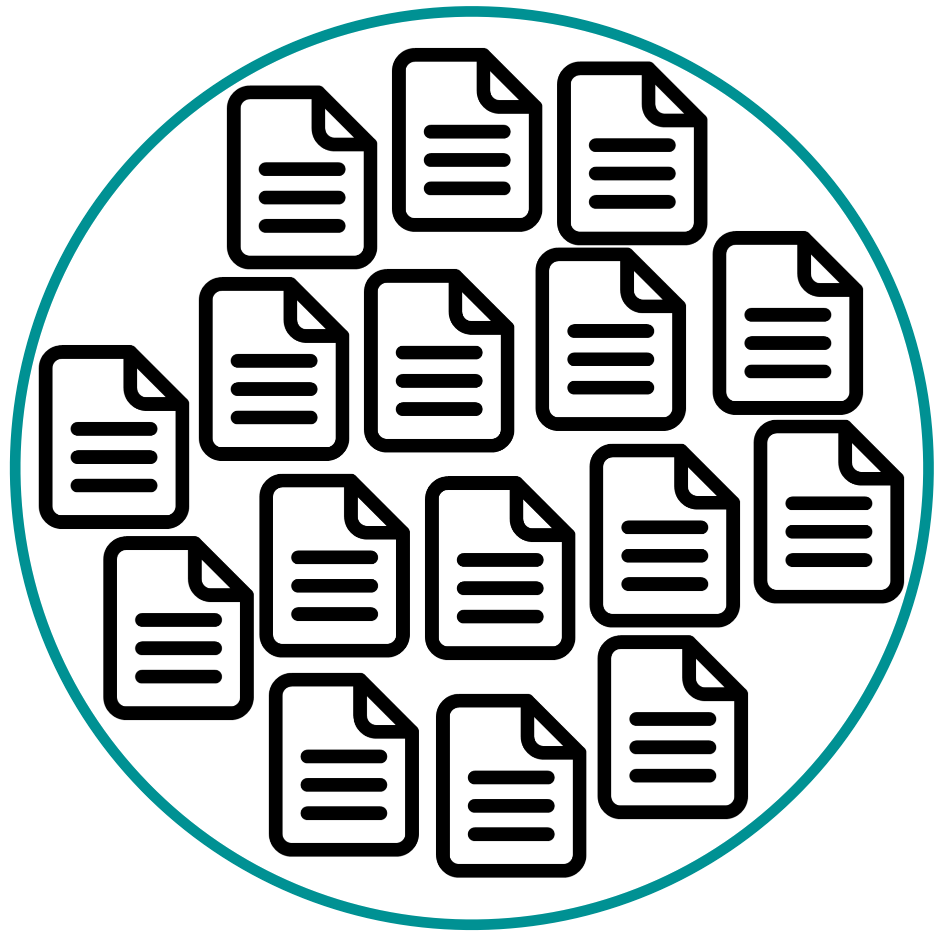
Simple model for word frequency: count-based language model



Predictions from self-training

Modeling word frequency

Simple model for word frequency: count-based language model

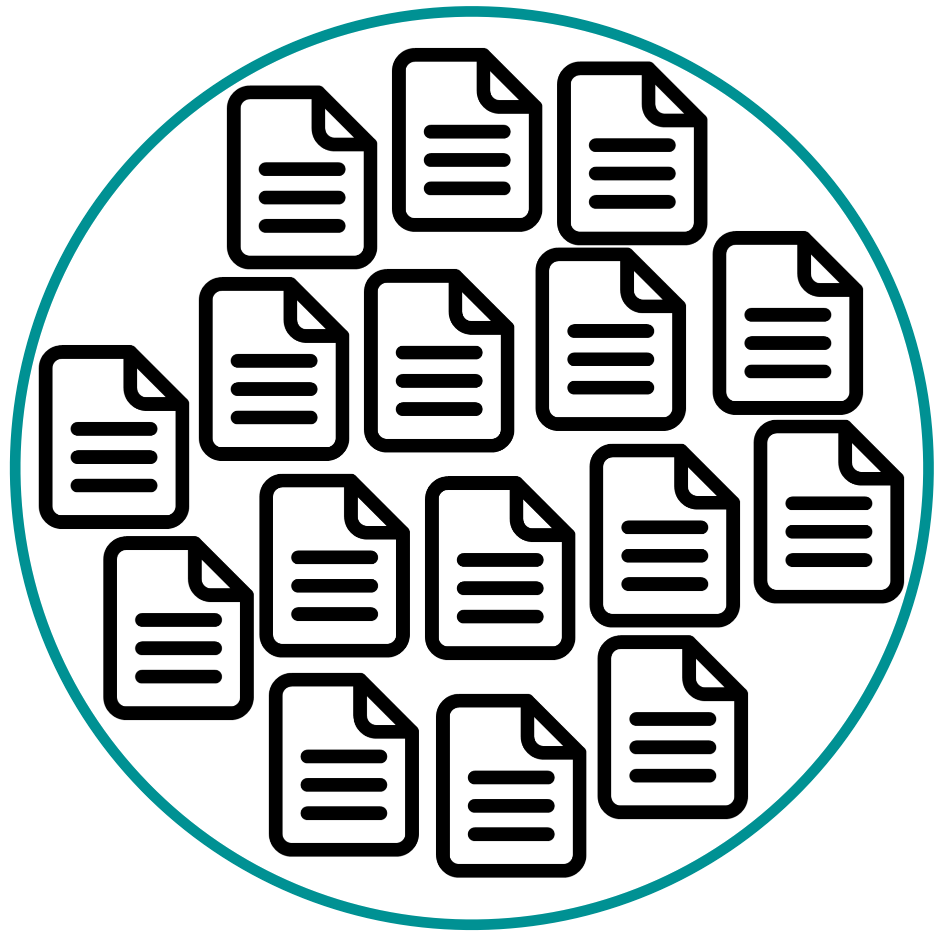


Predictions from self-training

Train a smoothed unigram LM

Modeling word frequency

Simple model for word frequency: count-based language model



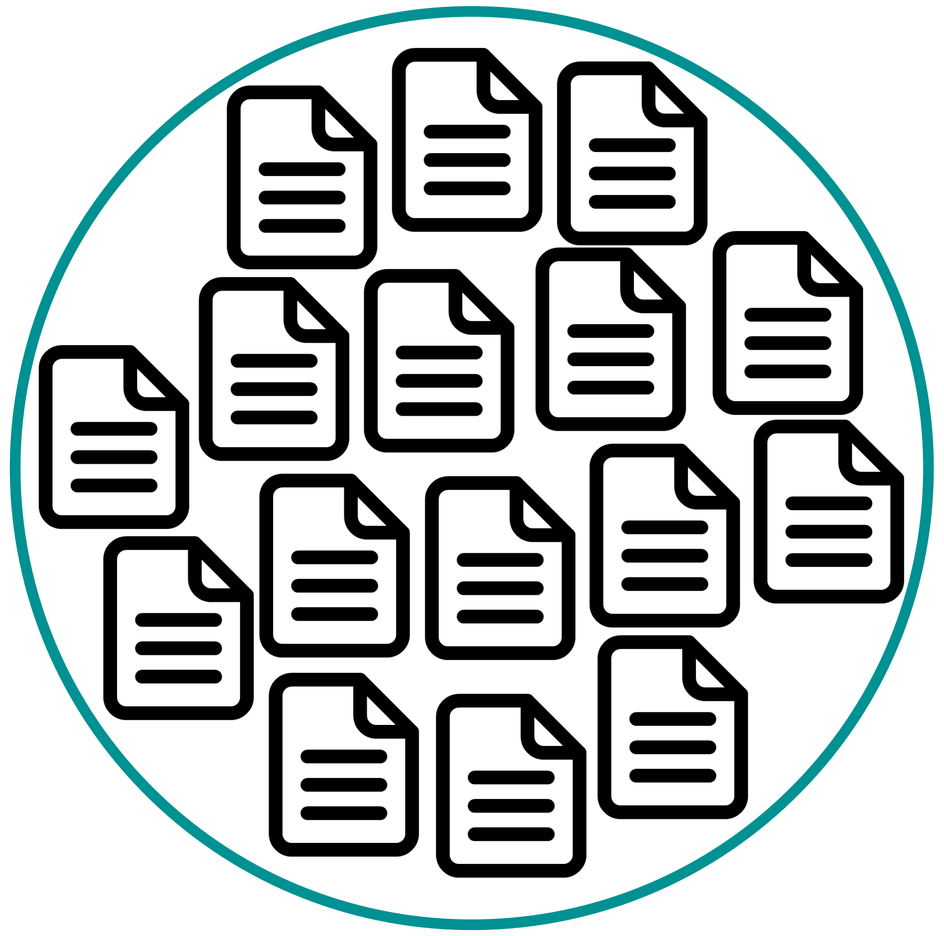
Train a smoothed unigram LM

Predictions from self-training

Using counts of the word forms

Modeling word frequency

Simple model for word frequency: count-based language model



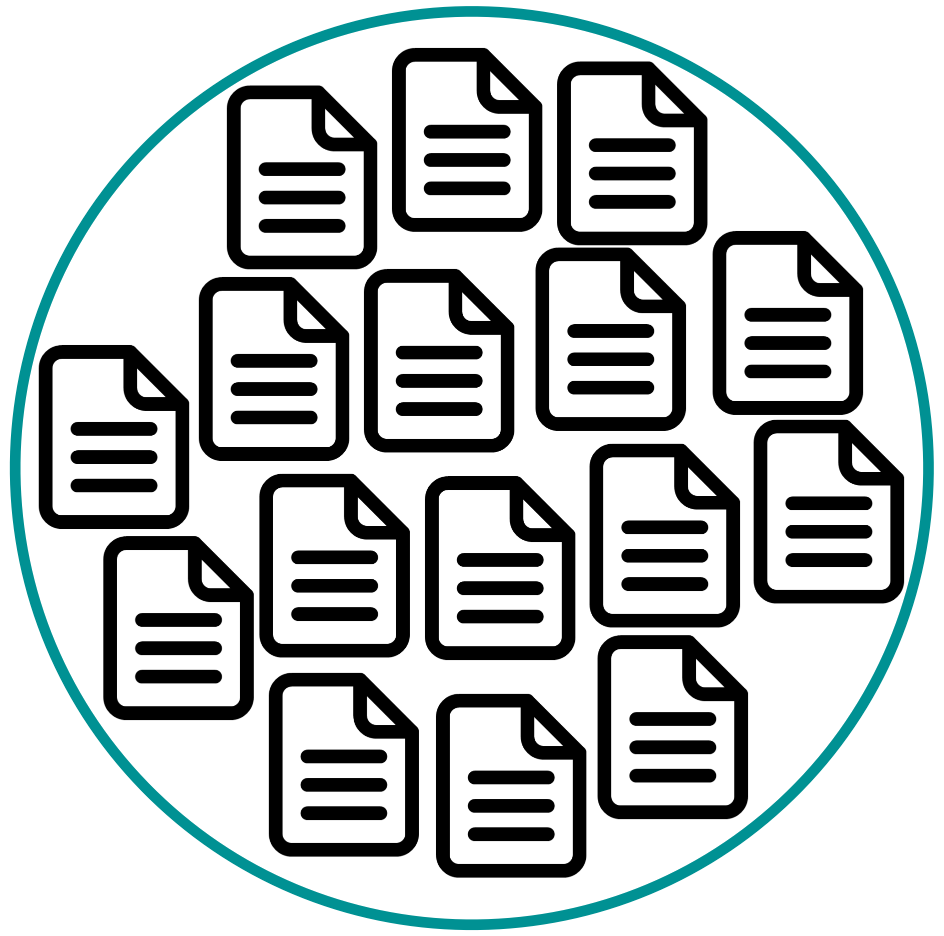
Predictions from self-training

Train a smoothed unigram LM

Frequency-based word-level probabilities

Modeling word frequency

Simple model for word frequency: count-based language model



Predictions from self-training

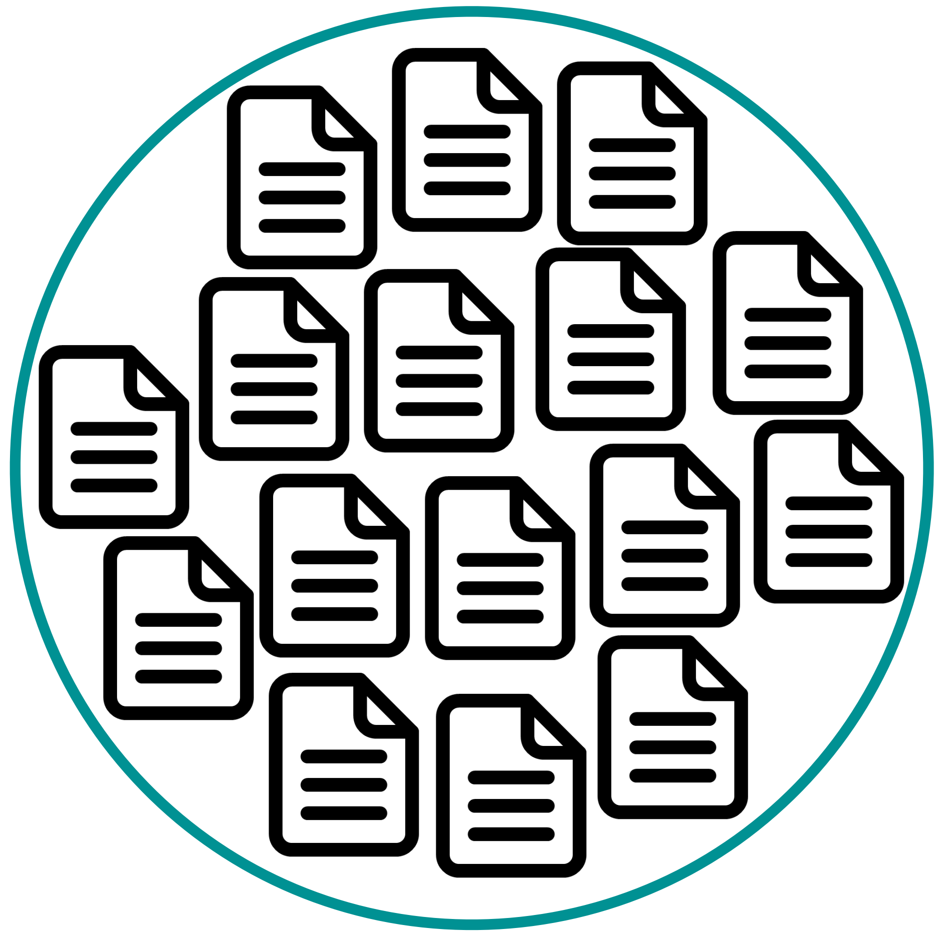
Train a smoothed unigram LM

Frequency-based word-level probabilities

Noisy weighted lexicon of the words in the predictions

Modeling word frequency

Simple model for word frequency: count-based language model



Predictions from self-training

Train a smoothed unigram LM

Frequency-based word-level probabilities

Lexically-aware decoding for post-correction

Lexically-aware decoding for post-correction

$$P(y) = p_{\mathrm{lstm}}(y) \qquad\qquad p_{\mathrm{freq}}$$

Lexically-aware decoding for post-correction

$$P(y) = p_{\text{lstm}}(y) \qquad\qquad p_{\text{freq}}$$

We have frequency-based probabilities from the unigram LM!

Lexically-aware decoding for post-correction

$$P(y) = p_{\text{lstm}}(y) \qquad\qquad p_{\text{freq}}$$

But these are at the word-level: how we get character-level scores?

Scoring at the character-level

Example LM with two words:

- P('dog') = 0.75
- P('door') = 0.2
- P(<unk>) = 0.05

Scoring at the character-level

Weighted Finite State Automaton (WFSA) representation of the LM

←

Example LM with two words:

- P('dog') = 0.75
- P('door') = 0.2
- P(<unk>) = 0.05

Scoring at the character-level

Weighted Finite State Automaton (WFSA) representation of the LM

Set of states with weighted transitions

Example LM with two words:

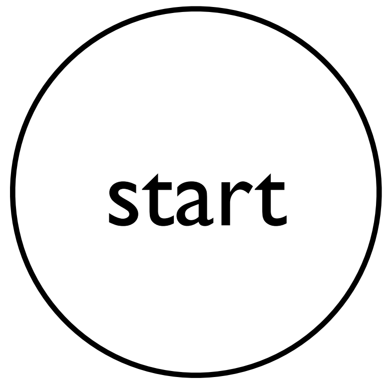
- P('dog') = 0.75
- P('door') = 0.2
- P(<unk>) = 0.05

Scoring at the character-level

Weighted Finite State Automaton (WFSA) representation of the LM

Example LM with two words:

- P('dog') = 0.75
- P('door') = 0.2
- P(<unk>) = 0.05

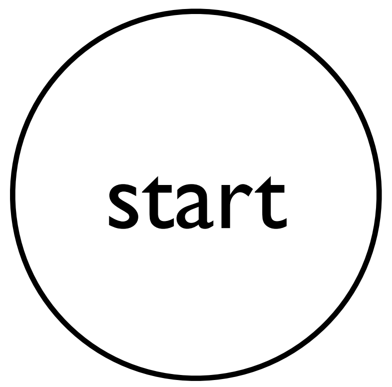


Scoring at the character-level

Weighted Finite State Automaton (WFSA)
representation of the LM

Example LM with two words:

- P('dog') = 0.75
- P('door') = 0.2
- P(<unk>) = 0.05

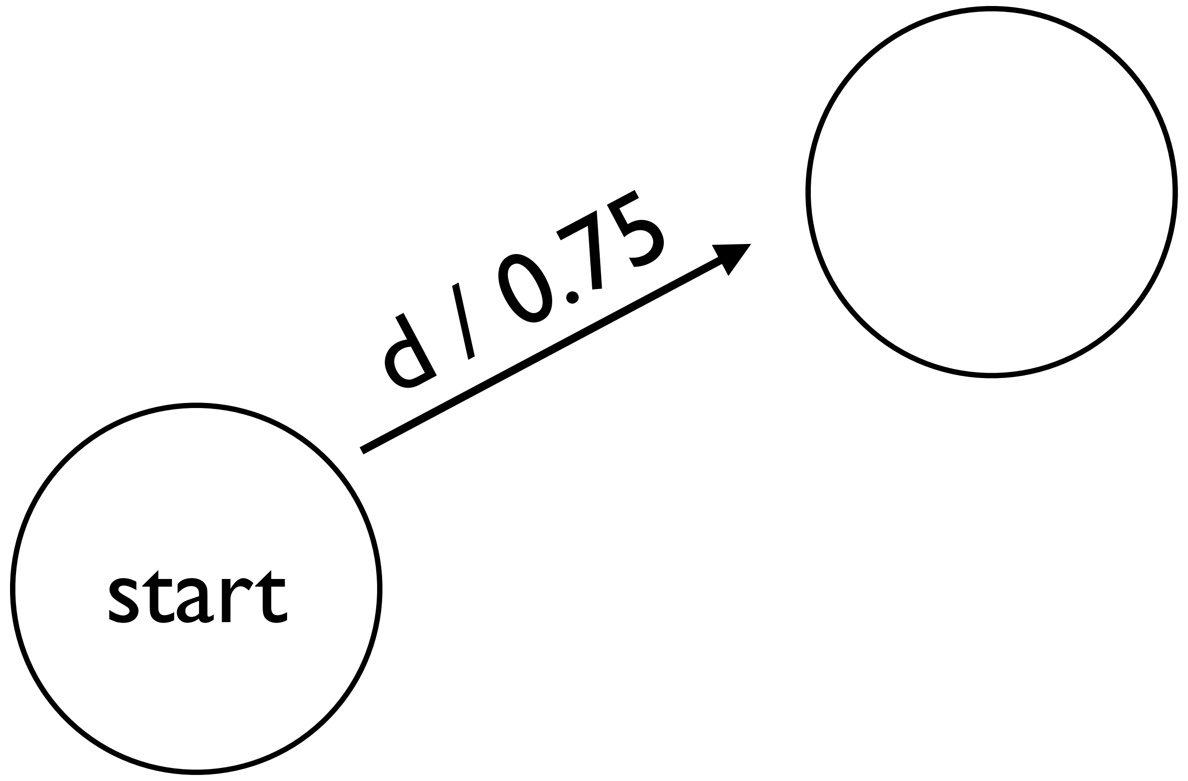


Scoring at the character-level

Weighted Finite State Automaton (WFSA) representation of the LM

Example LM with two words:

- P(‘dog’) = 0.75
- P(‘door’) = 0.2
- P(<unk>) = 0.05

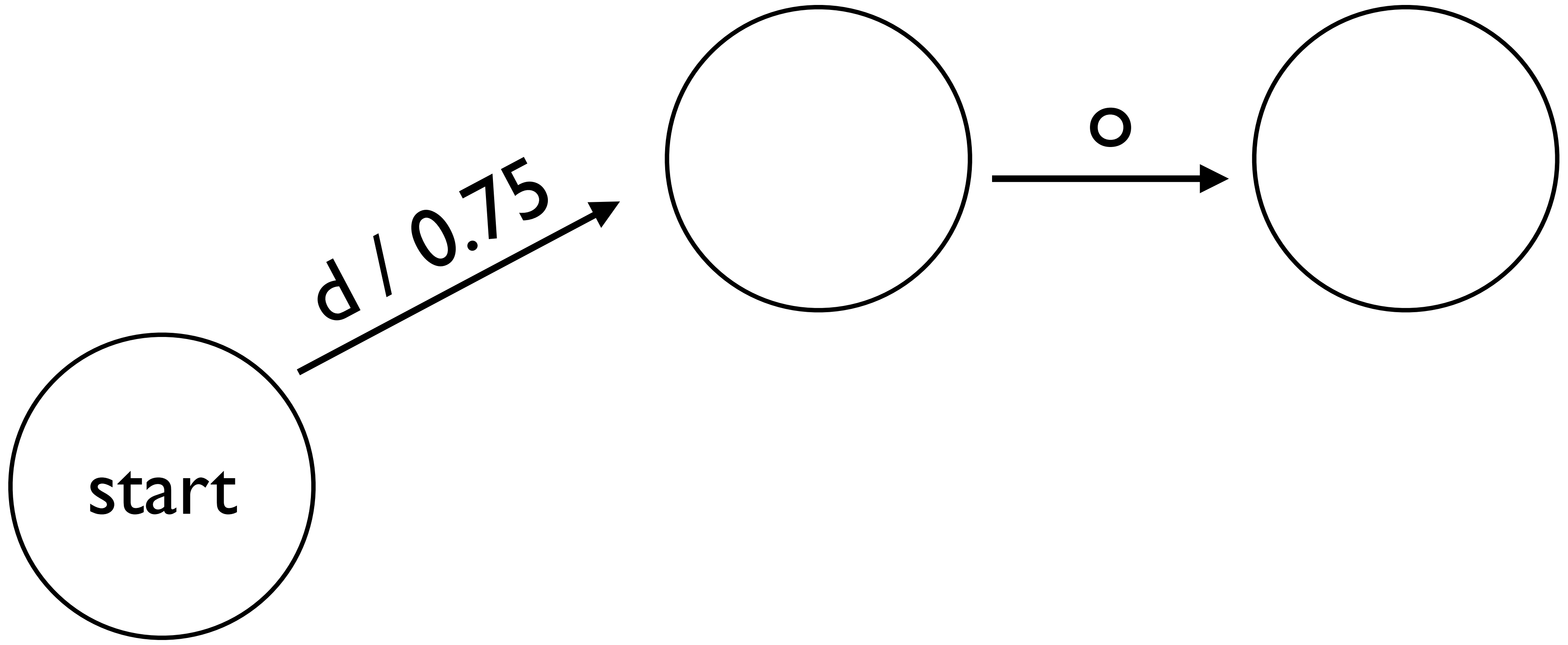


Scoring at the character-level

Weighted Finite State Automaton (WFSA) representation of the LM

Example LM with two words:

- P('dog') = 0.75
- P('door') = 0.2
- P(<unk>) = 0.05

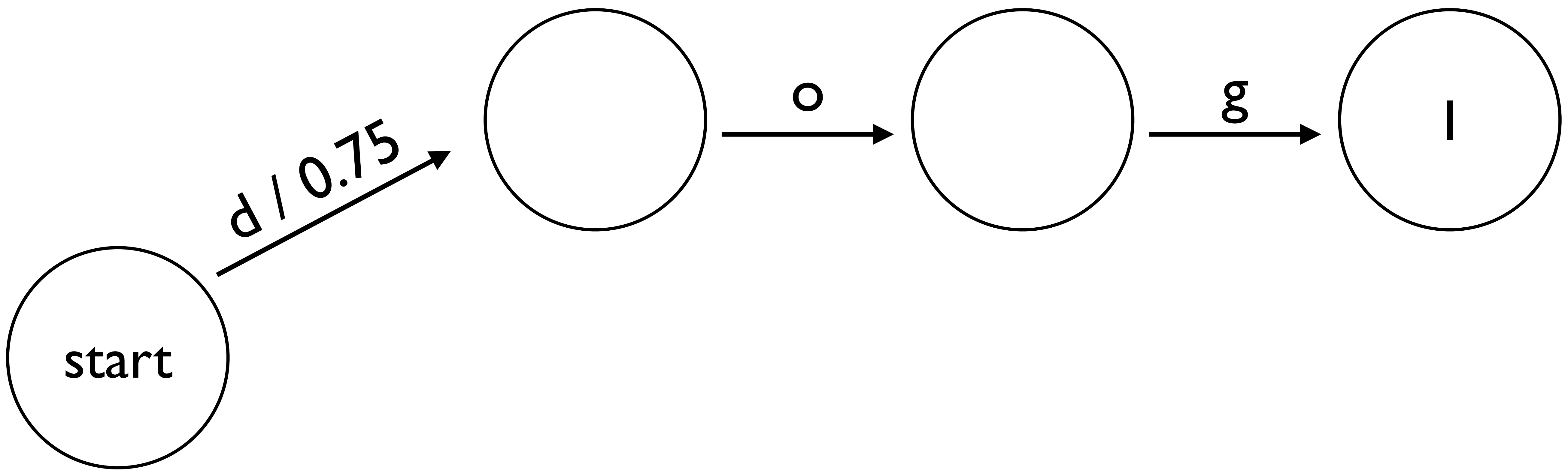


Scoring at the character-level

Weighted Finite State Automaton (WFSA) representation of the LM

Example LM with two words:

- P('dog') = 0.75
- P('door') = 0.2
- P(<unk>) = 0.05

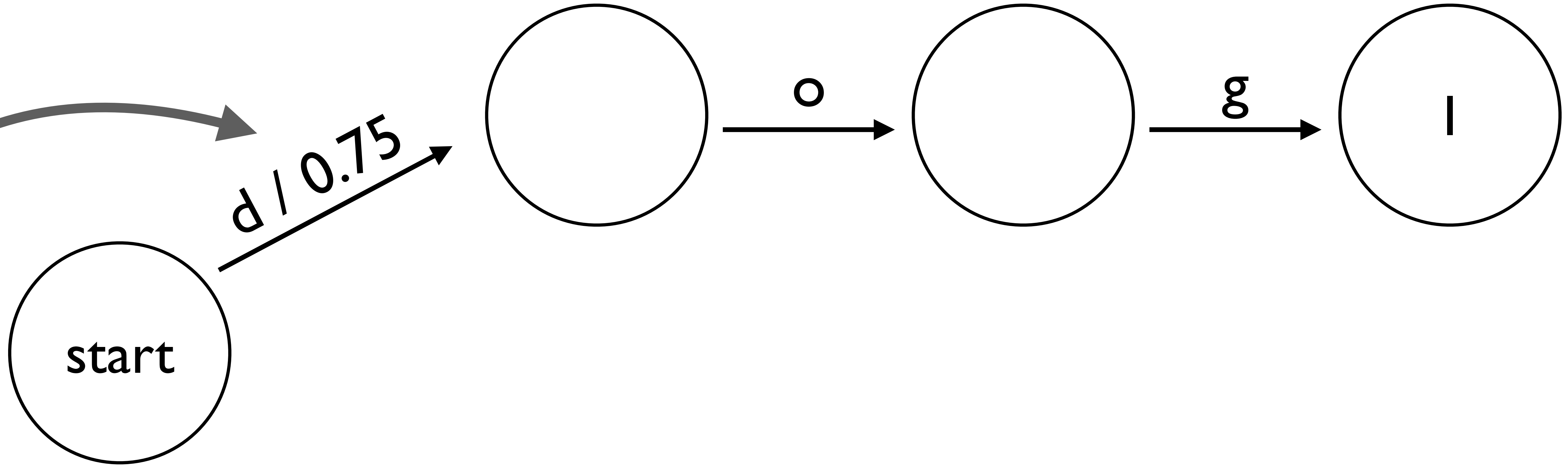


Scoring at the character-level

Weighted Finite State Automaton (WFSA) representation of the LM

Example LM with two words:

- $\boxed{\text{P('dog') = 0.75}}$
- P('door') = 0.2
- P(<unk>) = 0.05



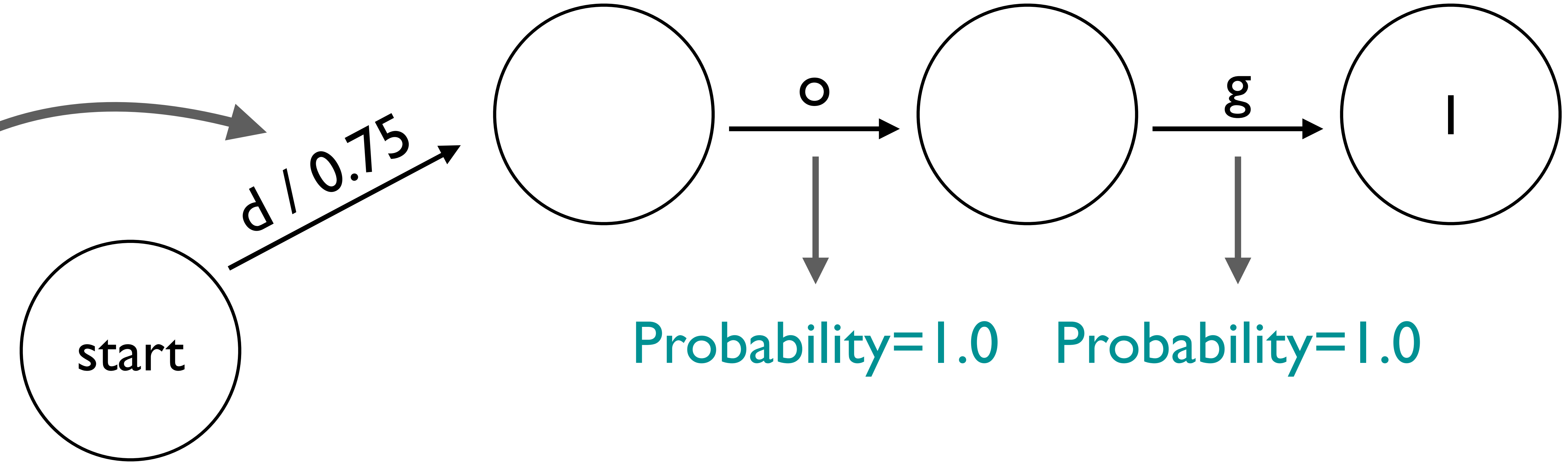
Probability from the count-based LM

Scoring at the character-level

Weighted Finite State Automaton (WFSA) representation of the LM

Example LM with two words:

- P(‘dog’) = 0.75
- P(‘door’) = 0.2
- P(<unk>) = 0.05



Probability=1.0 Probability=1.0

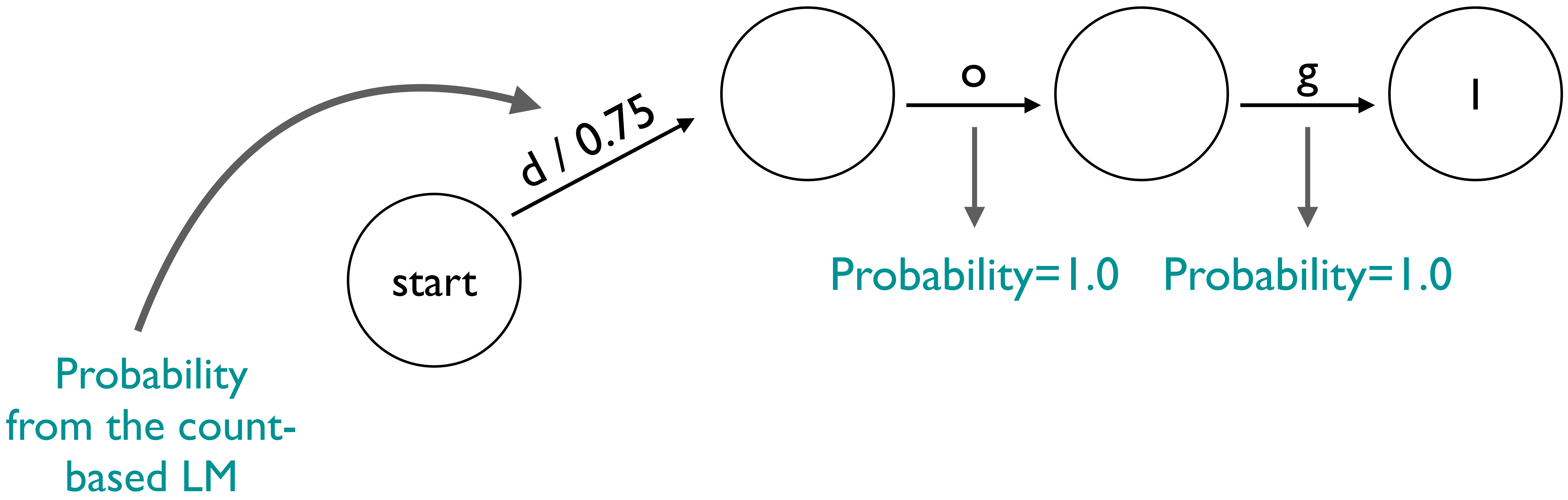
Probability from the count-based LM

Scoring at the character-level

Weighted Finite State Automaton (WFSA) representation of the LM

Example LM with two words:

- $\boxed{\text{P('dog') = 0.75}}$
- P('door') = 0.2
- P(<unk>) = 0.05



Weight of path for "dog" = 0.75

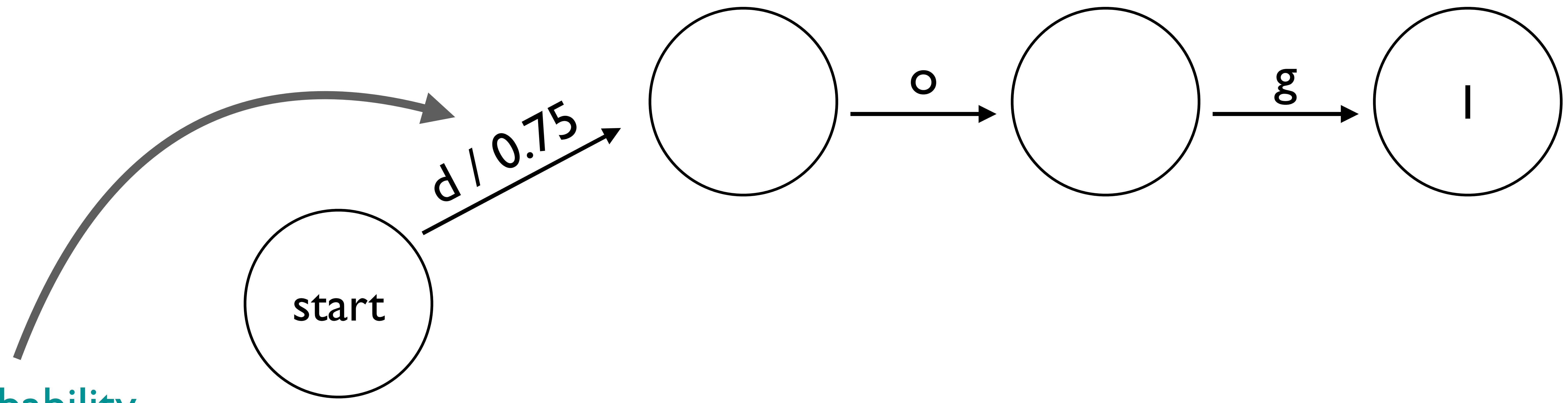
Same as the word-level LM!

Scoring at the character-level

Weighted Finite State Automaton (WFSA) representation of the LM

Example LM with two words:

- P('dog') = 0.75
- P('door') = 0.2
- P(<unk>) = 0.05



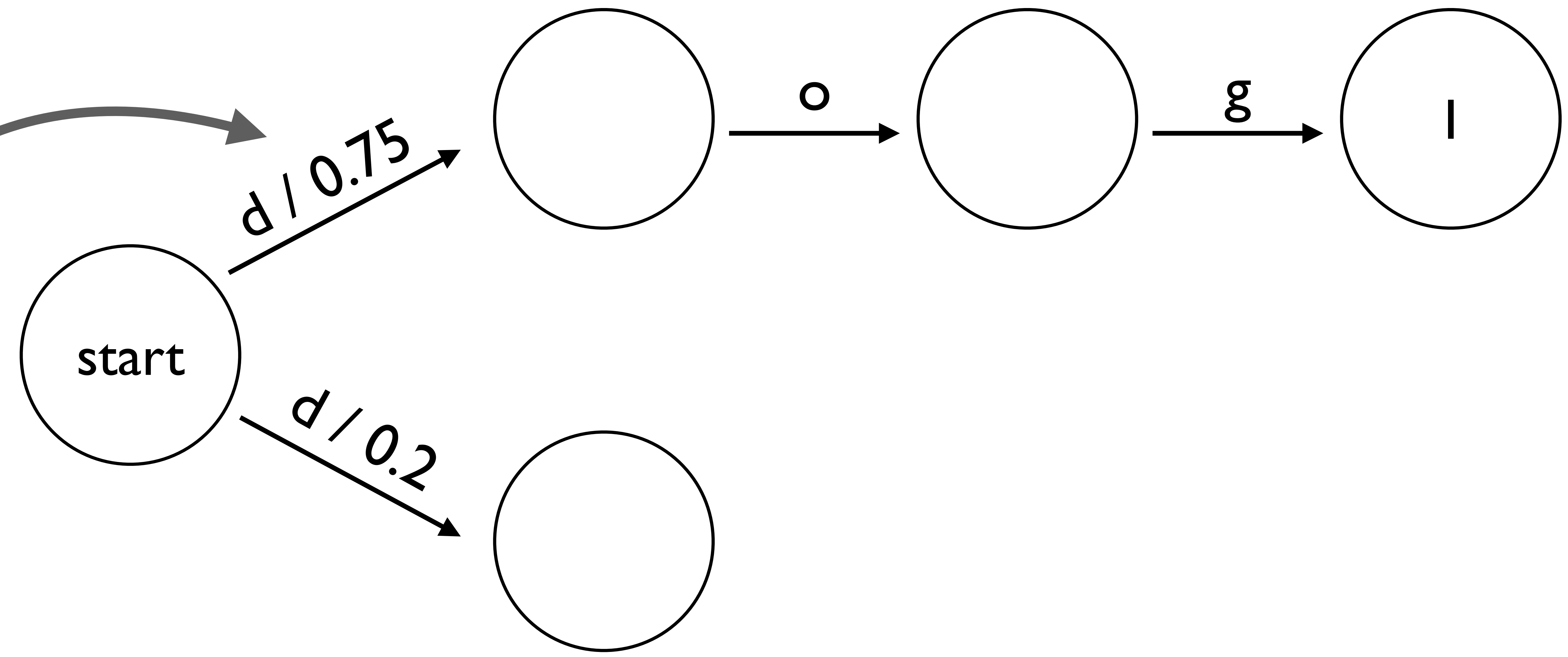
Probability from the count-based LM

Scoring at the character-level

Weighted Finite State Automaton (WFSA) representation of the LM

Example LM with two words:

- P('dog') = 0.75
- $\boxed{\text{P('door') = 0.2}}$
- P(<unk>) = 0.05



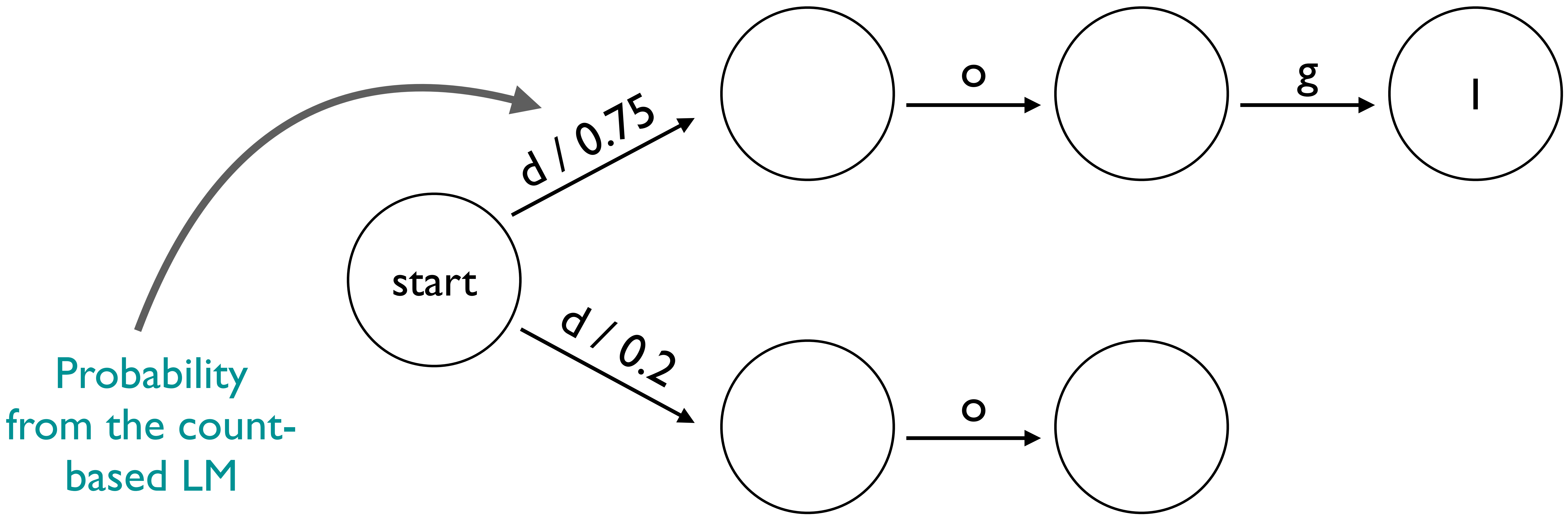
Probability from the count-based LM

Scoring at the character-level

Weighted Finite State Automaton (WFSA) representation of the LM

Example LM with two words:

- P(‘dog’) = 0.75
- P(‘door’) = 0.2
- P(<unk>) = 0.05

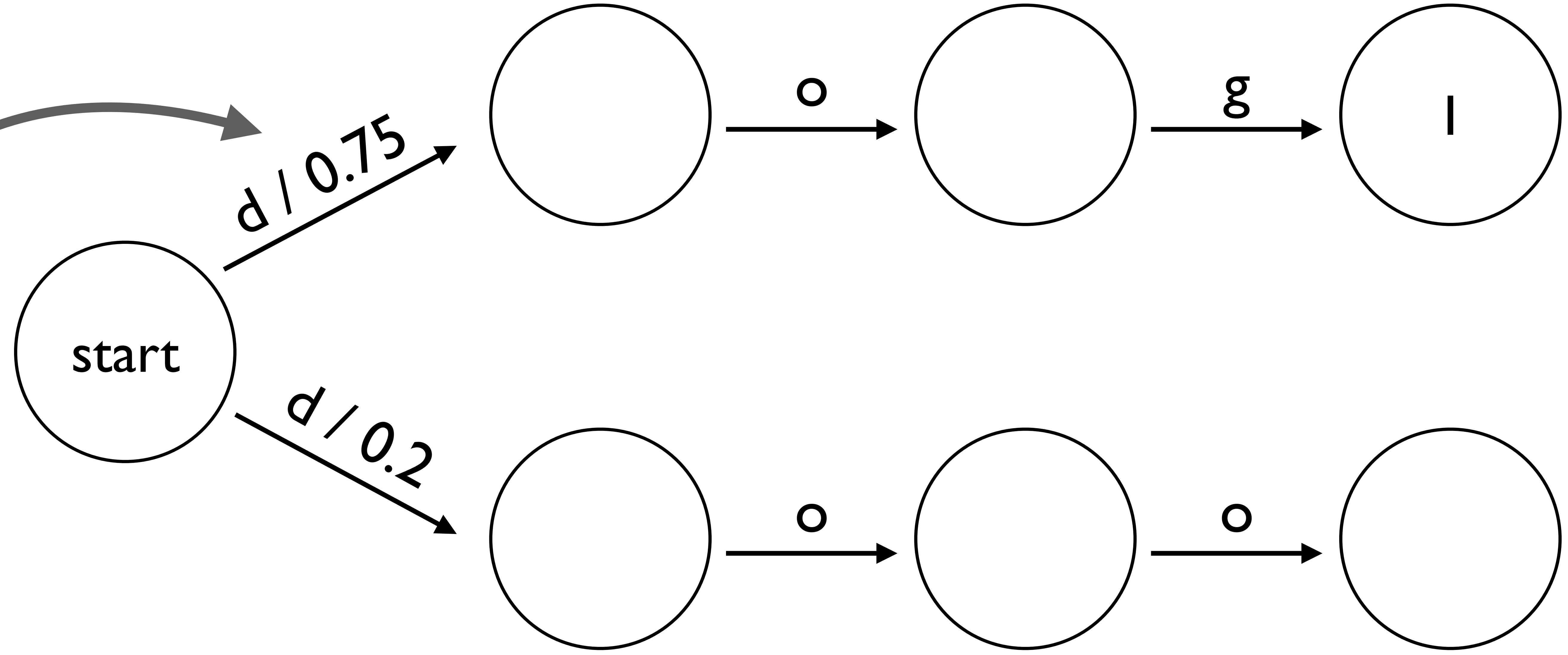


Scoring at the character-level

Weighted Finite State Automaton (WFSA) representation of the LM

Example LM with two words:

- P('dog') = 0.75
- P('door') = 0.2
- P(<unk>) = 0.05



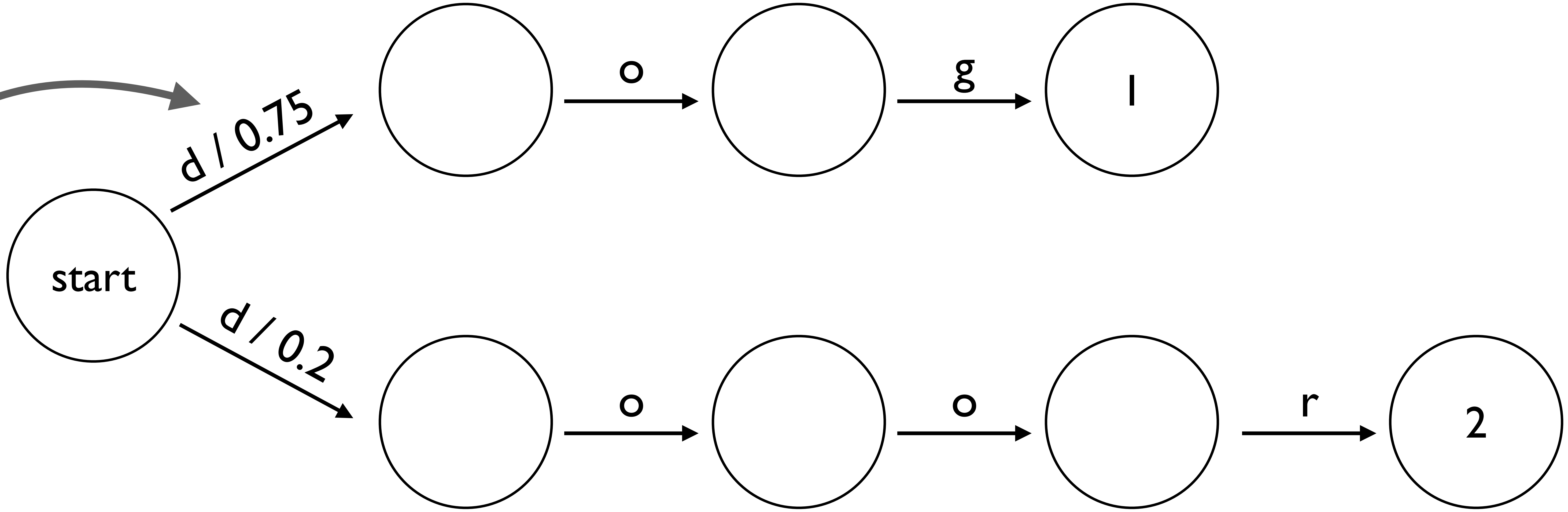
Probability from the count-based LM

Scoring at the character-level

Weighted Finite State Automaton (WFSA) representation of the LM

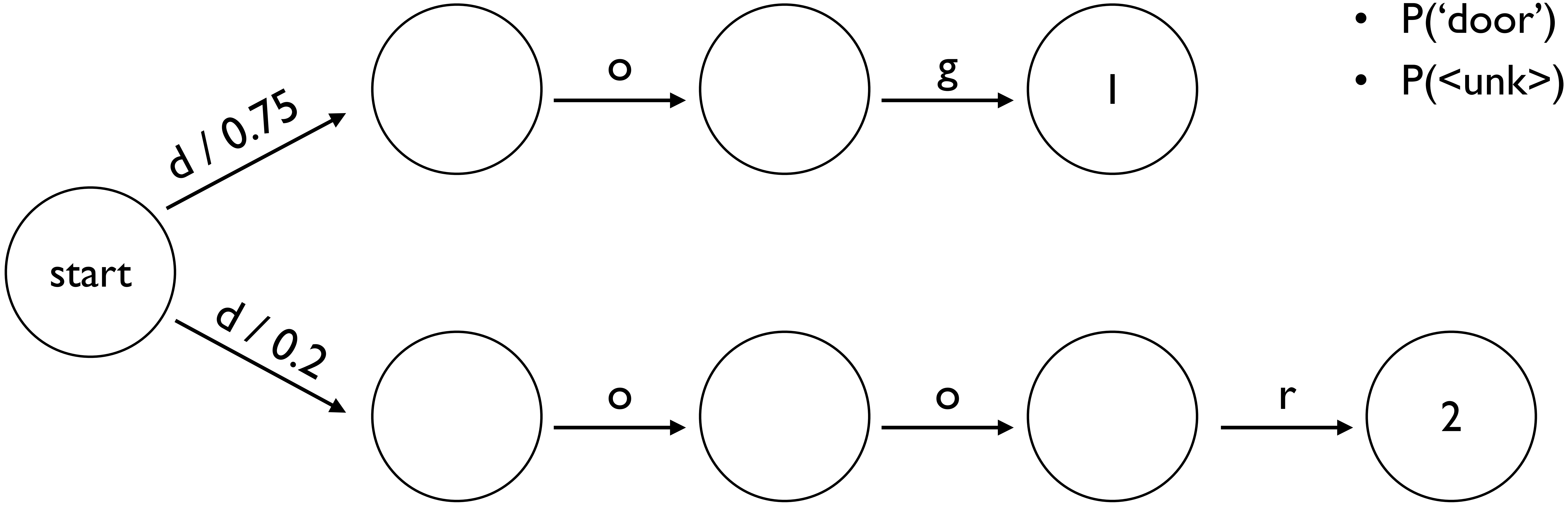
Example LM with two words:

- P('dog') = 0.75
- P('door') = 0.2
- P(<unk>) = 0.05



Probability from the count-based LM

Scoring at the character-level



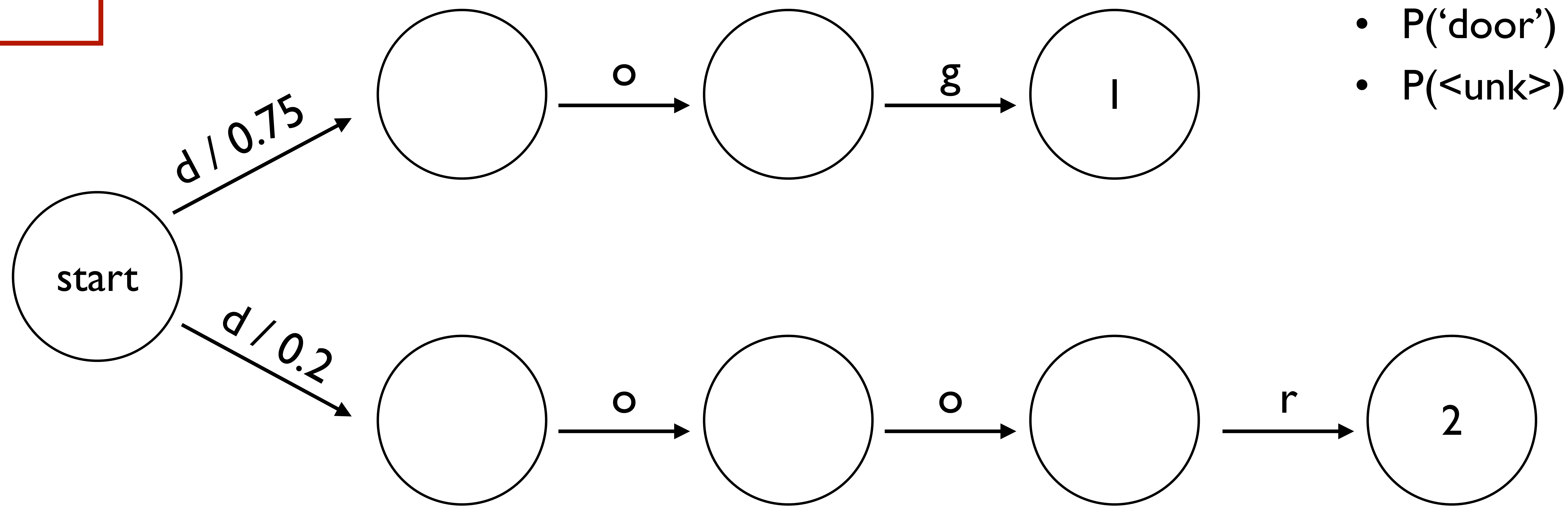
Example LM with two words:

- P('dog') = 0.75
- P('door') = 0.2
- P(<unk>) = 0.05

Scoring at the character-level

Output:

d o o r



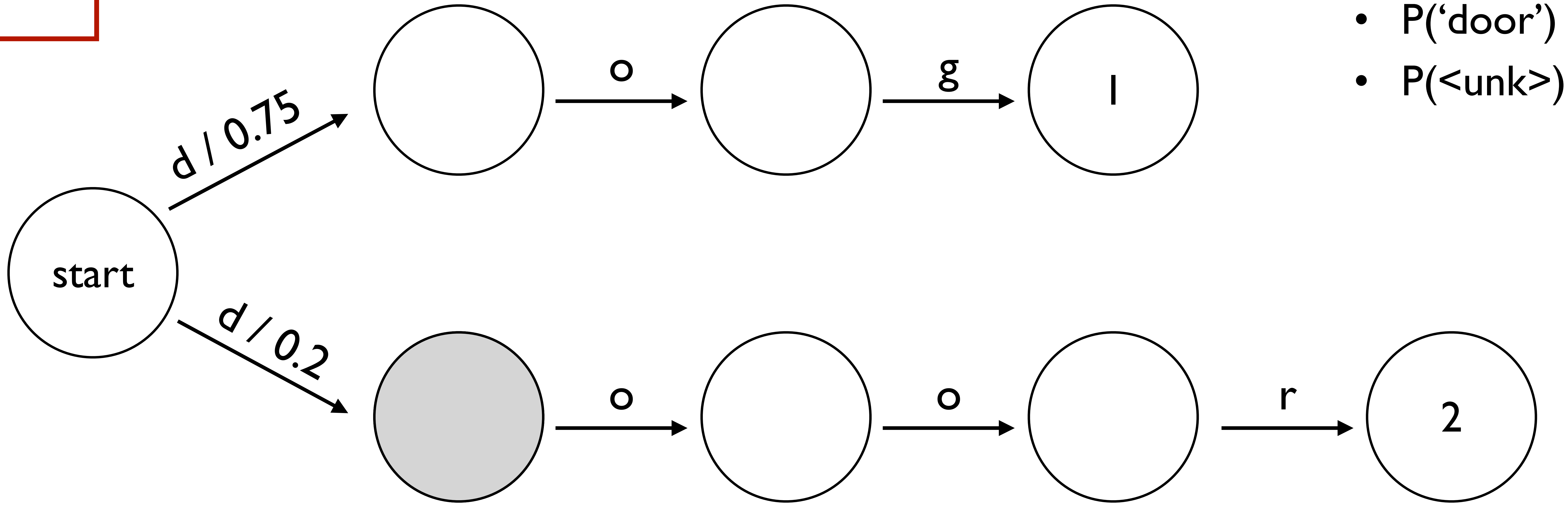
Example LM with two words:

- P(‘dog’) = 0.75
- P(‘door’) = 0.2
- P(<unk>) = 0.05

Scoring at the character-level

Output:

d o o r



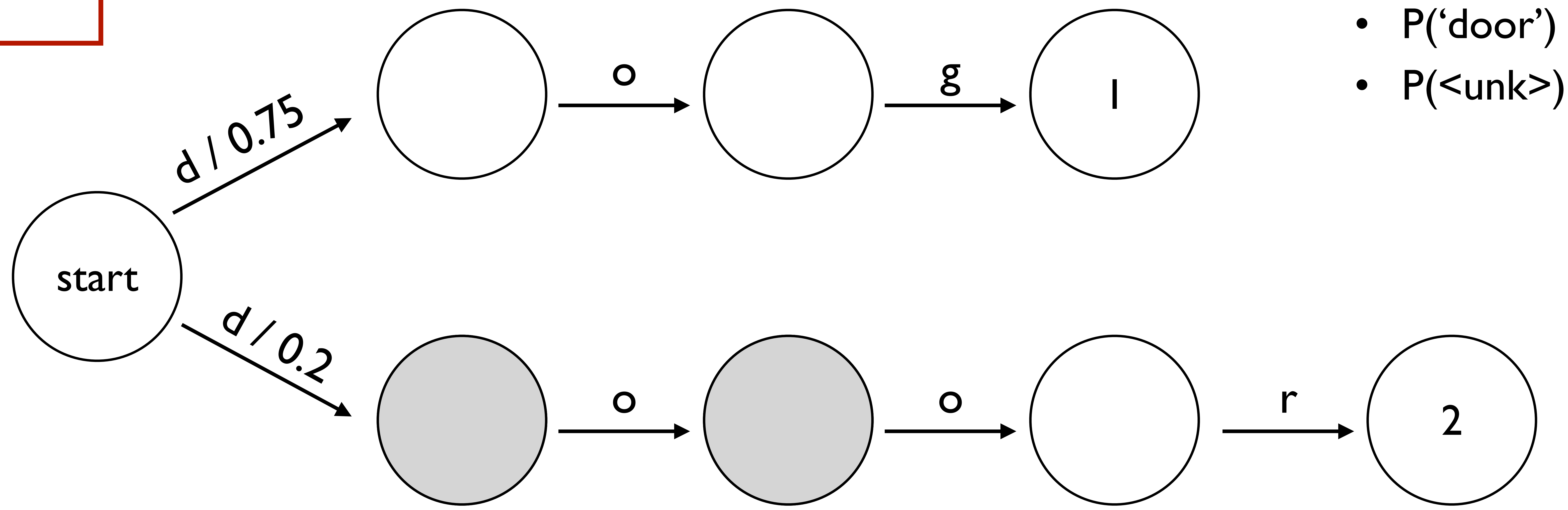
Example LM with two words:

- P('dog') = 0.75
- P('door') = 0.2
- P(<unk>) = 0.05

Scoring at the character-level

Output:

d o o r



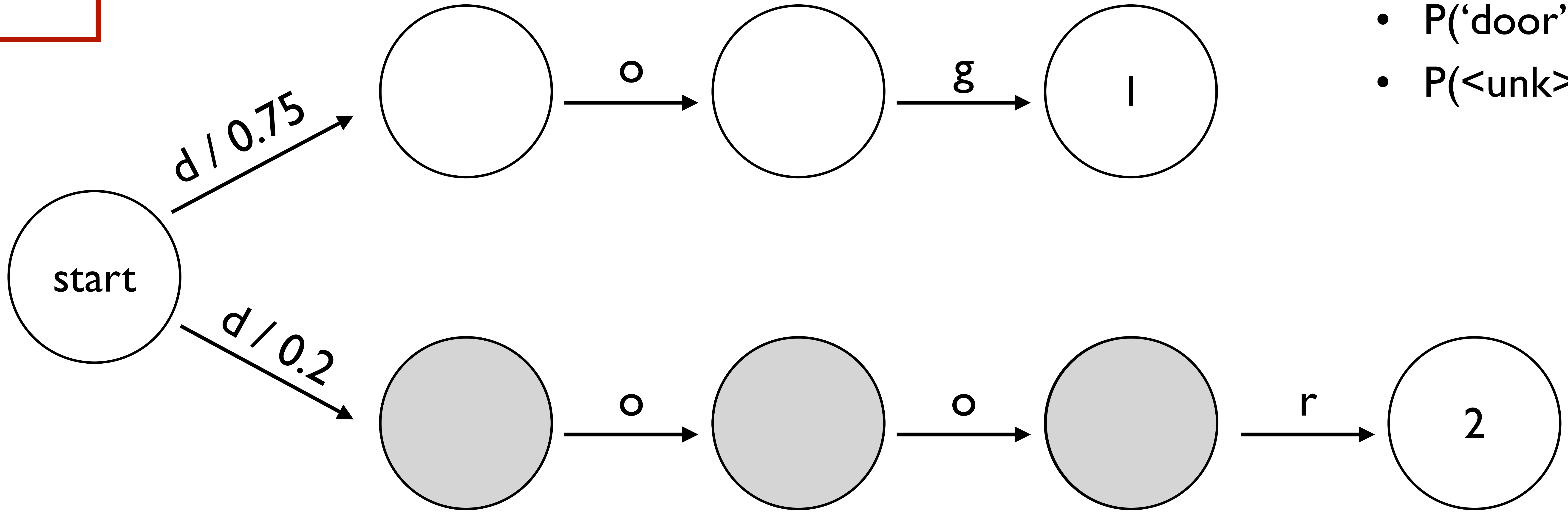
Example LM with two words:

- P('dog') = 0.75
- P('door') = 0.2
- P(<unk>) = 0.05

Scoring at the character-level

Output:

d o o r



Example LM with two words:

- P('dog') = 0.75
- P('door') = 0.2
- P(<unk>) = 0.05

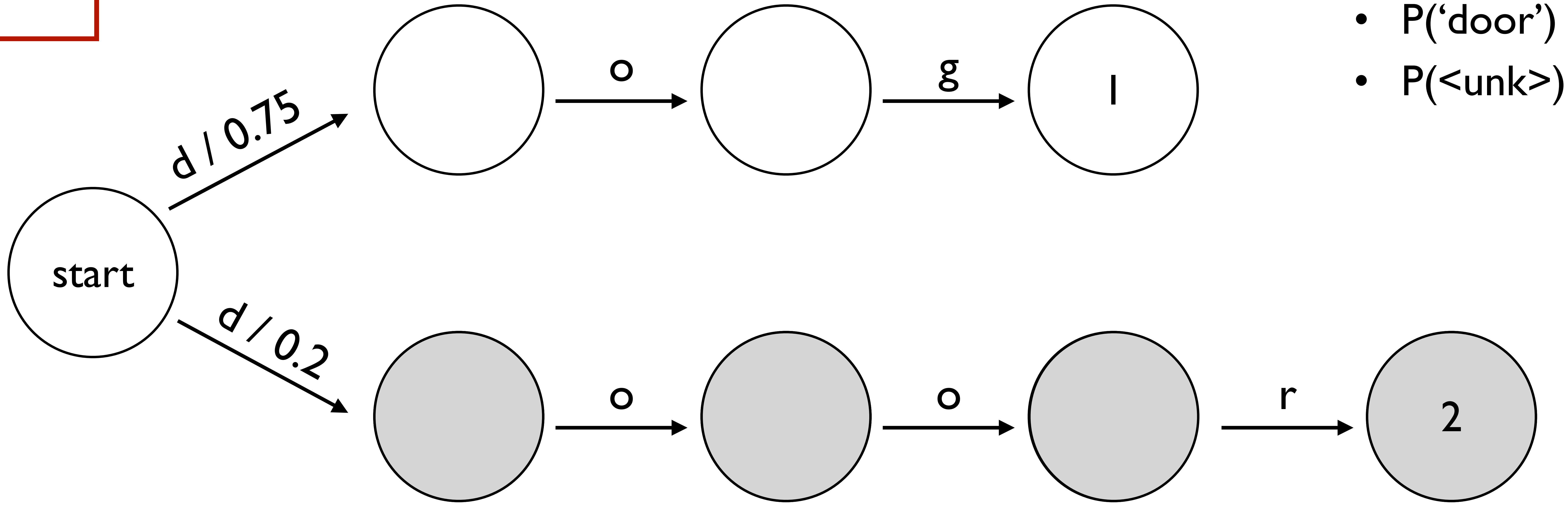
Scoring at the character-level

Output:

d o o r

Example LM with two words:

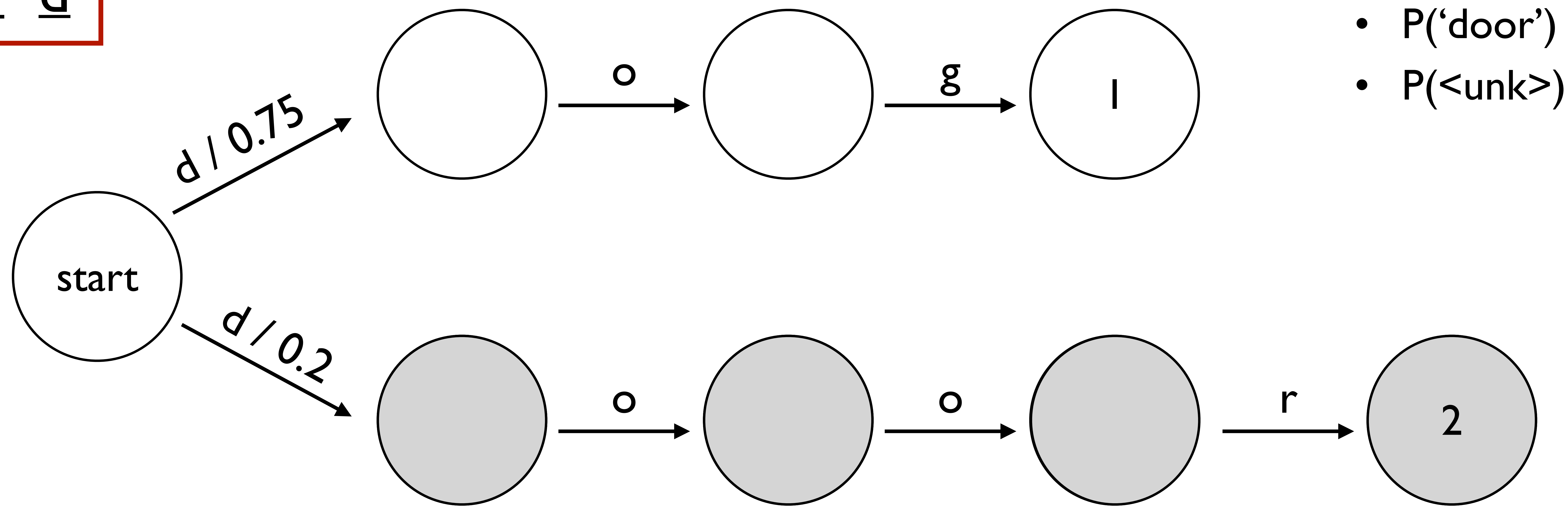
- P('dog') = 0.75
- P('door') = 0.2
- P(<unk>) = 0.05



Scoring at the character-level

Output:

d o o r _ d



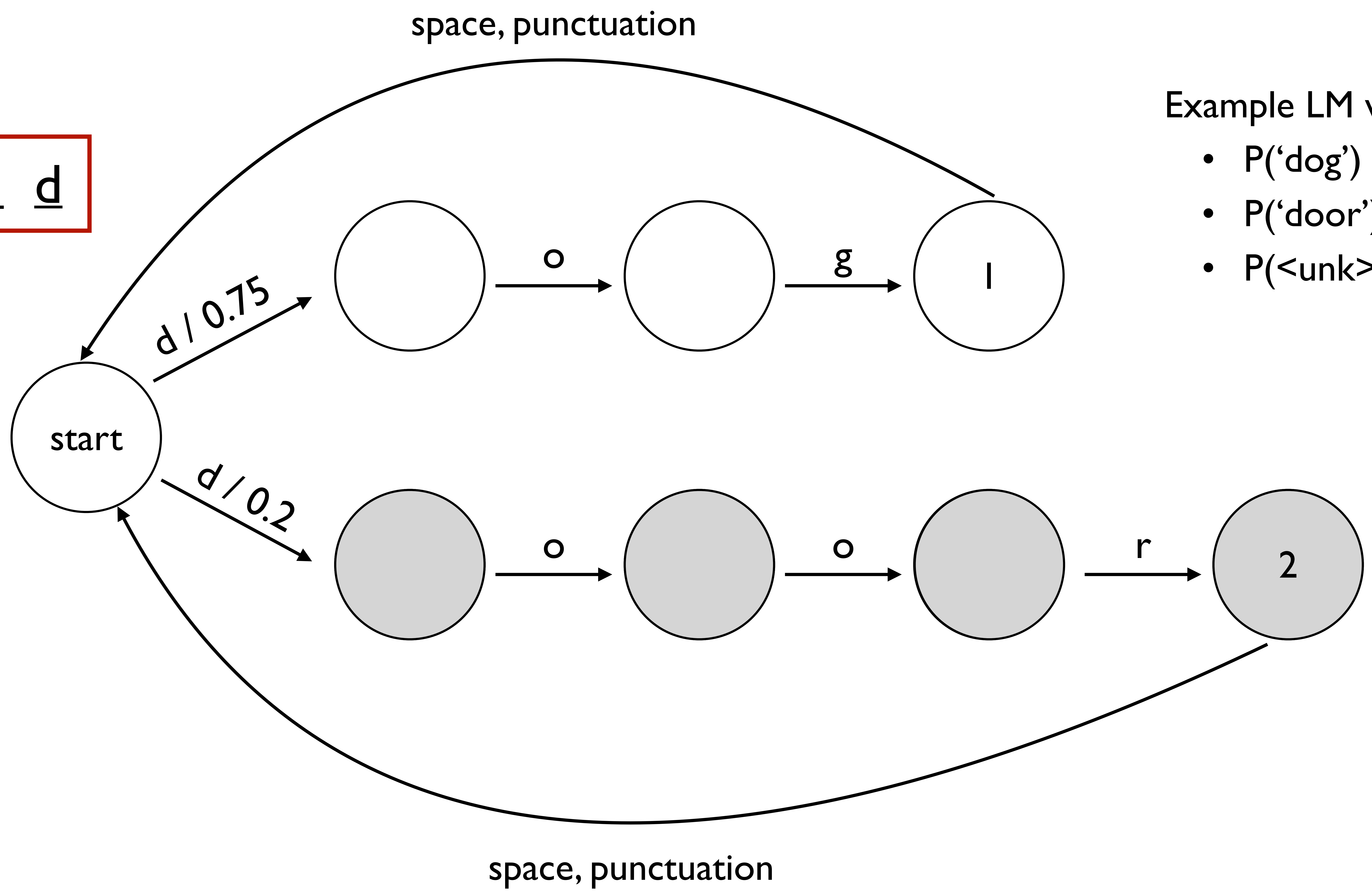
Example LM with two words:

- P('dog') = 0.75
- P('door') = 0.2
- P(<unk>) = 0.05

Scoring at the character-level

Output:

d o o r _ d



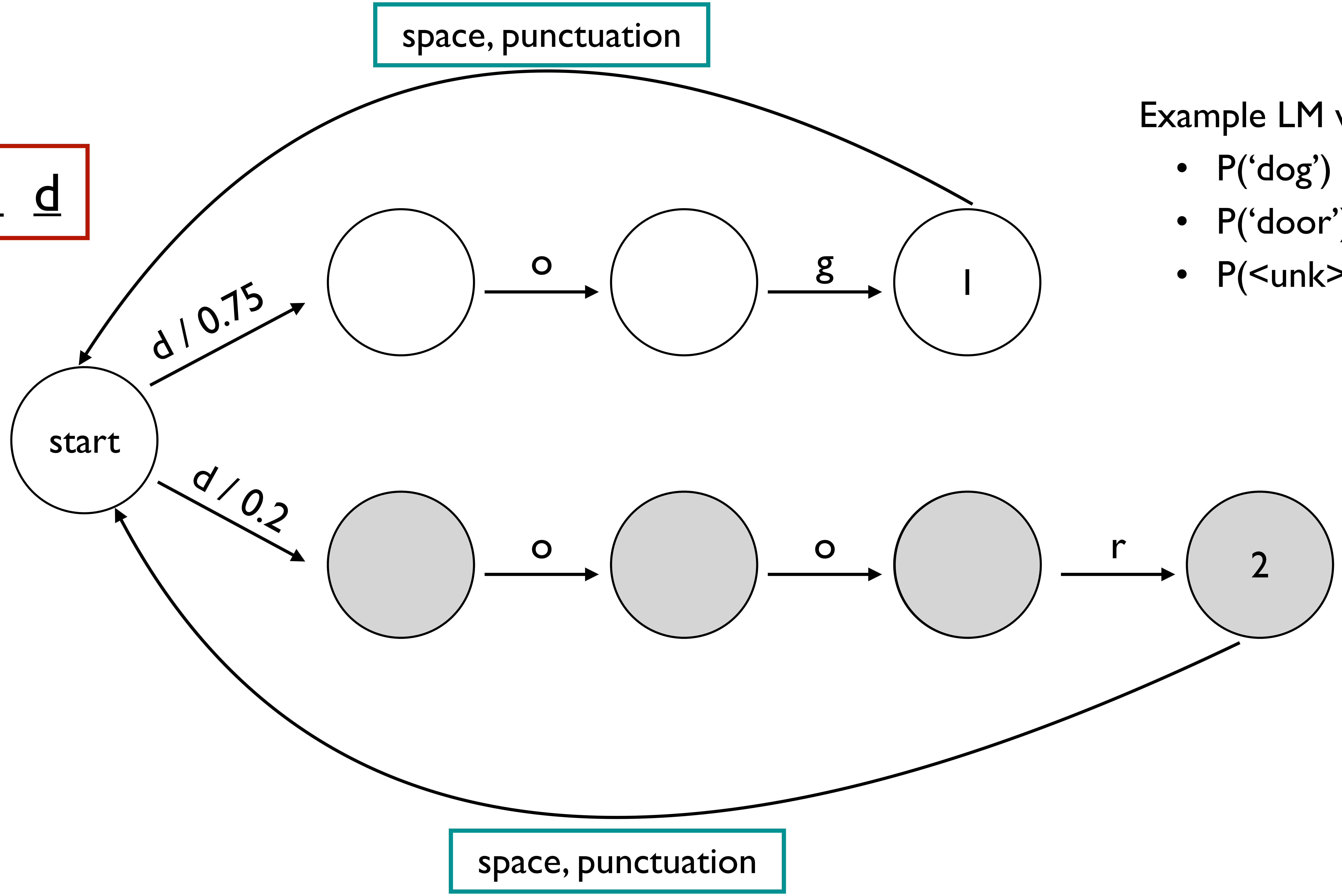
Example LM with two words:

- P(‘dog’) = 0.75
- P(‘door’) = 0.2
- P(<unk>) = 0.05

Scoring at the character-level

Output:

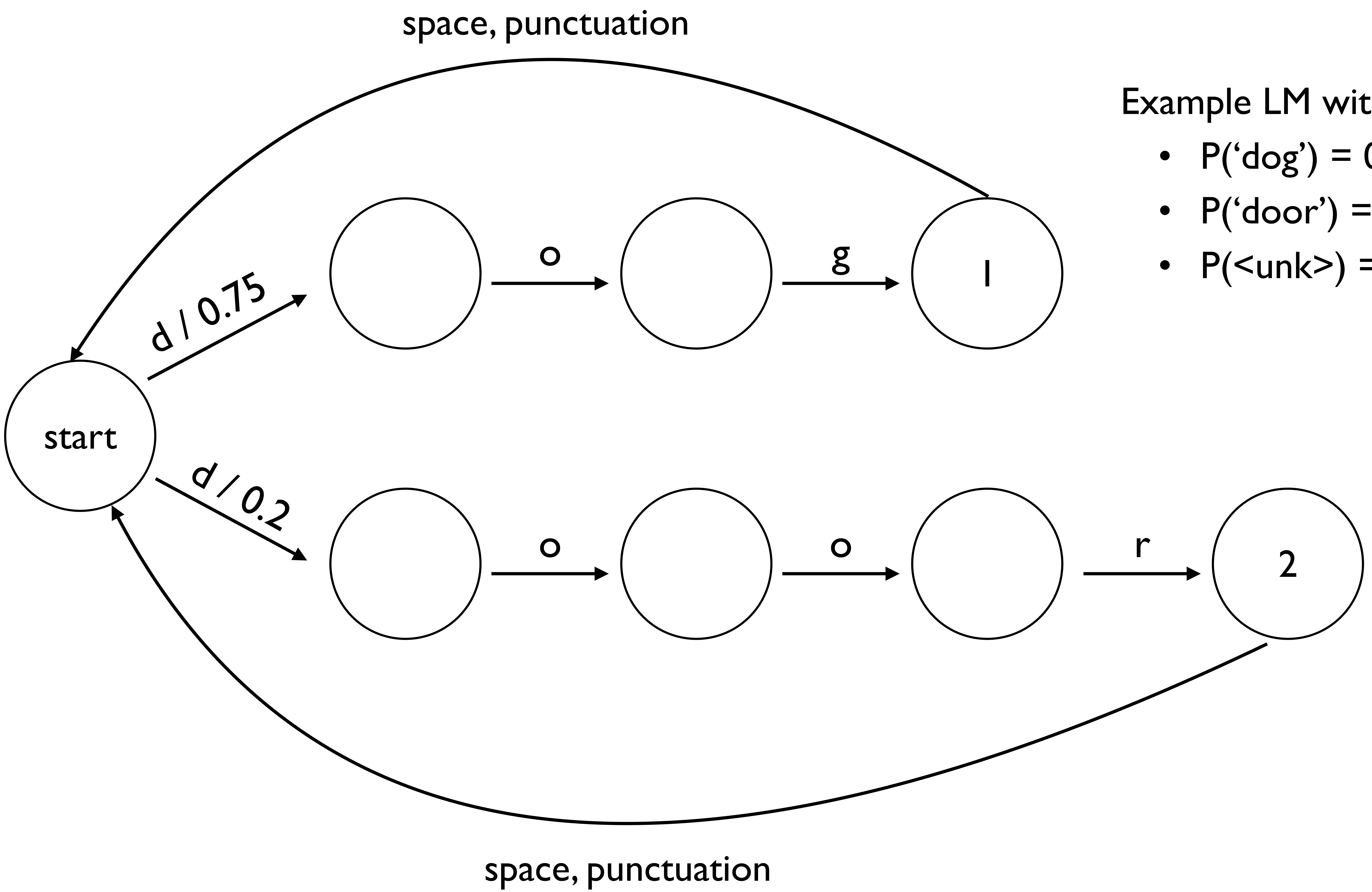
d o o r _ d



Example LM with two words:

- P('dog') = 0.75
- P('door') = 0.2
- P(<unk>) = 0.05

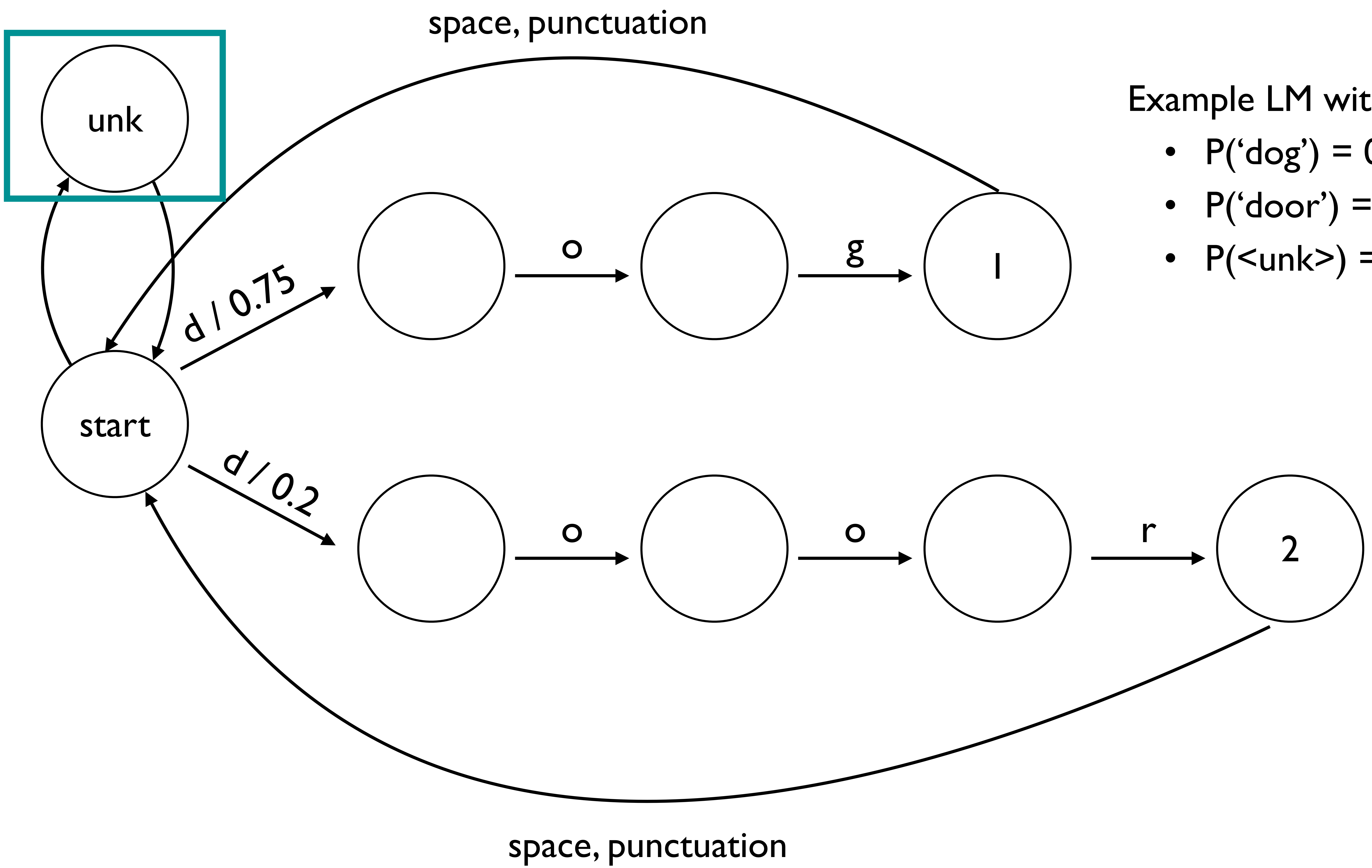
Scoring at the character-level



Example LM with two words:

- P('dog') = 0.75
- P('door') = 0.2
- P(<unk>) = 0.05

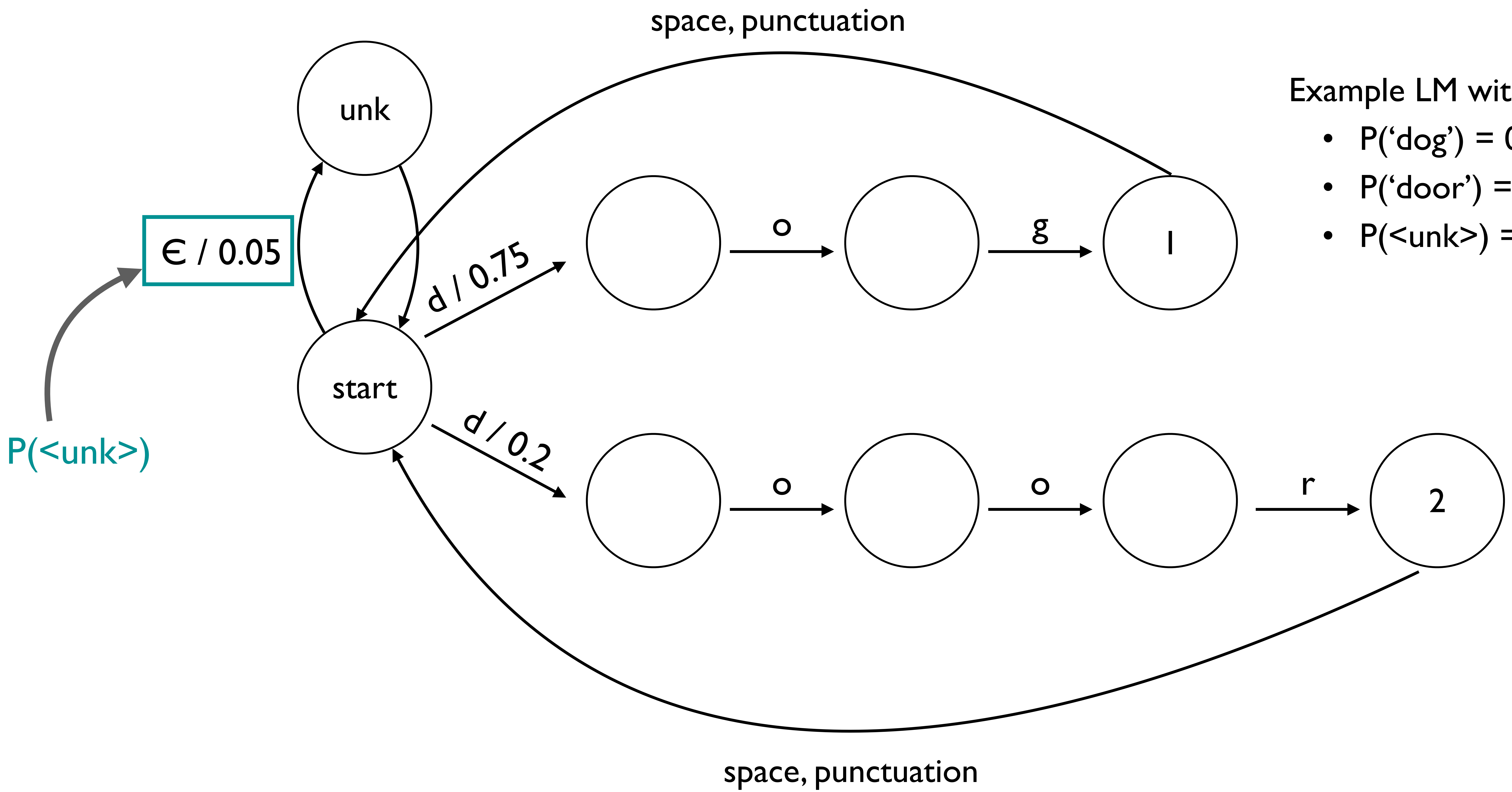
Scoring at the character-level



Example LM with two words:

- P('dog') = 0.75
- P('door') = 0.2
- P(<unk>) = 0.05

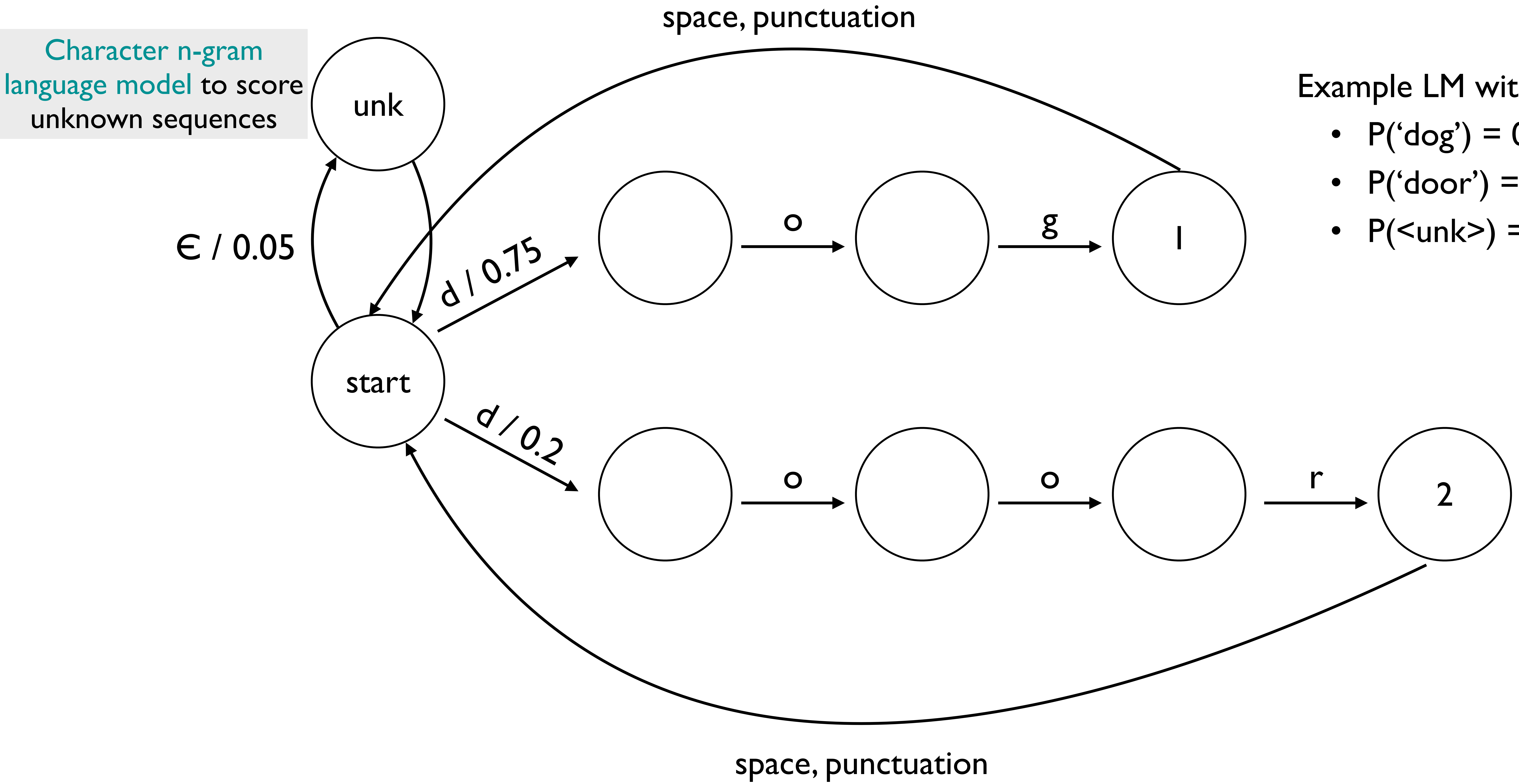
Scoring at the character-level



Example LM with two words:

- P('dog') = 0.75
- P('door') = 0.2
- P(<unk>) = 0.05

Scoring at the character-level

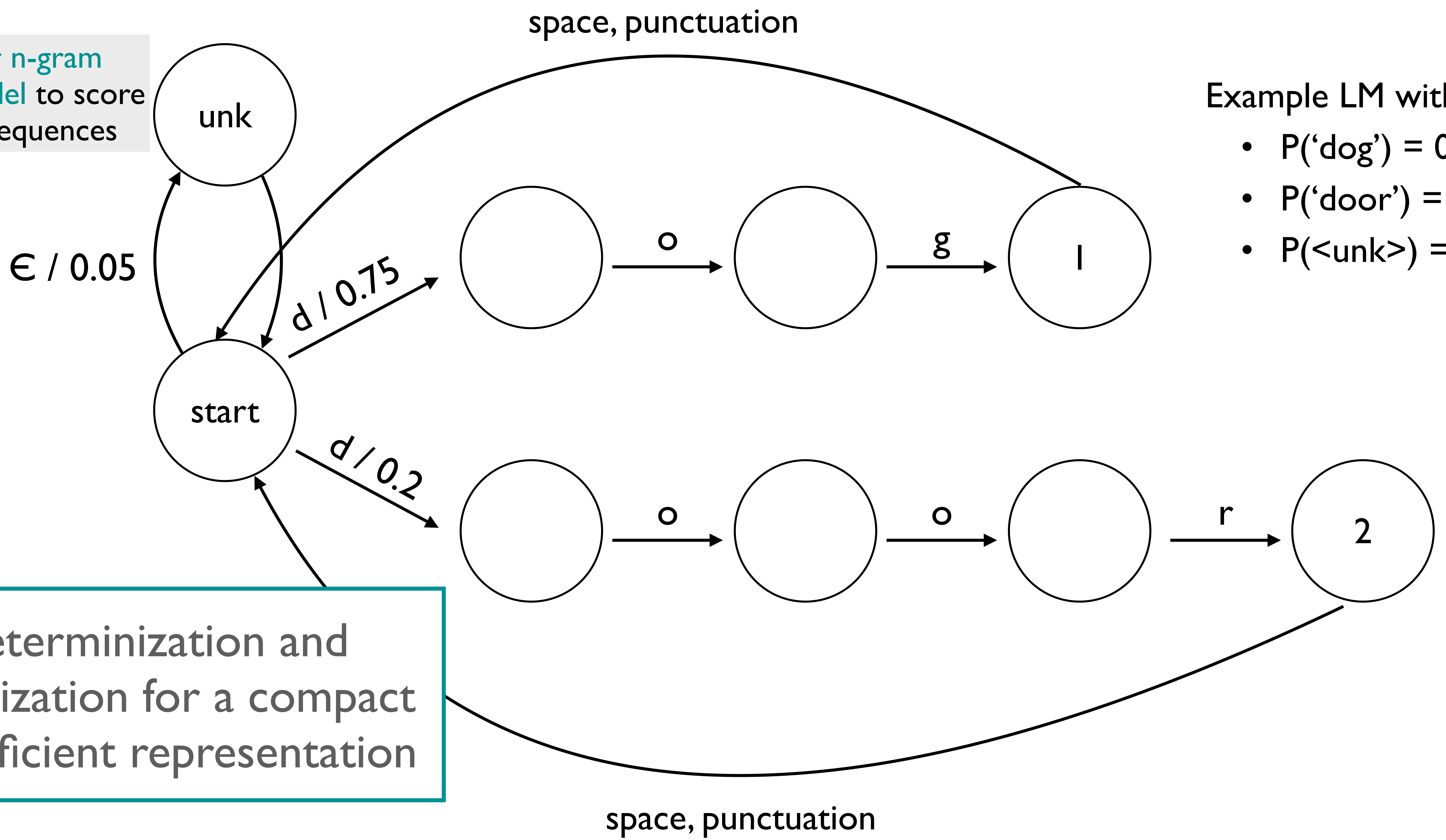


Example LM with two words:

- P('dog') = 0.75
- P('door') = 0.2
- P(<unk>) = 0.05

Scoring at the character-level

Character n-gram language model to score unknown sequences



Determinization and minimization for a compact and efficient representation

Example LM with two words:

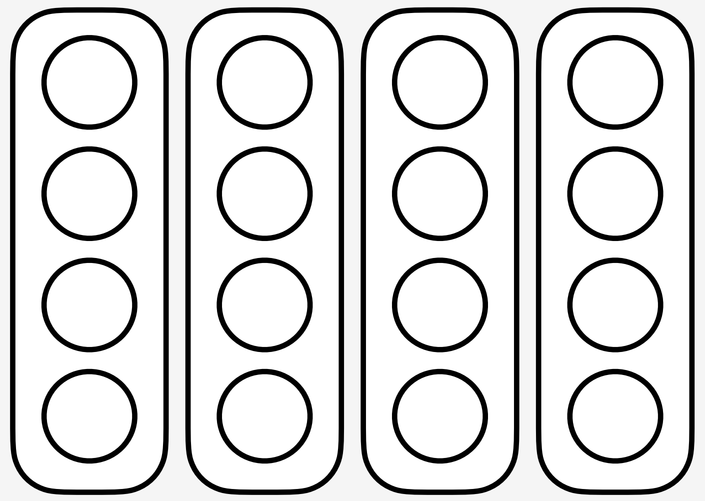
- P('dog') = 0.75
- P('door') = 0.2
- P(<unk>) = 0.05

Lexically-aware decoding for post-correction

Lexically-aware decoding for post-correction

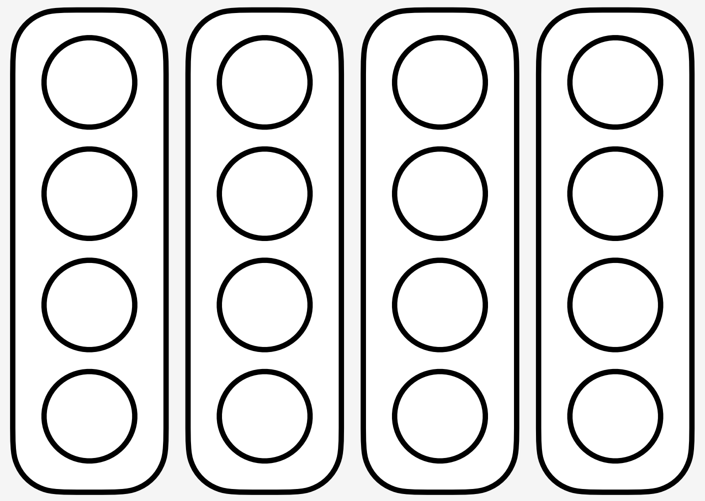
$$P(y) = \quad p_{\text{lstm}}(y) \qquad\qquad\qquad p_{\text{freq}}(y)$$

Lexically-aware decoding for post-correction



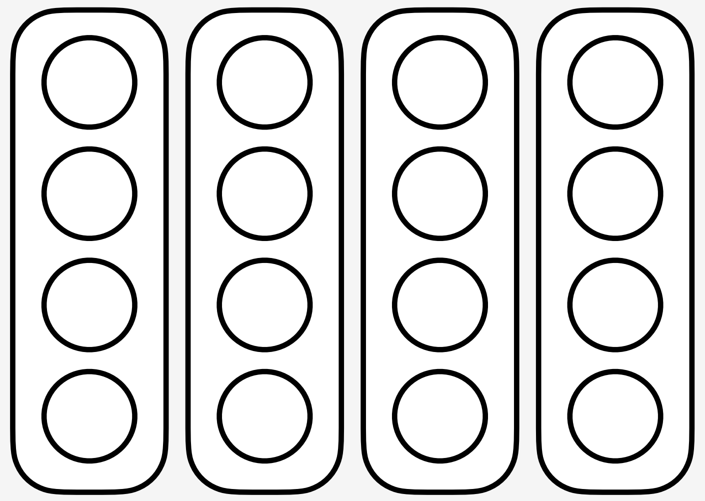
$$P(y) = p_{\text{lstm}}(y) \qquad p_{\text{freq}}(y)$$

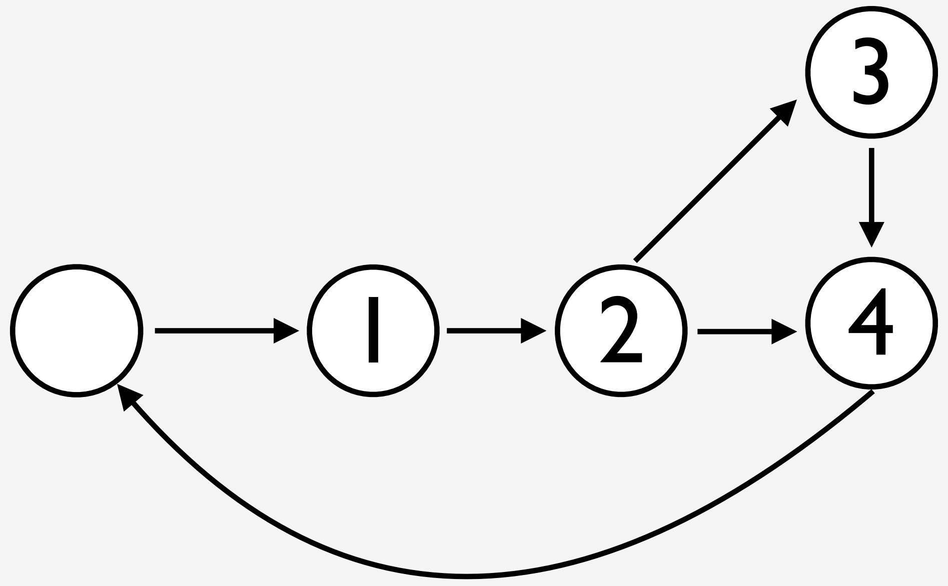
Lexically-aware decoding for post-correction



$$P(y) = p_{\text{lstm}}(y) \quad \boxed{p_{\text{freq}}(y)}$$

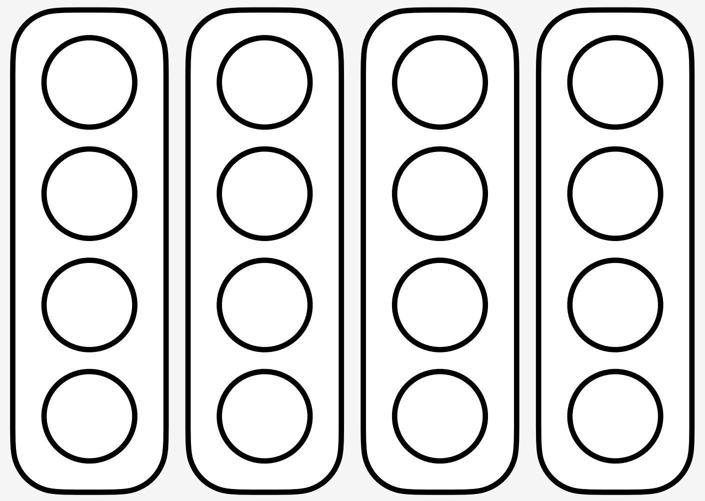
Lexically-aware decoding for post-correction

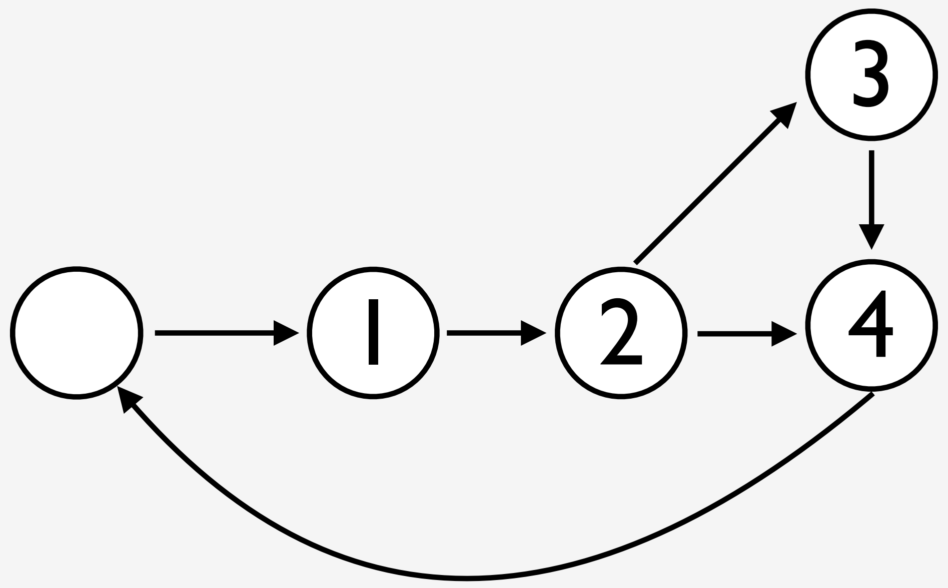




$$P(y) = p_{\mathrm{lstm}}(y) \quad p_{\mathrm{wfsa}}(y)$$

Lexically-aware decoding for post-correction



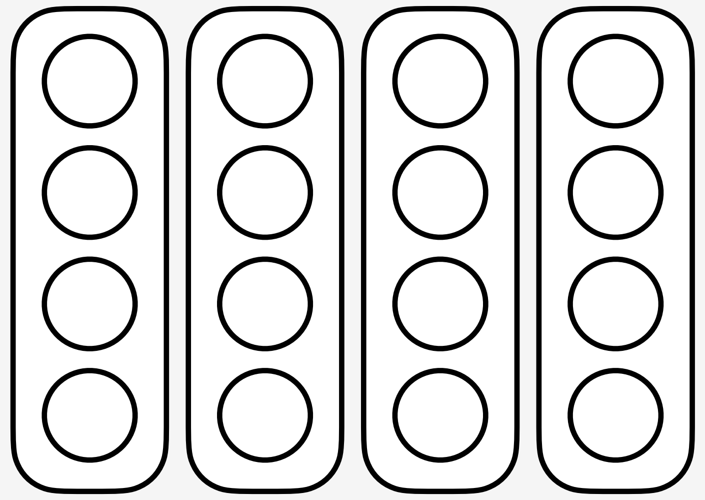


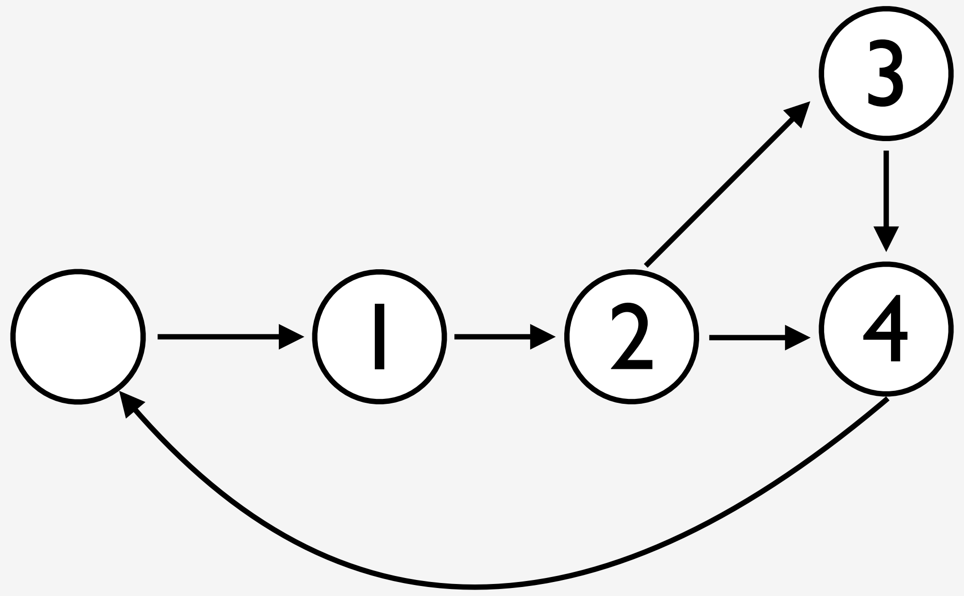
$$P(y) = p_{\text{lstm}}(y) \quad p_{\text{wfsa}}(y)$$

WFSA representation gives character-level scores

Lexically-aware decoding for post-correction

Linear interpolation to combine the probabilities for joint inference

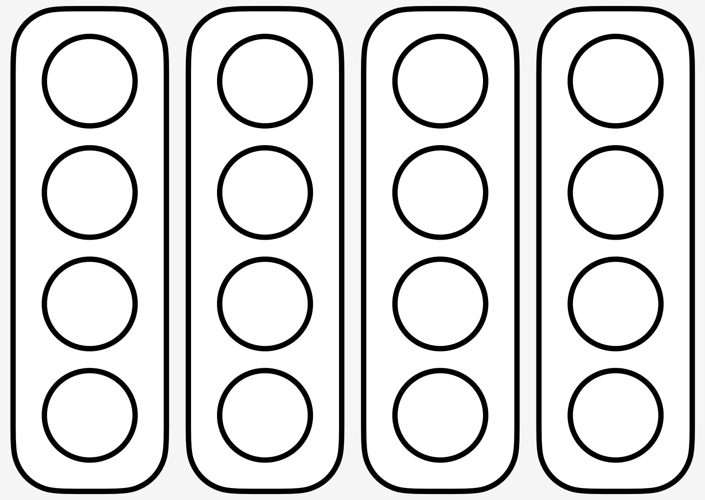


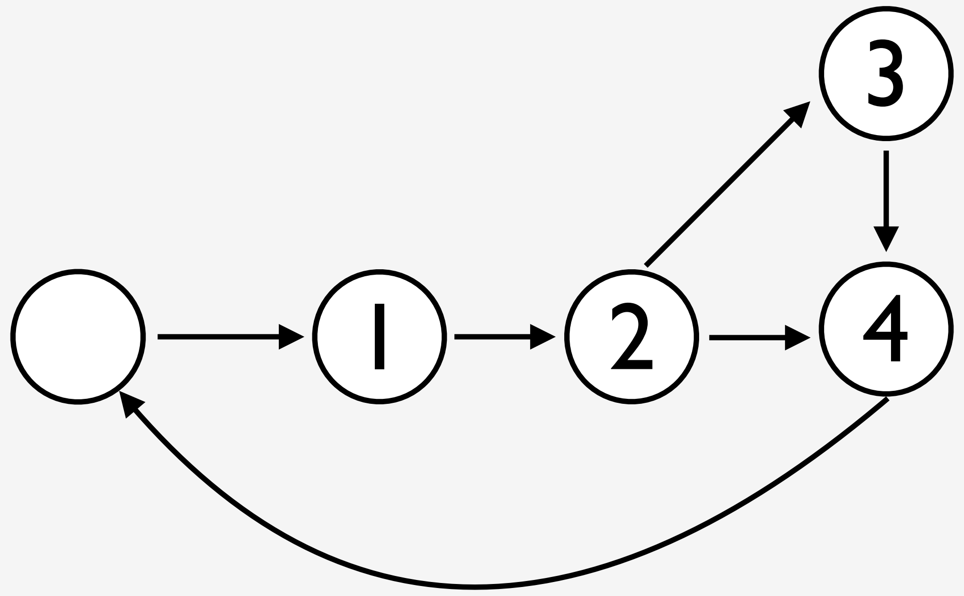


$$P(y) = (1 - \lambda) * p_{\mathrm{lstm}}(y) + \lambda * p_{\mathrm{wfsa}}(y)$$

Lexically-aware decoding for post-correction

Linear interpolation to combine the probabilities for joint inference



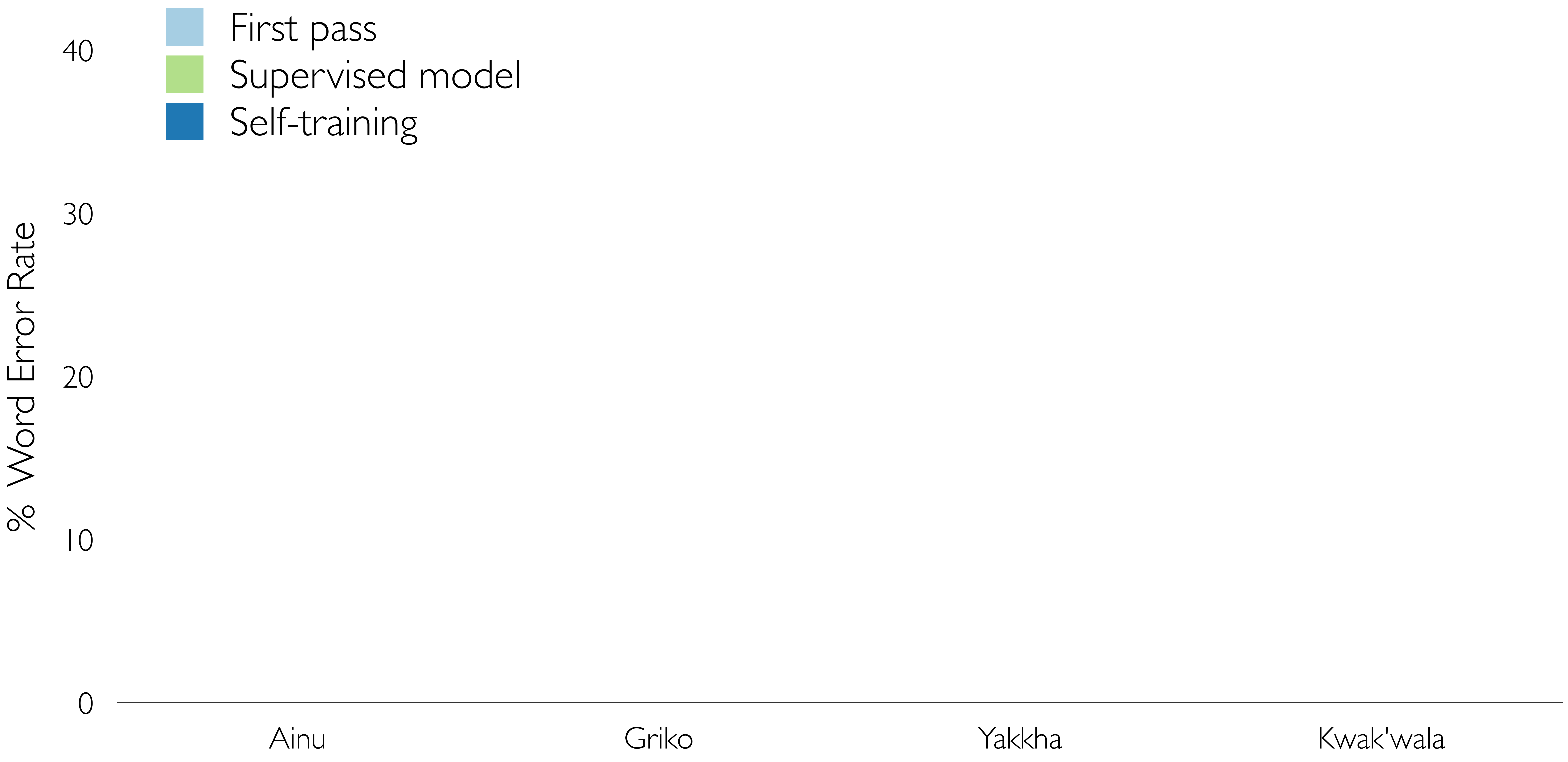


$$P(y) = \boxed{(1-\lambda)} * p_{\text{lstm}}(y) + \boxed{\lambda} * p_{\text{wfsa}}(y)$$

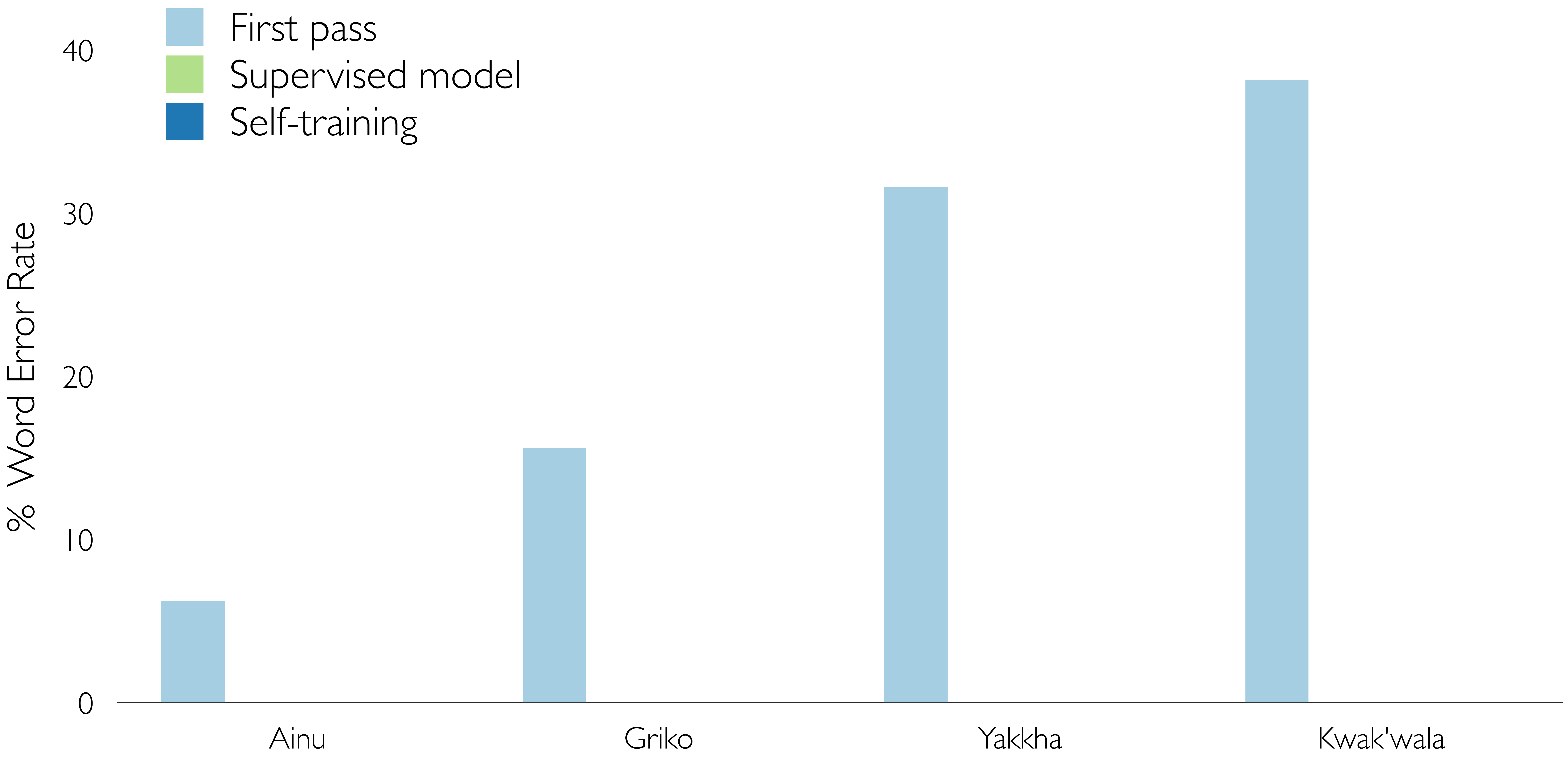
Hyperparameter for
linear interpolation

Experiments: does self-training improve performance?

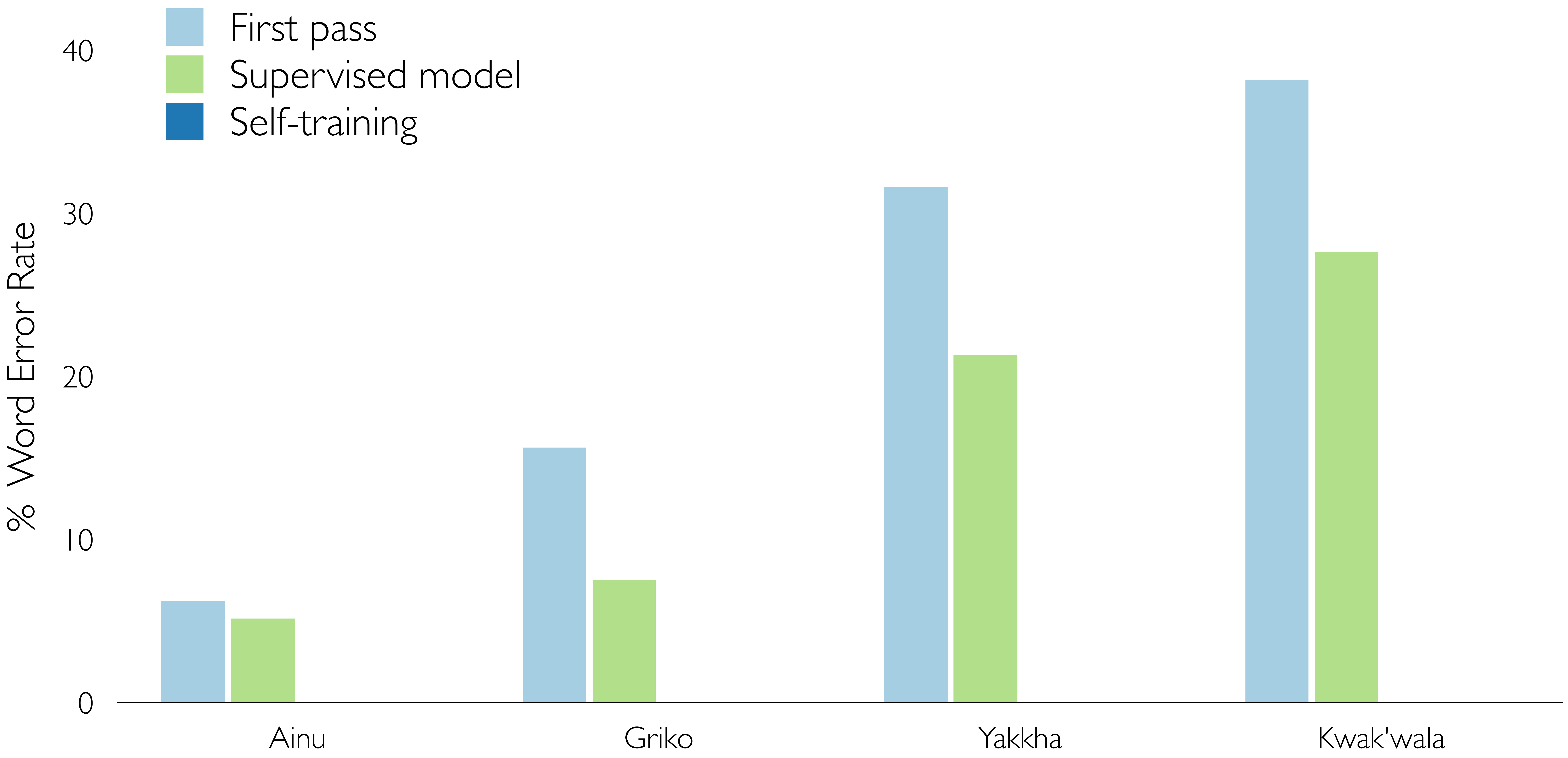
Experiments: does self-training improve performance?



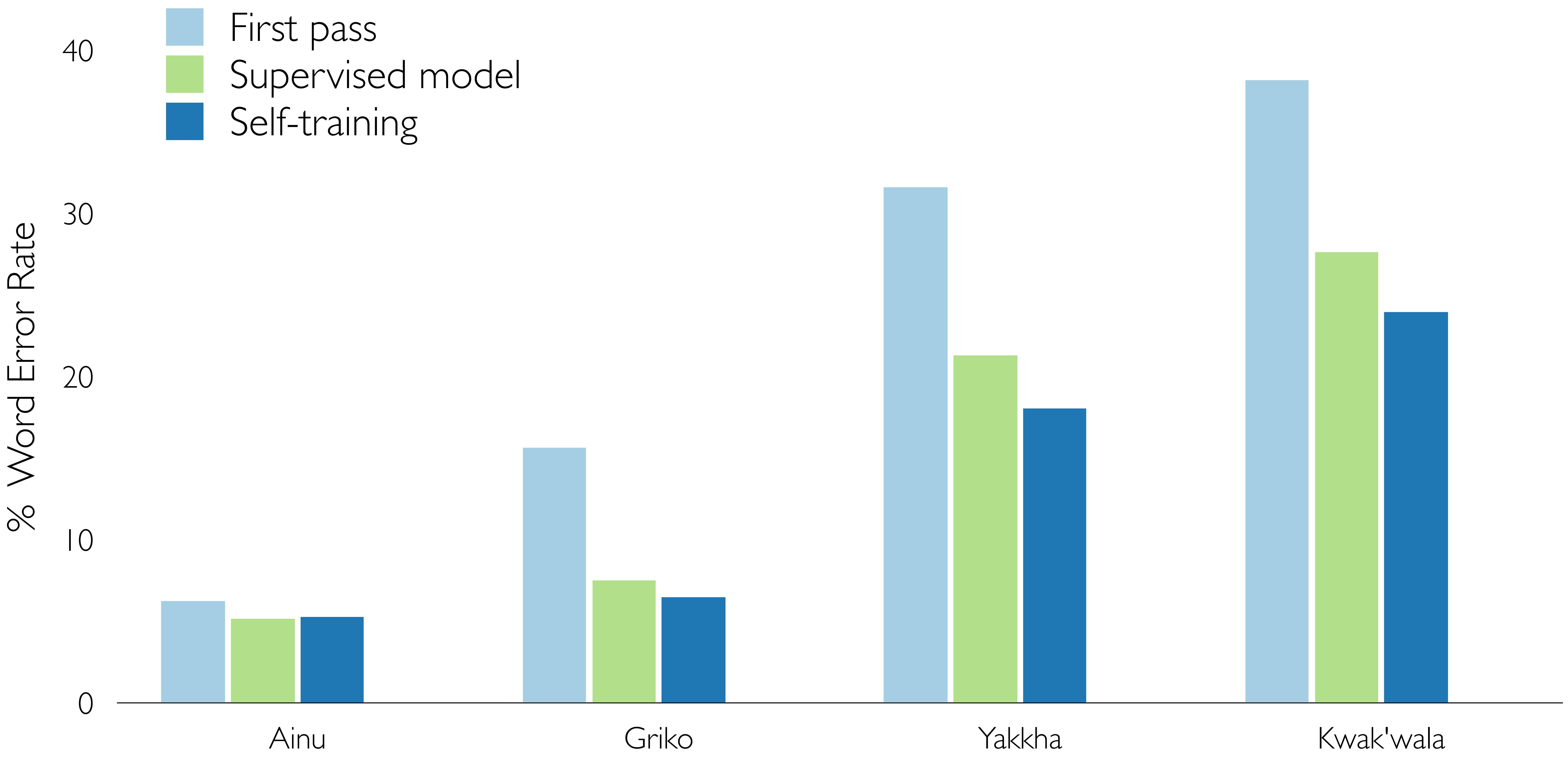
Experiments: does self-training improve performance?



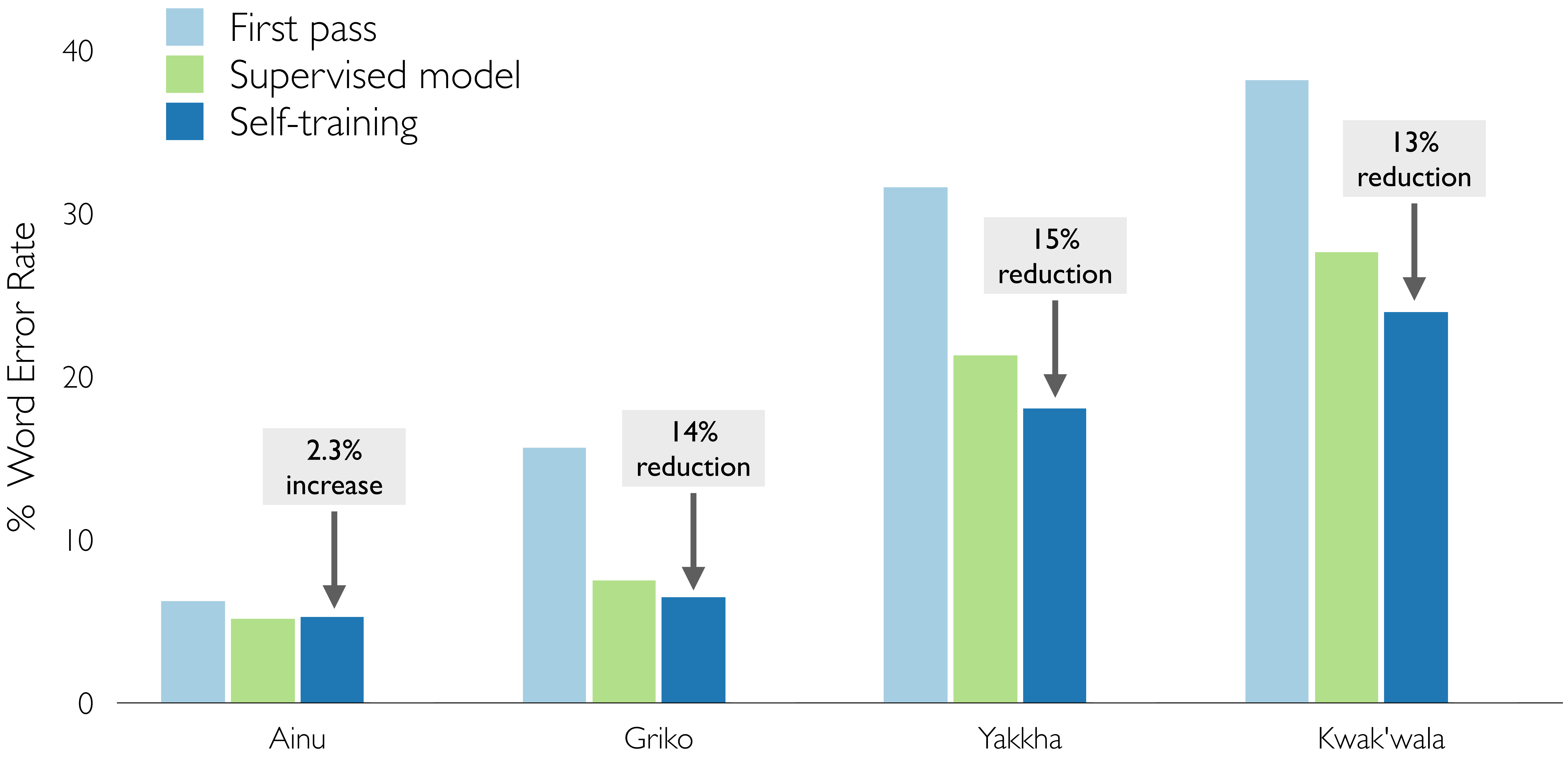
Experiments: does self-training improve performance?



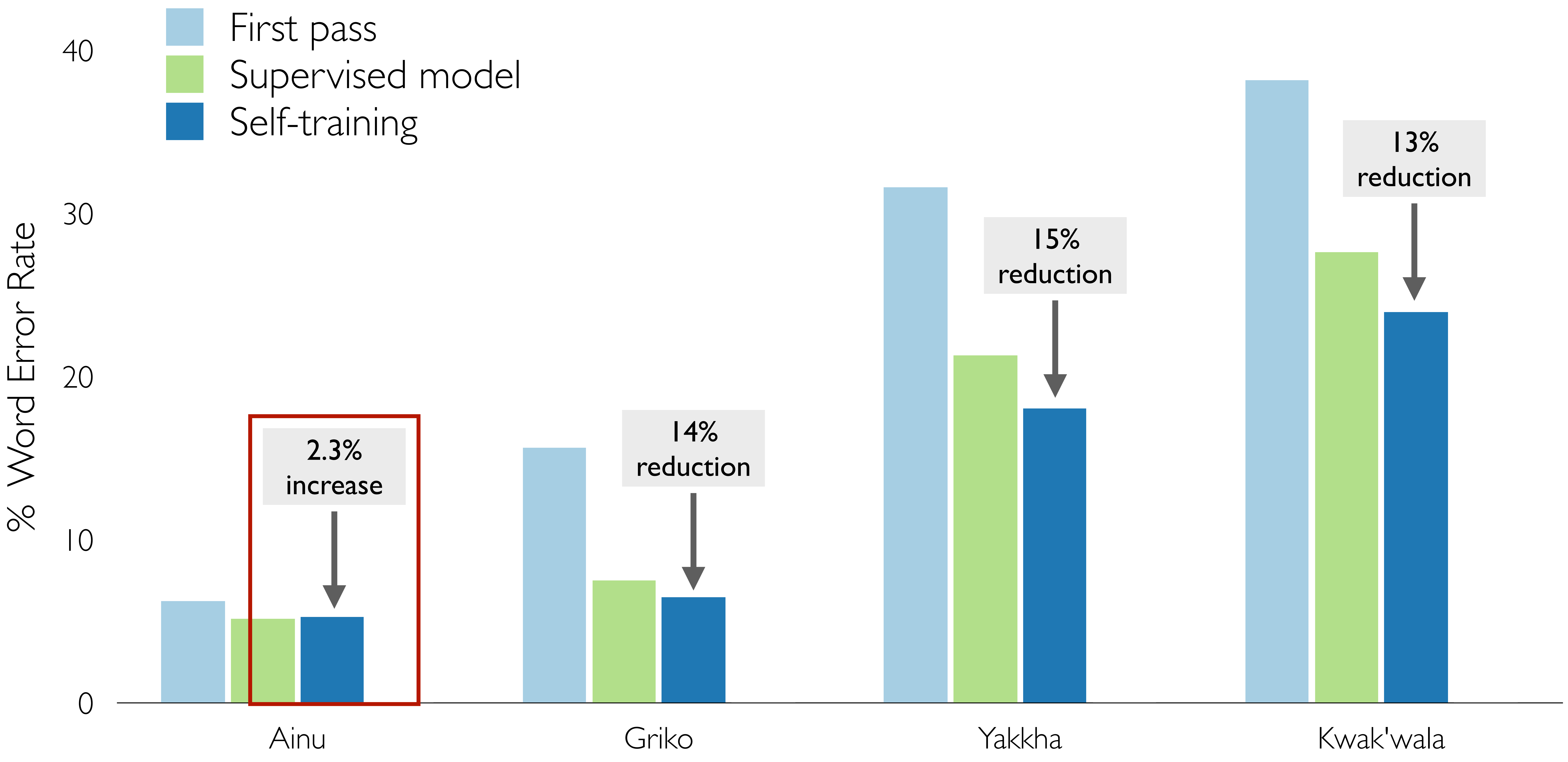
Experiments: does self-training improve performance?



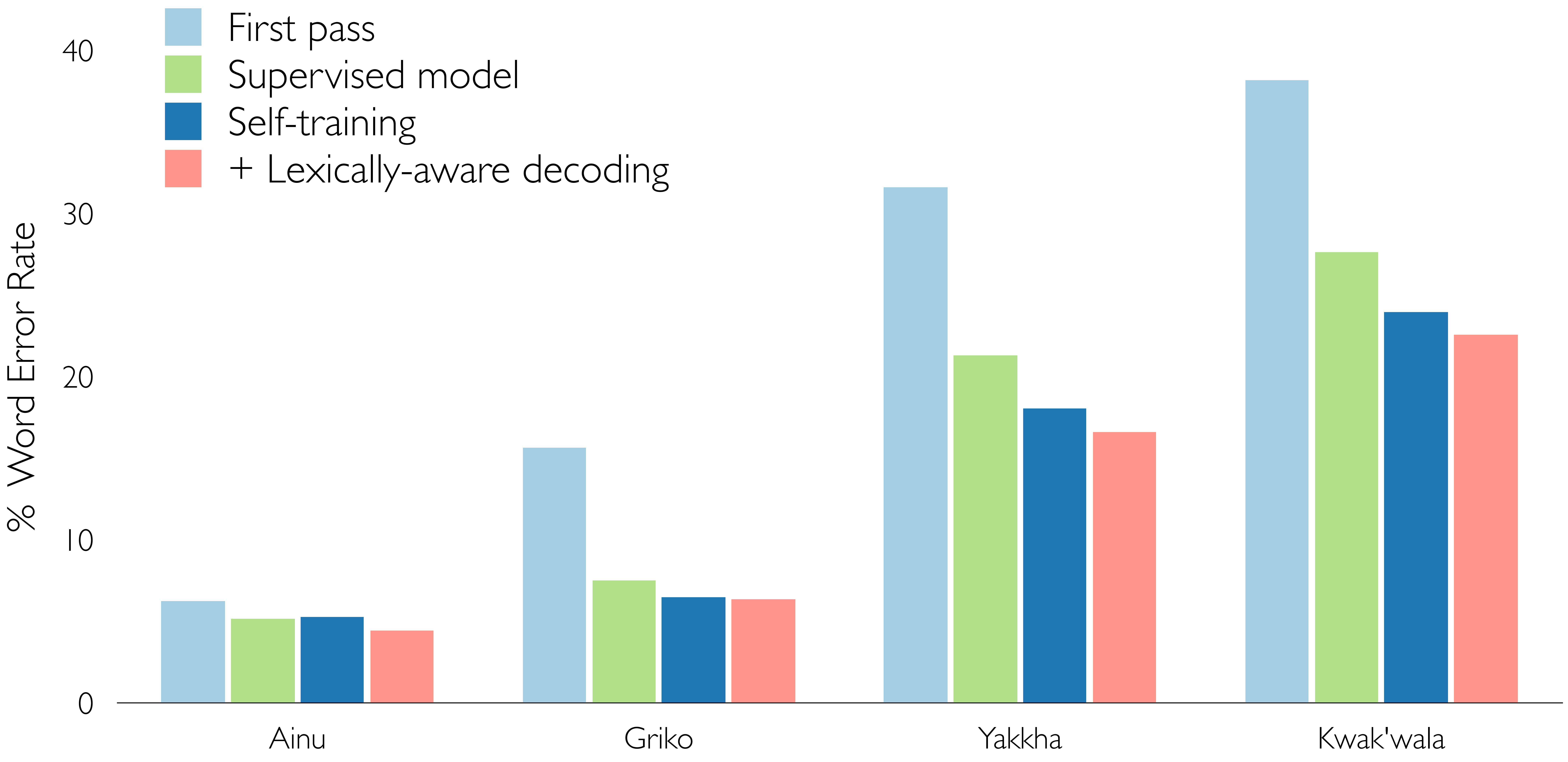
Experiments: does self-training improve performance?



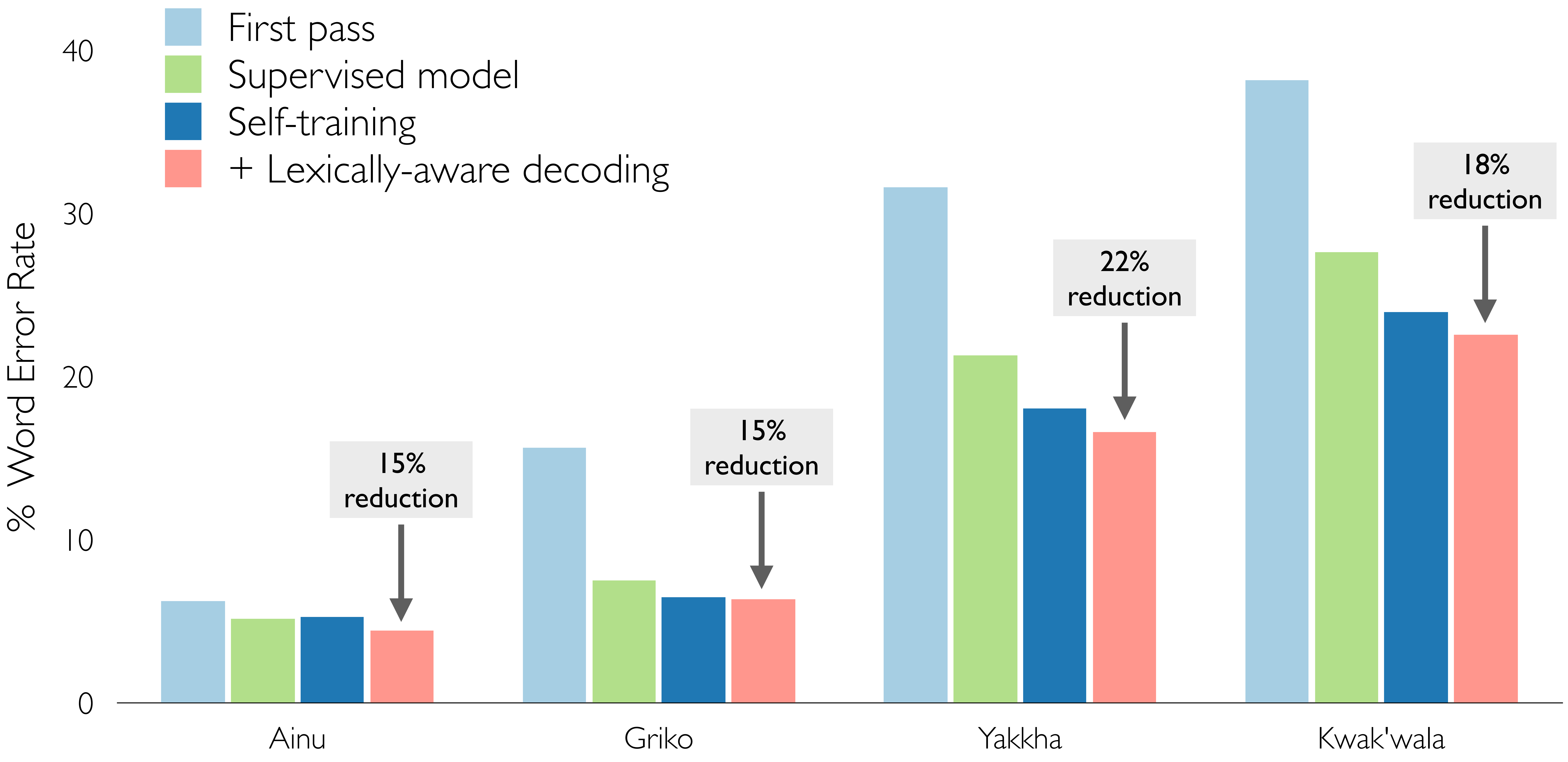
Experiments: does self-training improve performance?



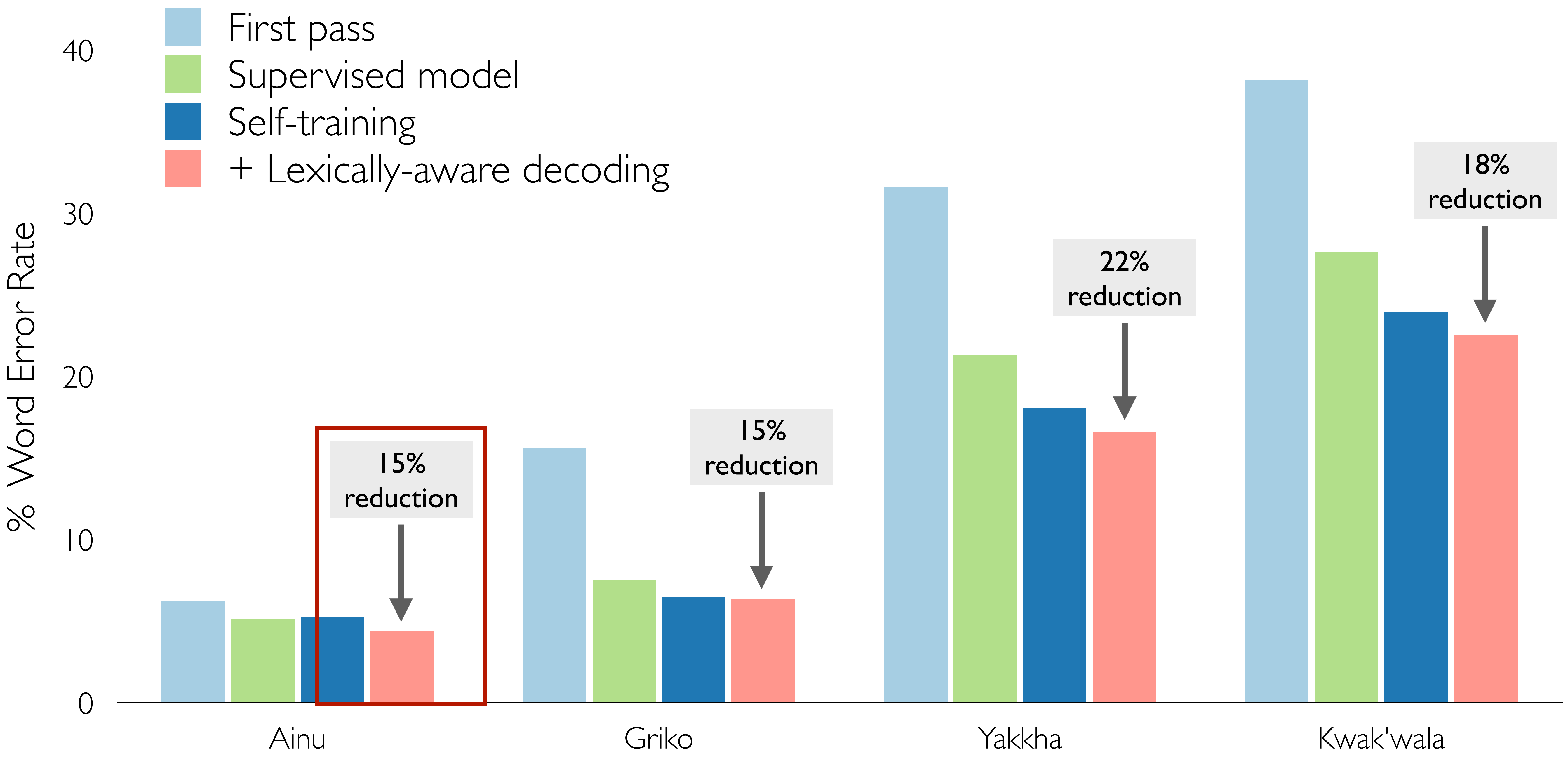
Experiments: does lexically-aware decoding counteract noise?



Experiments: does lexically-aware decoding counteract noise?



Experiments: does lexically-aware decoding counteract noise?



Summary

Summary

- Thousands of languages do not have easily accessible text to build NLP models

Summary

- Thousands of languages do not have easily accessible text to build NLP models
  - Text data does exist in many of these languages!

Summary

- Thousands of languages do not have easily accessible text to build NLP models
  - Text data does exist in many of these languages!
  - Locked away in non-machine-readable formats like printed books

Summary

- Thousands of languages do not have easily accessible text to build NLP models
  - Text data does exist in many of these languages!
  - Locked away in non-machine-readable formats like printed books
- OCR post-correction improves text extraction in very low-resourced settings

Summary

- Thousands of languages do not have easily accessible text to build NLP models
  - Text data does exist in many of these languages!
  - Locked away in non-machine-readable formats like printed books

- OCR post-correction improves text extraction in very low-resourced settings

Multi-source model: ↓ WER 17% – 52%

Summary

- Thousands of languages do not have easily accessible text to build NLP models
  - Text data does exist in many of these languages!
  - Locked away in non-machine-readable formats like printed books

- OCR post-correction improves text extraction in very low-resourced settings

  Multi-source model: ↓WER 17% – 52%

  Semi-supervised with lexically-aware decoding: ↓WER 29% – 59%

Impact case study: Kwak’wala

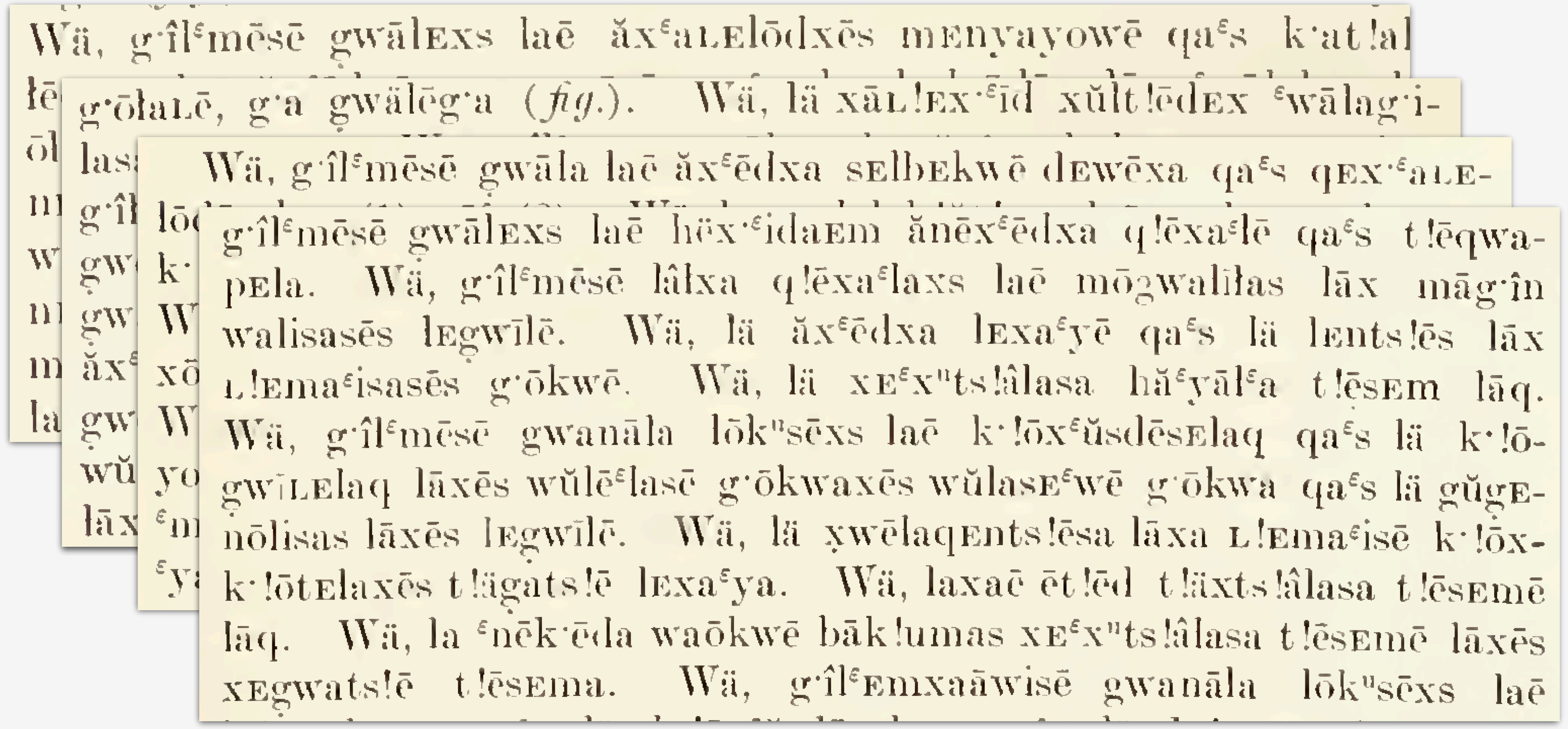
Impact case study: Kwak'wala

- Collaborating with documentary linguists and Kwak'wala speakers
  - Identify documents that would be most useful to extract text from

Impact case study: Kwak'wala

- Collaborating with documentary linguists and Kwak'wala speakers
  - Identify documents that would be most useful to extract text from

Boas texts: 10 volumes of Kwak'wala language and community documentation

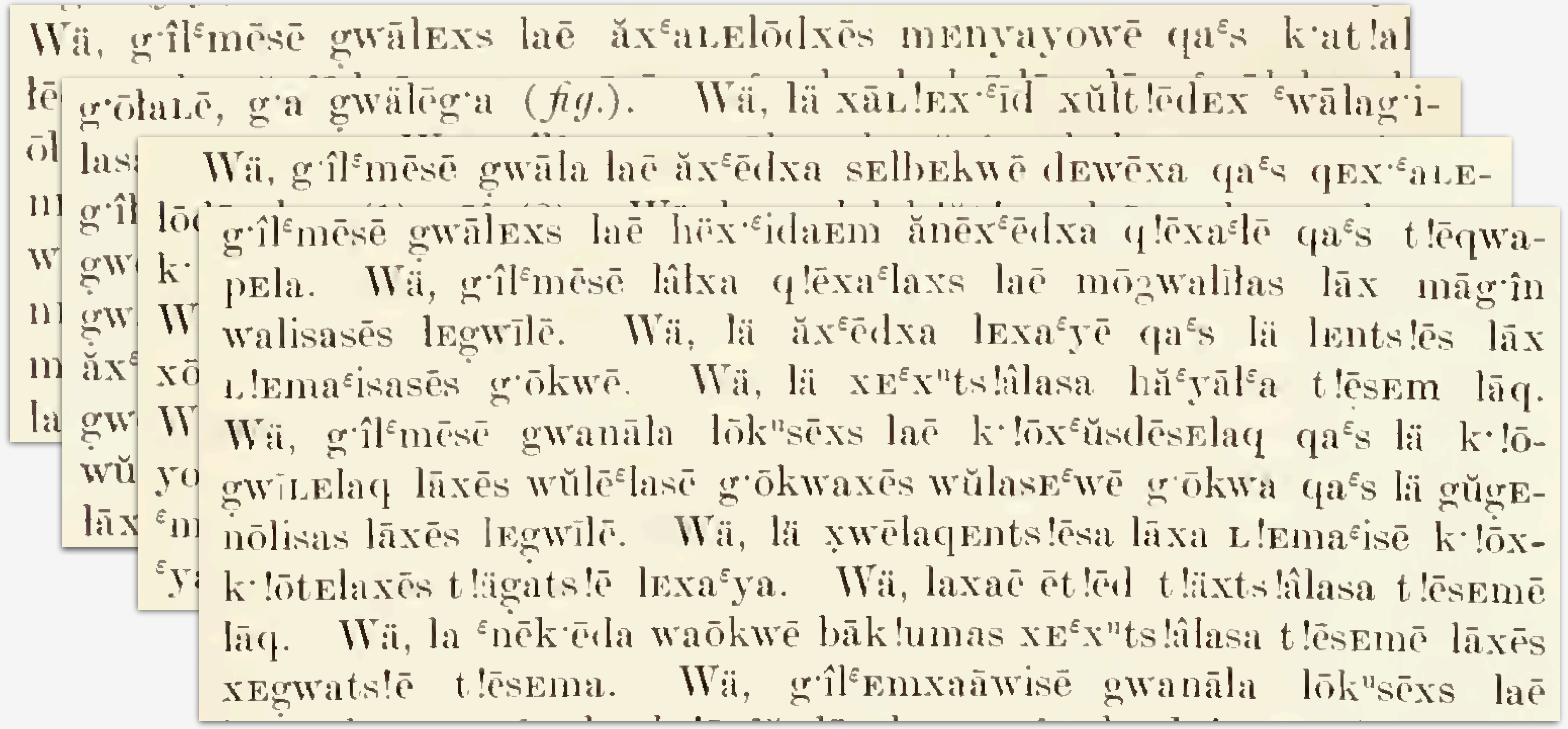

g·îlᵋmēsē gwālᴇxs laē hēx·ᵋidaᴇm ănēxᵋēdxa q!ēxaᵋlē qaᵋs t!ēqwa-
pᴇla. Wä, g·îlᵋmēsē lâlxa q!ēxaᵋlaxs laē mōgwalīłas lāx māg·în
walisasēs lᴇgwīlē. Wä, lä ăxᵋēdxa lᴇxaᵋyē qaᵋs lä lᴇnts!ēs lāx
ʟ!ᴇmaᵋisasēs g·ōkwē. Wä, lä xᴇᵋx"ts!âlasa hăᵋyālᵋa t!ēsᴇm lāq.
Wä, g·îlᵋmēsē gwanāla lōk"sēxs laē k·!ōxᵋŭsdēsᴇlaq qaᵋs lä k·!ō-
gwīʟᴇlaq lāxēs wŭlēᵋlasē g·ōkwaxēs wŭlasᴇᵋwē g·ōkwa qaᵋs lä gŭgᴇ-
nōlisas lāxēs lᴇgwīlē. Wä, lä xwēlaqᴇnts!ēsa lāxa ʟ!ᴇmaᵋisē k·!ōx-
k·!ōtᴇlaxēs t!ägats!ē lᴇxaᵋya. Wä, laxaē ēt!ēd t!äxts!âlasa t!ēsᴇmē
lāq. Wä, la ᵋnēk·ēda waōkwē băk!umas xᴇᵋx"ts!âlasa t!ēsᴇmē lāxēs
xᴇgwats!ē t!ēsᴇma. Wä, g·îlᵋᴇmxaāwisē gwanāla lōk"sēxs laē

Produced by Franz Boas in 1921

Impact case study: Kwak'wala

- Collaborating with documentary linguists and Kwak'wala speakers
  - Identify documents that would be most useful to extract text from

Boas texts: 10 volumes of Kwak'wala language and community documentation

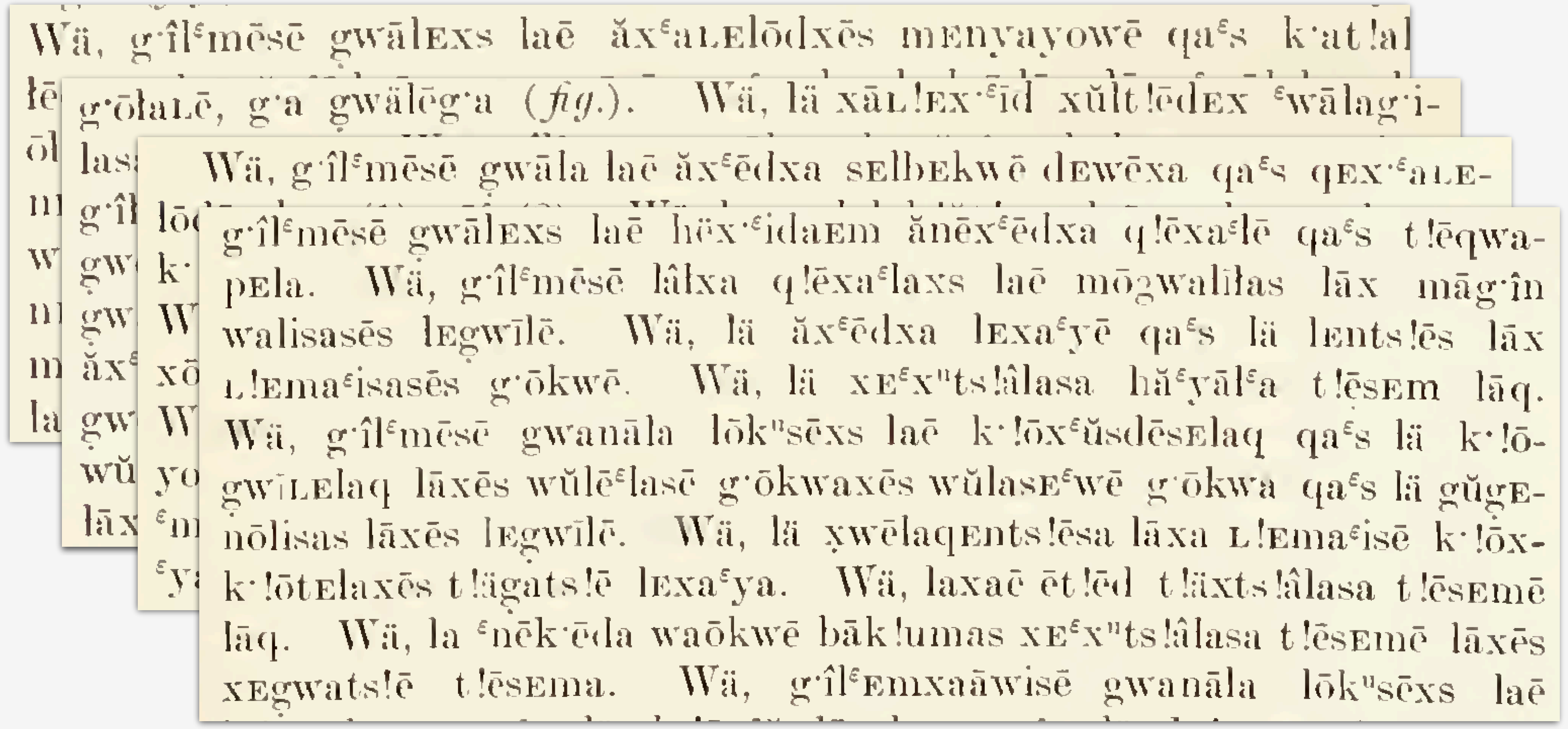

g·îlᵋmēsē gwālExs laē hëx·ᵋidaEm ănēxᵋēdxa q!ēxaᵋlē qaᵋs t!ēqwa-
pEla. Wä, g·îlᵋmēsē lâlxa q!ēxaᵋlaxs laē mōgwalīłas lāx māg·în
walisasēs łEgwīlē. Wä, lä ăxᵋēdxa łExaᵋyē qaᵋs lä łEnts!ēs lāx
L!Emaᵋisasēs g·ōkwē. Wä, lä xEᵋx·ts!âlasa hăᵋyālᵋa t!ēsEm lāq.
Wä, g·îlᵋmēsē gwanāla lōkᵘsēxs laē k·!ōxᵋŭsdēsElaq qaᵋs lä k·!ō-
gwiLElaq lāxēs wŭlēᵋlasē g·ōkwaxēs wŭlasEᵋwē g·ōkwa qaᵋs lä gŭgE-
nōlisas lāxēs łEgwīlē. Wä, lä xwēlaqEnts!ēsa lāxa L!Emaᵋisē k·!ōx-
k·!ōtElaxēs t!ägats!ē łExaᵋya. Wä, laxaē ēt!ēd t!äxts!âlasa t!ēsEmē
lāq. Wä, la ᵋnēk·ēda waōkwē bāk!umas xEᵋx·ts!âlasa t!ēsEmē lāxēs
xEgwats!ē t!ēsEma. Wä, g·îlᵋEmxaāwisē gwanāla lōkᵘsēxs laē

- Tremendous cultural and linguistic value!

Impact case study: Kwak'wala

- Collaborating with documentary linguists and Kwak'wala speakers
  - Identify documents that would be **most useful to extract text from**

Boas texts: 10 volumes of Kwak'wala language and community documentation

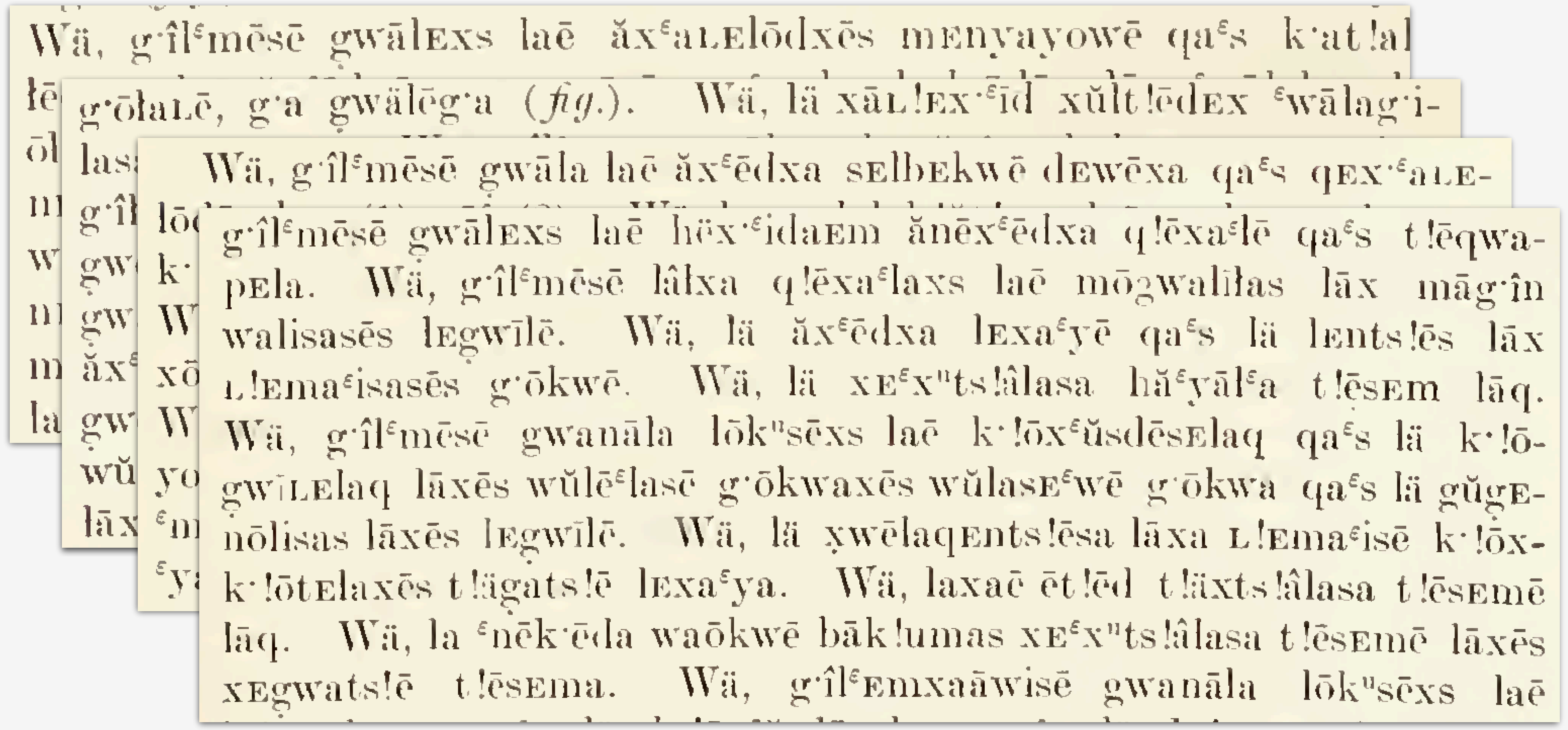


- Tremendous cultural and linguistic value!
- Minimally accessible to researchers

Impact case study: Kwak'wala

- Collaborating with documentary linguists and Kwak'wala speakers
  - Identify documents that would be most useful to extract text from

Boas texts: 10 volumes of Kwak'wala language and community documentation

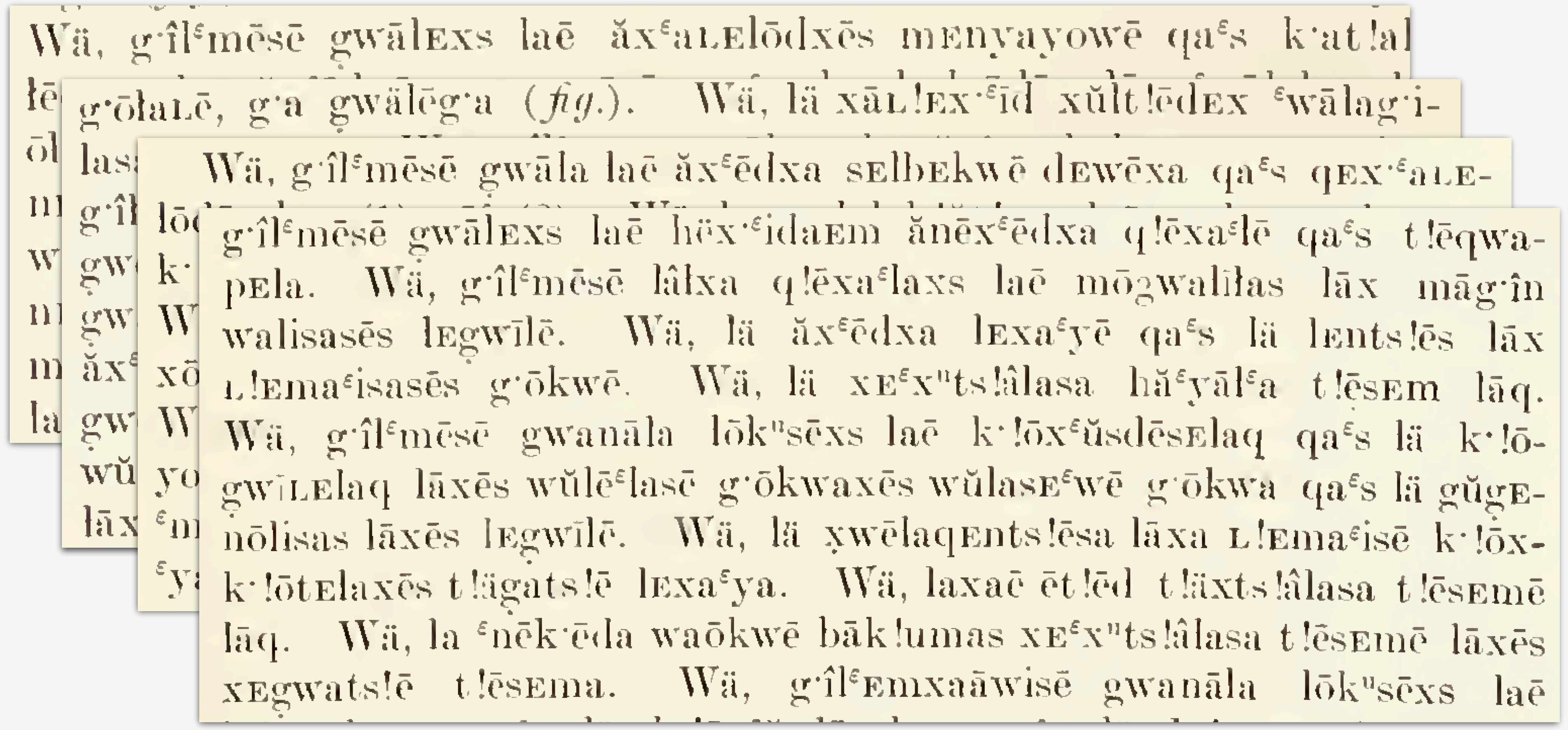

g·îlᵋmēsē gwālᴇxs laē hēx·ᵋidaᴇm ănēx·ᵋēdxa q!ēxaᵋlē qaᵋs t!ēqwa-
pᴇla. Wä, g·îlᵋmēsē lâlxa q!ēxaᵋlaxs laē mōgwalīłas lāx māg·în
walisasēs łᴇgwīlē. Wä, lä ăx·ᵋēdxa łᴇxaᵋyē qaᵋs lä łᴇnts!ēs lāx
ʟ!ᴇmaᵋisasēs g·ōkwē. Wä, lä xᴇᵋx"ts!âlasa hăᵋyālᵋa t!ēsᴇm lāq.
Wä, g·îlᵋmēsē gwanāla lōk"sēxs laē k·!ōxᵋŭsdēsᴇlaq qaᵋs lä k·!ō-
gwīʟᴇlaq lāxēs wŭlēᵋlasē g·ōkwaxēs wŭlasᴇᵋwē g·ōkwa qaᵋs lä gŭgᴇ-
nōlisas lāxēs łᴇgwīlē. Wä, lä xwēlaqᴇnts!ēsa lāxa ʟ!ᴇmaᵋisē k·!ōx-
k·!ōtᴇlaxēs t!ägats!ē łᴇxaᵋya. Wä, laxaē ēt!ēd t!äxts!âlasa t!ēsᴇmē
lāq. Wä, la ᵋnēk·ēda waōkwē bāk!umas xᴇᵋx"ts!âlasa t!ēsᴇmē lāxēs
xᴇgwats!ē t!ēsᴇma. Wä, g·îlᵋᴇmxaāwisē gwanāla lōk"sēxs laē

- Tremendous cultural and linguistic value!
- Minimally accessible to researchers
- Manual search in scanned images

Impact case study: Kwak’wala

- Collaborating with documentary linguists and Kwak’wala speakers
  - Identify documents that would be **most useful to extract text from**

Boas texts: 10 volumes of Kwak’wala language and community documentation

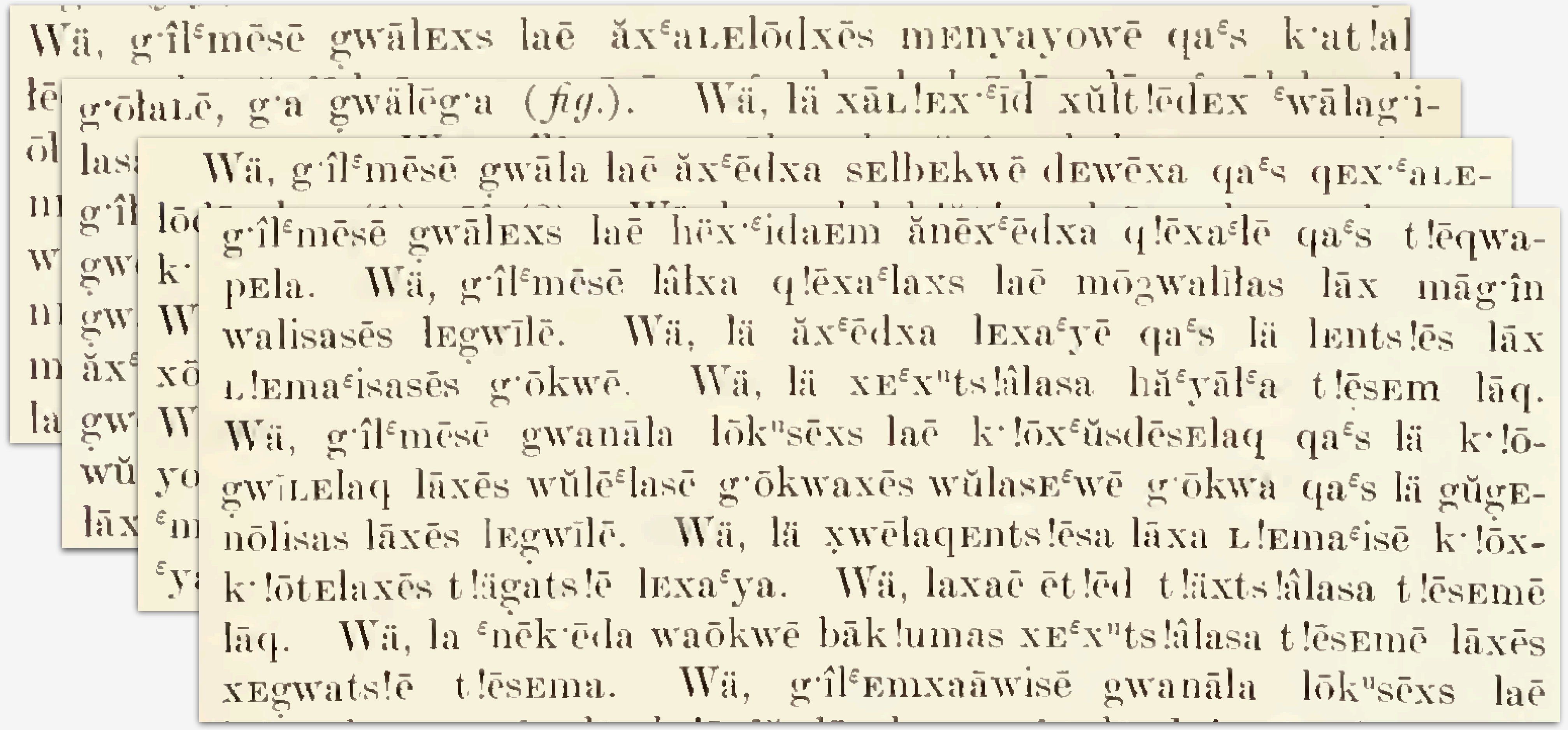


- Tremendous cultural and linguistic value!
- Minimally accessible to researchers
- Manual search in scanned images
- Legacy orthography that is hard to read

Impact case study: Kwak’wala

- Collaborating with documentary linguists and Kwak’wala speakers
  - Identify documents that would be most useful to extract text from

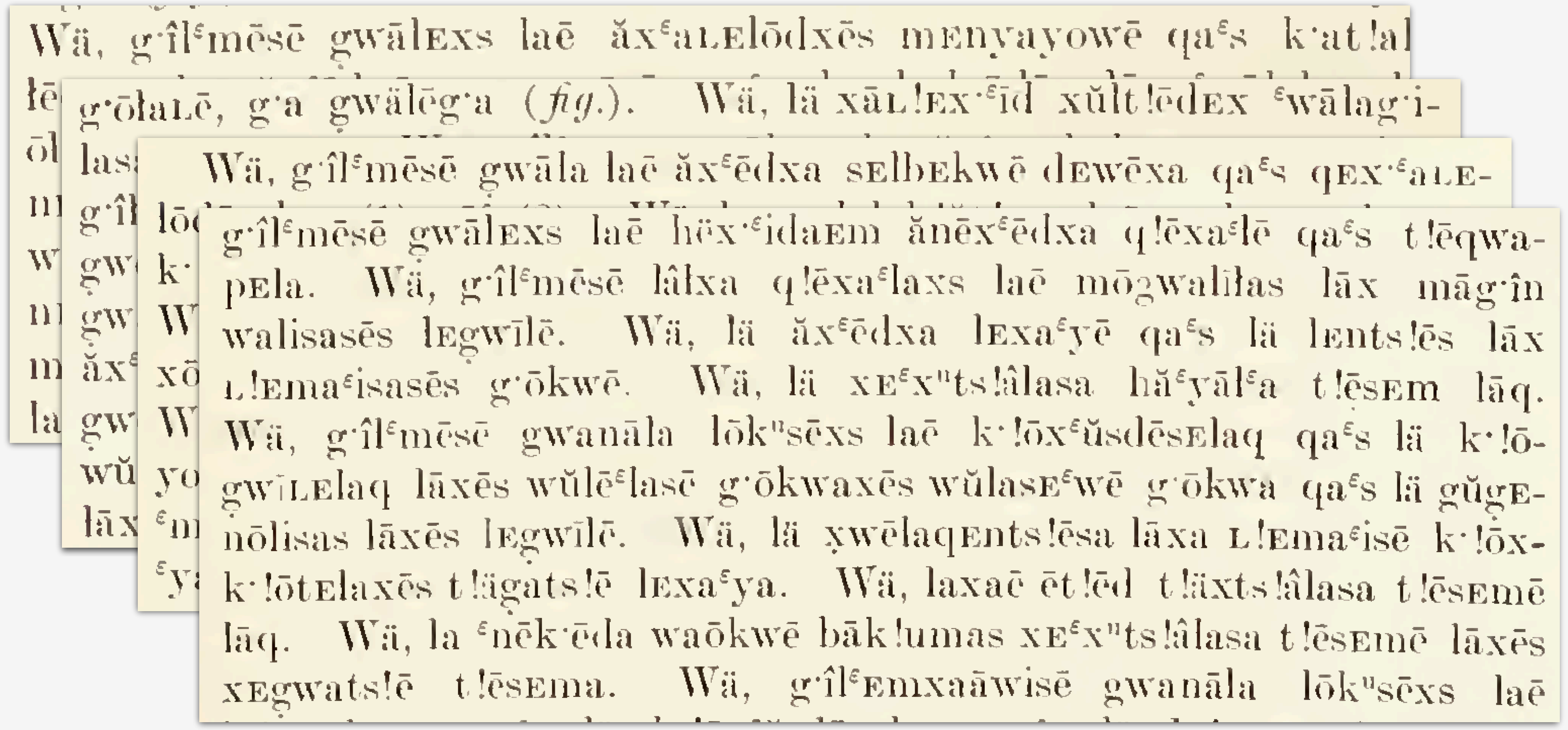
Boas texts: 10 volumes of Kwak’wala language and community documentation


g·îlᵋmēsē gwālᴇxs laē hēx·ᵋidaᴇm ănēxᵋēdxa q!ēxaᵋlē qaᵋs t!ēqwa-
pᴇla. Wä, g·îlᵋmēsē lâlxa q!ēxaᵋlaxs laē mōgwalīłas lāx māg·în
walisasēs łᴇgwīlē. Wä, lä ăxᵋēdxa łᴇxaᵋyē qaᵋs lä łᴇnts!ēs lāx
ʟ!ᴇmaᵋisasēs g·ōkwē. Wä, lä xᴇᵋx"ts!âlasa hăᵋyālᵋa t!ēsᴇm lāq.
Wä, g·îlᵋmēsē gwanāla lōk"sēxs laē k·!ōxᵋŭsdēsᴇlaq qaᵋs lä k·!ō-
gwiʟᴇlaq lāxēs wŭlēᵋlasē g·ōkwaxēs wŭlasᴇᵋwē g·ōkwa qaᵋs lä gŭgᴇ-
nōlisas lāxēs łᴇgwīlē. Wä, lä xwēlaqᴇnts!ēsa lāxa ʟ!ᴇmaᵋisē k·!ōx-
k·!ōtᴇlaxēs t!āgats!ē łᴇxaᵋya. Wä, laxaē ēt!ēd t!āxts!âlasa t!ēsᴇmē
lāq. Wä, la ᵋnēk·ēda waōkwē bāk!umas xᴇᵋx"ts!âlasa t!ēsᴇmē lāxēs
xᴇgwats!ē t!ēsᴇma. Wä, g·îlᵋᴇmxaāwisē gwanāla lōk"sēxs laē

Impact case study: Kwak'wala

- Collaborating with documentary linguists and Kwak'wala speakers
  - Identify documents that would be most useful to extract text from

Boas texts: 10 volumes of Kwak'wala language and community documentation

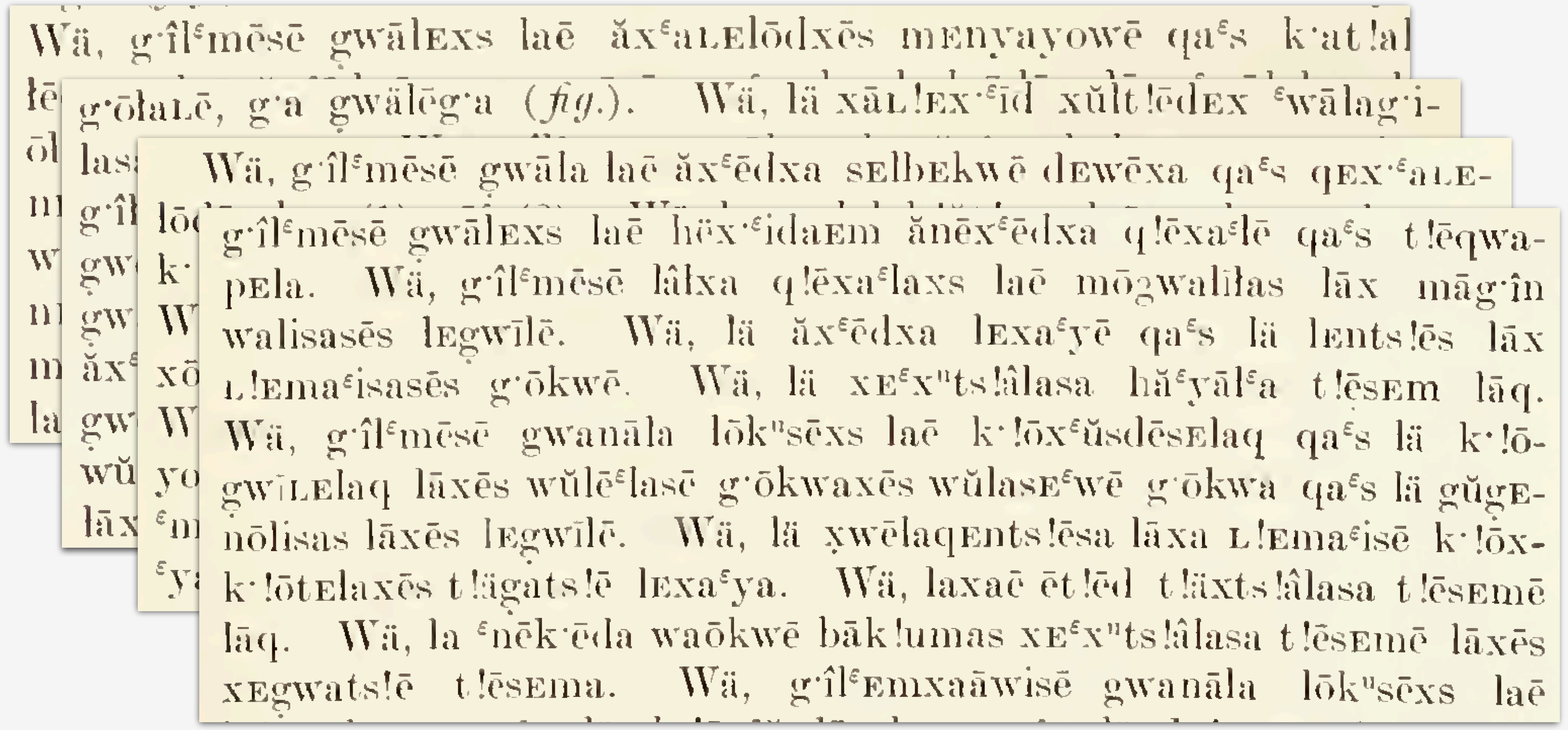

g·îlᵋmēsē gwālᴇxs laē hēx·ᵋidaᴇm ănēxᵋēdxa q!ēxaᵋlē qaᵋs t!ēqwa-
pᴇla. Wä, g·îlᵋmēsē lâlxa q!ēxaᵋlaxs laē mōgwalīlas lāx māg·în
walisasēs lᴇgwīlē. Wä, lä ăxᵋēdxa lᴇxaᵋyē qaᵋs lä lᴇnts!ēs lāx
ʟ!ᴇmaᵋisasēs g·ōkwē. Wä, lä xᴇᵋxᵘts!âlasa hăᵋyālᵋa t!ēsᴇm lāq.
Wä, g·îlᵋmēsē gwanāla lōkᵘsēxs laē k·!ōxᵋŭsdēsᴇlaq qaᵋs lä k·!ō-
gwīʟᴇlaq lāxēs wŭlēᵋlasē g·ōkwaxēs wŭlasᴇᵋwē g·ōkwa qaᵋs lä gŭgᴇ-
nōlisas lāxēs lᴇgwīlē. Wä, lä xwēlaqᴇnts!ēsa lāxa ʟ!ᴇmaᵋisē k·!ōx-
k·!ōtᴇlaxēs t!ägats!ē lᴇxaᵋya. Wä, laxaē ēt!ēd t!äxts!âlasa t!ēsᴇmē
lāq. Wä, la ᵋnēk·ēda waōkwē bāk!umas xᴇᵋxᵘts!âlasa t!ēsᴇmē lāxēs
xᴇgwats!ē t!ēsᴇma. Wä, g·îlᵋᴇmxaāwisē gwanāla lōkᵘsēxs laē

- 1500+ pages converted to a machine-readable format

Impact case study: Kwak’wala

- Collaborating with documentary linguists and Kwak’wala speakers
  - Identify documents that would be **most useful to extract text from**

Boas texts: 10 volumes of Kwak’wala language and community documentation

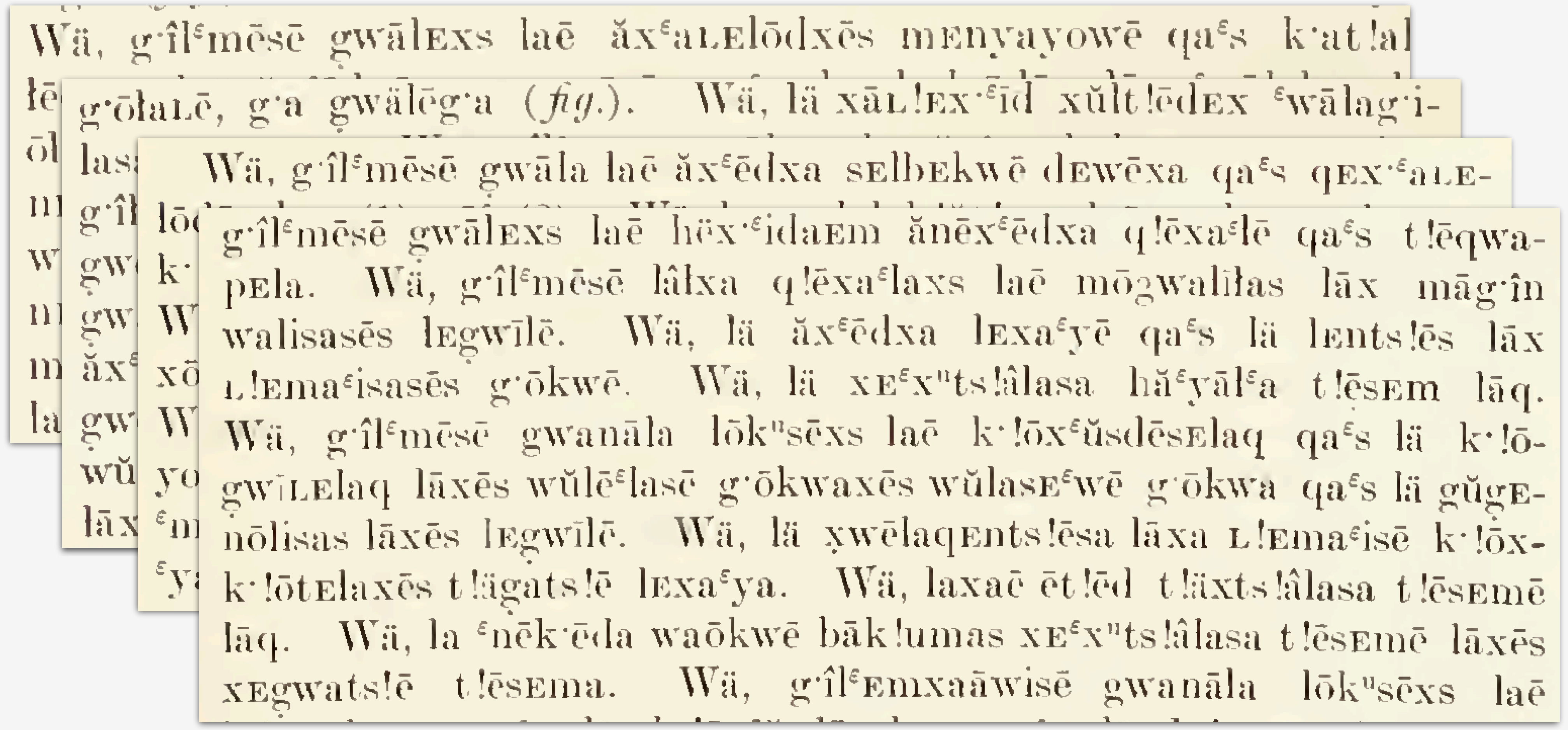


- **1500+ pages converted** to a machine-readable format
- **Searchable!**

Impact case study: Kwak'wala

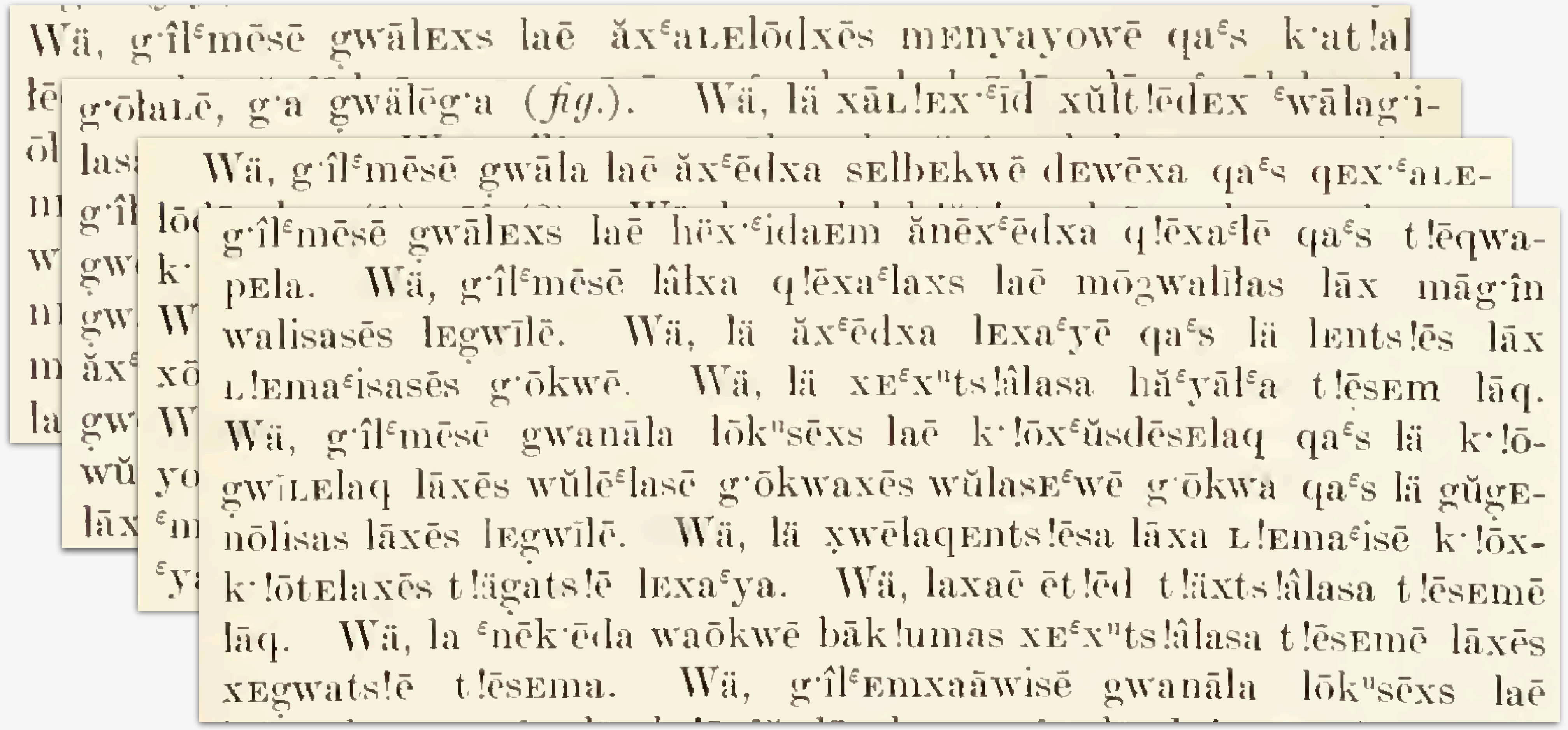
- Collaborating with documentary linguists and Kwak'wala speakers
  - Identify documents that would be most useful to extract text from

Boas texts: 10 volumes of Kwak'wala language and community documentation



- 1500+ pages converted to a machine-readable format
- Searchable!
- Legacy orthography can be automatically transliterated to modern writing systems

Impact case study: Kwak'wala



- 1500+ pages converted to a machine-readable format
- Searchable!
- Legacy orthography can be automatically transliterated to modern writing systems

Impact and applications: beyond this talk

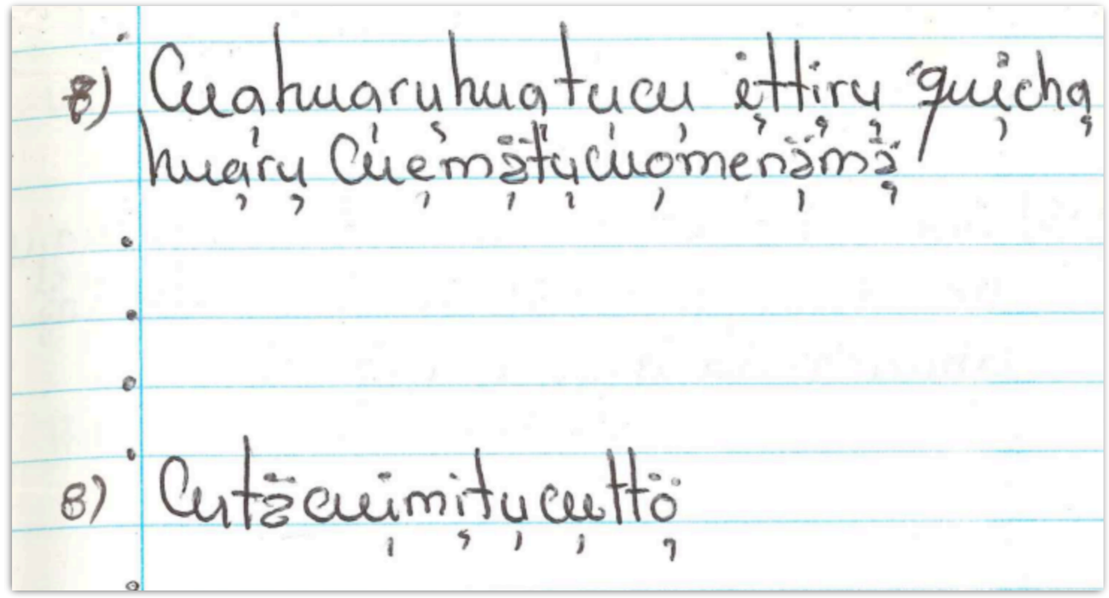
Impact and applications: beyond this talk

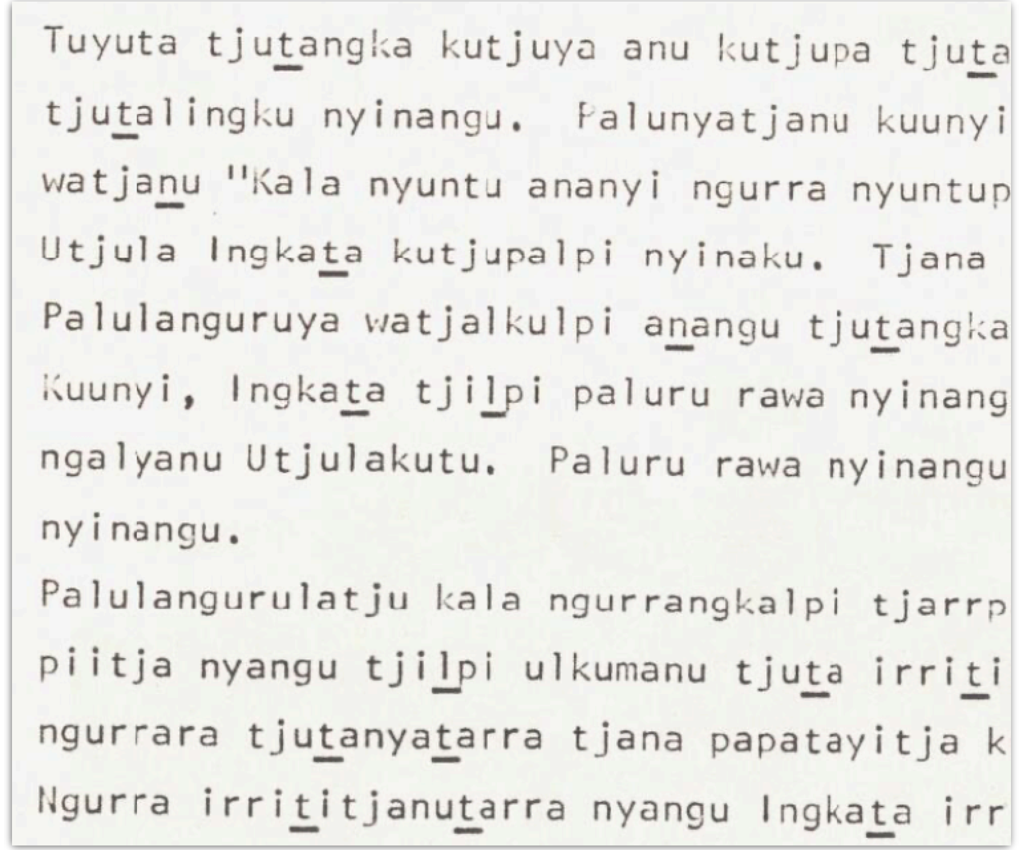
Our software is open-source and has been used on many other languages!

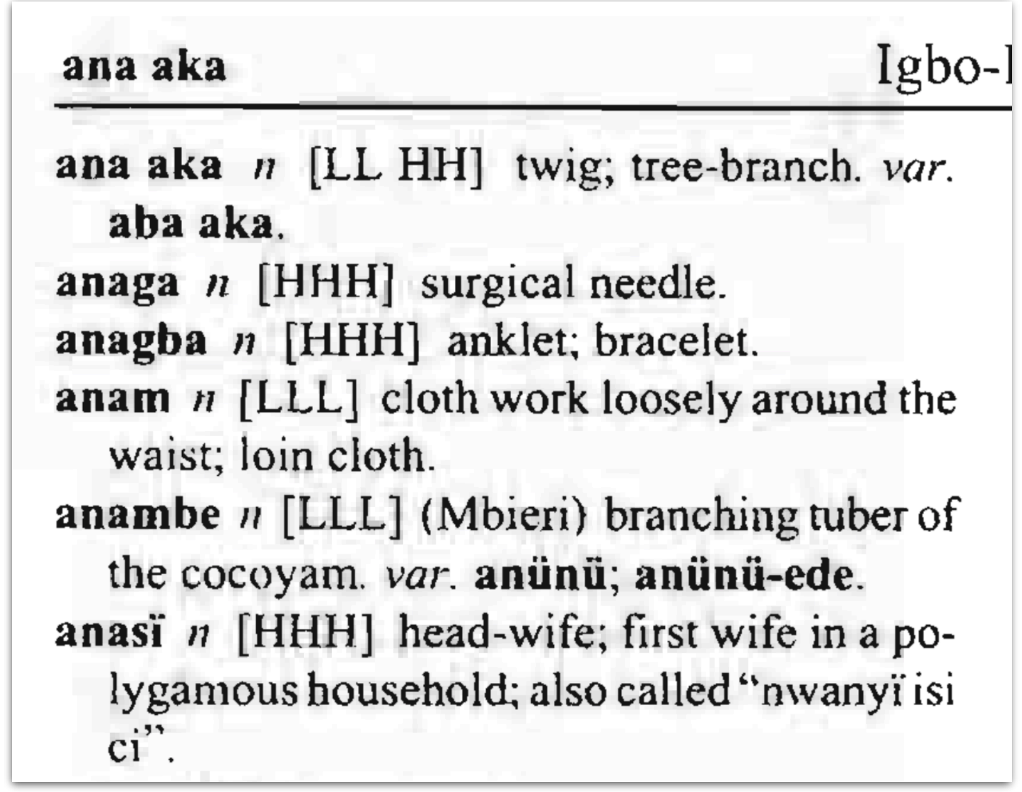
Impact and applications: beyond this talk

Our software is open-source and has been used on many other languages!

Bhutia
Sanskrit Quechua
Igbo Tibetan
Piaroa Secwepemctsín
Pintupi-Luritja




Tuyuta tjutangka kutjuya anu kutjupa tjuta
tjutalingku nyinangu. Palunyatjanu kuunyi
watjanu "Kala nyuntu ananyi ngurra nyuntup
Utjula Ingkata kutjupalpi nyinaku. Tjana
Palulanguruya watjalkulpi anangu tjutangka
Kuunyi, Ingkata tjilpi paluru rawa nyinang
ngalyanu Utjulakutu. Paluru rawa nyinangu
nyinangu.
Palulangurulatju kala ngurrangkalpi tjarrp
piitja nyangu tjilpi ulkumanu tjuta irriti
ngurrara tjutanyatarra tjana papatayitja k
Ngurra irrititjanutarra nyangu Ingkata irr


ana aka Igbo-I

ana aka *n* [LL HH] twig; tree-branch. *var.* **aba aka**.
anaga *n* [HHH] surgical needle.
anagba *n* [HHH] anklet; bracelet.
anam *n* [LLL] cloth work loosely around the waist; loin cloth.
anambe *n* [LLL] (Mbieri) branching tuber of the cocoyam. *var.* **anünü**; **anünü-ede**.
anasï *n* [HHH] head-wife; first wife in a polygamous household; also called "nwanyï isi ci".

Image credits: Ijemma Onwuzulike, Jorge Labrada, Ben Foley

Impact and applications: beyond this talk

Our software is open-source and has been used on many other languages!

Bhutia

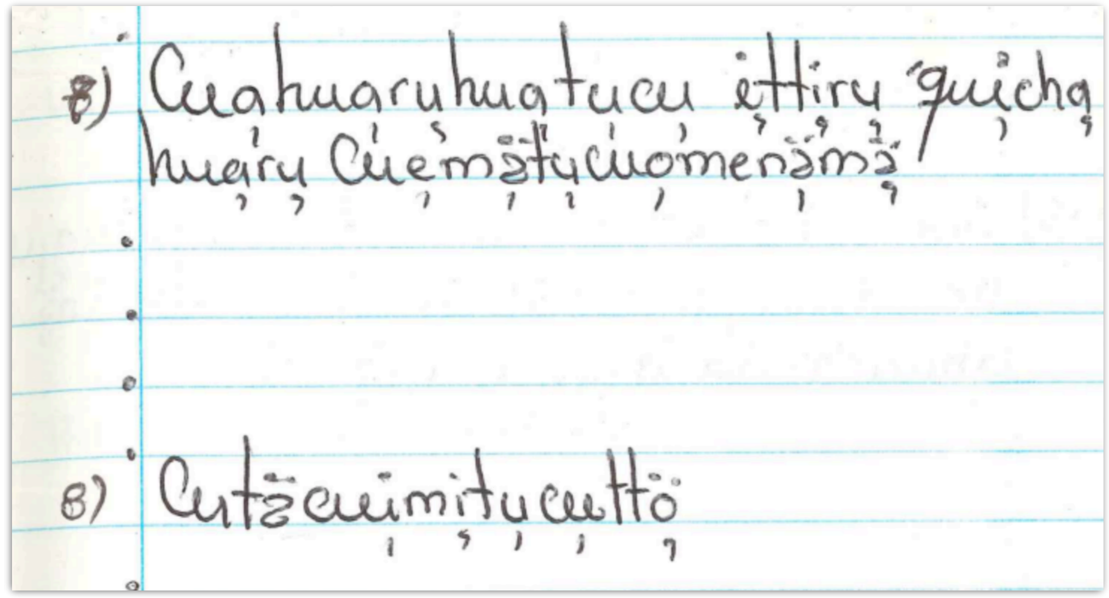
Sanskrit Quechua

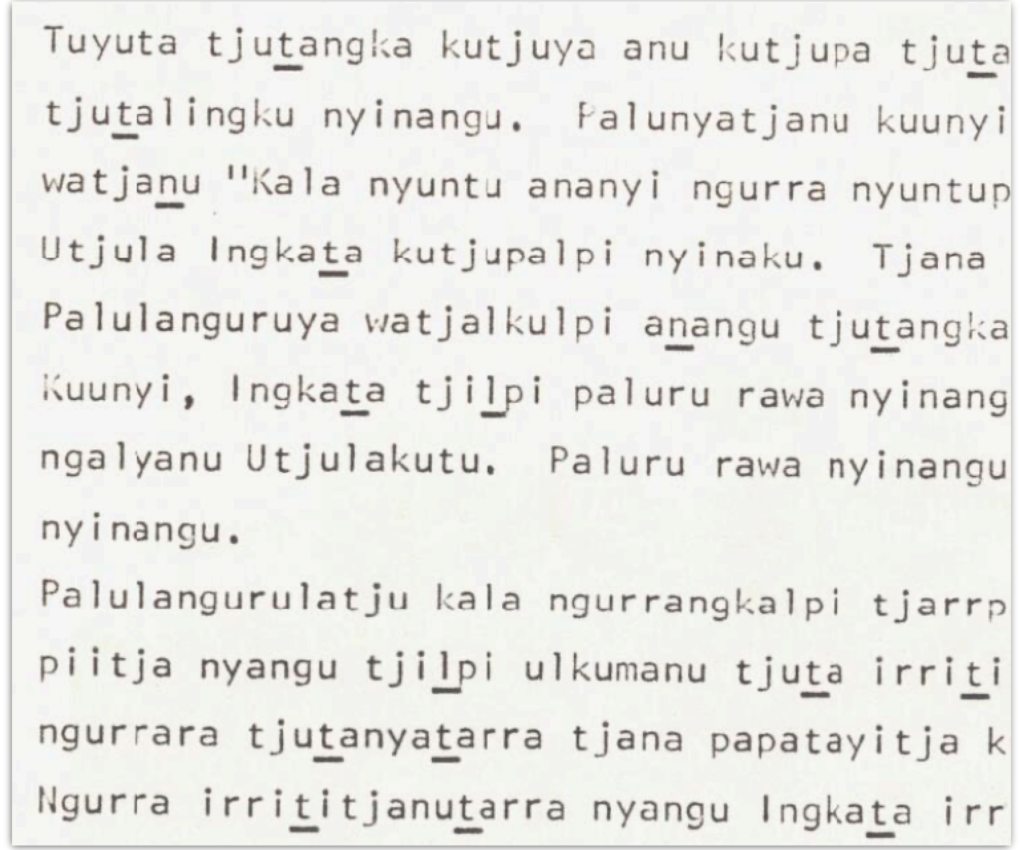
Igbo Tibetan

Piaroa Secwepemctsín

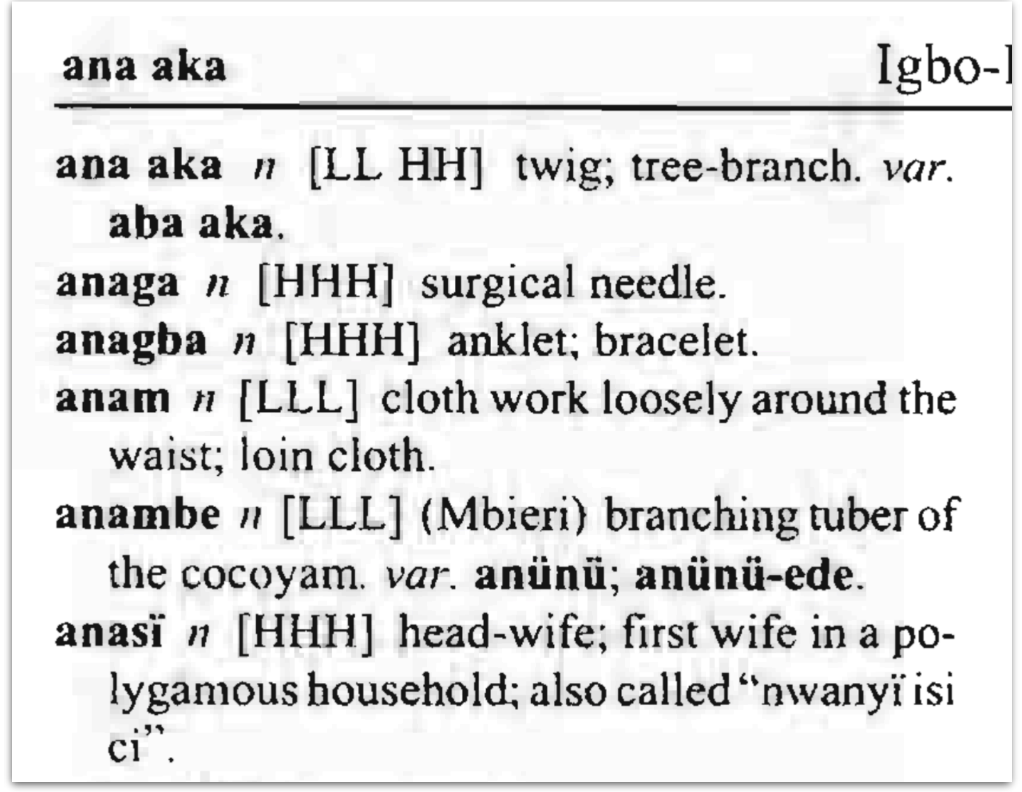
Pintupi-Luritja

Extracting text to train machine translation for Pintupi-Luritja





Tuyuta tjutangka kutjuya anu kutjupa tjuta
tjutalingku nyinangu. Palunyatjanu kuunyi
watjanu "Kala nyuntu ananyi ngurra nyuntup
Utjula Ingkata kutjupalpi nyinaku. Tjana
Palulanguruya watjalkulpi anangu tjutangka
Kuunyi, Ingkata tjilpi paluru rawa nyinang
ngalyanu Utjulakutu. Paluru rawa nyinangu
nyinangu.
Palulangurulatju kala ngurrangkalpi tjarrp
piitja nyangu tjilpi ulkumanu tjuta irriti
ngurrara tjutanyatarra tjana papatayitja k
Ngurra irrititjanutarra nyangu Ingkata irr



ana aka Igbo-

ana aka *n* [LL HH] twig; tree-branch. *var.* **aba aka**.
anaga *n* [HHH] surgical needle.
anagba *n* [HHH] anklet; bracelet.
anam *n* [LLL] cloth work loosely around the waist; loin cloth.
anambe *n* [LLL] (Mbieri) branching tuber of the cocoyam. *var.* **anünü**; **anünü-ede**.
anasï *n* [HHH] head-wife; first wife in a polygamous household; also called "nwanyï isi ci".

Image credits: Ijemma Onwuzulike, Jorge Labrada, Ben Foley

Impact and applications: beyond this talk

Our software is open-source and has been used on many other languages!

Bhutia

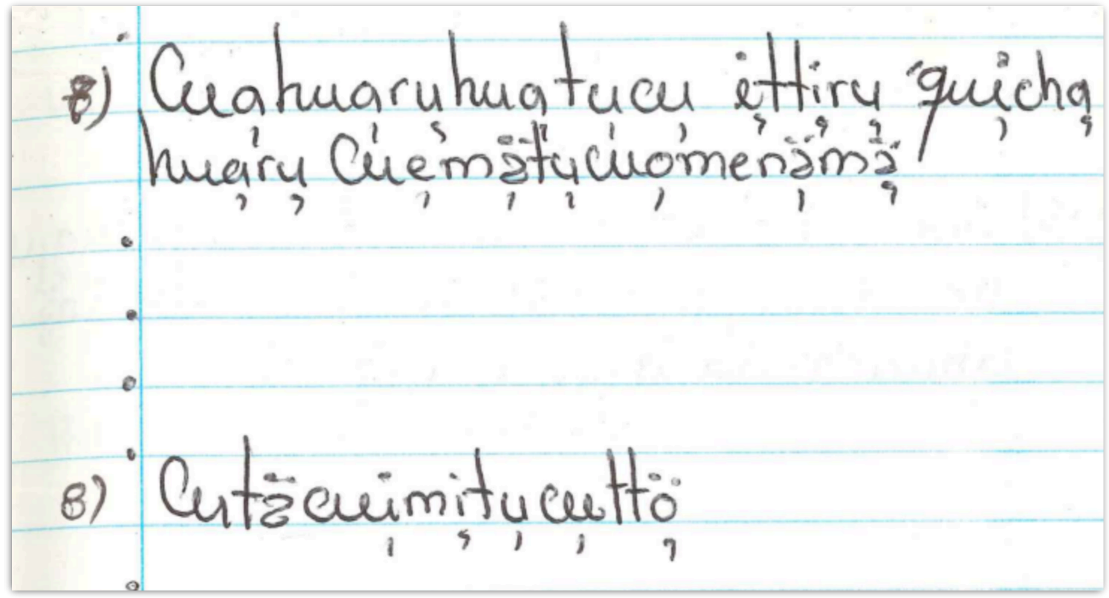
Sanskrit Quechua

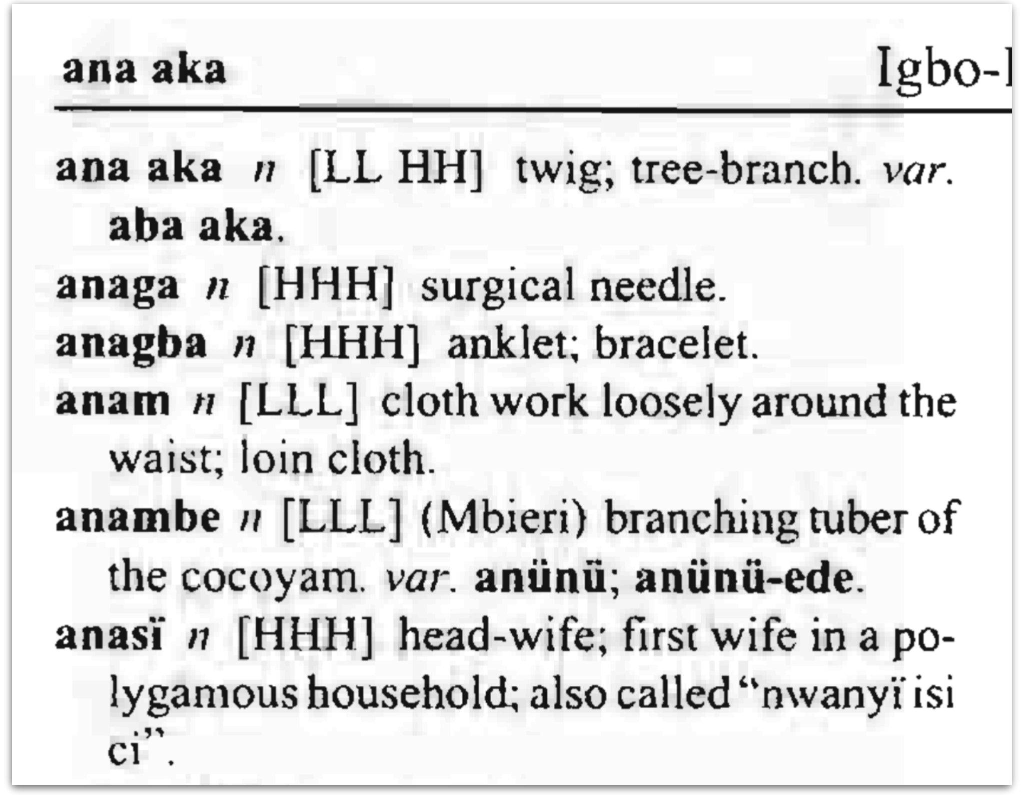
Igbo Tibetan

Piaroa Secwepemctsín

Pintupi-Luritja

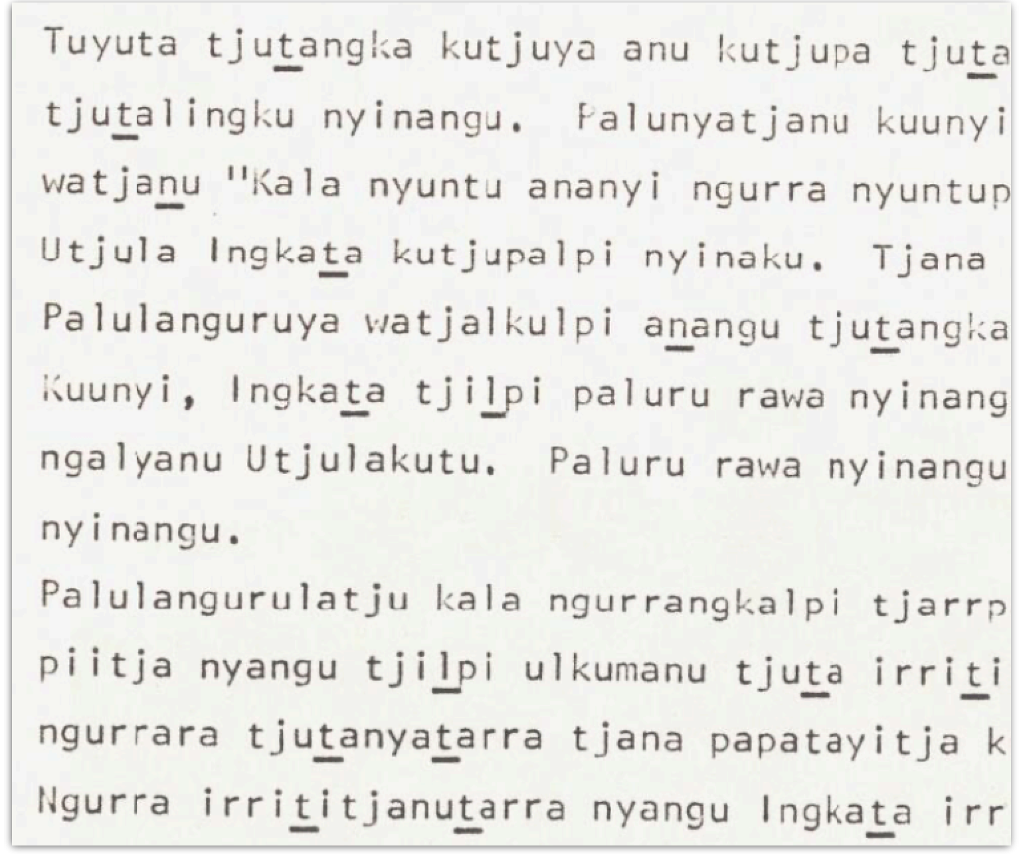
Automatic extraction of handwritten speech transcriptions in Piaroa





ana aka Igbo-I

ana aka *n* [LL HH] twig; tree-branch. *var.* **aba aka**.
anaga *n* [HHH] surgical needle.
anagba *n* [HHH] anklet; bracelet.
anam *n* [LLL] cloth work loosely around the waist; loin cloth.
anambe *n* [LLL] (Mbieri) branching tuber of the cocoyam. *var.* **anünü**; **anünü-ede**.
anasï *n* [HHH] head-wife; first wife in a polygamous household; also called "nwanyï isi ci".



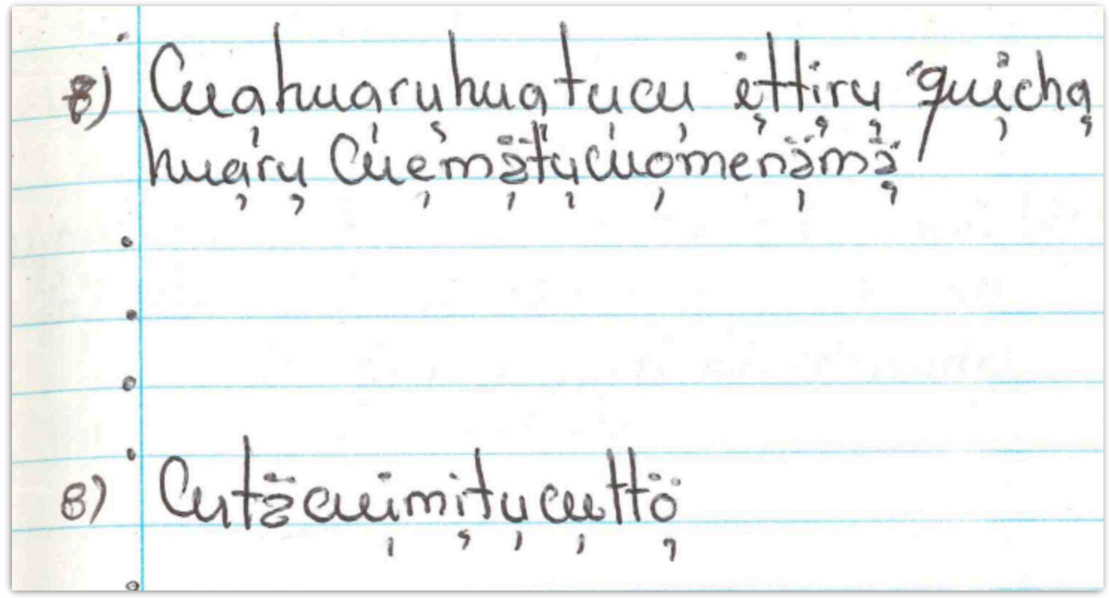
Tuyuta tjutangka kutjuya anu kutjupa tjuta
tjutalingku nyinangu. Palunyatjanu kuunyi
watjanu "Kala nyuntu ananyi ngurra nyuntup
Utjula Ingkata kutjupalpi nyinaku. Tjana
Palulanguruya watjalkulpi anangu tjutangka
Kuunyi, Ingkata tjilpi paluru rawa nyinang
ngalyanu Utjulakutu. Paluru rawa nyinangu
nyinangu.
Palulangurulatju kala ngurrangkalpi tjarrp
piitja nyangu tjilpi ulkumanu tjuta irriti
ngurrara tjutanyatarra tjana papatayitja k
Ngurra irrititjanutarra nyangu Ingkata irr

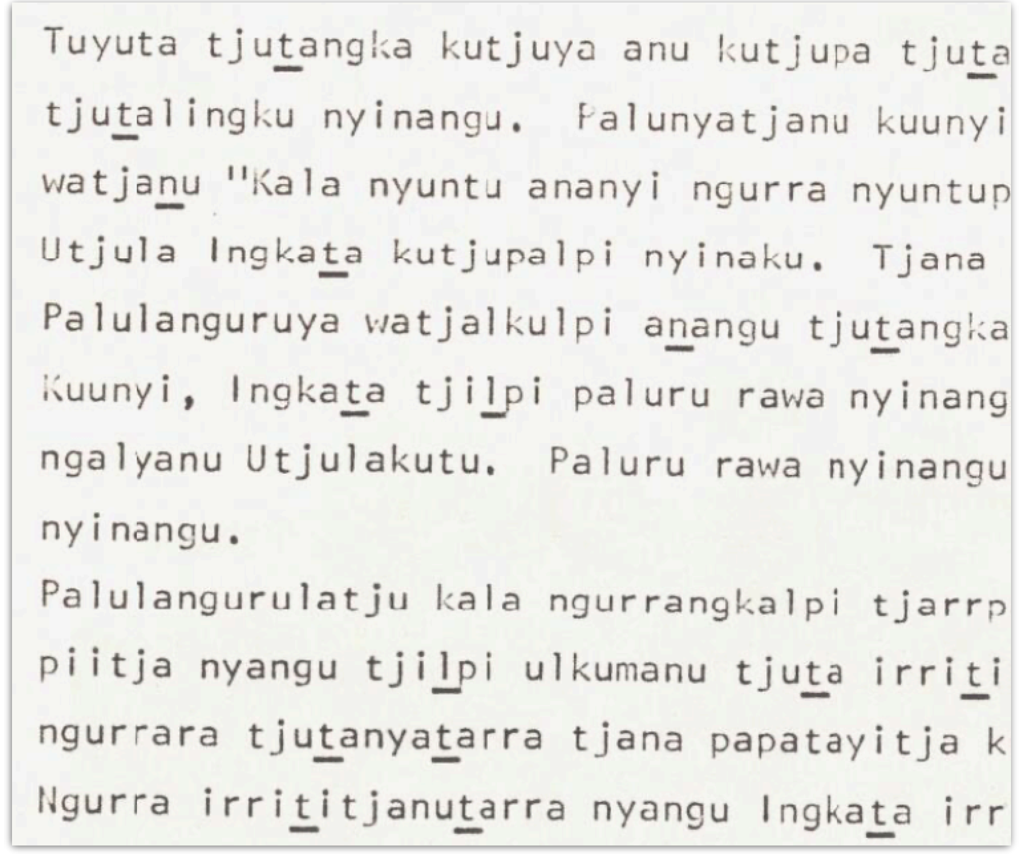
Image credits: Ijemma Onwuzulike, Jorge Labrada, Ben Foley

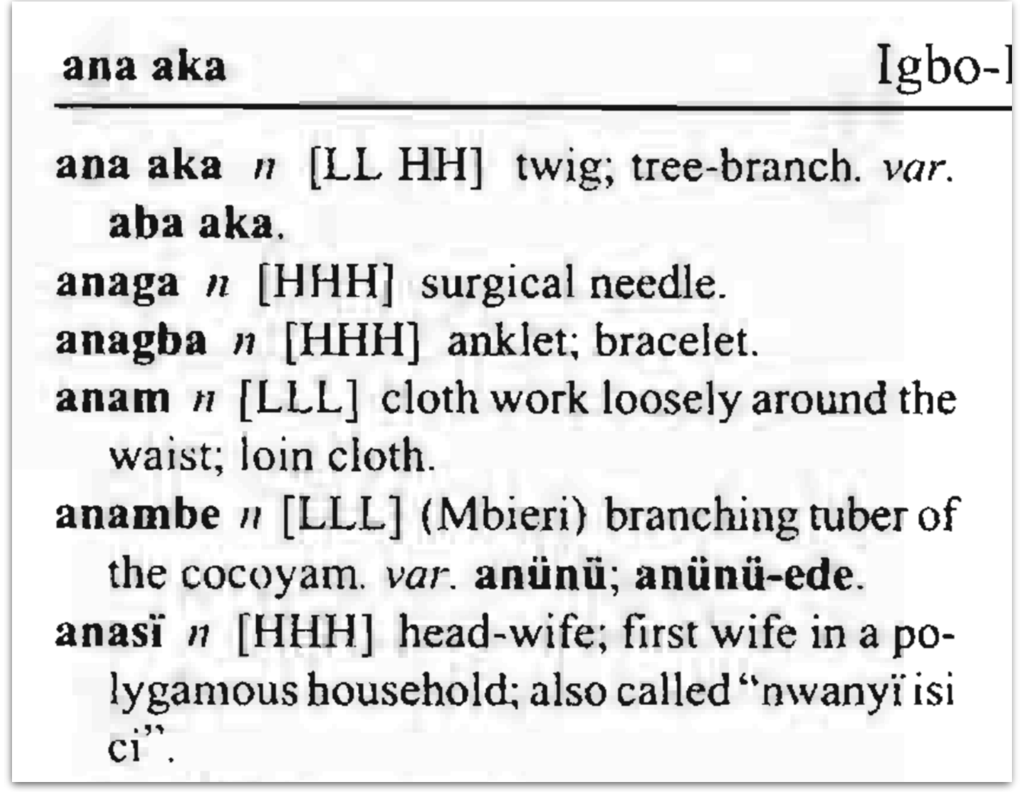
Impact and applications: beyond this talk

Our software is open-source and has been used on many other languages!

Bhutia
Sanskrit Quechua
Igbo Tibetan
Piaroa Secwepemctsín
Pintupi-Luritja




Tuyuta tjutangka kutjuya anu kutjupa tjuta
tjutalingku nyinangu. Palunyatjanu kuunyi
watjanu "Kala nyuntu ananyi ngurra nyuntup
Utjula Ingkata kutjupalpi nyinaku. Tjana
Palulanguruya watjalkulpi anangu tjutangka
Kuunyi, Ingkata tjilpi paluru rawa nyinang
ngalyanu Utjulakutu. Paluru rawa nyinangu
nyinangu.
Palulangurulatju kala ngurrangkalpi tjarrp
piitja nyangu tjilpi ulkumanu tjuta irriti
ngurrara tjutanyatarra tjana papatayitja k
Ngurra irrititjanutarra nyangu Ingkata irr


ana aka Igbo-

ana aka *n* [LL HH] twig; tree-branch. *var.* **aba aka**.
anaga *n* [HHH] surgical needle.
anagba *n* [HHH] anklet; bracelet.
anam *n* [LLL] cloth work loosely around the waist; loin cloth.
anambe *n* [LLL] (Mbieri) branching tuber of the cocoyam. *var.* **anünü**; **anünü-ede**.
anasï *n* [HHH] head-wife; first wife in a polygamous household; also called "nwanyï isi ci".

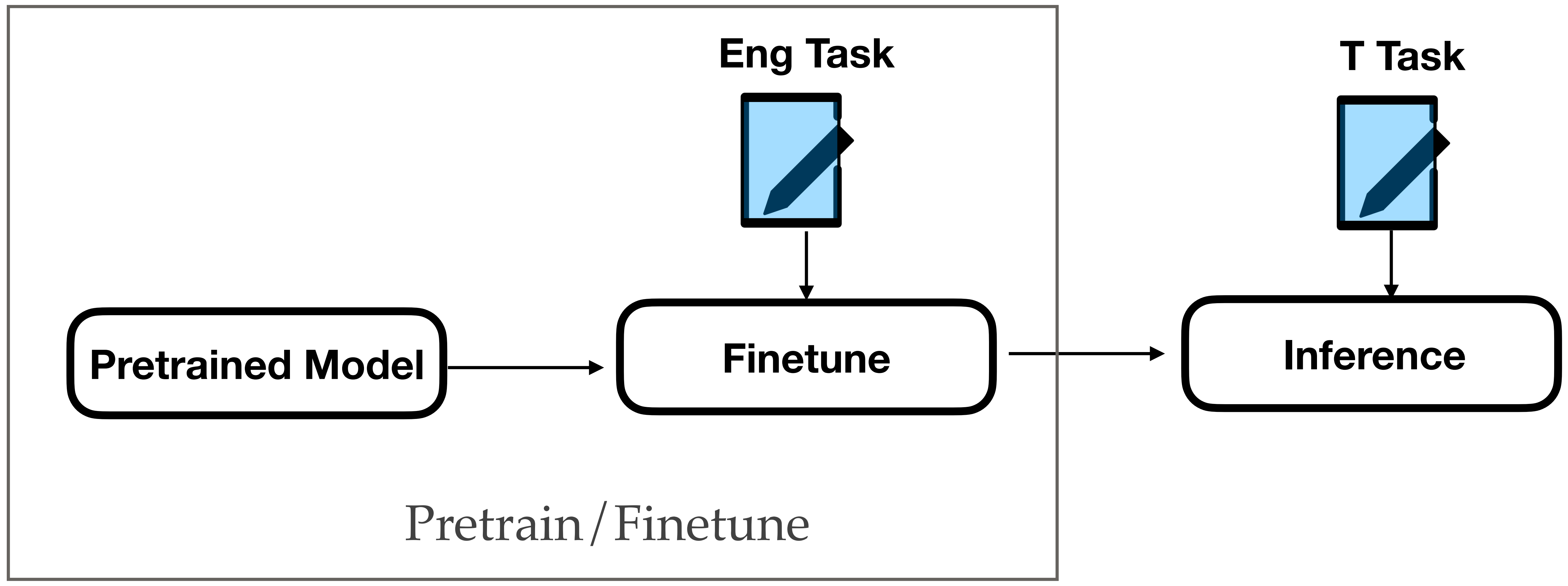
Print dictionaries in Igbo are high-coverage, but not digitized

Image credits: Ijemma Onwuzulike, Jorge Labrada, Ben Foley

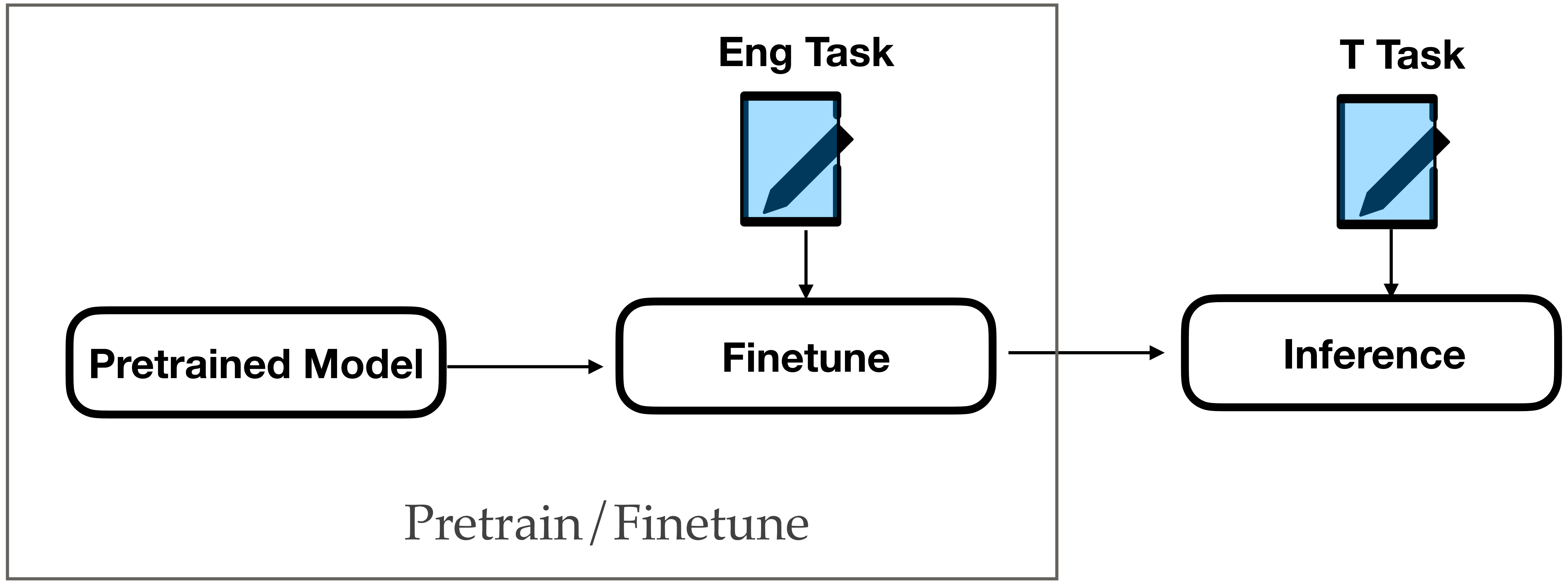
Unlocking Bi-lingual Lexicons

Xinyi Wang, Sebastian Ruder, Graham Neubig.
Expanding Pretrained Models to Thousands More Languages via Lexicon-based Adaptation.
ACL 2022.

Multilingual Pretrained Models

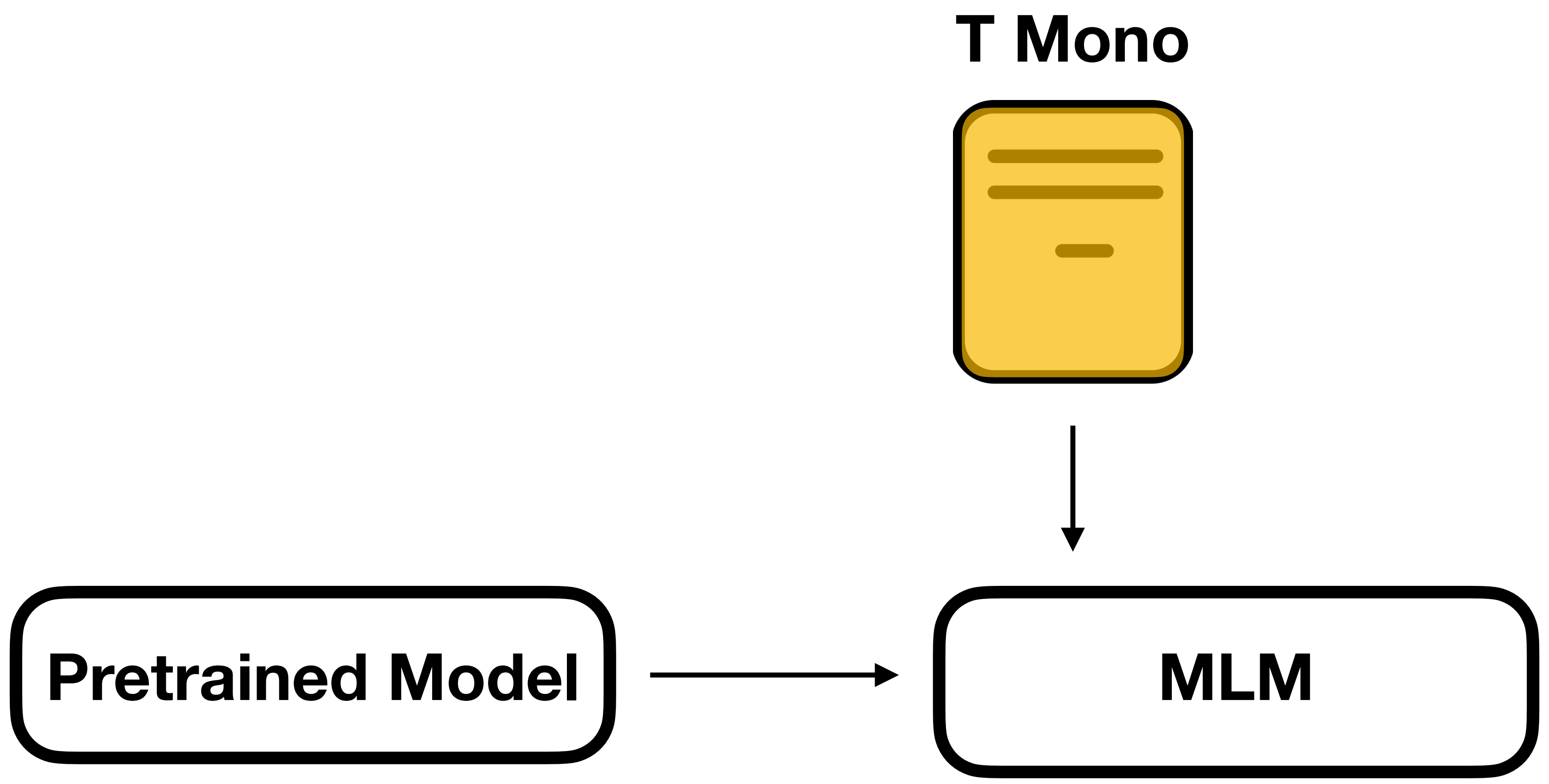


Multilingual Pretrained Models



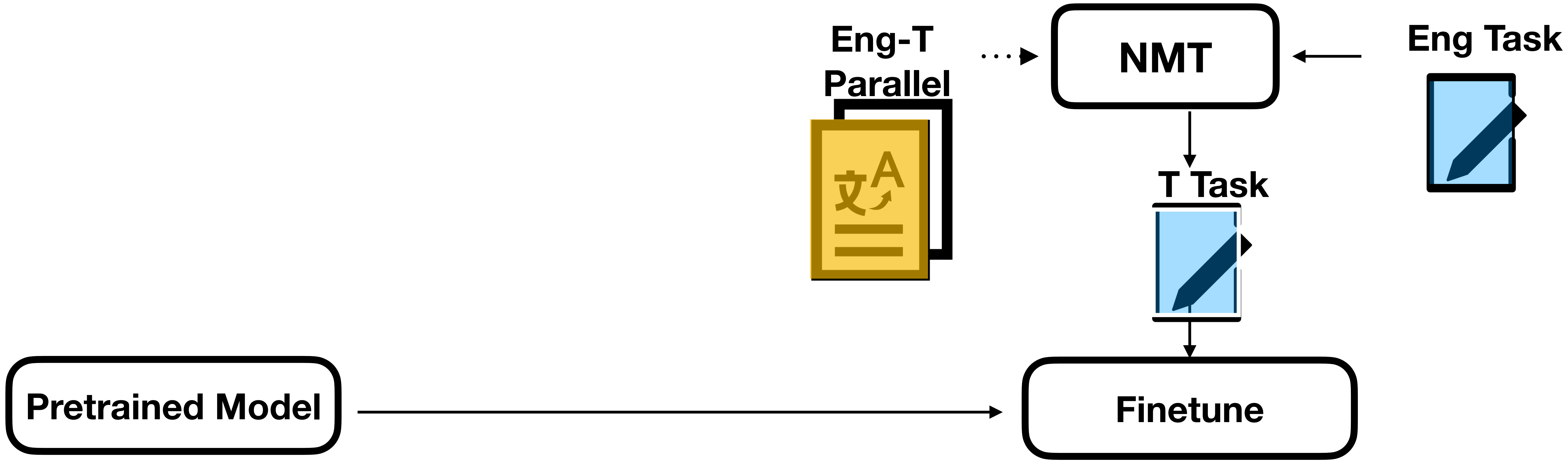
How to adapt the model for the language T?

Adaptation: Monolingual Data



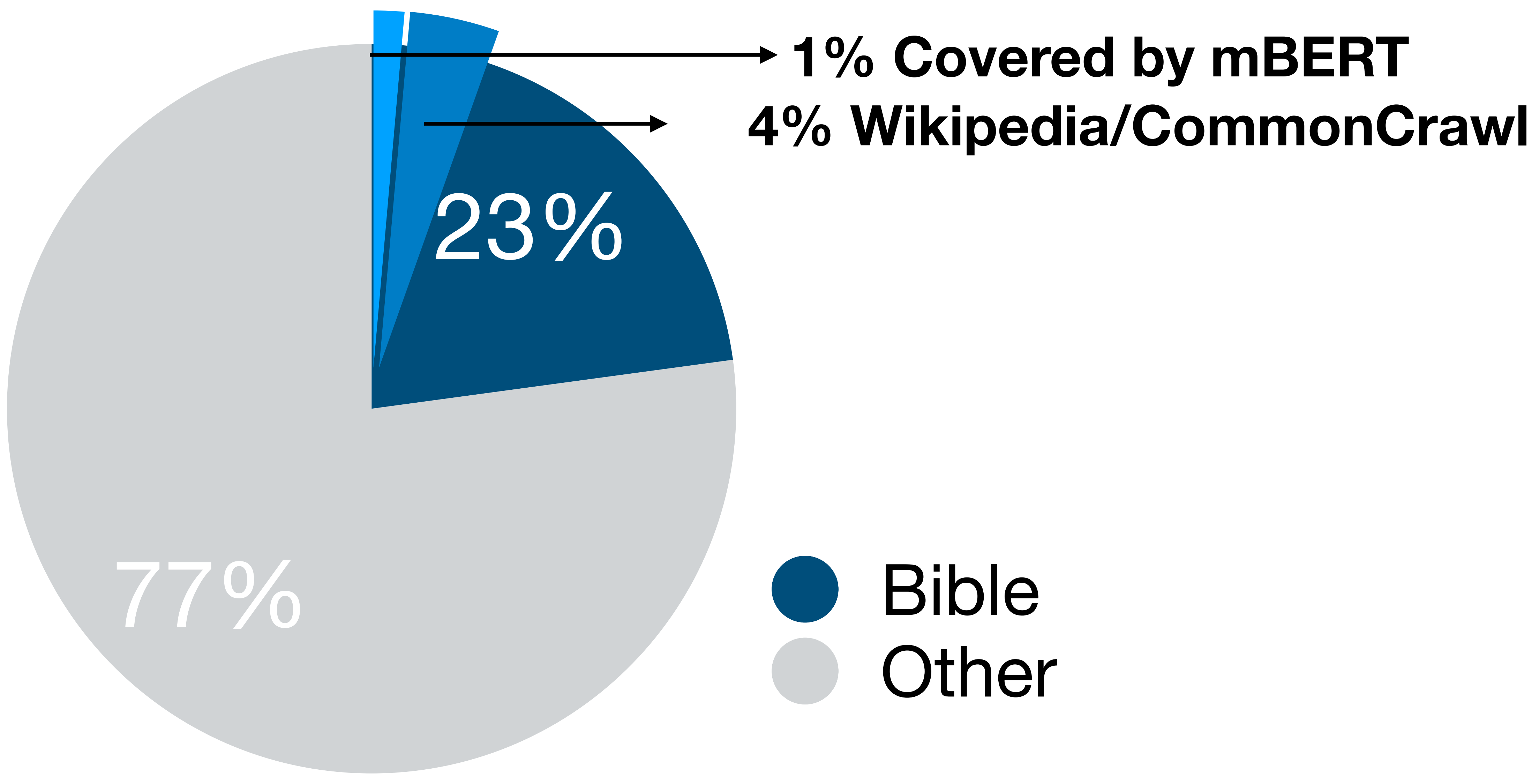
- e.g. Continued Masked Language Modeling(MLM) using monolingual data in the target language T

Adaptation: Parallel Data

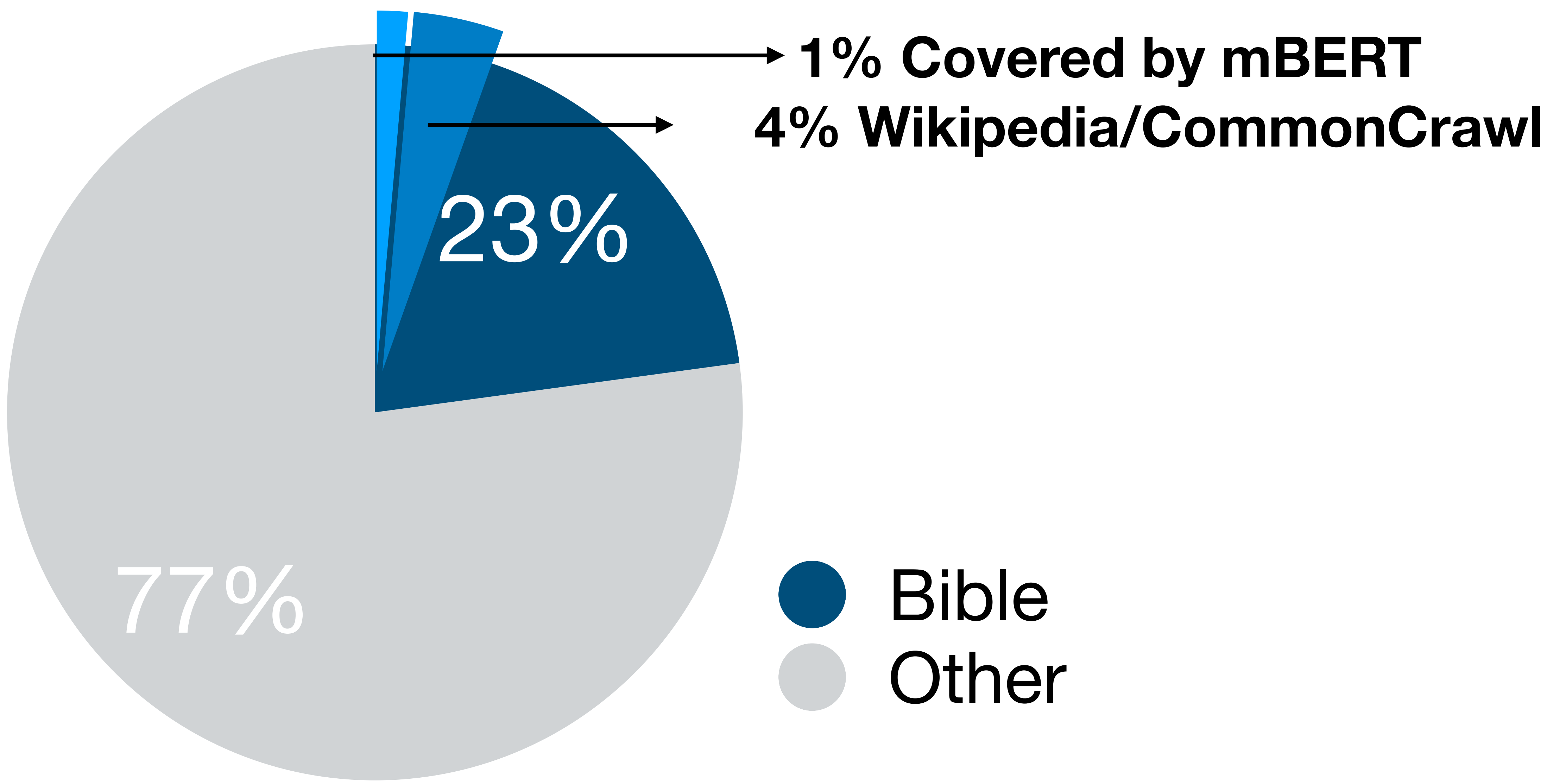


- e.g. Parallel Data: use best NMT system available to translate English task data into the target language T

Languages without Conventional Data

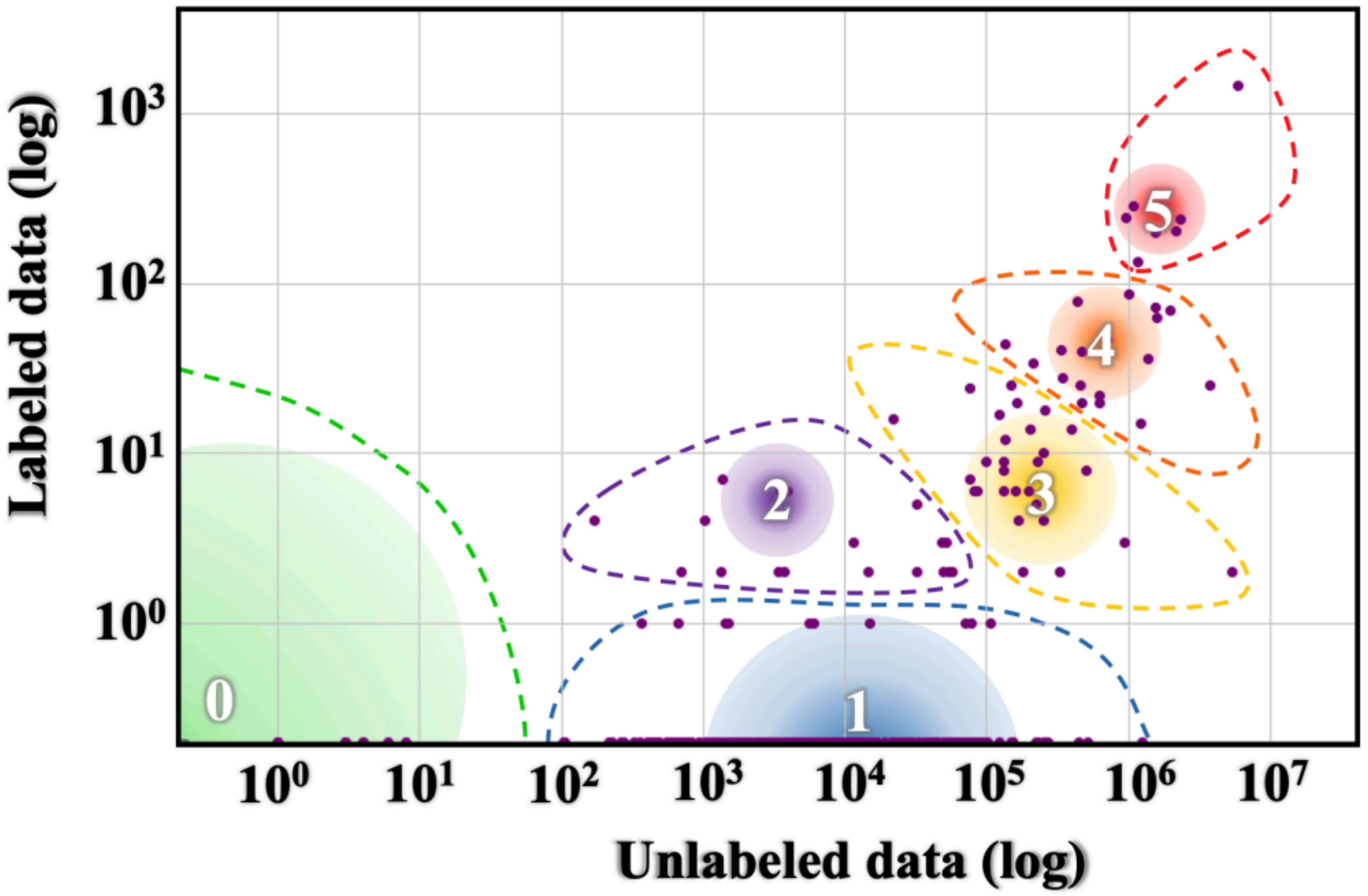


Languages without Conventional Data

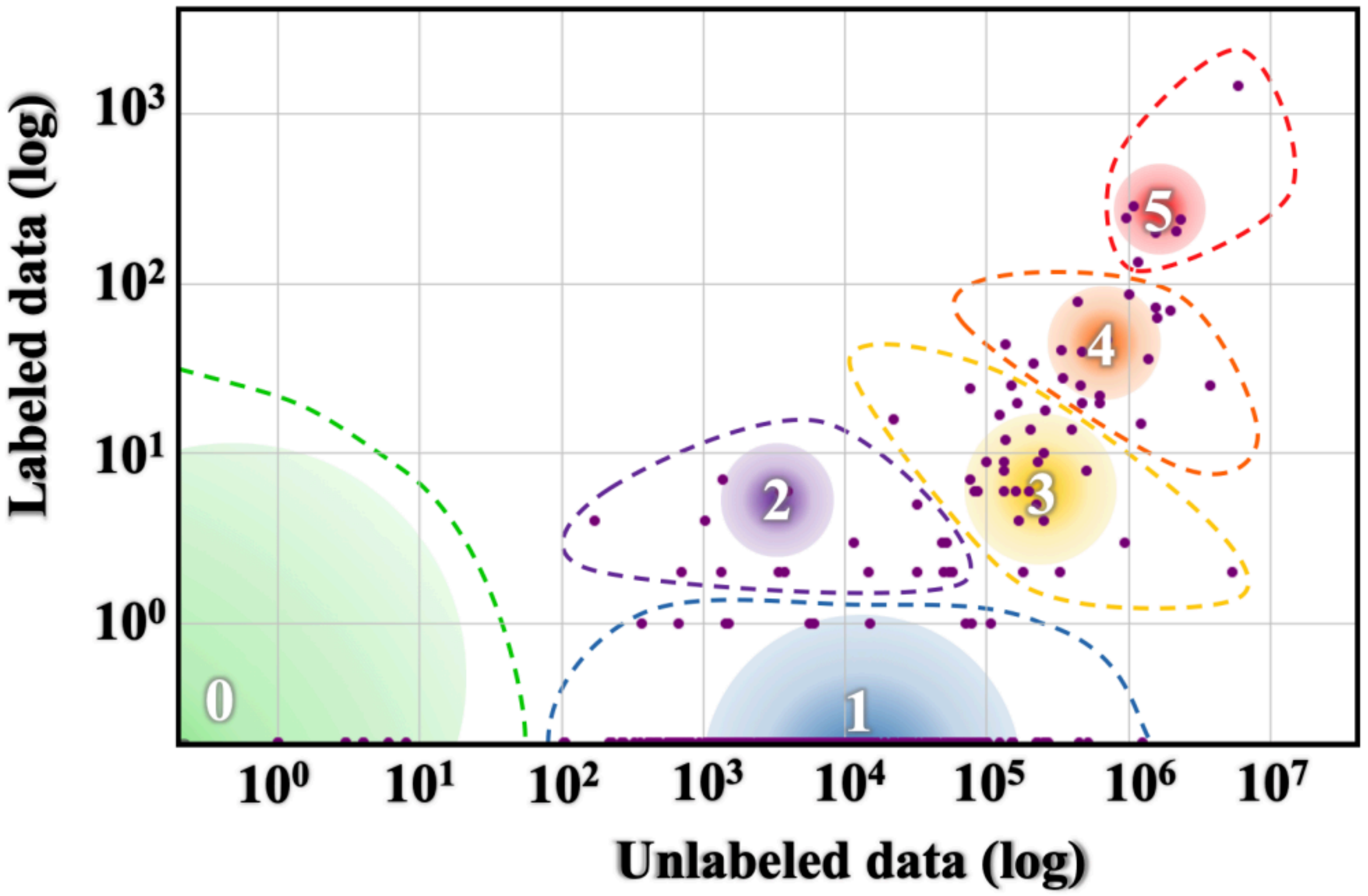


- Majority of languages in the world cannot benefit from progress in NLP due to lack of data

Two Groups of Low-resource Languages



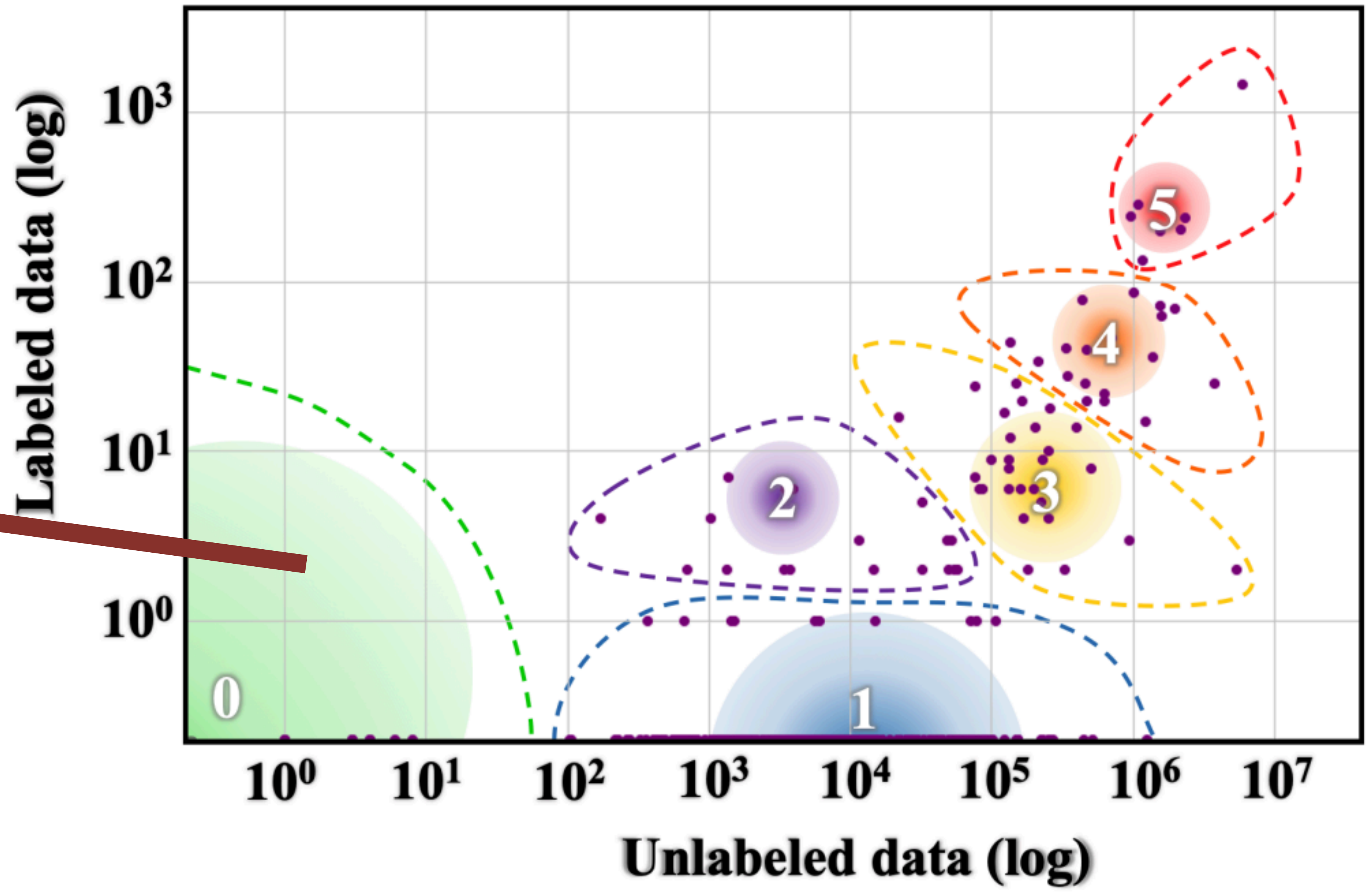
Two Groups of Low-resource Languages



- Majority of World's languages cannot benefit from progress in NLP (Joshi et al. 2020)

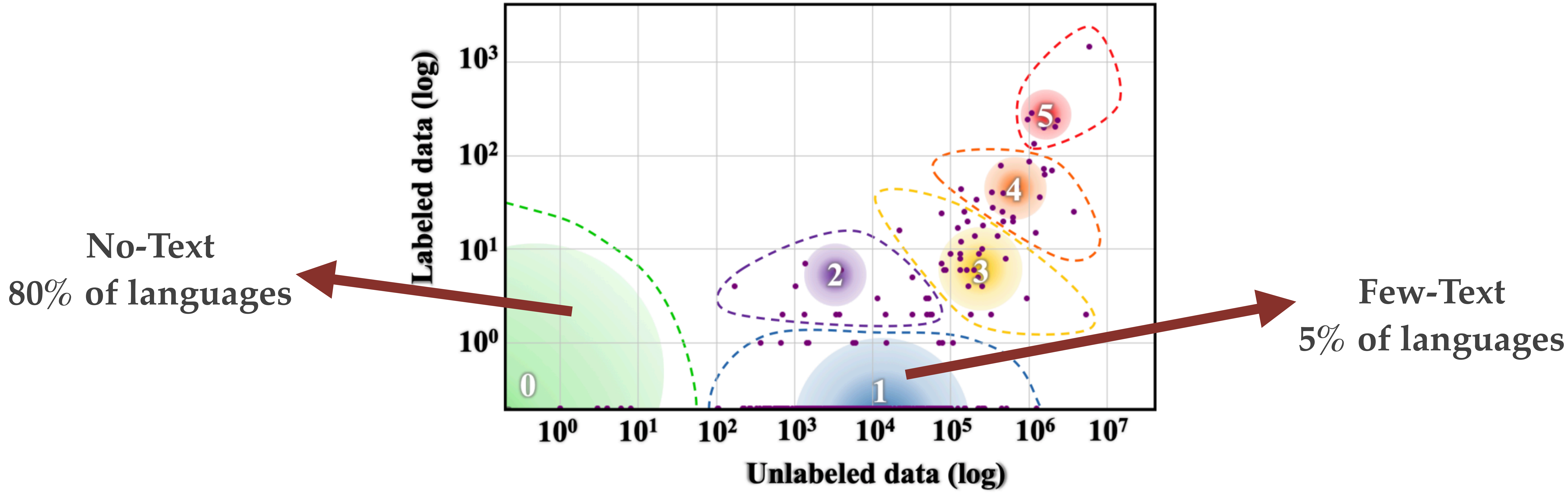
Two Groups of Low-resource Languages

No-Text
80% of languages



- Majority of World's languages cannot benefit from progress in NLP (Joshi et al. 2020)
  - No-Text: virtually no resource

Two Groups of Low-resource Languages



- Majority of World's languages cannot benefit from progress in NLP (Joshi et al. 2020)
  - No-Text: virtually no resource
  - Few-Text: very limited resource

Alternative Data Source

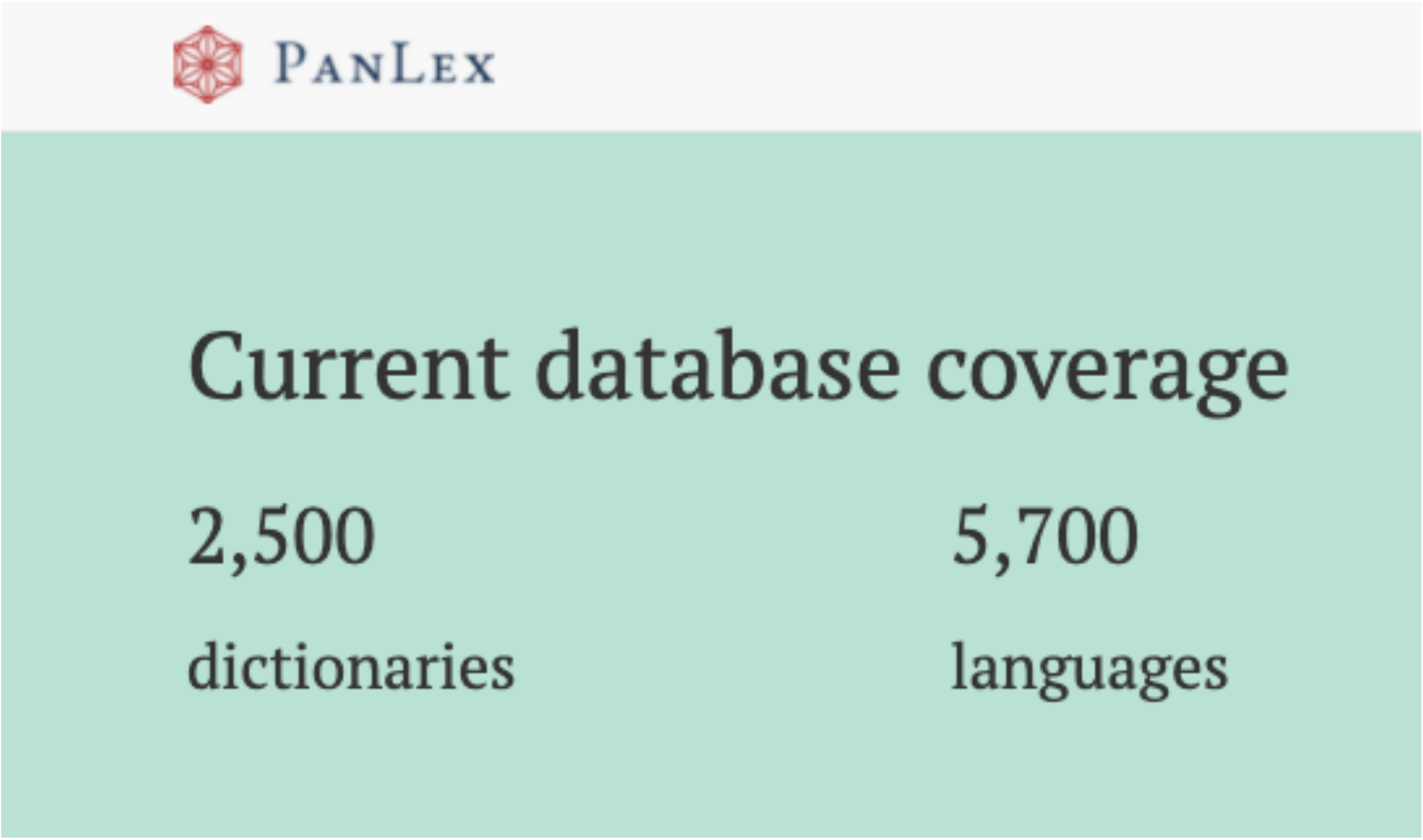
https://panlex.org/

Alternative Data Source

- Linguists have been documenting languages for years in formats such as lexicons

https://panlex.org/

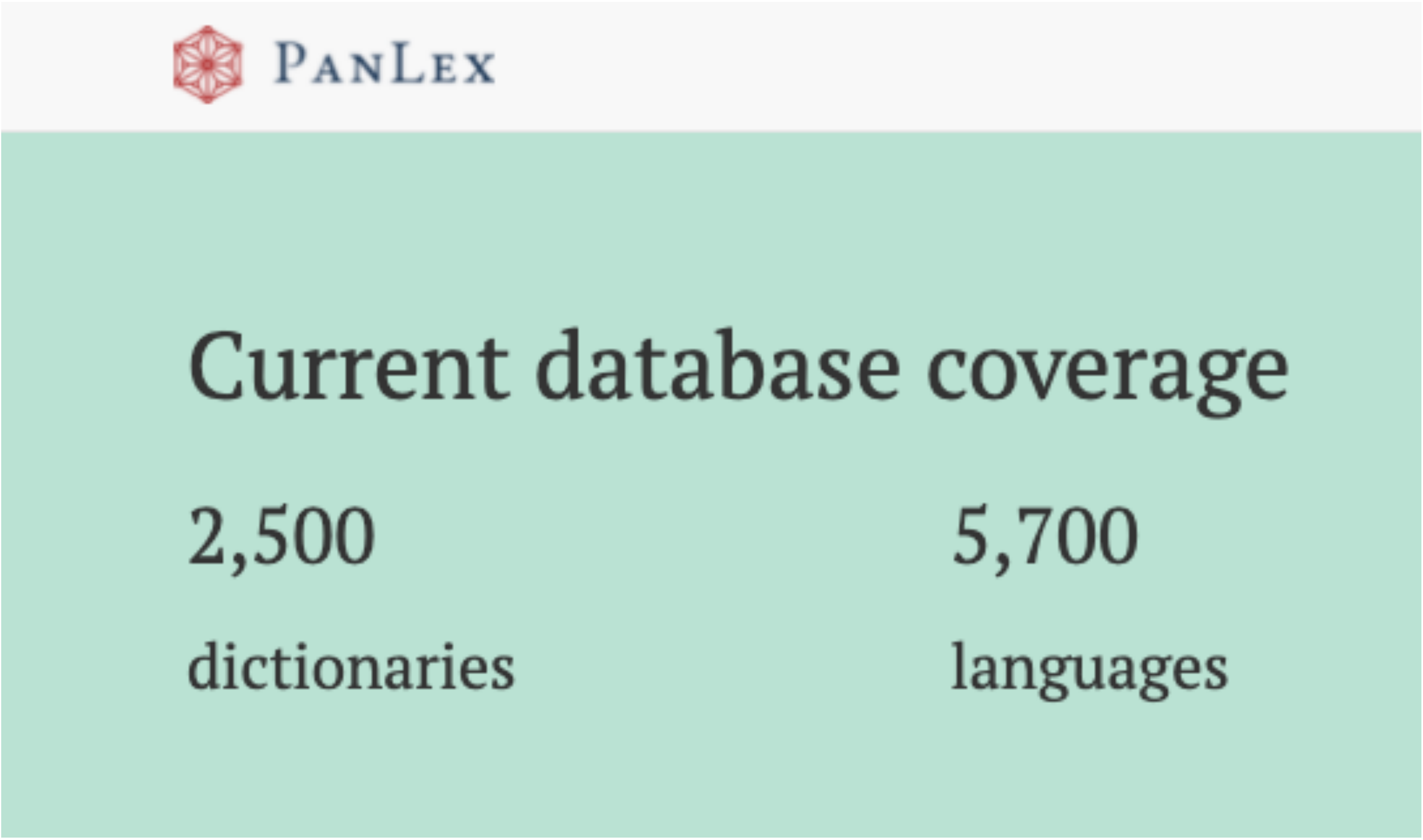
Alternative Data Source

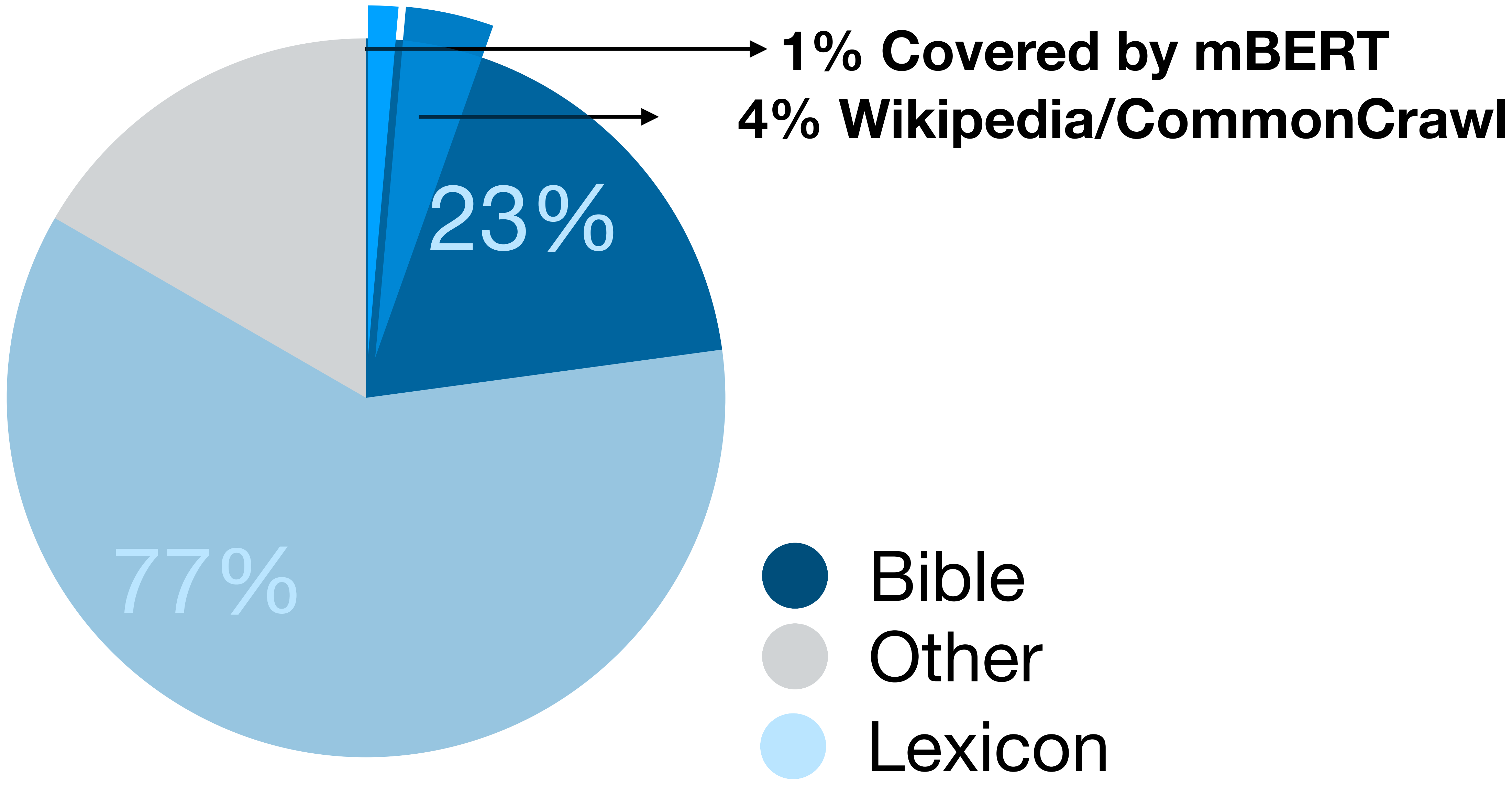


- Linguists have been documenting languages for years in formats such as lexicons
- PanLex: open-sourced database of lexicons with much better language coverage

https://panlex.org/

Alternative Data Source

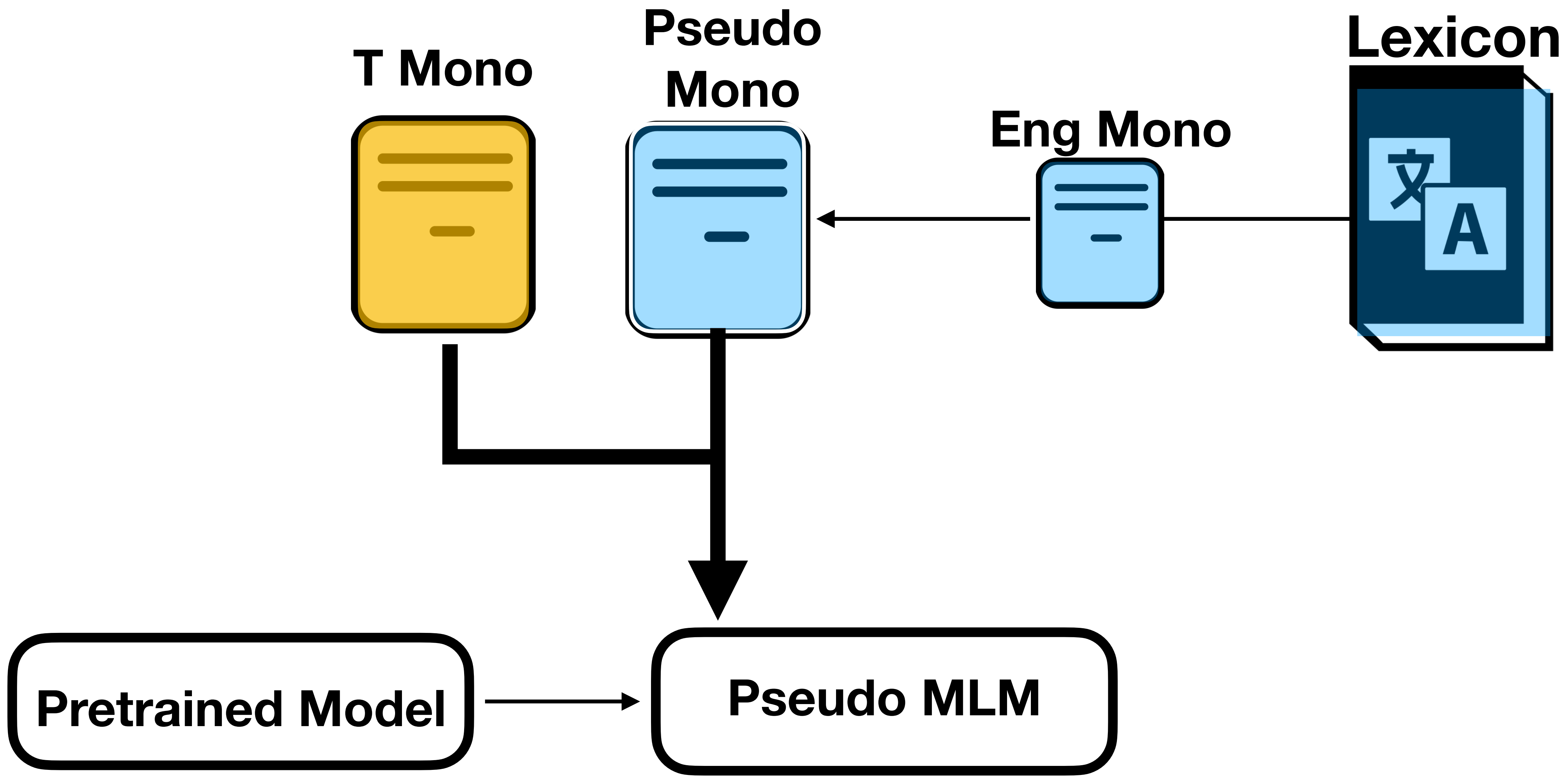




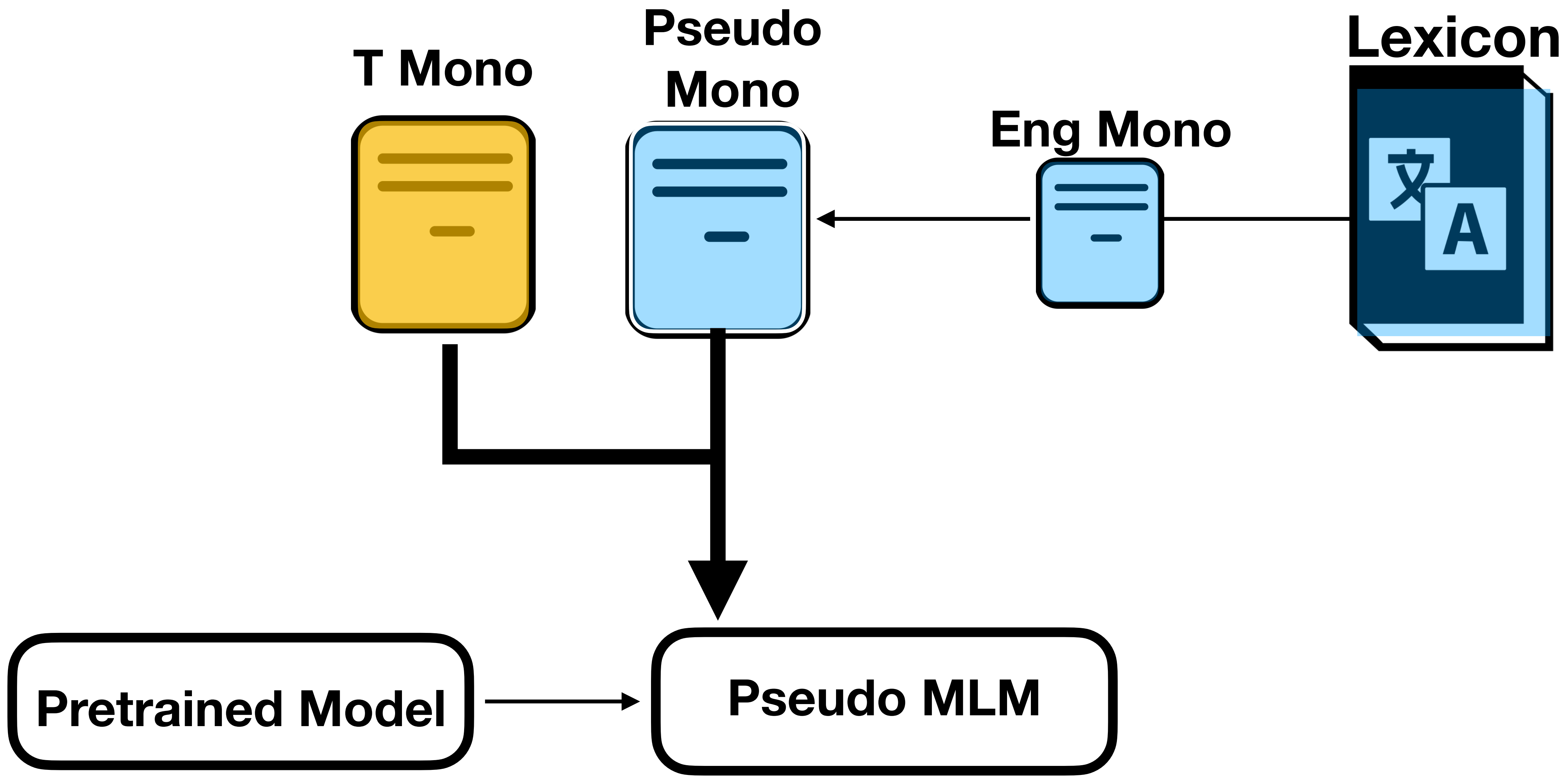
- Linguists have been documenting languages for years in formats such as lexicons
- PanLex: open-sourced database of lexicons with much better language coverage

https://panlex.org/

Synthesizing Data Using Lexicons

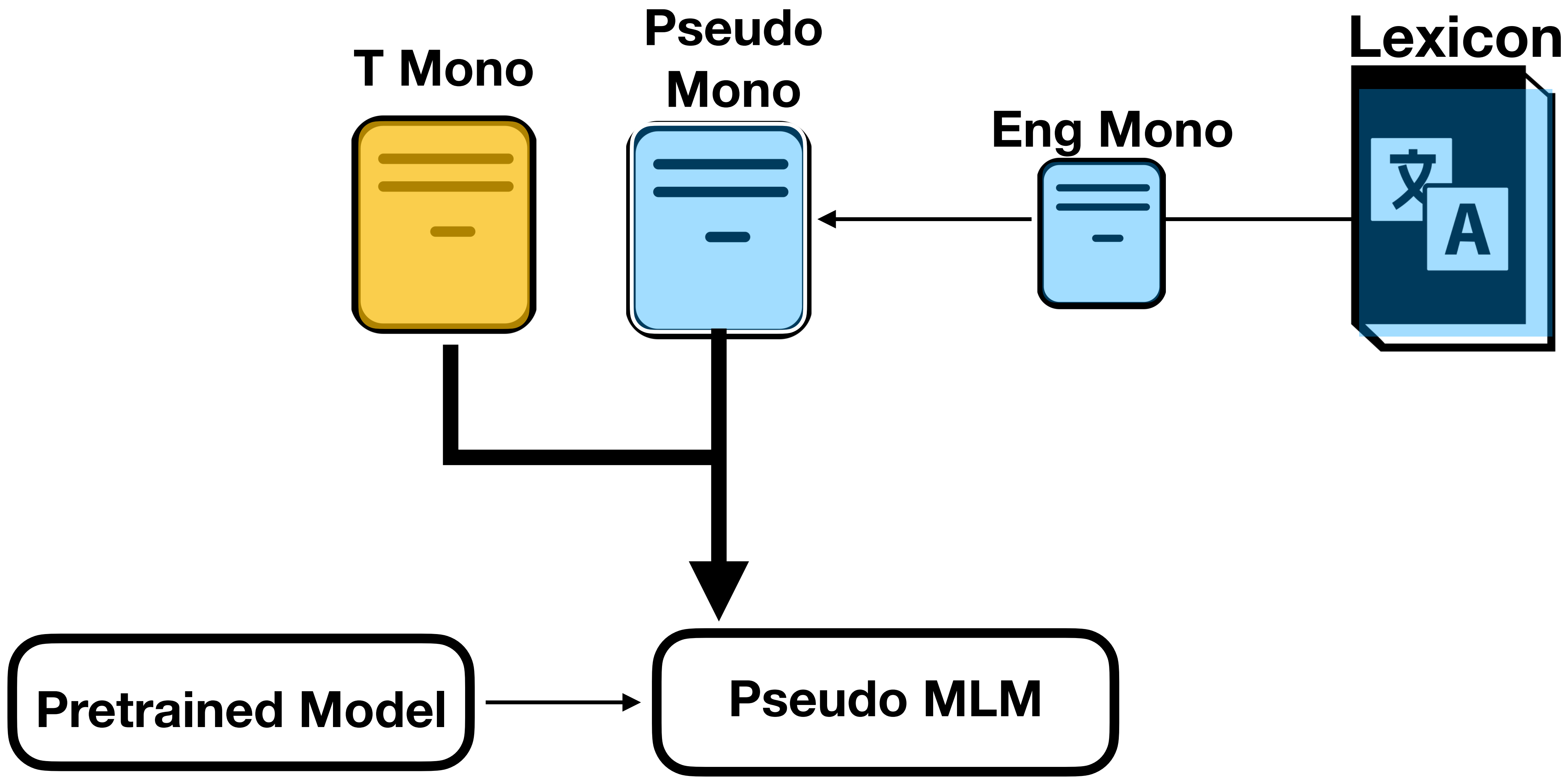


Synthesizing Data Using Lexicons



- Pseudo Mono Data: replace words in **English monolingual data** to its corresponding translation in the target language T

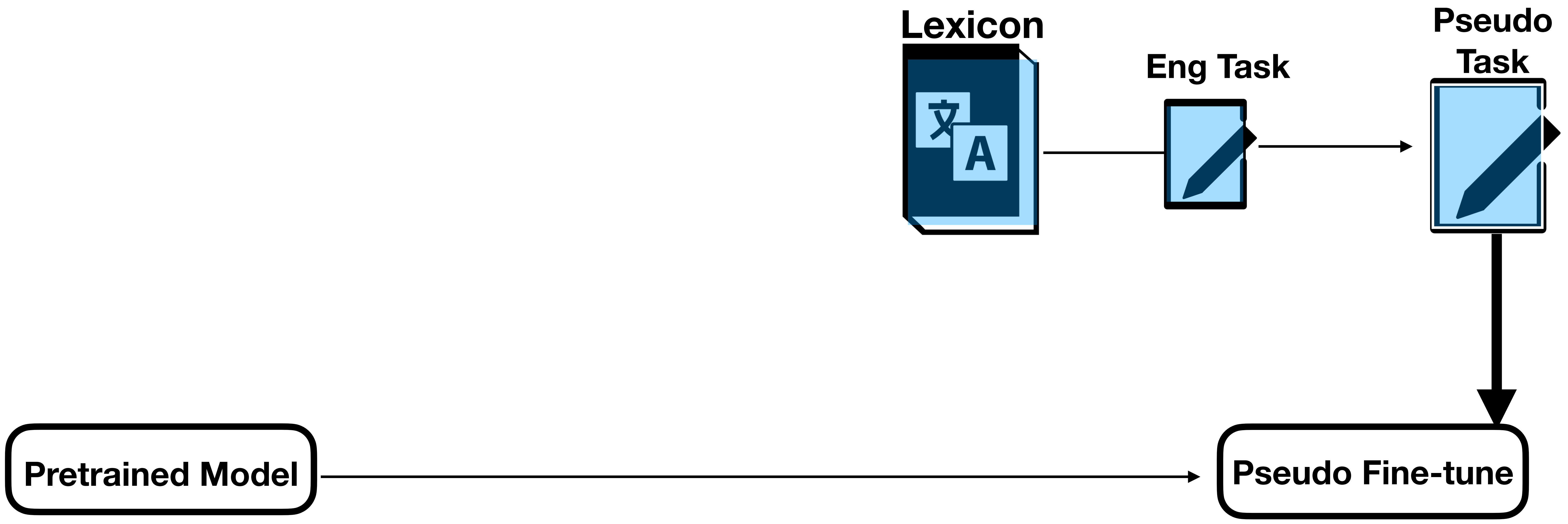
Synthesizing Data Using Lexicons



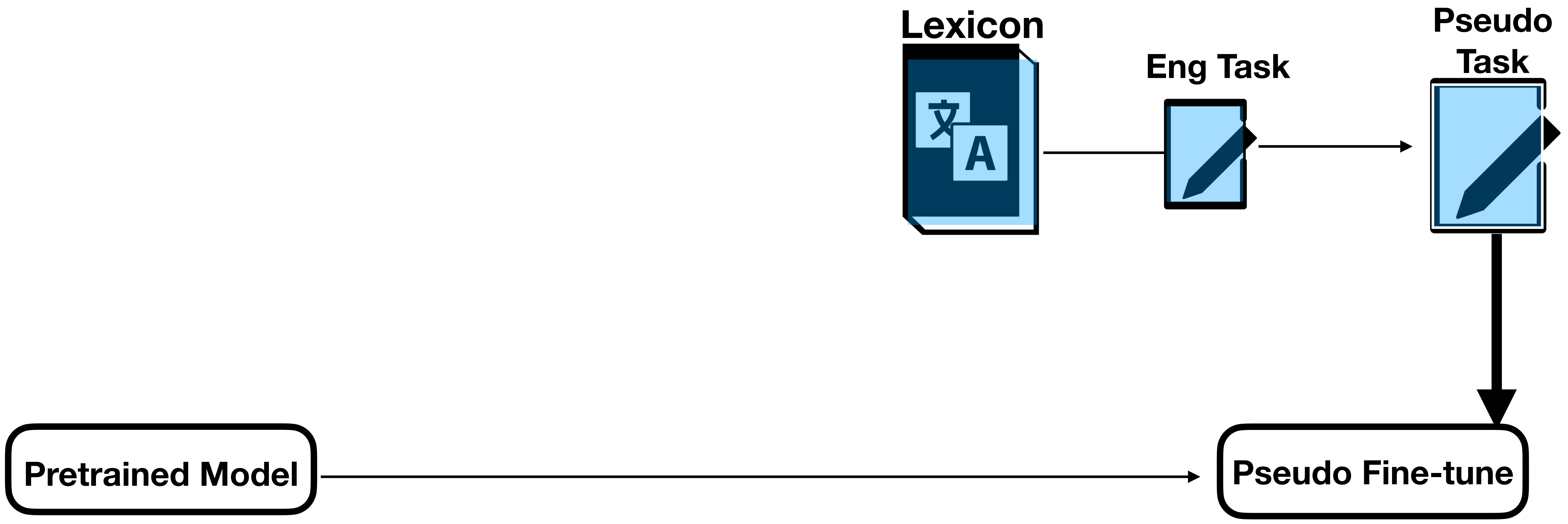
Eng Mono	Anarchism calls for the abolition of the state , which it holds to be undesirable , unnecessary , and harmful .
Pseudo Mono	Anarchism calls ghal il abolition ta' il stat , lima hi holds ghal tkun undesirable , bla bzonn , u harmful .

- Pseudo Mono Data: replace words in **English monolingual data** to its corresponding translation in the target language T

Synthesizing Data Using Lexicons

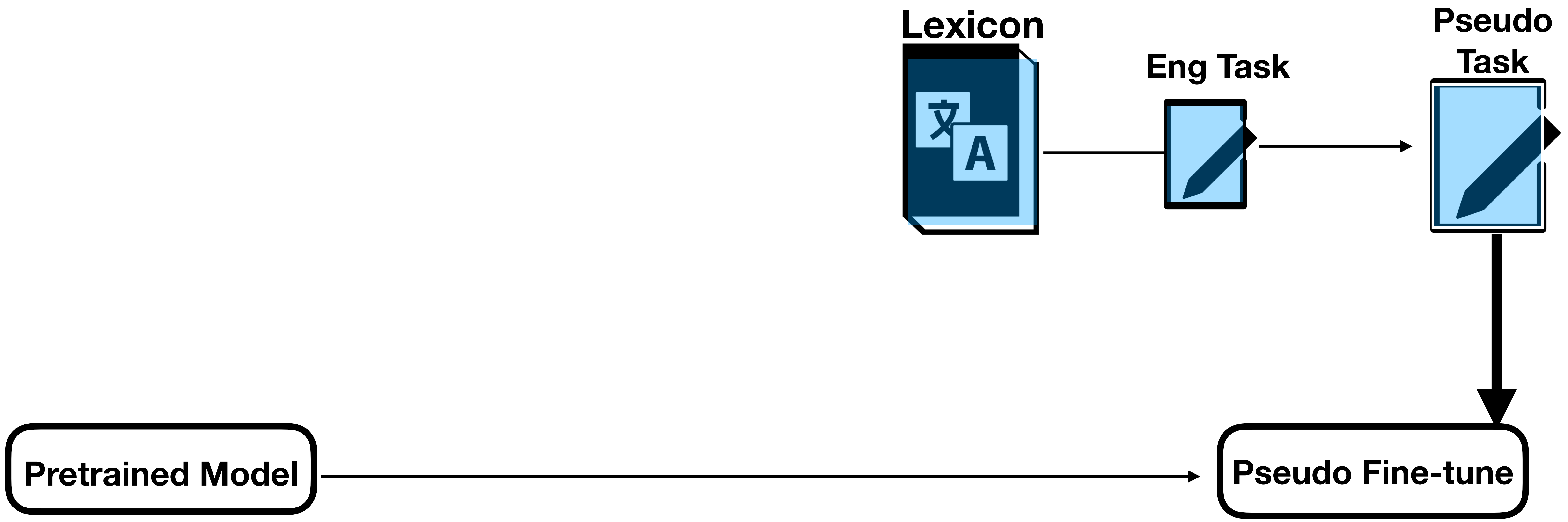


Synthesizing Data Using Lexicons



- Pseudo Task Data: replace words in **English task data** to its corresponding translation in the target language T

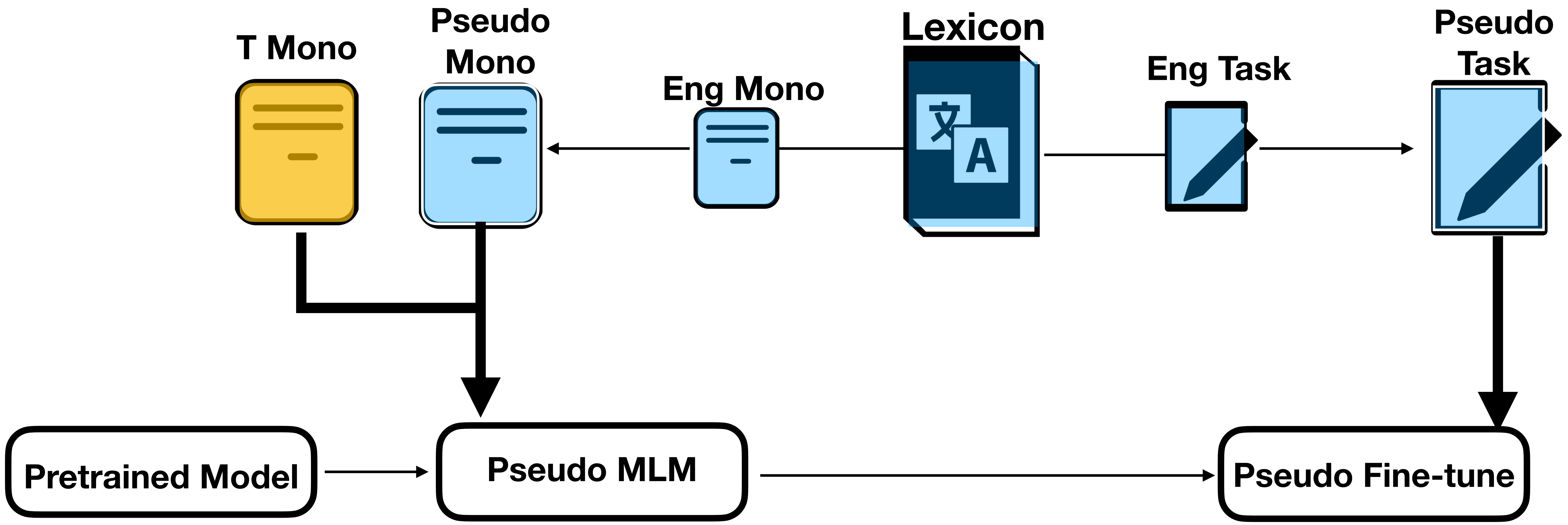
Synthesizing Data Using Lexicons



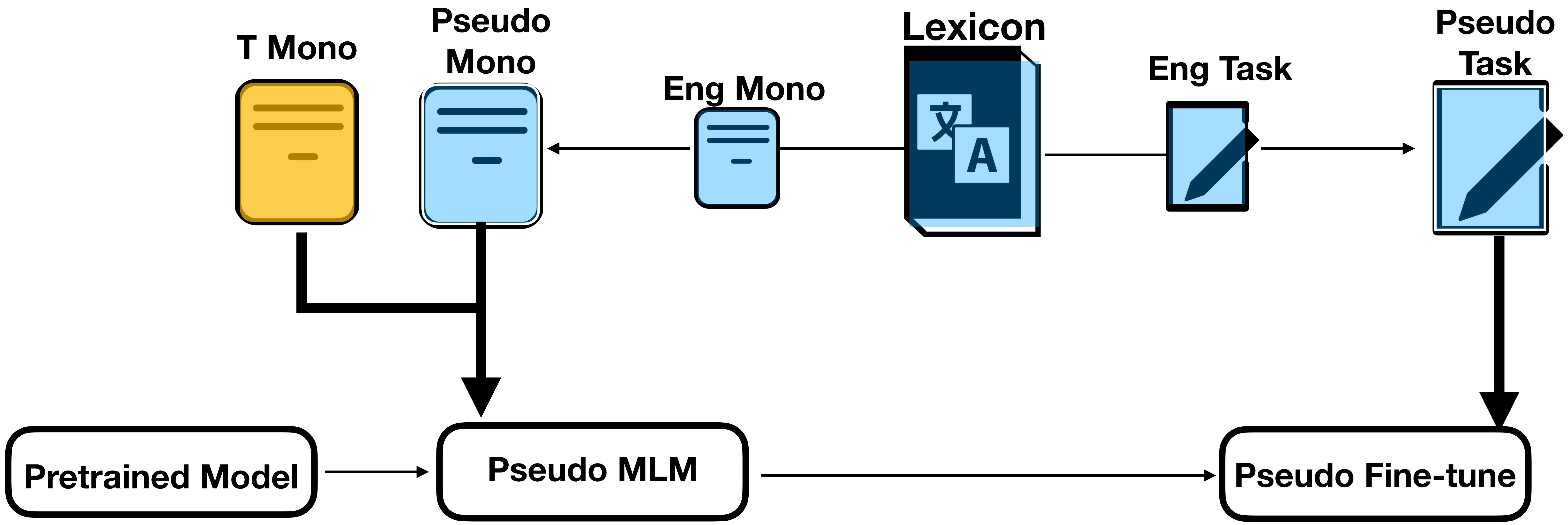
Eng Task	I suspect the streets of Baghdad will look as if a war is looming this week . PRON VERB DET NOUN ADP PROPN AUX VERB SCONJ SCONJ DET NOUN AUX VERB DET NOUN PUNCT
Pseudo Task	jien iddubita il streets ta' Bagdad xewqa hares kif jekk a gwerra is looming dan gimgha . PRON VERB DET NOUN ADP PROPN AUX VERB SCONJ SCONJ DET NOUN AUX VERB DET NOUN PUNCT

- Pseudo Task Data: replace words in **English task data** to its corresponding translation in the target language T

Synthesizing Data Using Lexicons



Synthesizing Data Using Lexicons



- Use either pseudo MLM or Pseudo Fine-tune, or both

Experiments

Experiments

- Model: mBERT

Experiments

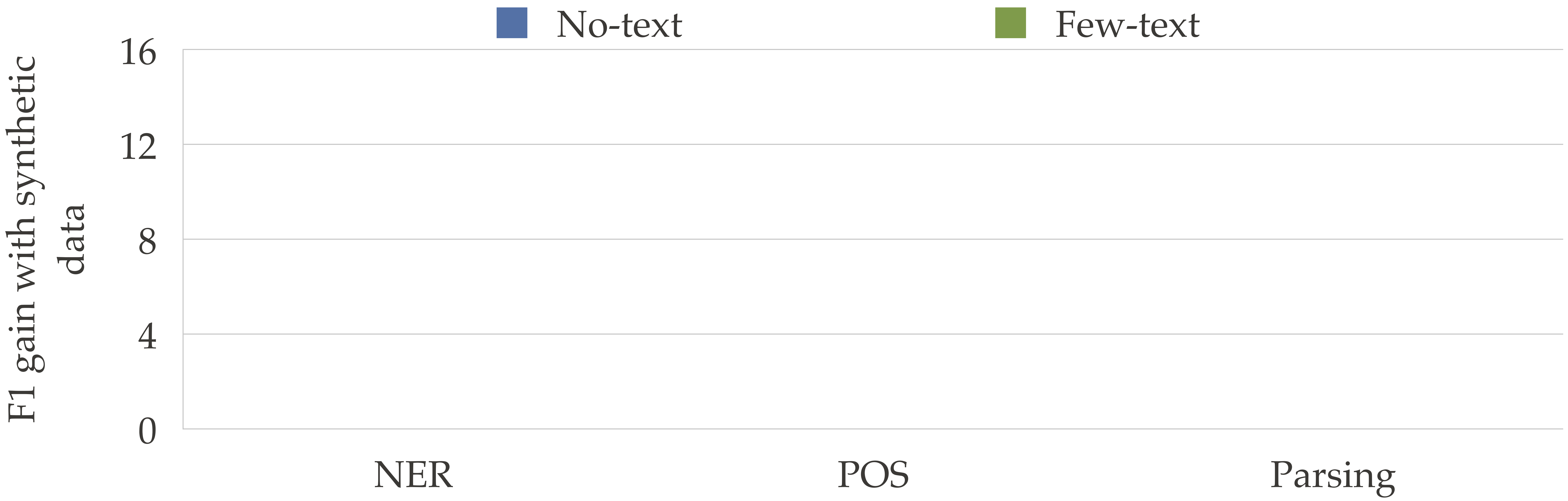
- Model: mBERT
- Tasks:
  - NER
  - POS tagging
  - Dependency Parsing

Experiments

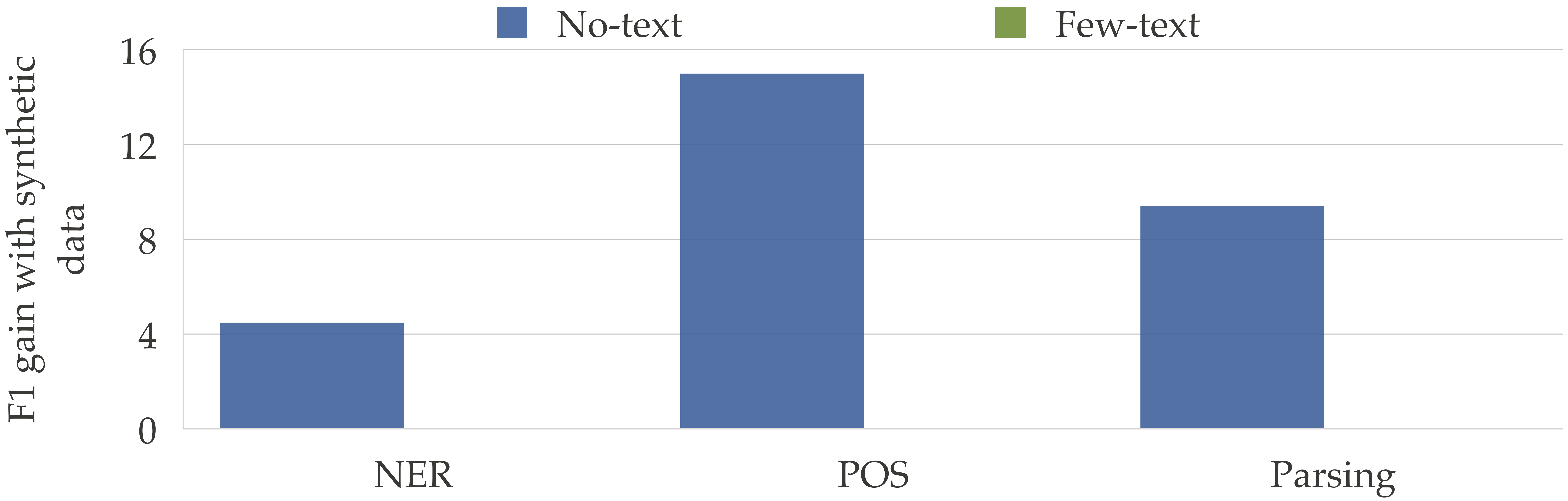
- Model: mBERT
- Tasks:
  - NER
  - POS tagging
  - Dependency Parsing
- Languages: 19 languages not covered by mBERT pretraining

Results

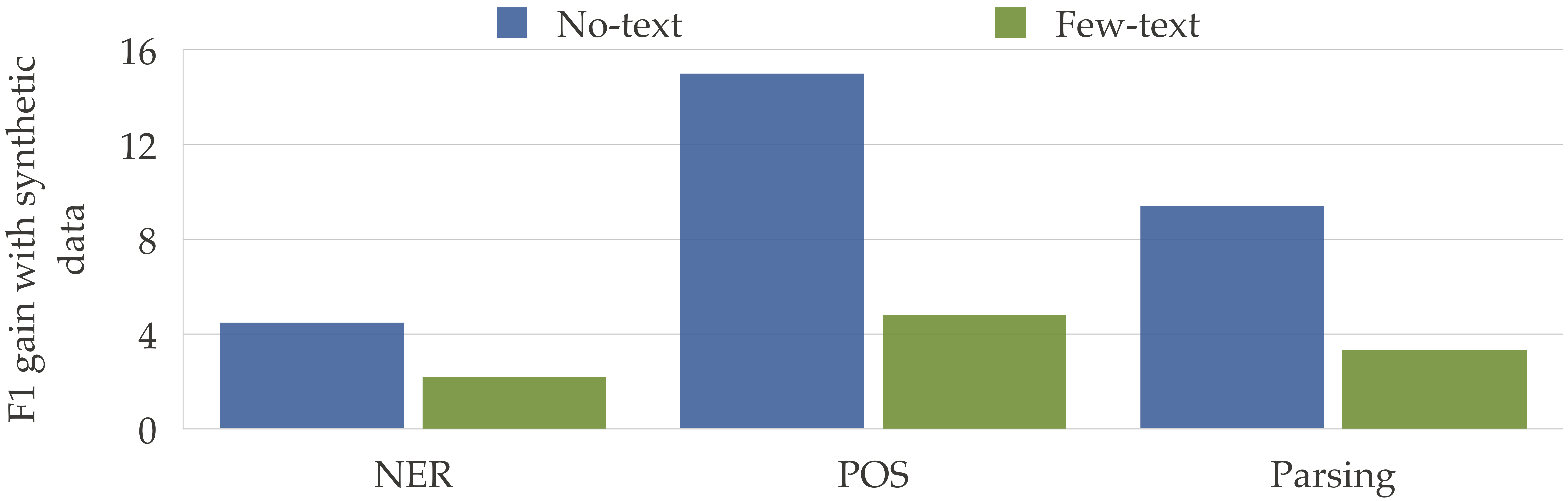
Results


No-text
Few-text
F1 gain with synthetic data
16
12
8
4
0
NER
POS
Parsing

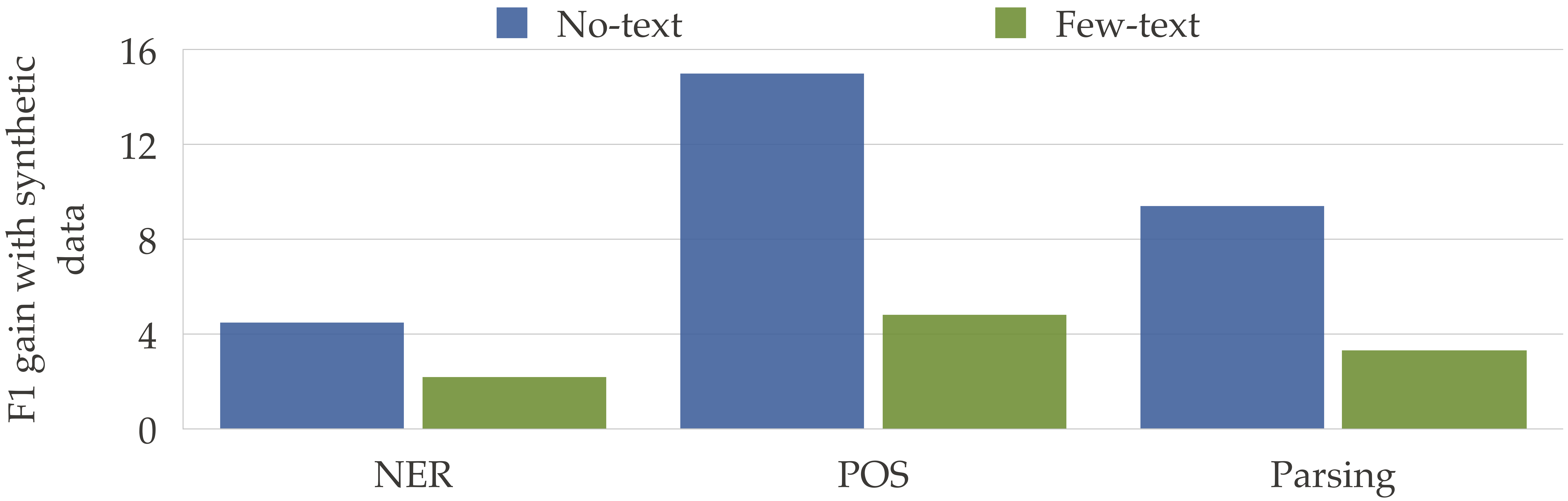
Results


No-text
Few-text
F1 gain with synthetic data
16
12
8
4
0
NER
POS
Parsing

Results


No-text
Few-text
F1 gain with synthetic data
16
12
8
4
0
NER
POS
Parsing

Results



- Using synthetic data leads to significant improvements for both no-text and few-text setting

Label Noise

Eng Task	I suspect the streets of Baghdad will look as if a war is looming this week . PRON VERB DET NOUN ADP PROPN AUX VERB SCONJ SCONJ DET NOUN AUX VERB DET NOUN PUNCT
Pseudo Task	jien iddubita il streets ta' Bagdad xewqa hares kif jekk a gwerra is looming dan gimgha . PRON VERB DET NOUN ADP PROPN AUX VERB SCONJ SCONJ DET NOUN AUX VERB DET NOUN PUNCT

Label Noise

Eng Task	I suspect the streets of Baghdad will look as if a war is looming this week . PRON VERB DET NOUN ADP PROPN AUX VERB SCONJ SCONJ DET NOUN AUX VERB DET NOUN PUNCT
Pseudo Task	jien iddubita il streets ta' Bagdad xewqa hares kif jekk a gwerra is looming dan gimgha . PRON VERB DET NOUN ADP PROPN AUX VERB SCONJ SCONJ DET NOUN AUX VERB DET NOUN PUNCT

- "xewqa" is a noun meaning "desire, will"

Label Noise

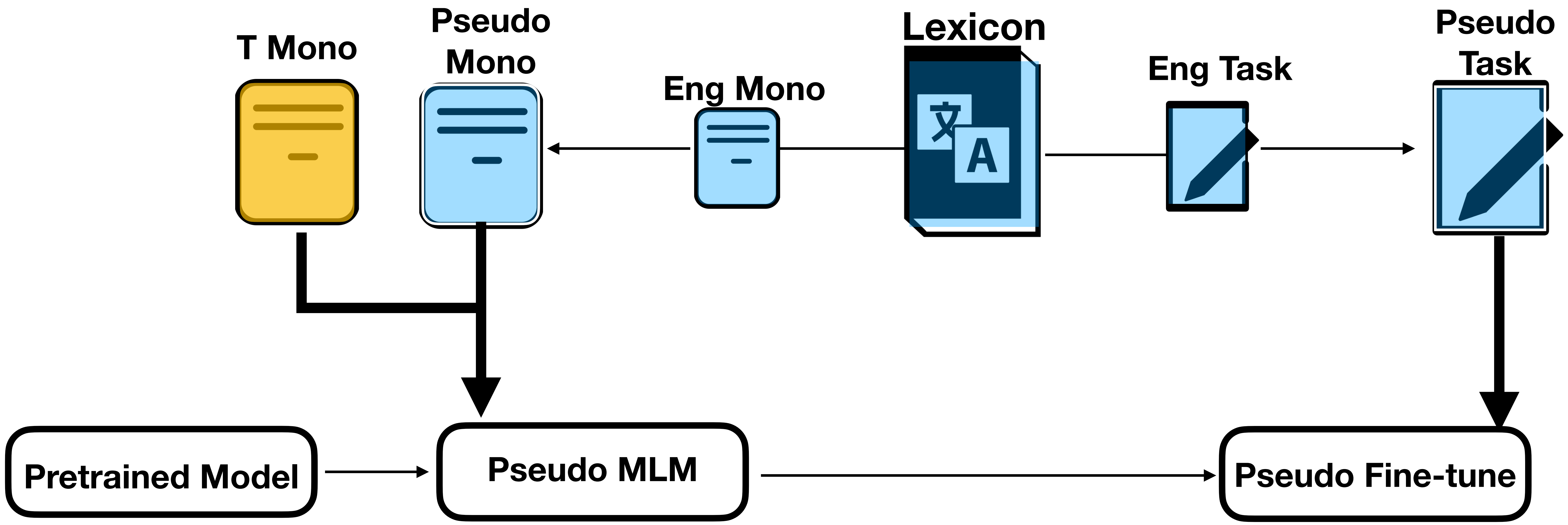
Eng Task	I suspect the streets of Baghdad will look as if a war is looming this week . PRON VERB DET NOUN ADP PROPN AUX VERB SCONJ SCONJ DET NOUN AUX VERB DET NOUN PUNCT
Pseudo Task	jien iddubita il streets ta' Bagdad xewqa hares kif jekk a gwerra is looming dan gimgha . PRON VERB DET NOUN ADP PROPN AUX VERB SCONJ SCONJ DET NOUN AUX VERB DET NOUN PUNCT

- "xewqa" is a noun meaning "desire, will"
- But the original English POS tag is inconsistent with the replaced word

Label Noise

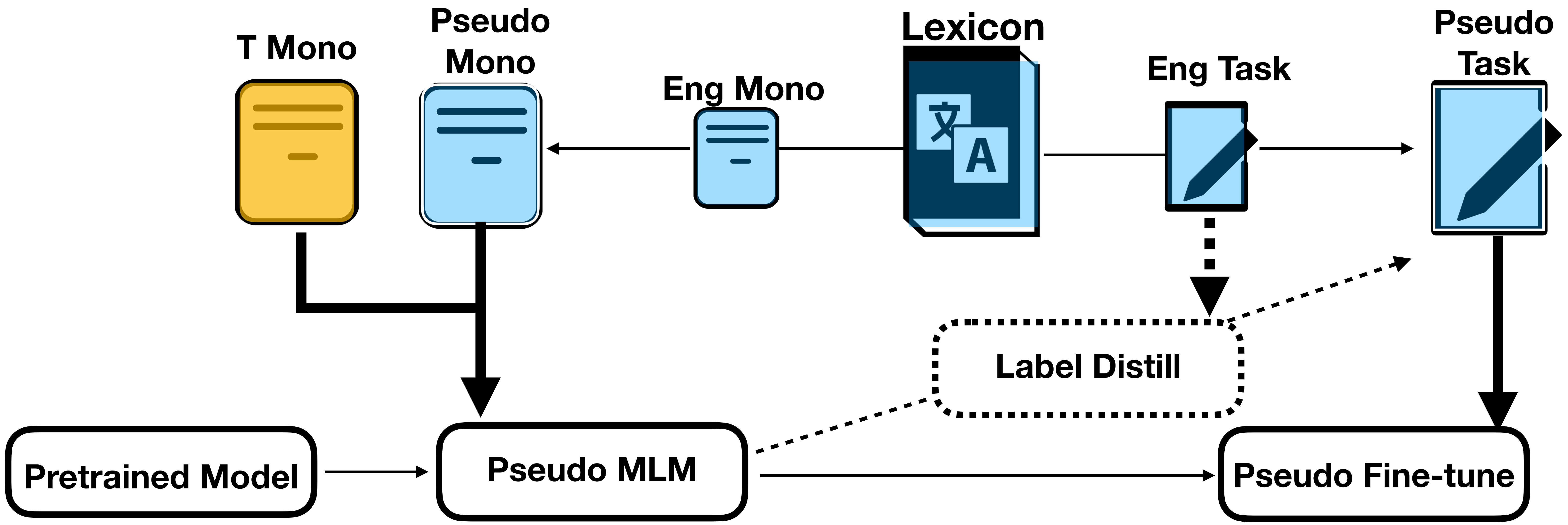
- Use the fine-tuned model to “correct” the labels for the Pseudo task data

Label Noise



- Use the fine-tuned model to “correct” the labels for the Pseudo task data

Label Noise



- Use the fine-tuned model to "correct" the labels for the Pseudo task data

Label Noise

Eng Task	I suspect the streets of Baghdad will look as if a war is looming this week . PRON VERB DET NOUN ADP PROPN AUX VERB SCONJ SCONJ DET NOUN AUX VERB DET NOUN PUNCT
Pseudo Task	jien iddubita il streets ta' Bagdad xewqa hares kif jekk a gwerra is looming dan gimgha . PRON VERB DET NOUN ADP PROPN AUX VERB SCONJ SCONJ DET NOUN AUX VERB DET NOUN PUNCT

- "xewqa" is a noun meaning "desire,will"

Label Noise

Eng Task	I suspect the streets of Baghdad will look as if a war is looming this week . PRON VERB DET NOUN ADP PROPN AUX VERB SCONJ SCONJ DET NOUN AUX VERB DET NOUN PUNCT
Pseudo Task	jien iddubita il streets ta' Bagdad xewqa hares kif jekk a gwerra is looming dan gimgha . PRON VERB DET NOUN ADP PROPN AUX VERB SCONJ SCONJ DET NOUN AUX VERB DET NOUN PUNCT
Label Distilled	PRON VERB DET NOUN ADP PROPN NOUN NOUN SCONJ SCONJ DET NOUN AUX VERB DET NOUN PUNCT

- "xewqa" is a noun meaning "desire, will"

Label Noise

Eng Task	I suspect the streets of Baghdad will look as if a war is looming this week . PRON VERB DET NOUN ADP PROPN AUX VERB SCONJ SCONJ DET NOUN AUX VERB DET NOUN PUNCT
Pseudo Task	jien iddubita il streets ta' Bagdad xewqa hares kif jekk a gwerra is looming dan gimgha . PRON VERB DET NOUN ADP PROPN AUX VERB SCONJ SCONJ DET NOUN AUX VERB DET NOUN PUNCT
Label Distilled	PRON VERB DET NOUN ADP PROPN NOUN NOUN SCONJ SCONJ DET NOUN AUX VERB DET NOUN PUNCT

- "xewqa" is a noun meaning "desire, will"
- The model is able to assign the correct label of noun

Label Noise

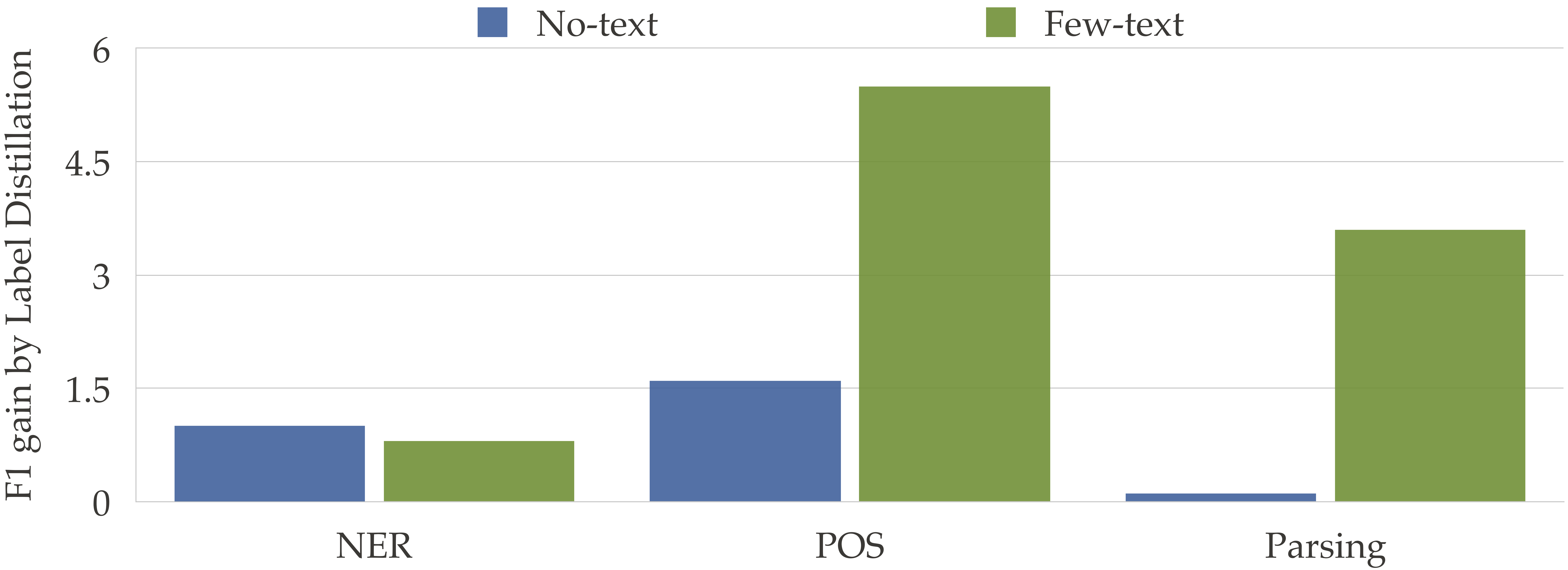
Label Noise


No-text
Few-text
F1 gain by Label Distillation
6
4.5
3
1.5
0
NER
POS
Parsing

Label Noise


No-text
Few-text
F1 gain by Label Distillation
6
4.5
3
1.5
0
NER
POS
Parsing

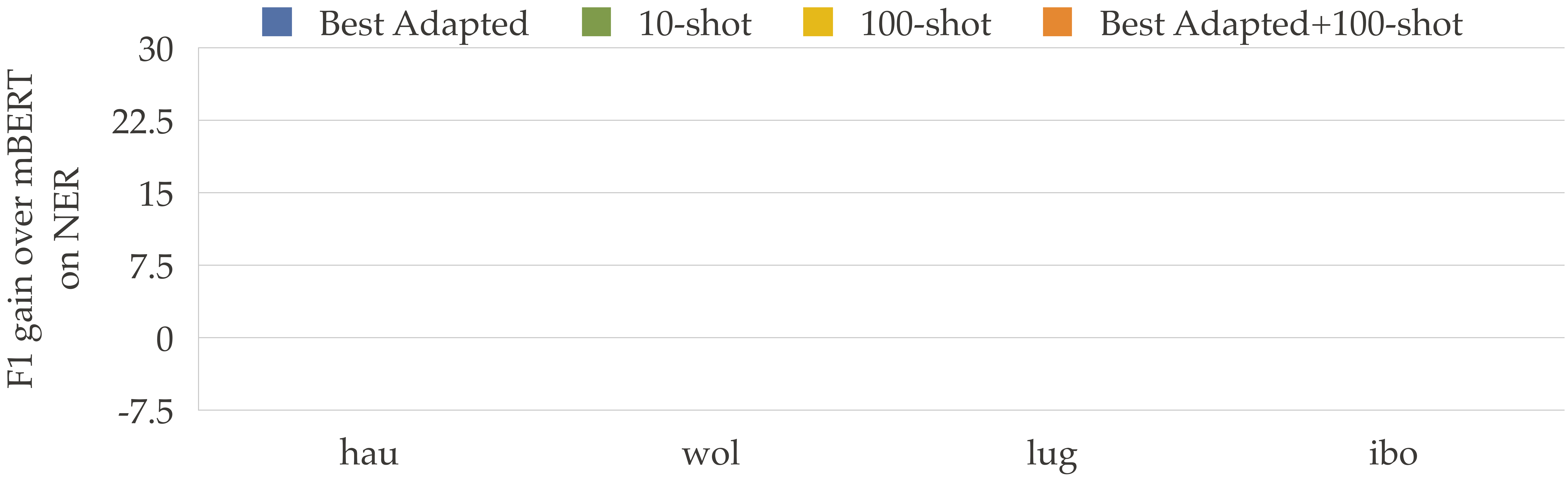
Label Noise



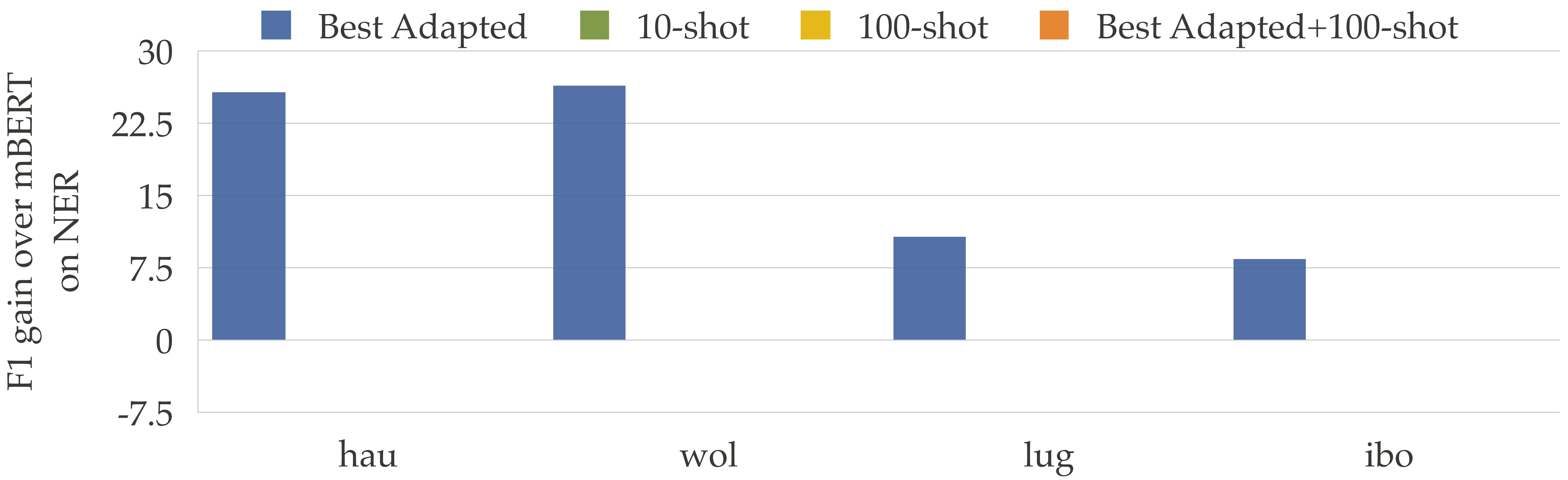
- Label Distillation is especially helpful for syntactic tasks

Comparison to Few-shot Learning

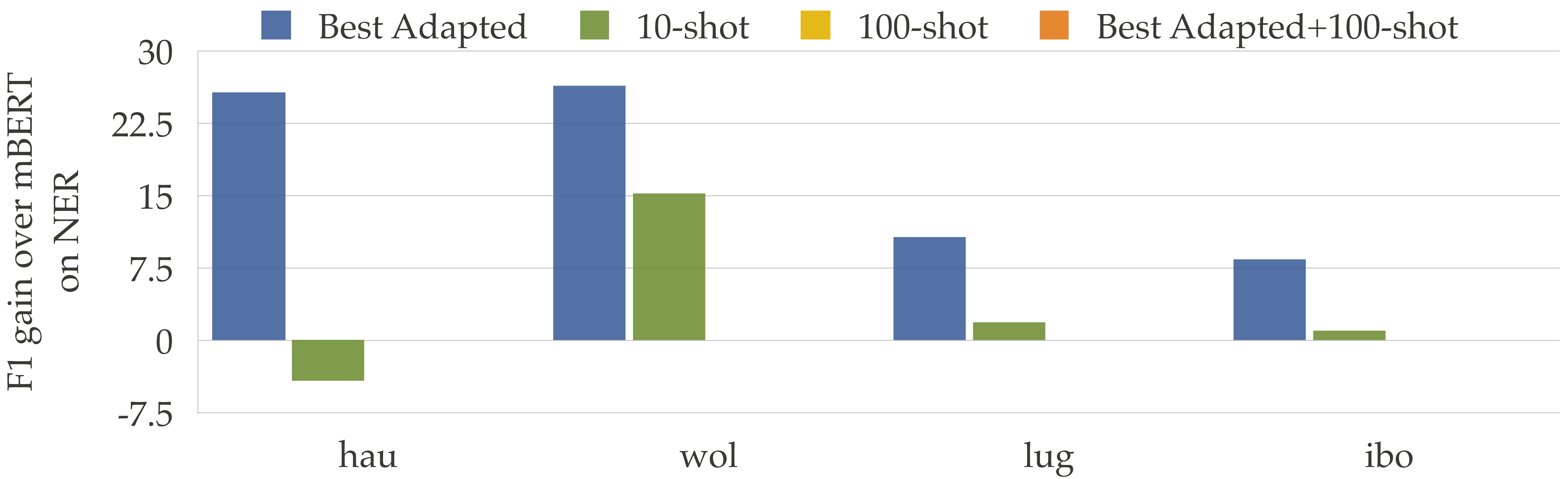
Comparison to Few-shot Learning



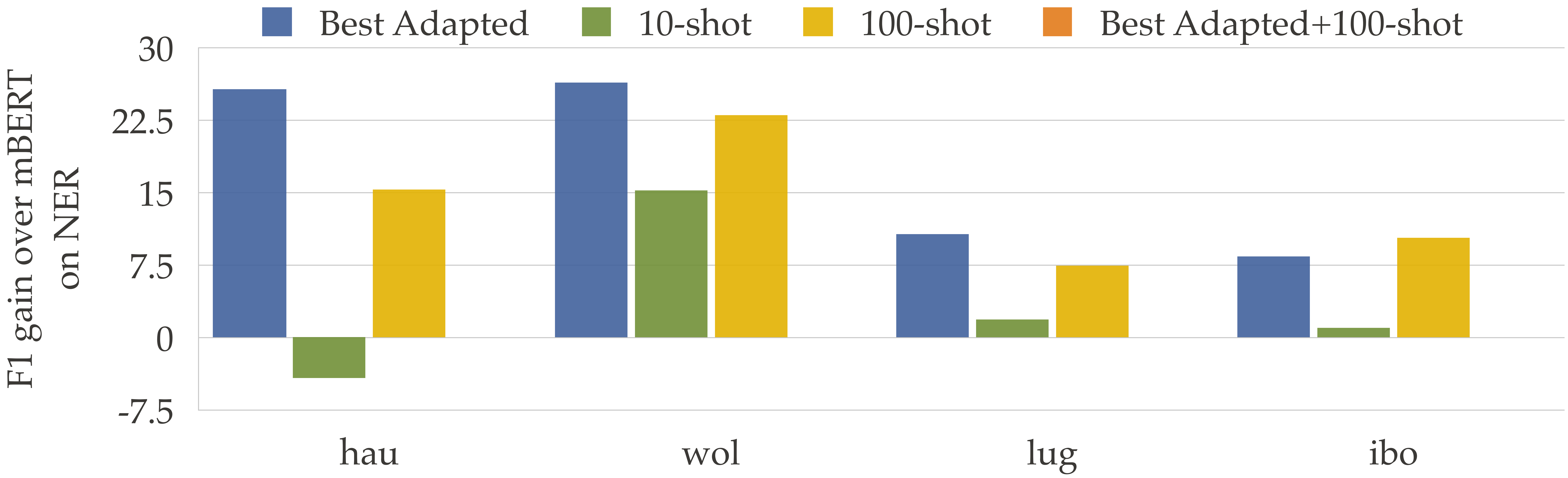
Comparison to Few-shot Learning



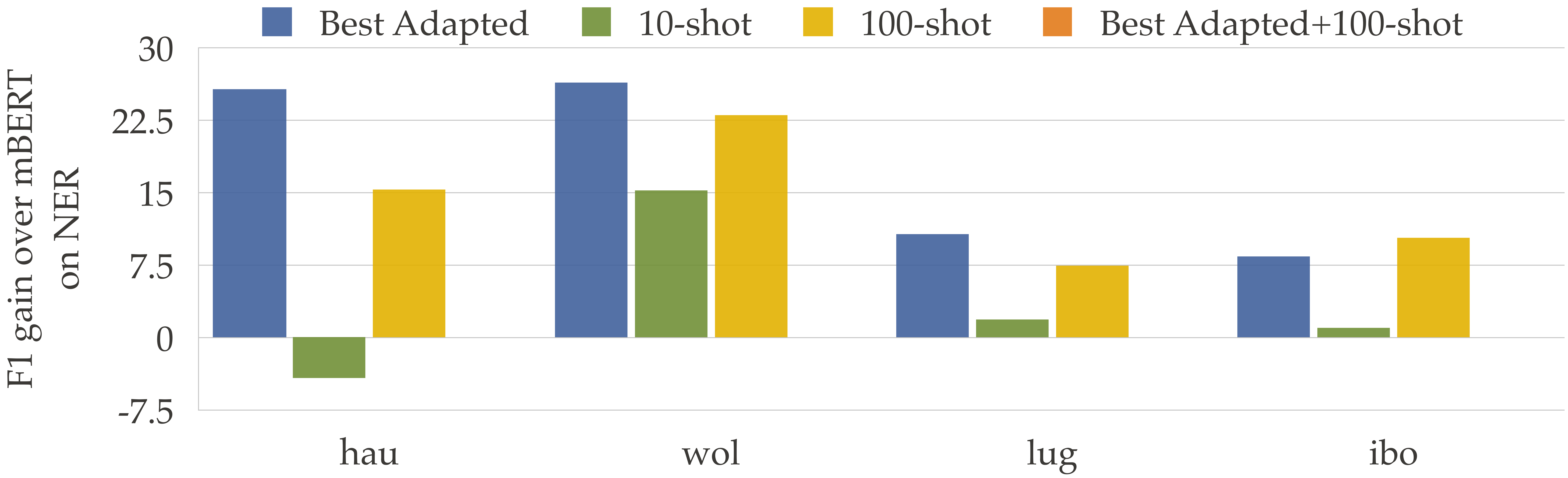
Comparison to Few-shot Learning



Comparison to Few-shot Learning

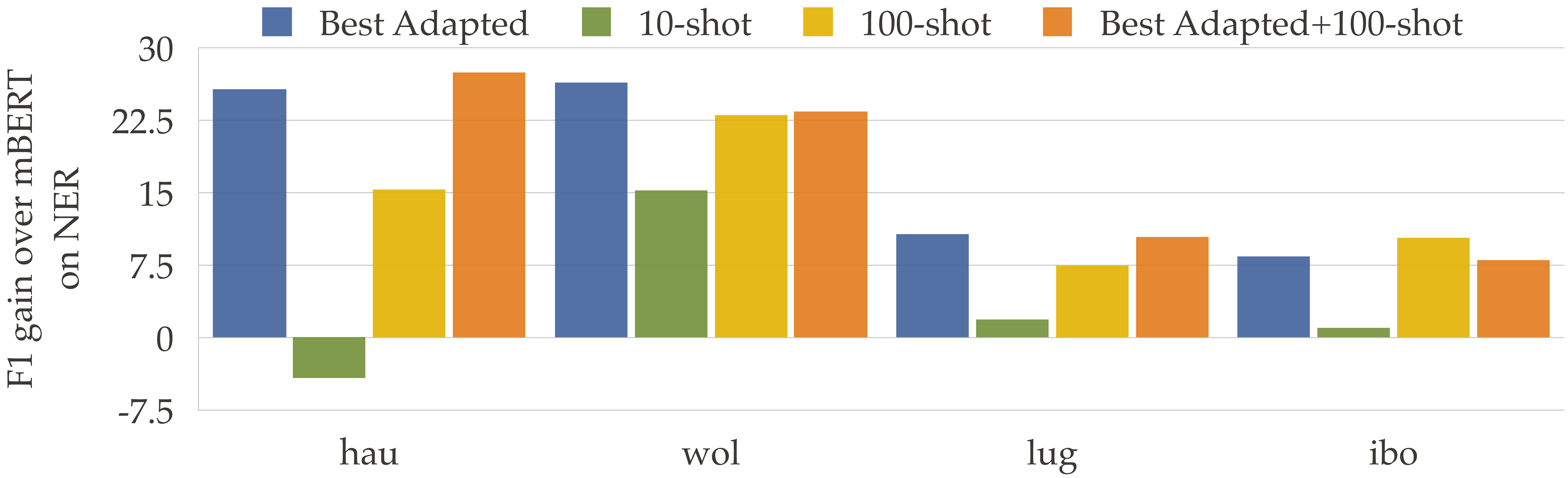


Comparison to Few-shot Learning



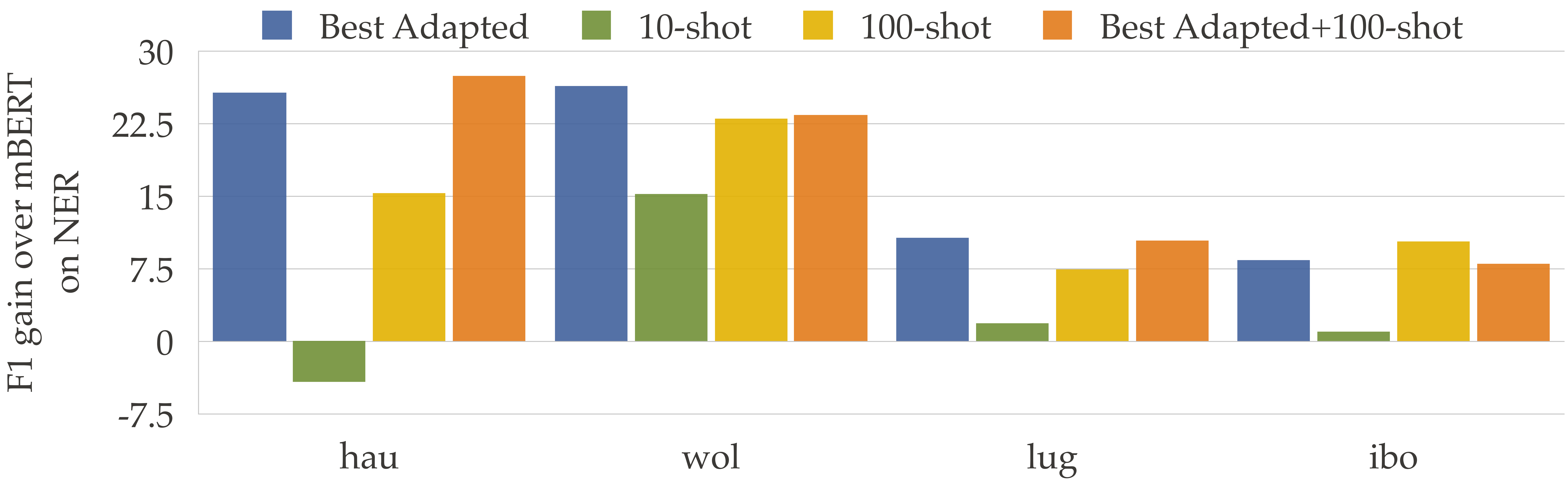
- Few-shot learning needs more annotated data for languages with limited text

Comparison to Few-shot Learning



- Few-shot learning needs more annotated data for languages with limited text

Comparison to Few-shot Learning

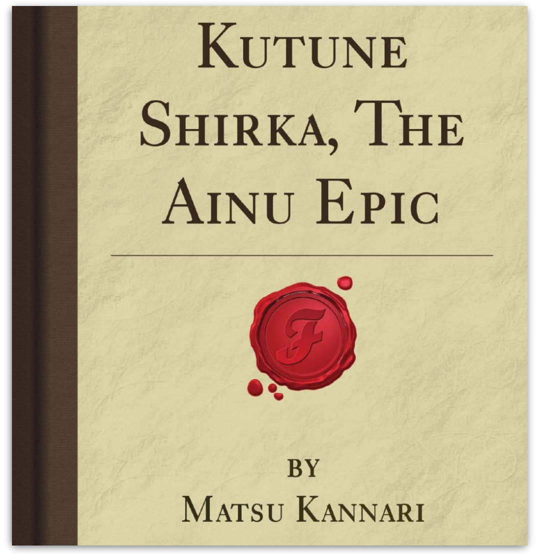


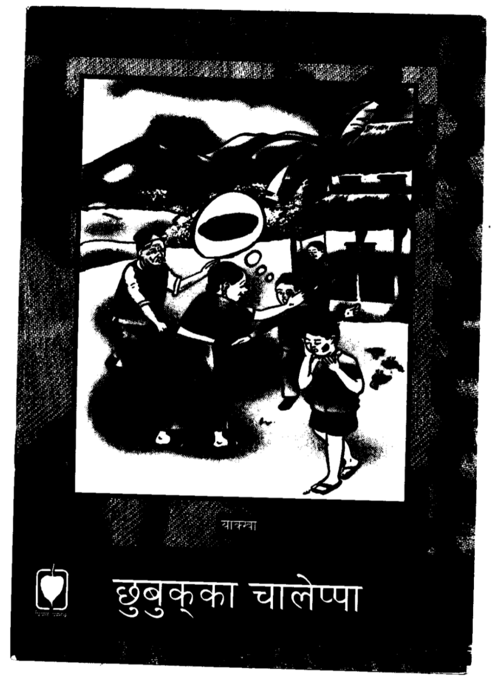
- Few-shot learning needs more annotated data for languages with limited text
- Combining adaptation and few-shot doesn't bring consistent improvements

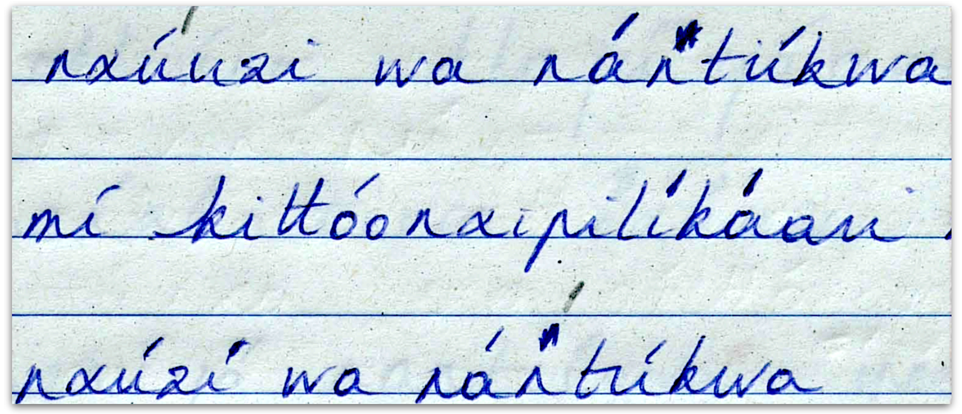
Conclusion

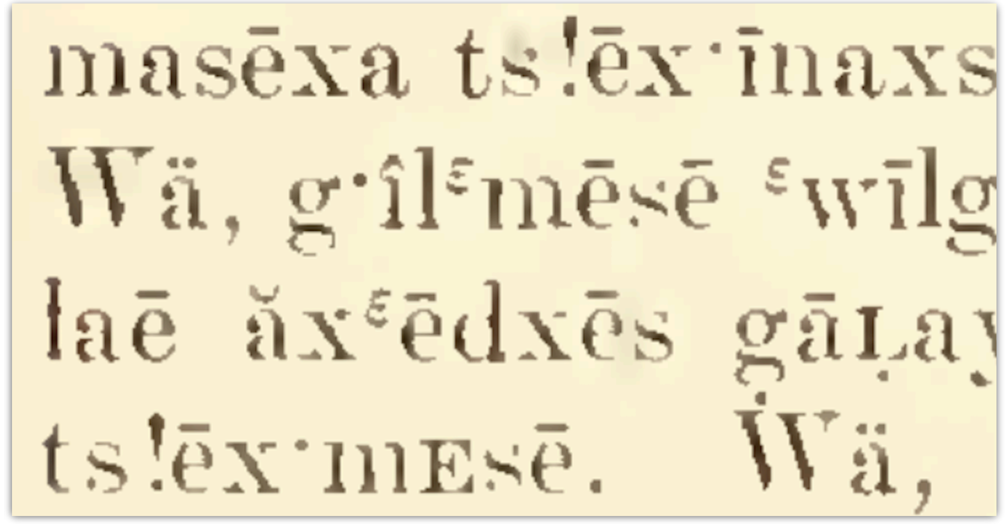
Conclusion

Conclusion



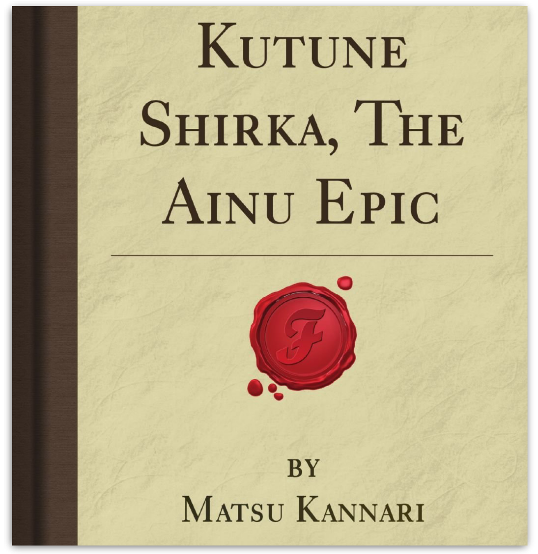


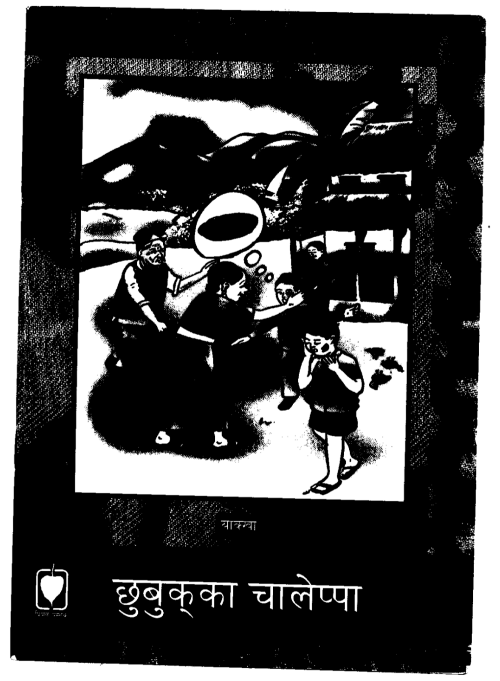


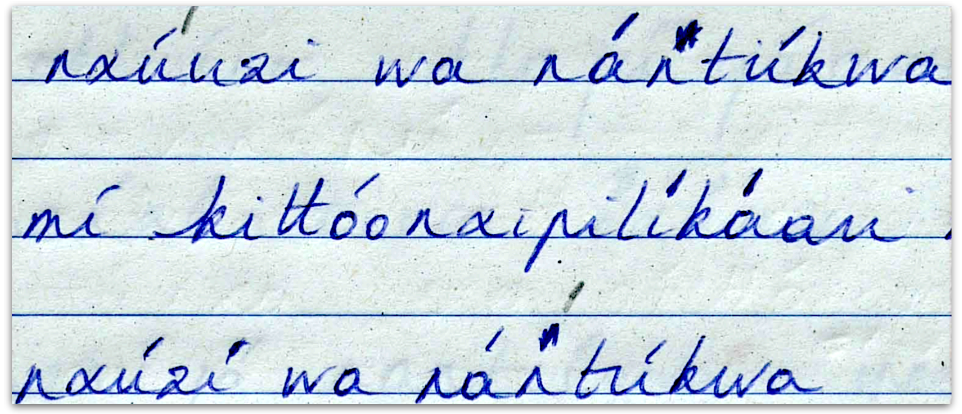


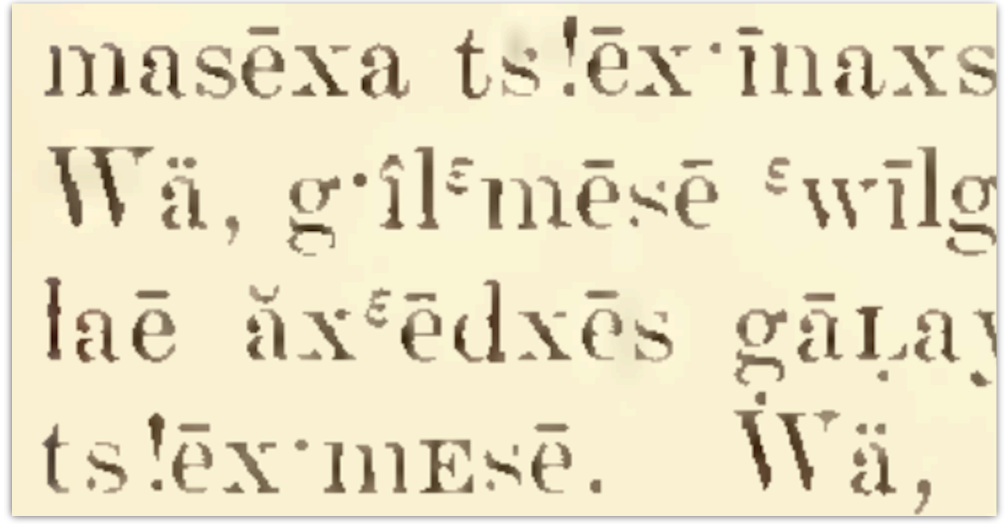
PanLex

Conclusion





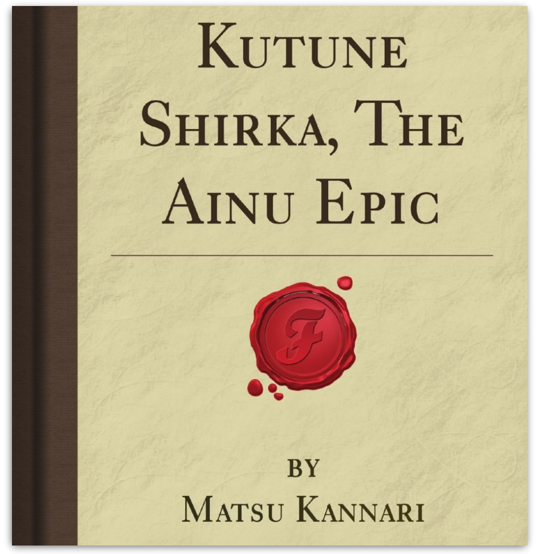


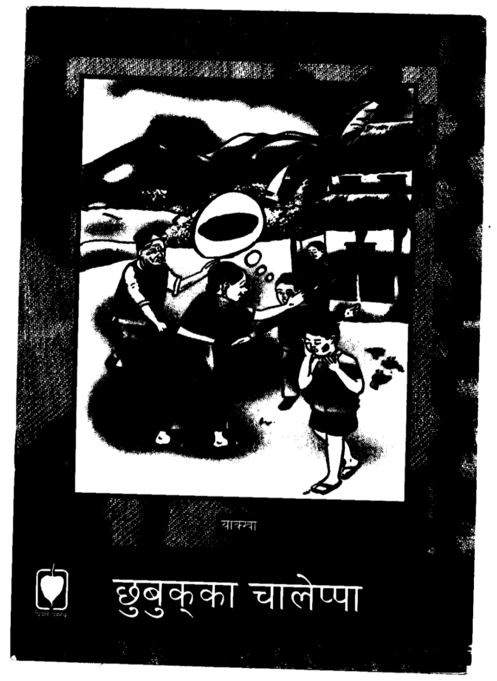


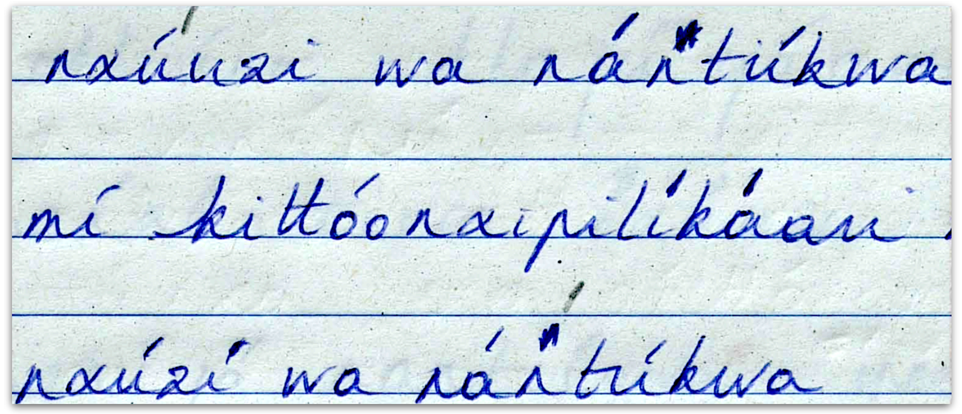
PanLex

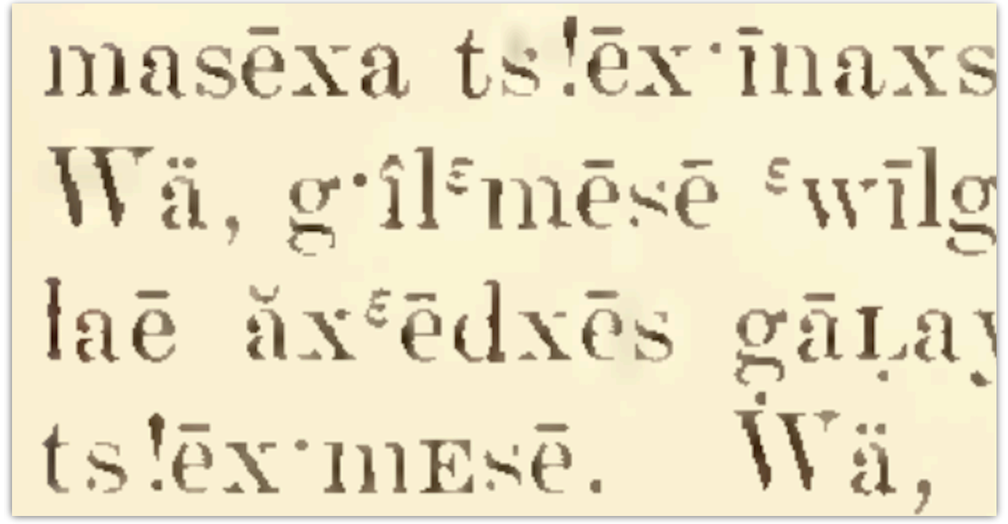
- Methods to **unlock new resources** for human or machine use in under-resourced languages

Conclusion





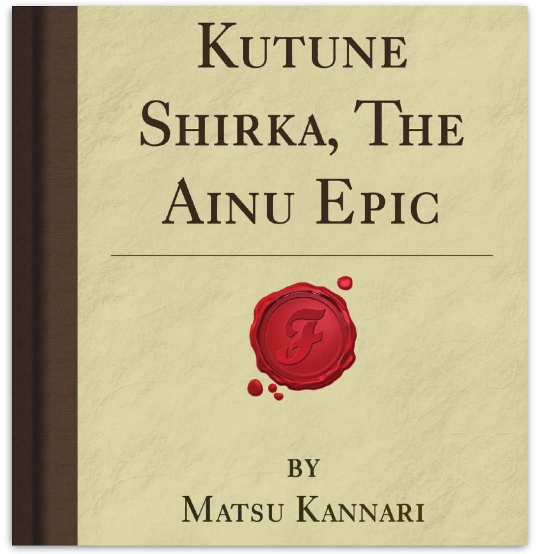


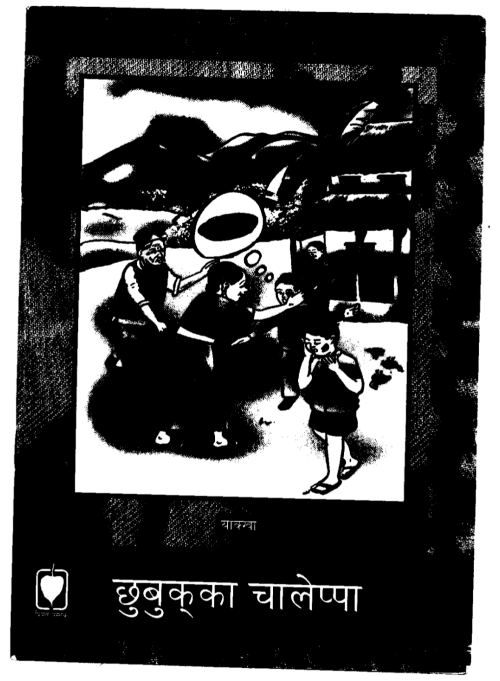


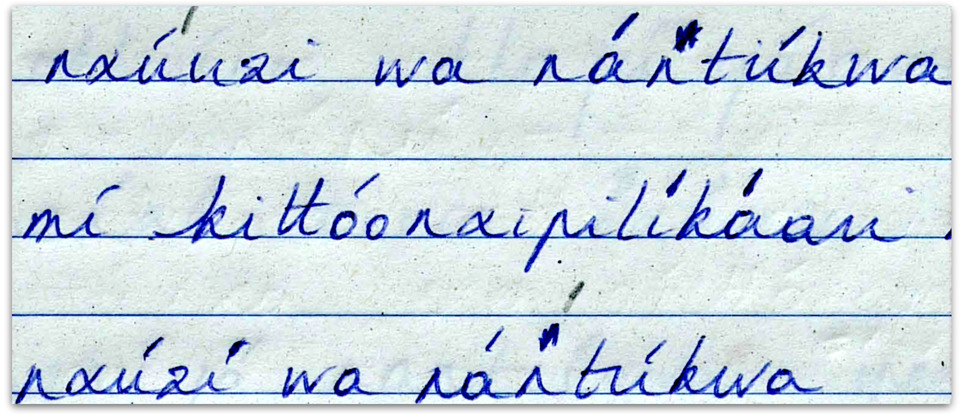
PanLex

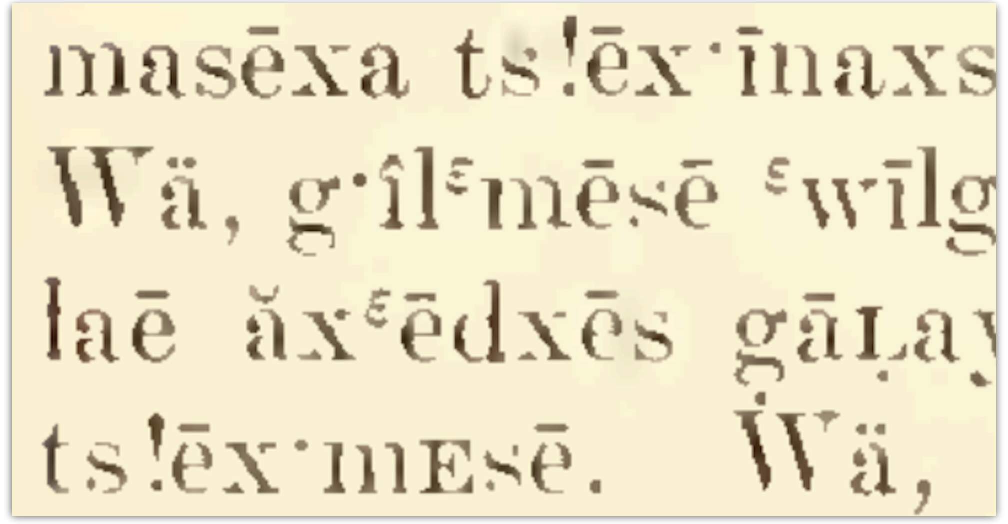
- Methods to **unlock new resources** for human or machine use in under-resourced languages
- What's next?

Conclusion





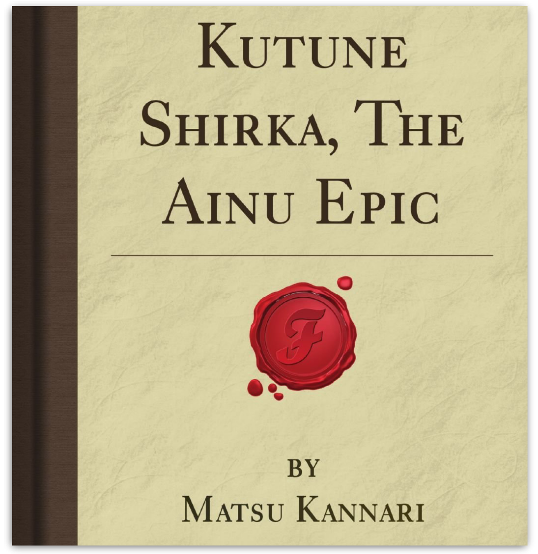


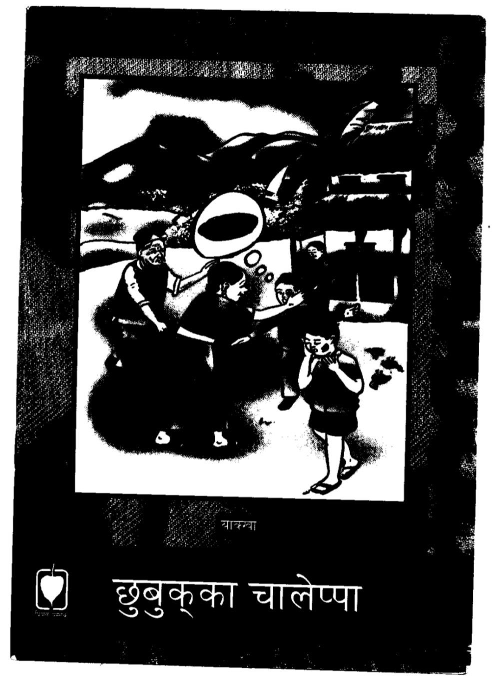


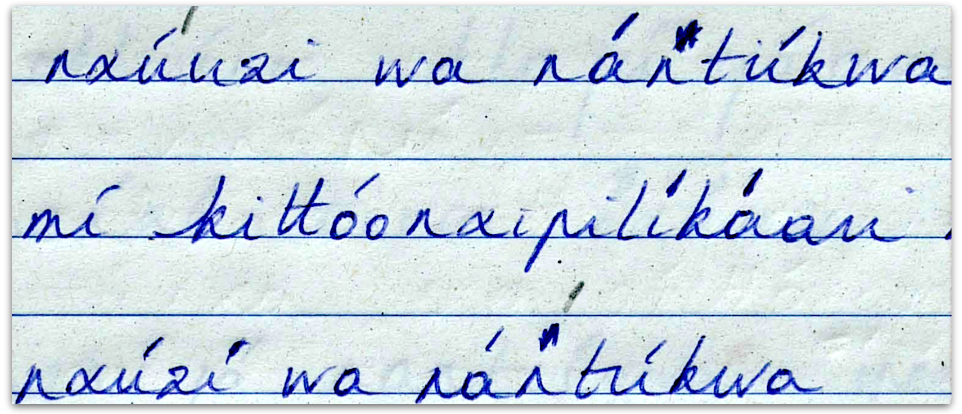
PanLex

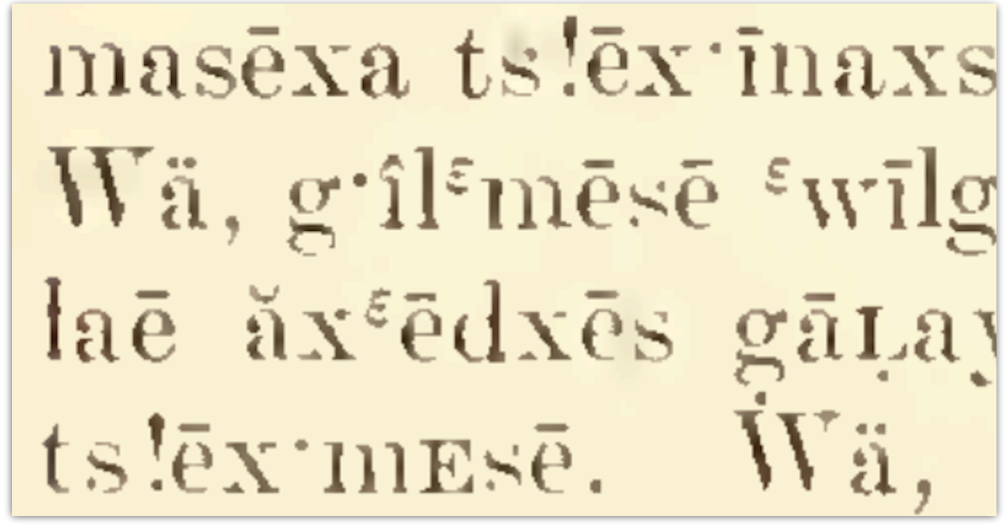
- Methods to **unlock new resources** for human or machine use in under-resourced languages
- What's next?
- Should we **put some linguistics in the models**?

Conclusion



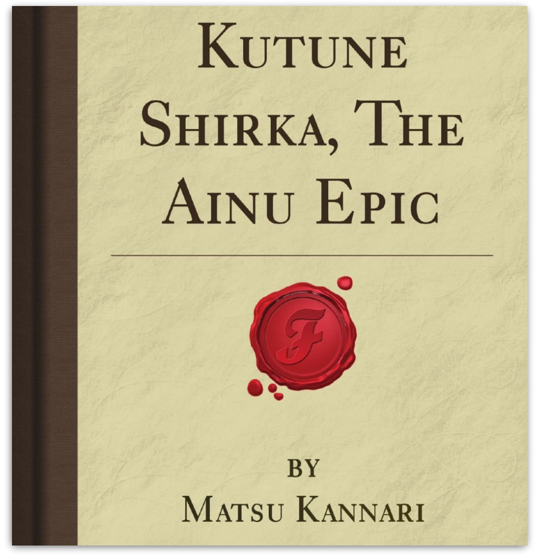


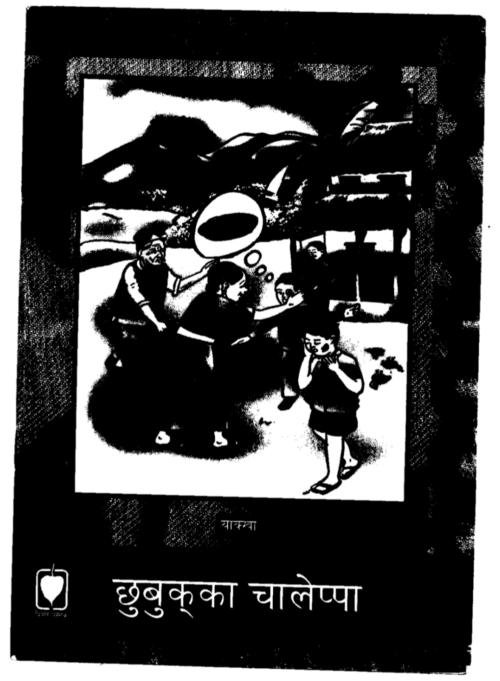


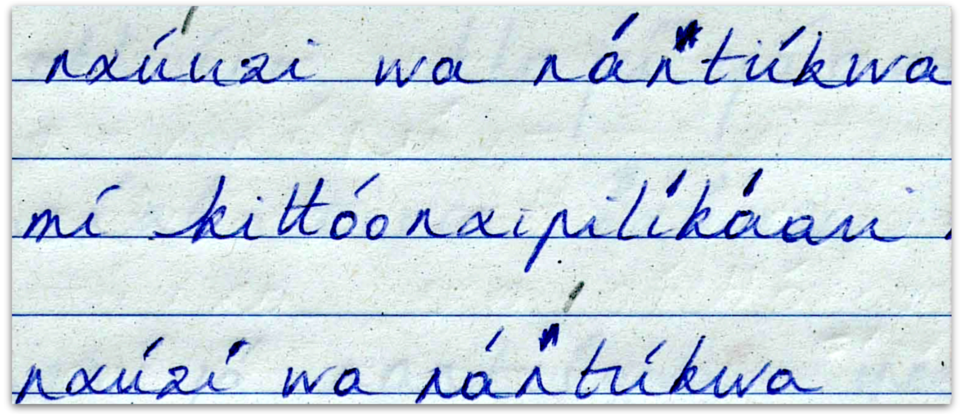


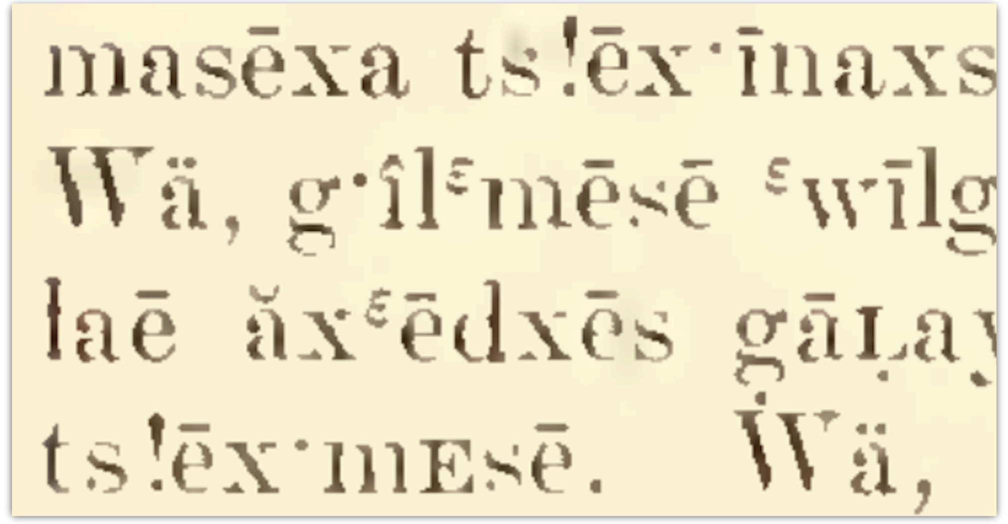
- Methods to **unlock new resources** for human or machine use in under-resourced languages
- What's next?
- Should we **put some linguistics in the models**?
  - Morphologically aware soft constraints for OCR?

Conclusion





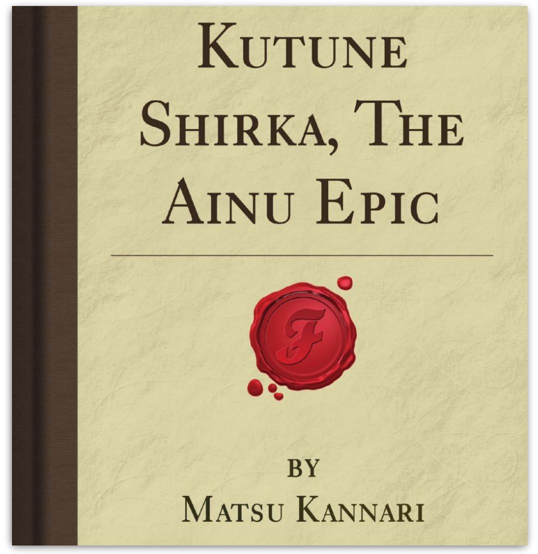


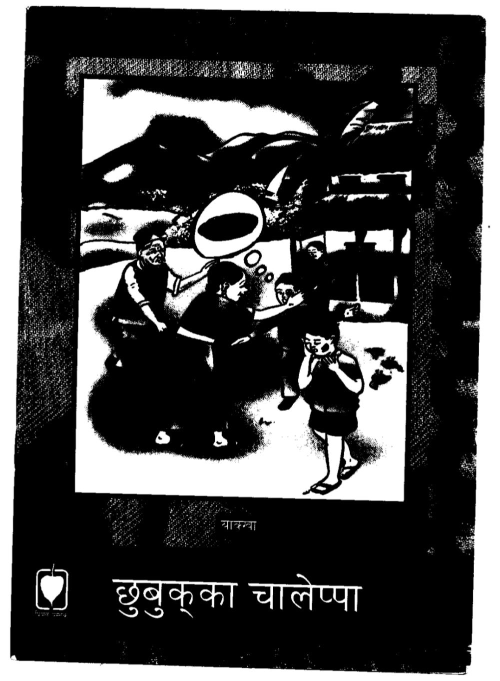


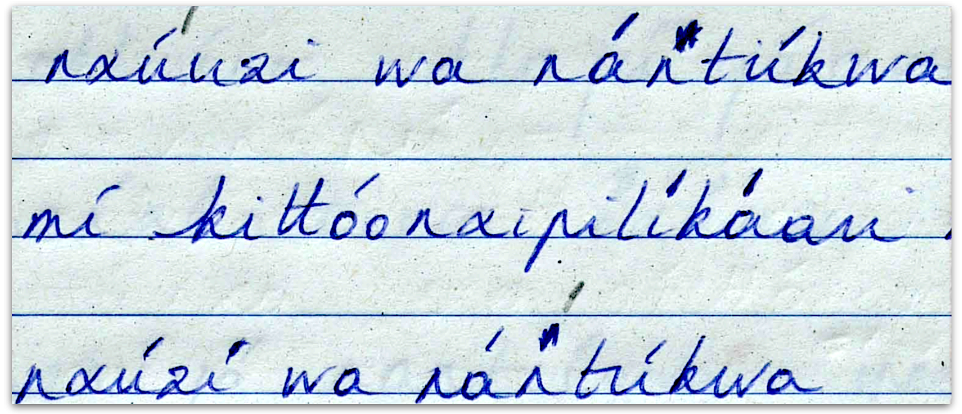
PanLex

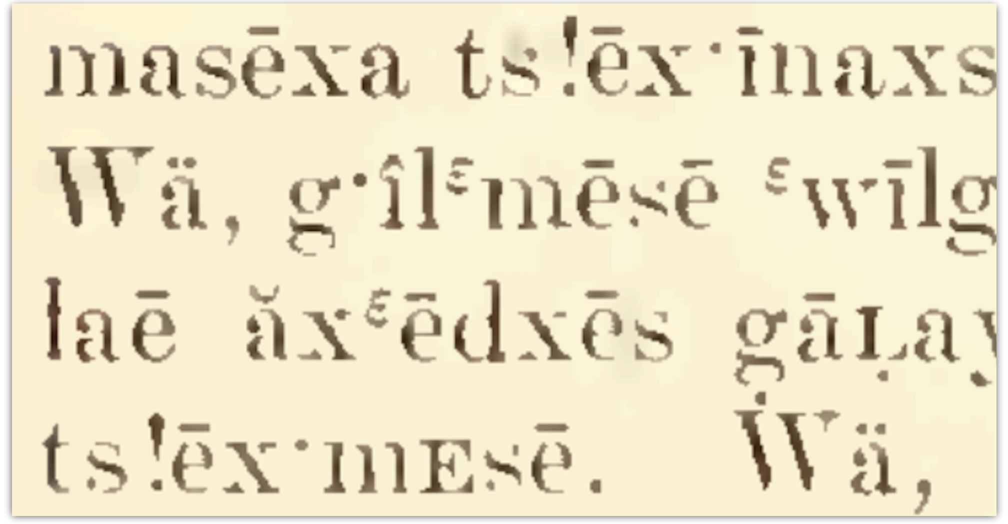
- Methods to **unlock new resources** for human or machine use in under-resourced languages
- What's next?
- Should we **put some linguistics in the models**?
  - Morphologically aware soft constraints for OCR?
  - Morphologically/syntactically aware data synthesis using lexicons?

Conclusion



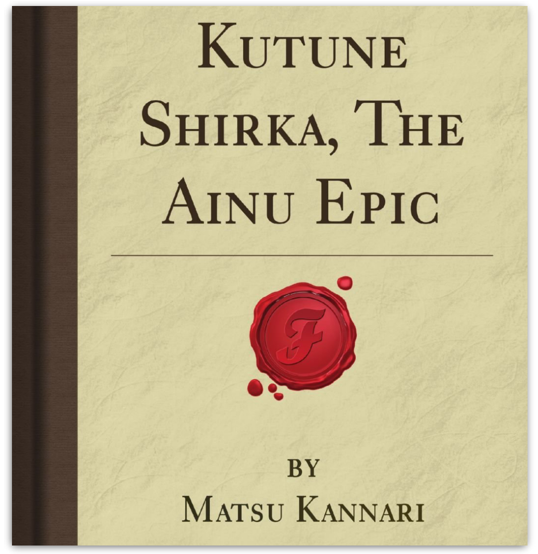


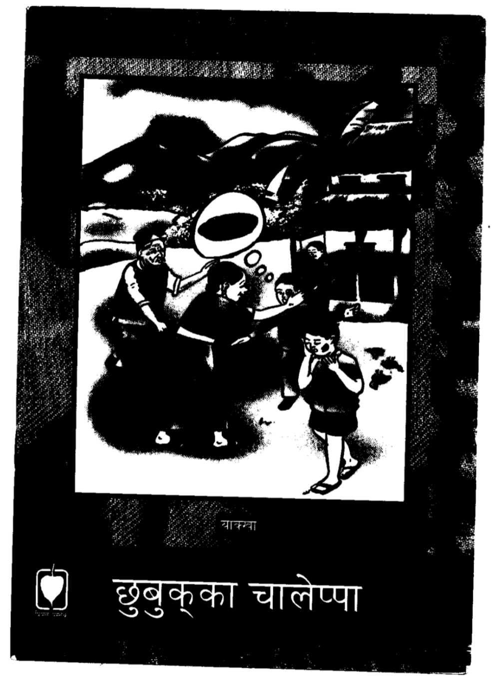


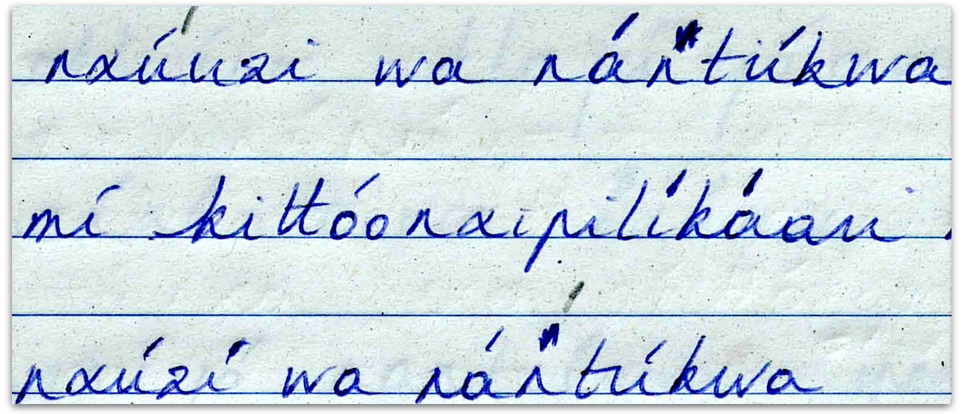


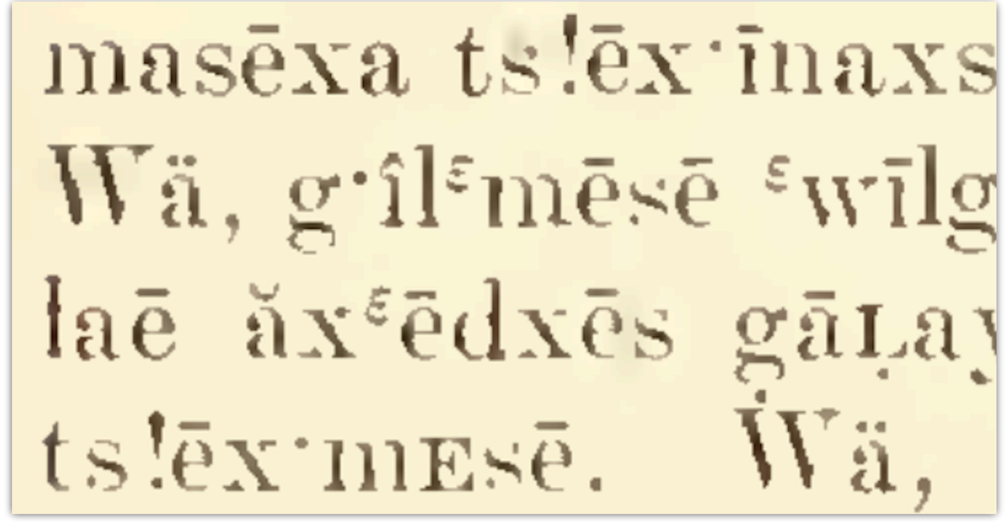
- Methods to **unlock new resources** for human or machine use in under-resourced languages
- What's next?
- Should we **put some linguistics in the models**?
  - Morphologically aware soft constraints for OCR?
  - Morphologically/syntactically aware data synthesis using lexicons?
- Should we **use the models in language learning or linguistics**?

Conclusion









- Methods to **unlock new resources** for human or machine use in under-resourced languages
- What's next?
- Should we **put some linguistics in the models**?
  - Morphologically aware soft constraints for OCR?
  - Morphologically/syntactically aware data synthesis using lexicons?
- Should we **use the models in language learning or linguistics**?
  - Large-scale extraction of text or inter-linear glosses for use in developing language materials?