THE UNIVERSITY *of* EDINBURGH

University of Sheffield

Limitations of Religious Data and the Importance of the Target Domain:

Towards Machine Translation for Guinea-Bissau Creole

Jacqueline Rowe, Edward Gow-Smith, Mark Hepple

SLT

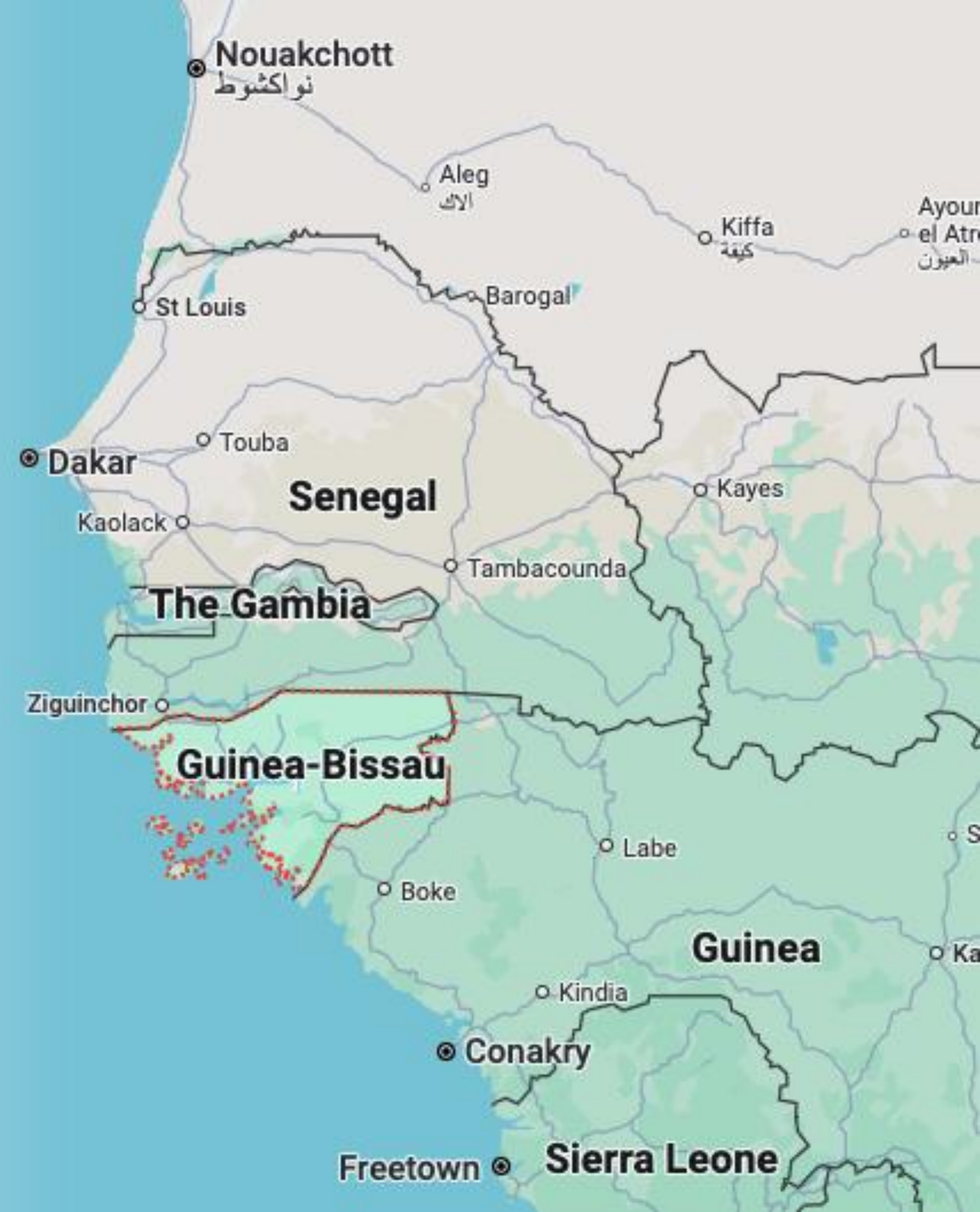
Designing Responsible Natural Language Processing

UK Research and Innovation

Guinea-Bissau Creole (Kiriol)

>1.5 million
L2 speakers

>350k
L1 speakers



Christoph Kohl. 2016. Limitations and Ambiguities of Colonialism in Guinea-Bissau: Examining the Creole and "Civilized" Space in Colonial Society. *History in Africa*, 43:169–203.

Guinea-Bissau Creole (Kiriol)

Kiriol	English
Timótiu mistiba juda jintis	Timothy wanted to help people
Sangi ta kuri na arteria.	Blood flows in the artery.
Un ermon sta ku si fiju, i na piskaba.	A brother was fishing with his son.

Available at https://www.bible.com. Copyright © 1993-2020, Instituto de Tradução e Alfabetização.

Available at https://www.jw.org/en/library/ magazines/. Copyright
© 2025 Watch Tower Bible and Tract Society of Pennsylvania.

Guinea-Bissau Creole (Kiriol)

Kiriol	Portuguese
Timótiu mistiba juda jintis	Timóteo queria ajudar os outros.
Sangi ta kuri na arteria.	O sangue corre na artéria.
Un ermon sta ku si fiju, i na piskaba.	Um irmão estava pescando junto com seu filho.

Available at https://www.bible.com. Copyright © 1993-2020, Instituto de Tradução e Alfabetização.

Available at https://www.jw.org/en/library/ magazines/. Copyright © 2025 Watch Tower Bible and Tract Society of Pennsylvania.

Guinea-Bissau Creole (Kiriol)

Kiriol	Portuguese
Timótiu mistiba juda jintis	Timóteo queria ajudar os outros.
Sangi ta kuri na arteria.	O sangue corre na artéria.
Un ermon sta ku si fiju, i na piskaba.	Um irmão estava pescando junto com seu filho.

Available at https://www.bible.com. Copyright © 1993-2020, Instituto de Tradução e Alfabetização.

Available at https://www.jw.org/en/library/ magazines/. Copyright © 2025 Watch Tower Bible and Tract Society of Pennsylvania.

Guinea-Bissau Creole (Kiriol)

Kiriol	Portuguese
Timótiu misti**ba** juda jintis	Timóteo quer**ia** ajud**ar** os outros.
Sangi **ta** kuri na arteria.	**O** sangue cor**re** na artéria.
Un ermon **sta** ku si fiju, **i na** piska**ba**.	Um irmão **estava** pescando junto com seu filho.

Available at https://www.bible.com. Copyright © 1993-2020, Instituto de Tradução e Alfabetização.

Available at https://www.jw.org/en/library/ magazines/. Copyright © 2025 Watch Tower Bible and Tract Society of Pennsylvania.

Creole NLP

- Work on individual creole languages
- Robinson et al. (2024). Kreyòl-MT: Building MT for Latin American, Caribbean and Colonial African Creole Languages.
- Lent et al. (2024). CreoleVal: Multilingual multitask benchmarks for creoles.

Google Translate, https://translate.google.com/, accessed 6 April 2025.

Creole NLP

- Work on individual creole languages

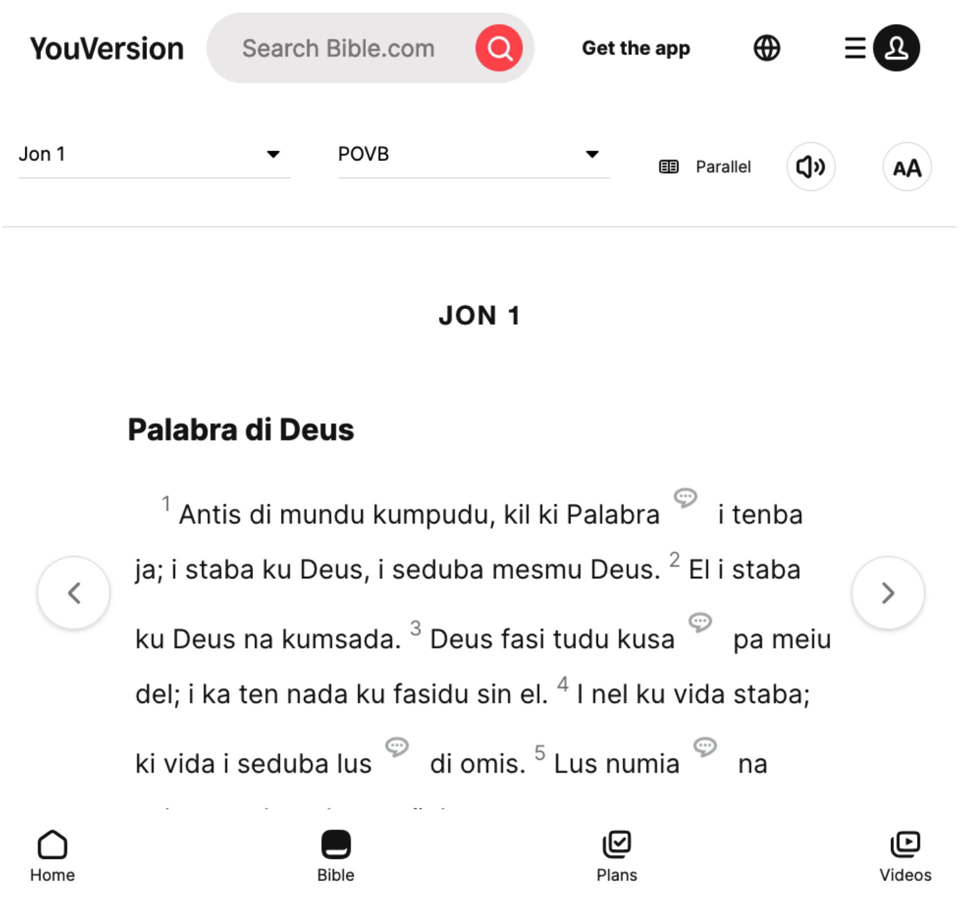
- Robinson et al. (2024). Kreyòl-MT: Building MT for Latin American, Caribbean and Colonial African Creole Languages.

- Lent et al. (2024). CreoleVal: Multilingual multitask benchmarks for creoles.

- MT performance shown to depend on vocabulary overlap (Birch et al 2008) and morphological complexity (*Koehn, 2005; Park et al., 2021; Cotterell et al., 2018; Arnett & Burgen (2024)*

Google Translate, https://translate.google.com/, accessed 6 April 2025.

Machine translation for Guinea-Bissau Creole (Kiriol)

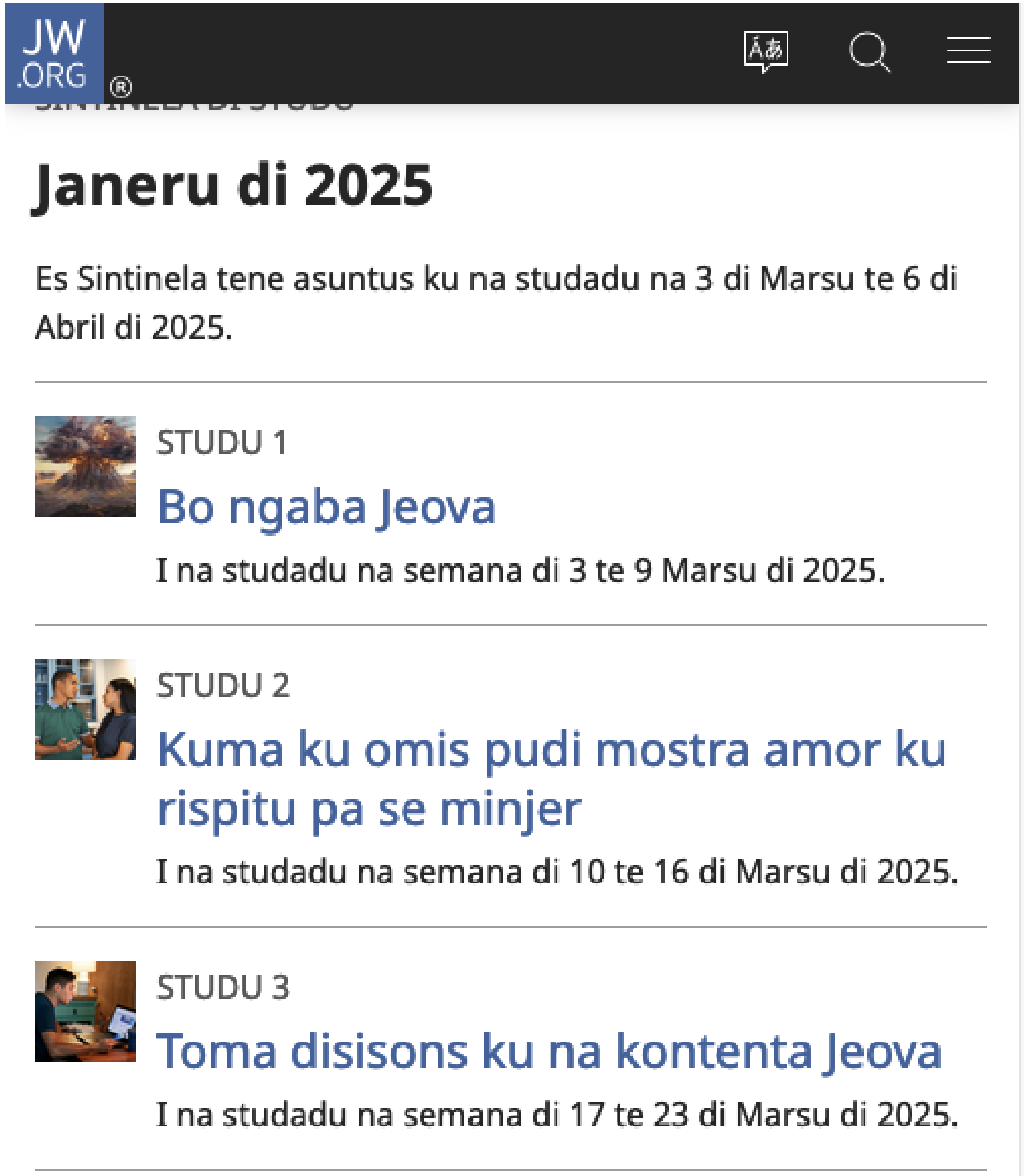


JON 1

Palabra di Deus

[1] Antis di mundu kumpudu, kil ki Palabra i tenba ja; i staba ku Deus, i seduba mesmu Deus. [2] El i staba ku Deus na kumsada. [3] Deus fasi tudu kusa pa meiu del; i ka ten nada ku fasidu sin el. [4] I nel ku vida staba; ki vida i seduba lus di omis. [5] Lus numia na

Available at https://www.bible.com. Copyright © 1993-2020, Instituto de Tradução e Alfabetização.



SINTINELA DI STUDU

Janeru di 2025

Es Sintinela tene asuntus ku na studadu na 3 di Marsu te 6 di Abril di 2025.

STUDU 1

Bo ngaba Jeova

I na studadu na semana di 3 te 9 Marsu di 2025.

STUDU 2

Kuma ku omis pudi mostra amor ku rispitu pa se minjer

I na studadu na semana di 10 te 16 di Marsu di 2025.

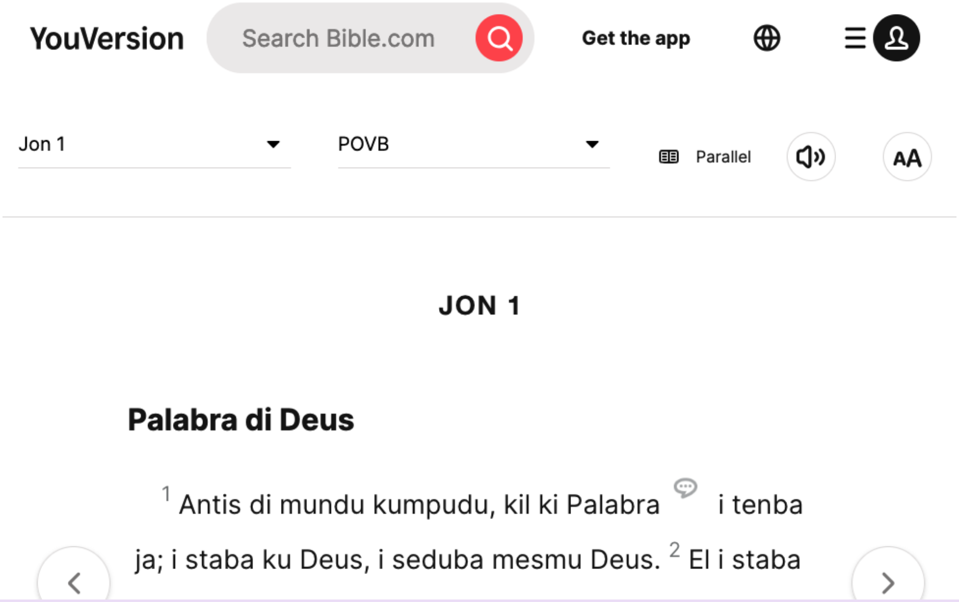
STUDU 3

Toma disisons ku na kontenta Jeova

I na studadu na semana di 17 te 23 di Marsu di 2025.

Available at https://www.jw.org/en/library/ magazines/. Copyright © 2025 Watch Tower Bible and Tract Society of Pennsylvania.

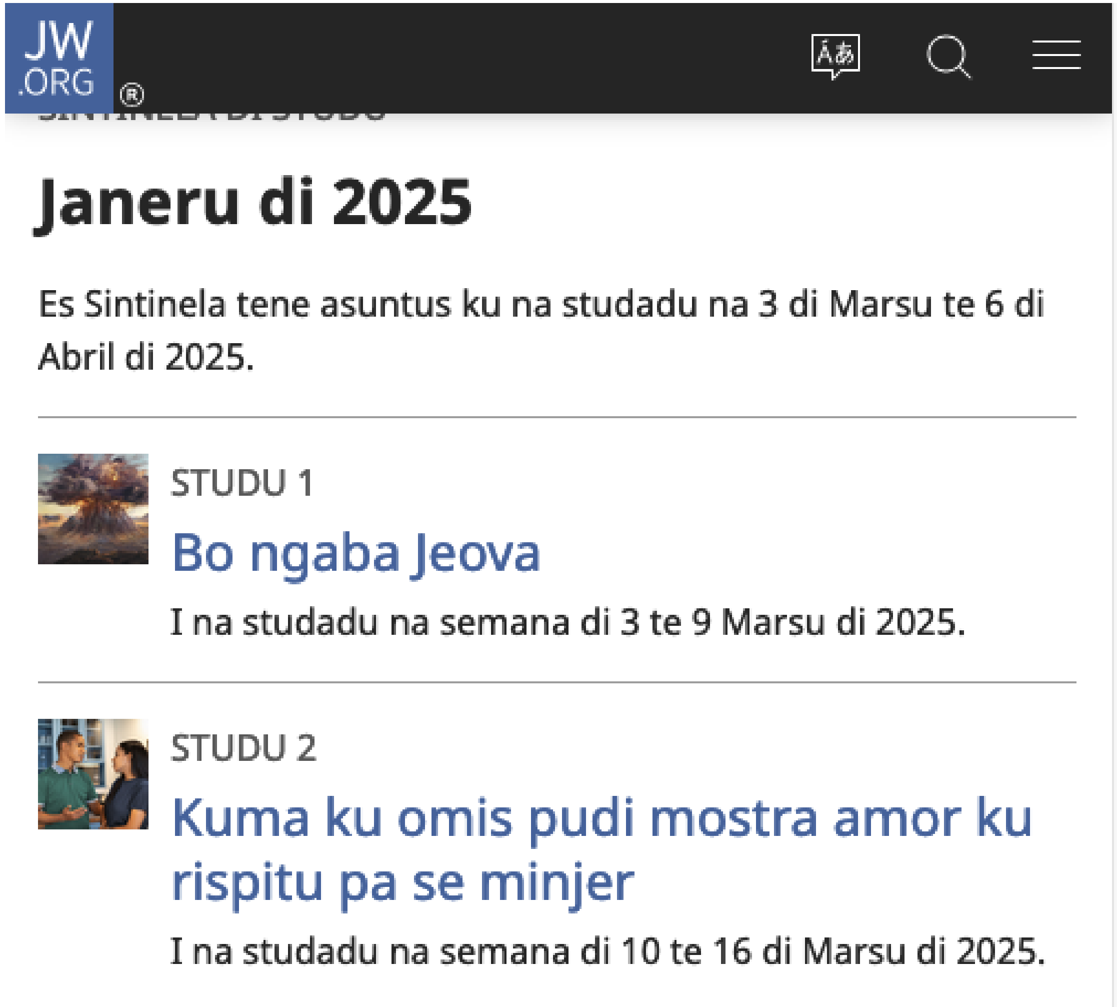
Machine translation for Guinea-Bissau Creole (Kiriol)



JON 1

Palabra di Deus

[1] Antis di mundu kumpudu, kil ki Palabra i tenba ja; i staba ku Deus, i seduba mesmu Deus. [2] El i staba



SINTINELA DI STUDU

Janeru di 2025

Es Sintinela tene asuntus ku na studadu na 3 di Marsu te 6 di Abril di 2025.

STUDU 1

Bo ngaba Jeova

I na studadu na semana di 3 te 9 Marsu di 2025.

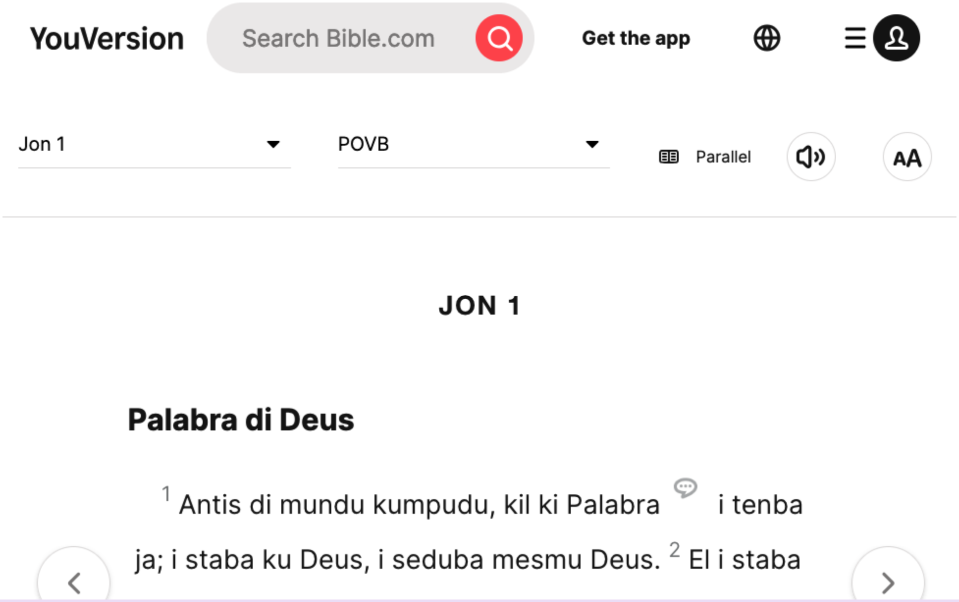
STUDU 2

Kuma ku omis pudi mostra amor ku rispitu pa se minjer

I na studadu na semana di 10 te 16 di Marsu di 2025.

1: How can we best leverage religious data to improve Kiriol MT in other domains?

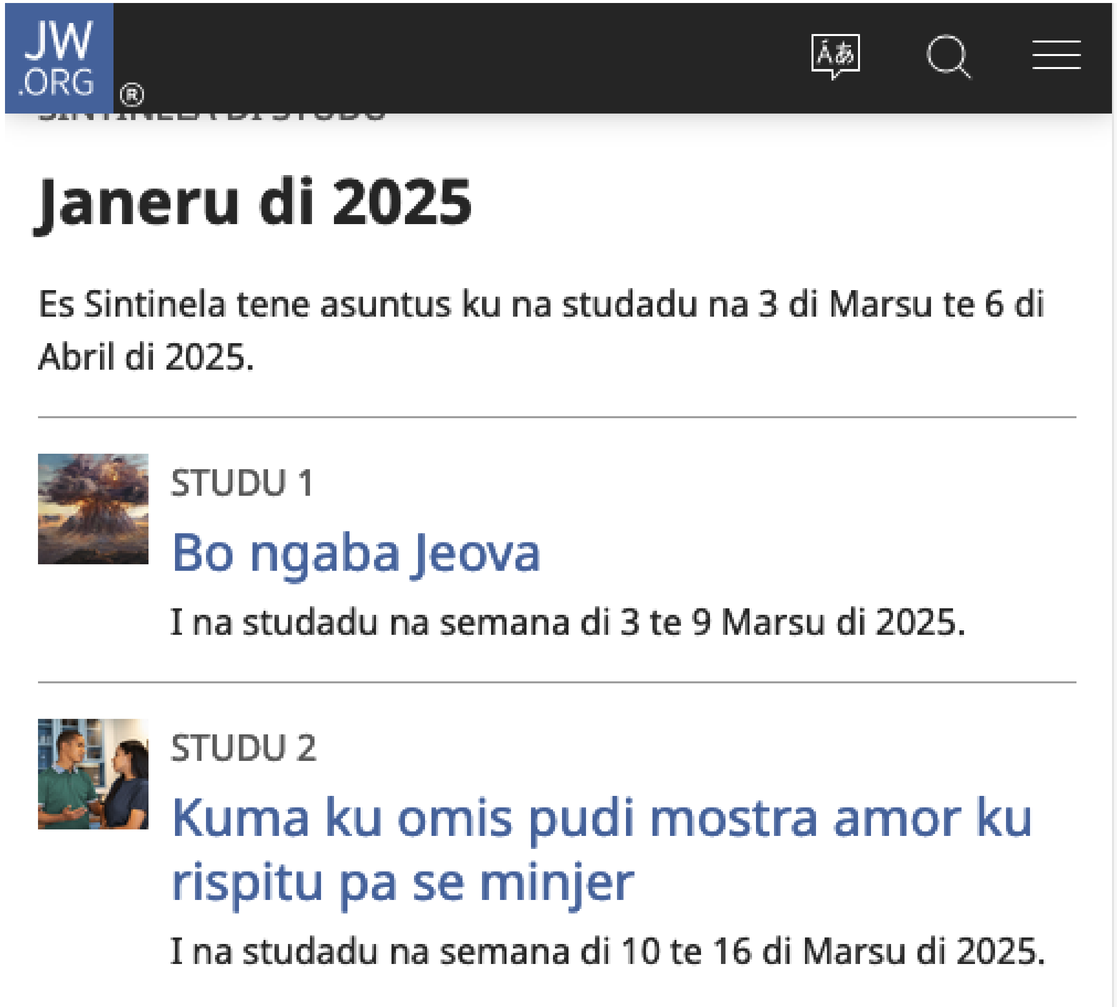
Machine translation for Guinea-Bissau Creole (Kiriol)



JON 1

Palabra di Deus

[1] Antis di mundu kumpudu, kil ki Palabra i tenba ja; i staba ku Deus, i seduba mesmu Deus. [2] El i staba



SINTINELA DI STUDU

Janeru di 2025

Es Sintinela tene asuntus ku na studadu na 3 di Marsu te 6 di Abril di 2025.

STUDU 1

Bo ngaba Jeova

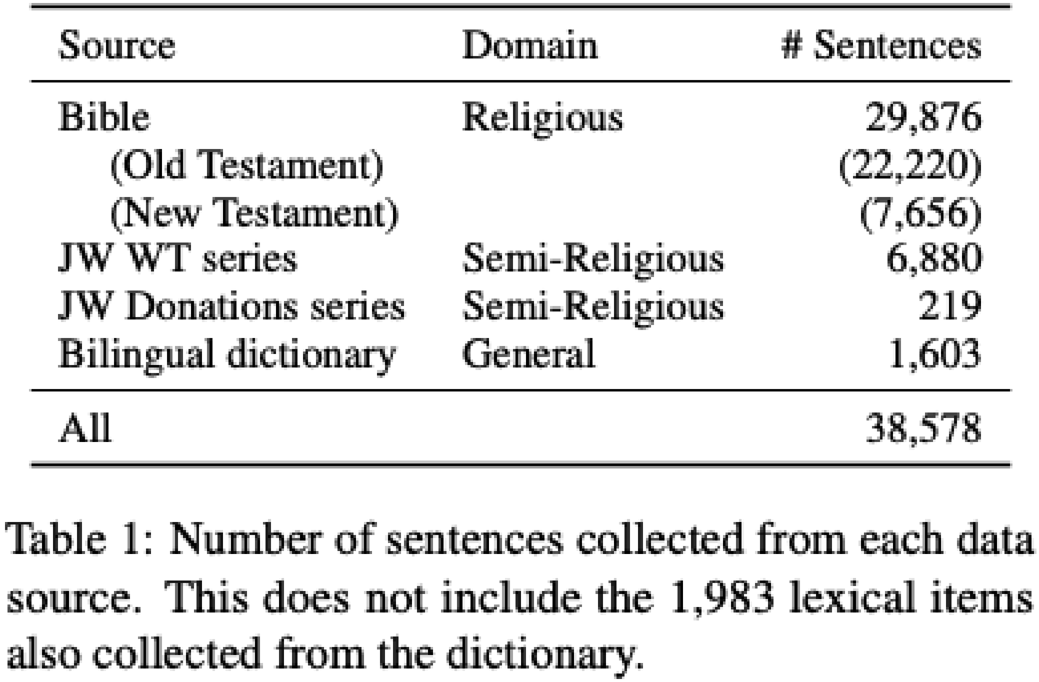
I na studadu na semana di 3 te 9 Marsu di 2025.

STUDU 2

Kuma ku omis pudi mostra amor ku rispitu pa se minjer

I na studadu na semana di 10 te 16 di Marsu di 2025.

2: Does the linguistic relationship between Portuguese and Kiriol impact MT?

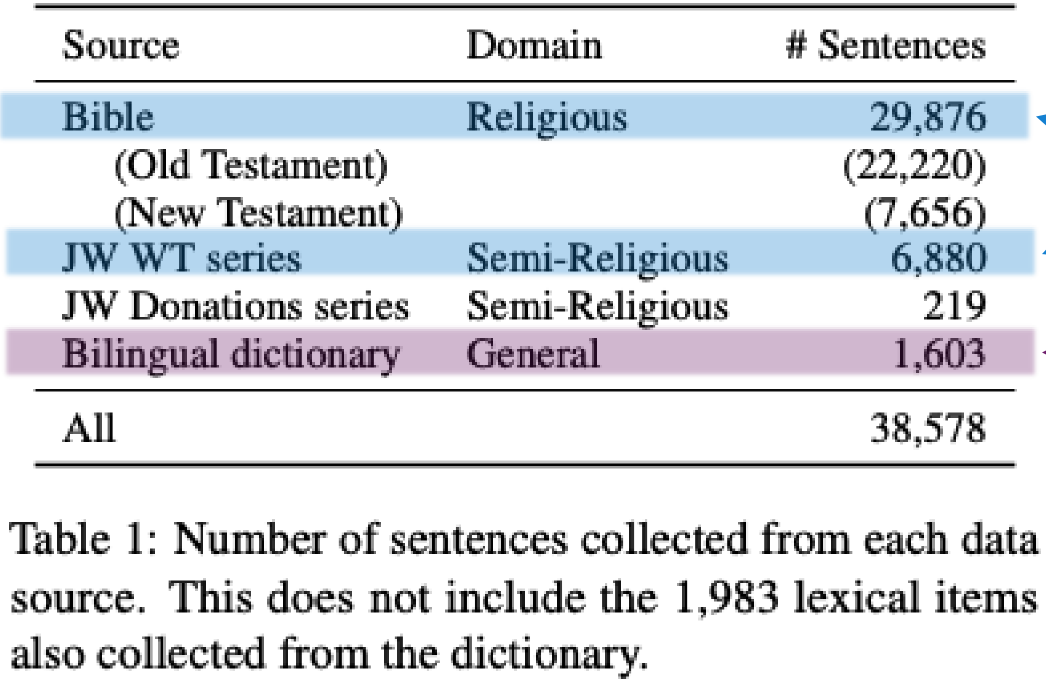


Source	Domain	# Sentences
Bible	Religious	29,876
(Old Testament)		(22,220)
(New Testament)		(7,656)
JW WT series	Semi-Religious	6,880
JW Donations series	Semi-Religious	219
Bilingual dictionary	General	1,603
All		38,578

Table 1: Number of sentences collected from each data source. This does not include the 1,983 lexical items also collected from the dictionary.

WT = Watchtower (a Jehovahs Witnesses monthly publication)

1: How can we best leverage religious data to improve Kiriol MT for the general domain?



Source	Domain	# Sentences
Bible	Religious	29,876
(Old Testament)		(22,220)
(New Testament)		(7,656)
JW WT series	Semi-Religious	6,880
JW Donations series	Semi-Religious	219
Bilingual dictionary	General	1,603
All		38,578

Table 1: Number of sentences collected from each data source. This does not include the 1,983 lexical items also collected from the dictionary.

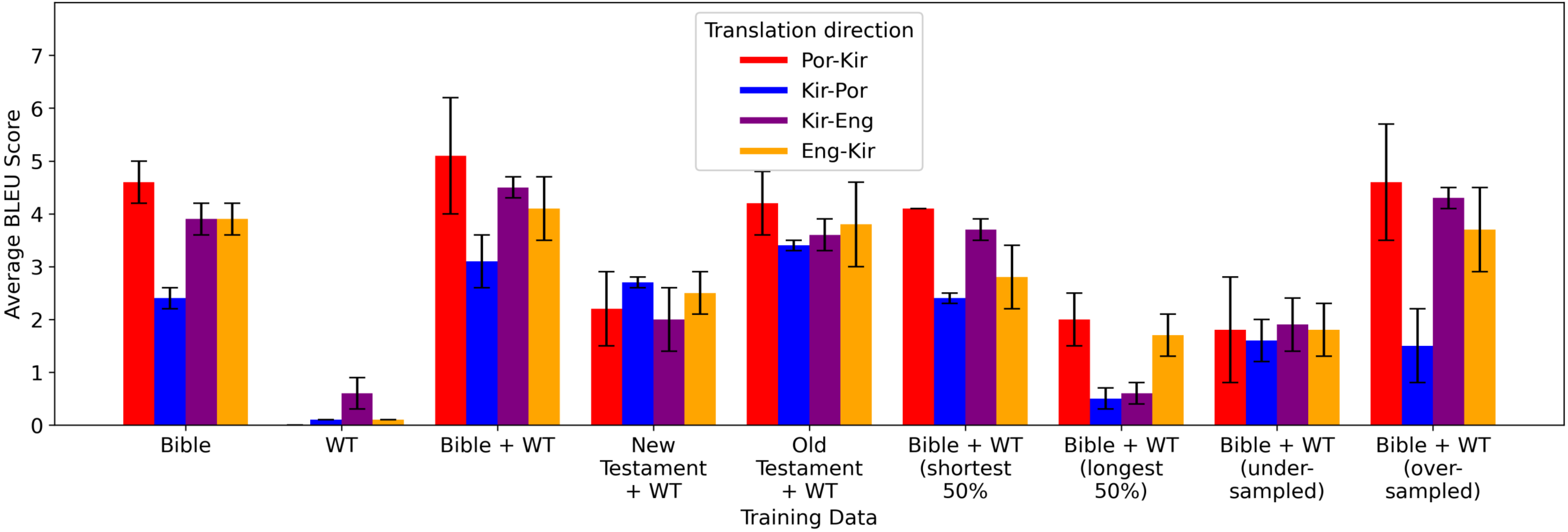
500 Bible + 500 WT sentences as validation set

1,000 dictionary sentences as test set

WT = Watchtower (a Jehovahs Witnesses monthly publication)

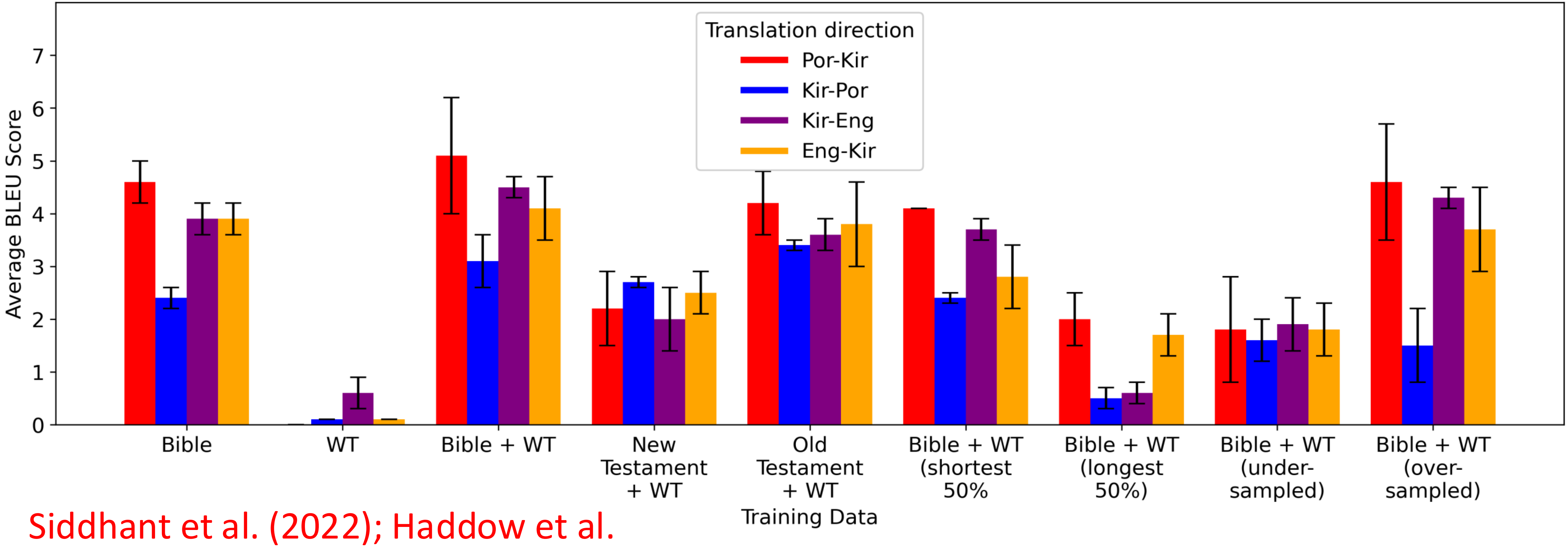
1: How can we best leverage religious data to improve Kiriol MT for the general domain?

Figure 1: Average performance of Portuguese-Kiriol, Kiriol-Portuguese, Kiriol-English and English-Kiriol models trained on different portions of Bible and Watchtower data when used to translate a test set of 1,000 domain-general dictionary sentences. Standard errors across model sets shown with error bars.



1: How can we best leverage religious data to improve Kiriol MT for the general domain?

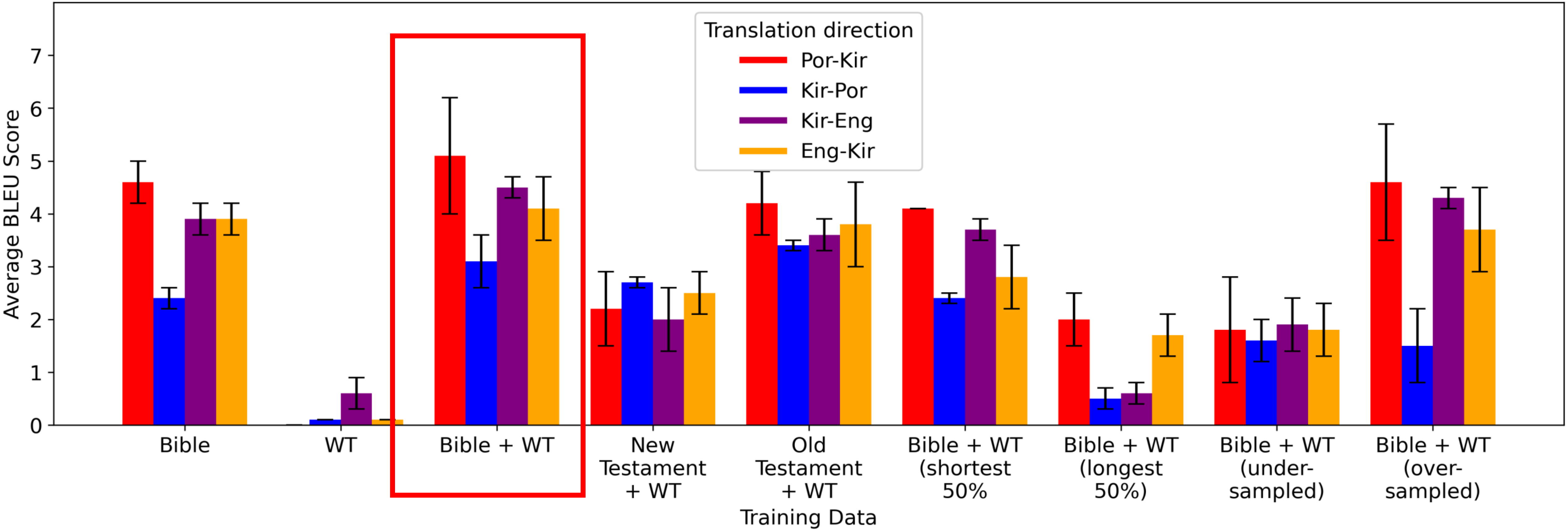
Figure 1: Average performance of Portuguese-Kiriol, Kiriol-Portuguese, Kiriol-English and English-Kiriol models trained on different portions of Bible and Watchtower data when used to translate a test set of 1,000 domain-general dictionary sentences. Standard errors across model sets shown with error bars.



Siddhant et al. (2022); Haddow et al. (2022); Kho et al. (2024)

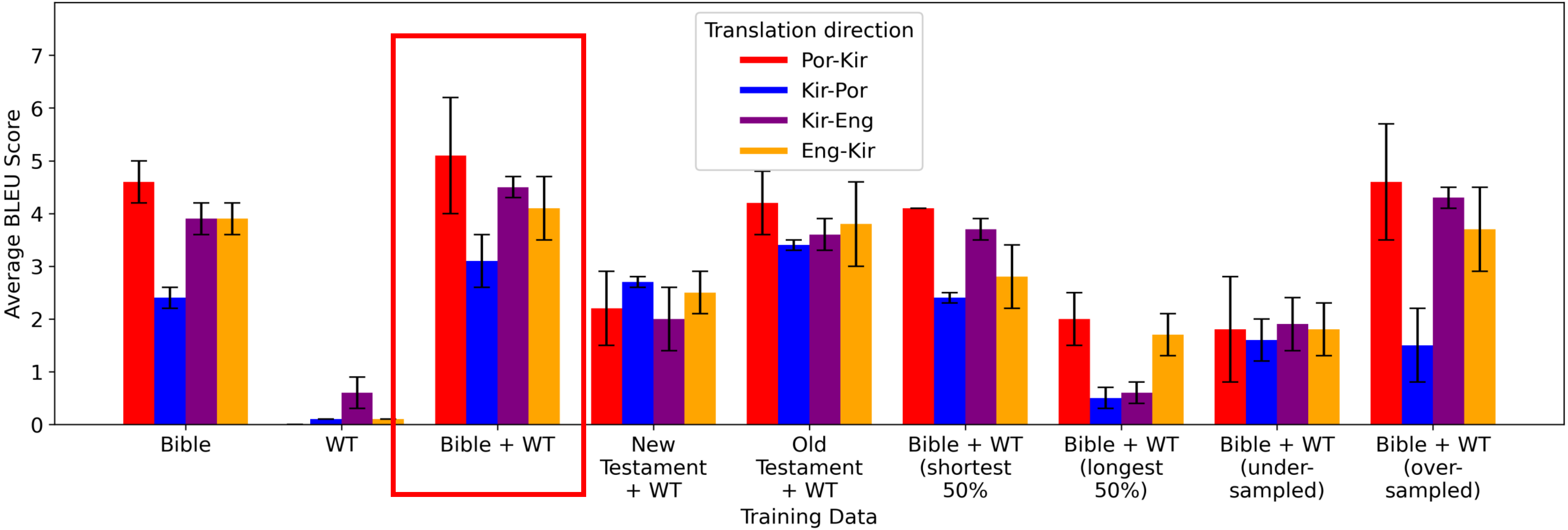
1: How can we best leverage religious data to improve Kiriol MT for the general domain?

Figure 1: Average performance of Portuguese-Kiriol, Kiriol-Portuguese, Kiriol-English and English-Kiriol models trained on different portions of Bible and Watchtower data when used to translate a test set of 1,000 domain-general dictionary sentences. Standard errors across model sets shown with error bars.



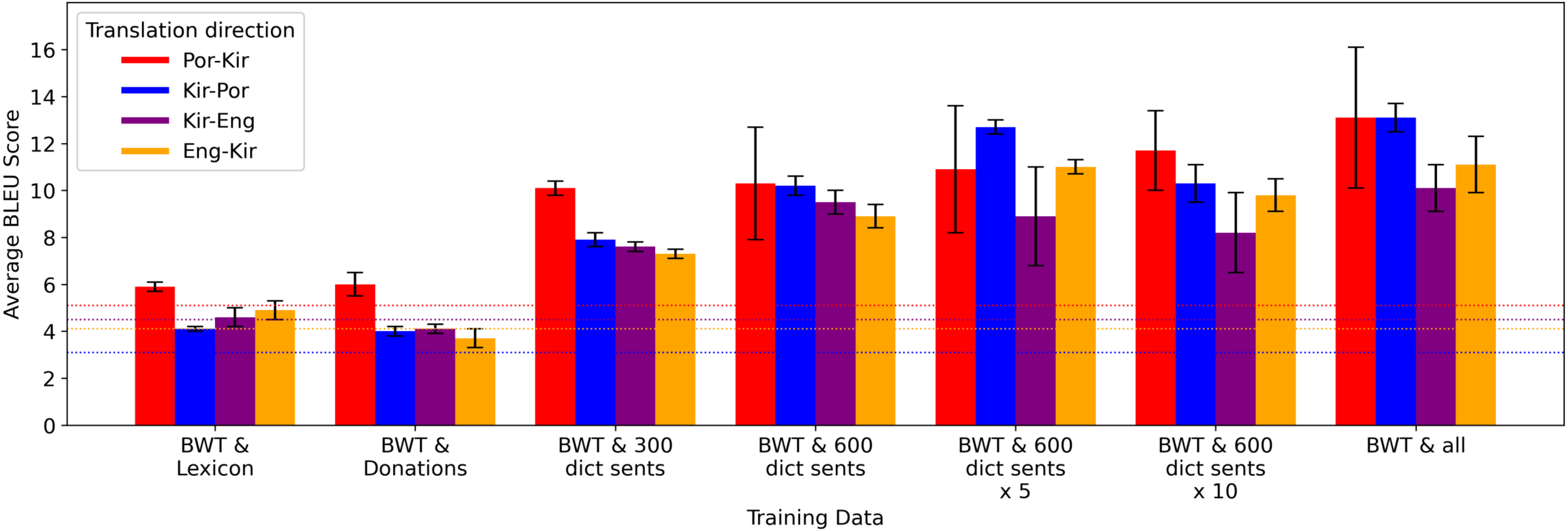
1: How can we best leverage religious data to improve Kiriol MT for the general domain?

Figure 1: Average performance of Portuguese-Kiriol, Kiriol-Portuguese, Kiriol-English and English-Kiriol models trained on different portions of Bible and Watchtower data when used to translate a test set of 1,000 domain-general dictionary sentences. Standard errors across model sets shown with error bars.



1: How can we best augment religious data **(+ small amounts of domain-general data)** to improve Kiriol MT for the general domain?

Figure 2: Average performance of Portuguese-Kiriol, Kiriol-Portuguese, Kiriol-English and English-Kiriol models trained on Bible, Watchtower and different combinations of domain-general data when used to translate test set of 1,000 domain-general dictionary sentences. Standard errors across model sets shown with error bars, and **baseline average performance of models trained only on Bible and WT data is shown with dotted lines.**



1: How can we best augment religious data **(+ small amounts of domain-general data)** to improve Kiriol MT for the general domain?

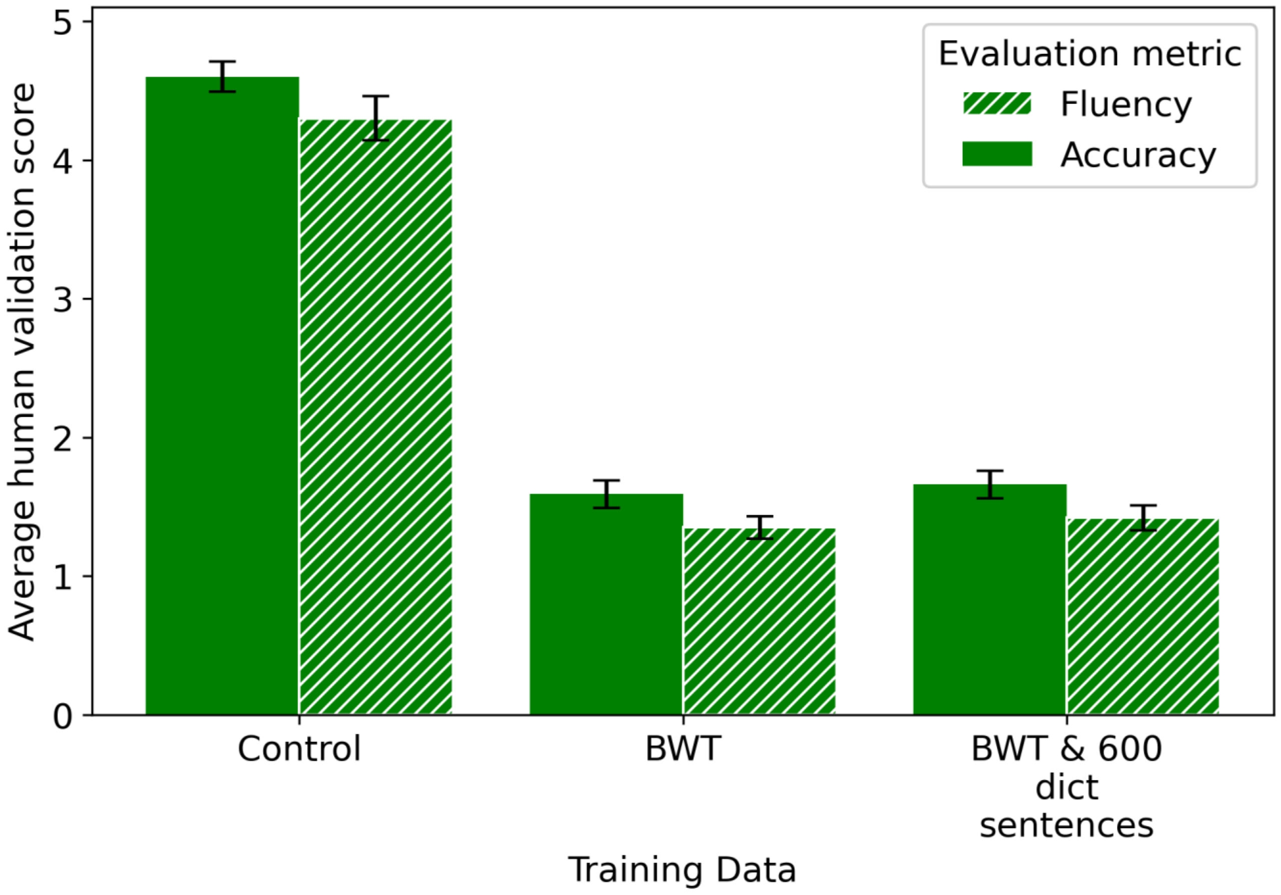


Figure 3: Average scores across all language directions of human judgements for accuracy (solid) and fluency (hatched) of translated sentences from the reference sets (control) and from models trained on Bible and WT data (BWT) and Bible, WT and 600 dictionary sentences. Standard errors across all judgements for each condition are shown with error bars.

1: How can we best augment religious data **(+ small amounts of domain-general data)** to improve Kiriol MT for the general domain?

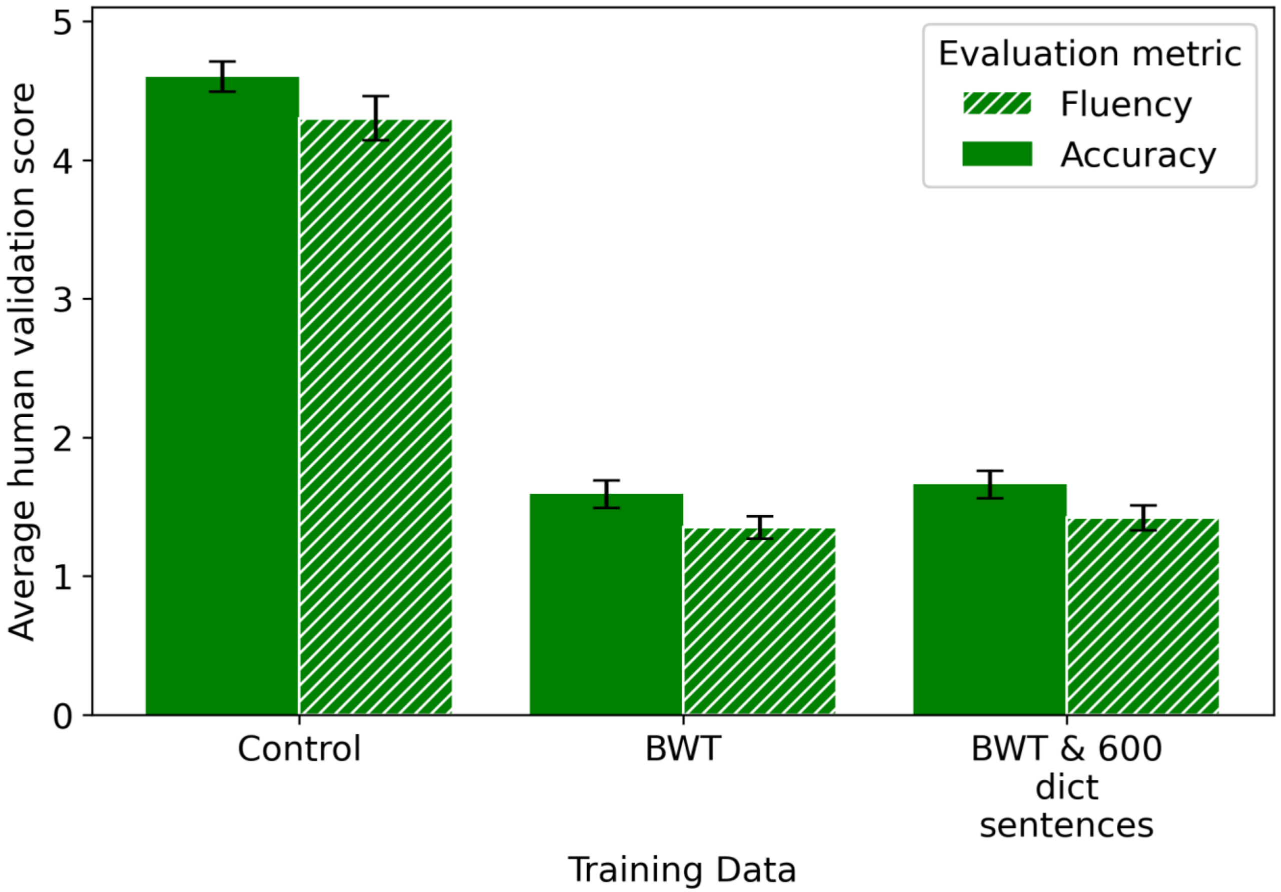


Figure 3: Average scores across all language directions of human judgements for accuracy (solid) and fluency (hatched) of translated sentences from the reference sets (control) and from models trained on Bible and WT data (BWT) and Bible, WT and 600 dictionary sentences. Standard errors across all judgements for each condition are shown with error bars.

- the increases in BLEU scores were still too low overall to be perceptible to the human eye
- the overall utility of the models trained on mostly religious data remains limited

1: How can we best augment religious data **(+ small amounts of domain-general data)** to improve Kiriol MT for the general domain?

Bianda sufisienti pa tudu djintis. →

There's enough food for everyone.

Give me the word to all the people.

N'baiba bisita ña primu →

I was going to visit my cousin.

I went to the Mishonite.

Para mima bu ermon. →

Stop spoiling your brother.

Paradise, be your brother.

- the increases in BLEU scores were still too low overall to be perceptible to the human eye
- the overall utility of the models trained on mostly religious data remains limited

1: How can we best augment religious data **(+ small amounts of domain-general data)** to improve Kiriol MT for the general domain?

Bianda sufisienti pa tudu djintis. → There's enough food for everyone.

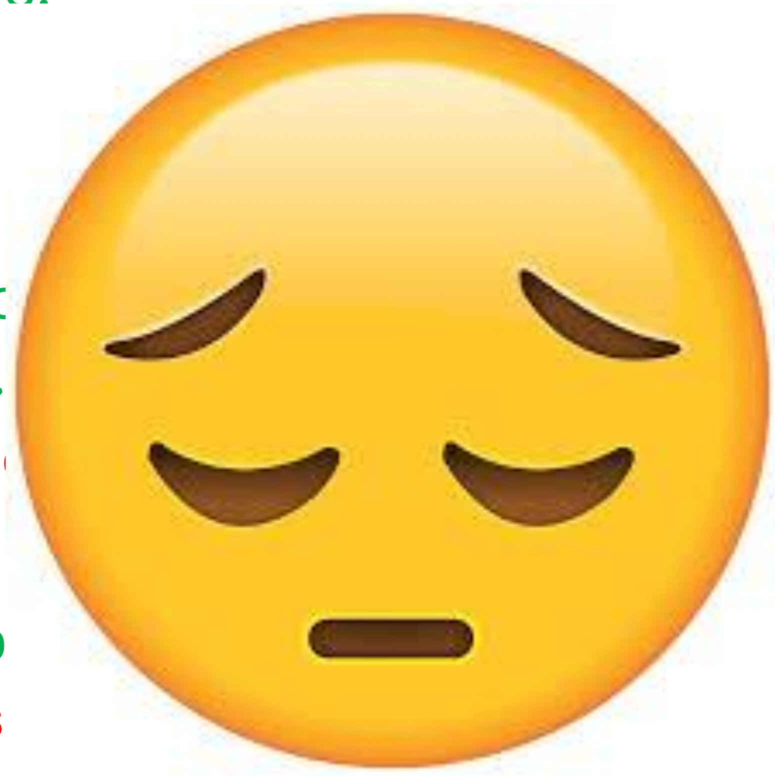
Give me people.

N'baiba bisita ña primu → I was gc cousin.

I went t

Para mima bu ermon. → Stop sp

Paradis



creases in BLEU scores were still w overall to be perceptible to the ın eye

verall utility of the models trained ɔstly religious data remains limited

1: How can we best augment religious data **(+ small amounts of domain-general data)** to improve Kiriol MT for the general domain?

2: Does the linguistic relationship between Portuguese and Kiriol impact MT?

Yes!

1. **Kiriol-Portuguese tokenisers have more overlapping vocabulary items than Kiriol-English tokenisers**

2: Does the linguistic relationship between Portuguese and Kiriol impact MT?

Yes!

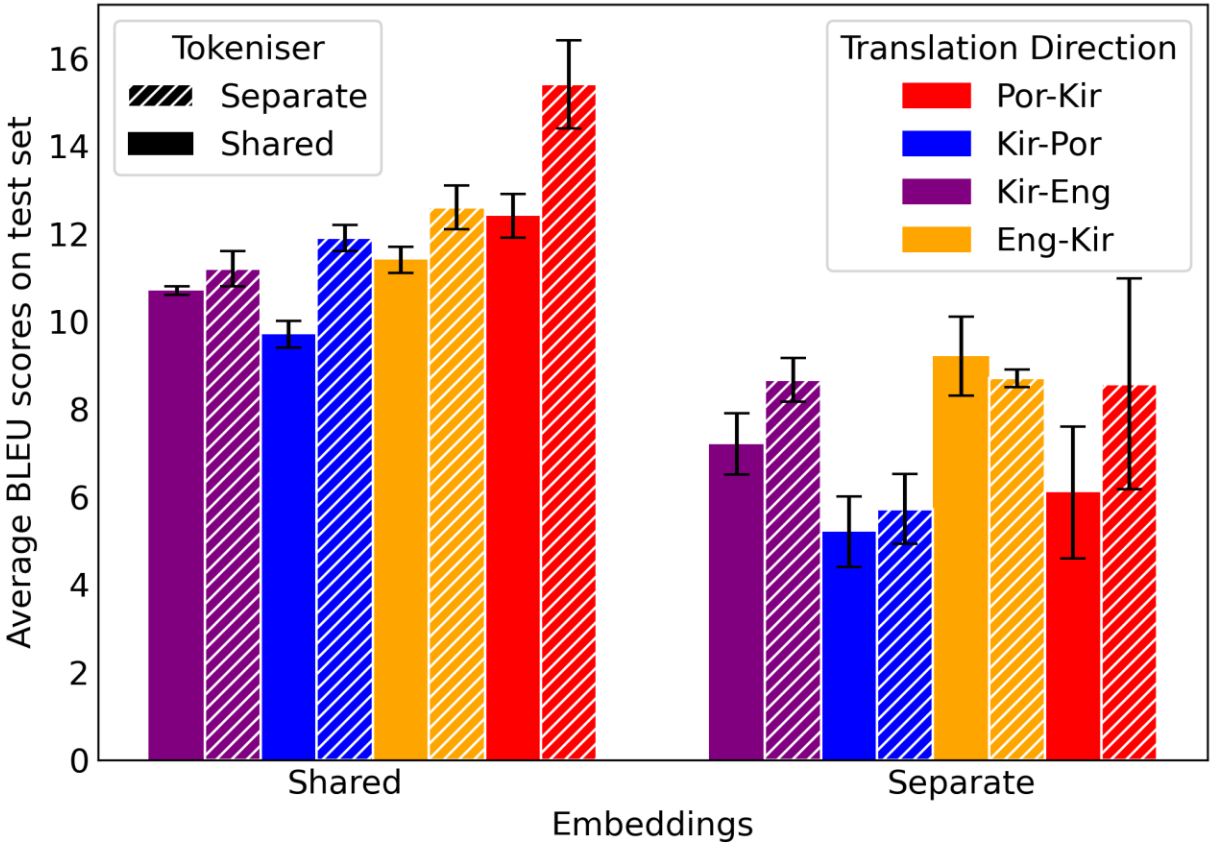
1. Kiriol-Portuguese tokenisers have more overlapping vocabulary items than Kiriol-English tokenisers
2. **Compared to using shared tokenisers, using separate tokenisers worsens fertility on Portuguese texts more than on English texts**

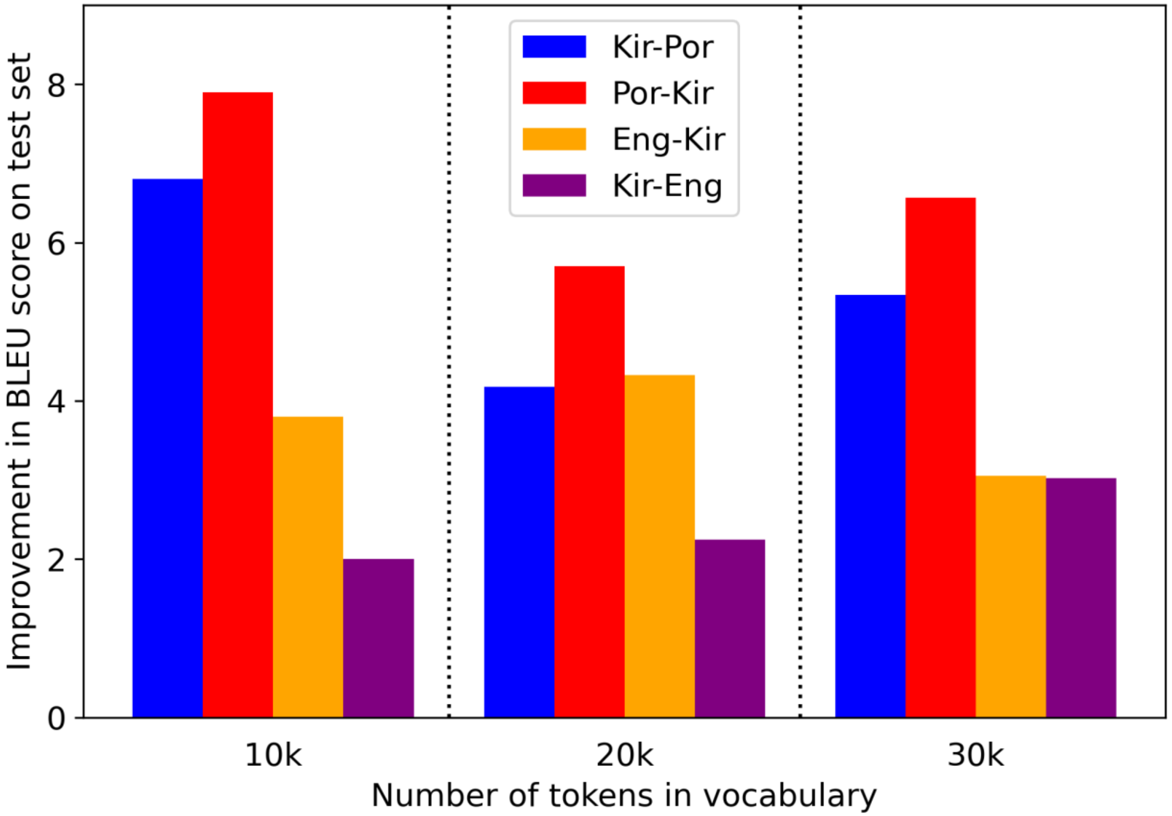
2: Does the linguistic relationship between Portuguese and Kiriol impact MT?

Yes!

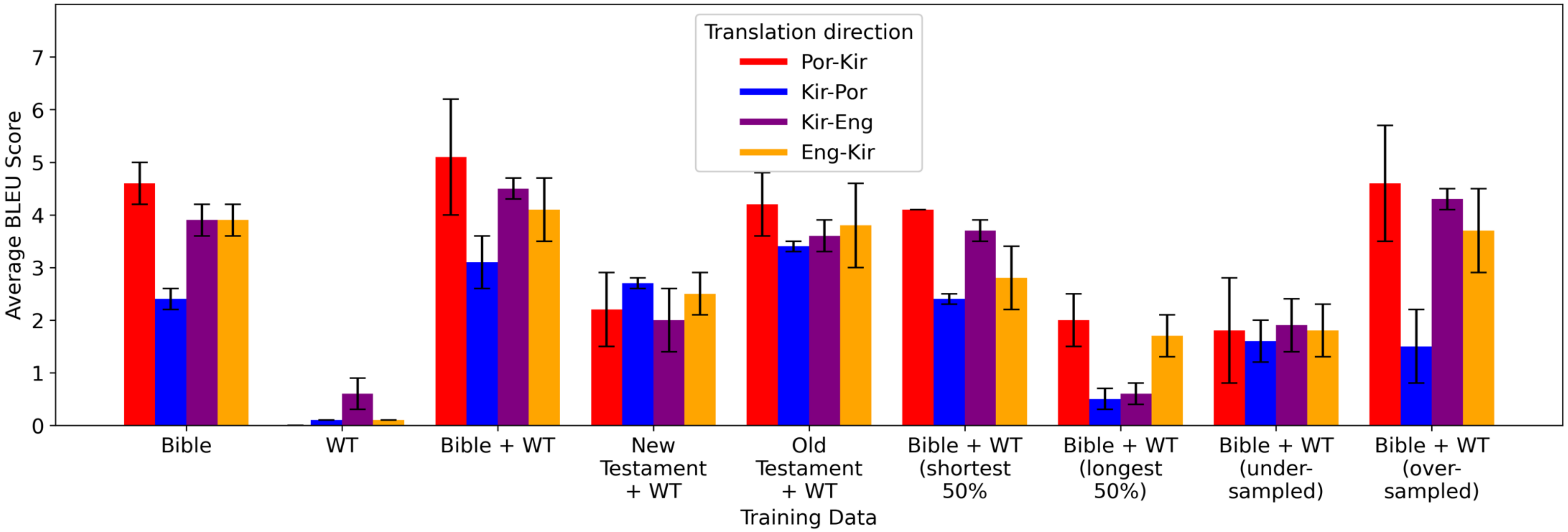
1. Kiriol-Portuguese tokenisers have more overlapping vocabulary items than Kiriol-English tokenisers
2. Compared to using shared tokenisers, using separate tokenisers worsens fertility on Portuguese texts more than on English texts
3. **Compared to using shared embeddings, using separate embeddings reduces performance for Kiriol-Portuguese and Portuguese-Kiriol more than Kiriol-English and English-Kiriol**

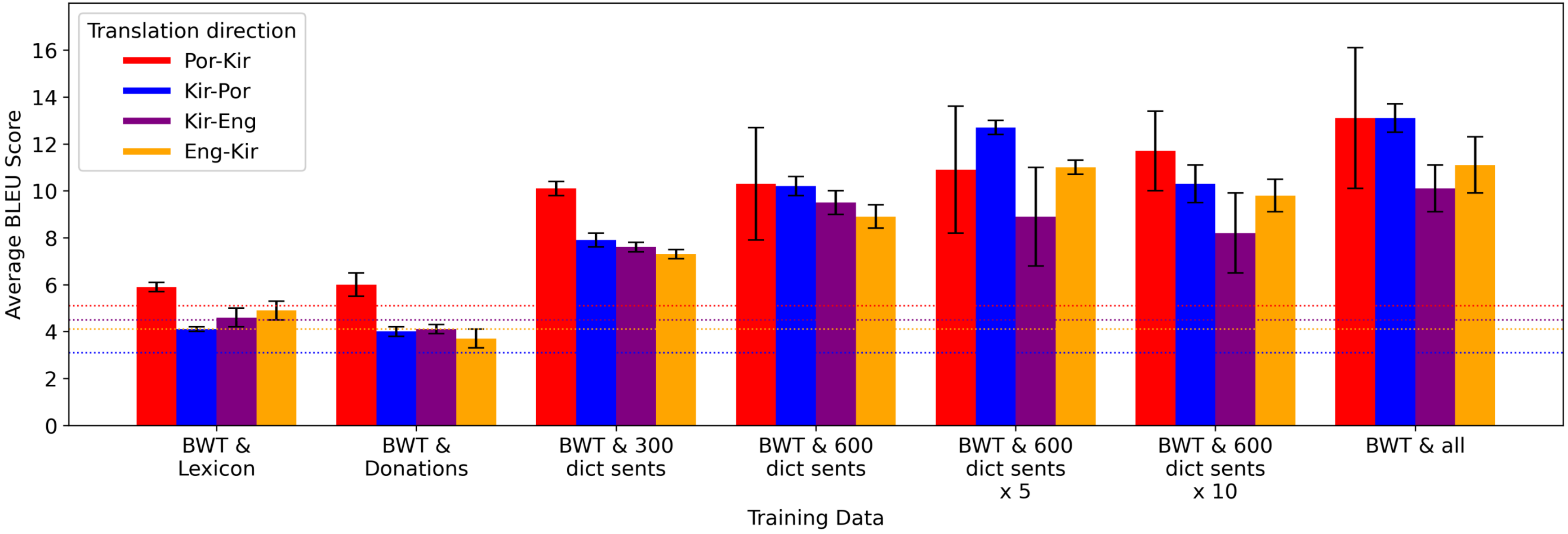
2: Does the linguistic relationship between Portuguese and Kiriol impact MT?





2: Does the linguistic relationship between Portuguese and Kiriol impact MT?


Translation direction
Por-Kir
Kir-Por
Kir-Eng
Eng-Kir
Average BLEU Score
0
1
2
3
4
5
6
7
Bible
WT
Bible + WT
New Testament + WT
Old Testament + WT
Bible + WT (shortest 50%
Bible + WT (longest 50%)
Bible + WT (under-sampled)
Bible + WT (over-sampled)
Training Data


Translation direction
Por-Kir
Kir-Por
Kir-Eng
Eng-Kir
Average BLEU Score
0
2
4
6
8
10
12
14
16
BWT & Lexicon
BWT & Donations
BWT & 300 dict sents
BWT & 600 dict sents
BWT & 600 dict sents x 5
BWT & 600 dict sents x 10
BWT & all
Training Data

To sum up:

Adding small amounts of target domain data to religious training datasets considerably increase BLEU scores on the target domain…

To sum up:

Adding small amounts of target domain data to religious training datasets considerably increase BLEU scores on the target domain…

…but these are not reflected in human judgements as BLEU scores remain low overall.

To sum up:

Adding small amounts of target domain data to religious training datasets considerably increase BLEU scores on the target domain…

…but these are not reflected in human judgements as BLEU scores remain low overall.

Shared vocabulary between Kiriol and Portuguese does improve tokenization efficiency and model training…

To sum up:

Adding small amounts of target domain data to religious training datasets considerably increase BLEU scores on the target domain…

…but these are not reflected in human judgements as BLEU scores remain low overall.

Shared vocabulary between Kiriol and Portuguese does improve tokenization efficiency and model training…

…but these effects are somewhat complicated by the differences in morphological complexity of the languages involved.

Designing Responsible Natural Language Processing
THE UNIVERSITY of EDINBURGH
UK Research and Innovation
University of Sheffield
SLT

What do you think would be the next best steps for this work?

What other methods should we look at for investigating impact of lexical overlap on creole-lexifier MT?

What role for participation by creole communities?

Thank you!

Documentary

Paper

Me