The University Of Sheffield.

Sample Efficient Approaches for Idiomaticity Detection

Dylan Phelps, **Xuan-Rui Fan**, **Edward Gow-Smith**, **Harish Tayyar Madabushi**, **Carolina Scarton**, and **Aline Villavicencio**

Motivation

Out of the box language models struggle to detect idiomaticity without fine-tuning on large amounts of labelled data

Motivation

Out of the box language models struggle to detect idiomaticity without fine-tuning on large amounts of labelled data

However, individual idioms are uncommon in general text, meaning that training data is hard to come by

Motivation

Out of the box language models struggle to detect idiomaticity without fine-tuning on large amounts of labelled data

However, individual idioms are uncommon in general text, meaning that training data is hard to come by

Therefore it important to find and develop methods for idiomaticity detection with relatively small amounts of data

Contributions

We explore two low resource methods to address this challenge:

- **BERT for Attentive Mimicking (BERTRAM)** [1] - to create single-token embeddings for idioms within existing pre-trained models

[1] Schick, Timo, and Hinrich Schütze. 'BERTRAM: Improved Word Embeddings Have Big Impact on Contextualized Model Performance'. arXiv, 29 April 2020. https://doi.org/10.48550/arXiv.1910.07181.

Contributions

We explore two low resource methods to address this challenge:

- **BERT for Attentive Mimicking (BERTRAM)** [1] - to create single-token embeddings for idioms within existing pre-trained models
- **Pattern Exploit Training (PET)** [2] - to introduce task knowledge into the training and test examples, reducing the amount of labelled data needed

[1] Schick, Timo, and Hinrich Schütze. 'BERTRAM: Improved Word Embeddings Have Big Impact on Contextualized Model Performance'. arXiv, 29 April 2020. https://doi.org/10.48550/arXiv.1910.07181.

[2] Schick, Timo, and Hinrich Schütze. 'Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference'. arXiv, 25 January 2021. https://doi.org/10.48550/arXiv.2001.07676.

SemEval-2022 Task 2

We evaluate our models on Subtask A of SemEval-2022 Task 2

The task encourages the creation of models that can better detect (subtask A) and represent (subtask B) idioms

The data spans 3 languages: English, Portuguese and Galician, with labeled data being available for English and Galician

Sentence	Idiomatic
When removing a big fish from a net, it should be held in a manner that supports the girth.	No
It was still a respectable finish for both Fadol and Nayre, who were ranked outside the top 500 in the world, but caught some big fish along the way.	Yes
To pay attention only to new housing and houses I think skewers the big picture.	Yes

SemEval-2022 Task 2

The MWEs featured in the task are all either noun-noun or noun-adjective compounds

For this paper we focus on the zero-shot split, where the MWEs in the test are disjoint from those in the training set

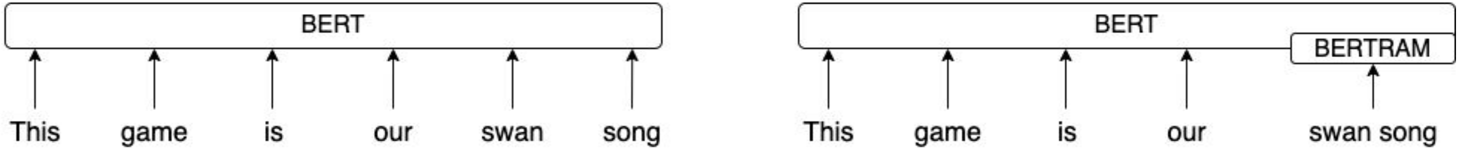
Sentence	Idiomatic
When removing a big fish from a net, it should be held in a manner that supports the girth.	No
It was still a respectable finish for both Fadol and Nayre, who were ranked outside the top 500 in the world, but caught some big fish along the way.	Yes
To pay attention only to new housing and houses I think skewers the big picture.	Yes

Methodology:

BERTRAM Idiom Representations

Background: Word Piece

Within the tokenizers for most LLMs, idioms are broken up into word or sub-word tokens



This forces the model to rely on compositionality (something not seen in all MWEs), and dilutes the idiomatic sense within the individual word embeddings

Our BERTRAM Embeddings

BERTRAM allows us to create embeddings for new tokens within an existing embedding space, from a small number of contexts

We use the embeddings created by Phelps, 2022 [1], which are created using 150 contexts for each idiom, taken from the CC-100 corpus

These embeddings were shown to perform well in SemEval-2022 Task 2 Subtask B, which was an STS featuring idioms, coming 1st and 2nd in the two leaderboards

[1] Phelps, Dylan. 'Drsphelps at SemEval-2022 Task 2: Learning Idiom Representations Using BERTRAM'. arXiv, 25 May 2022. https://doi.org/10.48550/arXiv.2204.02821.

BERTRAM for idiom detection

We transform the examples to use the BERTRAM embeddings

‘“I was in a ***<BERTRAM:IDpandacarID>*** *and we chased two of the robbers through a few streets.’*

Monolingual BERT models are used as our classifiers with a softmax layer to added to predict idiomaticity from the [CLS] token

Methodology:

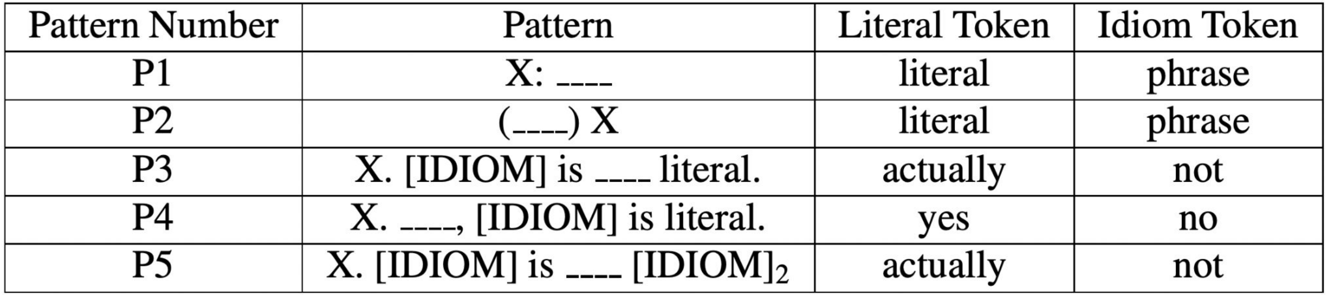
PET Idiom Detection

Pattern Exploit Training (PET)

We can reduce the amount of training examples, by giving more information about the task in the training/test examples

Pattern Exploit Training (PET)

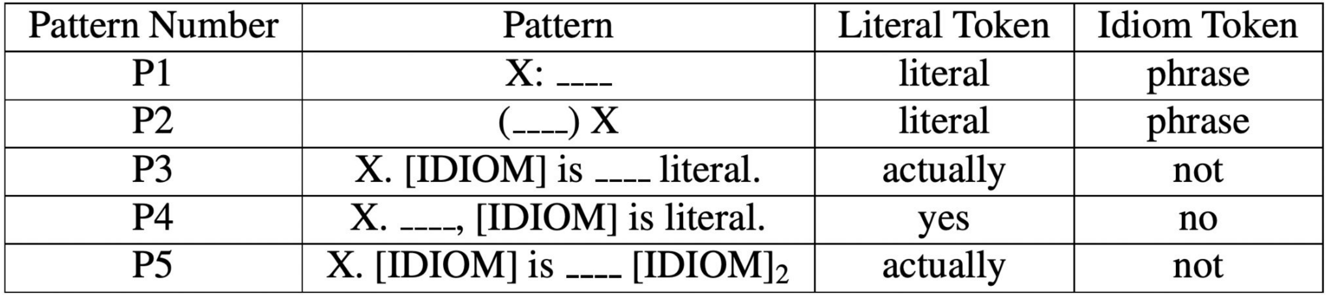
Examples can be transformed by using Pattern Verbaliser Pairs (PVPs)



Pattern Number	Pattern	Literal Token	Idiom Token
P1	X: ____	literal	phrase
P2	(____) X	literal	phrase
P3	X. [IDIOM] is ____ literal.	actually	not
P4	X. ____, [IDIOM] is literal.	yes	no
P5	X. [IDIOM] is ____ [IDIOM]$_2$	actually	not

Pattern Exploit Training (PET)

Examples can be transformed by using Pattern Verbaliser Pairs (PVPs)

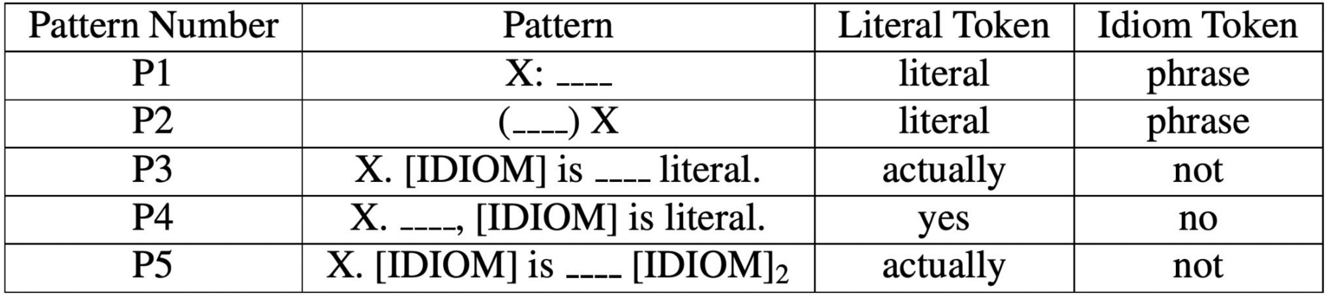


Pattern Number	Pattern	Literal Token	Idiom Token
P1	X: ____	literal	phrase
P2	(____) X	literal	phrase
P3	X. [IDIOM] is ____ literal.	actually	not
P4	X. ____, [IDIOM] is literal.	yes	no
P5	X. [IDIOM] is ____ [IDIOM]$_2$	actually	not

Transform each example using a pattern

Pattern Exploit Training (PET)

Examples can be transformed by using Pattern Verbaliser Pairs (PVPs)



Pattern Number	Pattern	Literal Token	Idiom Token
P1	X: ____	literal	phrase
P2	(____) X	literal	phrase
P3	X. [IDIOM] is ____ literal.	actually	not
P4	X. ____, [IDIOM] is literal.	yes	no
P5	X. [IDIOM] is ____ [IDIOM]$_2$	actually	not

Transform each example using a pattern

Mask output logits are used to predict class

Pattern Exploit Training (PET)

Example with P4:

I was in a **panda car** and we chased two of the robbers through a few streets. Panda car is [MASK] literal.

actually → *non idiomatic*

not → *idiomatic*

Pattern Exploit Training (PET)

Knowledge can be distilled from multiple PVPs, by training each PVP on a small labelled set

Then we combine the predictions on a larger unlabelled set, which can be used to train another classifier

This allows us to use all PVPs and not choose just the best one

PET Experiment Settings

From the 4000 labelled examples available for the task, we experiment with PET on labelled dataset sizes of 10, 100, 1000

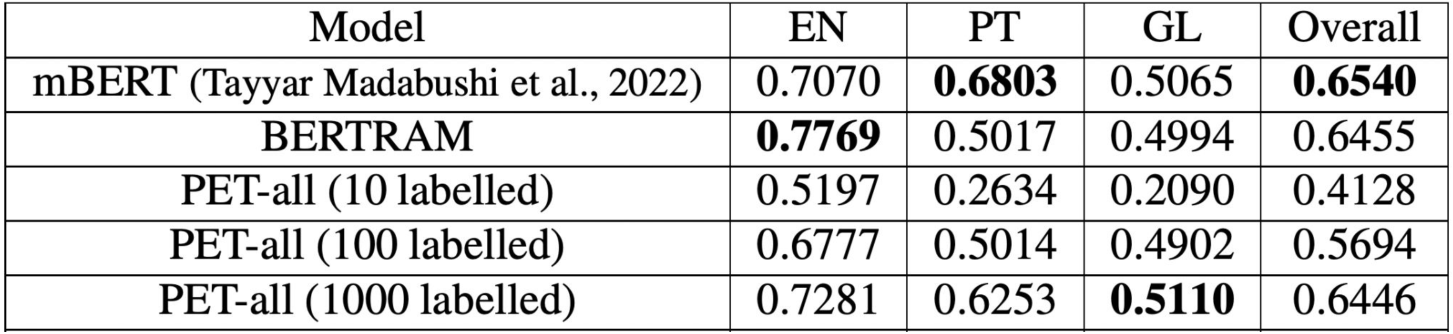
In each of these settings we use the distilled PET variant with all 5 patterns, and label a set of 3000 unlabeled examples

Results:

Idiom Detection

Results

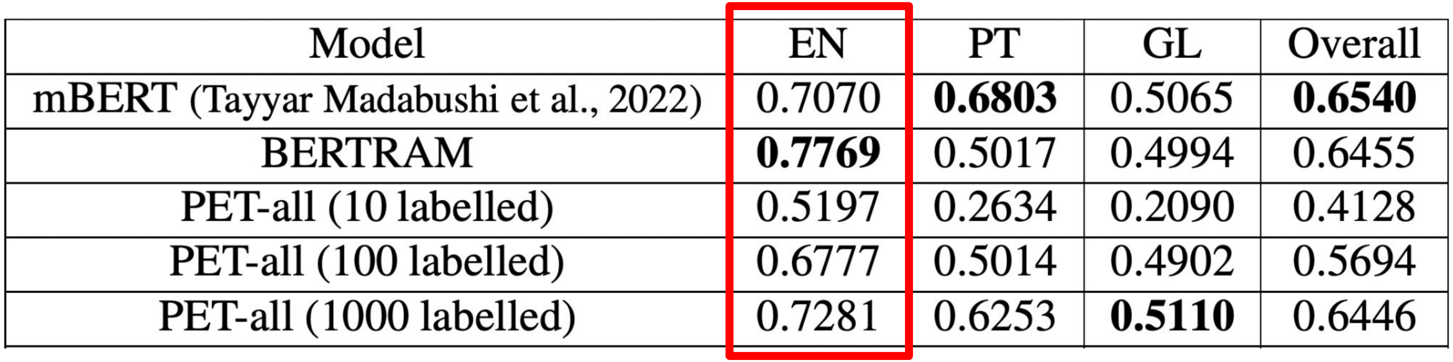
The table shows the F1 Score (Macro) on the test set, broken down into each language, for each of the models



Model	EN	PT	GL	Overall
mBERT (Tayyar Madabushi et al., 2022)	0.7070	**0.6803**	0.5065	**0.6540**
BERTRAM	**0.7769**	0.5017	0.4994	0.6455
PET-all (10 labelled)	0.5197	0.2634	0.2090	0.4128
PET-all (100 labelled)	0.6777	0.5014	0.4902	0.5694
PET-all (1000 labelled)	0.7281	0.6253	**0.5110**	0.6446

Results

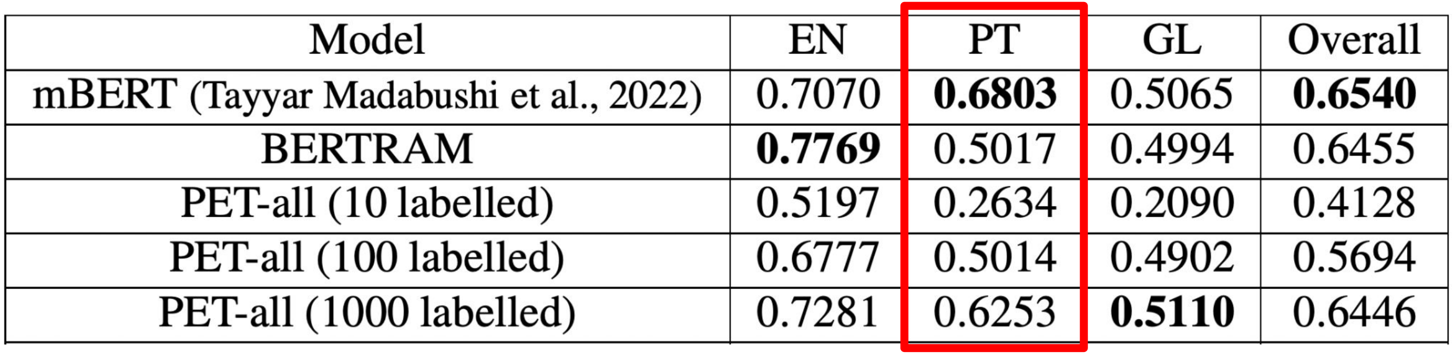
Our performance on English is encouraging, with improvements made via both methods



Model	EN	PT	GL	Overall
mBERT (Tayyar Madabushi et al., 2022)	0.7070	**0.6803**	0.5065	**0.6540**
BERTRAM	**0.7769**	0.5017	0.4994	0.6455
PET-all (10 labelled)	0.5197	0.2634	0.2090	0.4128
PET-all (100 labelled)	0.6777	0.5014	0.4902	0.5694
PET-all (1000 labelled)	0.7281	0.6253	**0.5110**	0.6446

Results

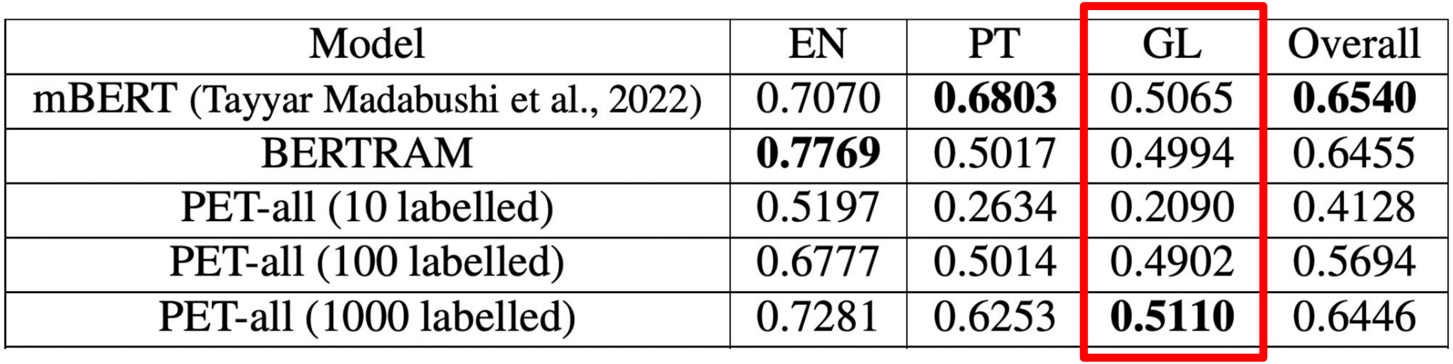
However, we see lower scores on Portuguese something we will analyse later



Model	EN	PT	GL	Overall
mBERT (Tayyar Madabushi et al., 2022)	0.7070	**0.6803**	0.5065	**0.6540**
BERTRAM	**0.7769**	0.5017	0.4994	0.6455
PET-all (10 labelled)	0.5197	0.2634	0.2090	0.4128
PET-all (100 labelled)	0.6777	0.5014	0.4902	0.5694
PET-all (1000 labelled)	0.7281	0.6253	**0.5110**	0.6446

Results

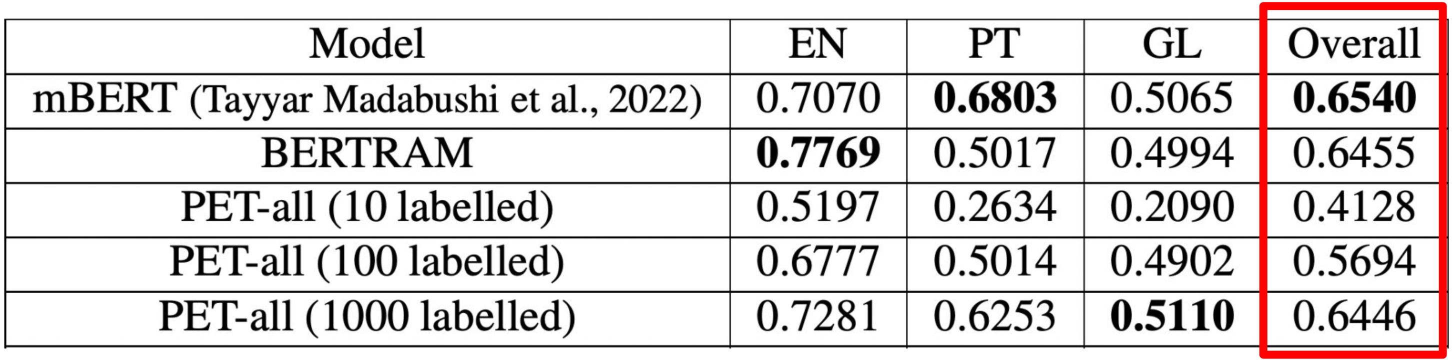
Galician performance is lower than English and Portuguese, but inline with that seen in the mBERT model



Model	EN	PT	GL	Overall
mBERT (Tayyar Madabushi et al., 2022)	0.7070	**0.6803**	0.5065	**0.6540**
BERTRAM	**0.7769**	0.5017	0.4994	0.6455
PET-all (10 labelled)	0.5197	0.2634	0.2090	0.4128
PET-all (100 labelled)	0.6777	0.5014	0.4902	0.5694
PET-all (1000 labelled)	0.7281	0.6253	**0.5110**	0.6446

Results

Overall, we see no significant increase between the mBERT model and our models



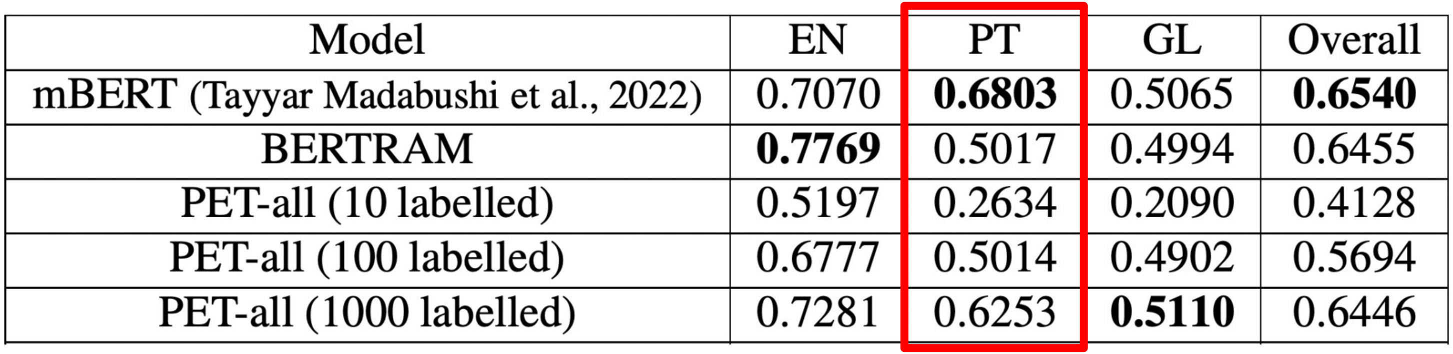
Model	EN	PT	GL	Overall
mBERT (Tayyar Madabushi et al., 2022)	0.7070	**0.6803**	0.5065	**0.6540**
BERTRAM	**0.7769**	0.5017	0.4994	0.6455
PET-all (10 labelled)	0.5197	0.2634	0.2090	0.4128
PET-all (100 labelled)	0.6777	0.5014	0.4902	0.5694
PET-all (1000 labelled)	0.7281	0.6253	**0.5110**	0.6446

Error Analysis:

Portuguese Idiom Detection

Results

Portuguese performs much worse than English within our models



Model	EN	PT	GL	Overall
mBERT (Tayyar Madabushi et al., 2022)	0.7070	**0.6803**	0.5065	**0.6540**
BERTRAM	**0.7769**	0.5017	0.4994	0.6455
PET-all (10 labelled)	0.5197	0.2634	0.2090	0.4128
PET-all (100 labelled)	0.6777	0.5014	0.4902	0.5694
PET-all (1000 labelled)	0.7281	0.6253	**0.5110**	0.6446

Portuguese Prompts

We are using English prompts for all languages, so this may lead to lower performance in languages other than English

We find that translating the prompts into portuguese/galician does not improve the performance with the F1 score falling from **0.6373** to **0.6260**

Multilingual Model

We also investigate whether the multilingual model just performs better for English than Portuguese

We see that using a single language Portuguese model and only training on the portuguese data also does not significantly improve the performance when compared with the multilingual model, with the F1 score rising from **0.4541** to **0.4621**

Conclusion

Conclusion

Low resource methods and frameworks show promise for idiom detection, even on unseen idioms

The applied methods only showed improvement in English, however we cannot find a simple explanation for poorer performance in Portuguese

Future work will focus on identifying the reasons for poorer performance and improving performance on Portuguese