Multi-Lingual Effects on Hebrew Idiom Recognition

Or Finkelstein ID. 211805155
Noa Bendet ID. 206397994
Omer Dayan ID. 319112322

Submitted as final project report for the NLP course,
Reichman University, 2025

1 Introduction

Idiomatic expressions are a special part of the language that combines culture, images, and shared meanings. They make communication richer by expressing ideas that go far beyond the meaning of each individual word. However, for computer systems, idioms are one of the hardest parts of language to understand [1].

This challenge is especially difficult in low-resource languages like Hebrew, where there is very little annotated idiom data available. Although multilingual language models such as mBERT and XLM-R can work in different languages, we still do not know how well they handle figurative language, particularly idioms[2]. Recent research on transferring knowledge about figurative speaking between languages shows that models can learn from one language and apply it to another, but success depends mainly on training data coverage, dataset structure, and language similarity rather than just how related the languages are [3].

Our project tries to fill this gap by testing multilingual transformer models on idiom tasks in Hebrew. We want to find out whether exposure to idioms in multiple languages enhances a model's ability to understand idioms in Hebrew, despite the linguistic and cultural distance. Through this research, we aim to understand how multilingual training works with figurative language and which methods work best for moving from high-resource to low-resource languages.

This work has both practical and theoretical reasons. From a practical perspective, better idiom recognition can improve applications like machine translation, chatbots, and sentiment analysis in Hebrew. From a theoretical perspective, idioms provide a good test case for cross-language learning because they combine word choice, sentence structure, and cultural differences. By studying idioms, we tackle a real-world problem and also learn more about how multilingual models capture and transfer complex language patterns.

2 Methodology

2.1 General Approach

The primary objective of this project is to determine whether cross-lingual fine-tuning of multilingual models can familiarize the models with the idiomatic concept and improve its' performance on a low-resource languages, and which tasks react best for such operation. To test this, we broke down the experiment into three variables - the model, the task and the level of fine-tuning.

For each of the tasks, we tested each of the models with each of the fine-tuning levels, in order to gain best grasp of the highest performing solution for each of the tasks.

2.2 Models

We selected three multilingual Transformer models to evaluate. For computational feasibility (RAM and GPU constraints), we used smaller or distilled variants of the standard architectures:

- **Distil-mBERT** (`distilbert-base-multilingual-cased`) — a distilled, lighter version of Multilingual BERT, originally trained on 104 languages with a masked language modeling (MLM) objective. The distilled version retains much of the performance while being faster and more memory efficient.
- **XLM-R (base)** (`xlm-roberta-base`) — a base-size variant of the multilingual RoBERTa model trained on 2.5TB of filtered CommonCrawl data across 100 languages. We chose the base configuration over large to reduce memory consumption.
- **mT5-small** (`google/mt5-small`) — a small variant of mT5, the multilingual version of the T5 model, pre-trained on 101 languages with a text-to-text objective. The small version provides a manageable trade-off between capacity and resource requirements.

Throughout the experiments, we therefore refer to these models by their family names (mBERT, XLM-R, mT5), but note that we actually used the smaller/distilled variants for feasibility.

2.3 Datasets

We combined several datasets in order to have the most relevant training sets for each task, while also having the widest variety of languages possible.

- **ID10M** [4] - a dataset composed of 10 languages, in the form of a word and its POS[1], spanning more than 100,000 tokens and containing more than 7,000 idioms.

[1]POS refers to the following set of tags for idiom detection - ['O', 'B-IDIOM', 'I-IDIOM']

- **LIdioms** [5] - a dataset comprised of 6 languages, in a .ttl format containing, among others, idioms and their English translations. Overall, it contains roughly 600 idioms.
- **EPIE** [6] - an all-English dataset containing idioms (whether static or prone to lexical changes), sentences using these idioms, and the tagging of those sentences. Overall, the dataset covers 717 idioms.
- **idiomem** [7] - an idiom-only dataset containing 800 idioms in English, to each we have created a sentence containing it, along with its tagging [8].
- **Manually-Generated Hebrew Dataset** [9] - a custom dataset created specifically for this project, providing idioms in Hebrew along with their meanings, a sentence containing each idiom, and POS tagging.

The manually-generated Hebrew dataset was a key component for both fine-tuning and evaluation. Its creation process involved several stages to ensure quality and diversity: idioms were selected from a comprehensive list of common Hebrew expressions to ensure relevance. For each idiom, a grammatically correct and contextually natural sentence was authored. The POS tagging for each sentence was performed by one of the authors and subsequently reviewed by another to ensure accuracy and consistency.

For the evaluation, the dataset, comprising 230 sentences, was partitioned into a training set and a test set using a 80/20 split. Crucially, this split was performed at the *idiom level*: all sentences containing a specific idiom were assigned exclusively to either the training set or the test set. This methodology prevents data leakage and ensures that the model is evaluated on its ability to generalize to entirely unseen idioms, rather than familiar ones in new contexts.

Language	Sentences	Tokens	Idioms	B	I	O
Chinese	9543	244422	1301	5272	3823	235327
Dutch	20935	548872	189	4530	10543	533799
English	37919	1199492	4568	10102	19884	1169506
French	35588	939161	188	12112	25248	901801
German	26963	722109	819	8311	11500	702298
Italian	29523	813445	452	8768	12353	792324
Japanese	6388	211437	165	2534	1662	207241
Polish	36333	862265	648	12971	14364	834930
Portuguese	30942	764017	559	5824	8871	749322
Spanish	28647	648776	1229	9994	13927	624855

Table 1: Statistics of the ID10M dataset across ten languages. The columns 'B' (Begin), 'I' (Inside), and 'O' (Outside) represent token counts according to the BIO annotation scheme used for identifying idiomatic expressions.

2.4 Levels of Fine-Tuning

To test the exact effect of the multilingual training on the performance, we divided the process to a few levels of fine-tuning:

Multilingual - Fine-tuning each model on datasets of all available languages. This is where we expect to see the actual impact on the tasks.

English only - Fine-tuning each model only on the English dataset. We use it as a middle-point to make sure the impact we see in the previous section doesn't rely mainly on English, and that there's value in multilinguality.

Hebrew only - Fine-tuning using very small number of Hebrew examples (up to 150). This is essentially a "no-fine-tuning" fine-tuning, mainly dictates the behavior we expect of the tasks by giving appropriate examples.

2.5 Tasks

We test each of the models' performance in several tasks, each examines a different aspect of idiomatic understanding:

2.5.1 Detection (Tagging)

Given a Hebrew sentence containing an idiom, tag the POS-s of the sentence correctly. This shows the model's ability to recognize an idiom when it sees one in use.

The training dataset for this task contains sentences incorporating the idioms, and their POS tagging.

2.5.2 Translation

Given the string "The English definition of <idiom> is: ", have the model complete the sentence with the correct definition of the idiom. This examines the model's ability to understand the meaning idioms and figures of speech.

The training dataset for this task contains sentences of the form above, followed by the correct definition of the idioms.

For this task we also added Meta's mBART and M2M100 models, since mBERT and RoBERTa perform poorly in text generation tasks.

3 Experimental results

For the detection task, we fine-tuned multiple multilingual models (`XLM-R`, `mBERT`, `mT5`) under three main training configurations:

- **Hebrew-only** — training only on Hebrew idiom data.
- **English-only** — training on English idiom and non-idiom sentences.
- **All-languages** — training on idiom data from multiple languages (Chinese, Dutch, English, French, German, Italian, Japanese, Polish, Portuguese, Spanish).

The evaluation was performed on the `hebrew_idioms` test set, which contains sentences with idioms labeled at the token level.

See metrics and qualitative examples under Appendix A

For the translation task, we fine-tuned multiple multilingual models under the same training configuration, but on data formatted as a text generation task.

See metrics and qualitative examples under Appendix B

4 Discussion

The results reveal several important patterns about cross-lingual tasks revolving around idioms:

Text Generation Limitations

The translation task, framed as a definition-generation problem, highlighted significant limitations in the generative capabilities of multilingual models when dealing with figurative language. The quantitative results were universally poor, with even the best model, M2M100, achieving a low ROUGE-1 score of just 0.2684. This weakness is further illuminated by the qualitative examples, which reveal a spectrum of failures. In the best cases, models produced "correct-ish" definitions that captured a general semantic field but missed the precise nuance, as seen with **"למען יראו ויירא"** ("so that they will see and be afraid") being translated vaguely as "to see and see; to foresee the future consequences". More severe errors included complete semantic hallucination, where idioms were given entirely unrelated meanings, such as **"אבן נגף"** (a stumbling block) being defined as "very short distance". At the extreme, the models exhibited a total breakdown in generation, producing incoherent, repetitive text like "very of of of". These findings suggest that while models like M2M100 and mBART are designed for text generation, their ability to grasp and articulate the abstract, culturally-specific meaning of idioms is profoundly limited, exposing a critical gap between recognizing linguistic patterns and performing high-fidelity semantic generation.

Given the poor translation results, the rest of the discussion section will refer to the detection task.

Cross-lingual transfer from English

XLM-R trained on English-only data achieved the highest idiom F1 score (0.571) among all settings, showing that idiomatic knowledge learned in English can transfer to Hebrew despite the linguistic distance between these languages. However, the recall (0.436) reveals an important limitation: cross-lingual transfer

works selectively and depends on structural similarity rather than just semantic meaning. The model succeeds when Hebrew idioms have direct parallels in English structure and style, but fails when encountering uniquely Hebrew expressions. This pattern is clear in our examples. Idioms like הישוב קפא על שמריו mirror the English "rest on one's laurels" in both structure and metaphor, and were detected correctly. Similarly, פרח לי מהזכרון ("flew from my memory"), which closely parallels "slipped my mind," was detected successfully. In contrast, Hebrew-specific idioms such as אין תוכו כברו were consistently missed, even though English has similar meanings like "two-faced" or "hypocrite." Even more telling, שמו נפשם בכפם ("put their souls in their palms") was completely missed despite having a clear English equivalent ("put their lives on the line"), suggesting that even semantic similarity is insufficient when the metaphorical imagery differs significantly between languages. The archaic Hebrew expression בא בימים was also consistently missed, as it has no direct idiomatic parallel in English. This suggests that successful cross-lingual transfer requires more than semantic equivalence. It needs similar metaphorical structures and linguistic patterns. For complete idiom detection in Hebrew, the model needs exposure to Hebrew-specific idiomatic patterns that cannot be learned through English training alone.

Precision vs. recall trade-off

mT5 trained on all-languages reached an extremely high precision (0.913) but a very low recall (0.158). This means it correctly identified idioms when it predicted them, but was overly conservative and avoided making positive predictions in ambiguous cases. Such behavior can be useful in some applications (where false positives are costly), but it limits general coverage. In practice, this cautious behavior means the model often marks tokens as "O" (non-idiom) even when the sentence partly contains an idiom. As a result, it misses many idioms (false negatives), even though its predictions are usually correct when it does decide to mark something as an idiom.

Interestingly, mBERT showed a different precision-recall pattern across training modes: Hebrew-only training achieved the highest recall (0.218) among all mBERT configurations, while maintaining reasonable precision (0.763). This contrasts with XLM-R, where Hebrew-only training resulted in perfect precision (1.0) but extremely low recall (0.068). This suggests that mBERT's architecture might be inherently better suited for leveraging limited target-language data. We hypothesize this could be due to its pre-training vocabulary and tokenization method, which may be more aligned with Hebrew's complex morphology compared to XLM-R's. While a full analysis of tokenization differences is beyond the scope of this project, this finding indicates a promising avenue for future research into model-architecture suitability for low-resource Semitic languages.

Limitations of Hebrew-only training

Both XLM-R and mT5 trained only on Hebrew idioms achieved very high precision (1.0 for XLM-R) but extremely low recall (0.068 for XLM-R, 0.180 for mT5). This indicates strong memorization of known idiom forms without generalizing to unseen variations. The cause is likely the small Hebrew idiom dataset, which does not provide enough variety for robust learning. In practice, this led to models performing perfectly on examples that exactly matched training idioms, but failing when idioms appeared in modified or extended syntactic forms.

However, mBERT Hebrew-only training showed a markedly different behavior, achieving the best idiom F1 score (0.339) among all mBERT configurations. This suggests that mBERT can better utilize limited target-language data compared to XLM-R and mT5, possibly due to its different pretraining approach or architectural characteristics.

Impact of multilingual training

Multilingual data slightly improved mT5's precision compared to English-only, but recall did not improve significantly. This suggests that idioms are culturally and linguistically specific, and simply adding more languages does not guarantee better recall. The model may learn many language-specific idiom structures without developing strong generalization to Hebrew. Furthermore, multilingual fine-tuning may dilute the representation space with unrelated idiomatic structures, causing the model to be more conservative in predicting idioms in Hebrew unless the match is very close.

For mBERT, multilingual training (F1=0.252) performed better than English-only training (F1=0.150) but worse than Hebrew-only training (F1=0.339). This reinforces the finding that direct target-language training can be more effective than cross-lingual transfer for certain model architectures.

Architecture differences

XLM-R consistently outperformed mBERT in idiom F1, despite both being encoder-only models, likely due to its larger size and more extensive multilingual pretraining. mT5, being an encoder–decoder model, may be less optimized for token-level classification tasks, which could explain its lower recall across all training configurations.

The updated results reveal that mBERT shows a unique pattern where Hebrew-only training outperforms both English-only and multilingual training, suggesting that different architectures may have varying capacities for utilizing limited target-language data effectively.

Error patterns

For all models, the most common misses (false negatives) happened when idioms:

- Were very figurative and had almost no similarity to English idioms.

- Had extra words in the middle that broke the idiom into separate parts.
- Used old or unusual Hebrew words that were not seen in the training data.

False positives often happened when the model marked a phrase as an idiom just because it sounded metaphorical, even though it wasn't a real idiom, especially in long sentences.

Overall insights

- Models can learn idiom recognition across languages, but transfer quality depends on similarity of idiom usage patterns between source and target languages.
- Small monolingual datasets lead to very high precision but poor recall for XLM-R and mT5, but mBERT demonstrates better recall (0.218) with Hebrew-only training, suggesting architectural differences in handling limited data.
- Multilingual training improves precision but can lower recall if the model becomes cautious in cross-language contexts.
- XLM-R appears to be the strongest choice for idiom detection overall, but mBERT Hebrew-only training shows promising results for scenarios with limited target-language data.
- Model architecture significantly affects the precision-recall trade-off: encoder-decoder models (mT5) tend toward high precision but low recall, while encoder-only models show more varied patterns depending on training data.

In future work, combining a high-precision multilingual model (such as mT5 all-languages) with a high-recall English-transfer model (such as XLM-R English-only) could potentially yield a more balanced system for Hebrew idiom detection. Additionally, the promising Hebrew-only results for mBERT suggest that investing in larger Hebrew idiom datasets, particularly for certain model architectures, could lead to significant improvements in both precision and recall. Given these findings, a hybrid fine-tuning strategy, starting with multilingual idiom exposure, followed by targeted Hebrew idiom augmentation, appears to be the most promising path forward.

5 Code

The code can be found in the following: Detection notebook, Translation notebook.

References

[1] Isuri Arachchige, Sachith Suraweera, and Dulip Herath. Transformer-based language models for the identification of idiomatic expressions. https://www.acl-bg.org/proceedings/2022/EUROPHRAS%202022/pdf/2022.europhras-1.15.pdf, 2022. Proceedings of EUROPHRAS 2022, pp. 119–127.

[2] Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. Natural language processing applications for low-resource languages. https://www.cambridge.org/core/journals/natural-language-processing/article/natural-language-processing-applications-for-lowresource-languages/7D3DA31DB6C01B13C6B1F698D4495951, 2025. Natural Language Processing, vol. 31, pp. 183–197.

[3] Julia Sammartino, Libby Barak, Jing Peng, and Anna Feldman. When does language transfer help? sequential fine-tuning for cross-lingual euphemism detection. https://arxiv.org/abs/2508.11831, 2024. arXiv preprint arXiv:2508.11831.

[4] Babelscape. Id10m: Idiom detection in 10 languages. https://github.com/Babelscape/ID10M. Accessed: 2025.

[5] DICE Group. Lidioms: A multilingual idiom dataset. https://github.com/dice-group/LIdioms. Accessed: 2025.

[6] Prateek Saxena. Epie corpus: English phrasal idiom expression corpus. https://github.com/prateeksaxena2809/EPIE_Corpus. Accessed: 2025.

[7] Adi Haviv. Idiomem: English idiom dataset. https://github.com/adihaviv/idiomem. Accessed: 2025.

[8] Idiomem corrected dataset. https://github.com/omerday/nlp-idiom-he/blob/main/idiomem-corrected.txt. Accessed: 2025.

[9] Hebrew idioms dataset. https://github.com/omerday/nlp-idiom-he. Accessed: 2025.

Appendix A Detection Metrics

Metrics include:

- **Accuracy** — overall token classification accuracy.
- **Weighted F1** — weighted by class frequency.
- **Idiom Precision / Recall / F1** — for binary idiom detection (any token tagged as part of an idiom vs. none).

Model	Training Mode	Acc.	Prec.	Recall	Idiom F1
XLM-R	All-languages	0.810	0.831	0.406	0.545
XLM-R	English-only	**0.819**	0.829	0.436	**0.571**
XLM-R	Hebrew-only	0.757	1.000	0.068	0.127
mBERT	All-languages	0.763	0.769	0.150	0.252
mBERT	English-only	0.753	0.786	0.083	0.150
mBERT	Hebrew-only	0.775	0.763	0.218	0.339
mT5	Hebrew-only	0.672	0.312	0.180	0.229
mT5	English-only	0.767	0.826	0.143	0.244
mT5	All-languages	0.777	**0.913**	0.158	0.269

Table 2: Detection task results on Hebrew idioms. Precision, recall, and F1 are computed for binary idiom-vs-non-idiom classification.

Qualitative Examples

Correct Detections

Example 1:
הישוב קפא על שמריו, ולא התפתח במשך שנים רבות.
```
True: ['O', 'B-IDIOM', 'I-IDIOM', 'I-IDIOM', 'O', 'O', 'O', 'O', 'O']
Pred: ['O', 'B-IDIOM', 'I-IDIOM', 'I-IDIOM', 'O', 'O', 'O', 'O', 'O']
```

Example 2:
משאת נפשם של הורי היתה לעלות לארץ ישראל
```
True: ['B-IDIOM', 'I-IDIOM', 'O', 'O', 'O', 'O', 'O', 'O']
Pred: ['B-IDIOM', 'I-IDIOM', 'O', 'O', 'O', 'O', 'O', 'O']
```

Example 3:
הבטחתי לצלצל אליך, אך הדבר פרח לי מהזכרון.
```
True: ['O', 'O', 'O', 'O', 'O', 'B-IDIOM', 'I-IDIOM', 'I-IDIOM']
Pred: ['O', 'O', 'O', 'O', 'O', 'B-IDIOM', 'I-IDIOM', 'I-IDIOM']
```

Incorrect / Missed Predictions

Example 1:
הלוחמים שמו נפשם בכפם בעת הגנתם על מדינת ישראל

True: ['O', 'B-IDIOM', 'I-IDIOM', 'I-IDIOM', 'O', 'O', 'O', 'O', 'O']
Pred: ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']

Example 2:

אדם זה, אין תוכו כברו, אומר את ההפך ממה שהוא חושב באמת.

True: ['O', 'O', 'B-IDIOM', 'I-IDIOM', 'I-IDIOM', 'O', 'O', 'O', 'O', 'O', 'O', 'O']
Pred: ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']

Example 3:

רב הקהלה בא בימים, ועדין דעתו צלולה.

True: ['O', 'O', 'B-IDIOM', 'I-IDIOM', 'O', 'O', 'O']
Pred: ['O', 'O', 'O', 'O', 'O', 'O', 'O']

Example 4:

השופט פקח את עיניו של השר, והוא הבין את טעותו.

True: ['O', 'B-IDIOM', 'I-IDIOM', 'I-IDIOM', 'O', 'O', 'O', 'O', 'O', 'O']
Pred: ['O', 'B-IDIOM', 'I-IDIOM', 'O', 'O', 'O', 'O', 'O', 'O', 'O']

Appendix B Translation Metrics

Measured Metric:

- **Rogue-1** — a metric used to evaluate the quality of automatic text summarization and machine translation.

Model	Training Mode	ROUGE-1
M2M100	Hebrew-only	0.2684
M2M100	English-only	0.1130
M2M100	All-languages	0.1191
mBART	Hebrew-only	0.1112
mBART	English-only	0.0848
mBART	All-languages	0.0908
mT5	Hebrew-only	0.0212
mT5	English-only	0.0327
mT5	All-languages	0.0529
mBERT	Hebrew-only	0.0000
mBERT	English-only	0.0005
mBERT	All-languages	0.0410
XLM-RoBERTa	Hebrew-only	0.0000
XLM-RoBERTa	English-only	0.0008
XLM-RoBERTa	All-languages	0.0852

Table 3: Translation task results on Hebrew idioms. ROUGE-1 is reported for different training scenarios.

Qualitative Examples

Correct-ish Translations

Example 1:

```
The English translation of למען יראו וייראו is:
True: so that they will see and be afraid; to make an example of someone
Pred: to see and see; to foresee the future consequences of an action
```

Incorrect Translations

Example 2:

```
The English translation of אבן נגף is:
True: a stumbling block; an obstacle
Pred: very short distance
```

Example 3:
The English translation of אבן שאין לה הופכין is:
True: a useless object; something that has no use
Pred: a person who has no patience or energy to change

Faulty Translations

Example 4:
The English translation of אוצר בלום is:
True: a hidden treasure; often used to describe a person with vast, untapped knowledge
Pred: very of of of