A Dataset for Metaphor Detection in Early Medieval Hebrew Poetry

TECHNION | The Henry and Marilyn Taub Faculty of Computer Science | האוניברסיטה הפתוחה THE OPEN UNIVERSITY OF ISRAEL الجامعة المفتوحة

Michael Toker[1]
tok@cs.technion.ac.il

Oren Mishali[1]
omishali@cs.technion.ac.il

Ophir Münz-Manor[2]
ophirmm@openu.ac.il

Benny Kimelfeld[1]
bennyk@cs.technion.ac.il

Yonatan Belinkov[1]
belinkov@technion.ac.il

[1]Technion – Israel Institute of Technology
[2]The Open University of Israel

The corpus of late antique and medieval Hebrew is vast and represents a crucial linguistic and cultural bridge between Biblical and modern Hebrew. Poetry is prominent in texts from these time periods and one of its characteristics is the use of metaphor. Distinguishing figurative and literal words or phrases is a major task for scholars in the Humanities, especially in the fields of literature, linguistics, and hermeneutics. We present a new, challenging dataset of early medieval Hebrew poetry with expert annotations of metaphors with baseline results.

The Dataset

- Early medieval Hebrew poems.
- 309 labeled texts, 13,000 sentences, 75,000 words.
- 15% of the words are metaphors.

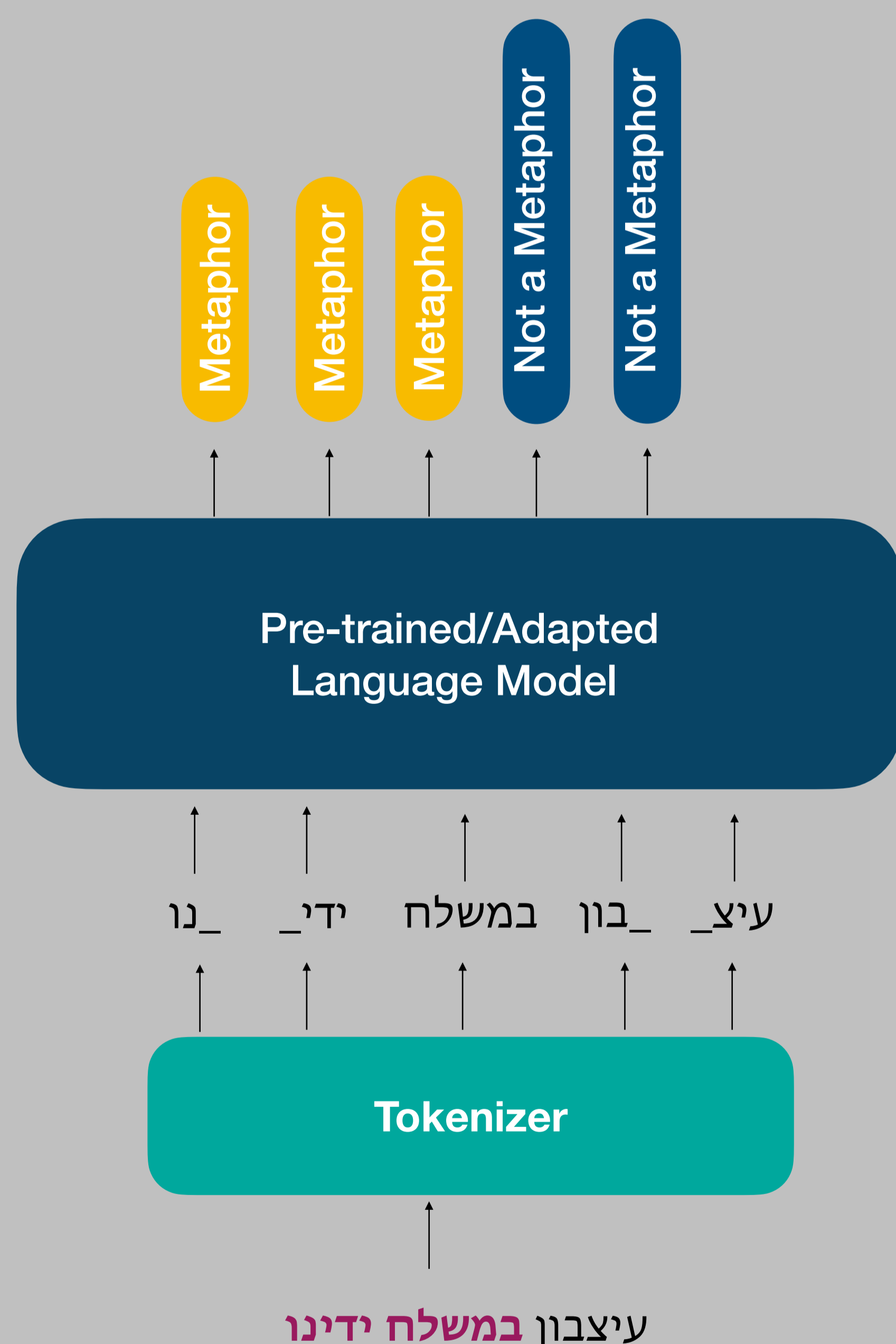| Hebrew	English	Meaning
טבענו בגזרות, כשטופי ים	We drowned in decrees, like sea washed	There are too many decrees
הצית נשמה ויבער נר	Ignite a soul, fire a candle	Create a soul
עד דור עשירי לא עשו פרי	Until the tenth generation they did not bear fruit	For ten generations, they did no good deeds

Inter Annotator Agreement

Inter-annotator agreement evaluates how consistently different annotators label data for NLP research (measured in Cohen's kappa). The score for our dataset is 0.618 (considered substantial).
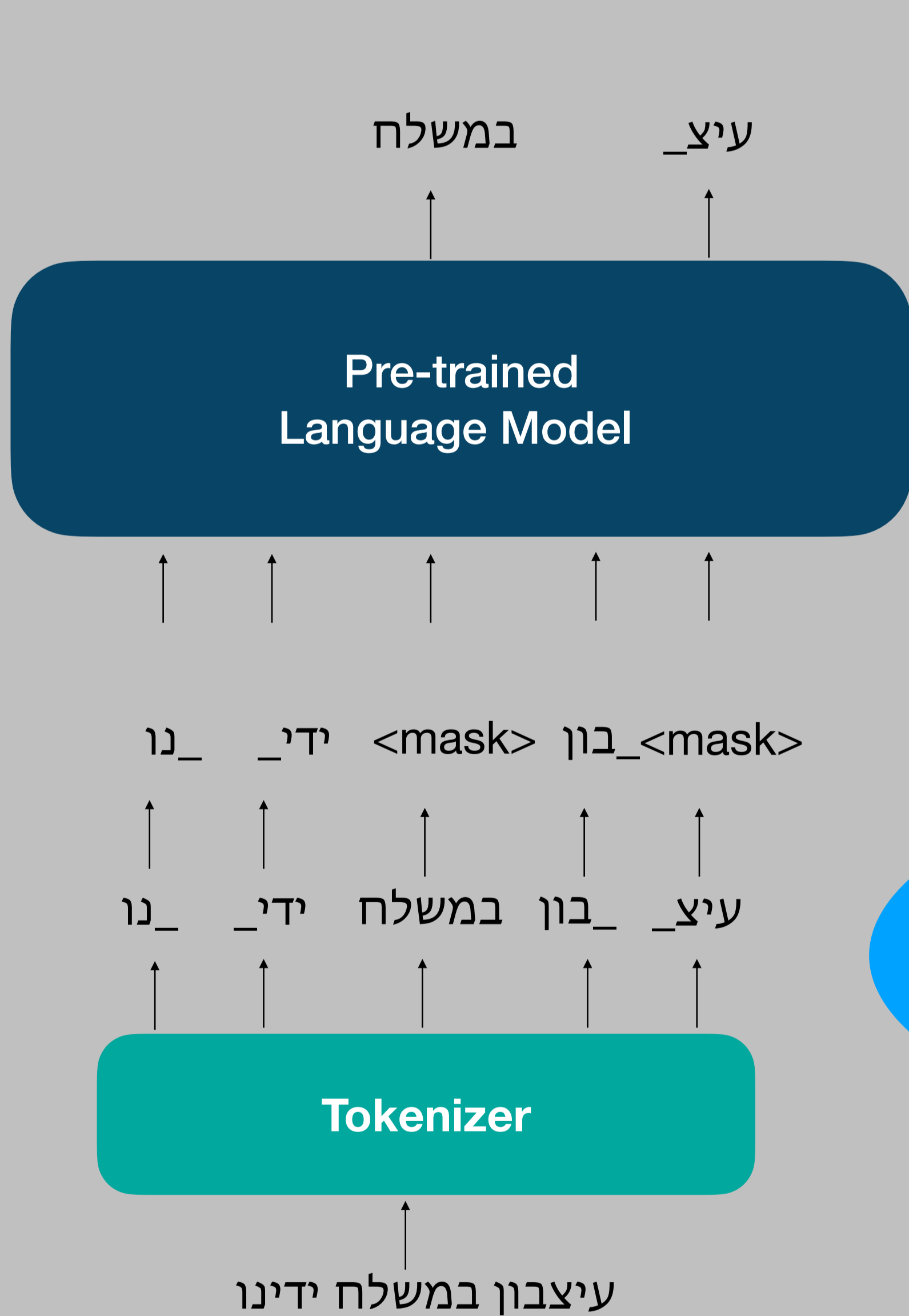

A medieval Hebrew poem from the Worms Prayerbook, Germany, 13th Century, The National Library of Israel, "Ktiv" Project

Methods

Fine-tuning for metaphor detection

Metaphor | Metaphor | Metaphor | Not a Metaphor | Not a Metaphor

Pre-trained/Adapted Language Model

נו_ | ידי_ | במשלה | _בון | עיצ_

Tokenizer

עיצבון במשלח ידינו

LM adaptation

במשלח | עיצ_

Pre-trained Language Model

נו_ | ידי_ | \<mask\> | _בון | \<mask\>_

נו_ | ידי_ | במשלח | _בון | עיצ_

Tokenizer

עיצבון במשלה ידינו

Trained on additional prose and poetic medieval texts (Project Ben-Yehuda*) + Piyyut dataset

*Project Ben-Yehuda is a volunteer-based online library of free Hebrew literature.

Language models (pre-trained & adapted)

	Data Size (words)	Data Type
BEREL	220 Milion	Medieval
AlephBERT	2 Bilion	Modern
AlephBERT Adapted	2 Bilion + 10 Milion	Modern + Medieval

Results

- AlephBERT is outperformed by AlephBERT Adapted.
- **BEREL**, pre-trained on medieval texts, shows the best results.

f1 Score on test set

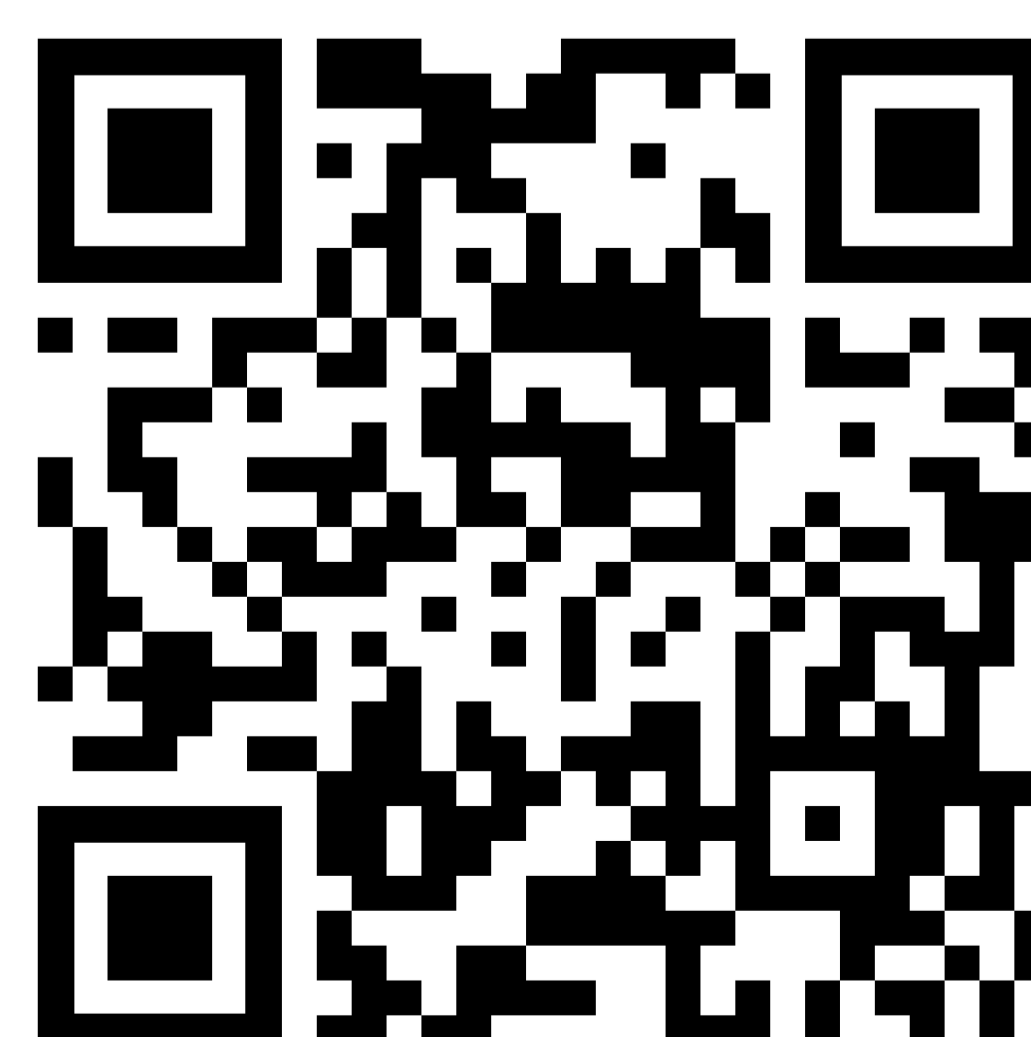|	f1 Score
AlephBERT	44.404
BEREL	48.3
AlephBERT - adapted	46.56
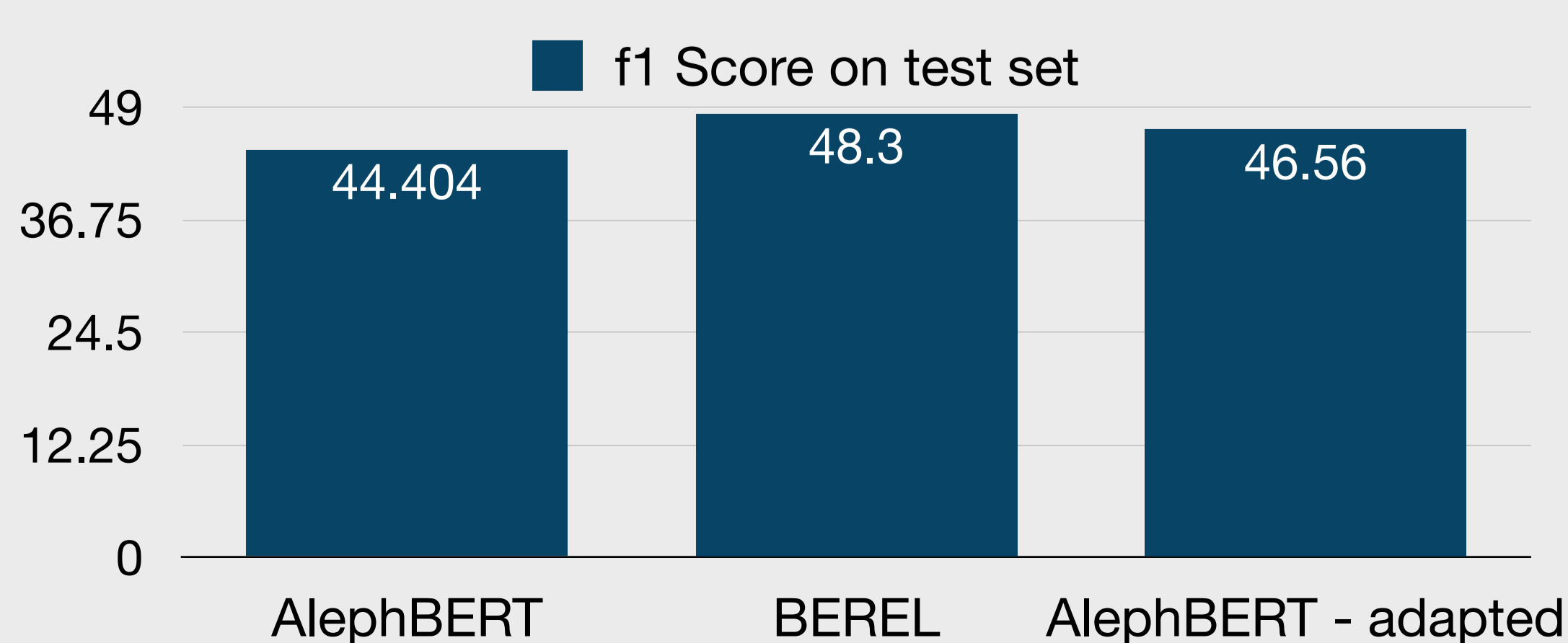
Error Analysis

ילד לניחוח חשקתי (A child for a scent - I have desired)

Berel

ילד לניחוח חשקתיו

Expert explanation: The "child" represents Isaac and "scent" stands for cultic sacrifice; both are non-metaphorical epithets.

Project Page