BLiMP: A Benchmark of Linguistic Minimal Pairs for English

Alex Warstadt  
*New York University*, alexwarstadt@gmail.com

Alicia Parrish  
*New York University*, alicia.v.parrish@nyu.edu

Haokun Liu  
*New York University*, haokunliu@nyu.edu

Anhad Mohananey  
*New York University*, amohananey@ea.com

Wei Peng  
*New York University*, weipeng@nyu.edu

*See next page for additional authors*

Follow this and additional works at: [https://scholarworks.umass.edu/scil](https://scholarworks.umass.edu/scil)  

Part of the Computational Linguistics Commons

**Recommended Citation**

Warstadt, Alex; Parrish, Alicia; Liu, Haokun; Mohananey, Anhad; Peng, Wei; Wang, Sheng-Fu; and Bowman, Samuel R. (2020) "BLiMP: A Benchmark of Linguistic Minimal Pairs for English," *Proceedings of the Society for Computation in Linguistics*: Vol. 3 , Article 43.  
DOI: [https://doi.org/10.7275/zejz-qs04](https://doi.org/10.7275/zejz-qs04)  
Available at: [https://scholarworks.umass.edu/scil/vol3/iss1/43](https://scholarworks.umass.edu/scil/vol3/iss1/43)

This Abstract is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Proceedings of the Society for Computation in Linguistics by an authorized editor of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
BLiMP: A Benchmark of Linguistic Minimal Pairs for English

Authors
Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman

This abstract is available in Proceedings of the Society for Computation in Linguistics: https://scholarworks.umass.edu/scil/vol3/iss1/43
BLiMP💡: A Benchmark of Linguistic Minimal Pairs for English

Alex Warstadt,1 Alicia Parrish,1 Haokun Liu,3 Anhad Mohananey,5 Wei Peng,3 Sheng-Fu Wang,1 and Samuel R. Bowman1,2,3

1NYU Linguistics 2NYU Data Science 3NYU Computer Science
Correspondence: warstadt@nyu.edu

Introduction & Prior Work

We introduce BLiMP (The Benchmark of Linguistic Minimal Pairs, or 📘), a large new benchmark dataset for the targeted evaluation of statistical language models’ knowledge of linguistic phenomena. The benchmark consists of 67 datasets, each containing 1000 minimal pairs isolating a specific grammatical contrast and collectively offering broad coverage of major phenomena in English grammar. Like the GLUE benchmark for reusable sentence understanding models (Wang et al., 2018), 📘 assigns a single numerical score to a language model (LM) measuring its overall mastery of grammar, enabling straightforward comparison of LMs. The dataset is ideal for fine grained analysis of an LM’s knowledge of different grammatical domains. For baselines, we evaluate four representative LMs from NLP literature. We find that 📘 is hard even for state-of-the-art models, though Transformers perform better than LSTM and n-gram LMs. Humans overwhelmingly agree with the generated minimal pair contrasts in 📘.

A growing body of work evaluates LSTM LMs’ knowledge of grammar by testing whether they prefer acceptable sentences over minimally different unacceptable ones (Linzen et al., 2016, a.o.). So far, results have been mixed, motivating the creation of this benchmark which scales up this kind of investigation to isolate dozens of grammatical contrasts within an otherwise-uniform controlled artificial dataset. Our results show that knowledge of grammar has increased as LM technology progressed from n-grams to LSTMs to Transformers. LSTMs and Transformers alike are very accurate in detecting morphological and agreement violations, but state-of-the-art Transformer LMs have an especially large advantage over LSTMs in contrasts where simple generalizations are difficult to find, such as NPI licensing and island effects.

Data 📘 consists of 67 datasets of 1000 minimal pairs each, grouped into twelve broader categories (Table 1). A minimal pair consists of two minimally different sentences where one is grammatically acceptable and the other is not. All minimal pairs in 📘 contain the same number of tokens and differ only in word order or the identity of one lexical item, following Marvin and Linzen (2018).

We include minimal pairs illustrating linguistic phenomena well known in morphology, syntax, and semantics. While this set is not exhaustive, it does cover a wide range of topics found in formal implementations of English grammar (e.g., HPSG; generative linguistics textbooks). To fully isolate the phenomena of interest, we use realistic artificially-generated sentences, following Marvin and Linzen, a.o. To generate text, we construct a vocabulary of over 3300 lexical items labeled with features reflecting morphology (e.g. singular/plural), syntax (e.g. transitive/intransitive), and semantics (e.g. animate/inanimate), and build a simple artificial grammar for each paradigm.

We validate the acceptability contrasts in the generated pairs with Mechanical Turk annotators, testing 5 randomly-selected pairs from each paradigm using the same forced-choice task models are presented with. Majority vote of 20 annotators agrees with 📘 on at least 4/5 examples from each paradigm and on 96.4% of pairs overall.

Baselines 📘 We evaluate 4 baselines: (1) An n-gram LM trained on the English Gigaword corpus (Graff et al., 2003), based on a modified Kneser Ney implementation by (Heafield, 2011), which considers up to 5-grams, restricting the model from learning dependencies spanning more than 5 words. (2) An LSTM recurrent neural network LM from Gulordava et al. (2018). (3) Transformer-XL (Dai et al., 2019), a transformer LM with additional features that enable it to model
We have offered a human-solvable challenge set that covers a broad overview of major grammatical phenomena in English. is hard even for SotA models, though recent large-scale Transformers outperform simple baselines.

References