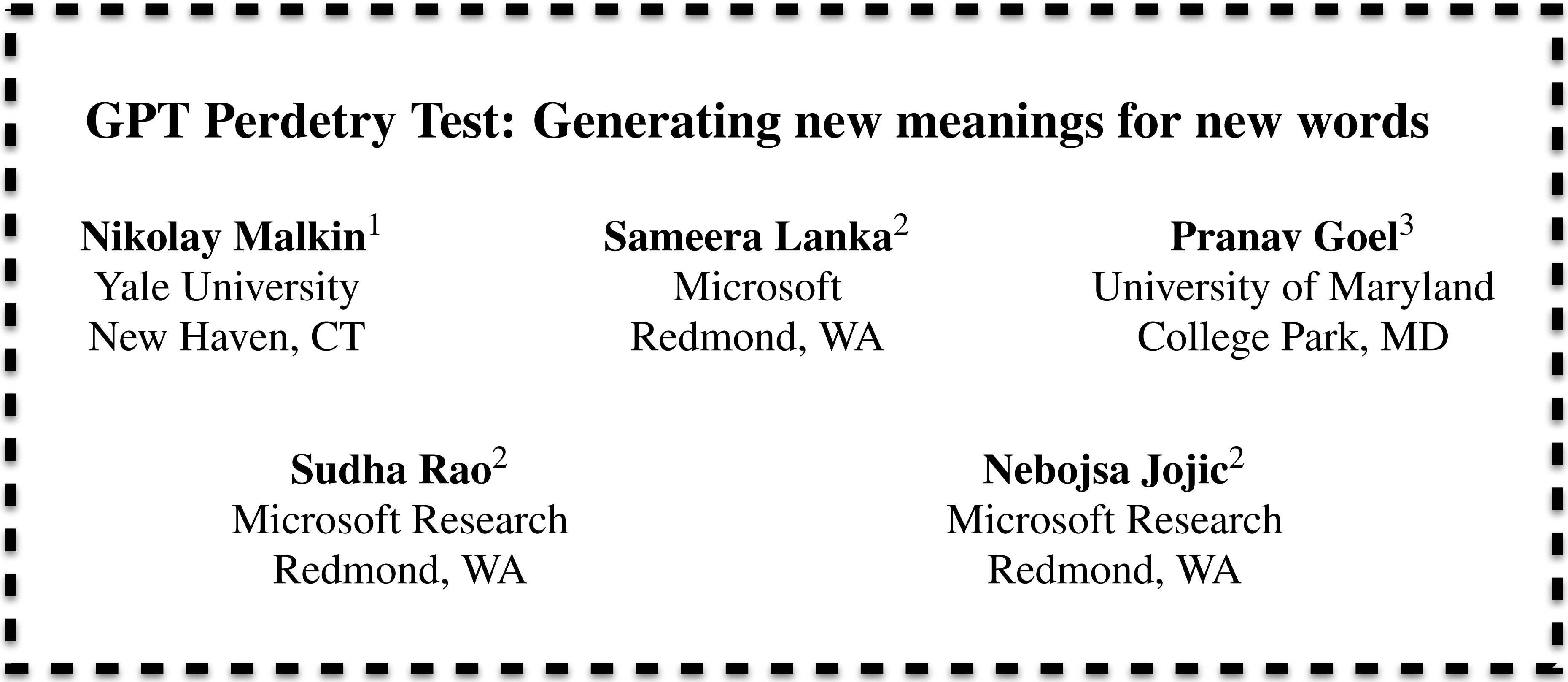
Computational morphology reading group



GPT Perdetry Test: Generating new meanings for new words

Nikolay Malkin[1]
Yale University
New Haven, CT

Sameera Lanka[2]
Microsoft
Redmond, WA

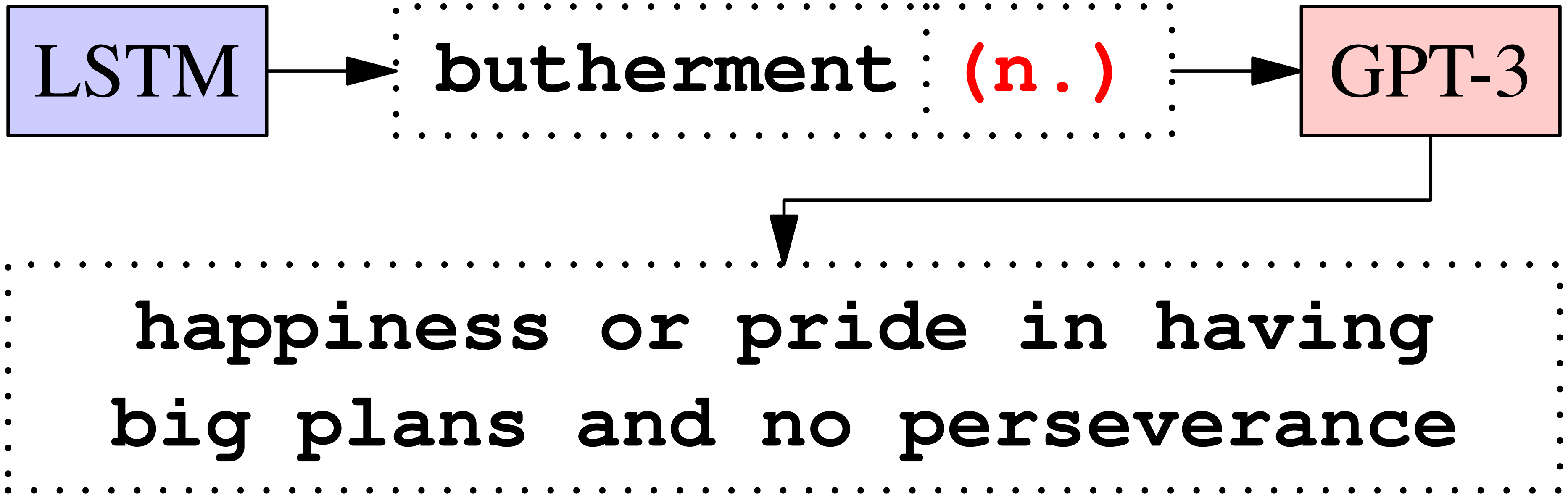
Pranav Goel[3]
University of Maryland
College Park, MD

Sudha Rao[2]
Microsoft Research
Redmond, WA

Nebojsa Jojic[2]
Microsoft Research
Redmond, WA

June 18, 2021
Miikka Silfverberg

General idea



Explore large pre-trained LM's to define neologisms.

A specific LM GPT-3 is shown to generate realistic, original definitions for new words

This finding sheds light on GPT-3's ability to adapt to and even extend a changing vocabulary

Background

Orthography functions as a clue for word meaning

Phonosemantic patterns are common in many languages:

glimmer, **gl**ow, **gl**itter, **gl**oss

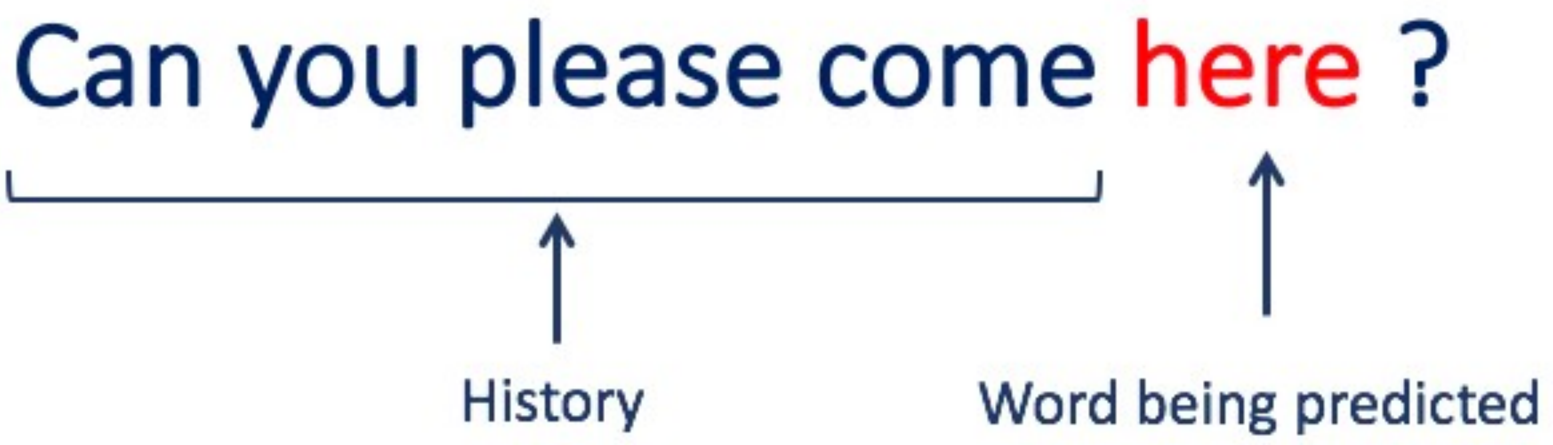
Meanings may be associated with lexical strata in English:

abstract nouns from Norman French,
concrete nouns from the Germanic substrate
artificially constructed terms with Greek or Latinate elements

This information might be learnable by language models like GPT-3

GPT-3

Generative pre-trained transformer or GPT(-3)
is a language model:



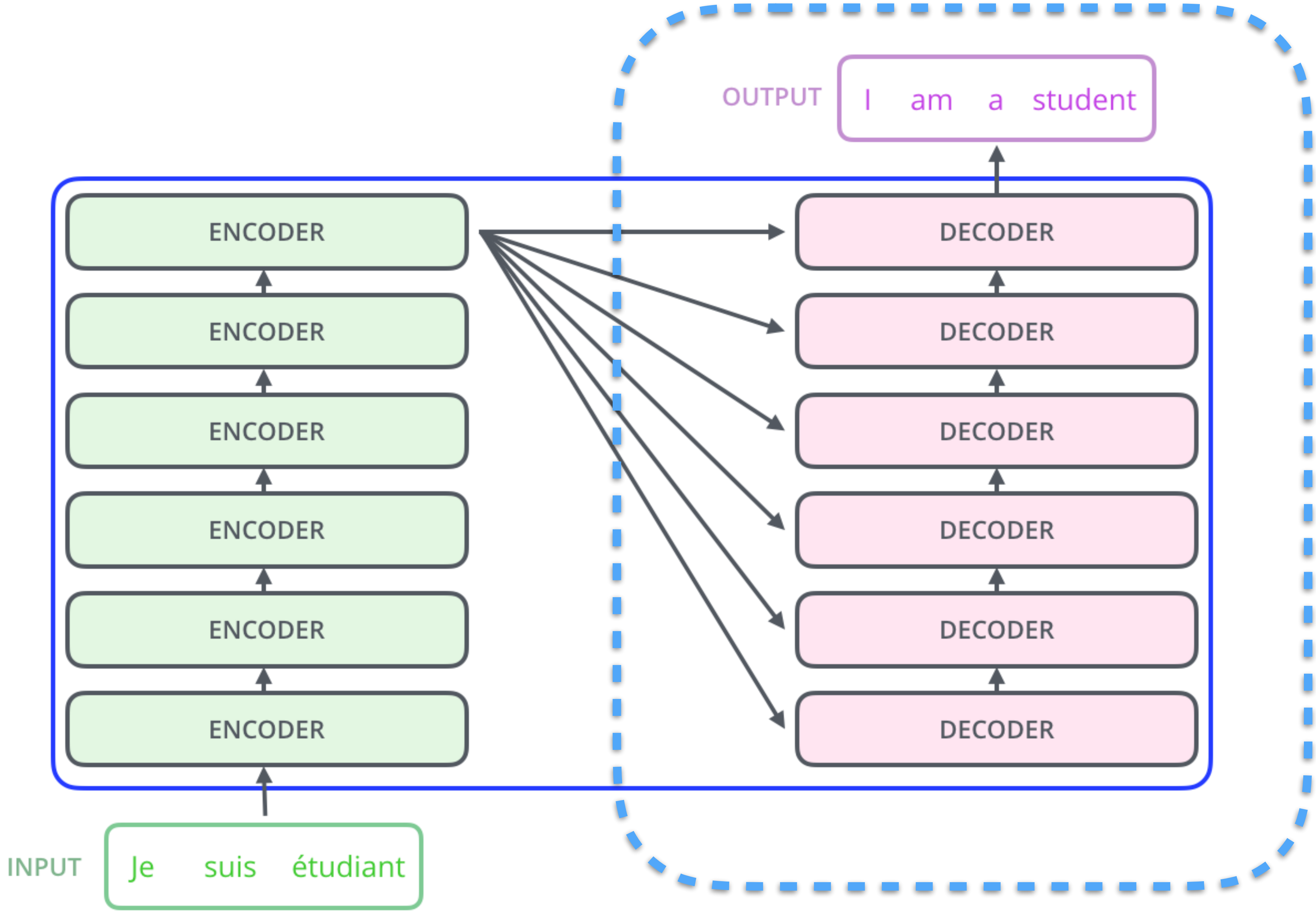
Can you please come here ?

History

Word being predicted

$$L_1(T) = \sum_i log\, P(t_i|t_{i-k}, \ldots, t_{i-1}; \theta)$$

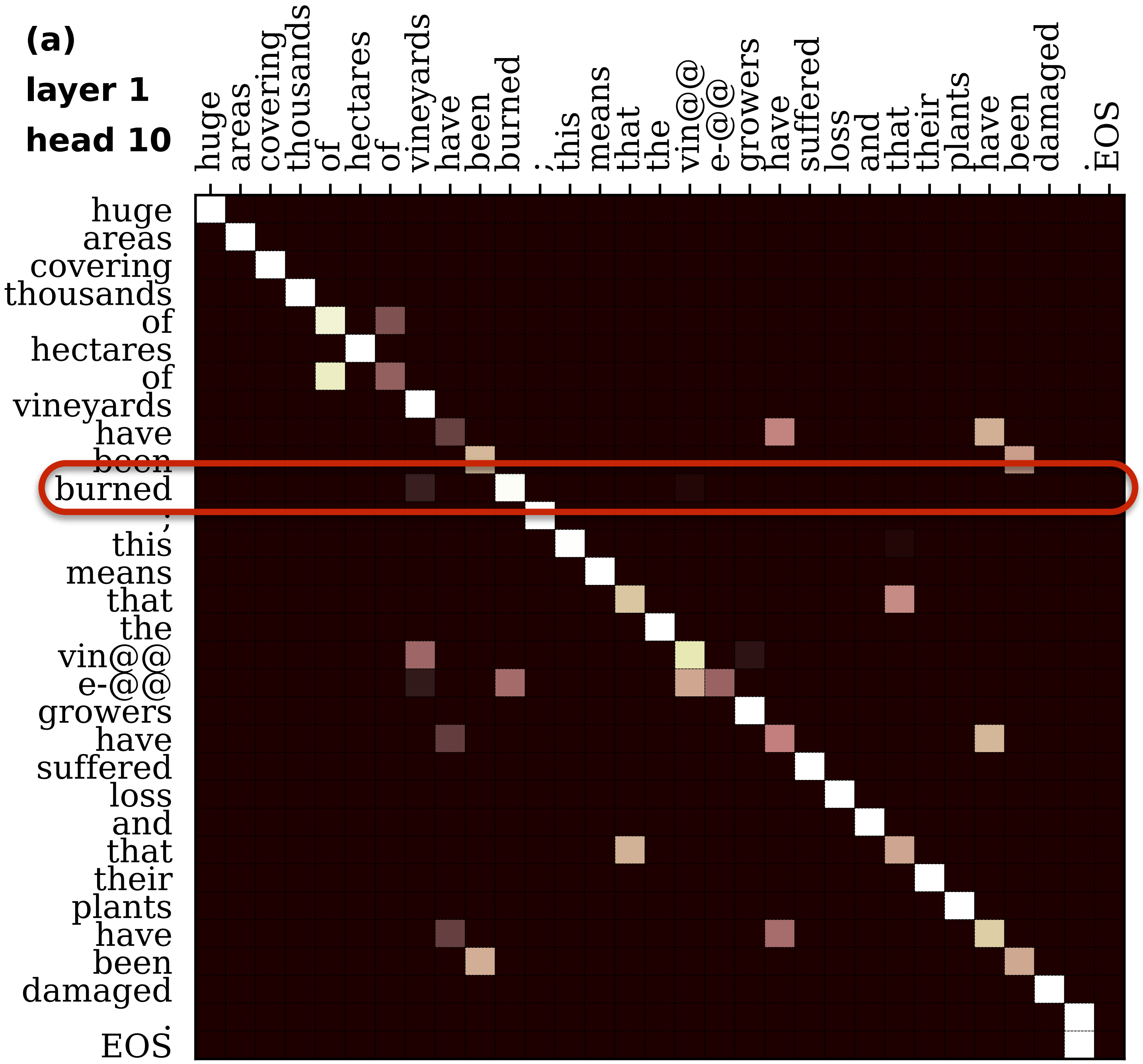
Short recap of transformers



Originally for machine translation

Layers and heads

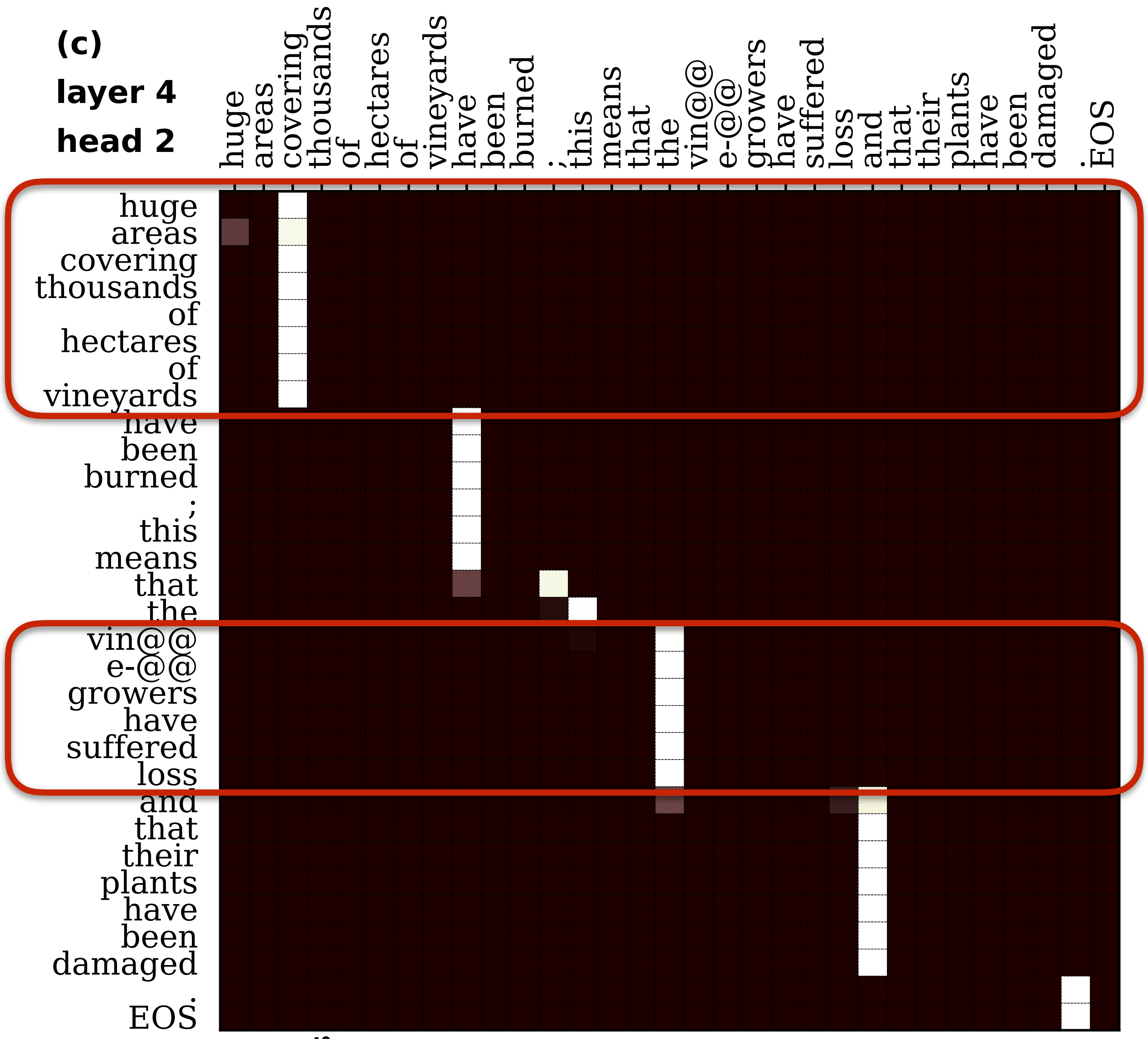
In lower layers many heads attend to very local context



D. Marecek and R. Rosa. *From Balustrades to Pierre Vinken: Looking for Syntax in Transformer Self-Attentions*, BlackboxNLP 2019

Layers and heads

Phrase-like units start to appear in later layers



D. Marecek and R. Rosa. *From Balustrades to Pierre Vinken: Looking for Syntax in Transformer Self-Attentions*, BlackboxNLP 2019

Difference between GPT-1, GPT-2 and GPT-3

GPT-3 is bigger:
GPT-1 (2018): 117M parameters
GPT-2 (2019): 1.5G parameters
GPT-3 (2020): 175G parameters

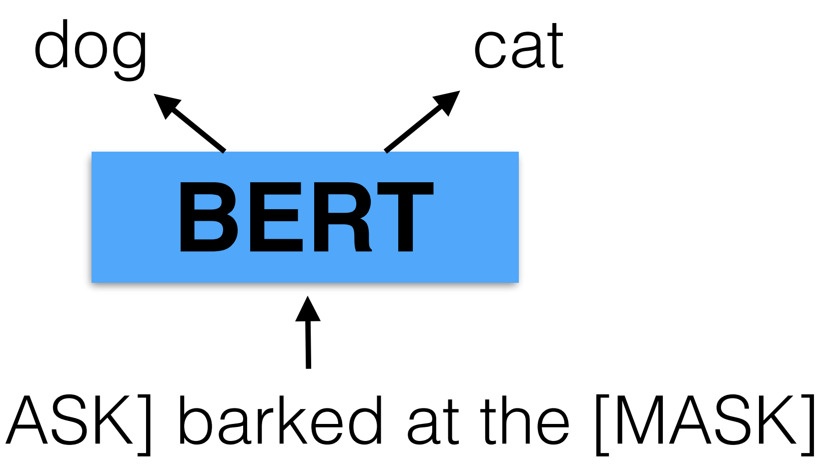
GPT-3 is deeper:
GPT-1: 12 layers,
GPT-2: 48 layers,
GPT-3: 96 layers

GPT-3 vs. BERT

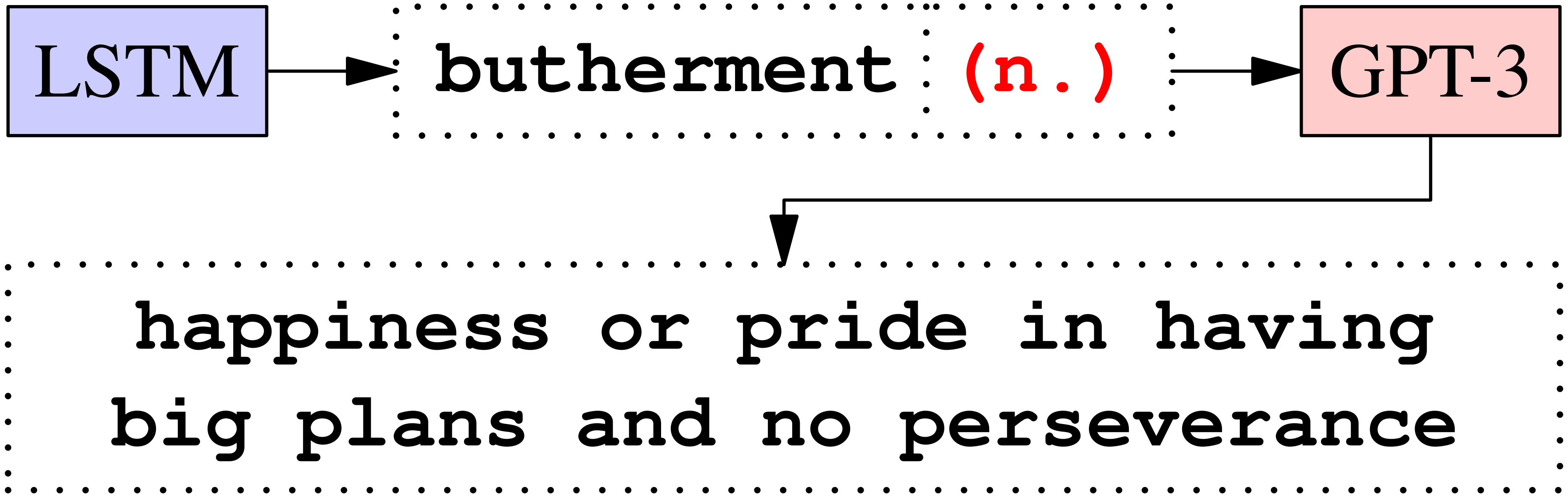
GPT-3 is a **generative** model. It can generate text:

"Buddhists are **divided into two main branches - Theravada and Mahayana. Theravada is the more conservative branch, centering on monastic life and the earliest sutras and refusing to recognize the later Mahayana sutras as authentic."**

BERT is a masked language model. It can't generate



Generate definitions using GPT-3

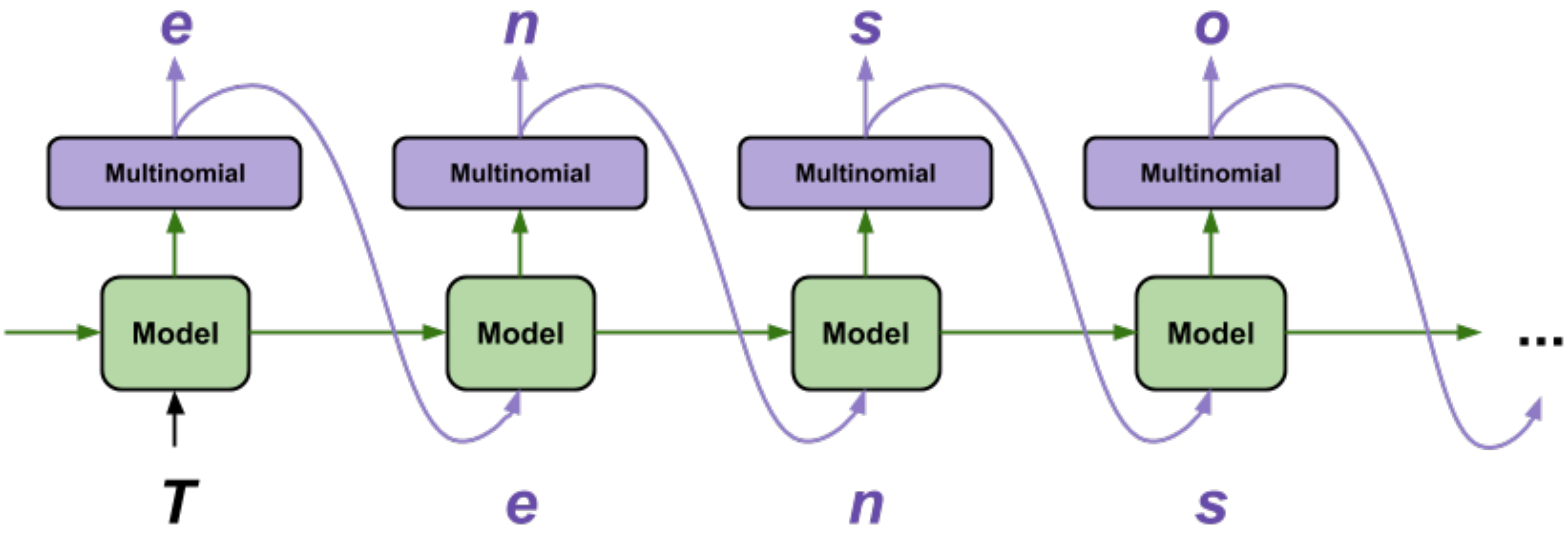


A character-based LSTM model first generates a nonce word “butherment”

GPT-3 then generates a definition for this word

Generating words using an LSTM

The authors train an LSTM model to generate English-like nonce words on a set of 466k existing words



The words were manually(?) lemmatized and assigned POS tags (n., v. or adj.)

Sub-word encoding

Internally, GPT-3 represents words as sequences of sub-word units:

`per|detry, har|bole|mic, sh|out|ze`

When GPT-3 is trained, its training data is first segmented into sub-word units using byte-pair encoding (BPE)

Common character sequences like “the” will be represented as complete units

Rare units are split into morpheme-like units

`unfriend -> un|friend`

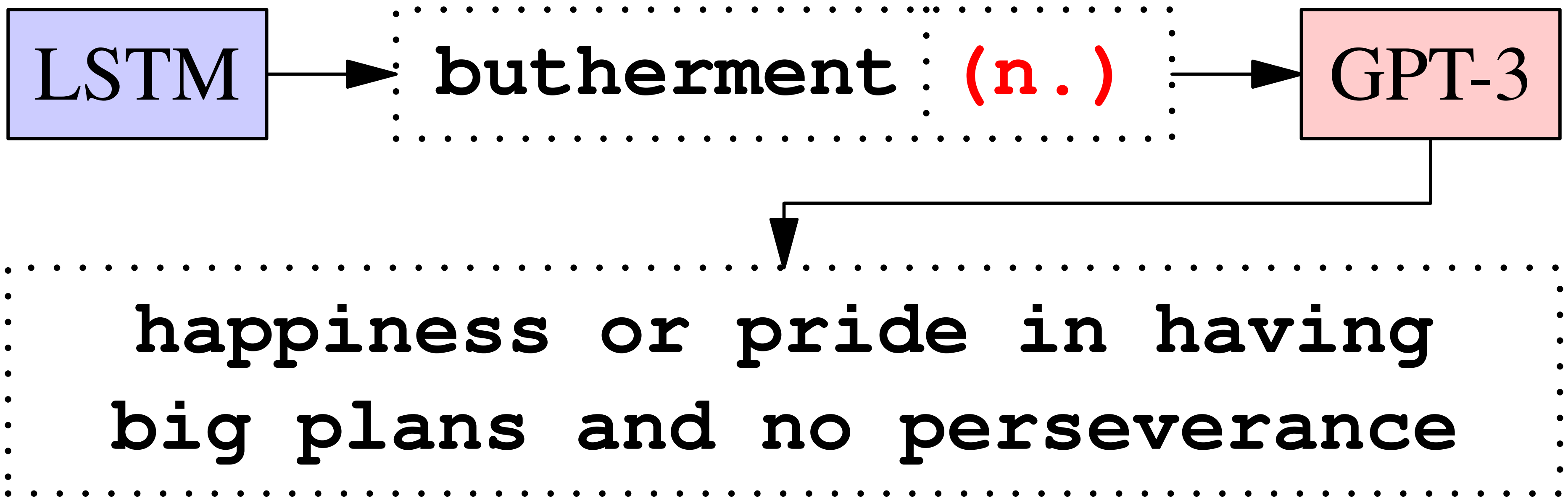
Sub-word encoding

These sub-word units allow us to input any character-sequence into GPT-3:

```
per|detry, har|bole|mic, sh|out|ze
```

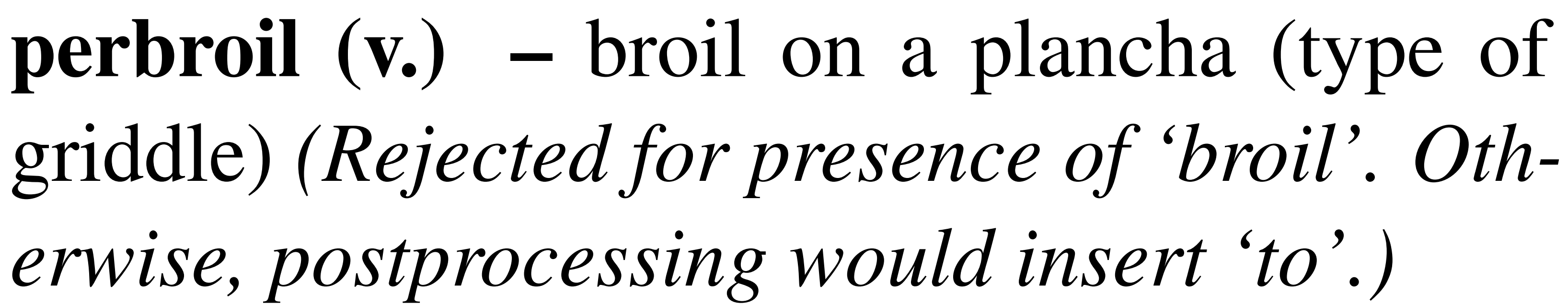
Generating definitions

In its immense training data, GPT-3 has seen enough examples of dictionary entries to generate convincing definitions:



Post-processing of definitions

The authors filtered out definitions which didn’t resemble dictionary definitions, circular definitions and output containing obscene or violent content:



perbroil (v.) – broil on a plancha (type of griddle) *(Rejected for presence of ‘broil’. Otherwise, postprocessing would insert ‘to’.)*

To clean the definitions, the authors removed parenthesized comments and alternative senses and made minor edits for consistent syntax and punctuation.

Example definitions

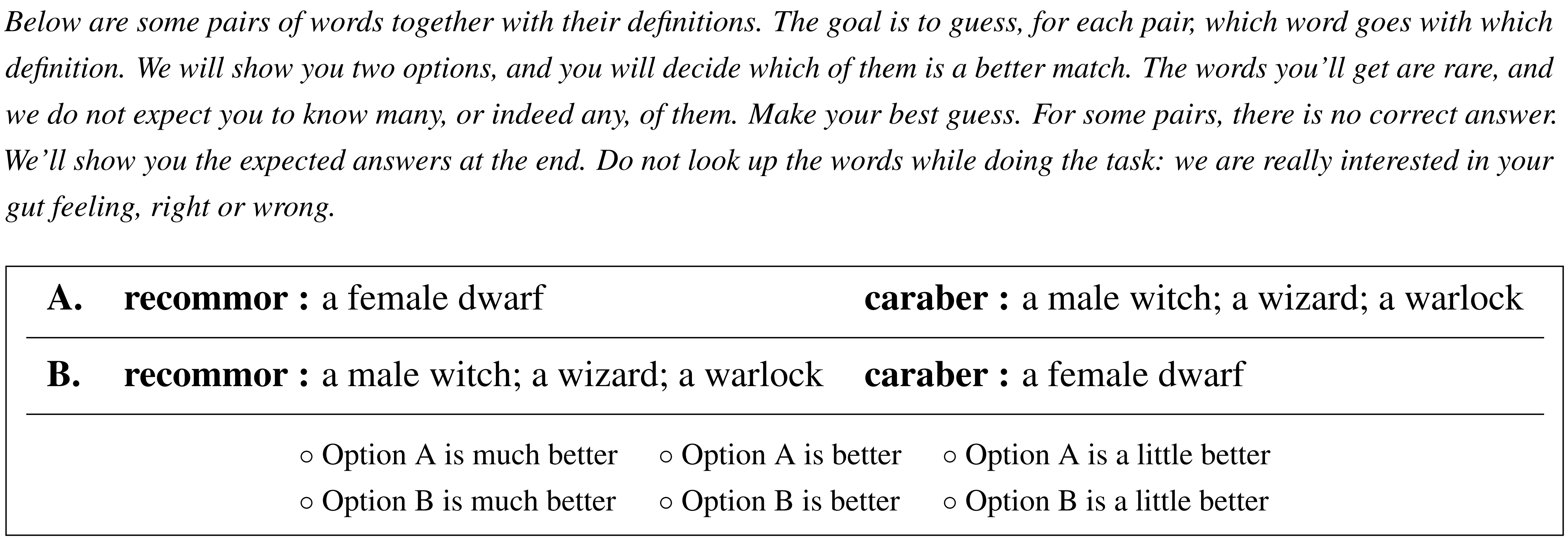
Fake by LSTM+GPT

29	**tagabism**	a tendency to be trapped in a view or a way of thinking
51	**transpanity**	excessive appetite for salt
36	**undergrism**	the controversial practice of participating in retrograde activities within a group
94	**wairl**	an Anglo-Saxon stanzaic poem which imitates the stress patterns of an earlier poem

Rare taken from a dictionary

93	**aeolipyle**	a steam engine powered by rocket propulsion due to escaping steam
100	**aroba**	a horse-drawn carriage once used for transportation
21	**boll**	the rounded seed-bearing capsule of a cotton or flax plant
92	**chott**	a dry salt lake that stays dry in the summer but receives some water in the winter
3	**cirrhopod**	any barnacle or similar crustacean

Experiment: matching word and definition



Below are some pairs of words together with their definitions. The goal is to guess, for each pair, which word goes with which definition. We will show you two options, and you will decide which of them is a better match. The words you'll get are rare, and we do not expect you to know many, or indeed any, of them. Make your best guess. For some pairs, there is no correct answer. We'll show you the expected answers at the end. Do not look up the words while doing the task: we are really interested in your gut feeling, right or wrong.

A.	**recommor :** a female dwarf	**caraber :** a male witch; a wizard; a warlock
B.	**recommor :** a male witch; a wizard; a warlock	**caraber :** a female dwarf

○ Option A is much better ○ Option A is better ○ Option A is a little better

○ Option B is much better ○ Option B is better ○ Option B is a little better

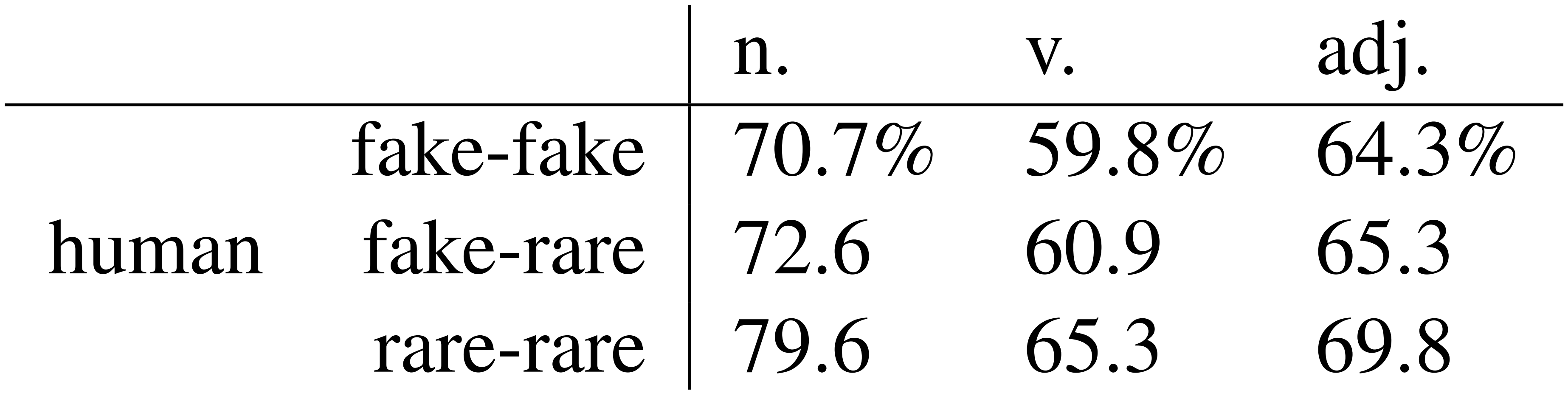
Generated by GPT-3

Given two words and definitions, guess which ones go together

Human performance

Both generated words and definitions (fake) and infrequent but genuine English words with definitions (rare) are used

The authors compare three settings using human test subjects:



		n.	v.	adj.
human	fake-fake	70.7%	59.8%	64.3%
	fake-rare	72.6	60.9	65.3
	rare-rare	79.6	65.3	69.8

Correlation across POS was high

Human performance

Test subjects would often prefer the same definitions regardless of type (fake-fake, fake-rare, rare-rare)

Users preferred the same assignment 61% of the time

LLD

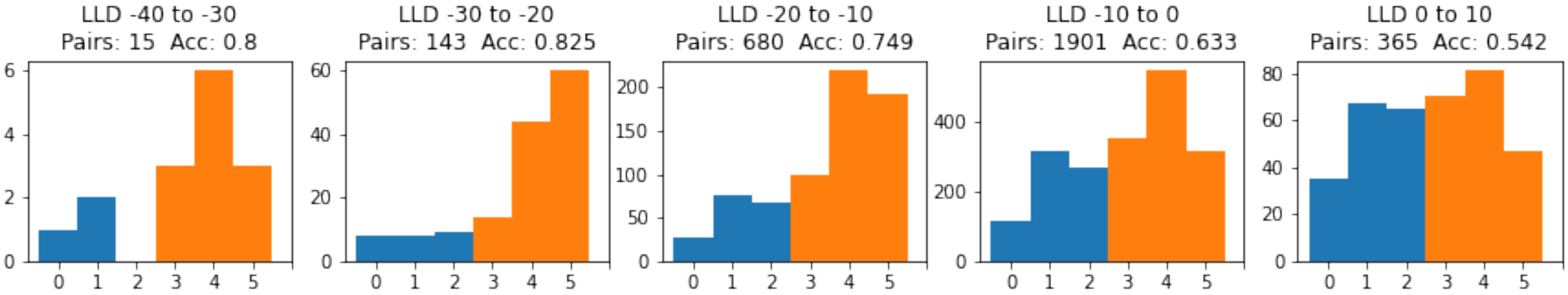
We can measure how much GPT-3 prefers given matches:

$$\mathrm{LLD}(w_1, w_2) = \log \frac{p(d_2|w_1)p(d_1|w_2)}{p(d_1|w_1)p(d_2|w_2)}.$$

LLD	word pair	
−33.3	**parademme :**	a person, party, or body that is not a participant in a dispute or controversy
	calcanism :	the study of extrusions of lava, volcanic rock, and ash
−23.2	**carburist :**	a person with an abnormally large capacity for consuming substances
	subacitide :	the longest possible decimal number, continued indefinitely by periodic repetition of 0
−13.2	**stucenium :**	a little roof, the soffit of a cornice, the median part of a pediment
	helliact :	a person with high level of knowledge and experience in a specific area
−9.4	**prosemer :**	a vessel with rough edges or projecting parts
	drobbler :	a person who enjoys listening to music
+0.2	**frequayer :**	a person who has given up hope for fun
	endosman :	a performer in a minstrel duo

LLD

Human confidence is inversely correlated with LLD:



0: Highly prefer swapped definitions

…

5: Highly prefer the definitions proposed by GPT-3

Human-generated vs. GPT-3 definitions

Human-generated neologism: **backmasking**

Human definition:
the instinctive tendency to see someone as you knew them in their youth

GPT-3 definition:
the act of disguising messages within recordings via sound effects

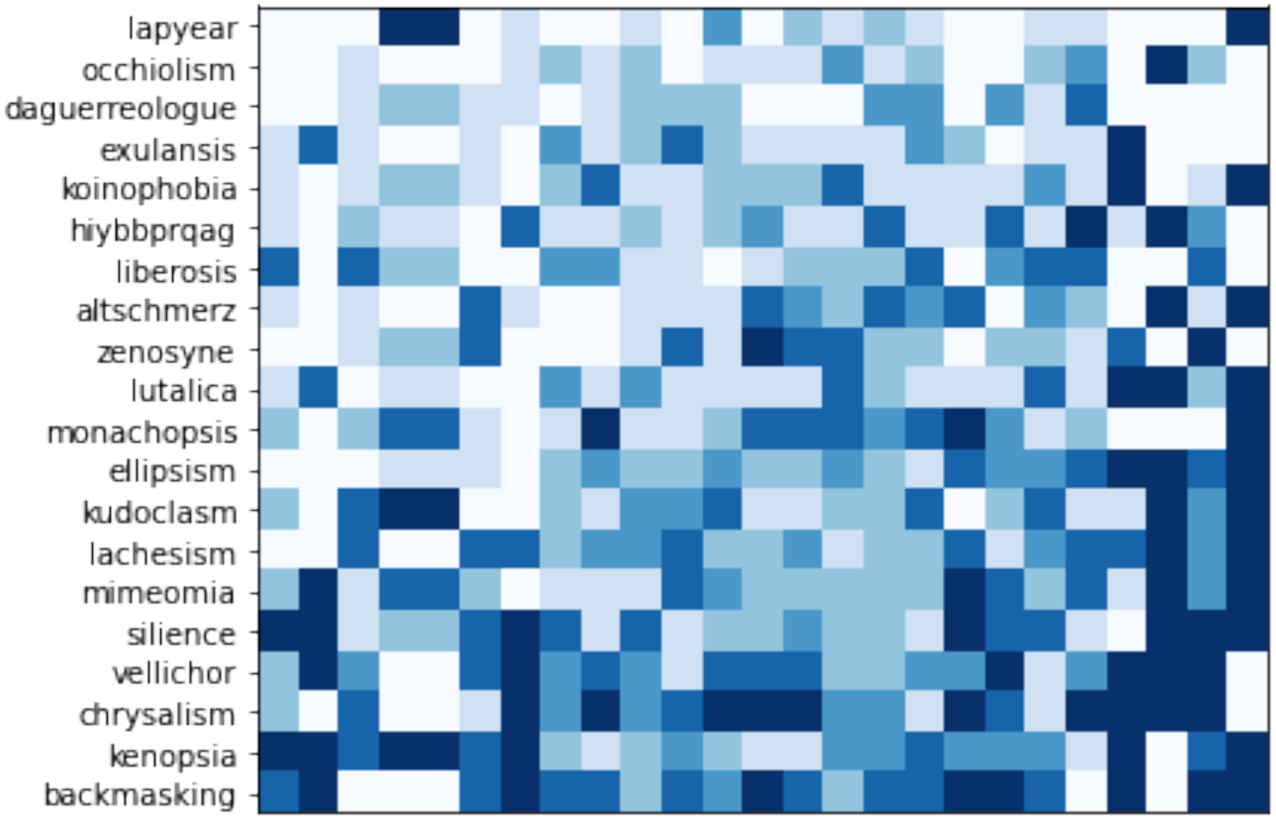
Participants were asked to rank the definitions:
0 (strongly prefer human) … 5 (strongly prefer GPT-3)

Human-generated vs. GPT-3 definitions

The GPT-3 definition was preferred 40% of the time

For certain examples, the human definition is preferred by most participants.

For other examples, most prefer GPT-3



How do you interpret this?

These human-coined neologisms have a bias towards meanings with an existential slant, which results in additional structure in our results, reflecting the population structure of the subjects.

Conclusions

GPT-3 does not utilize morphology exclusively: Many neologisms have no clear roots or derivational origin

Phonological (or orthographic) clues seem to play a role

GPT-3 seems to lean nuances of etymology and the correspondences of sound and meaning that lie at the very base of language understanding