What is Language?

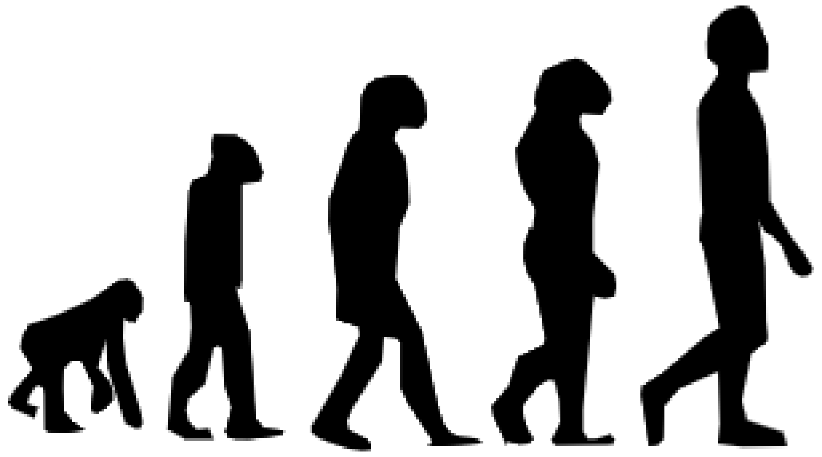
Properties of Language

A language is a system of communication used by a particular country, region, or community,
with a set of rules for grammar,
vocabulary, and pronunciation.

ChatGPT

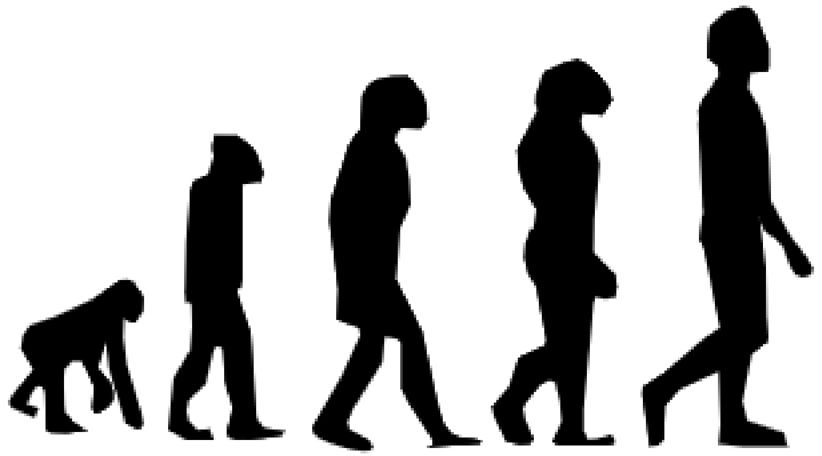
Language is an evolving system.

Steels

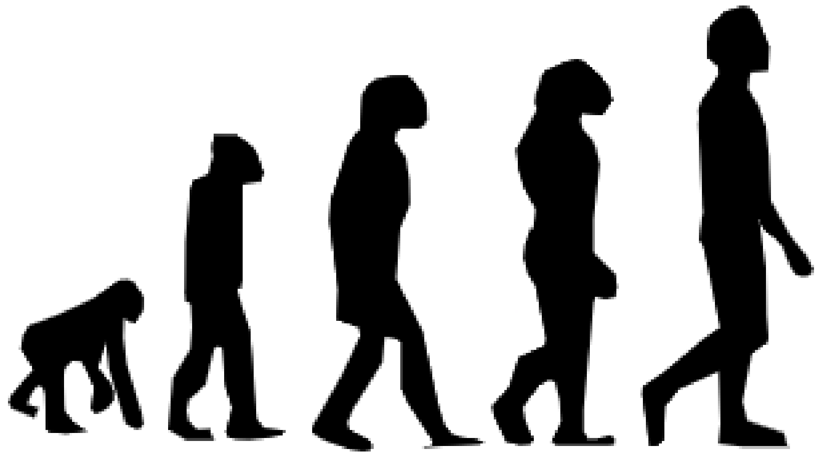


Language has been found in every society ever studied by anthropologists.

Pinker



Language is what distinguishes humans from other species on Earth.

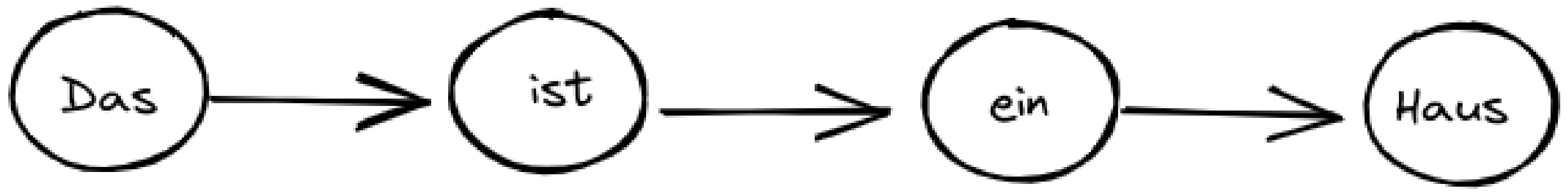


Language encodes information and acts as a medium for exchanging information.

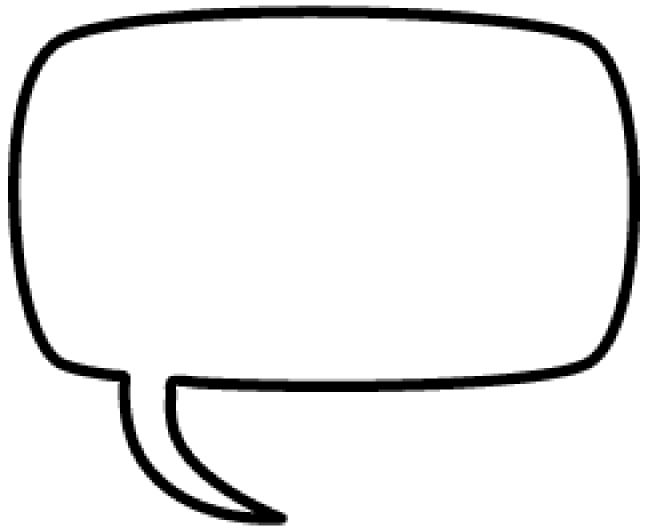
Pinker

10101010101010101010

Language is sequential in its nature.



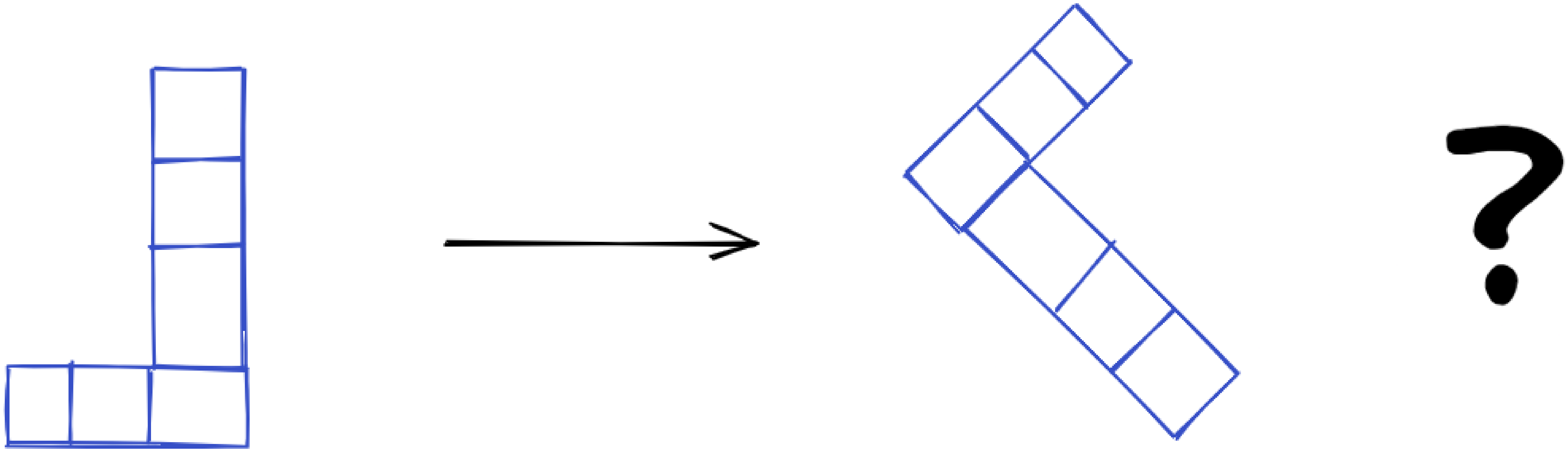
Language refers to spoken language NOT written language.



Man has an instinctive tendency to speak as we see in the babble
of our young children while no child has an instinctive tendency
to bake, brew or write.

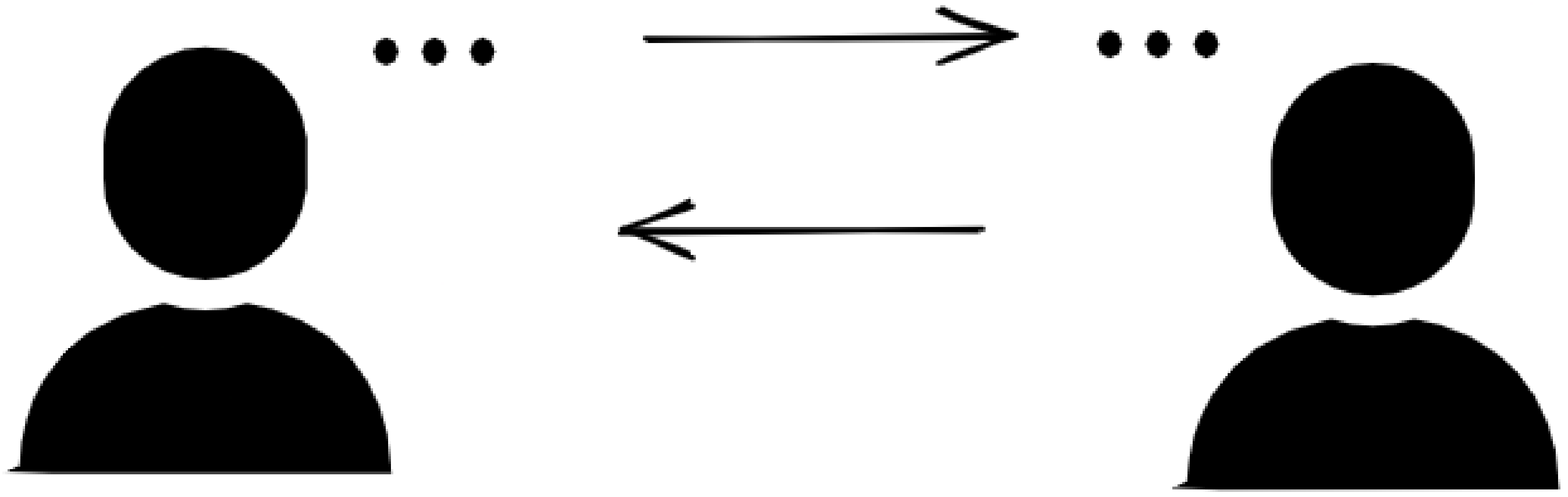
Darwin

Language is NOT thought!

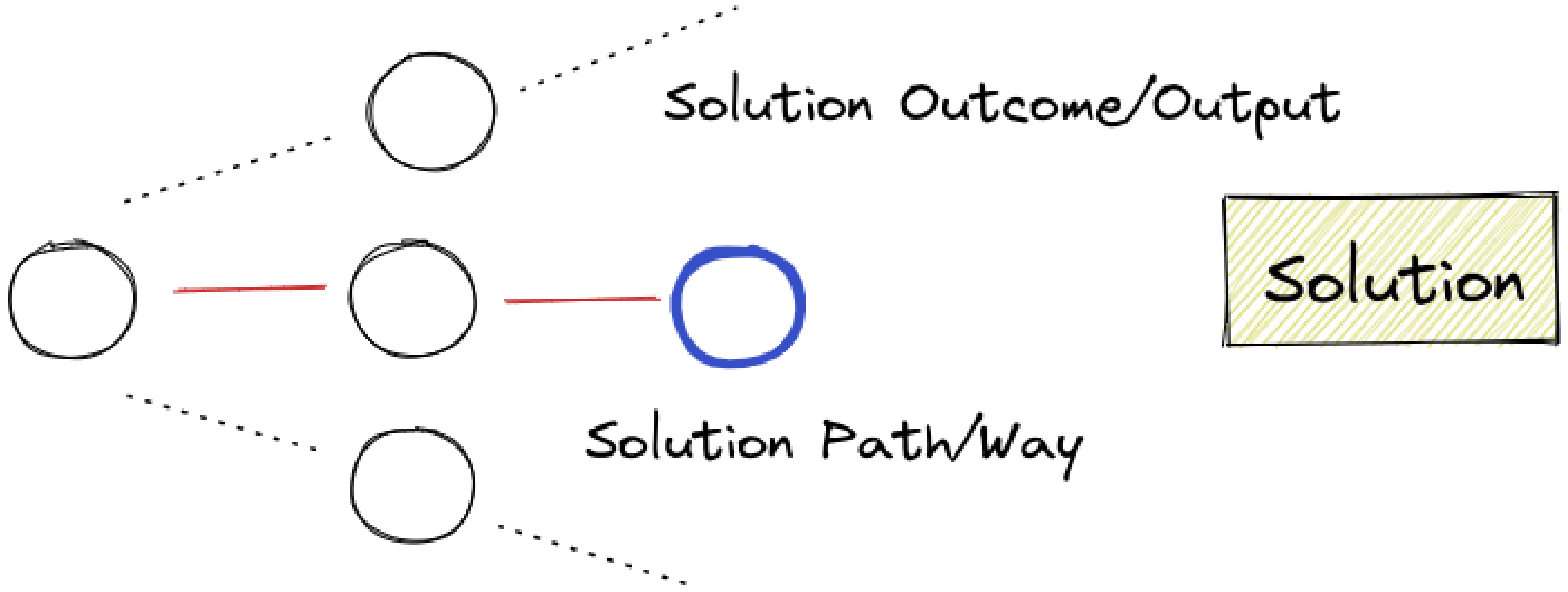


Functionalities of Language

Transfer of Information



Reasoning

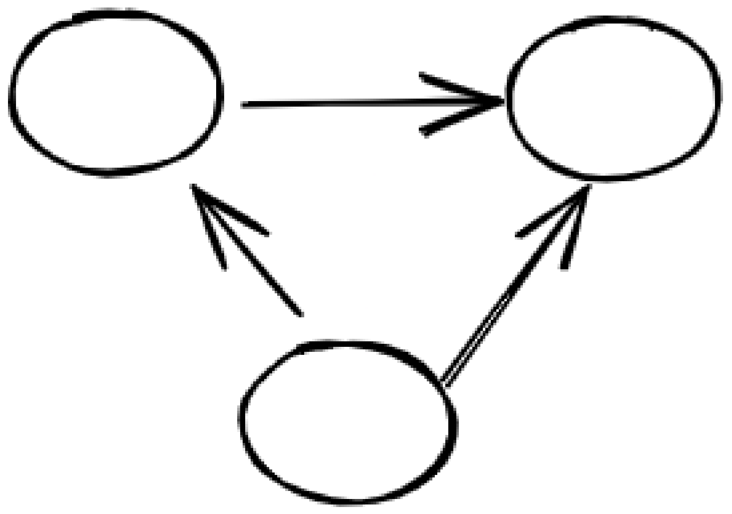


Solution Outcome/Output

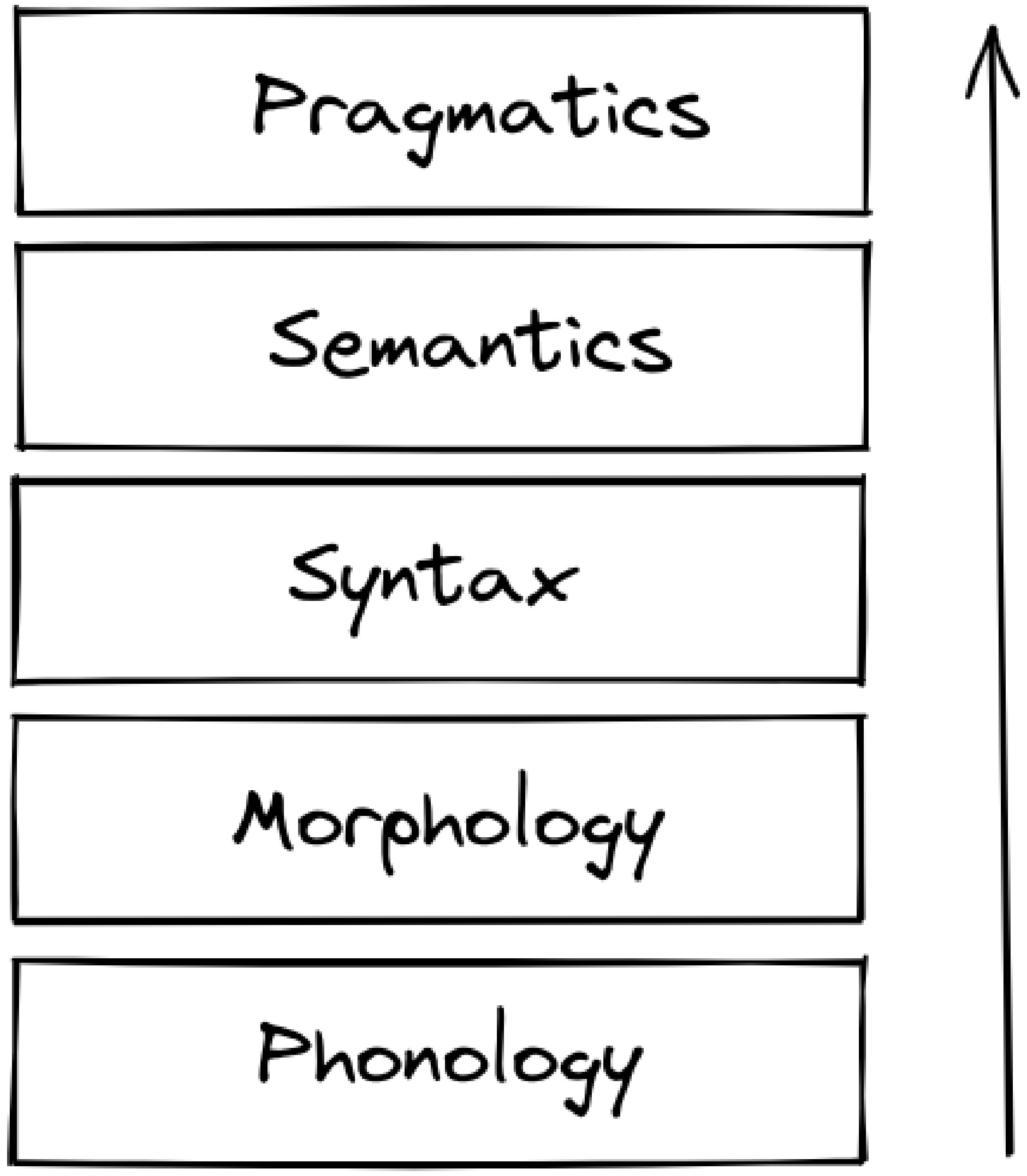
Solution

Solution Path/Way

Memory



The Language Technology Stack



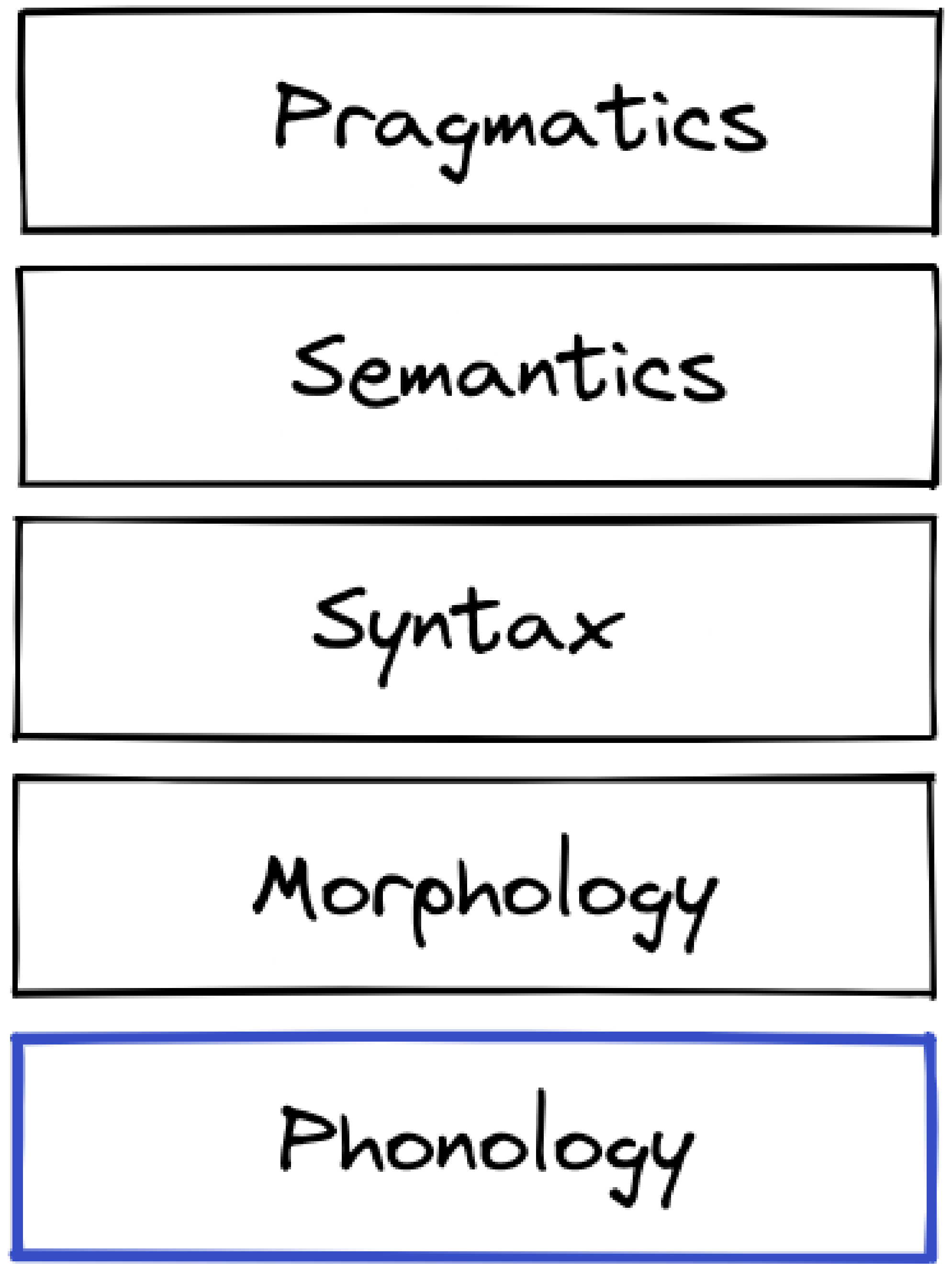
Pragmatics

Semantics

Syntax

Morphology

Phonology



Pragmatics

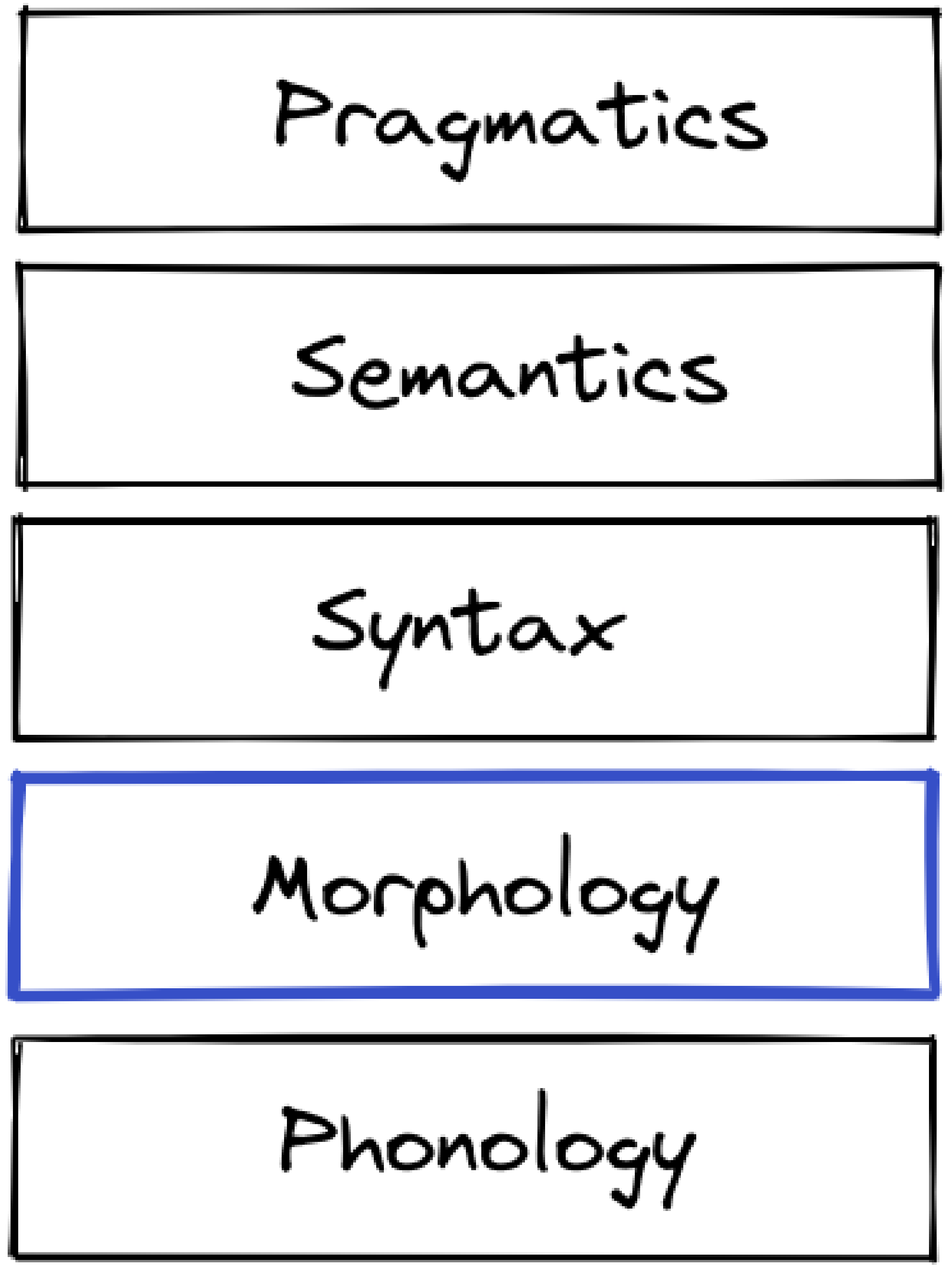
Semantics

Syntax

Morphology

Phonology

a	d[a]s
a:	j[a]hr
ɛ	w[e]nn
ə	ein[e]
ɐ	od[er]
ɛ:	k[ä]se
e:	g[e]gen
ɪ	w[i]rd
i:	d[ie]
ɔ	d[o]ch
o:	w[o]
œ	k[ö]nnen
ø:	l[ös]en
ʊ	m[u]ss
u:	g[u]t
ʏ	m[ü]cke
y:	f[ü]r


Pragmatics

Semantics

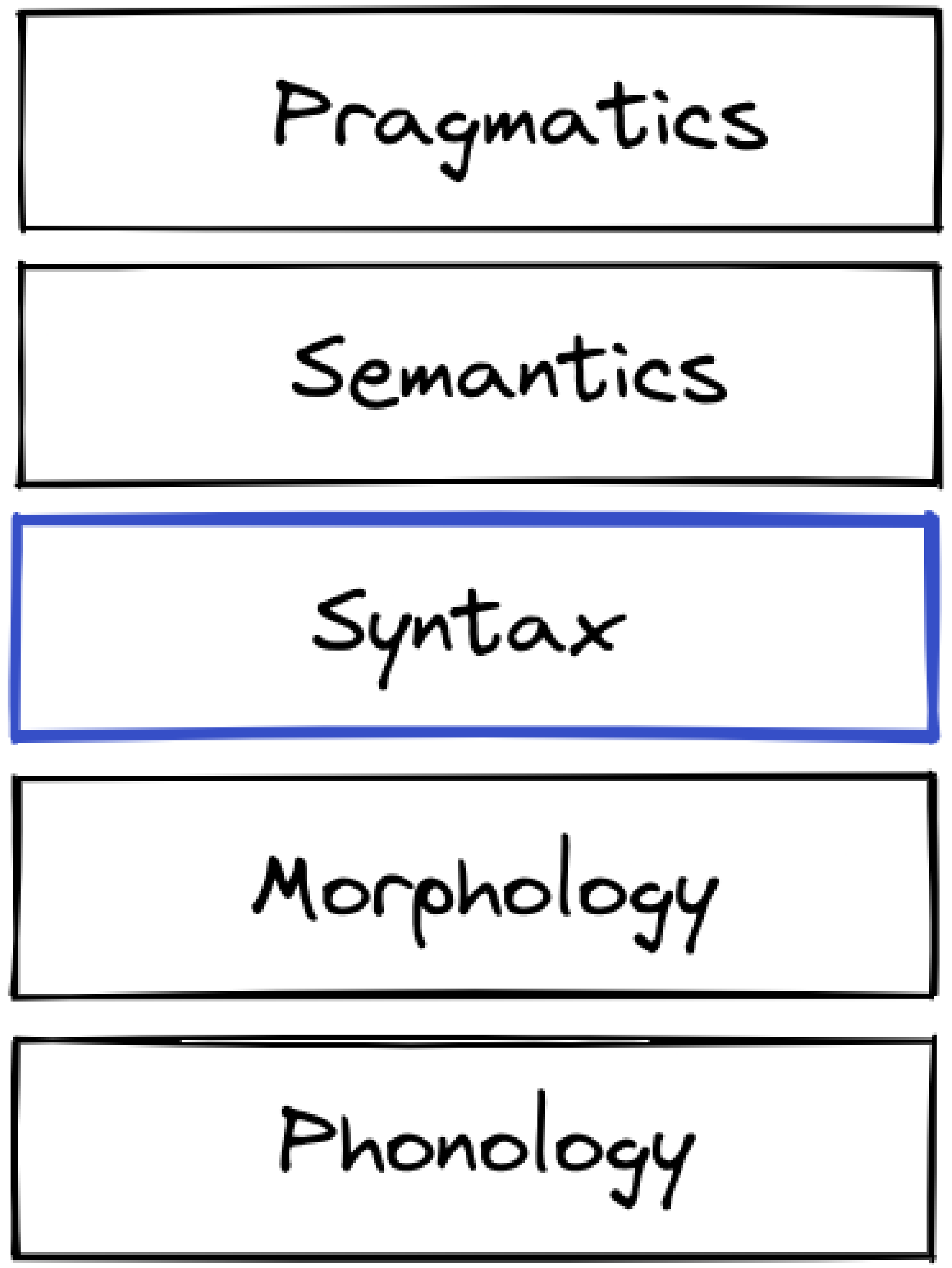
Syntax

Morphology

Phonology

prefix		suffix	
ver-	fɛɐ	-ung	ʊŋ

Veranstaltung fɛɐˈʃtaːltʊŋ



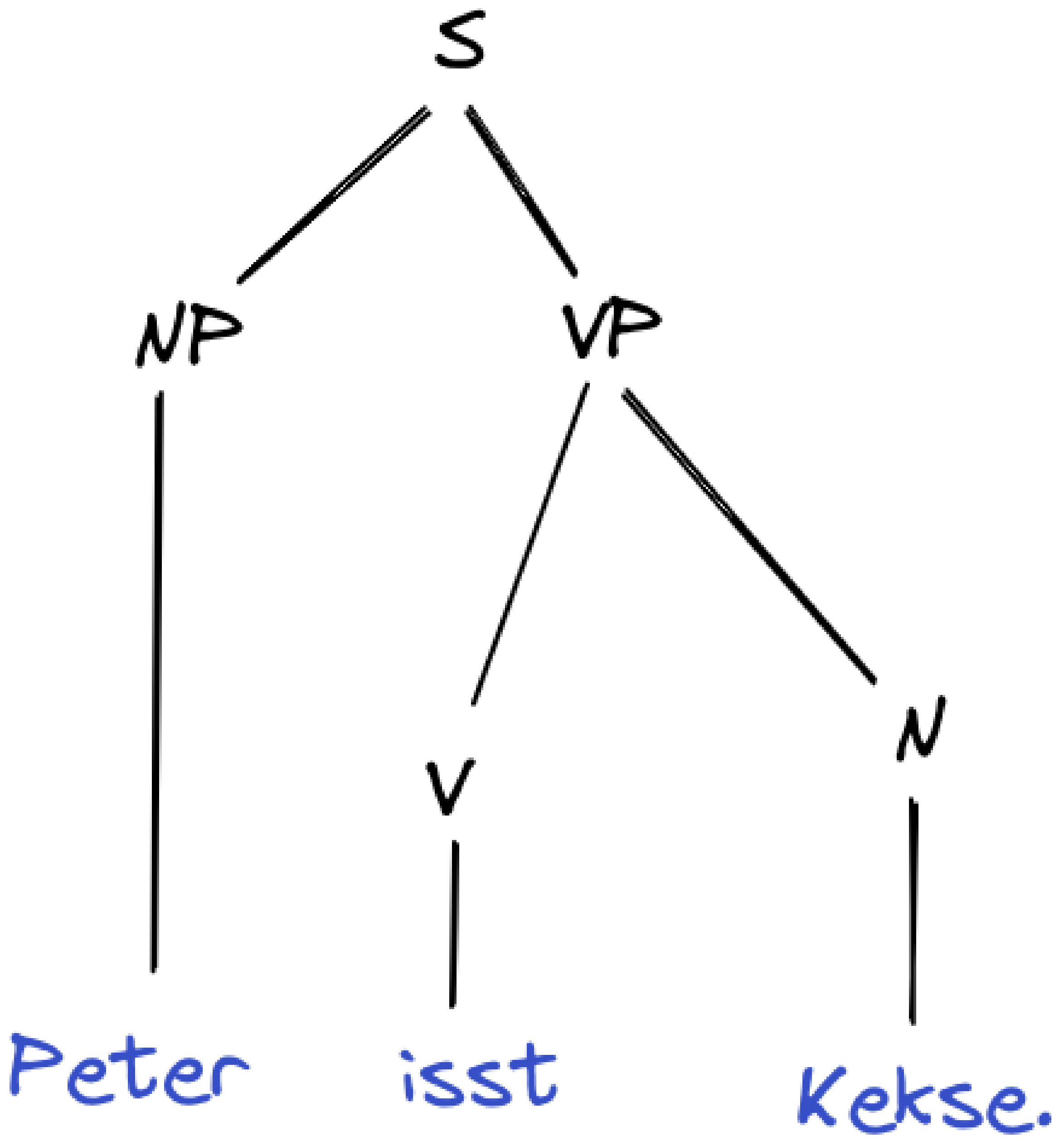
Pragmatics

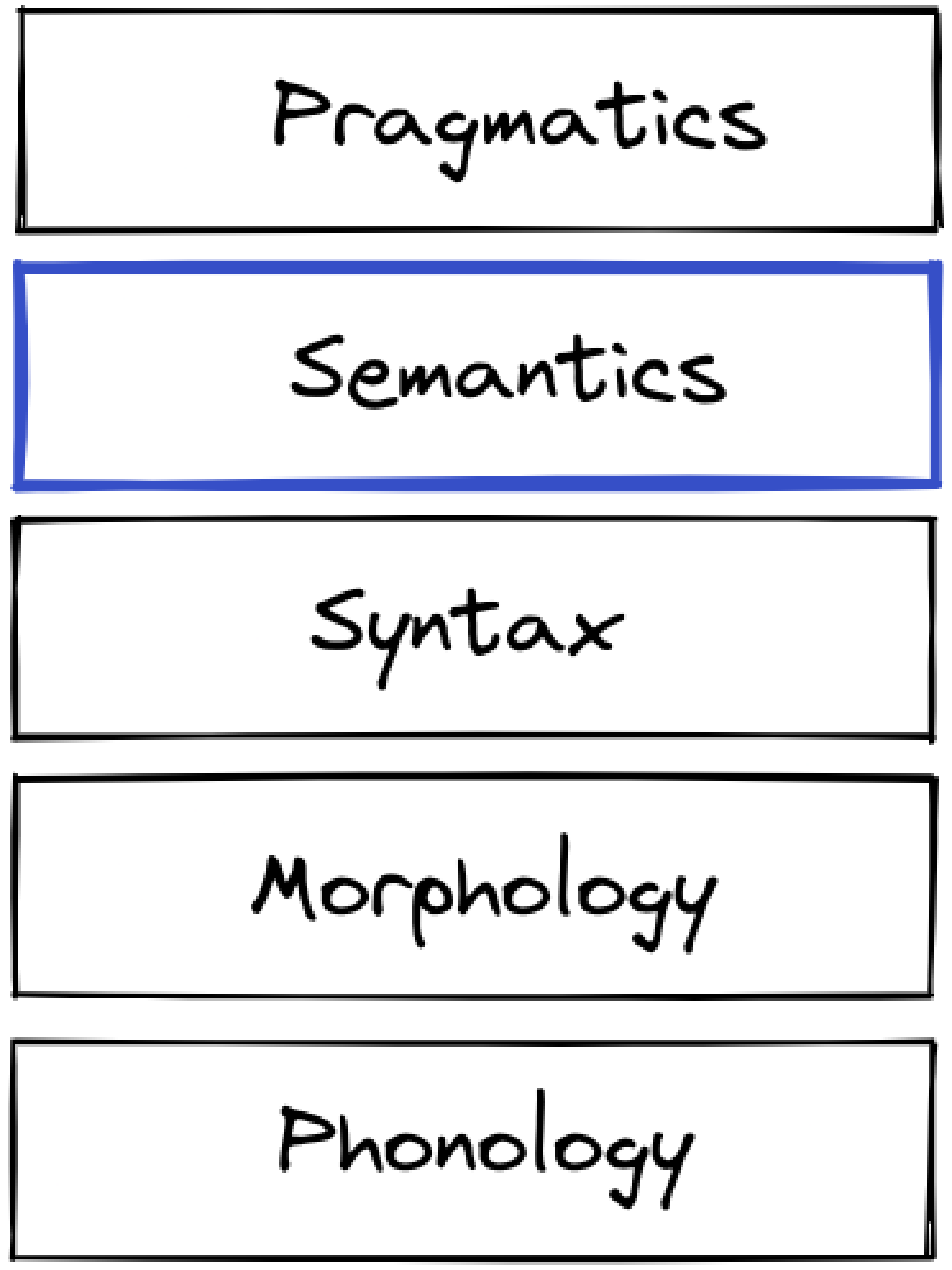
Semantics

Syntax

Morphology

Phonology





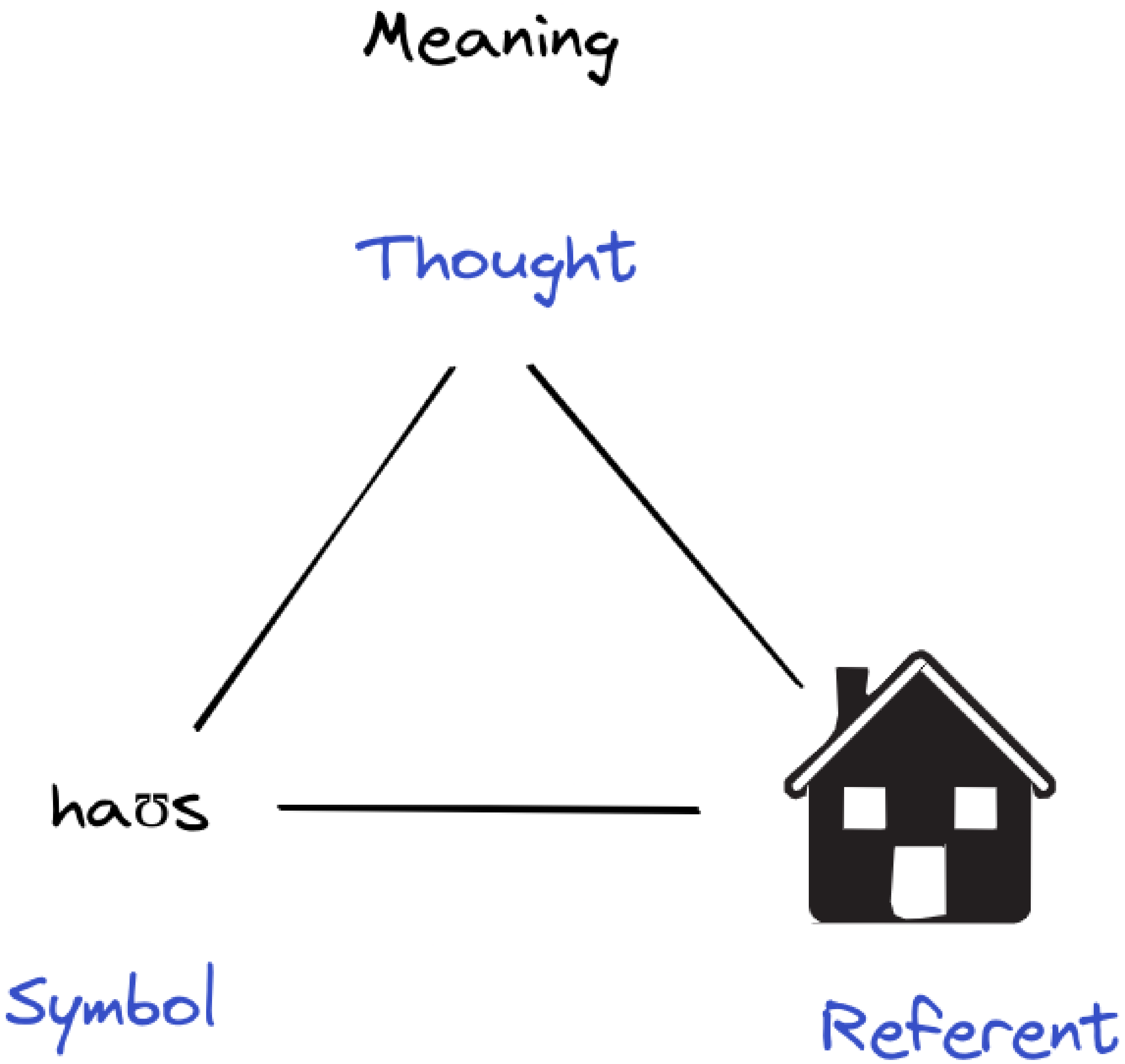
Pragmatics

Semantics

Syntax

Morphology

Phonology



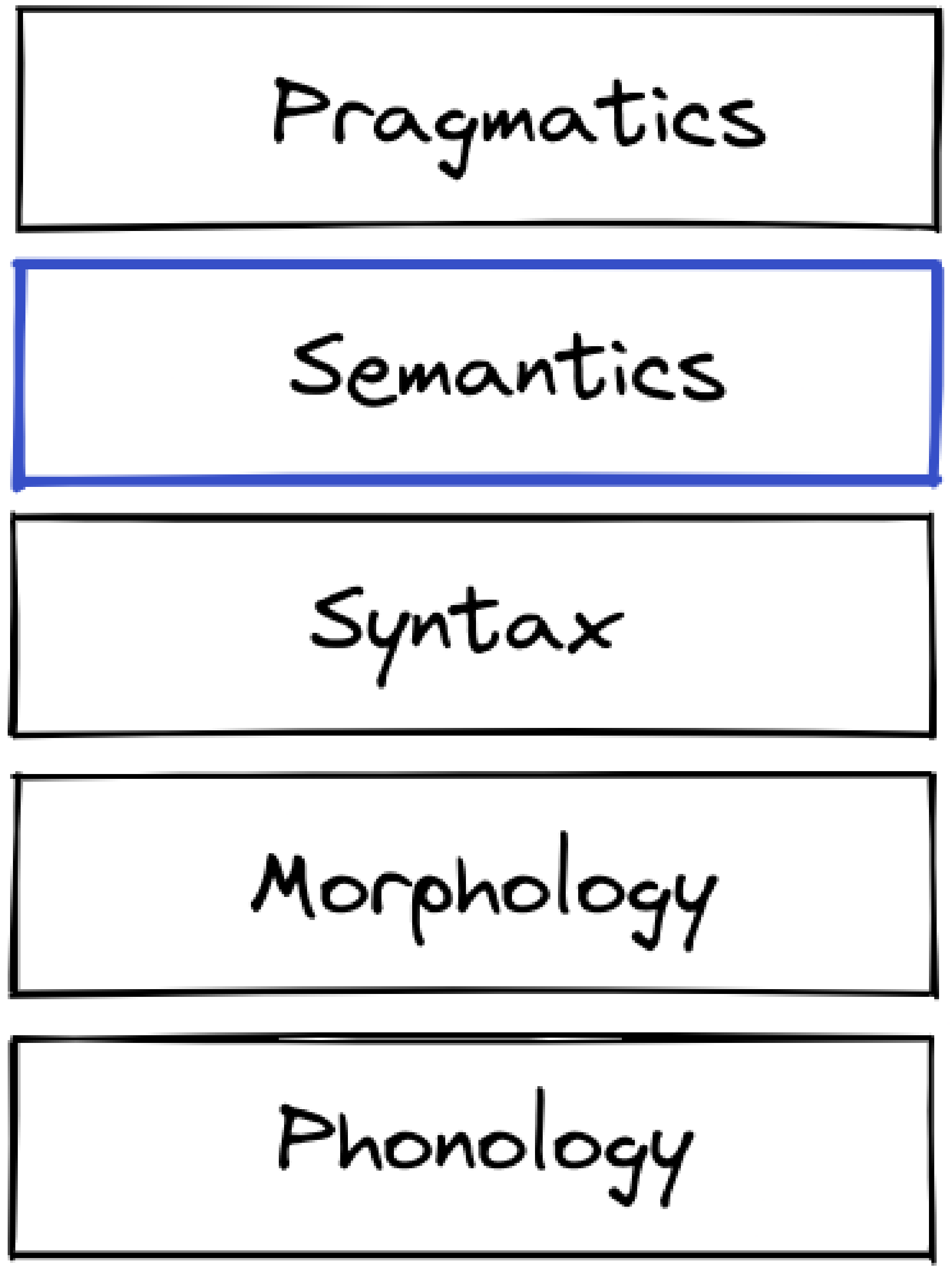
Meaning

Thought

haʊs

Symbol

Referent



Pragmatics

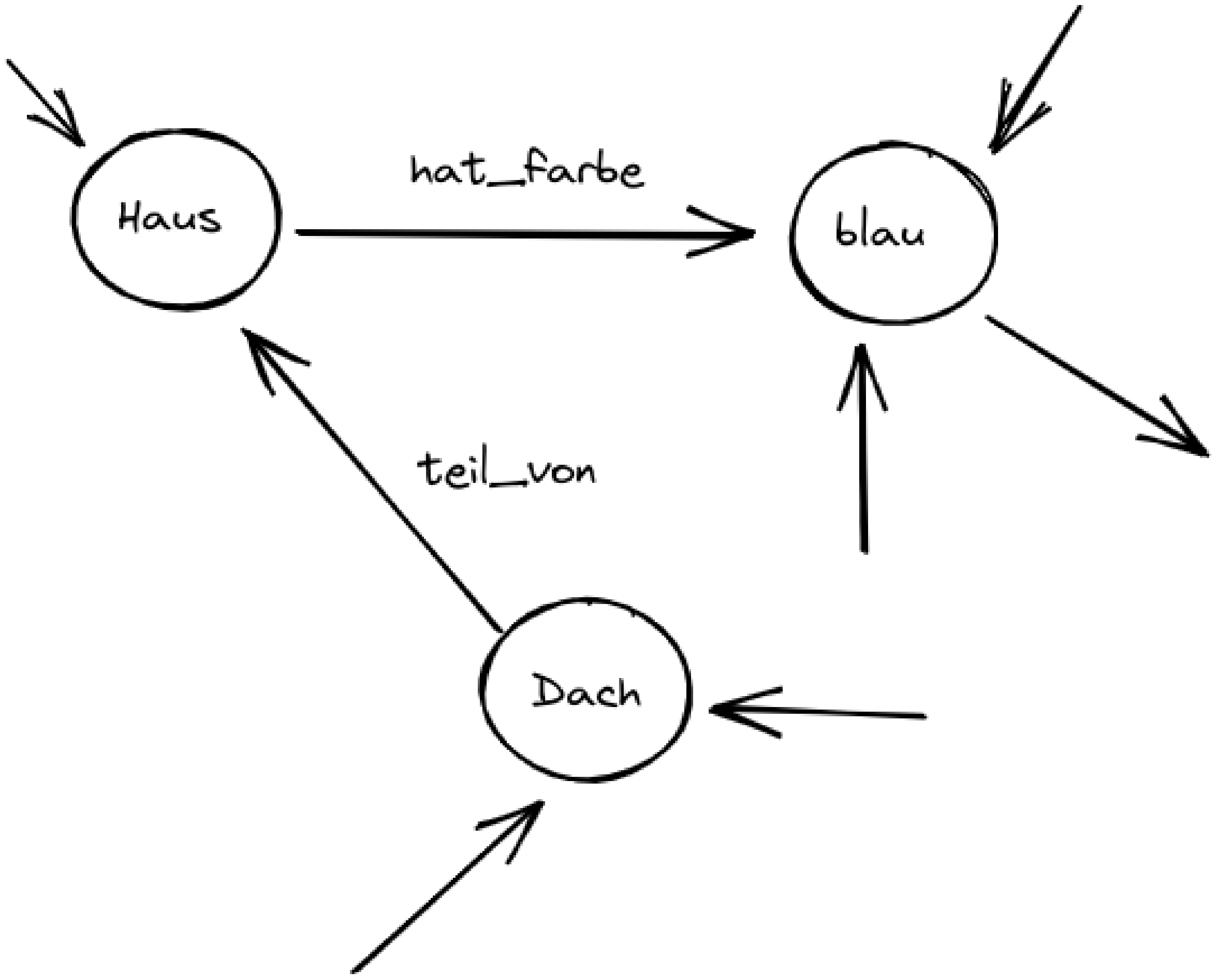
Semantics

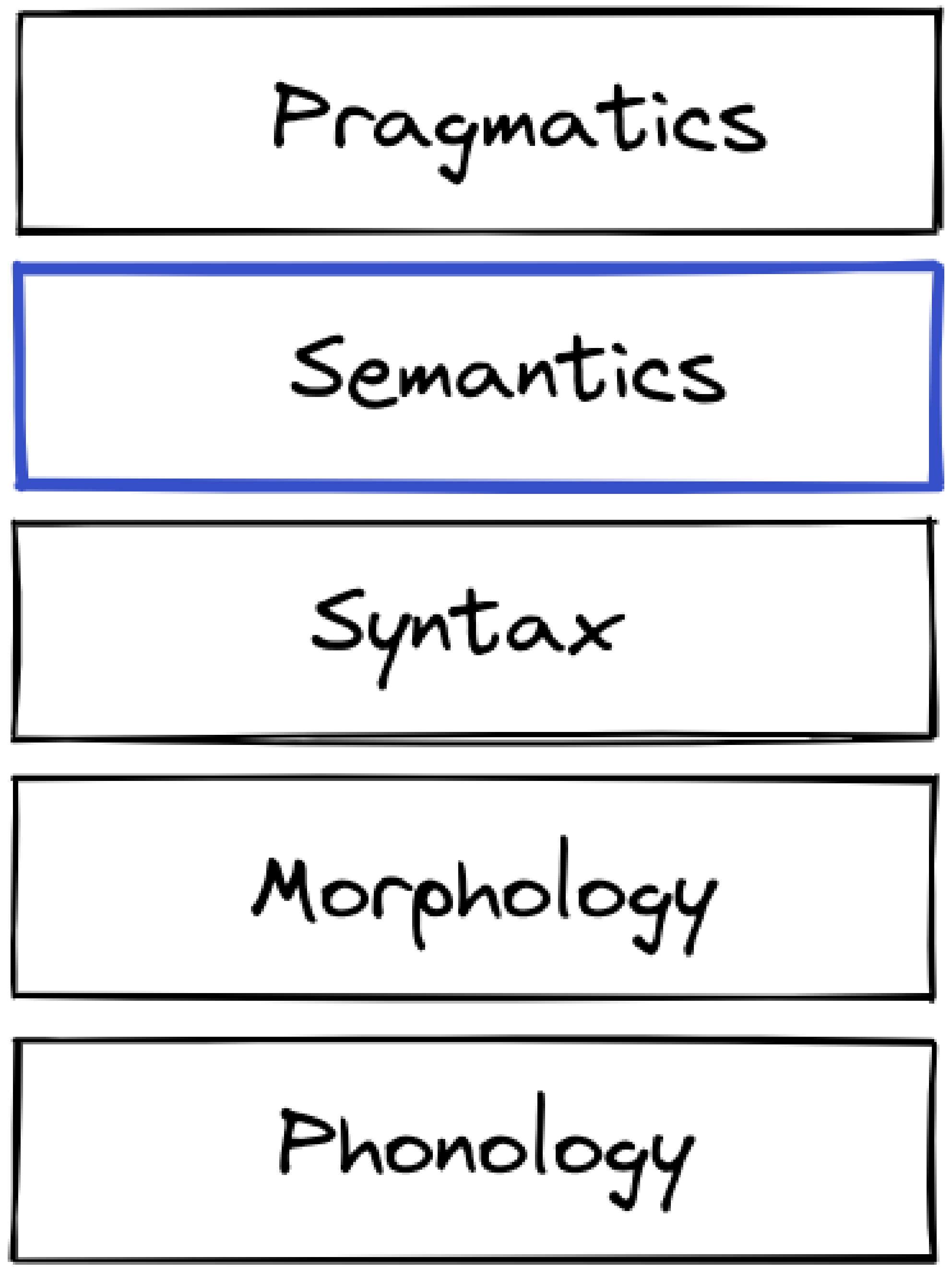
Syntax

Morphology

Phonology

Knowledge





Pragmatics

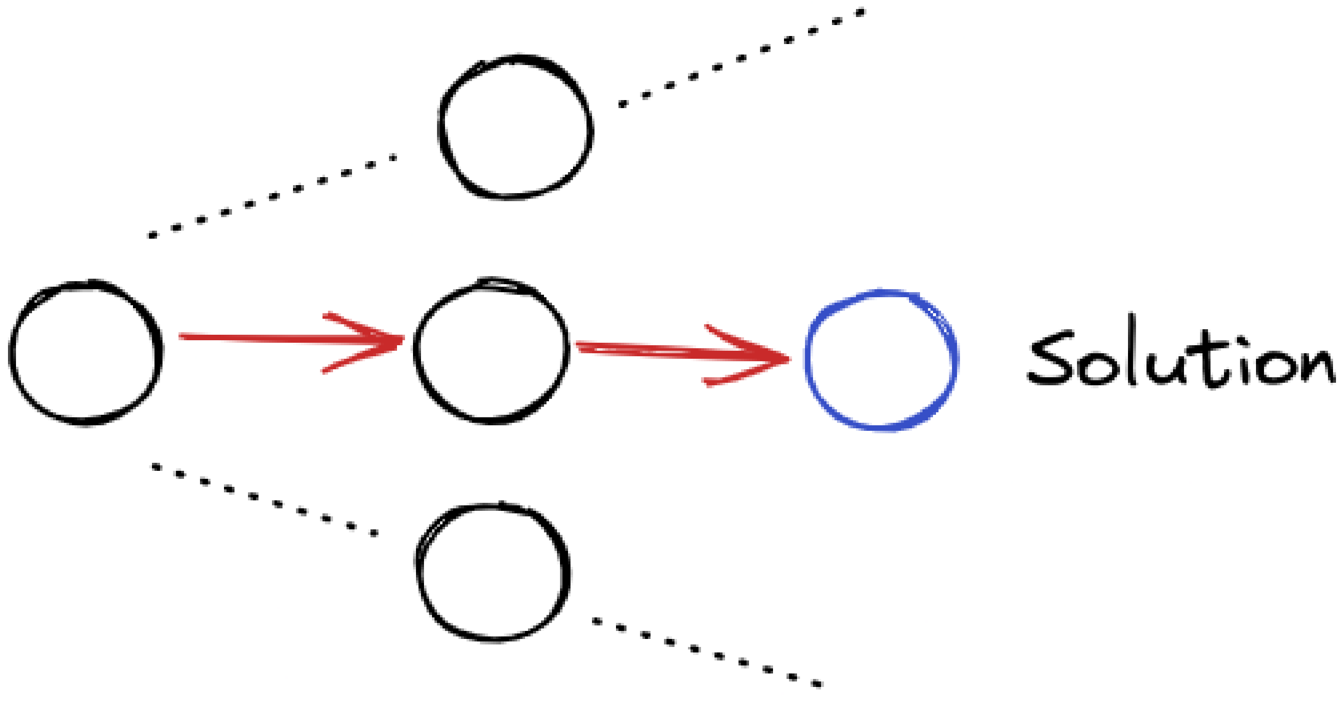
Semantics

Syntax

Morphology

Phonology

Reasoning

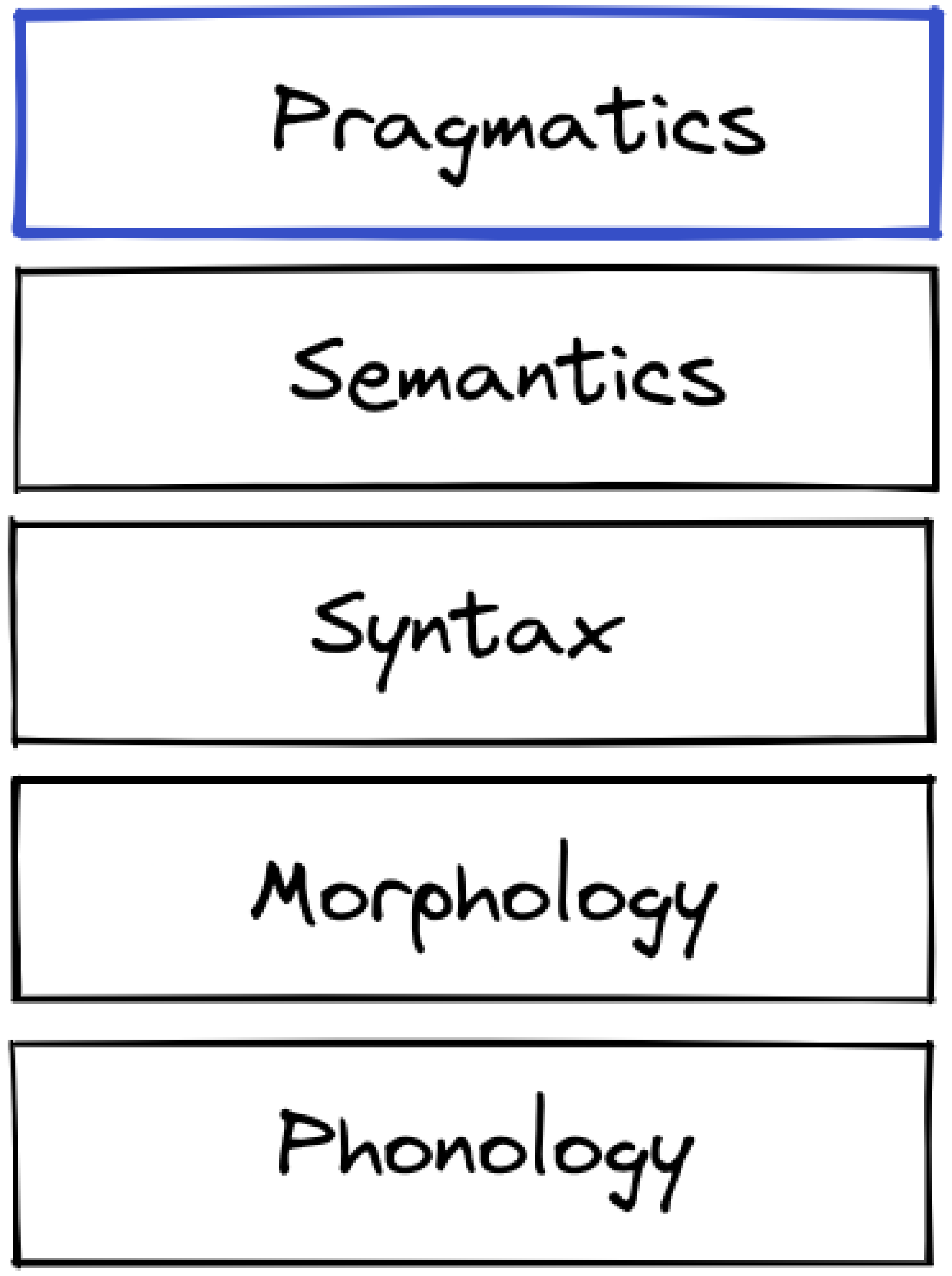


Solution

Given:
a > b
and
b > c

Then:
a > b > c

Therefore:
a > c



Pragmatics

Semantics

Syntax

Morphology

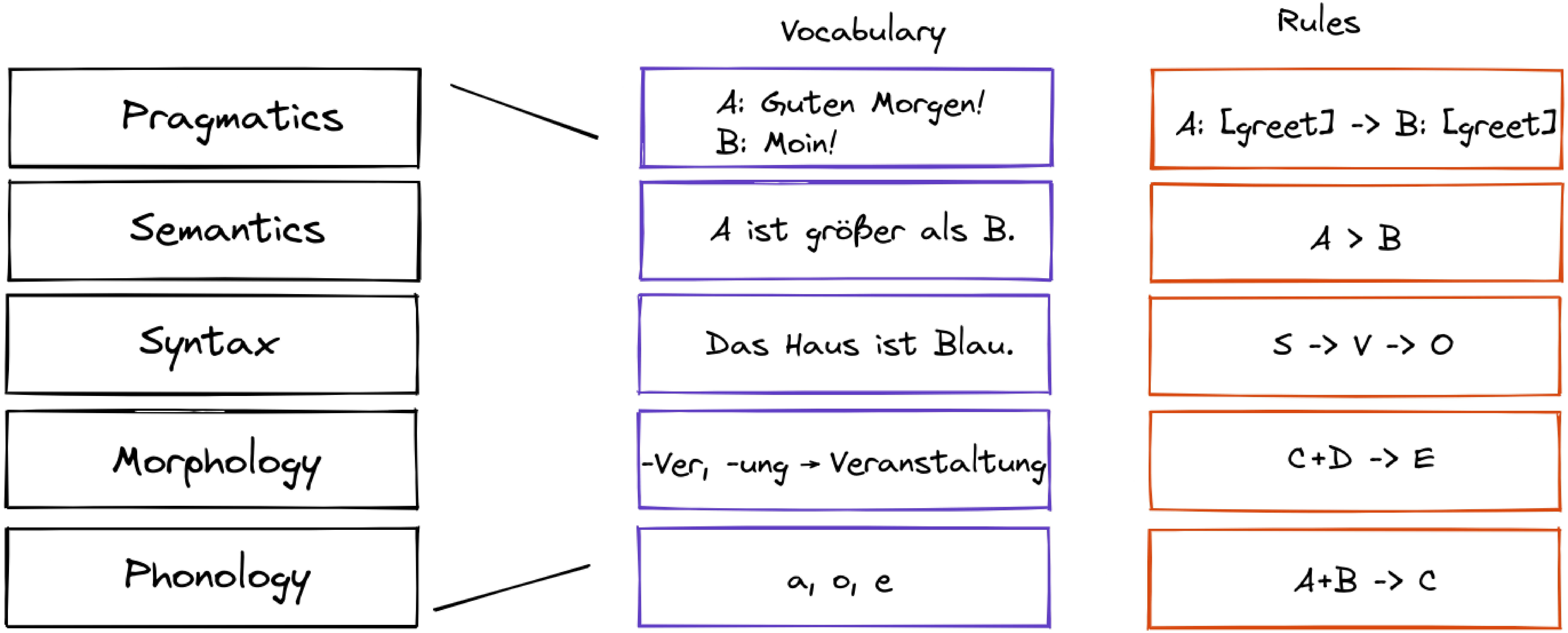
Phonology

...

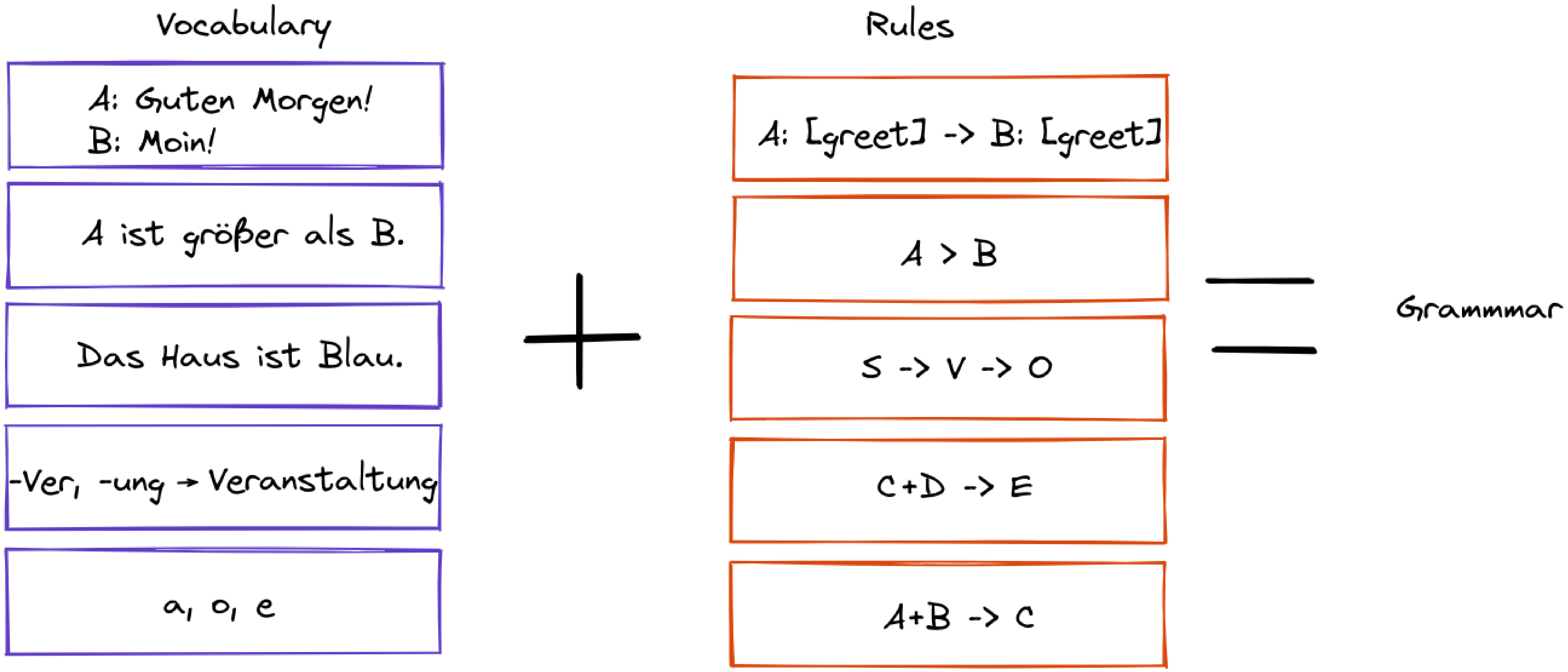
A: Welches Buch möchtest lesen?

B: Dieses!

How can we formally describe this system?



	Vocabulary	Rules
Pragmatics	A: Guten Morgen! B: Moin!	A: [greet] -> B: [greet]
Semantics	A ist größer als B.	A > B
Syntax	Das Haus ist Blau.	S -> V -> O
Morphology	-Ver, -ung → Veranstaltung	C+D -> E
Phonology	a, o, e	A+B -> C



Vocabulary

A: Guten Morgen!
B: Moin!

A ist größer als B.

Das Haus ist Blau.

-Ver, -ung → Veranstaltung

a, o, e

+

Rules

A: [greet] -> B: [greet]

A > B

S -> V -> O

C+D -> E

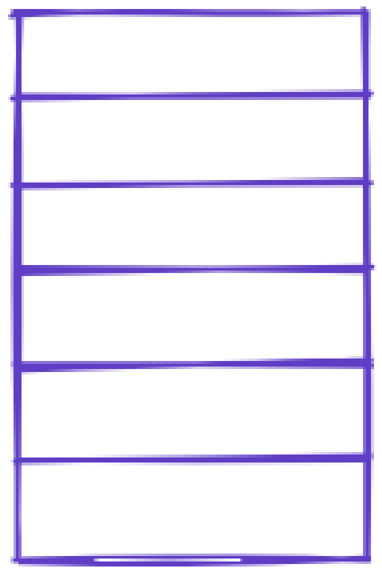
A+B -> C

=

Grammar

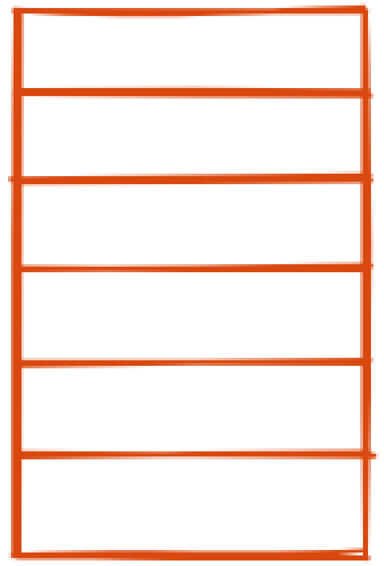
Vocabulary

V = {a, ..., z}

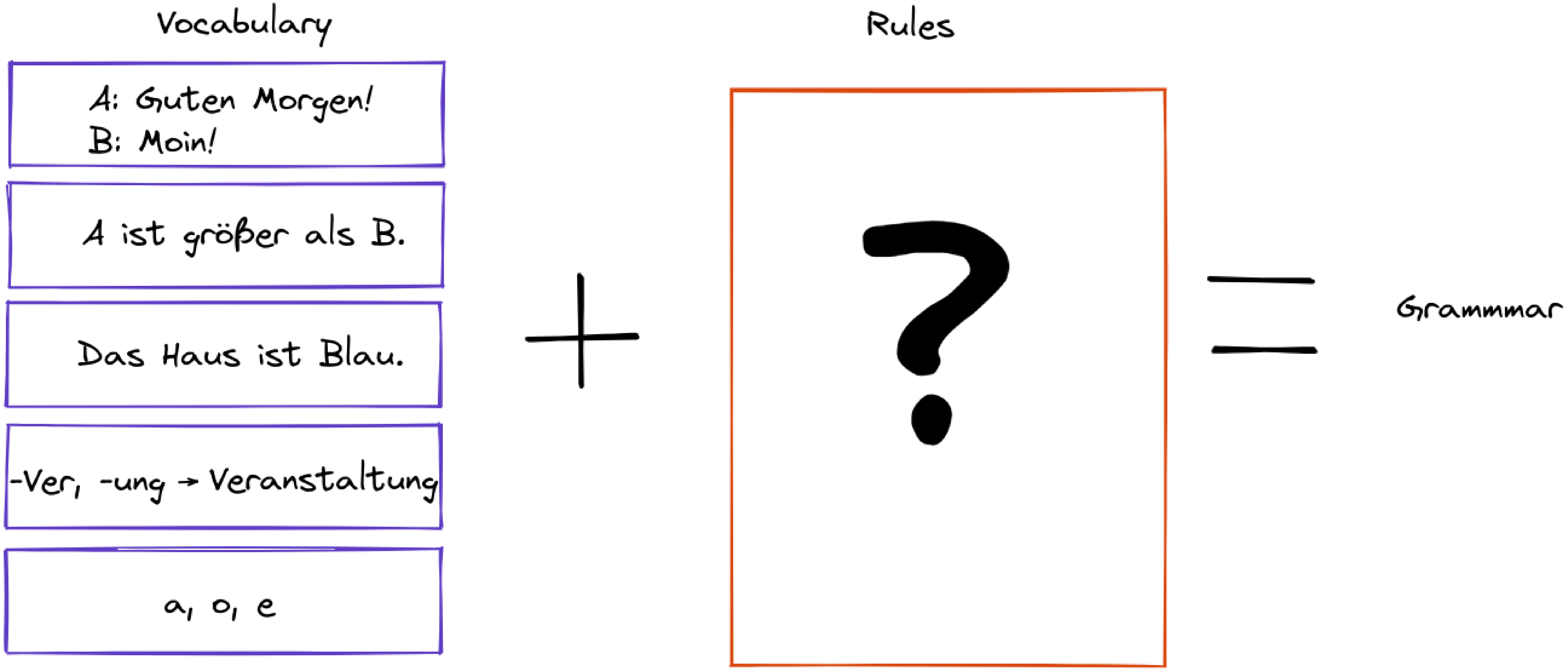


Rules

R = {A-> B+C, ..., C+D->E}



What is Language Modeling?



Vocabulary

A: Guten Morgen!
B: Moin!

A ist größer als B.

Das Haus ist Blau.

-Ver, -ung → Veranstaltung

a, o, e

+

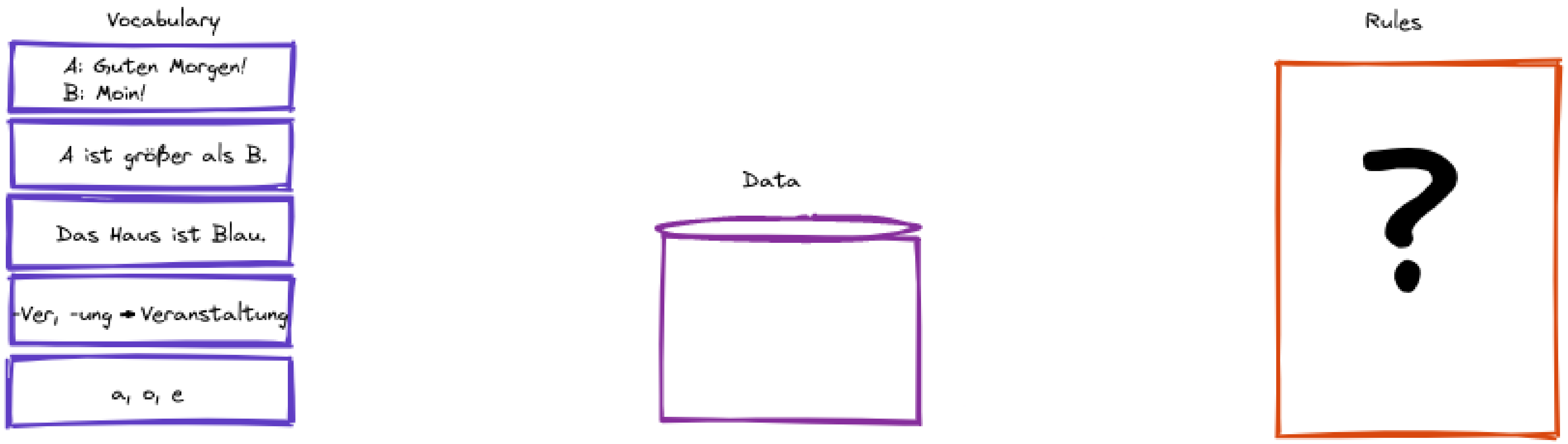
Rules

?

=

Grammar

Given a vocabulary and data but NO rules, how could we generate an expression in a given language?



Vocabulary

A: Guten Morgen!
B: Moin!

A ist größer als B.

Das Haus ist Blau.

-Ver, -ung → Veranstaltung

a, o, e

Data

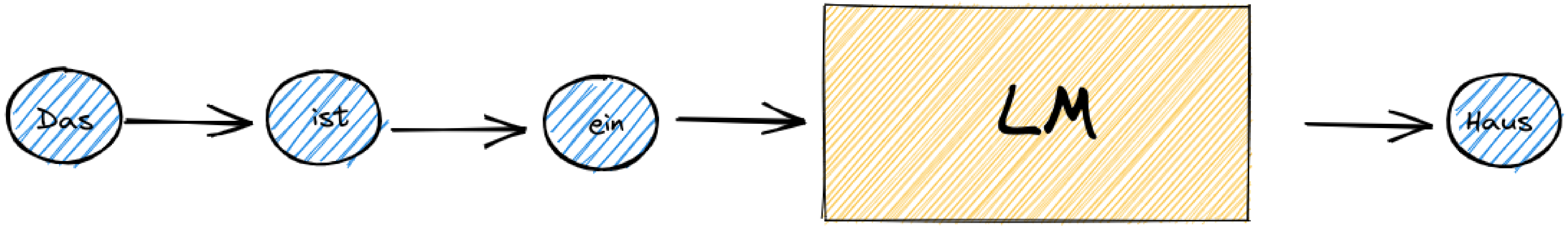
Rules

?

We start trying to learn the simplest possible rule:

predicting the next token/word/phoneme/symbol in a sequence!

GOAL: The goal of language modeling is to predict the next symbol/word/token in a sequence.



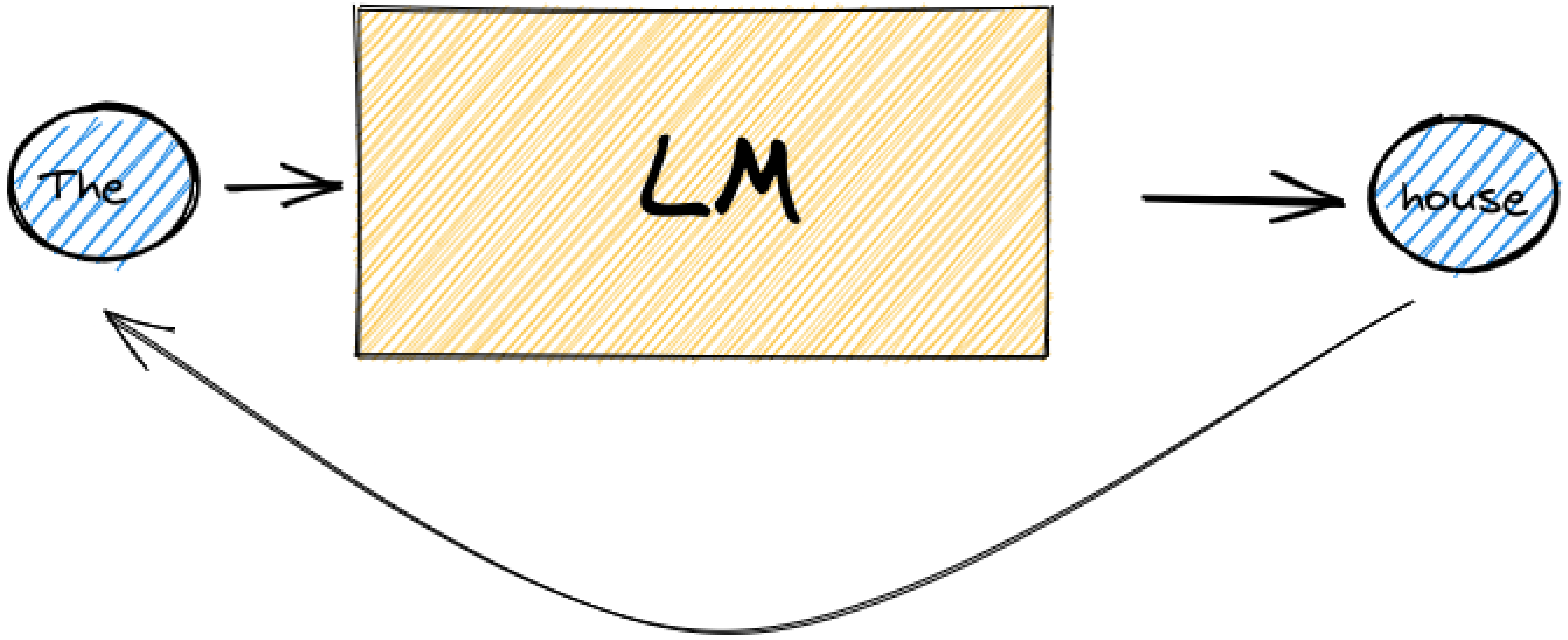
Language Model

A language model is a probability distribution over a sequence. It tells us how likely a given sequence is in a given language.

We can assign a probability to a sequence that tells us how often do we expect to see that sequence in the given language we are modeling.

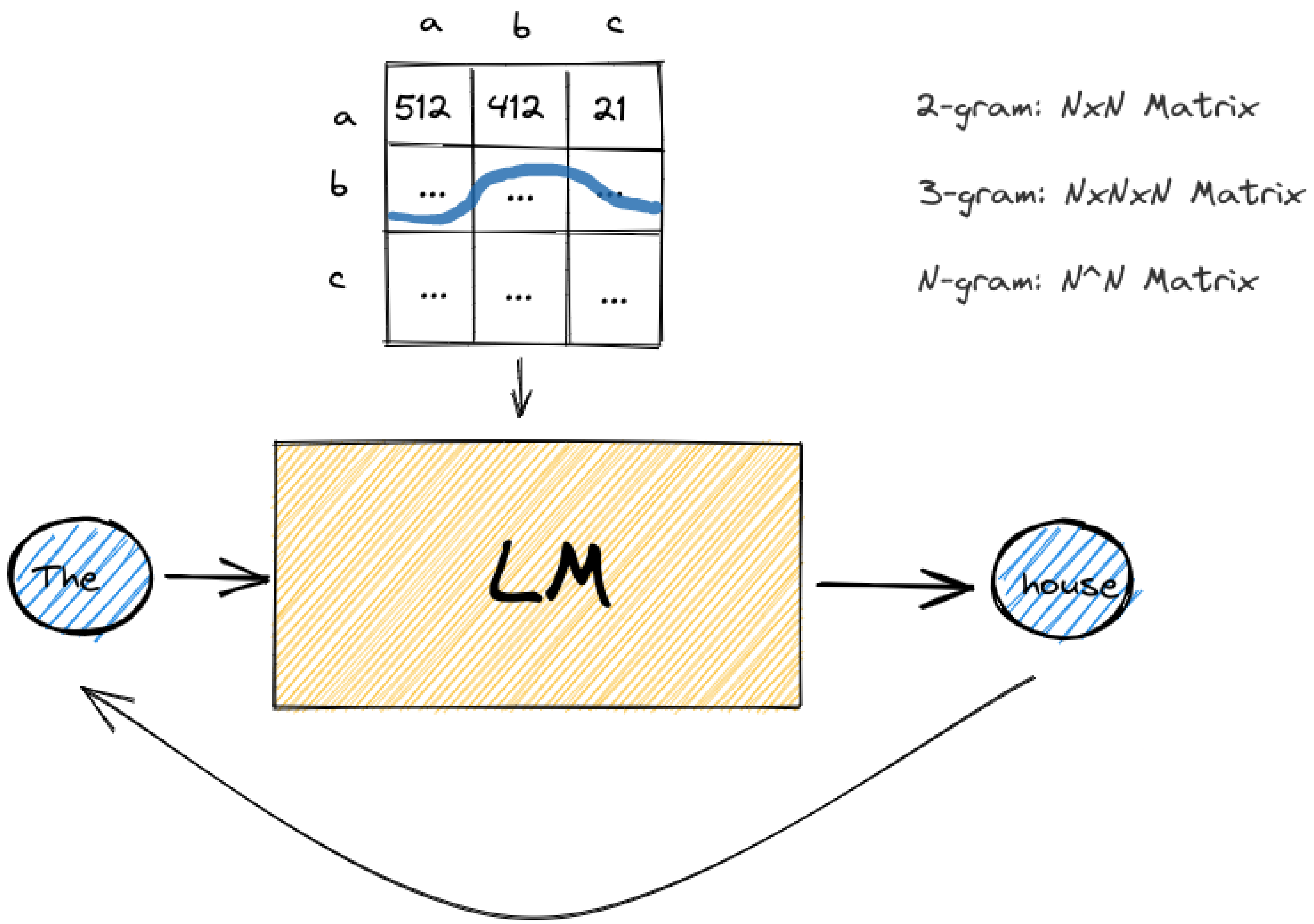
Why is a language model useful?

Once we have a good language model that approximates the 'true' distribution of a language very good, we can just sample from that distribution to generate very 'likely'/well sounding text of that language.



Naive 1st idea:

For every word, count its successor and then sample from that distribution.
= N-gram.



	a	b	c
a	512	412	21
b	...	...	...
c	...	...	...

2-gram: NxN Matrix

3-gram: NxNxN Matrix

N-gram: N^N Matrix

How do we generate a sentence?

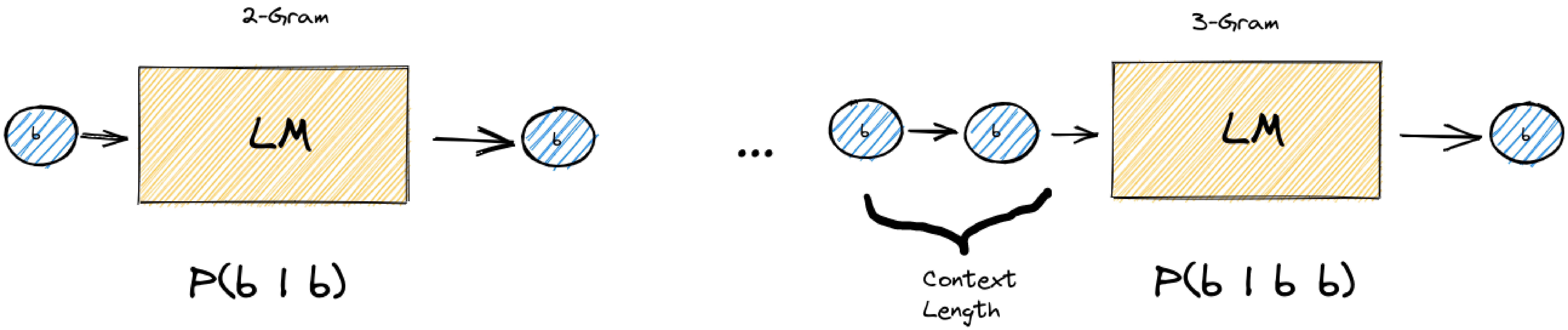
We just sample from the distribution of the table row.

	a	b	c
a	512	412	21
b	...	...	...
c	...	...	...

2-gram: NxN Matrix

3-gram: NxNxN Matrix

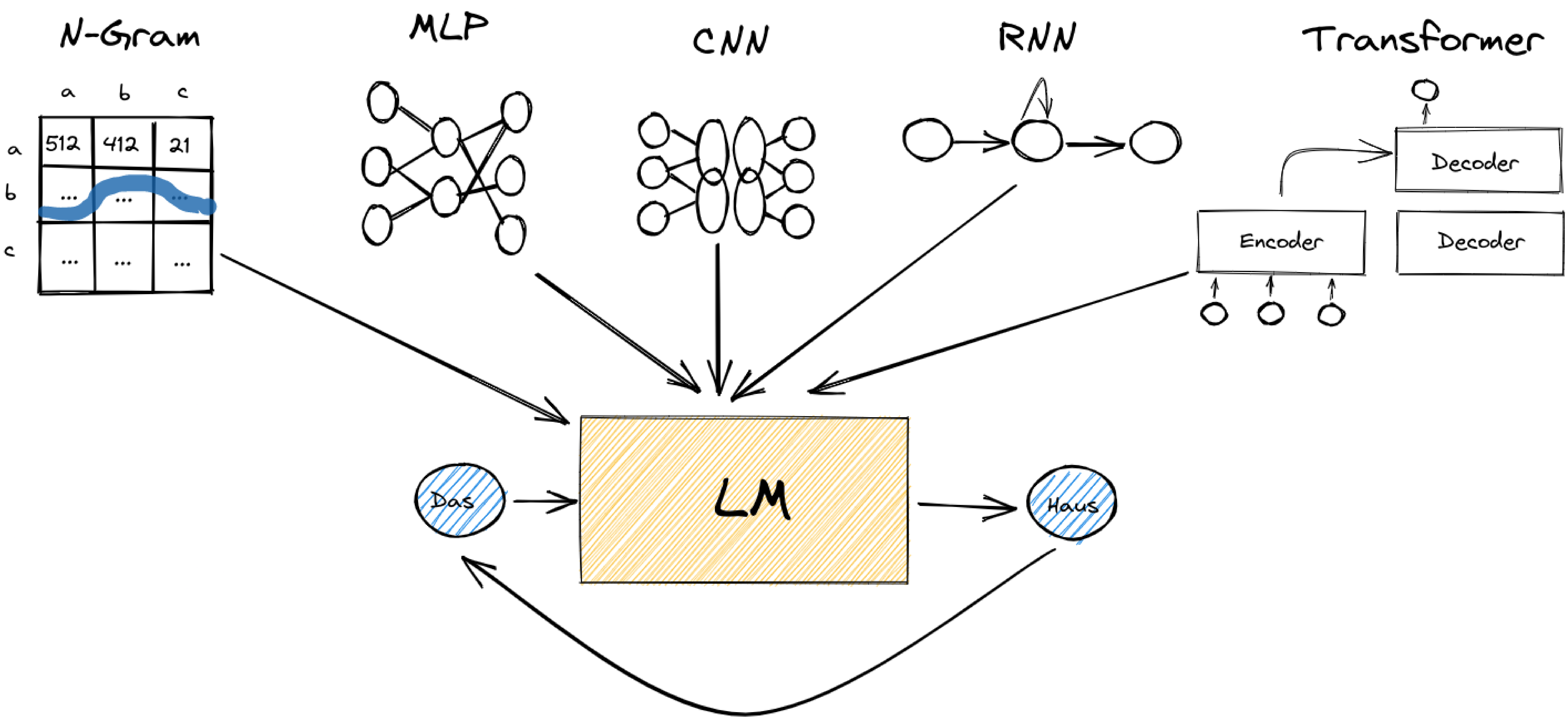
N-gram: N^N Matrix



P(b | b)

P(b | b b)

Future



	a	b	c
a	512	412	21
b	...	...	...
c	...	...	...

Applications of Language Models

- Text Suggestion in Messengers
- Text Suggestion in Code Editors

- Translation

- ...
- Question-Answering
- Image-Captioning
- Text Summarization
- Named Entity Recognition
- Text Classification
- ...

- Dialog

Announcements

Question List

...?
...?
...?
...?

Sources

Blogs
YouTube
github
twitter
papers

Thank you for listening!