Dr. Paula González Avalos

Head of Data Science SPICED

Lead Data Science Coach

@pga99

Big PyData Südwest BBQ 2022

Your first 1000 words

say more than 1000 words

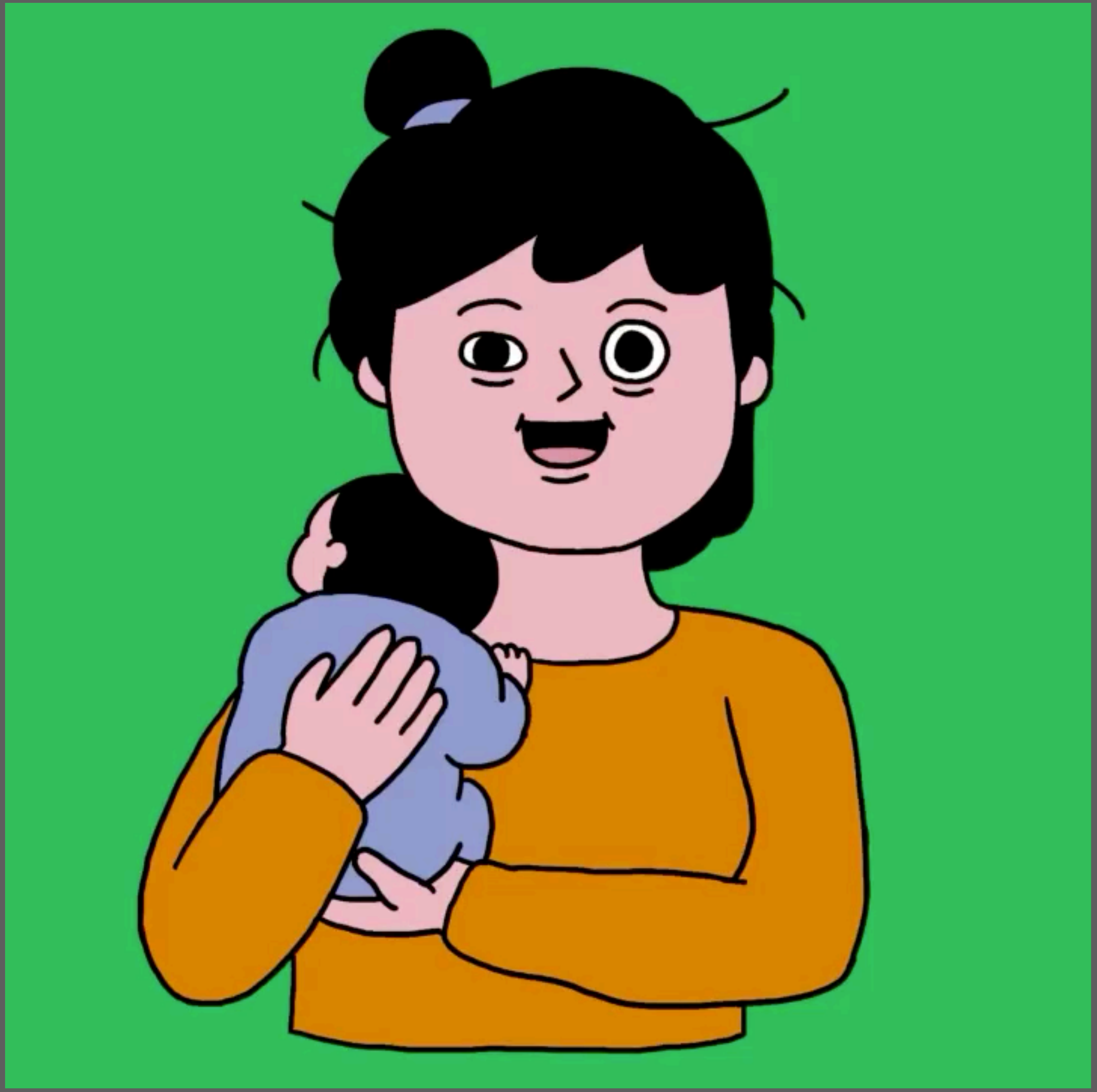
Reasons I ♥ my job

Head of Data Science SPICED

Lead Data Science Coach

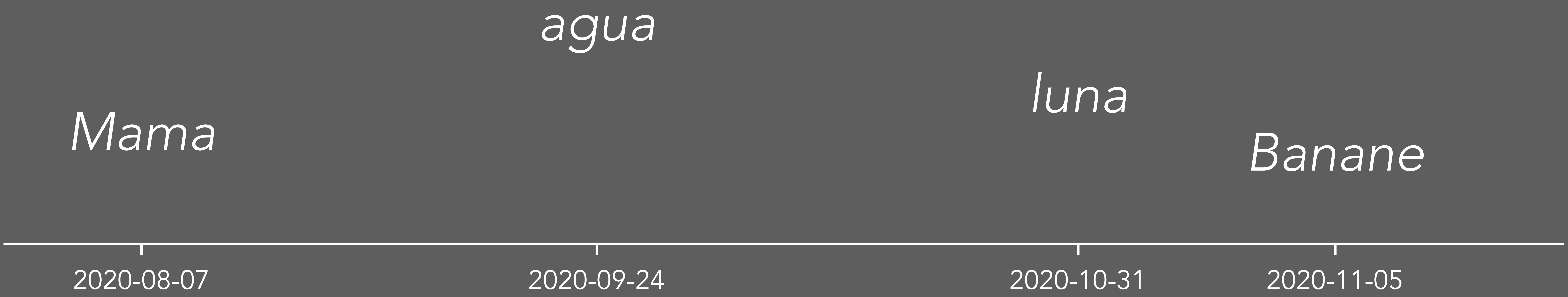
- ☑ Data
- ☑ Learning
- ☑ Teaching
- ☑ Watching others learn

Meanwhile in my other job…



Source: Sherchle

- □ Data
- ☑ Learning
- ☑ Teaching
- ☑ **Watching others learn**



Mama

2020-08-07

agua

2020-09-24

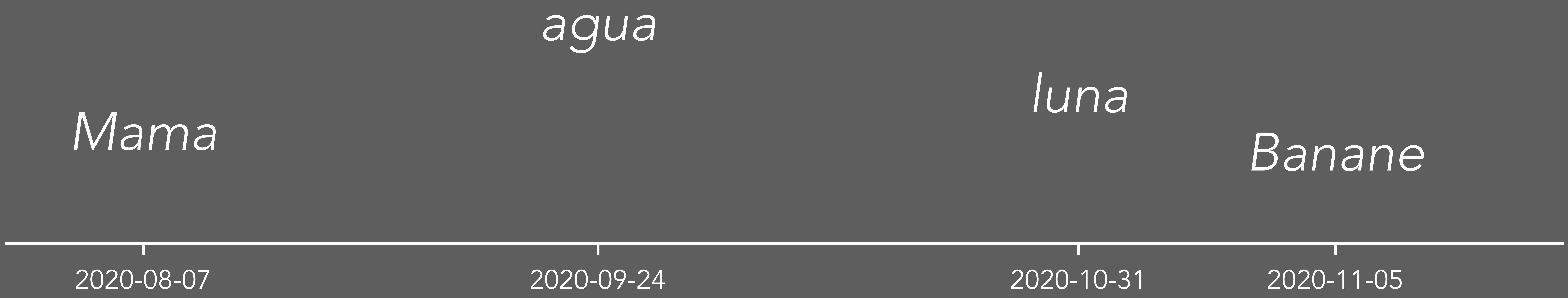
luna

2020-10-31

Banane

2020-11-05

☑ Data



Mama 2020-08-07

agua 2020-09-24

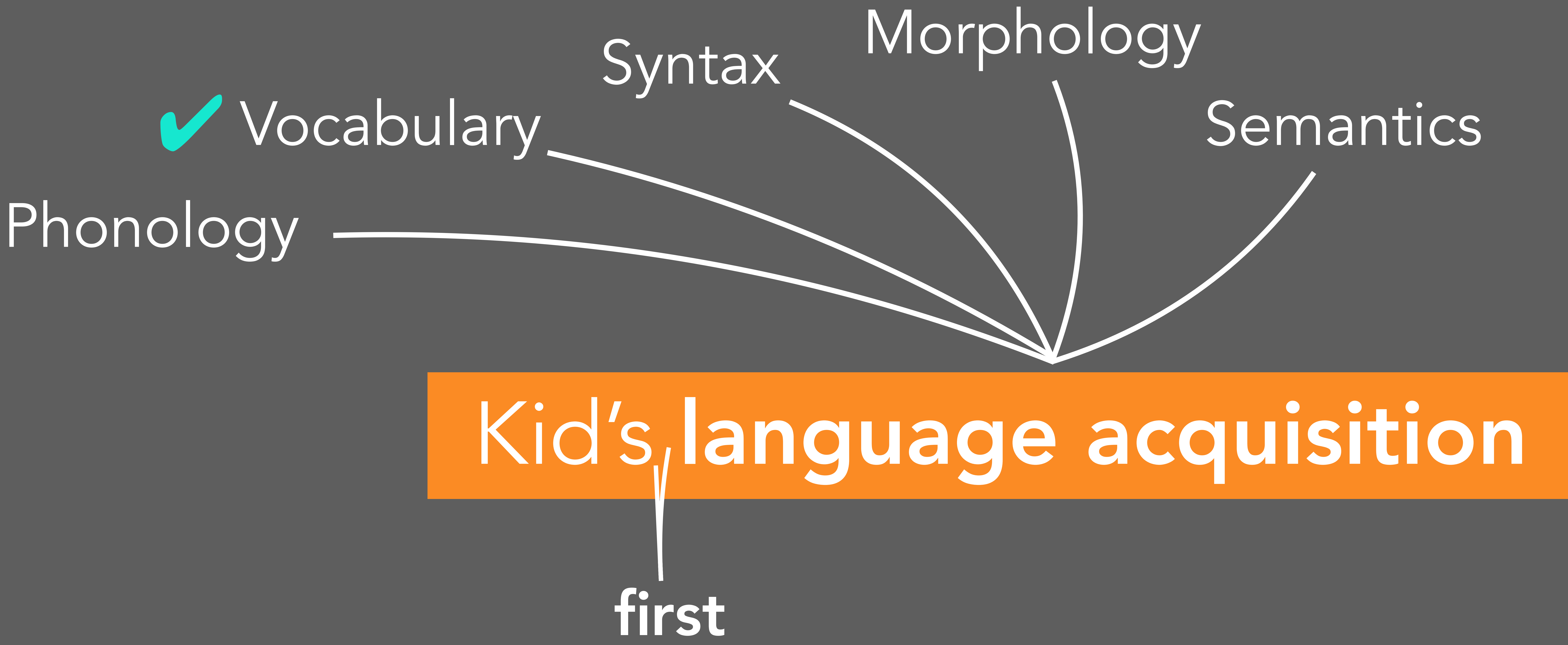
luna 2020-10-31

Banane 2020-11-05

Kid's **language acquisition**

☑ Data First 1000 words| 17 months

pandas seaborn spaCy



Phonology

✓ Vocabulary

Syntax

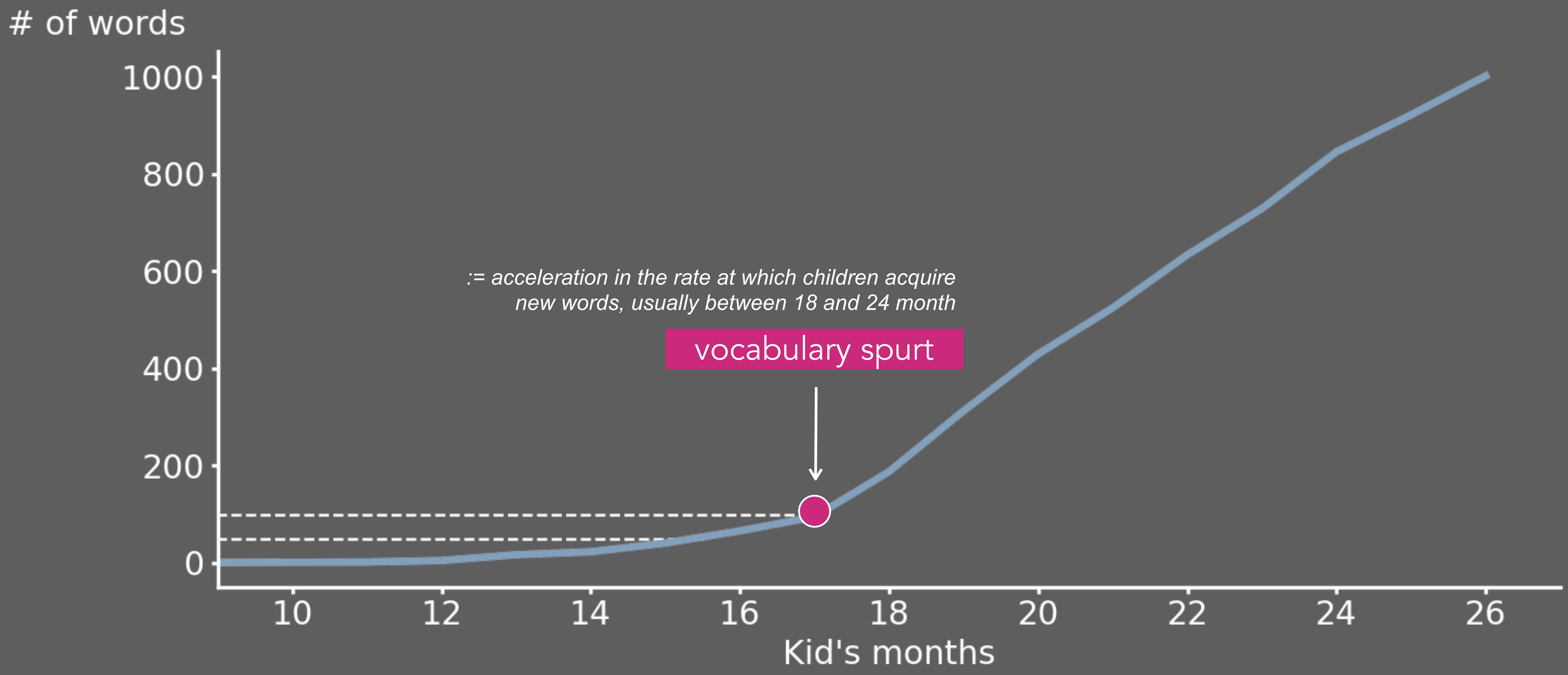
Morphology

Semantics

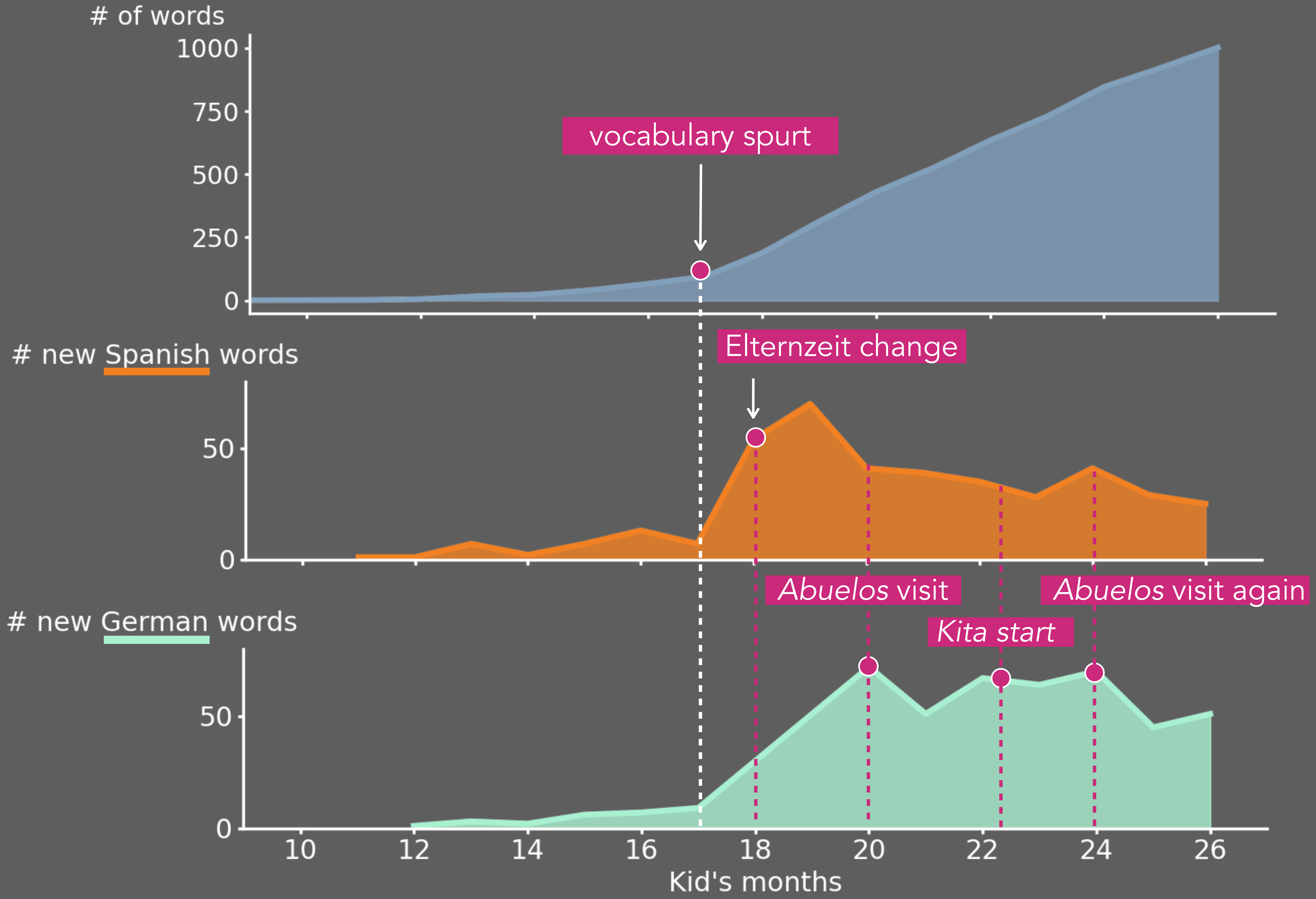
Kid's **language acquisition**

first

Vocabulary increase



:= acceleration in the rate at which children acquire new words, usually between 18 and 24 month



vocabulary spurt

Elternzeit change

Abuelos visit

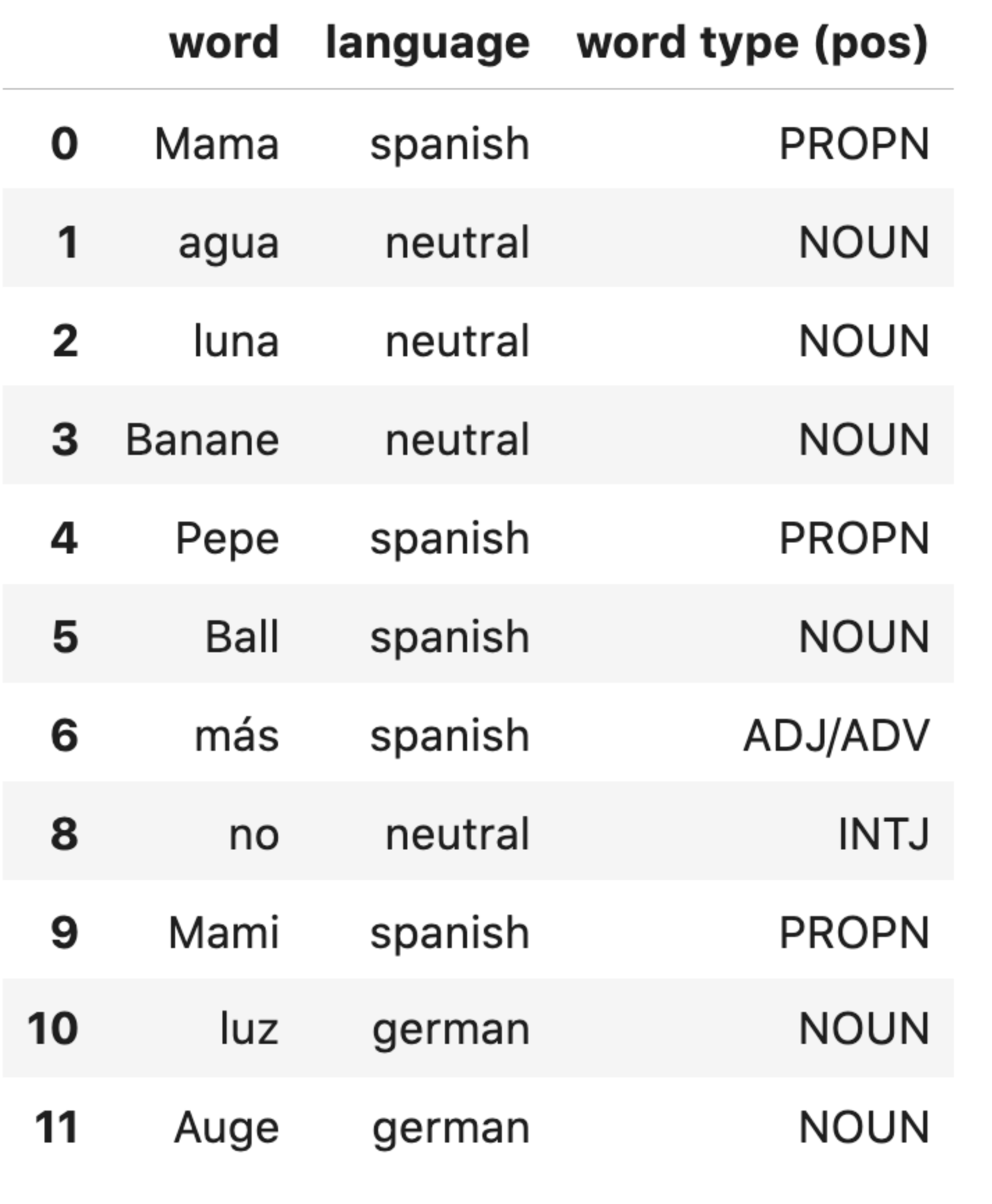
Kita start

Abuelos visit again

Acquiring **morpho-syntax** complexity

First appearance of new types of words spaCy

`token.pos_`



	word	language	word type (pos)
0	Mama	spanish	PROPN
1	agua	neutral	NOUN
2	luna	neutral	NOUN
3	Banane	neutral	NOUN
4	Pepe	spanish	PROPN
5	Ball	spanish	NOUN
6	más	spanish	ADJ/ADV
8	no	neutral	INTJ
9	Mami	spanish	PROPN
10	luz	german	NOUN
11	Auge	german	NOUN

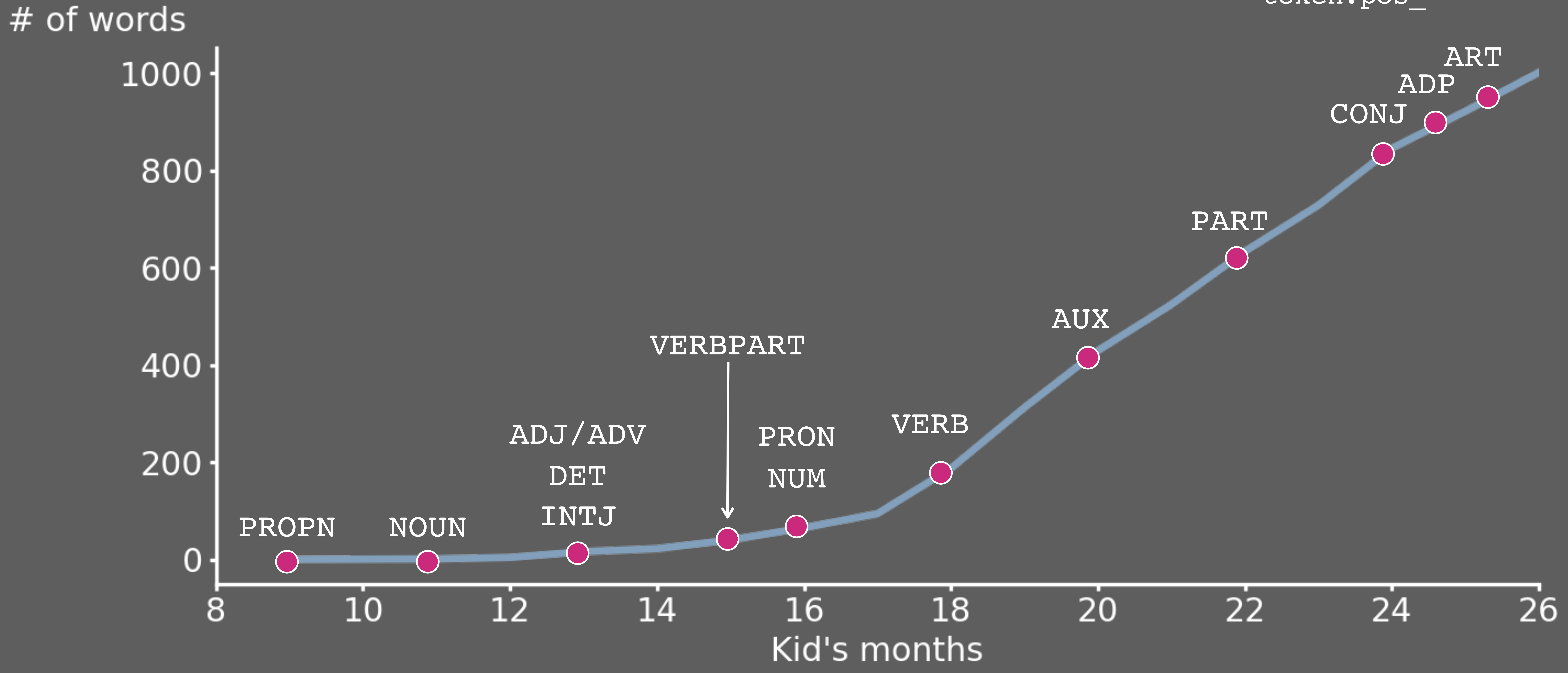
Part-of-speech categories

- `ADJ`: adjective
- `ADP`: adposition
- `ADV`: adverb
- `AUX`: auxiliary
- `CONJ`: conjunction
- `DET`: determiner
- `INTJ`: interjection
- `NOUN`: noun
- `NUM`: numeral
- `PART`: particle
- `PRON`: pronoun
- `PROPN`: proper noun
- `PUNCT`: punctuation
- `SYM`: symbol
- `VERB`: verb
- `X`: other

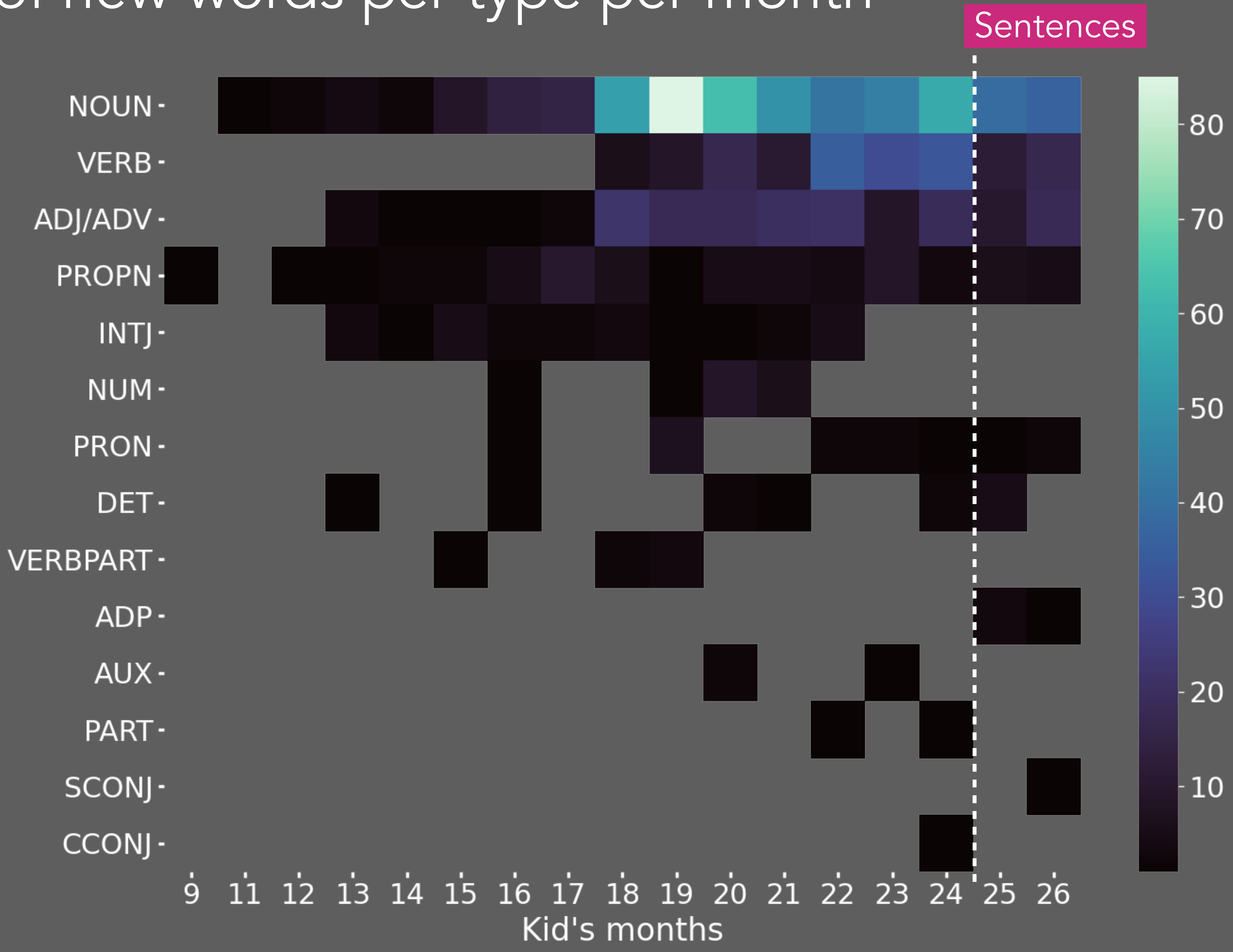
First appearance of new types of words

spaCy

token.pos_



Amount of new words per type per month

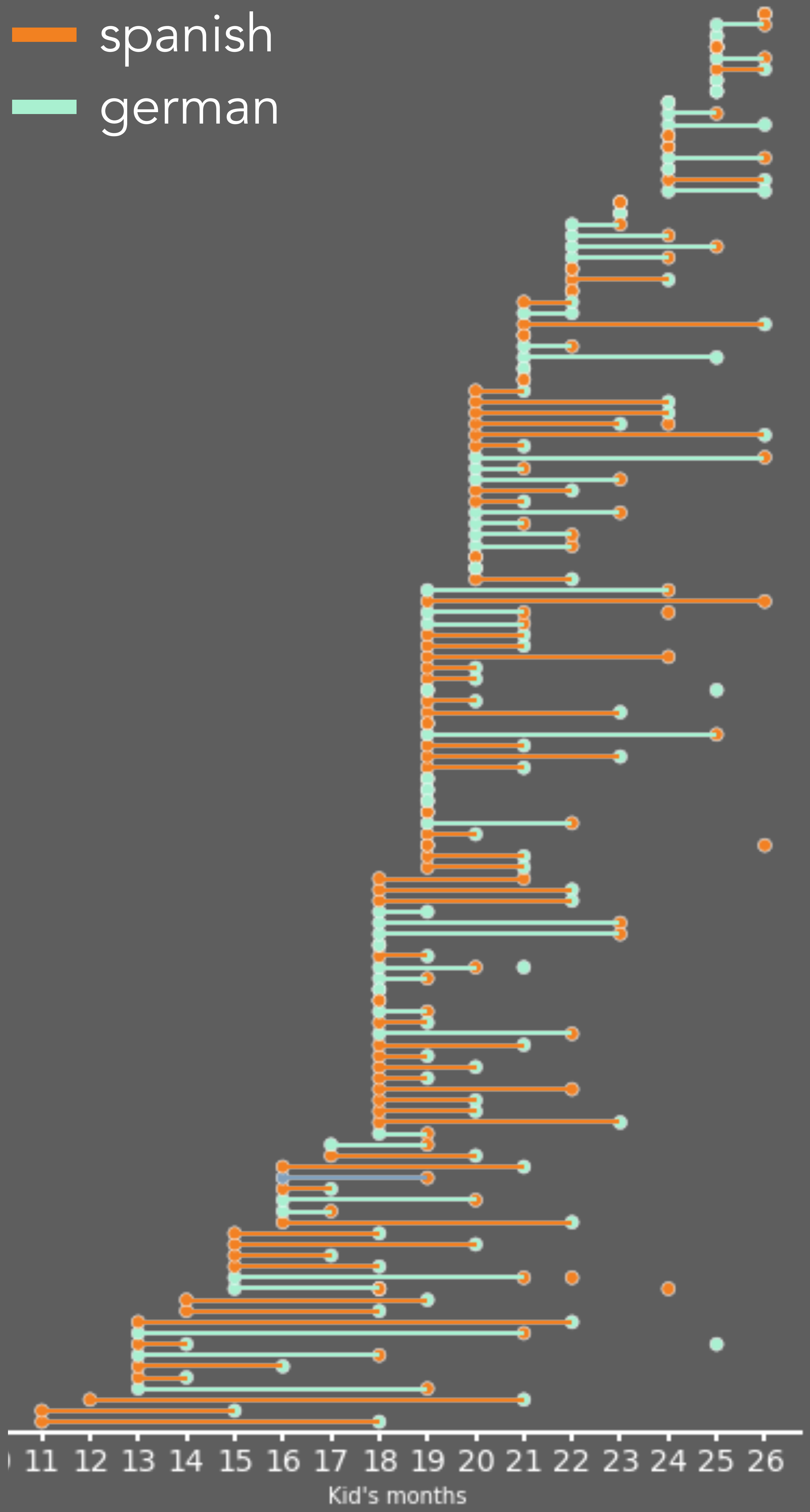


Q: How many of the 1000 words are unique concepts?

i.e. excluding concepts repeated in both languages

- □ ~50 %
- □ ~ 60 % are unique
- ☑ ~ 90 % are unique

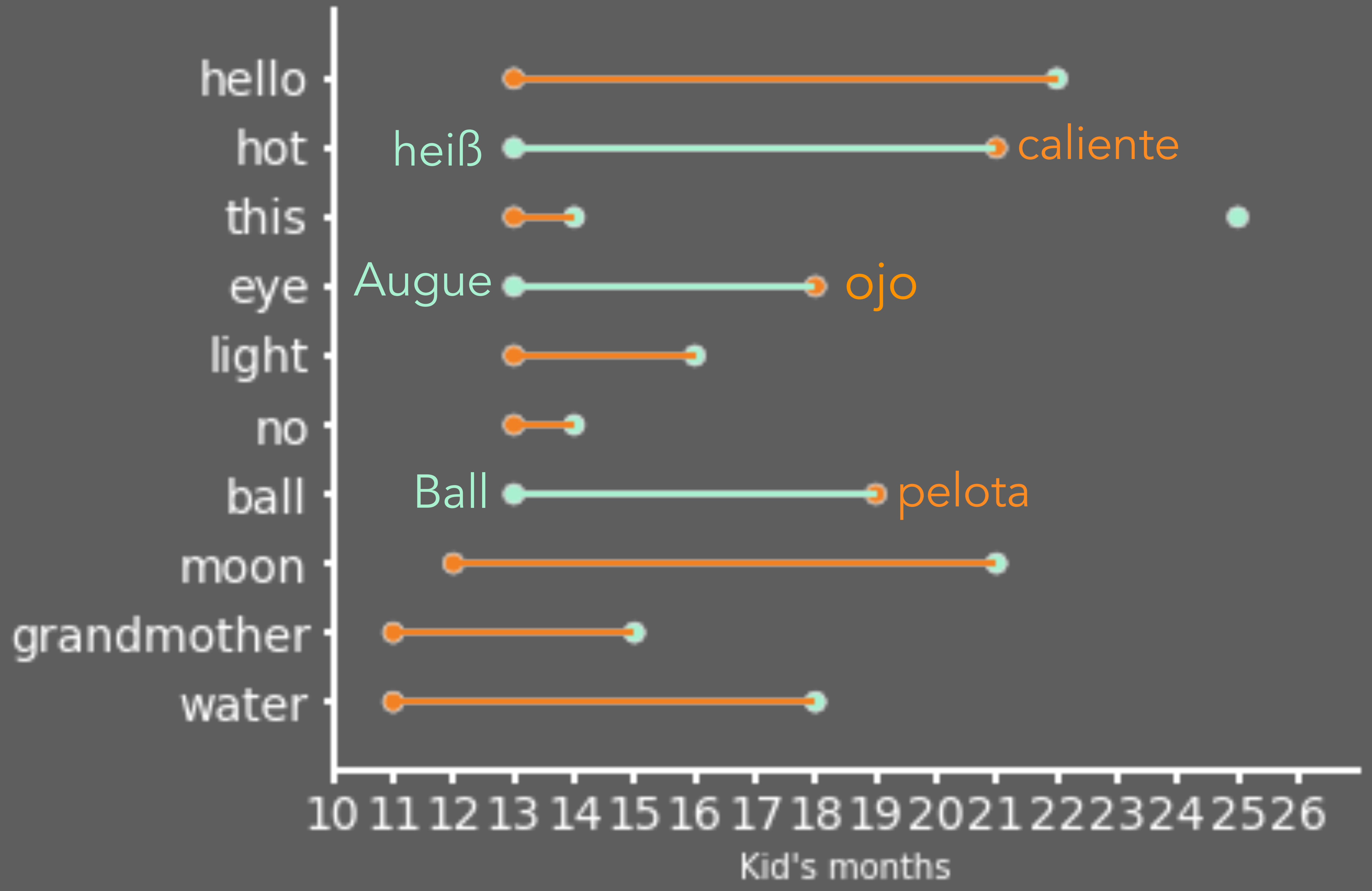
`translation (google translate API) -> remove duplicates`



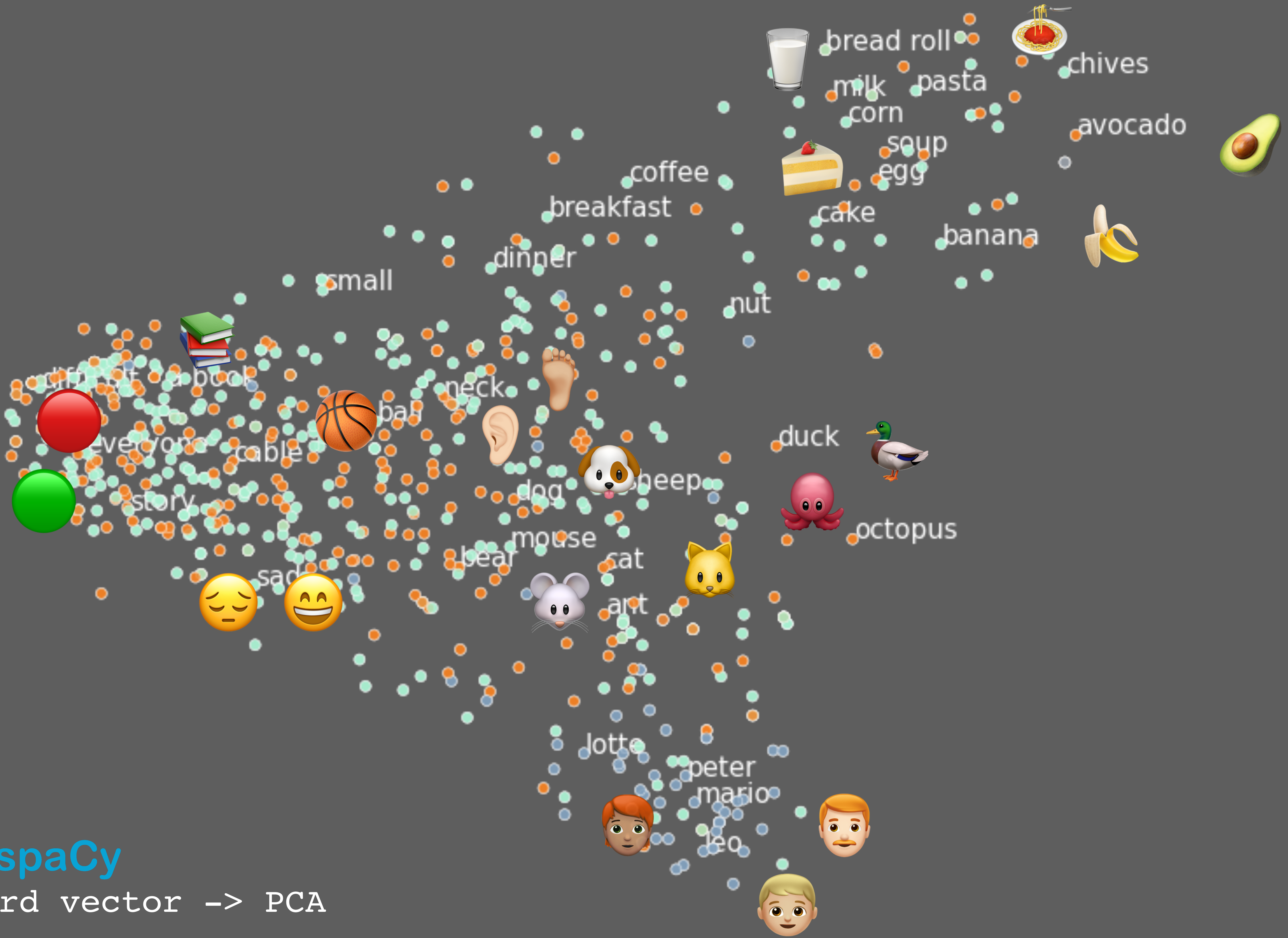
Duplicated concepts

semantics

Why? *Hypothesis: phonetics!*



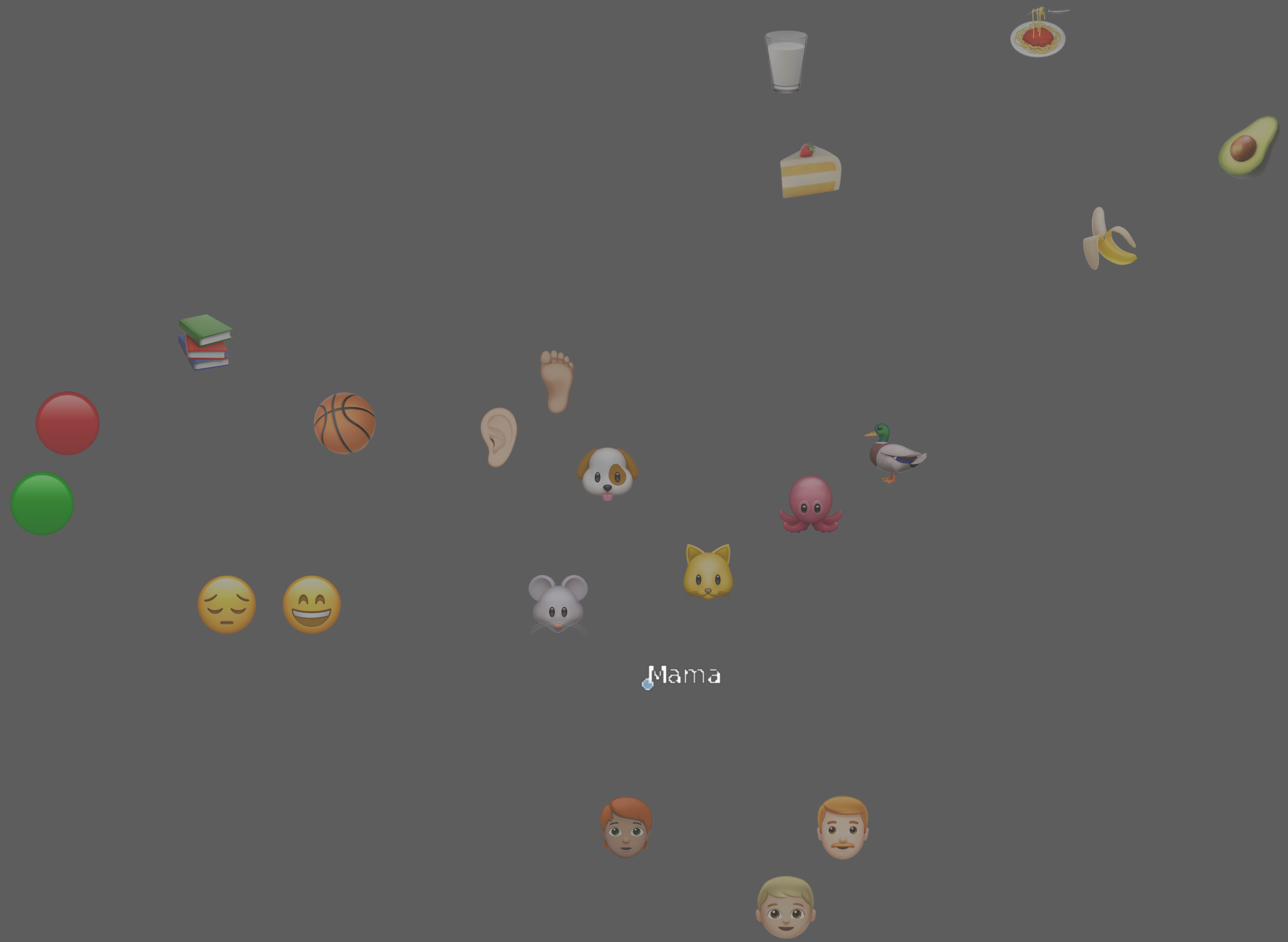
Conquering **semantic** space



spaCy

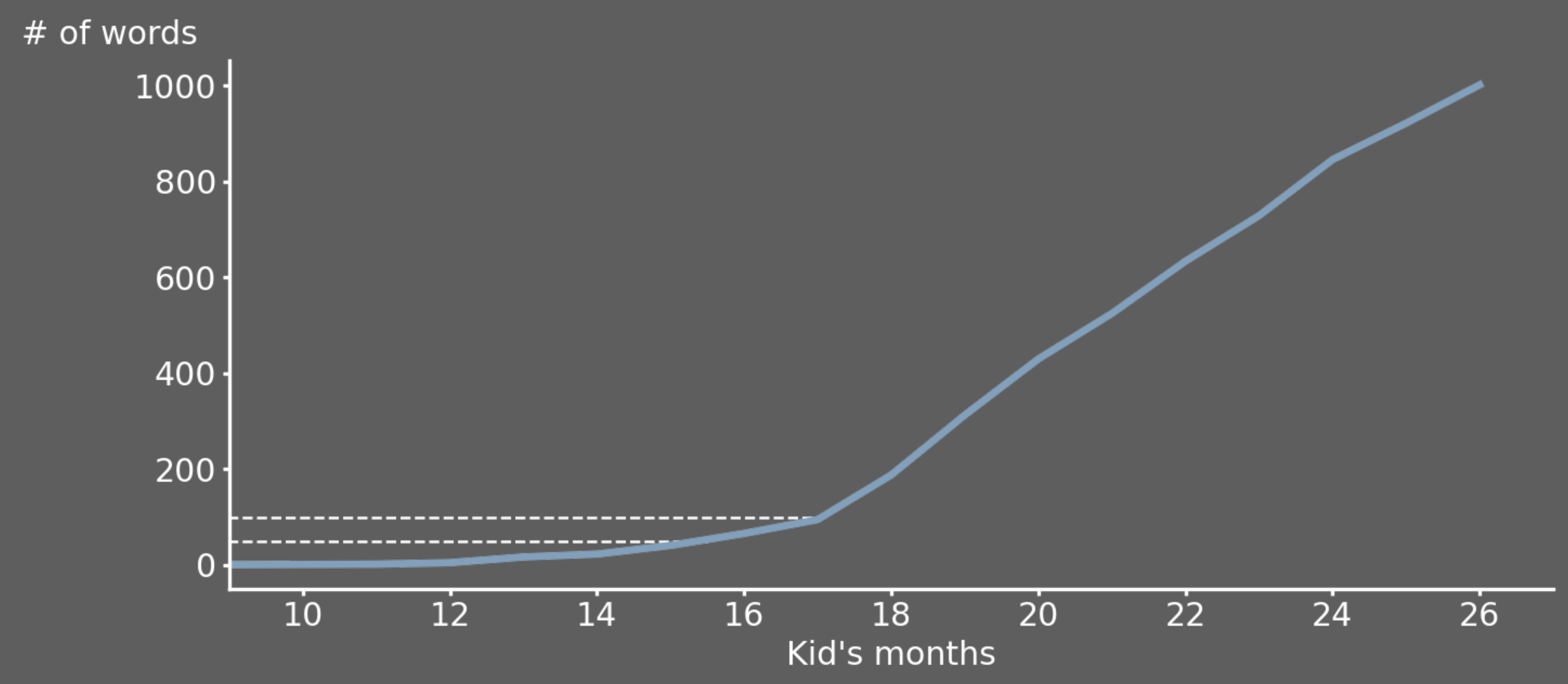
translation -> word vector -> PCA

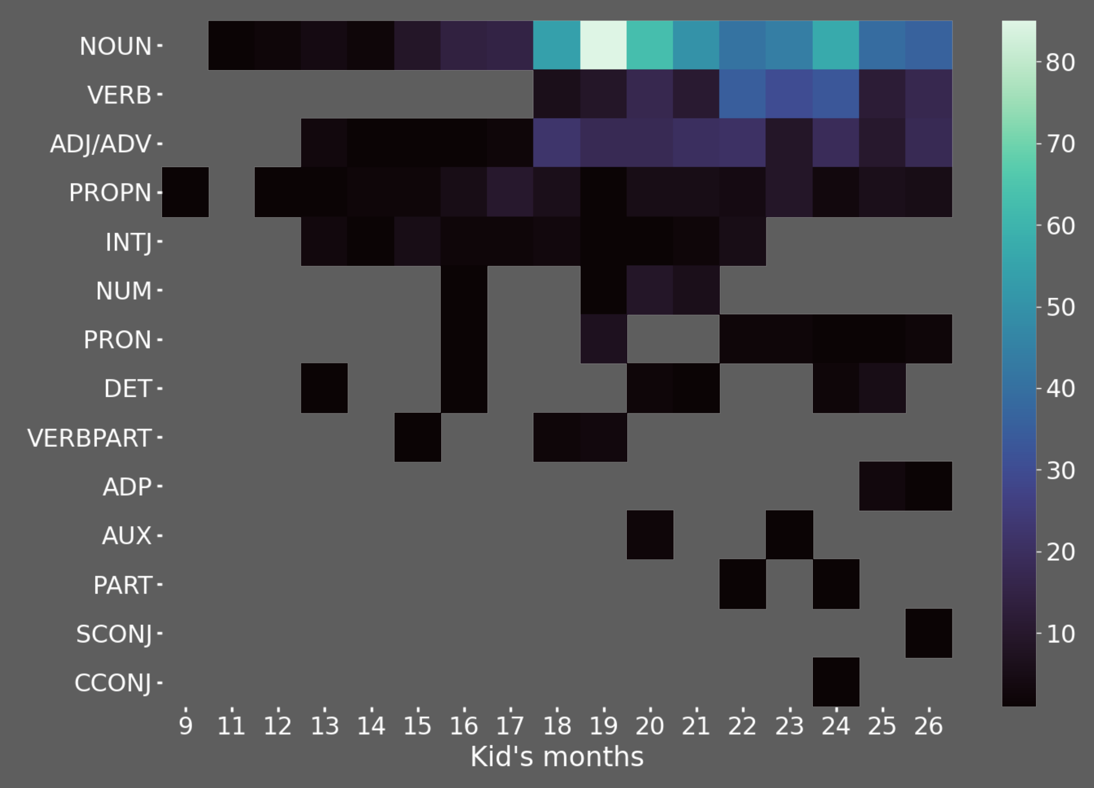
Conquering **semantic** space



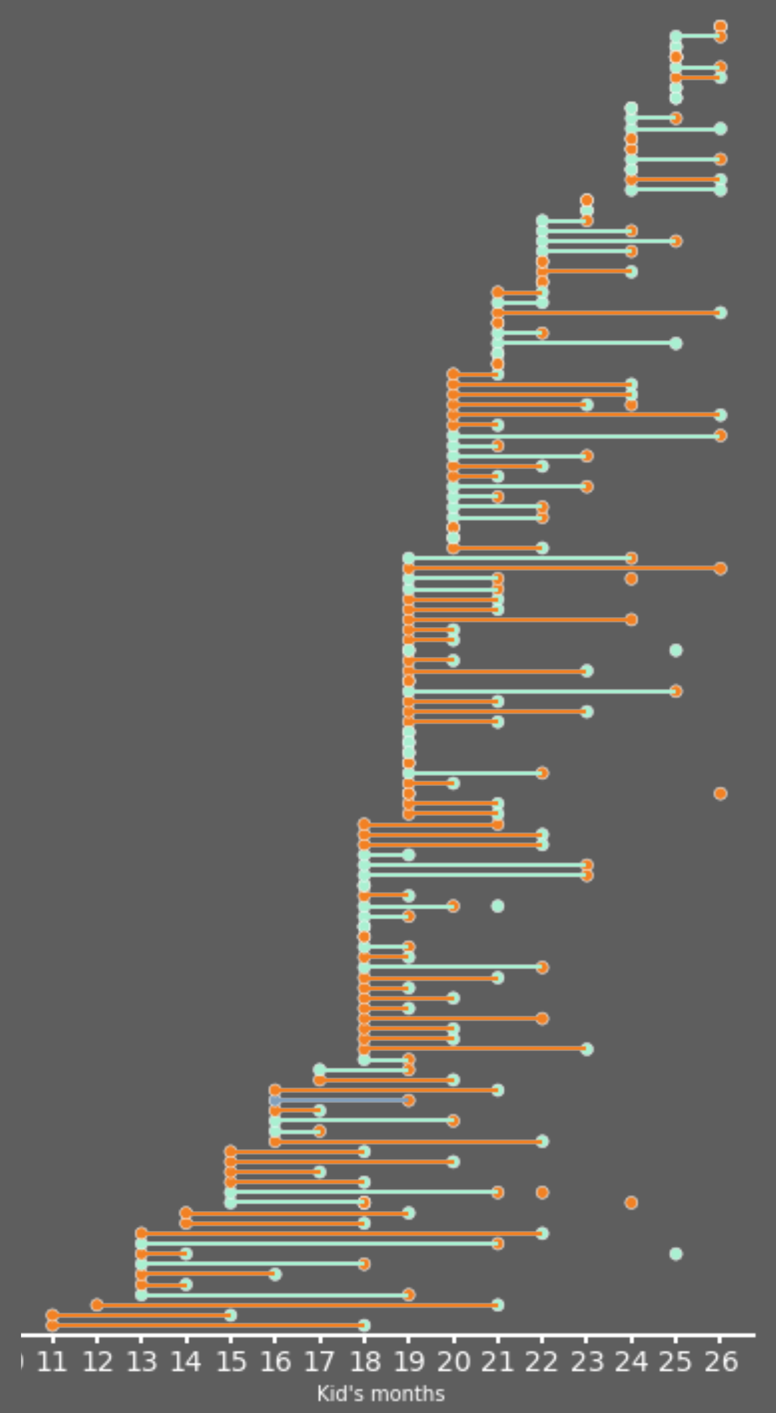
The end.

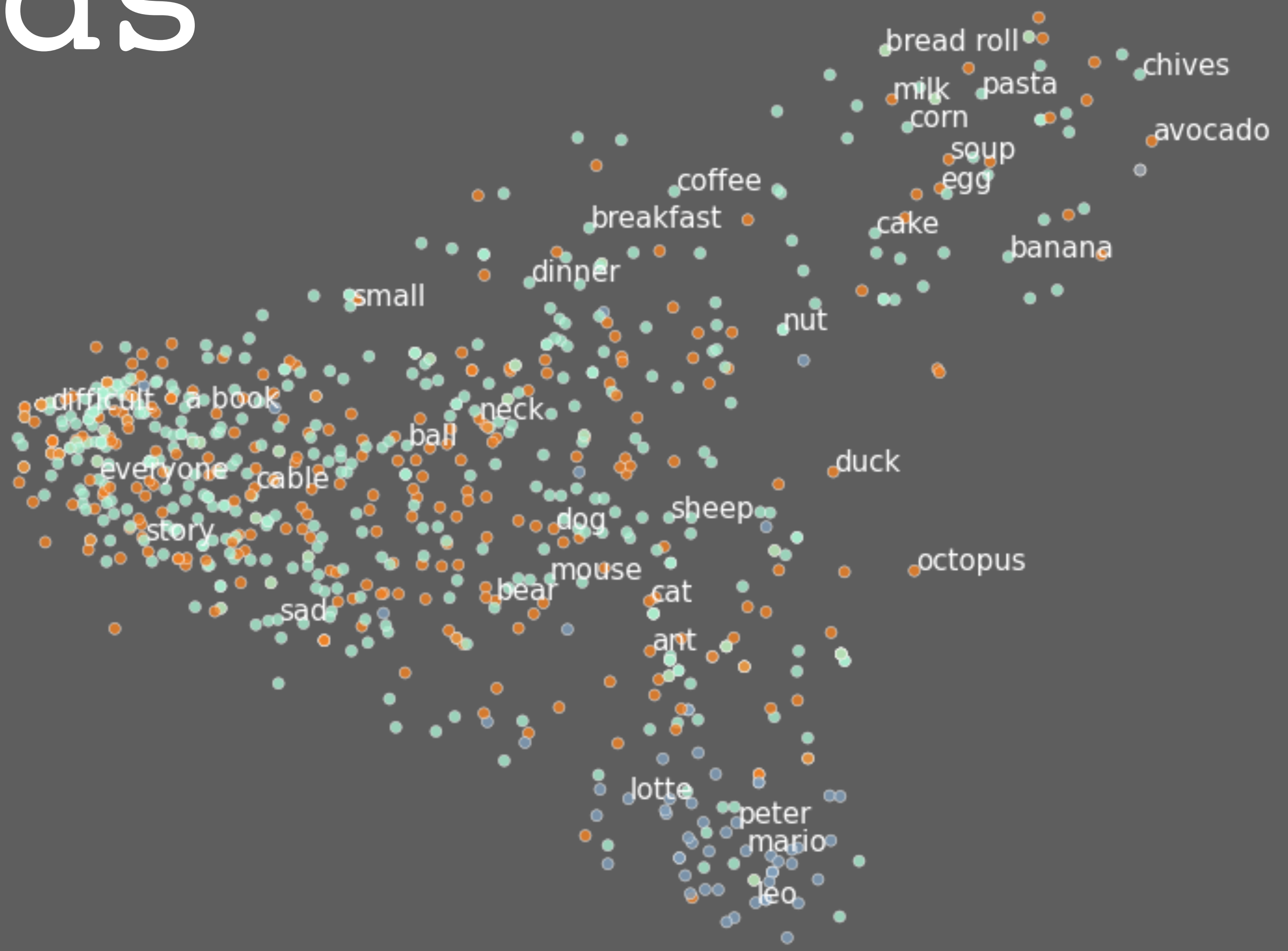
@pga99





First 1000 words





Call for Citizen Science?