Multi-lingual natural language understanding with spaCy

Matthew Honnibal

Explosion AI

Explosion AI is a digital studio specialising in Artificial Intelligence and Natural Language Processing.

spaCy

Open-source library for industrial-strength Natural Language Processing

THINC

spaCy’s next-generation Machine Learning library for deep learning with text

prodigy

A radically efficient data collection and annotation tool, powered by active learning

DataStore

Coming soon: pre-trained, customisable models for a variety of languages and domains

Matthew Honnibal

CO-FOUNDER

PhD in Computer Science in 2009.
10 years publishing research on state-of-the-art natural language understanding systems.
Left academia in 2014 to develop spaCy.

Ines Montani

CO-FOUNDER

Programmer and front-end developer with degree in media science and linguistics.
Has been working on spaCy since its first release. Lead developer of Prodigy.

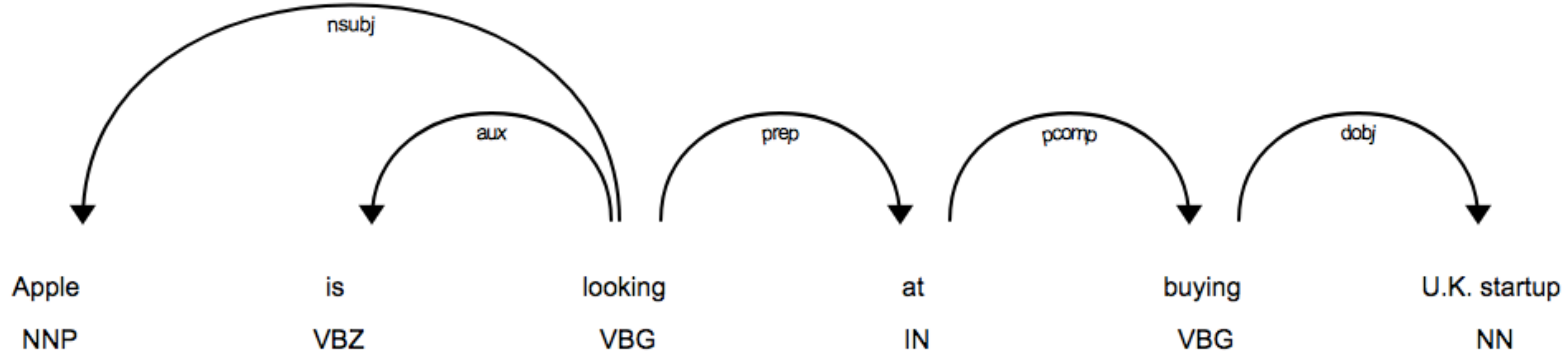
“I don’t get it. Can you explain like I’m five?”

Think of us as a boutique kitchen.

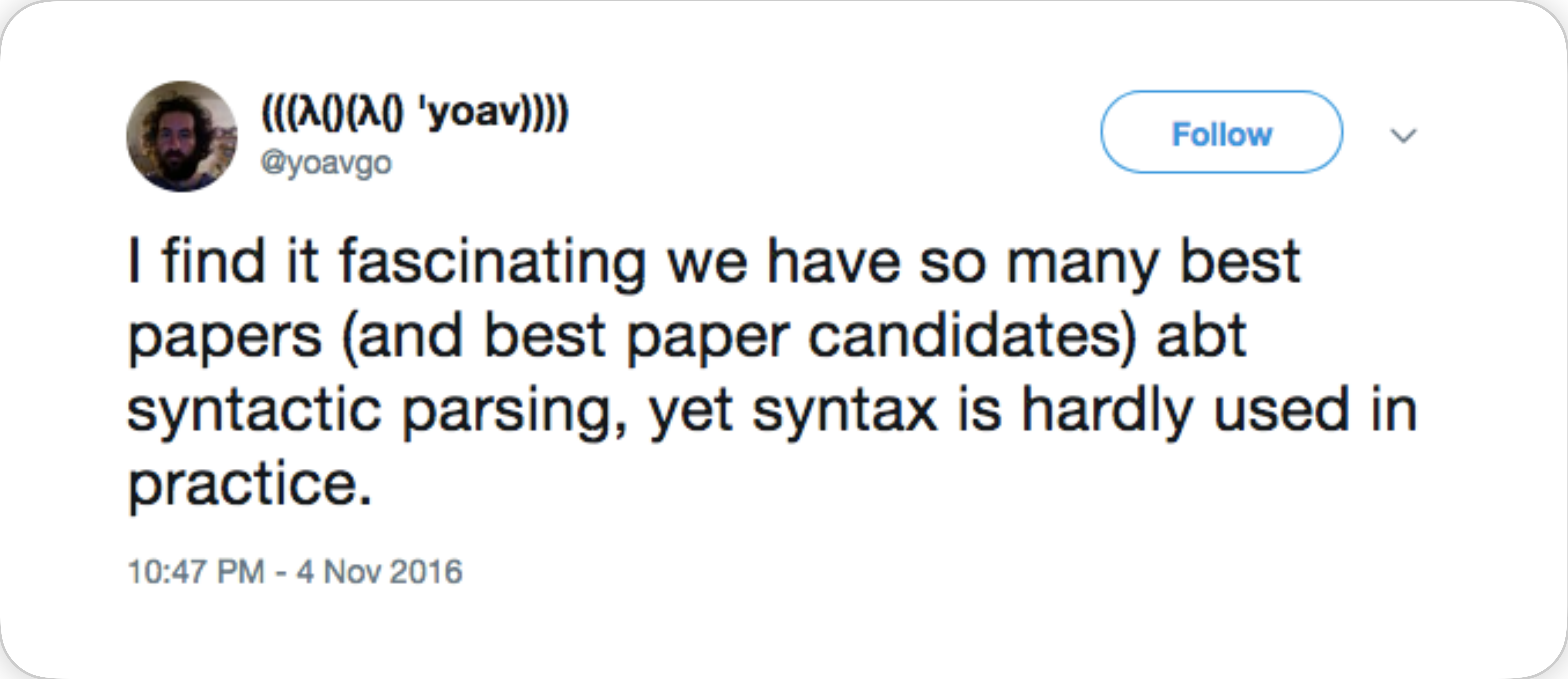
- **free recipes** published online **open-source software**
- **catering** for select events **consulting**
- a line of kitchen **gadgets** **downloadable tools**
- *soon:* a line of fancy **sauces** and **spice mixes** you can use at home **pre-trained models**

Joint transition-based segmentation and parsing

```
doc = nlp(u"Apple is looking at buying U.K. startup")

for token in doc:
    print(token.text, token.pos_, token.tag_, token.dep_,
          token.head.text, token.lefts, token.rights)
```



What’s parsing good for?



(((λ()(λ() 'yoav))))
@yoavgo

I find it fascinating we have so many best papers (and best paper candidates) abt syntactic parsing, yet syntax is hardly used in practice.

10:47 PM - 4 Nov 2016

🌳 Trees are the truth

- sentences are tree-structured
- dependencies can be **arbitrarily long** in string space
- syntax is **application-independent**

🌳 Trees are the truth

- ○ sentences are tree-structured

  ... but they’re read and written in order

- ○ dependencies can be **arbitrarily long** in string space

  ... but they’re usually short

- ○ syntax is **application-independent**

  Learn the language once, apply it many times.

Whitespace != Word

income tax return

Einkommensteuererklärung

C’est une phrase.

זהו משפט.

これは文章です。

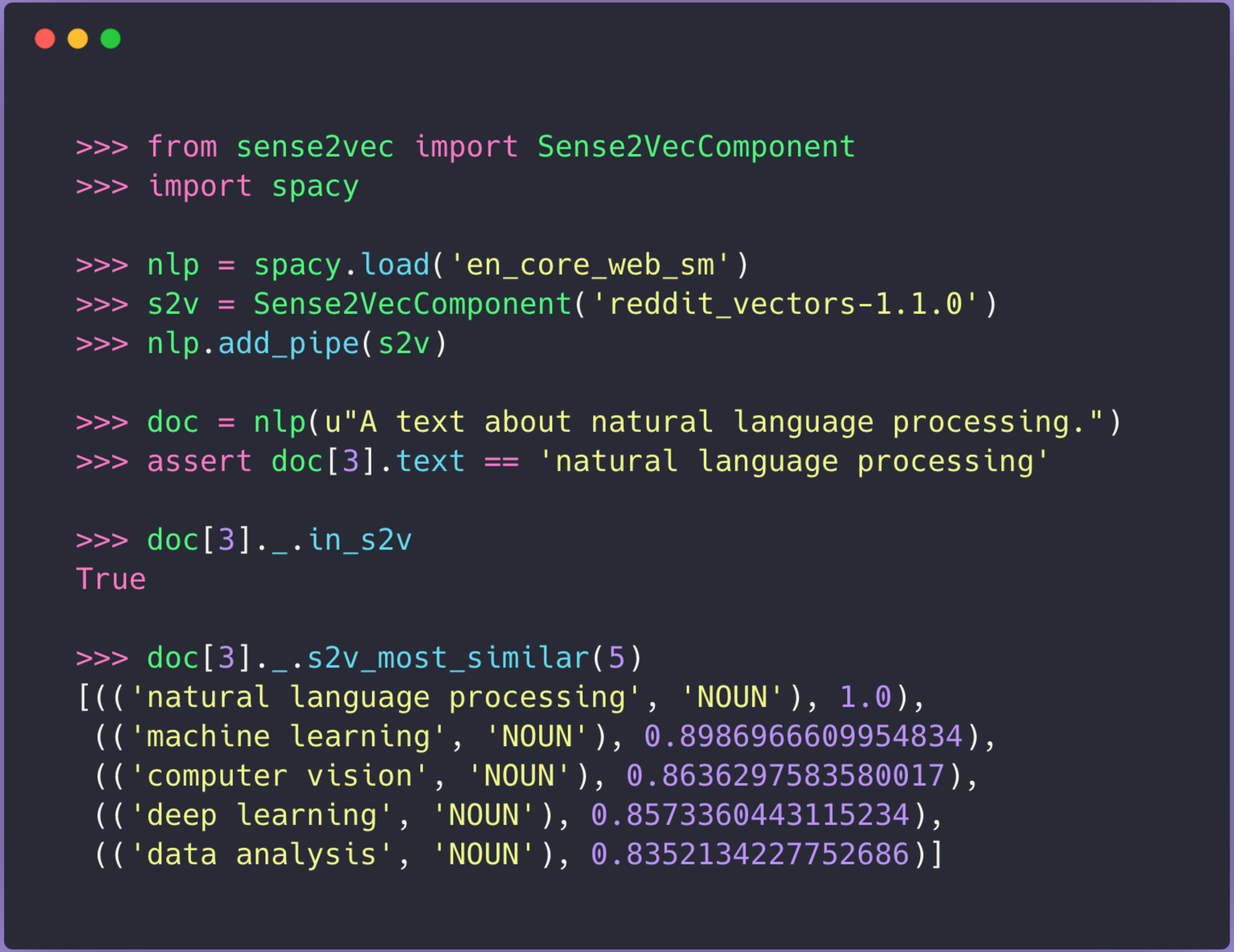
sense2vec: Semantic Analysis of the Reddit Hivemind

Our neural network read every comment posted to Reddit in 2015, and built a semantic map using word2vec and spaCy. Try searching for a phrase that's more than the sum of its parts to see what the model thinks it means. Try your favourite band, slang words, technical things, or something totally random.

Term: natural language processing

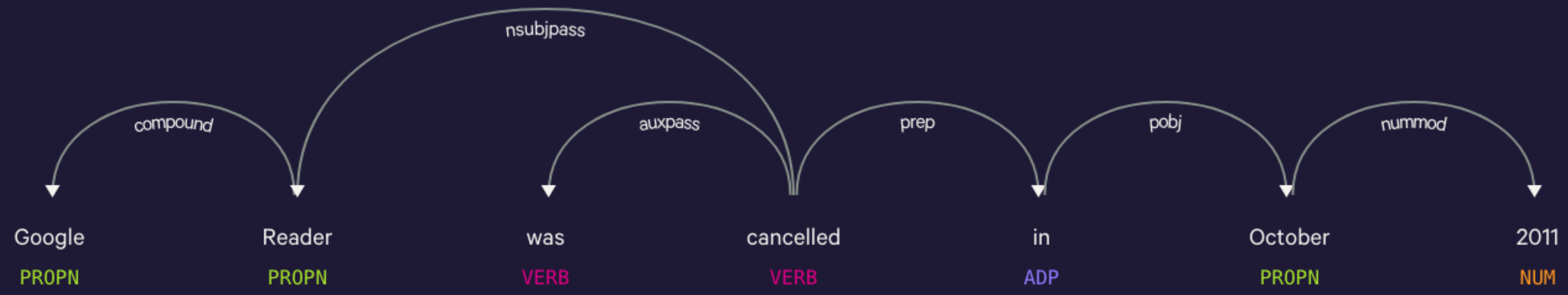
Sense: auto

machine learning ›	90%
computer vision ›	86%
data analysis ›	84%
neural nets ›	83%
relational databases ›	82%
algorithms ›	81%
neural networks ›	80%
image recognition ›	80%



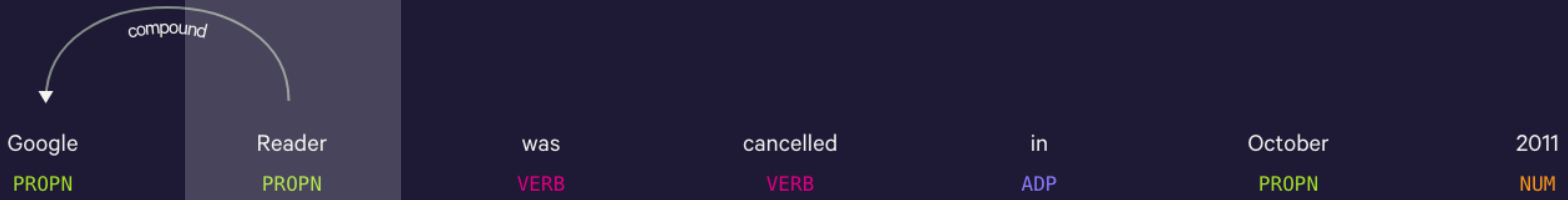
```
>>> from sense2vec import Sense2VecComponent
>>> import spacy

>>> nlp = spacy.load('en_core_web_sm')
>>> s2v = Sense2VecComponent('reddit_vectors-1.1.0')
>>> nlp.add_pipe(s2v)

>>> doc = nlp(u"A text about natural language processing.")
>>> assert doc[3].text == 'natural language processing'

>>> doc[3]._.in_s2v
True

>>> doc[3]._.s2v_most_similar(5)
[(('natural language processing', 'NOUN'), 1.0),
 (('machine learning', 'NOUN'), 0.8986966609954834),
 (('computer vision', 'NOUN'), 0.8636297583580017),
 (('deep learning', 'NOUN'), 0.8573360443115234),
 (('data analysis', 'NOUN'), 0.8352134227752686)]
```

How the parser works


nsubjpass
compound
auxpass
prep
pobj
nummod
Google
PROPN
Reader
PROPN
was
VERB
cancelled
VERB
in
ADP
October
PROPN
2011
NUM

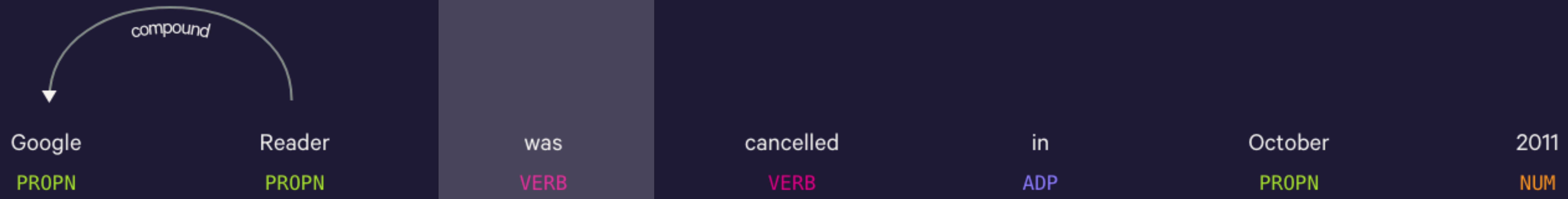
Google	Reader	was	cancelled	in	October	2011
PROPN	PROPN	VERB	VERB	ADP	PROPN	NUM

	Google					
Google	Reader	was	cancelled	in	October	2011
PROPN	PROPN	VERB	VERB	ADP	PROPN	NUM



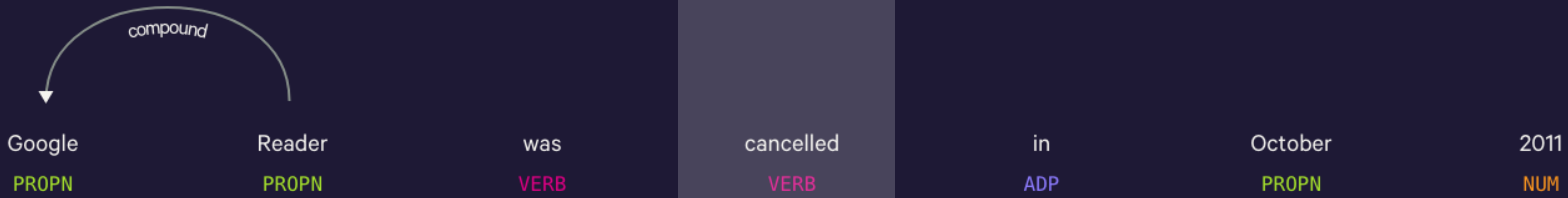
Google	Reader	was	cancelled	in	October	2011
PROPN	PROPN	VERB	VERB	ADP	PROPN	NUM

Reader


compound
Google
PROPN
Reader
PROPN
was
VERB
cancelled
VERB
in
ADP
October
PROPN
2011
NUM

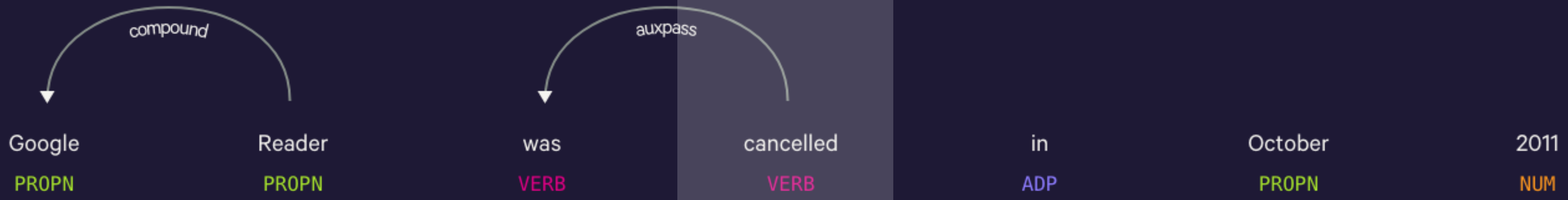
was

Reader



Google	Reader	was	cancelled	in	October	2011
PROPN	PROPN	VERB	VERB	ADP	PROPN	NUM

Reader



compound

auxpass

Google PROPN

Reader PROPN

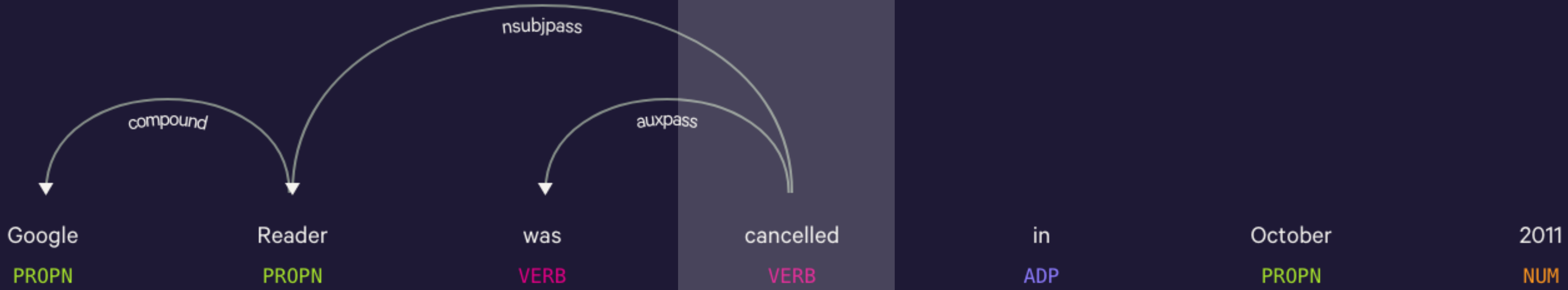
was VERB

cancelled VERB

in ADP

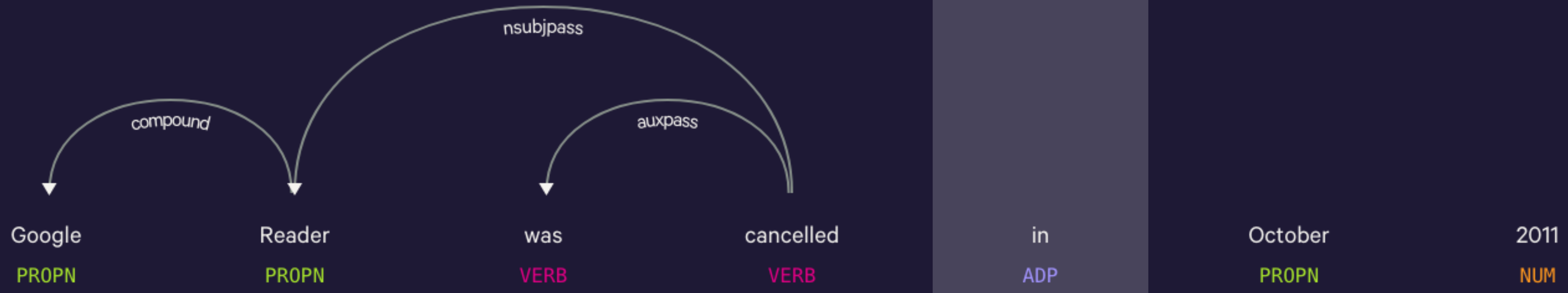
October PROPN

2011 NUM



Google	Reader	was	cancelled	in	October	2011
PROPN	PROPN	VERB	VERB	ADP	PROPN	NUM

cancelled



Google PROPN

Reader PROPN

was VERB

cancelled VERB

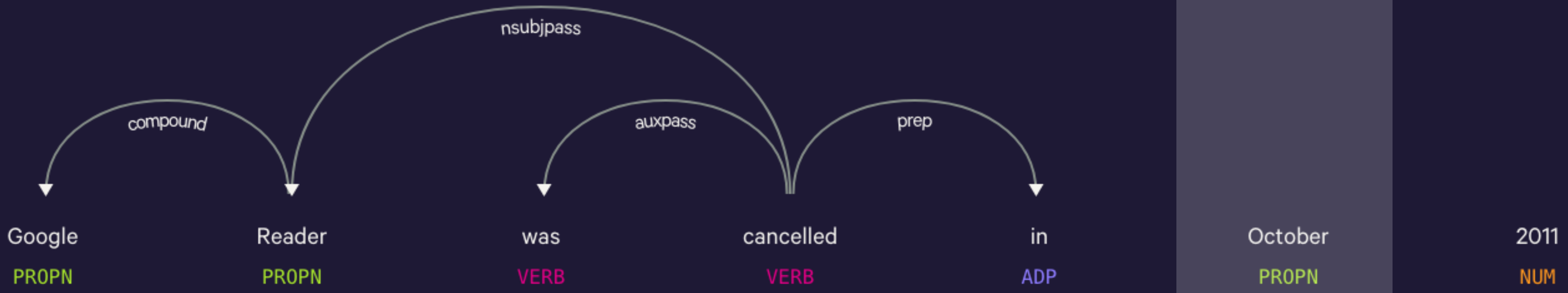
in ADP

October PROPN

2011 NUM

in

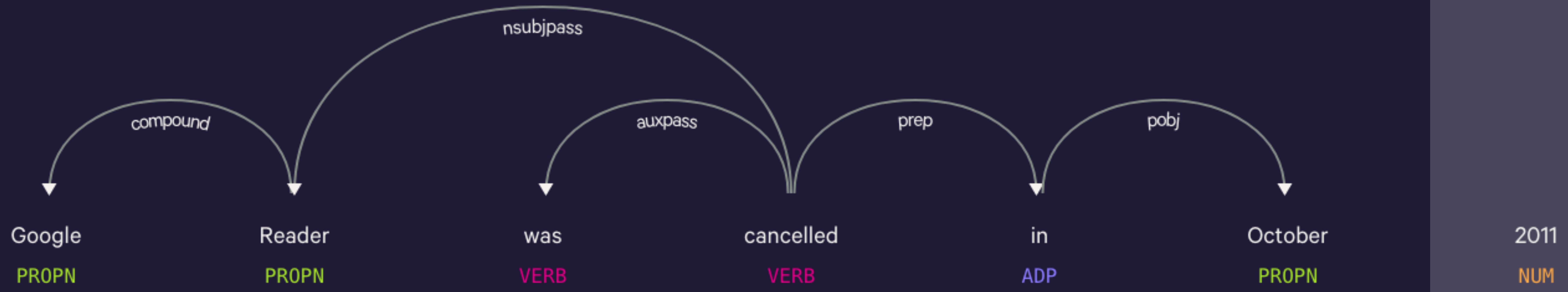
cancelled



October

in

cancelled



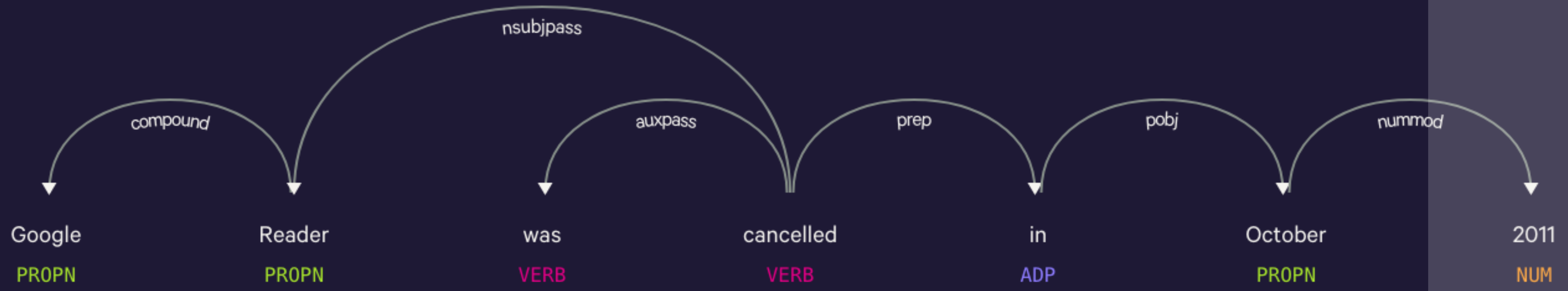
Google	Reader	was	cancelled	in	October	2011
PROPN	PROPN	VERB	VERB	ADP	PROPN	NUM

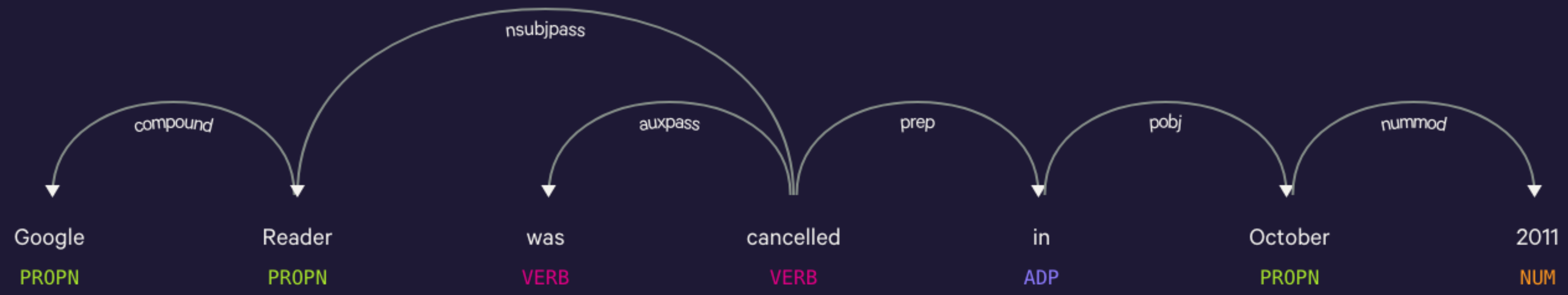
2011

October

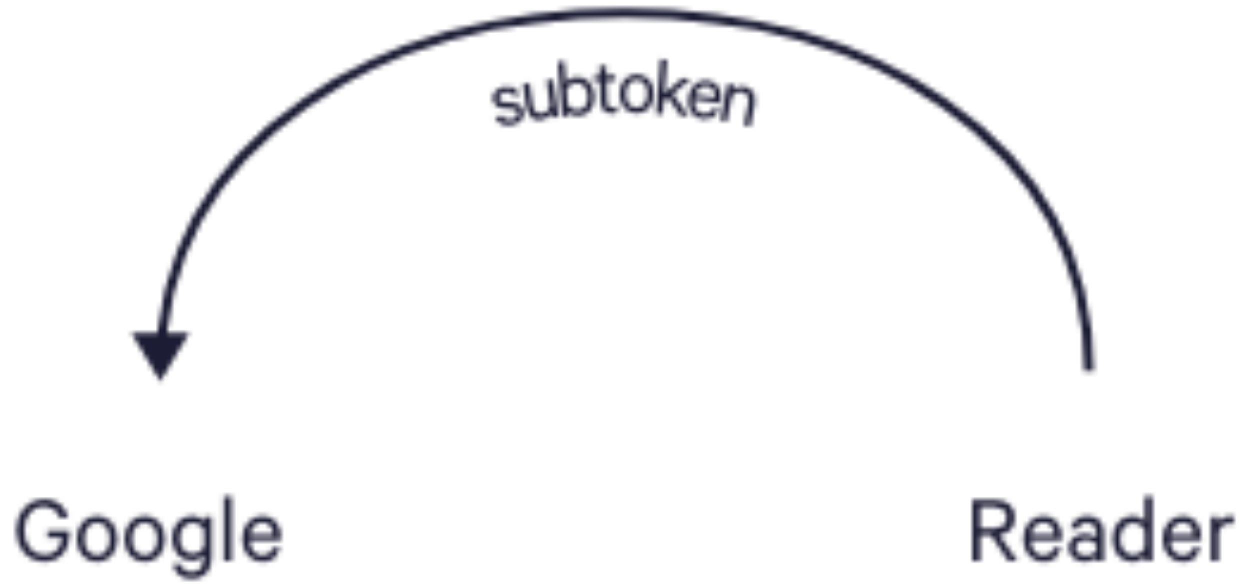
in

cancelled

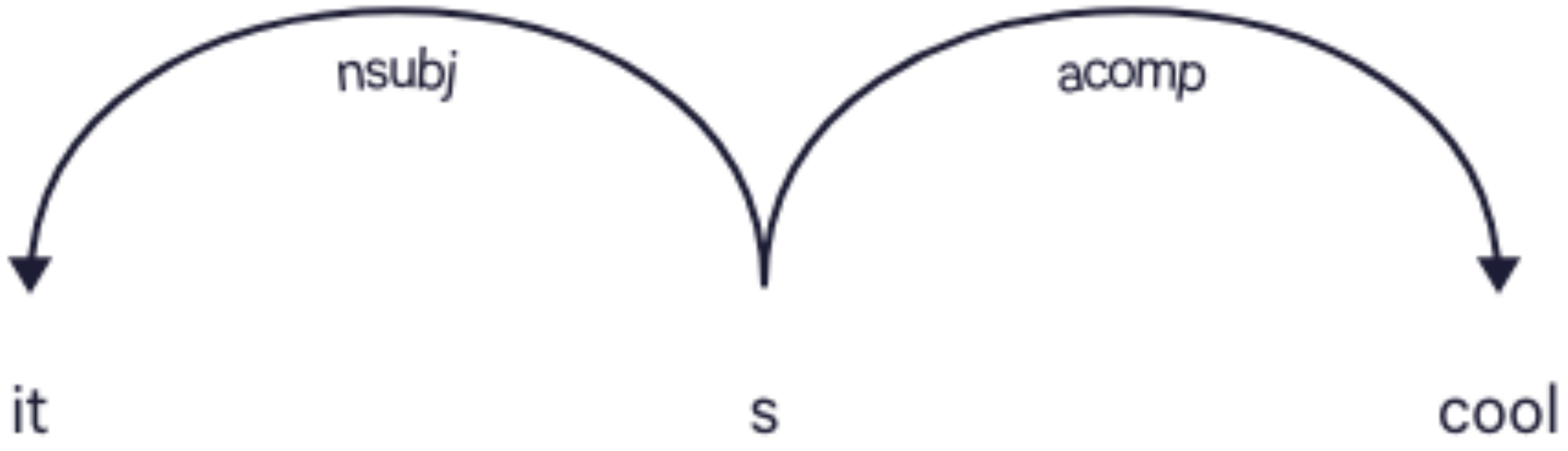



nsubjpass
compound
auxpass
prep
pobj
nummod
Google
PROPN
Reader
PROPN
was
VERB
cancelled
VERB
in
ADP
October
PROPN
2011
NUM

spaCy for other languages



learning to merge tokens



learning to split tokens

What’s the current progress?

- implemented **learning to merge**
- working on **learning to split**
- ranking **~2nd place** on the **CoNLL** 2017 benchmark
- great results for **Chinese**, **Vietnamese**, **Japanese**
- **joint model** consistently better than pipeline

Workflow of the future

- start with **pre-trained models**
- same representation **across languages**
- parse tree enables powerful rule-based **matching**
- updateable models for accuracy on **your domain**
- rapid iteration and **data annotation**

Thanks!

💥 **Explosion AI**
explosion.ai

📲 **Follow us on Twitter**
@honnibal
@_inesmontani
@explosion_ai