Tailoring a Natural Language Processing Pipeline for Mad Lib Auto-Generation

Lauryn Anderson, University of Calgary

1 April 2023

Motivation

What are Mad Libs?

Mad Libs:

a _____ word game
adjective

for _____
plural noun

- Fill-in-the-blank game
- Collaborative humour
- Accessible
- Educational
- Blanks are chosen deliberately so that resulting sentence is nonsensical but grammatical enough to be read fluidly

Goal: Identify optimal words for replacement in a Mad Lib

Method

What is Natural Language Processing?

- Intersection of Linguistics, Computer Science, and Artificial Intelligence
- Can computers process/understand human language?
- Uses mixture of statistical/rule-based algorithms and trained neural networks
- Many relevant sub-functionalities within the discipline

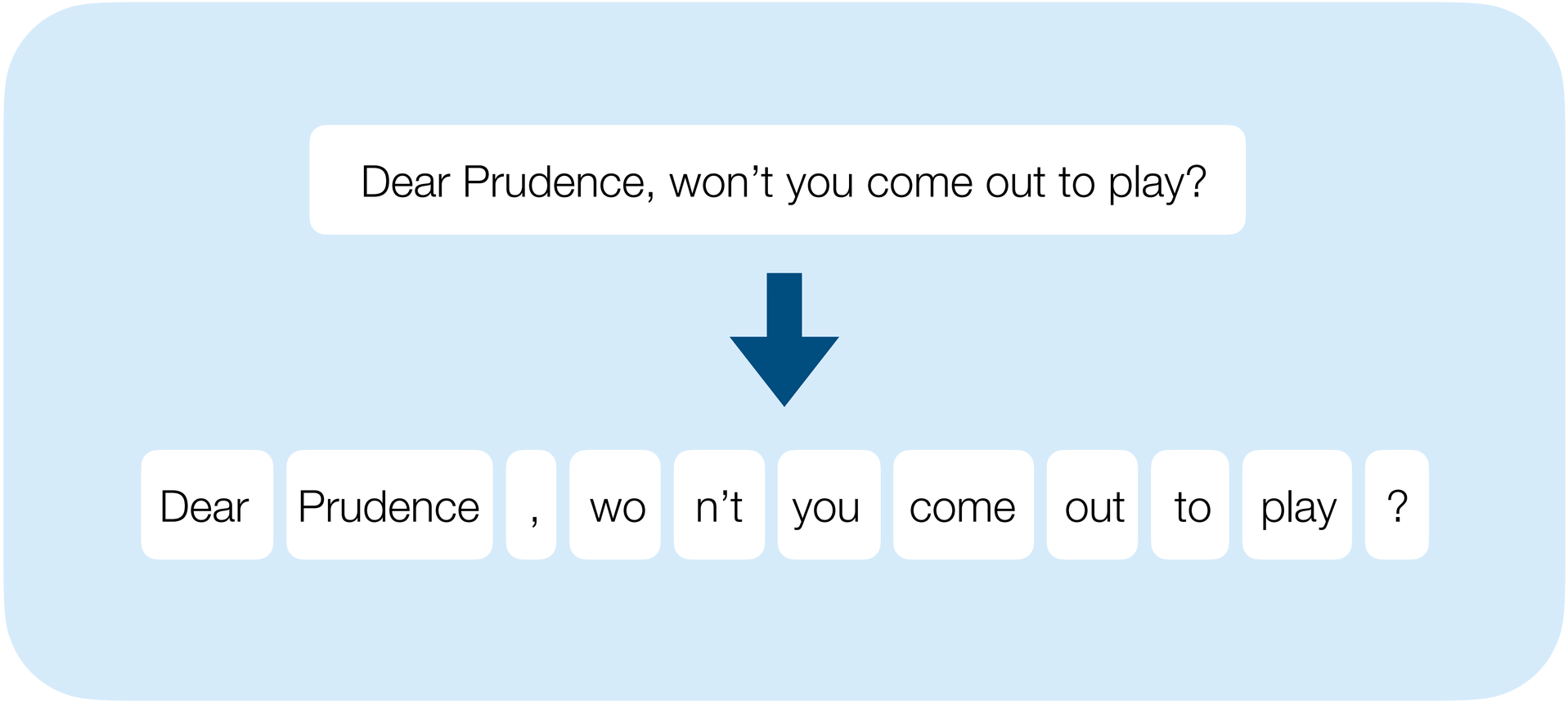
Method

spaCy library

- Open-source Python library
- Natural Language Processing with a focus on pre-processing
- Combines series of NLP functionalities in customizable pipeline

Tokenizer

Rule-based spaCy pipeline component



Problem

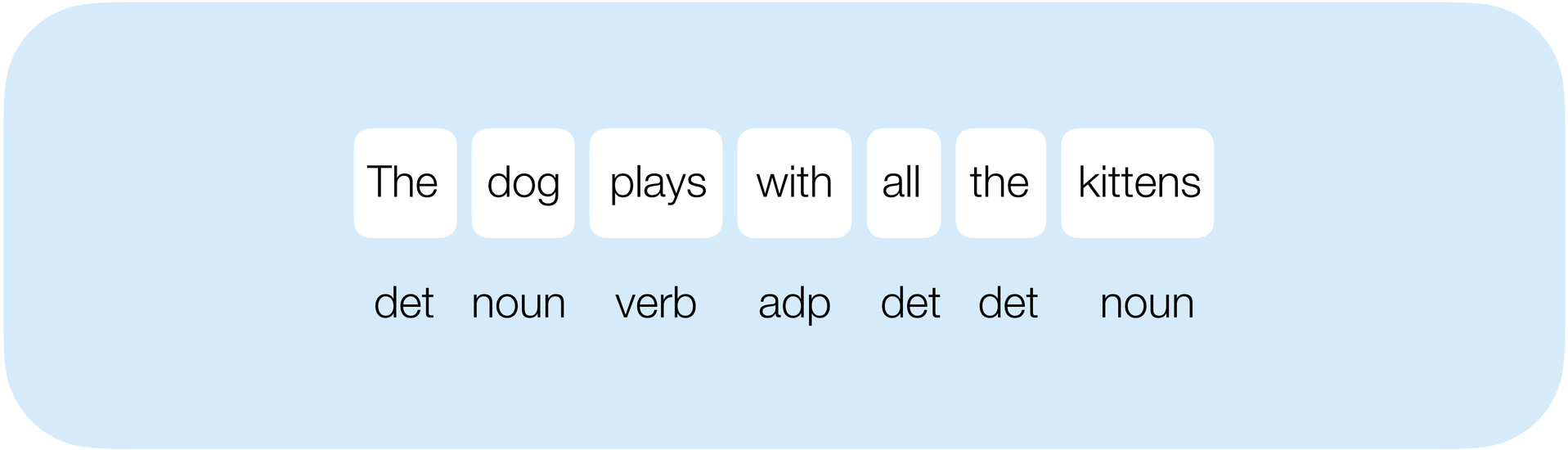
Parts of speech are difficult to identify

Noun or verb?

Dear	Prudence	,	wo	n’t	you	come	out	to	play	?

Tagger

Neural network spaCy pipeline component



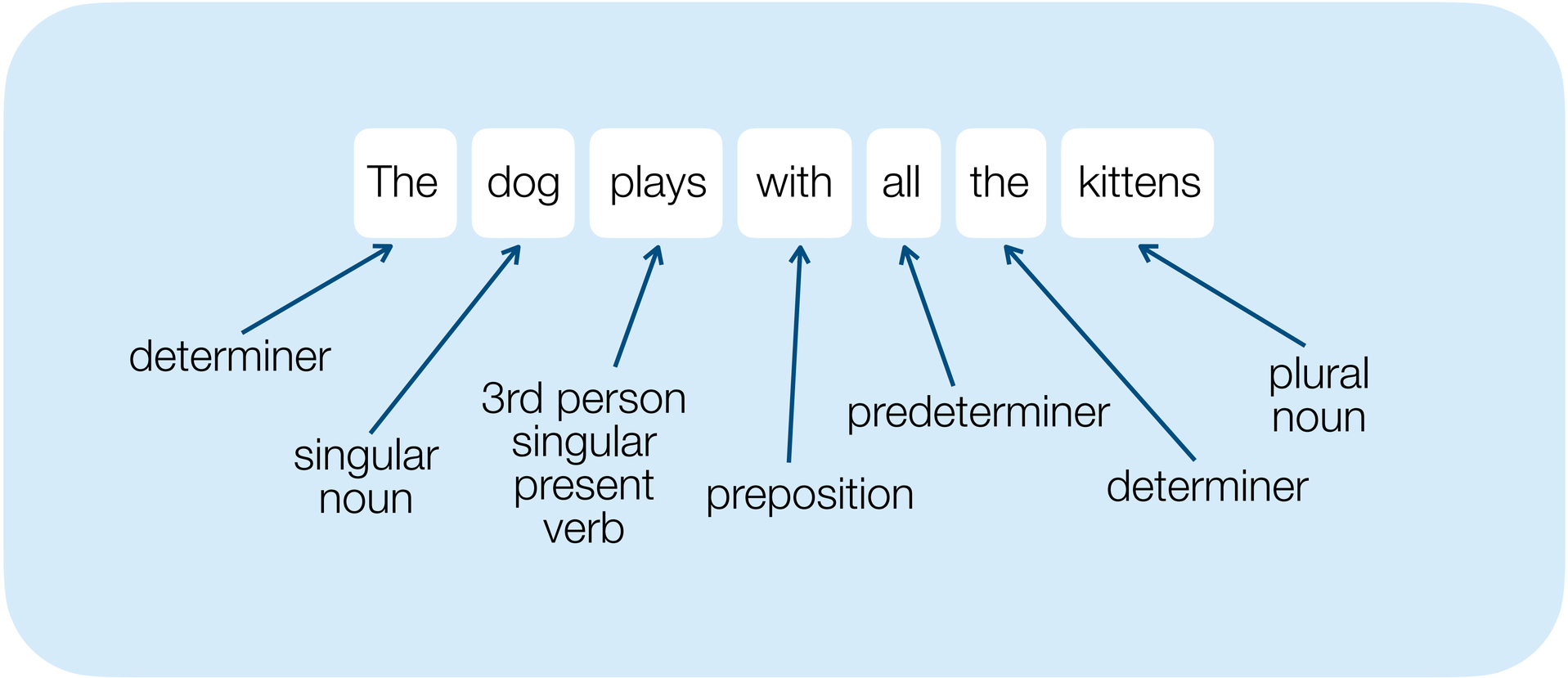
Problem

Tags do not capture inflectional properties

	The	dog	plays	with	all	the	kittens
	det	noun	verb	adp	det	det	noun
	The	dog	_____ (verb)	with	all	the	_____ (noun)
*	The	dog	**swim**	with	all	the	**piano**

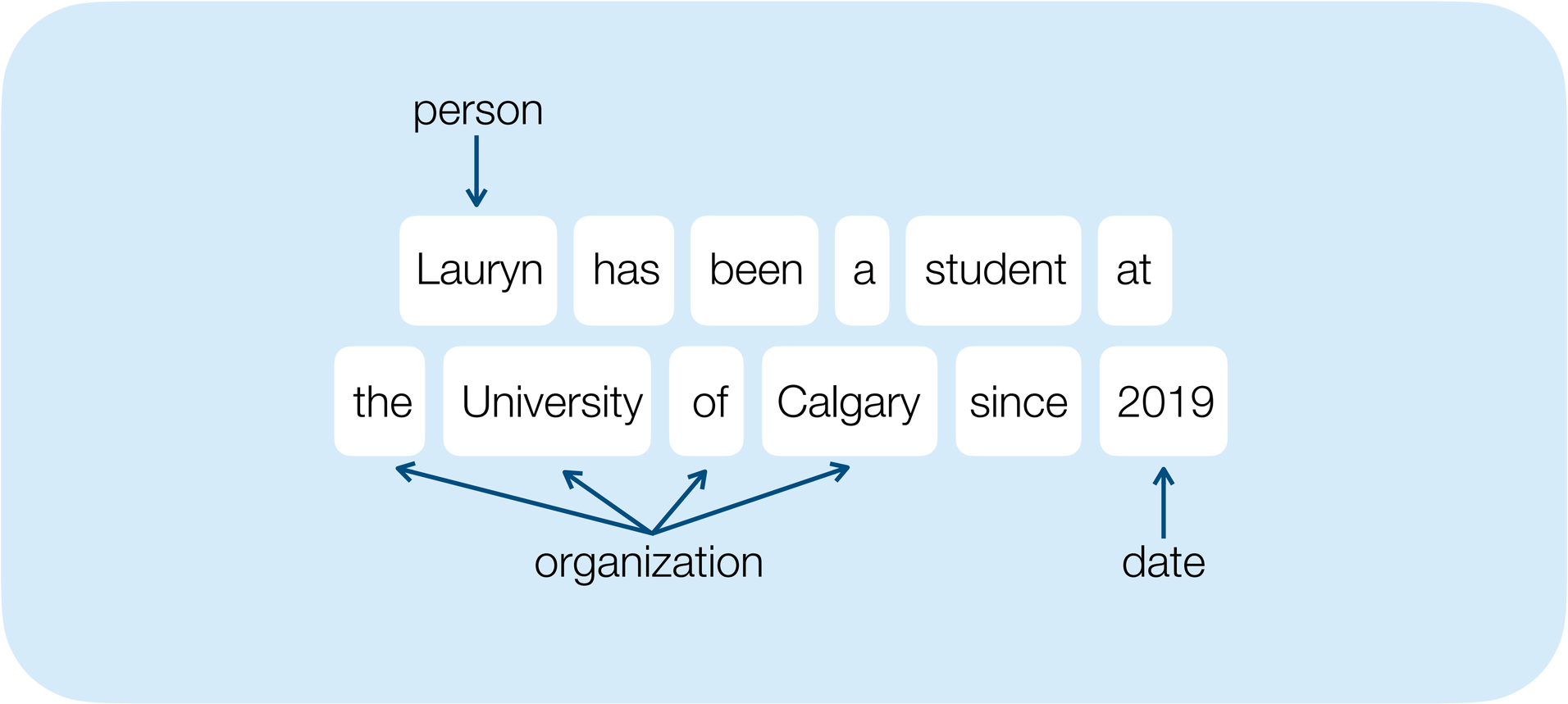
Tagger

Neural network spaCy pipeline component trained on the Penn Treebank Project dataset



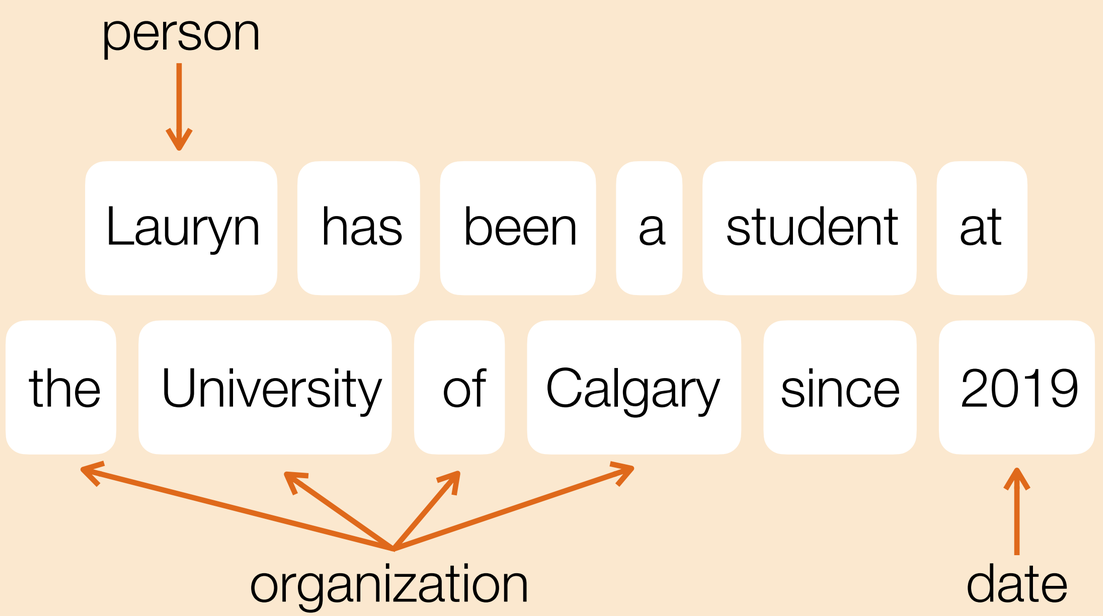
Entity Recognizer

Neural network spaCy pipeline component



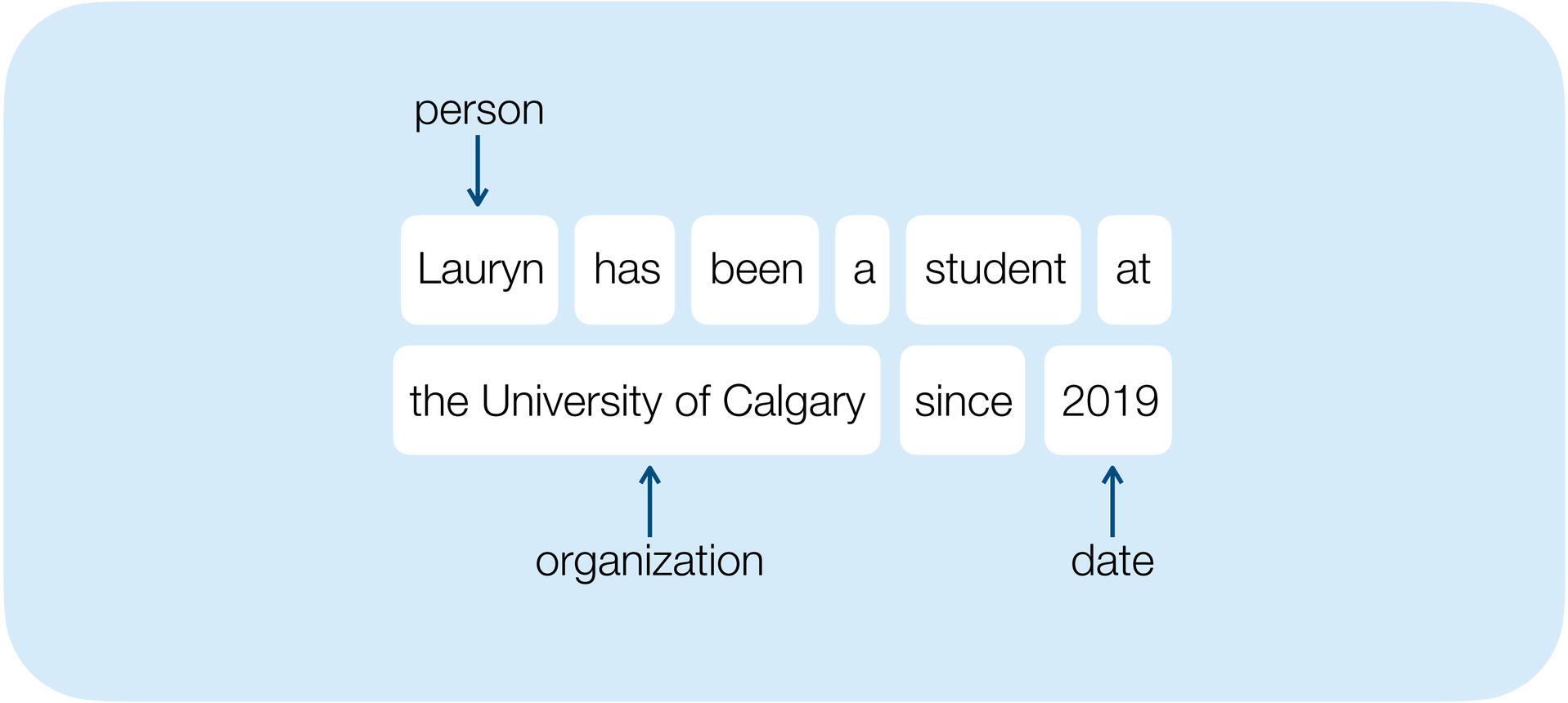
Problem

Named entities often span multiple tokens



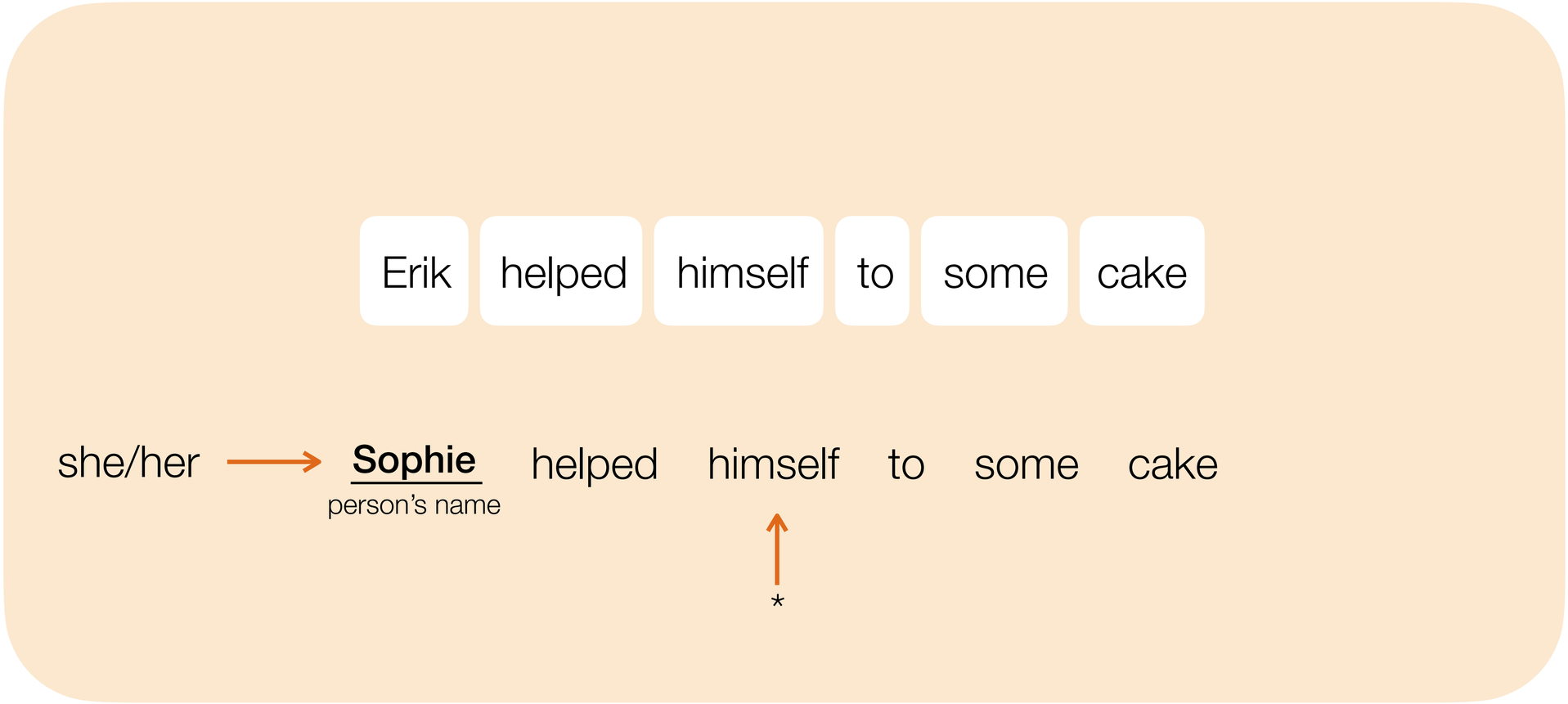
Retokenizer

spaCy content management tool



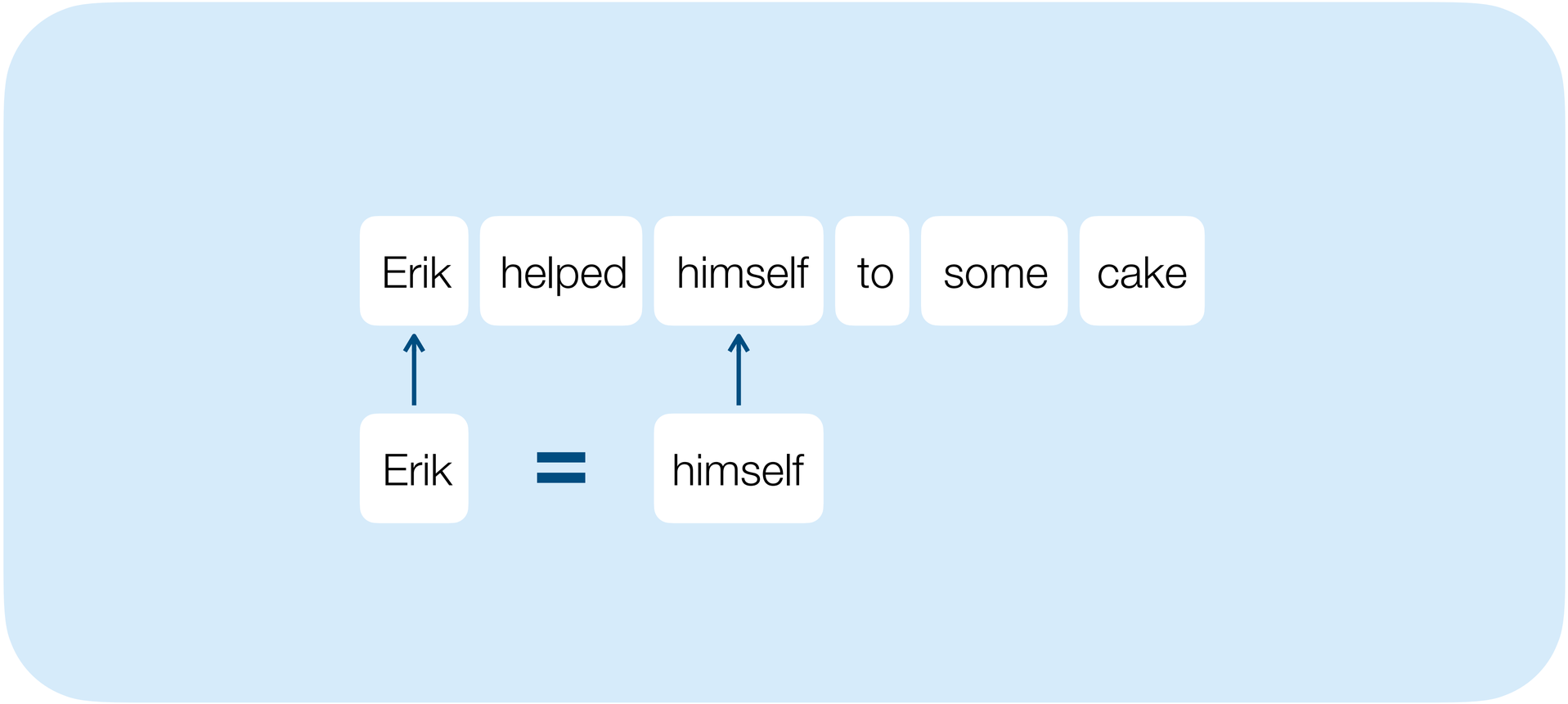
Problem

Pronoun mismatch after replacing person's name



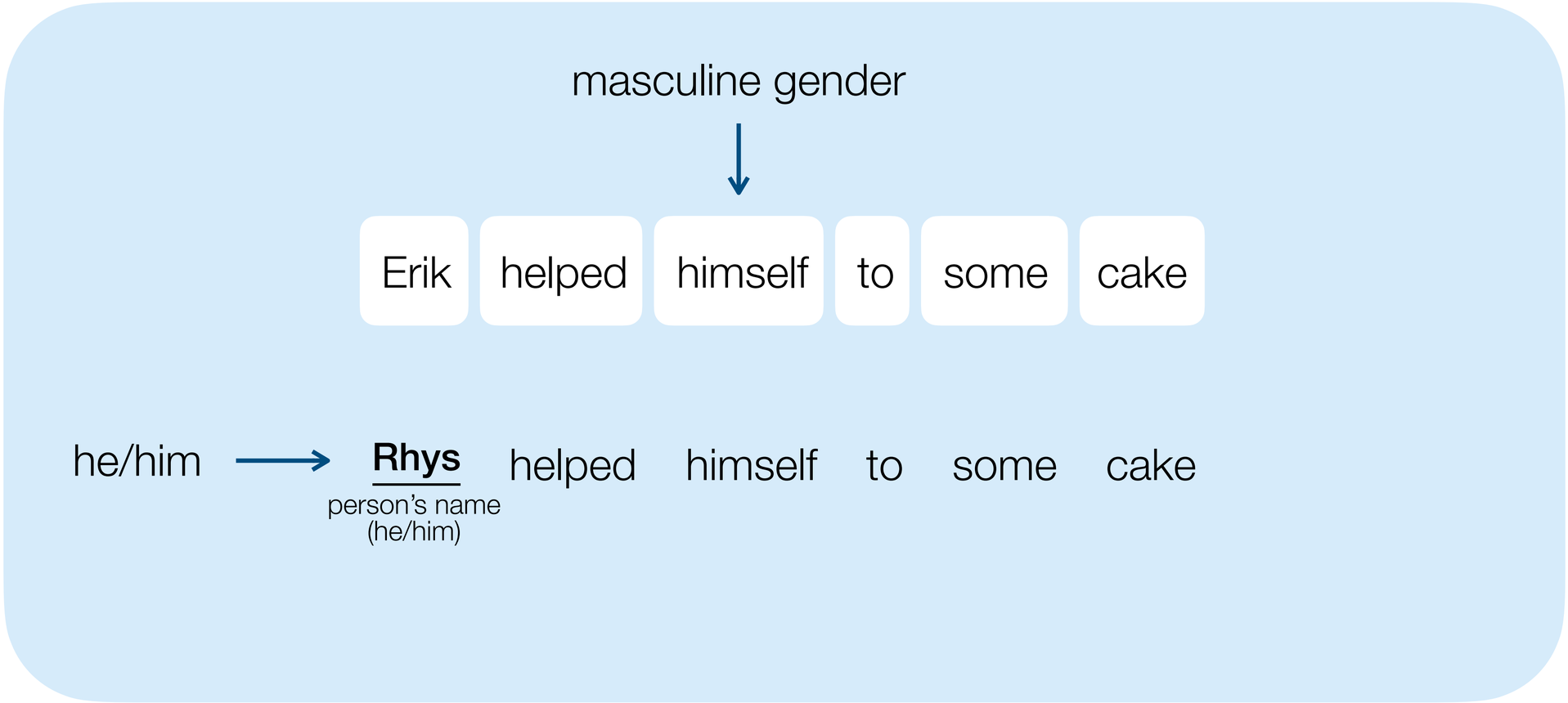
Coreference Resolution

Neural network and rule-based extension to spaCy pipeline



Morphologizer

Neural network spaCy pipeline component



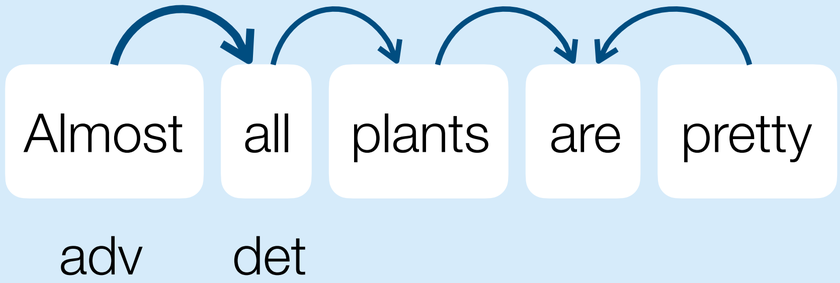
Problem

Many adverbs are not interchangeable

She **also** *ran* past the door	She **boldly** *ran* past the door
This is **more** *delicious*	This is **boldly** *delicious*
He spoke **really** *quietly*	* He spoke **boldly** *quietly*
Almost *all* plants are pretty	* **Boldly** *all* plants are pretty
We drove **nearly** *until* midnight	* We drove **boldly** *until* midnight

Dependency Parser

Neural network and rule-based spaCy pipeline component



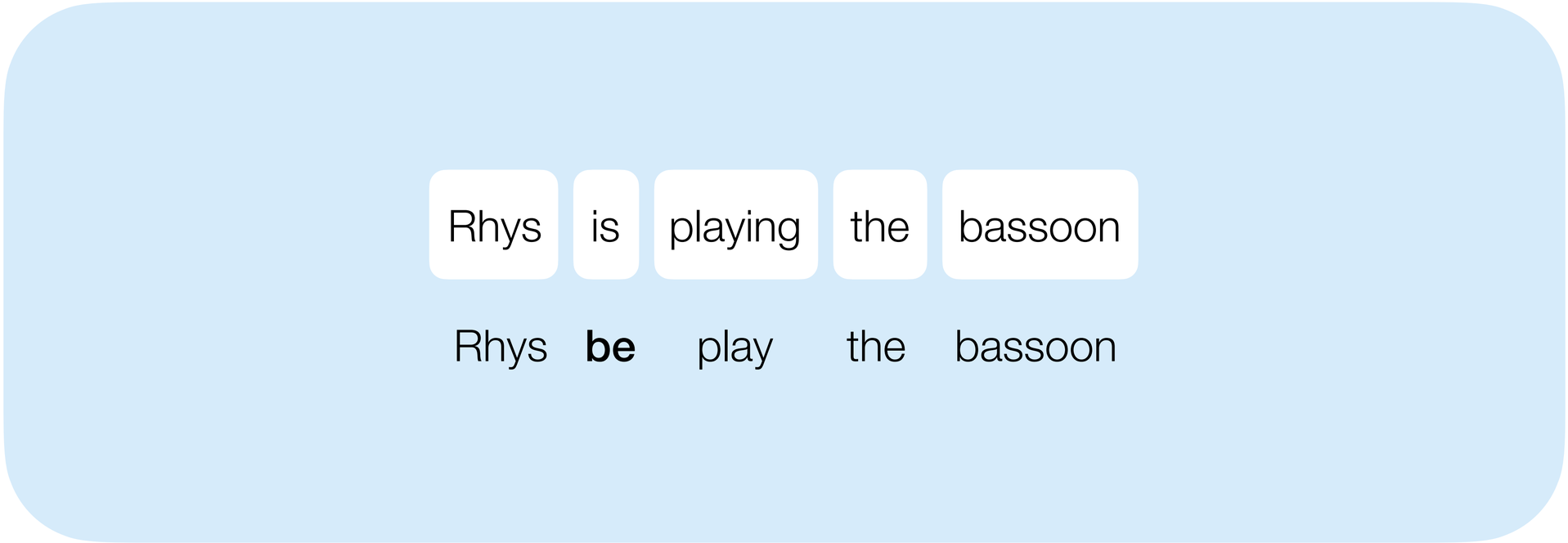
Problem

Verbs may be functional rather than meaningful

Rhys **is** playing the bassoon	* Rhys **runs** playing the bassoon
The birds **were** drawn by Erik	* The birds **ran** drawn by Erik
Sophie will **be** 19 years old	* Sophie will **run** 19 years old

Lemmatizer

Rule-based spaCy pipeline component



Discussion

Conclusion

Existing Natural Language Processing tools can be applied in a word game context to optimize the selection of words based on morphological, syntactic, and semantic criteria.

Discussion

Future Work

- Improve text processing
  - Post-processing to preserve capitalization, update articles
  - Refine verbs by valence characteristics
  - Categorize nouns by countability
- Improve user interface
- Apply existing tools to linguistic research and other text analyses

References

Explosion AI. 2015–2023. SpaCy. https://spacy.io.

Matthew Honnibal and Mark Johnson. 2015. An Improved Non-monotonic Transition System for Dependency Parsing. In *Proceedings of EMNLP 2015*. ACL, Lisbon, Portugal, 1373–1378.

Nabil Hossain, John Krumm, Lucy Vanderwende, Eric Horvitz, and Henry Kautz. 2017. Filling the Blanks (hint: plural noun) for Mad Libs Humor. In *Proceedings of EMNLP 2017*. ACL, Copenhagen, Denmark, 638–647.

Richard Hudson. 2022–2023. Coreferee. https://github.com/richardpaulhudson/coreferee.

Francisco Javier Mariño Arboleda and Verónica Elizabeth Chicaiza Redín. 2022. Mad Libs and The Parts of Speech Awareness. Bachelor's Thesis. Universidad Técnica de Ambato, Ambato, Tungurahua, Ecuador

Beatrice Santorini. 1990. Part-of-speech tagging guidelines for the penn treebank project. Technical report MS-CIS-90-47. University of Pennsylvania.