# Data

[NYC](https://drive.google.com/drive/folders/1NRnmfA6faQ4NucVR_qP9RNZJY_VXcGFW?usp=sharing): Now You’re Cooking dataset used in (Bosselut et al., 2019)

Antoine Bosselut, Omer Levy, Ari Holtzman, Corin Ennis, Dieter Fox, and Yejin Choi. 2018. Simulating Action Dynamics with Neural Process Networks. *Sixth International Conference on Learning Representations (ICLR)*, November. arXiv: 1711.05313

* Naoki annotated the dev split of the dataset to evaluate the quality.
* Very noisy (retrieved from a mailing list used in 90s)
  + There are many typos, grammatical errors, sentence segmentation errors, and non-English (e.g., URLs).
* My note: [Annotation Note](https://docs.google.com/document/d/1V1qzRqPh9Toatt5ic__DaBiEhbGbBimUXmErOZTCNNo/edit?usp=sharing)

[KitchenStories](https://drive.google.com/drive/folders/1RwSNJe726C1YuMXs4bfLUS-kP1Om7DWU?usp=sharing): Recipes sampled from <https://www.kitchenstories.com/en> **(main target)**

* We want to annotate this dataset.
* 108 recipes that are written by professional writers.
  + 1252 unique sentences.
* Include various recipe categories (pasta, dessert, etc.)
* Naoki is developing a semantic parser on this dataset, aiming to submit a paper this summer.

# Goal of treebanking

1. Develop a syntax-based semantic parser as a baseline system for my study.
   1. I was going to use an off-the-shelf parser to obtain syntax annotations, but it would be great if I could experiment with accurate annotations.
   2. I’d like to annotate at least 100/1252 sentences.
      1. More is better.
2. Develop (I think) the first UD treebank for procedural texts like a cooking recipe.
   1. Parsers trained on standard UD treebanks don’t perform well on cooking recipes. StanfordNLP dependency parser only achieved F1 scores of 0.87 and 0.66 in the identification of dobj and obl, respectively.
   2. c.f.: The research team in NYTimes made a named entity recognizer for cooking recipes. This is a very good NLP project. <https://open.blogs.nytimes.com/2015/04/09/extracting-structured-data-from-recipes-using-conditional-random-fields/>

Contact: Naoki Otani <notani@cs.cmu.edu>