Building an Annotated Dataset of Literary Entities and Events

David Bamman
School of Information, UC Berkeley
dbamman@berkeley.edu

In collaboration with Matt Sims, Jerry Park, Sejal Popat and Sheng Shen

Computational Humanities

Ted Underwood (2018), “Why Literary Time is Measured in Minutes”

Ryan Heuser, Franco Moretti, Erik Steiner (2016), The Emotions of London

Richard Jean So and Hoyt Long (2015), “Literary Pattern Recognition”

Ted Underwood, David Bamman and Sabrina Lee, The Transformation of Gender in English-Language Fiction (2018)

Franco Moretti (2005), Graphs, Maps, Trees

Holst Katsma (2014), Loudness in the Novel

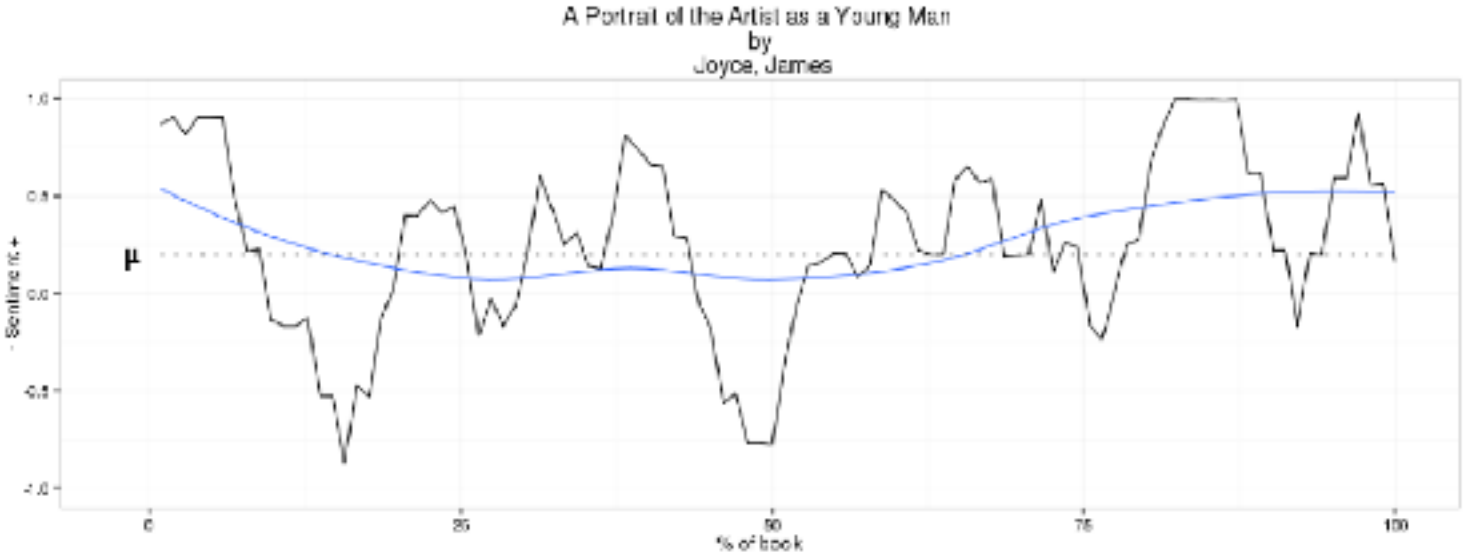
So et al (2014), “Cents and Sensibility”

Matt Wilkens (2013), “The Geographic Imagination of Civil War Era American Fiction”

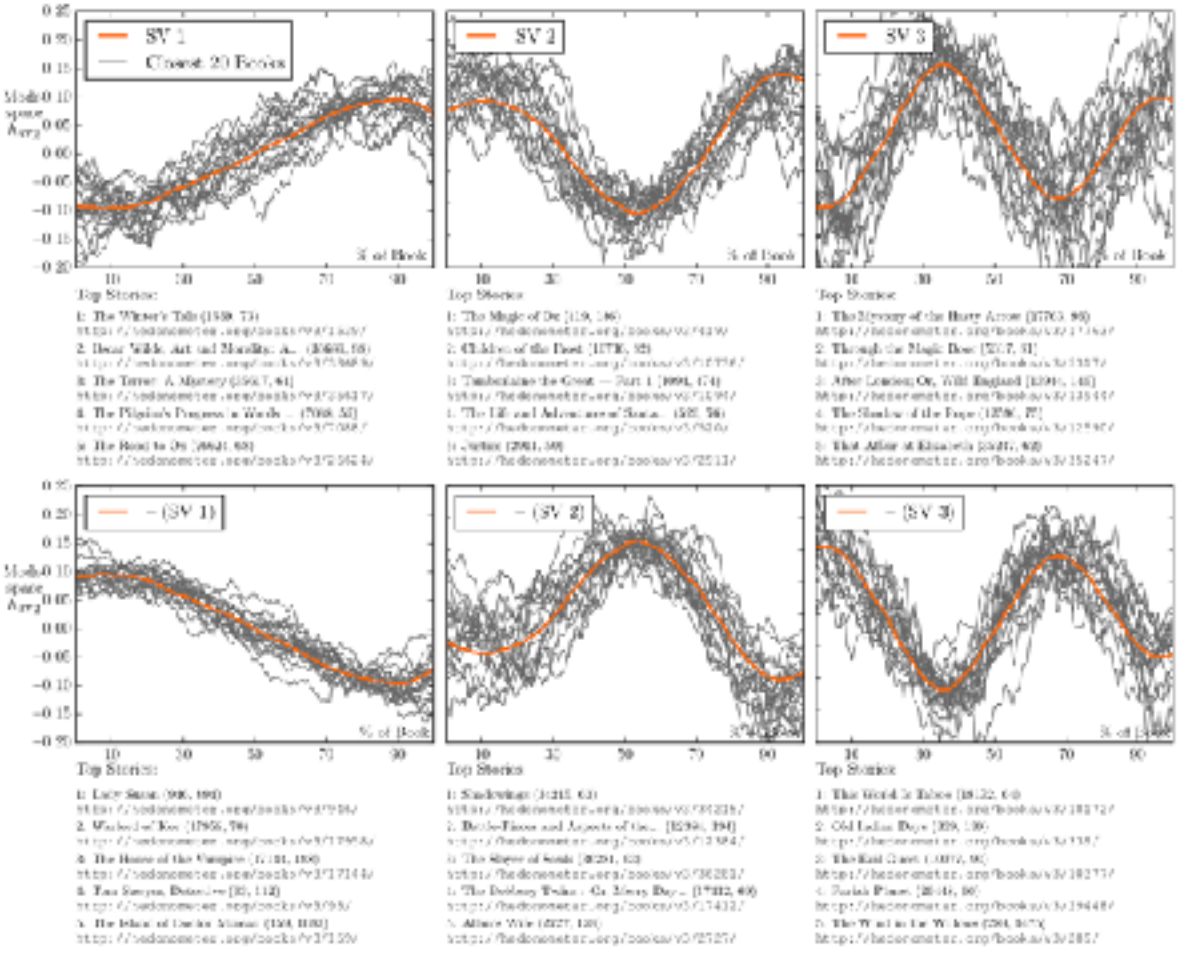
Jockers and Mimno (2013), “Significant Themes in 19th-Century Literature,”

Ted Underwood and Jordan Sellers (2012). “The Emergence of Literary Diction.” JDH

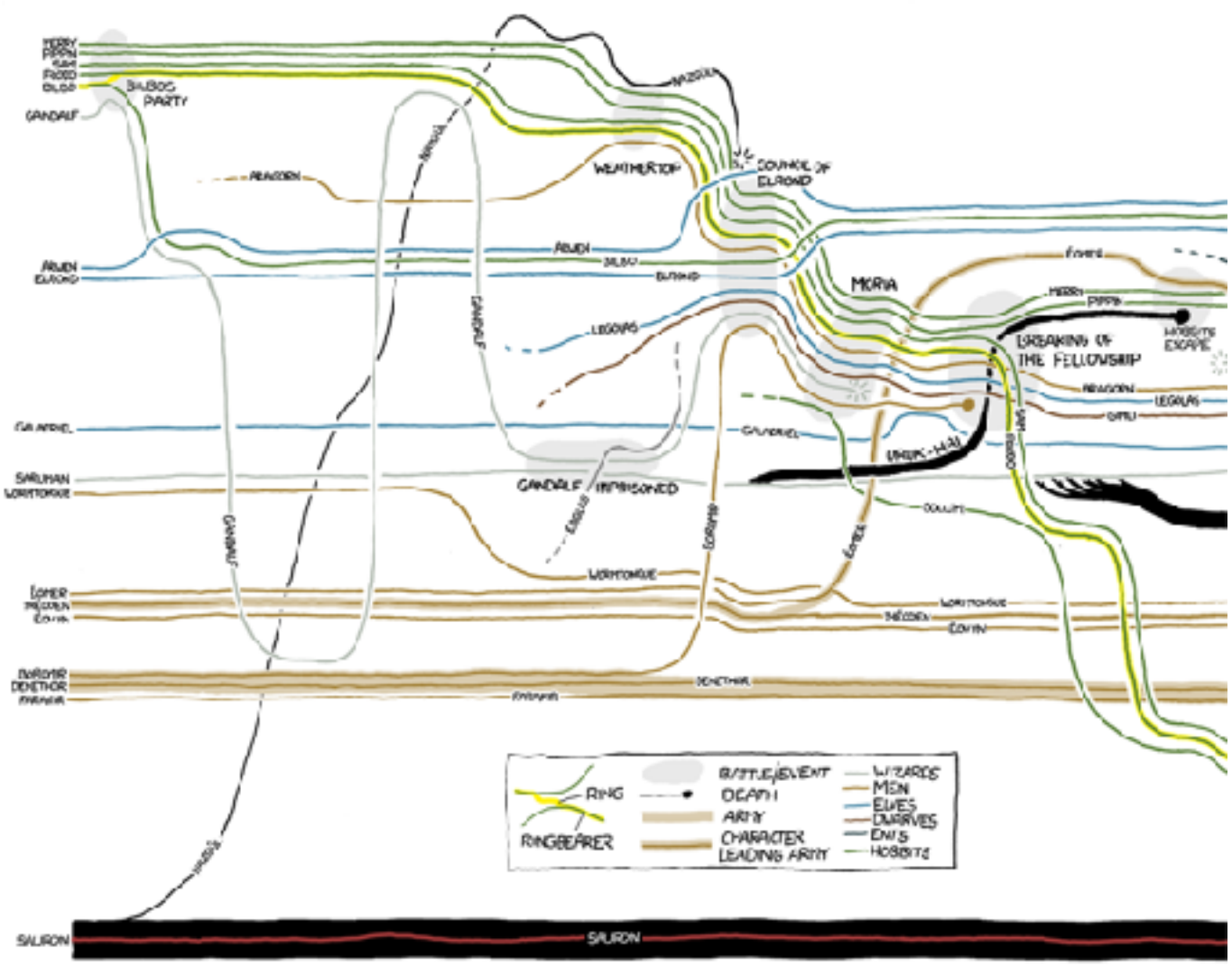
Plot



Jockers (2014), “A Novel Method for Detecting Plot”
http://www.matthewjockers.net/2014/06/05/a-novel-method-for-detecting-plot/



Reagan et al. (2016), “The emotional arcs of stories are dominated by six basic shapes”



https://xkcd.com/657/

Plot

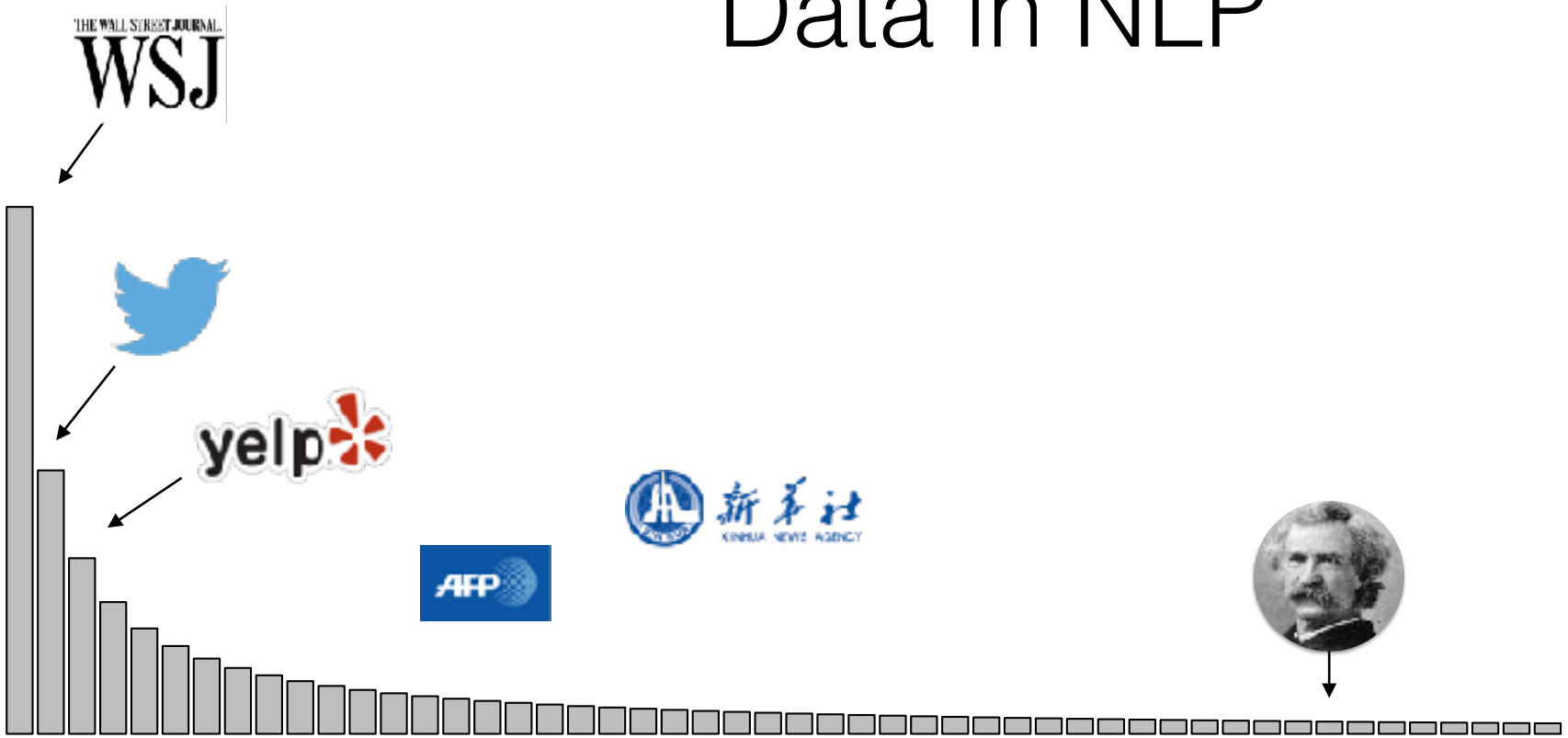
We’re decomposing plot into structured elements

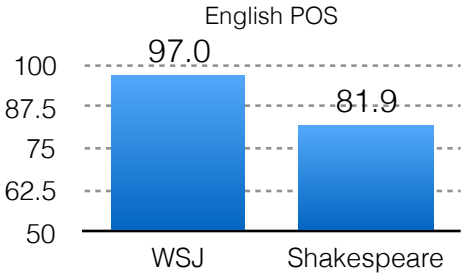
Element	Task
Characters	Entity recognition
Events	Event detection
Setting	Entity recognition, setting coreference
Objects	Object detection/coreference
Time	Temporal processing, event ordering

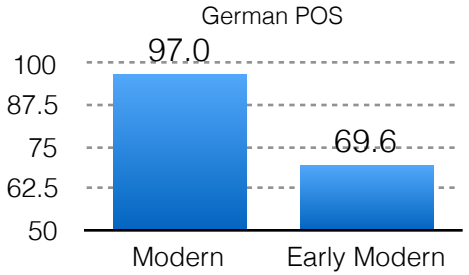
NLP Pipeline

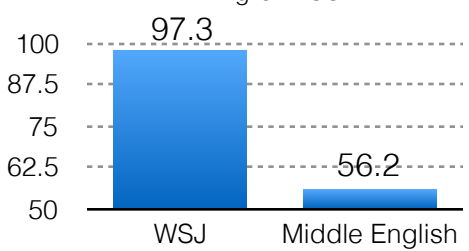
NLP Task	Accuracy
Tokenization	100%
Part-of-speech tagging	98.0% [Bohnet et al. 2018]
Named entity recognition	93.1 [Akbik et al. 2018]
Syntactic parsing	95.1 F [Kitaev and Klein 2018]
Coreference resolution	73.0 F [Lee et al. 2018]

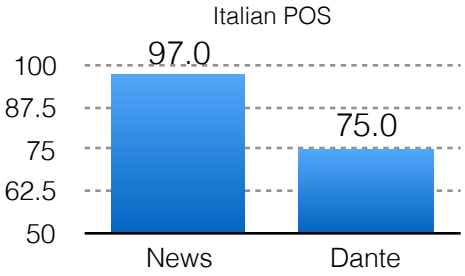
Data in NLP

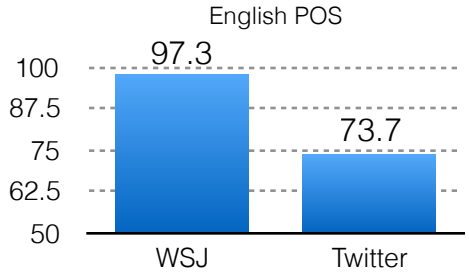


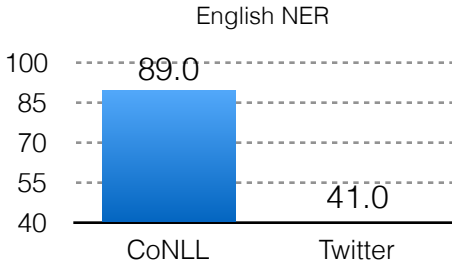

English POS
100
87.5
75
62.5
50
97.0
81.9
WSJ
Shakespeare

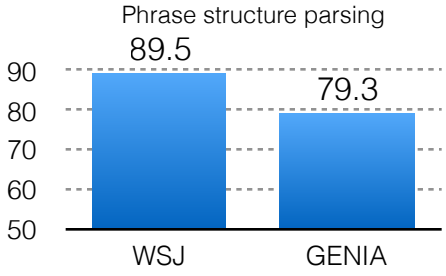

German POS
100
87.5
75
62.5
50
97.0
69.6
Modern
Early Modern

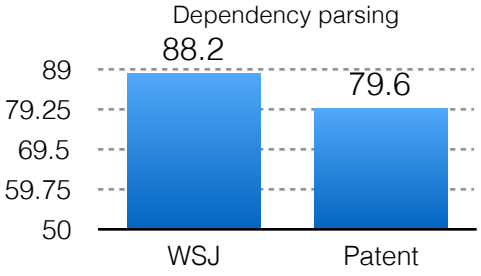

English POS
100
87.5
75
62.5
50
97.3
56.2
WSJ
Middle English

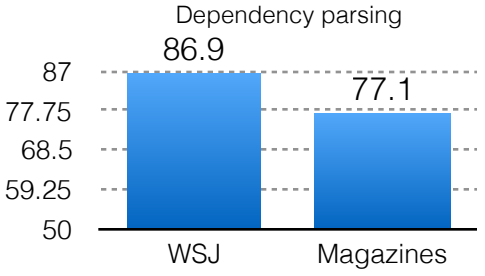

Italian POS
100
87.5
75
62.5
50
97.0
75.0
News
Dante


English POS
100
87.5
75
62.5
50
97.3
73.7
WSJ
Twitter

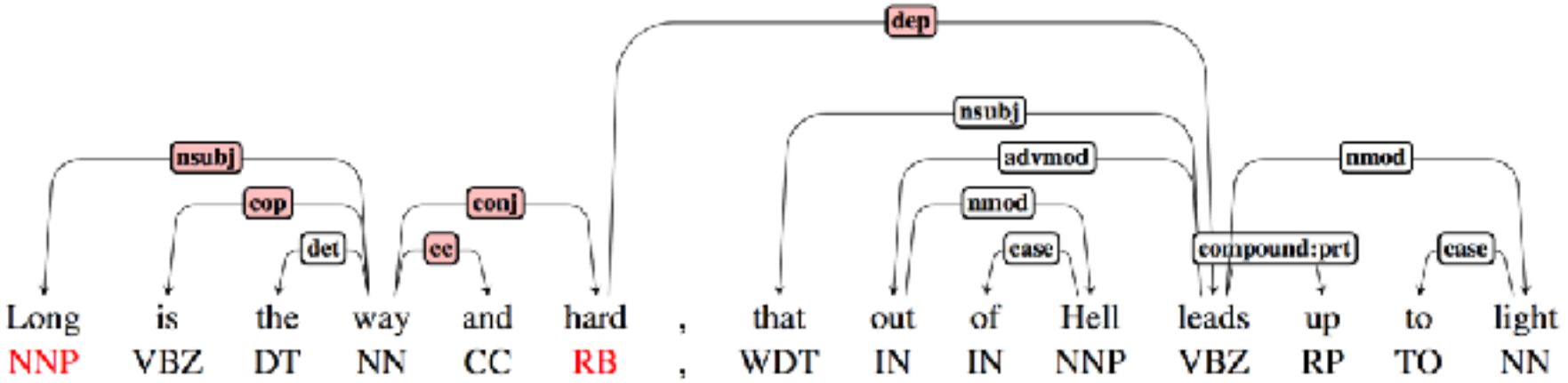

English NER
100
85
70
55
40
89.0
41.0
CoNLL
Twitter


Phrase structure parsing
90
80
70
60
50
89.5
79.3
WSJ
GENIA


Dependency parsing
89
79.25
69.5
59.75
50
88.2
79.6
WSJ
Patent


Dependency parsing
87
77.75
68.5
59.25
50
86.9
77.1
WSJ
Magazines

Syntax


dep
nsubj
nsubj
advmod
nmod
cop
conj
nmod
det
cc
case
compound:prt
case
Long is the way and hard , that out of Hell leads up to light
NNP VBZ DT NN CC RB , WDT IN IN NNP VBZ RP TO NN

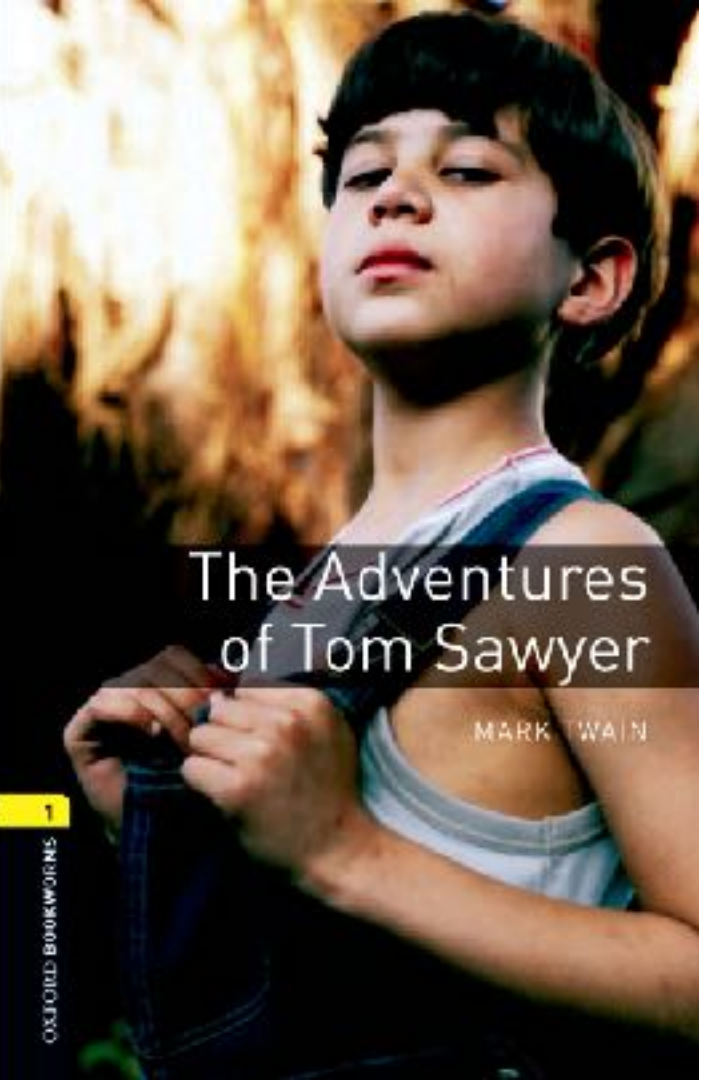
Active work

- Domain adaptation
  [Chelba and Acero, 2006; Daumé and Marcu, 2006; Daumé 2009; Duong et al. 2015; Glorot et al. 2011, Chen et al. 2012, Yang and Eisenstein 2014, Schnabel and Schütz 2014]

- Contextualized word representations
  [Peters et al. 2018; Devlin et al. 2018; Howard and Ruder 2018; Radford et al. 2019]

- Data annotation. 200,000 tokens from 100 different novels, annotated for:

  - Entities (person/place, etc.)
  - Events
  - Coreference
  - Quotation attribution



Literary entities

"TOM!"

No answer.

"TOM!"

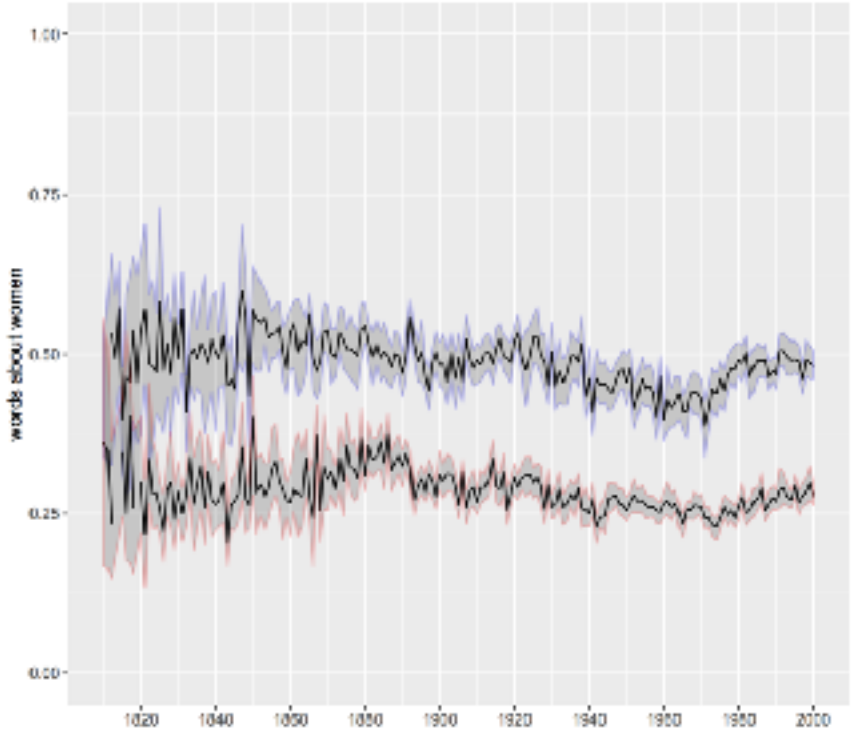
No answer.

"What's gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.

Literary entities



written by women

written by men

Underwood, Bamman and Lee (2018), "The Transformation of Gender in English-Language Fiction"

Literary entities

Most work in NLP focuses on *named* entity recognition — mentions of specific categories (person, place, organization) that are explicitly named.

> Mr. Knightley, a sensible man about seven or eight-and-thirty, was not only a very old and intimate friend of the family, but particularly connected with it, as the elder brother of Isabella's husband.

Austen, *Emma*

Entity recognition

- Mr. Knightley
- a sensible man about seven or eight-and-thirty
- a very old and intimate friend of the family
- the family
- Isabella
- Isabella's husband
- the elder brother of Isabella's husband

Mr. Knightley, a sensible man about seven or eight-and-thirty, was not only a very old and intimate friend of the family, but particularly connected with it, as the elder brother of Isabella's husband.

Austen, *Emma*

Entity recognition

- Mr. Knightley
- a sensible man about seven or eight-and-thirty
- a very old and intimate friend of the family
- the elder brother of Isabella's husband

- the family

- Isabella

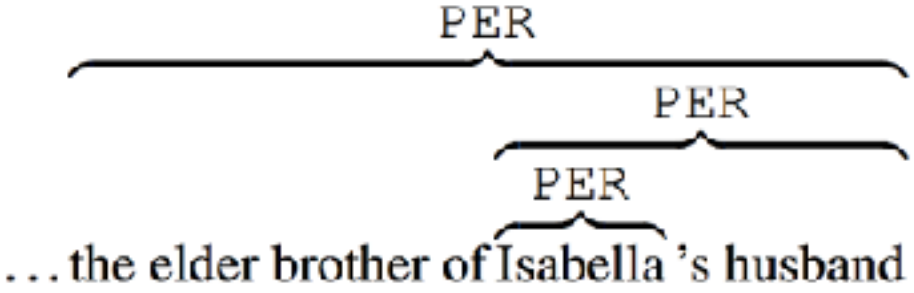
- Isabella's husband

Mr. Knightley, a sensible man about seven or eight-and-thirty, was not only a very old and intimate friend of the family, but particularly connected with it, as the elder brother of Isabella's husband.

Austen, *Emma*

Nested entity recognition

- Recognize spans of text that correspond to categories of entities (whether named or not).



Dataset

- 100 books from Project Gutenberg
- Mix of high literary style (e.g., Edith Wharton's *Age of Innocence,* James Joyce's *Ulysses)* and popular pulp (Haggard's *King Solomon's Mines*, Alger's *Ragged Dick).*
- Select first 2000 words from each text

Entity classes

- Person. Single person with proper name (Tom Sawyer) or common entity (the boy); set of people (her daughters).
- Organization. Formal association (the army, the Church as an administrative entity).
- Vehicle. Devices primarily designed to move an object from one location to another (ships, trains, carriages).

Entity classes

- GPE. Entities that contain a population, government, physical location and political boundaries (New York, the village)
- Location. Entities with physicality but w/o political status (New England, the South, Mars), including natural settings (the country, the valley, the forest)
- Facility. Functional, primarily built structure designed for habitation (buildings), storage (barns), transportation (streets) and maintained outdoor space (gardens).

Metaphor

- Only annotate phrases whose types denotes an entity class.

PER PER

John is a doctor

PER PER ???

the young man was not really a poet; but surely he was a poem

Chesterton, *The Man Who Was Thursday*

Personification

- Person includes characters who engage in dialogue or have reported internal monologue, regardless of human status (includes aliens and robots as well).

> As soon as I was old enough to eat grass my mother used to go out to work in the daytime, and come back in the evening.

Sewell, *Black Beauty*

Metonymy

- Describing one concept by a closely related one (e.g., “the White House said…”)
- Annotate the evoked entity class (PER here rather than FAC).

“Them men would eat and drink if we was all in our graves,” said the indignant cook, who indeed had a real grievance; and the outraged sentiment of the kitchen was avenged by a bad and hasty dinner

Oliphant, *Miss Marjoribanks*

Data

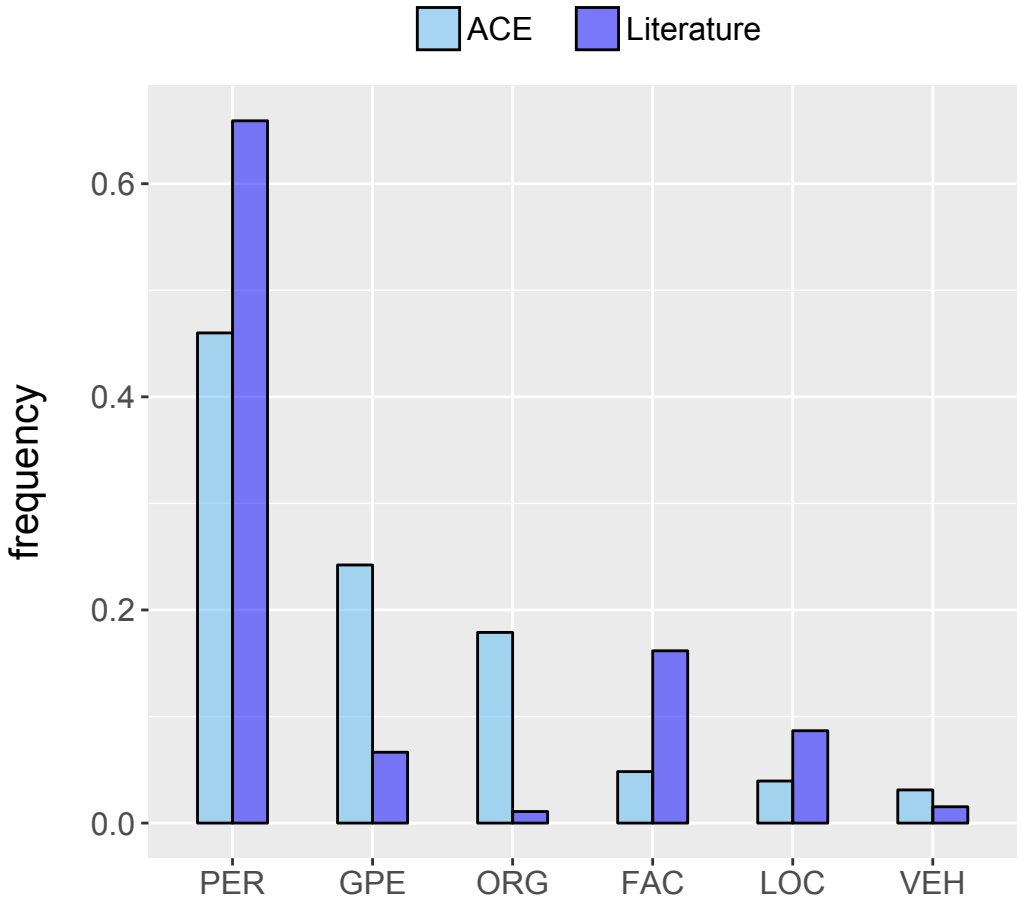
Cat	Count	Examples
PER	9,383	my mother, Jarndyce, the doctor, a fool, his companion
FAC	2,154	the house, the room, the gardne, the drawing-room, the library
LOC	1,170	the sea, the river, the country, the woods, the forest
GPE	878	London, England, the town, New York, the village
VEH	197	the ship, the car, the train, the boat, the carriage
ORG	130	the army, the Order of Elks, the Church, Blodgett College

Prediction

How well can find these entity mentions in text as a function of the training domain?

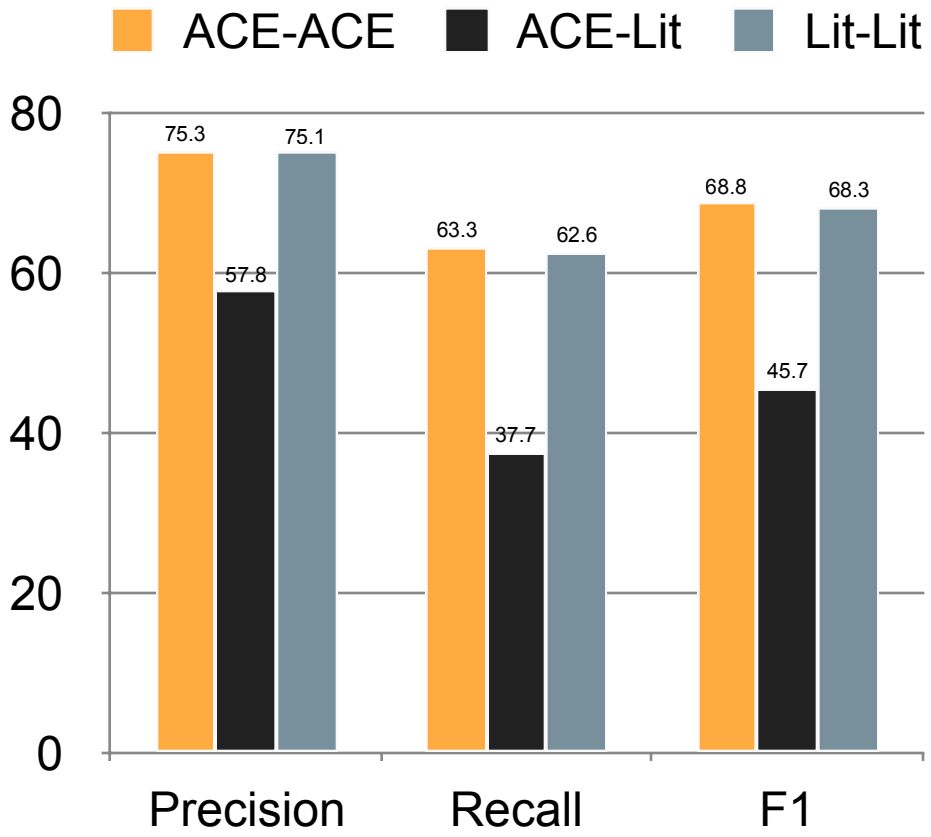
Data

- ACE (2005) data from newswire, broadcast news, broadcast conversation, weblogs



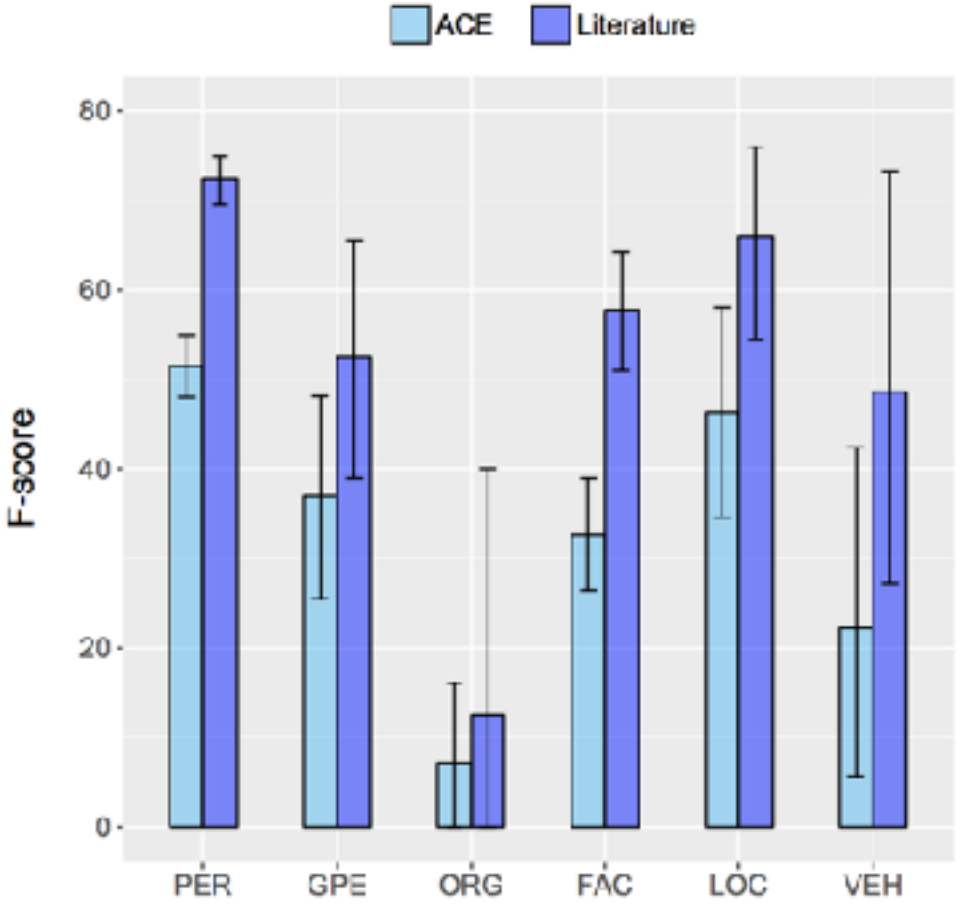
Prediction

- Ju et al. (2018): layered BiLSTM-CRF; state-of-the-art on ACE 2005.
- Evaluate performance difference when altering the training/test domain.



Prediction

- F1 by category (train on ACE/Literature) and test on Literature.



Analysis

- Tag entities in 1000 new Gutenberg texts (78M tokens) using the two models (ACE vs. LIT) and analyze the difference in frequencies with which a given string is tagged as `PER` under both models.

Mrs.
Miss
Lady
Aunt

MOSCOW, April 17 (AFP)

Silence is golden -- especially when your hand is weak -- top Moscow policy analysts said in an assessment of the fallout from Russia's vocal opposition to what turned out to be a swift US-led campaign in Iraq.

Several top diplomacy experts told a Kremlin-run forum that countries like China and India that said little about the conflict before its March 20 launch were already reaping the benefits.

Some suggested that Russian President Vladimir Putin will now be scrambling to contain the damage to his once-budding friendship with US President George W. Bush because he was poorly advised by his intelligence and defense aides.

AFP_ENG_20030417.0307

Chapter I: The Bertolini

“The Signora had no business to do it,” said Miss Bartlett, “no business at all. She promised us south rooms with a view close together, instead of which here are north rooms, looking into a courtyard, and a long way apart. Oh, Lucy!”

“And a Cockney, besides!” said Lucy, who had been further saddened by the Signora’s unexpected accent. “It might be London.”

Forster, *A Room with a View*

Analysis

- How well does each model identify entities who are men and women?
- We annotate the gender for all PER entities in the literary test data and measure the recall of each model with respect to those entities.

Training	Women	Men	Diff
ACE	38.0	49.6	-11.6
Literary	69.3	68.2	1.1

https://github.com/dbamman/litbank

README.md

LitBank

LitBank is an annotated dataset of 100 works of fiction to support tasks in natural language processing and the computational humanities, described in more detail in:

David Bamman, Sejal Popat and Sheng Shen, "An Annotated Dataset of Literary Entities", NAACL 2019.

LitBank is licensed under a Creative Commons Attribution 4.0 International License.

Events

- Event trigger detection, slot filling
  [MUC, ACE, DEFT ERE]

- Veridicality, factuality & committed belief
  [Saurí and Pustejovksy 2009; de Marneffe et al. 2012, Werner et al. 2015, Lee et al. 2015, Rudinger et al. 2018]

- Temporal grounding
  [Pustejovsky et al. 2003]

- Narrative event chains, schemas
  [Chambers and Jurafsky 2008, Cheung et al. 2013]

Events

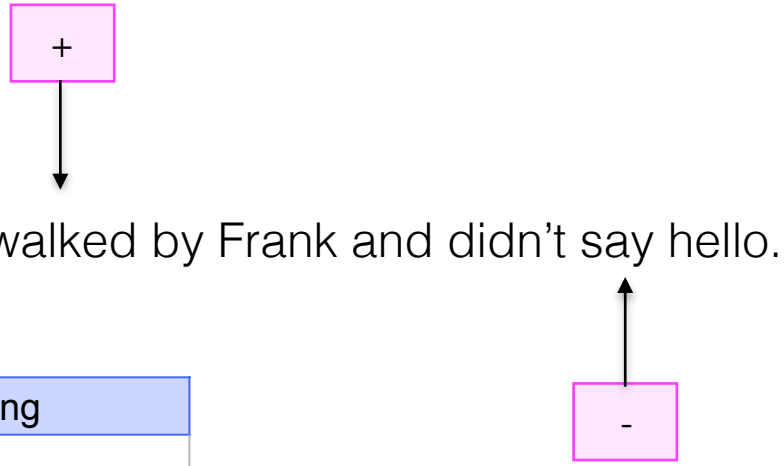
Realist view: events are things that happen.

Aristotle, Russell, Whitehead, Quine, Vendler, Montague, Davidson, Dowty, etc.

Events in language

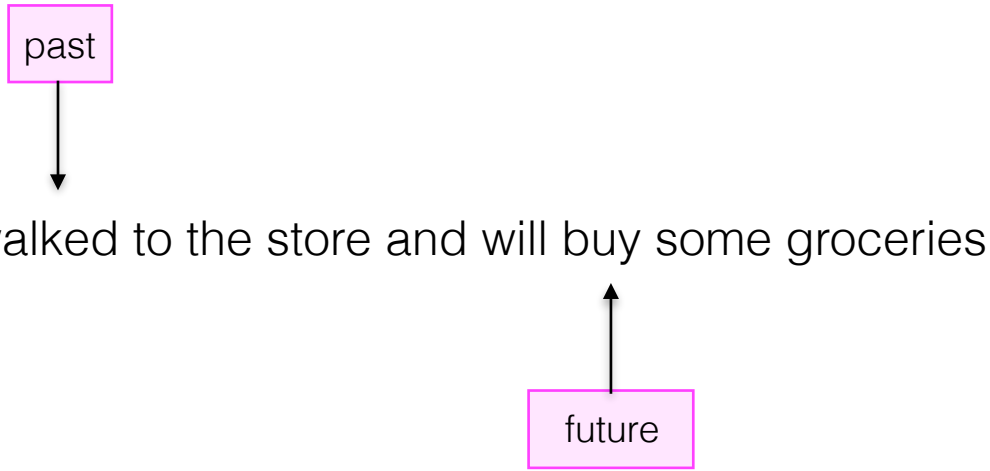
- Events are typically realized through verbs (and some nominalizations)
    - He walked down the street
    - He had a nice walk.

Polarity



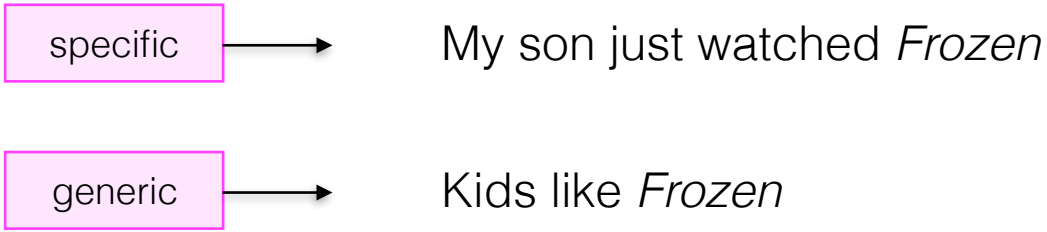
Value	Meaning
positive	depicted as taking place
negative	depicted as not taking place

Tense



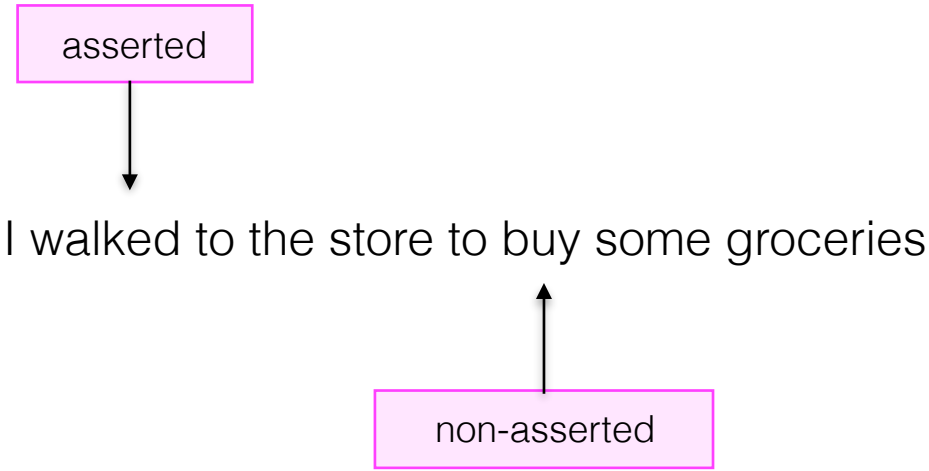
Value
past
present
future

Specificity



Value	Meaning
specific	singular occurrence at a particular place and time
general	claim about groups, abstractions

Modality



Value
asserted
non-asserted

Modality

Beliefs: Rumors of my demise have been greatly exaggerated

Hypotheticals: If you visit Rockridge, try Zachary's.

Commands: John was ordered to return the book or face a fine.

Threats: AMI threatened to release the photos if Bezos did not comply.

Desires: She wants to go to Rome.

Dataset

- 100 books from Project Gutenberg (same as for entity annotations)
- First 2000 words from each text
- 7,892 events

My father’s eyes had closed upon the light of this world six months, when mine opened on it.

Dickens, *David Copperfield*

Call me Ishmael

Melville, *Moby Dick*

Event detection

Can we detect events that have occurred?

0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0

My father's eyes had closed upon the light of this world six months, when mine opened on it.

Dickens, *David Copperfield*

0 0 0

Call me Ishmael

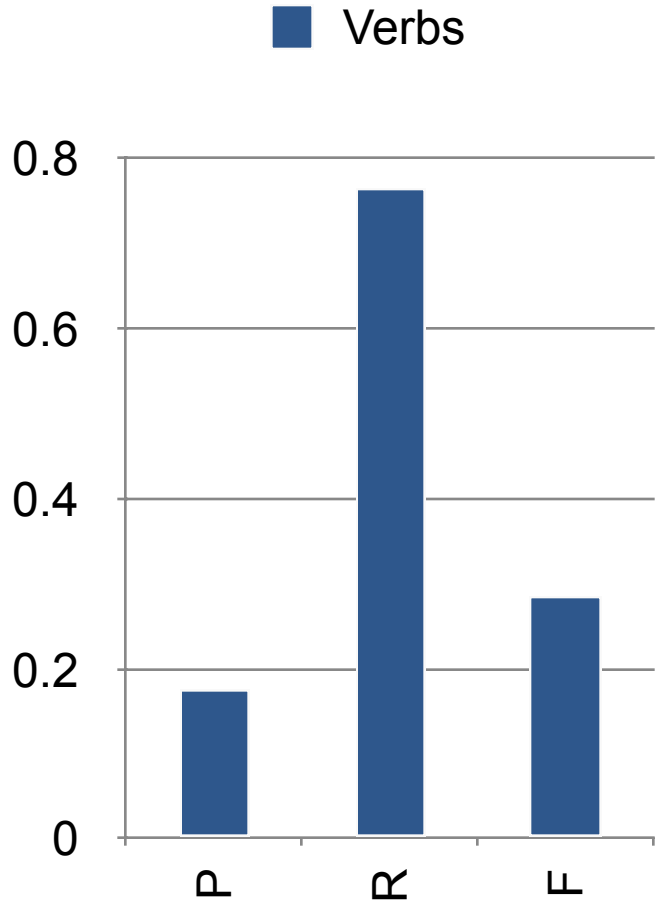
Melville, *Moby Dick*

Baseline

All (and only) verbs are events

VB PRP NNP

Call me Ishmael



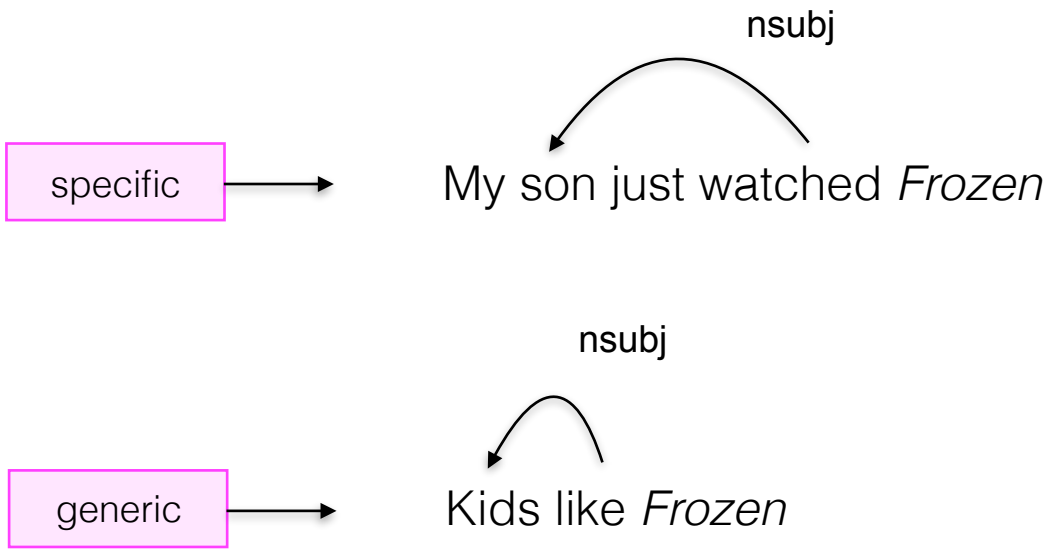
Featurized model

“Call me Ishmael”

Feature	Examples
Word identity	Call
Part of speech	VB
Context	L:∅, R:“me Ishmael”, L_POS:∅, R:“PRP NNP”
Wordnet synset	[name, call]
Dependency label + head	root

Featurized model

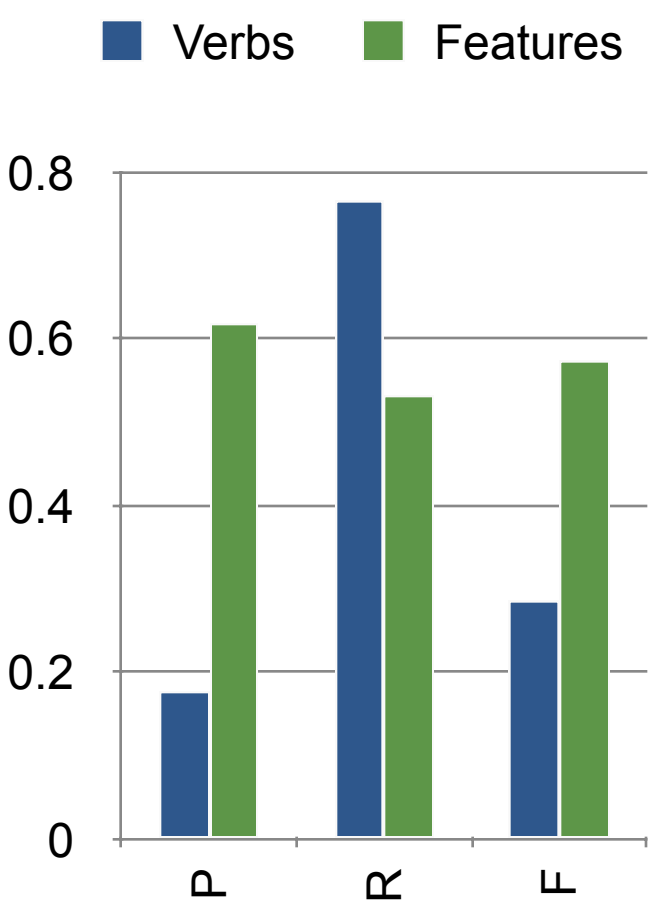
- Syntactic information: is subject a bare plural?
- Is the subject countable?



Reiter and Frank (2010)

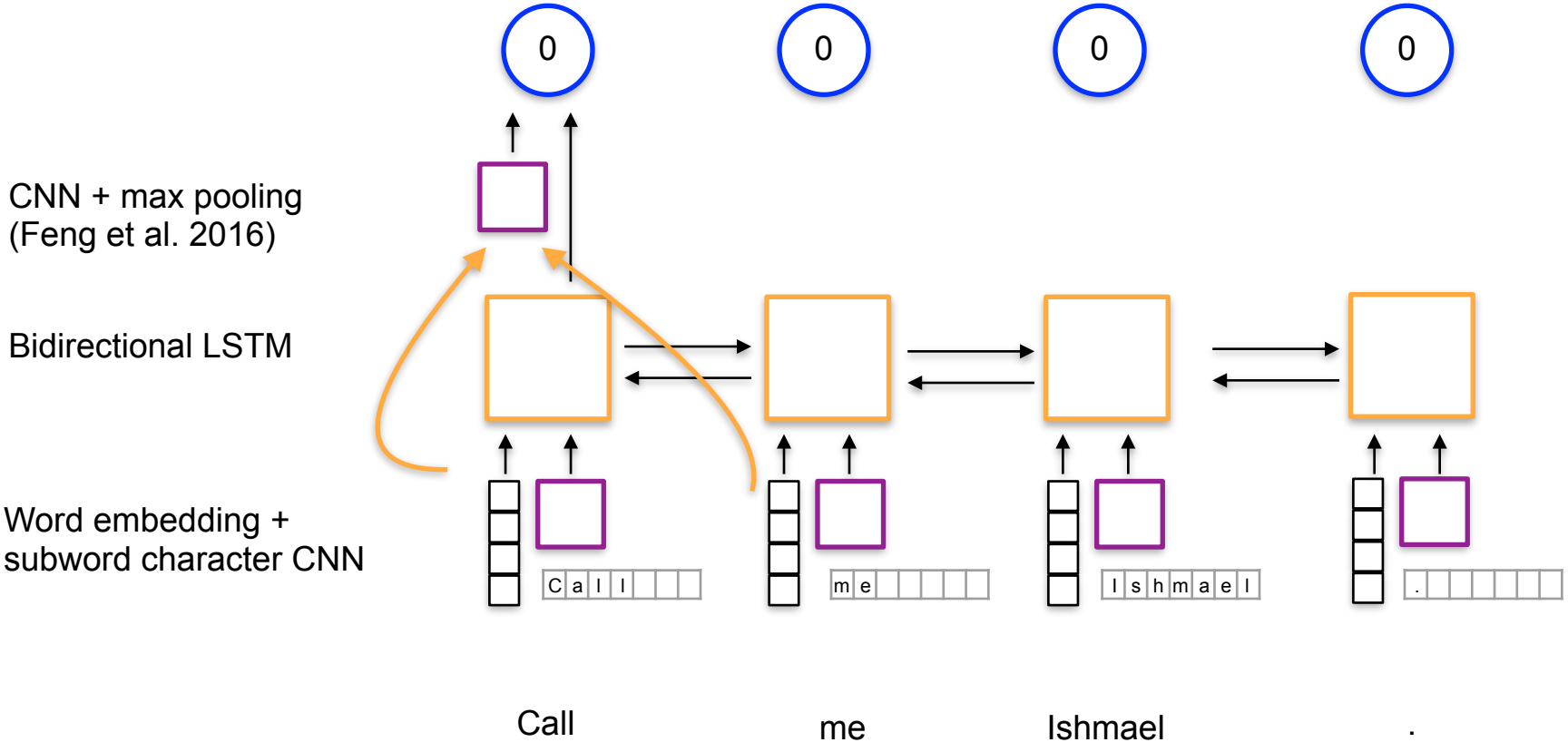
Featurized model

L2-regularized logistic regression



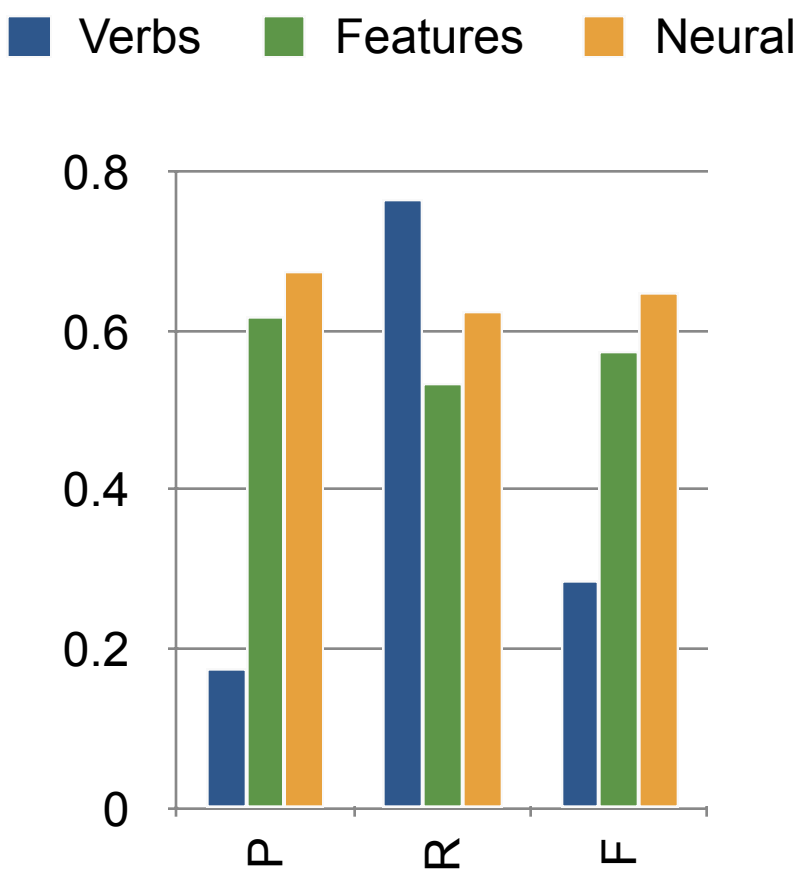
Neural

- Bidirectional LSTM + CNN (with subword CNN)
- Word embeddings learned from 15,000 Project Gutenberg texts
- Sentence-level CNN (Feng et al. 2016)


CNN + max pooling
(Feng et al. 2016)
Bidirectional LSTM
Word embedding +
subword character CNN
0
0
0
0
C a l l
m e
I s h m a e l
.
Call
me
Ishmael
.

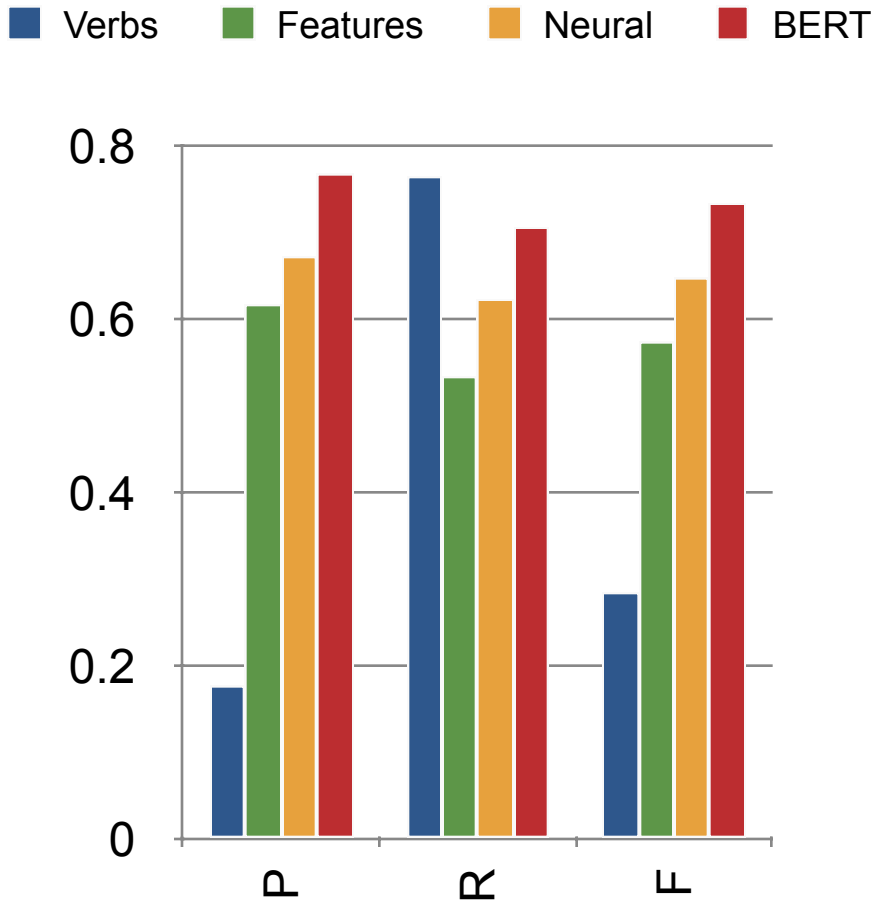
Neural

- Bidirectional LSTM + CNN (with subword CNN)
- Word embeddings learned from 15,000 Project Gutenberg texts
- Sentence-level CNN (Feng et al. 2016)



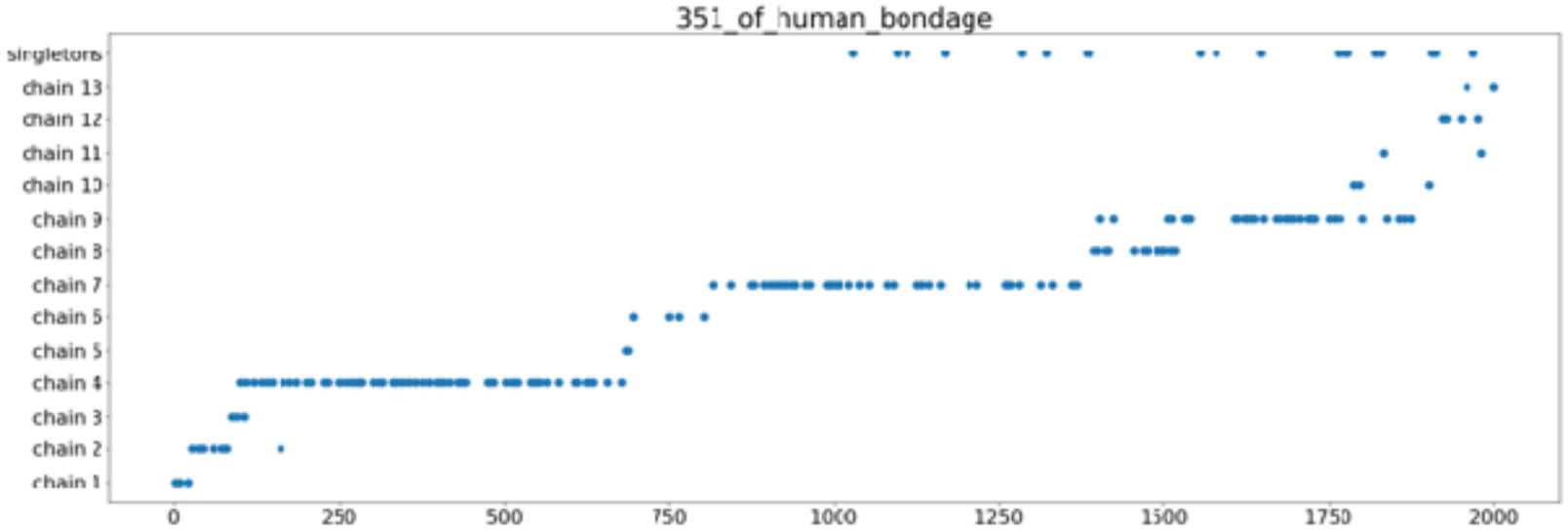
BERT

- Contextualized word representations from BERT base model + bidirectional LSTM.
- No fine-tuning on domain or task.



Setting coreference

- Which events take place at the same physical location in the narrative?


351_of_human_bondage
singletons
chain 13
chain 12
chain 11
chain 10
chain 9
chain 8
chain 7
chain 6
chain 5
chain 4
chain 3
chain 2
chain 1
0
250
500
750
1000
1250
1500
1750
2000

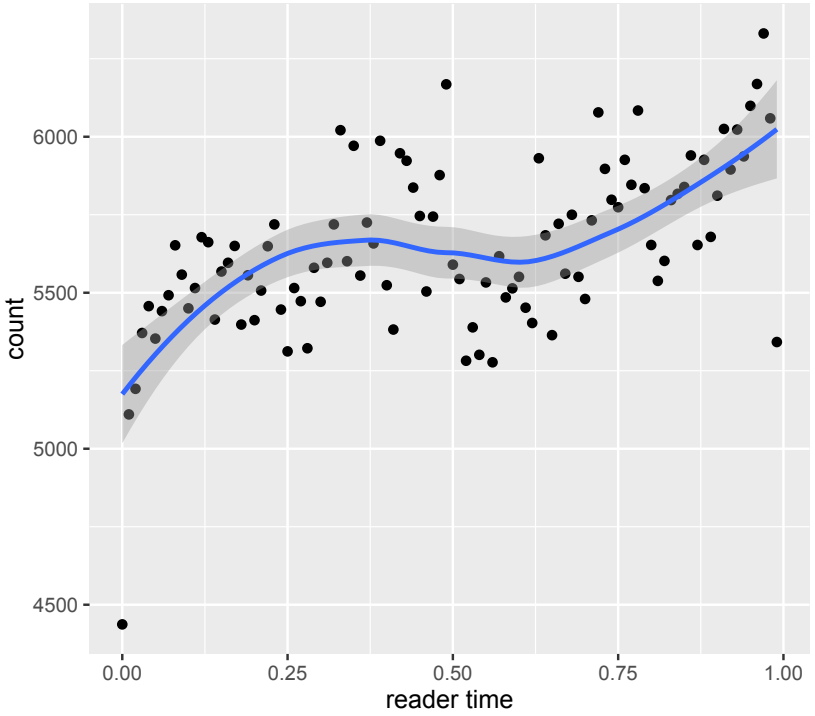
Questions

- Measuring the abstractness of a novel by the density of realis events.

Events	Book
0.109	The Quest of the Silver Fleece
0.091	The Invisible Man
0.083	Of Human Bondage
0.081	The Man of the Forest
0.078	Gullivers Travels
...	...
0.006	The Legend of Sleepy Hollow
0.004	The Mysteries of Udolpho
0.004	Moby Dick
0.003	Middlemarch
0	The Magnificent Ambersons

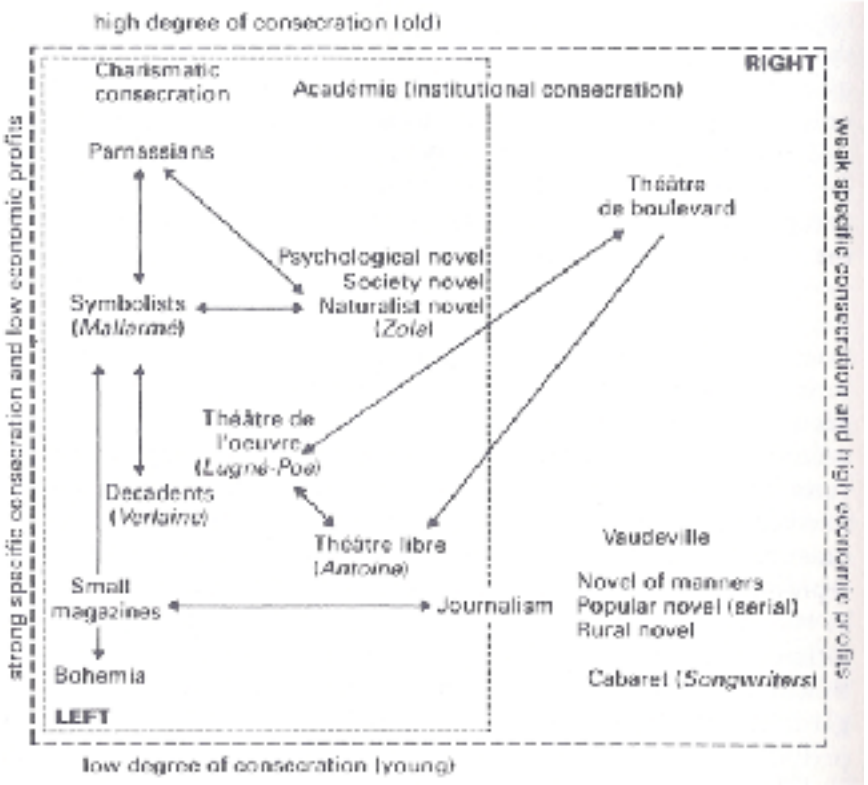
Questions

- Measuring the abstractness of a novel by the density of realis events.
- Modeling the distribution of settings over narrative time



Prestige

- What are the textual signals of authorial prestige?
- "Prestige" =
  - Inclusion in ODNB, MLA, Stanford exam lists (Algee-Hewitt et al. 2016)
  - Reviewed by elite literary journals (Underwood 2019)



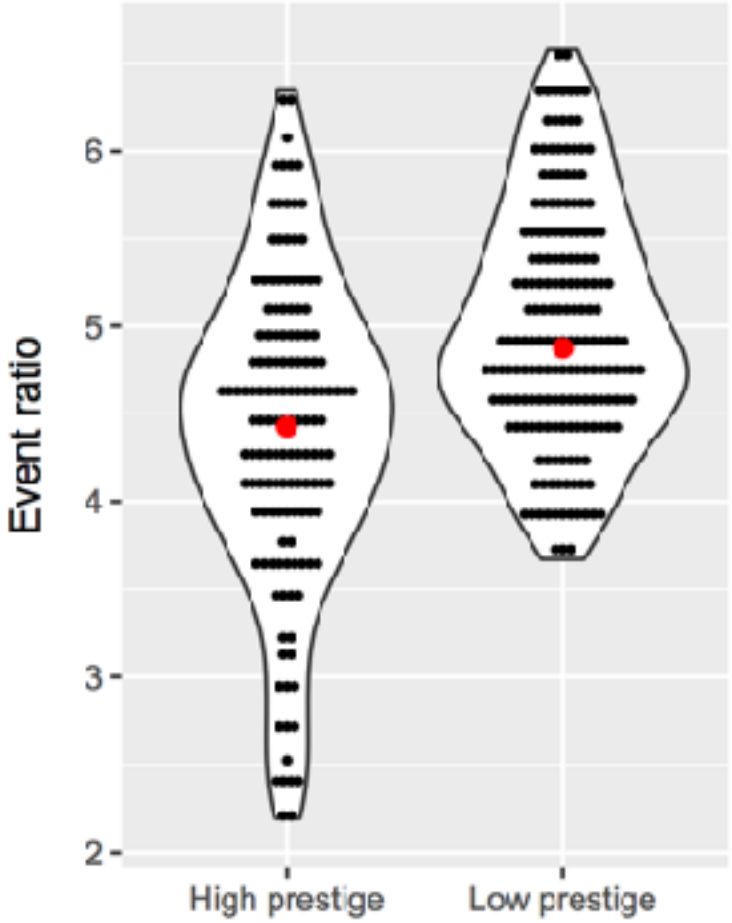
Prestige

- 100 authors identified in Underwood (2019) with highest and lowest prestige (as measured by the number of times their works were reviewed by elite literary journals); 150 high-prestige novels + 188 low-prestige novels.

- Metric = frequency of *realis* events (normalized by number of tokens)

Class	Ratio
High prestige	4.3 [4.2-4.5]
Low prestige	5.0 [4.9-5.1]

- Low prestige novels have less variability in depiction of realis events — something is always *happening*.



Thanks!

David Bamman
dbamman@berkeley.edu

Matt Sims, Jerry Park and David Bamman, “Literary Event Detection” (ACL 2019)

David Bamman, Sejal Popat and Sheng Shen, “An Annotated Dataset of Literary Entities” (NAACL 2019)

`https://github.com/dbamman/litbank`