Language Models for Law and Social Science

2. Tokenization

Any logistical questions about the course?

Tokenization: Overview

- Input:
  - A set of documents (e.g. text files), D.
- Output 1:
  - A sequence, W, containing a list of tokens – words or word pieces for use in natural language processing
- Output 2:
  - A document-term matrix, X, containing statistics about word/phrase frequencies in those documents.

Three Approaches to Tokenization

1. convert documents to count vectors, e.g. over pre-processed n-grams.
   - "bag of words" or "bag of terms" representation
   - used with topic models and classical ML
   - should be **informative/predictive** in the learning task, computationally **tractable**, and preferably somewhat **interpretable**.

Three Approaches to Tokenization

1. convert documents to count vectors, e.g. over pre-processed n-grams.
   - "bag of words" or "bag of terms" representation
   - used with topic models and classical ML
   - should be **informative/predictive** in the learning task, computationally **tractable**, and preferably somewhat **interpretable**.

2. segment documents into word pieces using byte pair encoding
   - maintain as much info from the original document as possible
   - for inputs to sequence models, i.e. transformers.

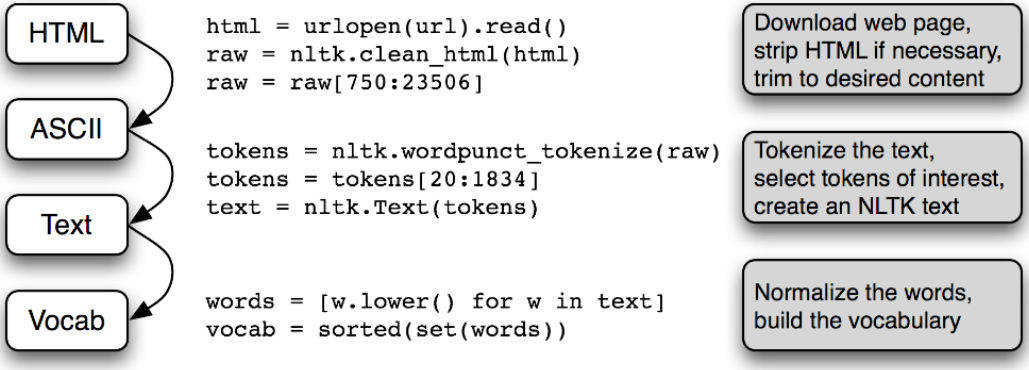
Three Approaches to Tokenization

1. convert documents to count vectors, e.g. over pre-processed n-grams.
   - "bag of words" or "bag of terms" representation
   - used with topic models and classical ML
   - should be **informative/predictive** in the learning task, computationally **tractable**, and preferably somewhat **interpretable**.

2. segment documents into word pieces using byte pair encoding
   - maintain as much info from the original document as possible
   - for inputs to sequence models, i.e. transformers.

3. enrich document with linguistics/grammar information
   - add more information to unprocessed doc based on sentence boundaries, parts of speech, syntax, etc
   - needed for specific tasks – eg relation extraction

A Standard Tokenization Pipeline



Source: NLTK Book, Chapter 3.

Pre-processing

- For many projects, the first question is: what data to throw out?
  - Uninformative data add noise and reduce statistical precision.
  - They are also computationally costly.
- Pre-processing choices can affect down-stream results, especially in unsupervised learning tasks (Denny and Spirling 2017), or self-supervised learning.
  - in particular: some features are more interpretable, e.g. ("discretion", "have", "judge") vs ("the judge has discretion").

Normalizing Text

1. Capitalization
   - usually the capitalized/non-capitalized version of a word are equivalent $\rightarrow$ capitalization not informative.

Normalizing Text

1. Capitalization
   - usually the capitalized/non-capitalized version of a word are equivalent → capitalization not informative.
   - On the other hand: what about “the first amendment” versus “the First Amendment”?

Normalizing Text

1. Capitalization
   - usually the capitalized/non-capitalized version of a word are equivalent → capitalization not informative.
   - On the other hand: what about "the first amendment" versus "the First Amendment"?
2. Punctuation
   - the number of periods or commas in a document is usually not that useful
   - so in a bag of terms approach, punctuation can be dropped.

Normalizing Text

1. Capitalization
   - usually the capitalized/non-capitalized version of a word are equivalent → capitalization not informative.
   - On the other hand: what about "the first amendment" versus "the First Amendment"?
2. Punctuation
   - the number of periods or commas in a document is usually not that useful
   - so in a bag of terms approach, punctuation can be dropped.
   - but what about "Let's eat, Grandpa", versus "Let's eat Grandpa"?

Normalizing Text

1. Capitalization
   - usually the capitalized/non-capitalized version of a word are equivalent $\rightarrow$ capitalization not informative.
   - On the other hand: what about "the first amendment" versus "the First Amendment"?
2. Punctuation
   - the number of periods or commas in a document is usually not that useful
   - so in a bag of terms approach, punctuation can be dropped.
   - but what about "Let's eat, Grandpa", versus "Let's eat Grandpa"?
3. Numbers
   - individual numbers are usually too specific to keep in the vocabulary
   - But how often numbers are mentioned might be important; can replace with a special character, e.g. #.

Stopwords

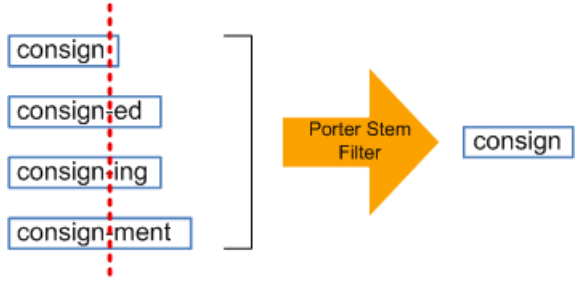
a	an	and	are	as	at	be	by	for	from
has	he	in	is	it	its	of	on	that	the
to	was	were	will	with					

Stopwords

a	an	and	are	as	at	be	by	for	from
has	he	in	is	it	its	of	on	that	the
to	was	were	will	with					

- What about "not guilty"?
- Legal terms often contain stopwords:
  - "beyond a reasonable doubt"
  - "with all deliberate speed"
- can drop stopwords by themselves, but keep them when part of phrases.

Stemming/lemmatizing



- Effective dimension reduction with little loss of information.
- Lemmatizer produces real words, but N-grams won't make grammatical sense
  - e.g., "judges have been ruling" would become "judge have be rule"

Pre-processing with `gensim`

gensim.parsing.preprocessing.preprocess_string(*s*, *filters=[<function <lambda>>, <function strip_tags>, <function strip_punctuation>, <function strip_multiple_whitespaces>, <function strip_numeric>, <function remove_stopwords>, <function strip_short>, <function stem_text>]*)

Apply list of chosen filters to *s*.

Default list of filters:

- `strip_tags()`,
- `strip_punctuation()`,
- `strip_multiple_whitespaces()`,
- `strip_numeric()`,
- `remove_stopwords()`,
- `strip_short()`,
- `stem_text()`.

Parameters:
- **s** (*str*) –
- **filters** (*list of functions, optional*) –

Returns: Processed strings (cleaned).

Return type: list of str

Bag-of-words representation

Say we want to convert a corpus D to a matrix X:

- In the "bag-of-words" representation, a row of X is just the frequency distribution over words in the document corresponding to that row.

Counts and frequencies:

- **Document counts**: number of documents where a word appears.
- **Term counts**: number of total appearances of a word in corpus.
- **Term frequency**:

$$\text{Term Frequency of } w \text{ in document } k = \frac{\text{Count of } w \text{ in document } k}{\text{Total tokens in document } k}$$

Building a vocabulary

- What are the columns of the document-term matrix X?
  - Assign numerical indices to words to increase speed and reduce disk usage.
- Pick a number:, e.g. 100,000 most frequent words.

Building a vocabulary

- What are the columns of the document-term matrix X?
  - Assign numerical indices to words to increase speed and reduce disk usage.

- Pick a number:, e.g. 100,000 most frequent words.
- Frequency threshold:
  - Compute document frequencies for all words
  - Inspect low-frequency words and determine a minimum document threshold.
    - e.g., 10 documents, or .25% of documents.

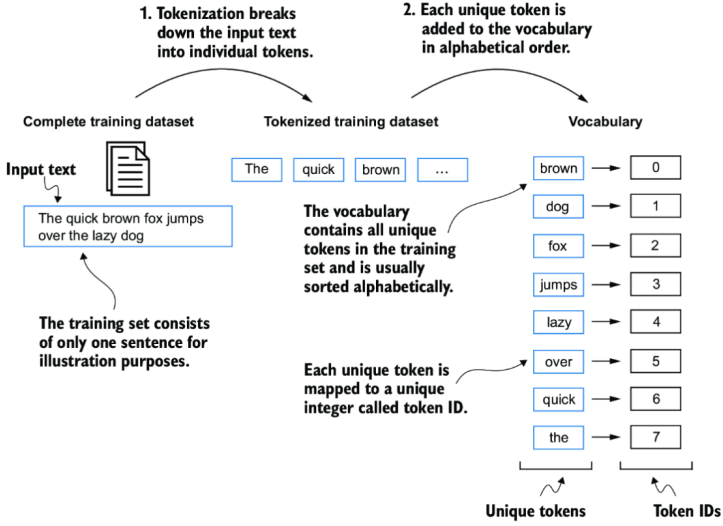


Figure 2.6 We build a vocabulary by tokenizing the entire text in a training dataset into individual tokens. These individual tokens are then sorted alphabetically, and duplicate tokens are removed. The unique tokens are then aggregated into a vocabulary that defines a mapping from each unique token to a unique integer value. The depicted vocabulary is purposely small and contains no punctuation or special characters for simplicity.

scikit-learn's CountVectorizer

https://scikit-learn.org/stable/modules/feature_extraction.html

`CountVectorizer` implements both tokenization and occurrence counting in a single class:

```
>>> from sklearn.feature_extraction.text import CountVectorizer
```

This model has many parameters, however the default values are quite reasonable (please see the reference documentation for the details):

```
>>> vectorizer = CountVectorizer()
>>> vectorizer
CountVectorizer()
```

Let's use it to tokenize and count the word occurrences of a minimalistic corpus of text documents:

```
>>> corpus = [
...     'This is the first document.',
...     'This is the second second document.',
...     'And the third one.',
...     'Is this the first document?',
... ]
>>> X = vectorizer.fit_transform(corpus)
>>> X
<4x9 sparse matrix of type '<... 'numpy.int64'>'
    with 19 stored elements in Compressed Sparse ... format>
```

The default configuration tokenizes the string by extracting words of at least 2 letters.

- `corpus` is a sequence of strings, e.g. pandas data-frame columns.
- pre-processing options: strip accents, lowercase, drop stopwords,
- vocab options: min/max frequency, vocab size
- n-grams: can produce phrases up to length n (words or characters).

What about out-of-vocab words?

What about out-of-vocab words?

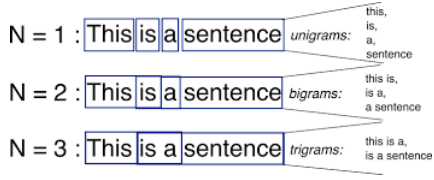
- in bag-of-words model:
  - drop them
  - replace with "unknown" token (<unk>)
  - replace with part-of-speech tag
  - replace with in-vocab hypernym (from WordNet)
  - others?

What about out-of-vocab words?

- in bag-of-words model:
  - drop them
  - replace with "unknown" token (<unk>)
  - replace with part-of-speech tag
  - replace with in-vocab hypernym (from WordNet)
  - others?
- alternative approaches that don't have this problem (more below):
  - hashing vectorizer
  - byte pair encoding

N-grams are phrases, sequences of words up to length N

- e.g. bigrams, trigrams, quadgrams, etc.



- Baseline for text classification of long documents (Google Developers Guide):
  - X = counts over bigrams, with vocab size = 20,000

`https://developers.google.com/machine-learning/guides/text-classification/step-3`

Feature selection

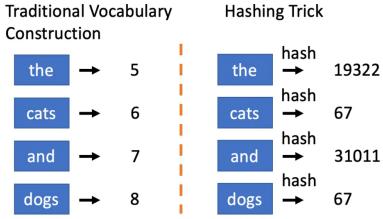
- N-grams quickly blow up the feature space.
  - filtering on frequency is easiest but not optimal – can filter on usefulness for a task instead.
- Text normalization is important (capitalization, punctuation, stopwords, stemming)

- For supervised learning tasks:
  - Use supervised feature selection to select predictive features (more on week 4)
- What about unsupervised learning (e.g. topic models)?
  - can use parts of speech / co-location statistics (week 3)

Feature selection

- N-grams quickly blow up the feature space.
  - filtering on frequency is easiest but not optimal – can filter on usefulness for a task instead.
- Text normalization is important (capitalization, punctuation, stopwords, stemming)
- For supervised learning tasks:
  - Use supervised feature selection to select predictive features (more on week 4)
- What about unsupervised learning (e.g. topic models)?
  - can use parts of speech / co-location statistics (week 3)
- In week 3, we explore more general problem of dimensionality reduction.

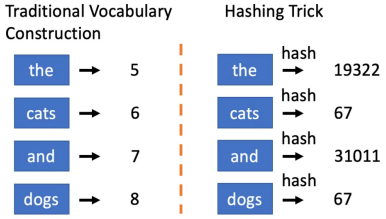
Hashing Vectorizer



- Rather than make a one-to-one lookup for each n-gram, put n-grams through a hashing function that takes an arbitrary string and deterministically outputs an integer in some range (e.g. 1 to 10,000).

```
>>> from sklearn.feature_extraction.text import HashingVectorizer
>>> hv = HashingVectorizer(n_features=10)
>>> hv.transform(corpus)
<4x10 sparse matrix of type '<... 'numpy.float64'>'
    with 16 stored elements in Compressed Sparse ... format>
```

Hashing Vectorizer



- Rather than make a one-to-one lookup for each n-gram, put n-grams through a hashing function that takes an arbitrary string and deterministically outputs an integer in some range (e.g. 1 to 10,000).

```
>>> from sklearn.feature_extraction.text import HashingVectorizer
>>> hv = HashingVectorizer(n_features=10)
>>> hv.transform(corpus)
<4x10 sparse matrix of type '<... 'numpy.float64'>'
    with 16 stored elements in Compressed Sparse ... format>
```

Pros:

- can have arbitrarilly small feature space
- handles out-of-vocabulary words – any word or n-gram gets assigned to an arbitrary integer based on the hash function.

Cons:

- harder to interpret features, at least not directly (eli5 implementation keeps track of the mapping)-
- collisions – n-grams will randomly be paired with each other in the feature map – in supervised learning, usually innocuous

Research Question

- What drives slant in print media in the United States?
  - **Consumer preferences?**
  - Politicians?
  - Newspaper owners?

- Example: What influences, whether a newspaper is more likely to use the term **death tax**, or the term **estate tax**?

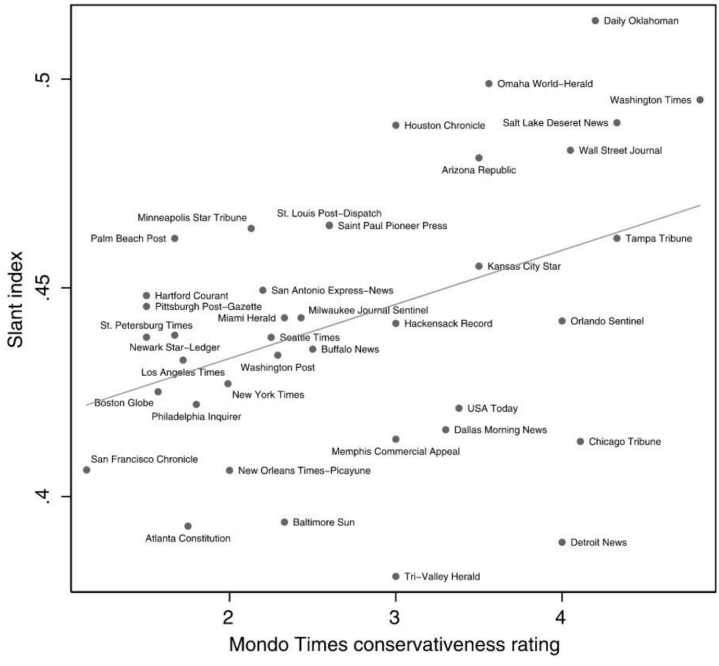
Corpus and data

- 2005 Congressional record and politicians' positions
  - https://www.congress.gov/congressional-record
    - Pre-processed to exclude common words
  - Vote share in the 2004 presidential election in politicians' constituencies
- Newspapers and their markets
  - newslibrary.com and proquest.com
    - Exclude opinion content
    - Exclude globally read newspapers (e.g. New York Times)
  - Owner information from E&P international yearbook (2000)
  - Demographic data from 2000 Census in the newspaper's PMSA (primary metropolitan statistical area)
  - Corporate political contributions (publicintegrity.org)

NLP Methods

$$\chi^2_{pl} = \frac{(f_{plr}f_{\sim pld} - f_{pld}f_{\sim plr})^2}{(f_{plr} + f_{pld})(f_{plr} + f_{\sim plr})(f_{pld} + f_{\sim pld})(f_{\sim plr} + f_{\sim pld})}.$$

- Pre-processing: remove common neutral words (the, if, …)
- Phrase selection
  - Select phrases which are used (a) frequently, and (b) significantly more by one party than the other.
  - Use 2- word phrases and 3-word phrases
  - 2x top 500 → 1000 phrases total
- Phrase to ideology mapping
  - Linear regression on congresspersons' ideologies
  - Use obtained slope parameters to estimate newspapers' ideologies
  - **0.4 correlation with Mondo times conservativeness rating of newspapers**
    - Could have repeated all following sections with this alternative rating


Daily Oklahoman
Omaha World-Herald
Washington Times
Houston Chronicle
Salt Lake Deseret News
Wall Street Journal
Arizona Republic
Minneapolis Star Tribune
St. Louis Post-Dispatch
Saint Paul Pioneer Press
Palm Beach Post
Tampa Tribune
Kansas City Star
Hartford Courant
San Antonio Express-News
Pittsburgh Post-Gazette
Miami Herald
Milwaukee Journal Sentinel
St. Petersburg Times
Hackensack Record
Orlando Sentinel
Seattle Times
Newark Star-Ledger
Buffalo News
Washington Post
Los Angeles Times
New York Times
Boston Globe
USA Today
Philadelphia Inquirer
Dallas Morning News
Chicago Tribune
Memphis Commercial Appeal
San Francisco Chronicle
New Orleans Times-Picayune
Baltimore Sun
Atlanta Constitution
Detroit News
Tri-Valley Herald
.5
.45
.4
Slant index
2
3
4
Mondo Times conservativeness rating

Panel A: Phrases Used More Often by Democrats

Two-Word Phrases

private accounts	Rosa Parks	workers rights
trade agreement	President budget	poor people
American people	Republican party	Republican leader
tax breaks	change the rules	Arctic refuge
trade deficit	minimum wage	cut funding
oil companies	budget deficit	American workers
credit card	Republican senators	living in poverty
nuclear option	privatization plan	Senate Republicans
war in Iraq	wildlife refuge	fuel efficiency
middle class	card companies	national wildlife

Three-Word Phrases

veterans health care	corporation for public broadcasting	cut health care
congressional black caucus	additional tax cuts	civil rights movement
VA health care	pay for tax cuts	cuts to child support
billion in tax cuts	tax cuts for people	drilling in the Arctic National
credit card companies	oil and gas companies	victims of gun violence
security trust fund	prescription drug bill	solvency of social security
social security trust	caliber sniper rifles	Voting Rights Act
privatize social security	increase in the minimum wage	war in Iraq and Afghanistan
American free trade	system of checks and balances	civil rights protections
central American free	middle class families	credit card debt

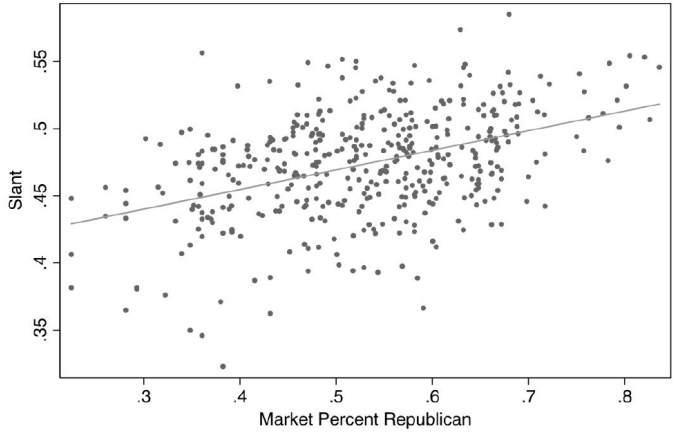
Panel B: Phrases Used More Often by Republicans

Two-Word Phrases

stem cell	personal accounts	retirement accounts
natural gas	Saddam Hussein	government spending
death tax	pass the bill	national forest
illegal aliens	private property	minority leader
class action	border security	urge support
war on terror	President announces	cell lines
embryonic stem	human life	cord blood
tax relief	Chief Justice	action lawsuits
illegal immigration	human embryos	economic growth
date the time	increase taxes	food program

Three-Word Phrases

embryonic stem cell	Circuit Court of Appeals	Tongass national forest
hate crimes legislation	death tax repeal	pluripotent stem cells
adult stem cells	housing and urban affairs	Supreme Court of Texas
oil for food program	million jobs created	Justice Priscilla Owen
personal retirement accounts	national flood insurance	Justice Janice Rogers
energy and natural resources	oil for food scandal	American Bar Association
global war on terror	private property rights	growth and job creation
hate crimes law	temporary worker program	natural gas natural
change hearts and minds	class action reform	Grand Ole Opry
global war on terrorism	Chief Justice Rehnquist	reform social security


Slant
.35
.4
.45
.5
.55
.3
.4
.5
.6
.7
.8
Market Percent Republican

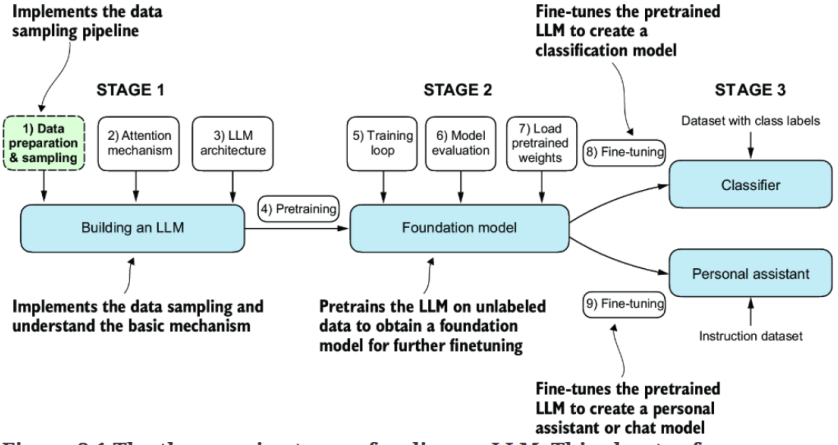


Figure 2.1 The three main stages of coding an LLM. This chapter focuses on step 1 of stage 1: implementing the data sample pipeline.

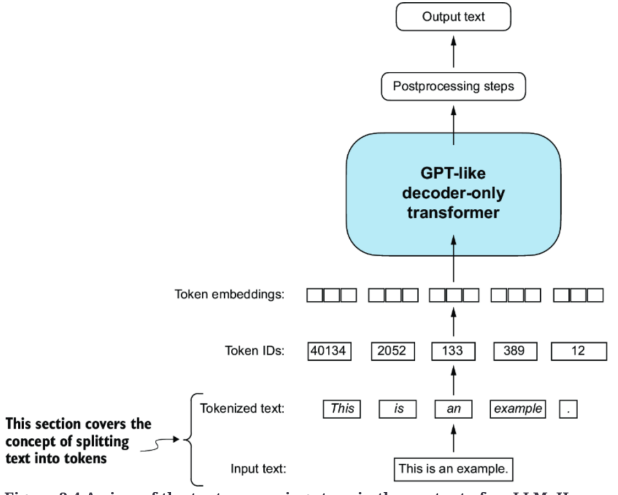


Figure 2.4 A view of the text processing steps in the context of an LLM. Here, we split an input text into individual tokens, which are either words or special characters, such as punctuation characters.

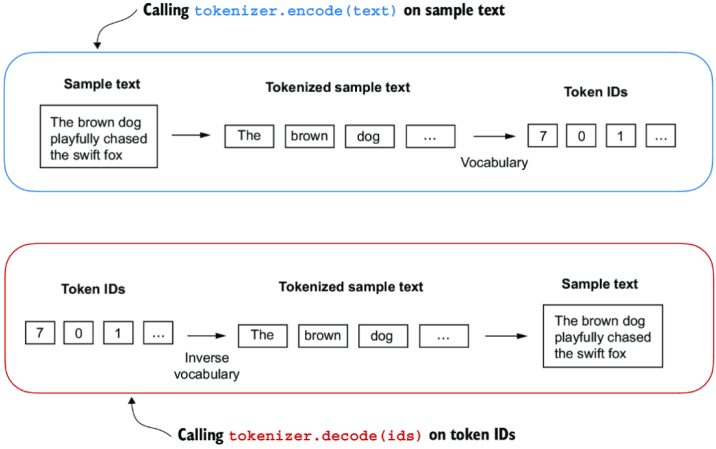


Figure 2.8 Tokenizer implementations share two common methods: an encode method and a decode method. The encode method takes in the sample text, splits it into individual tokens, and converts the tokens into token IDs via the vocabulary. The decode method takes in token IDs, converts them back into text tokens, and concatenates the text tokens into natural text.

Limitations of word tokenization

- modern language models are designed/intended to capture all meaning in texts.
  - with word tokenization, would need a massive vocabulary to capture all cases.
  - not impossible but computationally costly.
  - model might encounter new words in the test data.

Limitations of word tokenization

- modern language models are designed/intended to capture all meaning in texts.
  - with word tokenization, would need a massive vocabulary to capture all cases.
  - not impossible but computationally costly.
  - model might encounter new words in the test data.
- treats different word forms as separate words (e.g. “tax”, “taxes”, “taxed”)
  - or, with stemming, treats them as identical

Character tokenization

- alternative – tokenize characters rather than words:
  - "hello world" $\rightarrow$ {h,e,l,l,o, ,w,o,r,l,d}
  - by construction, no unknown words.

Character tokenization

- alternative – tokenize characters rather than words:
  - "hello world" $\rightarrow$ {h,e,l,l,o, ,w,o,r,l,d}
  - by construction, no unknown words.
- this actually works fine, and is used in some recent language models.

Character tokenization

- alternative – tokenize characters rather than words:
  - "hello world" $\rightarrow$ {h,e,l,l,o, ,w,o,r,l,d}
  - by construction, no unknown words.
- this actually works fine, and is used in some recent language models.
  - but not efficient: some single characters (e.g. "x") are much less frequent than some character subsequences (e.g. "tion") or even whole words (e.g. "the")

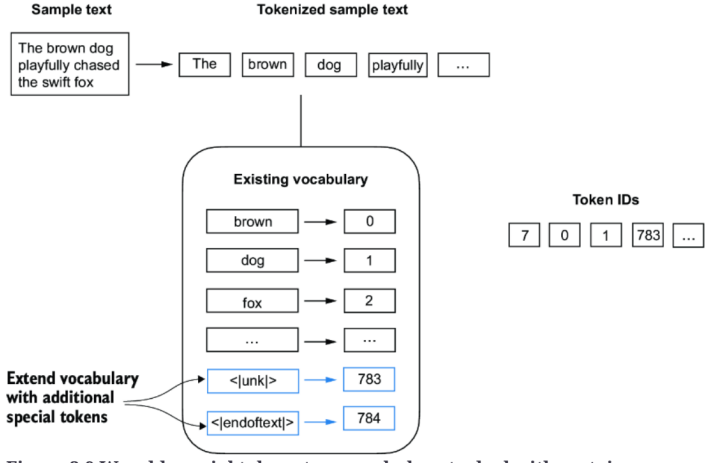


Figure 2.9 We add special tokens to a vocabulary to deal with certain contexts. For instance, we add an `<|unk|>` token to represent new and unknown words that were not part of the training data and thus not part of the existing vocabulary. Furthermore, we add an `<|endoftext|>` token that we can use to separate two unrelated text sources.

Subword Tokenization

Most modern language models (e.g. BERT, GPT) use subword tokenization:

- construct character-level n-grams using byte pair encoding (frequent character sequences treated as a token)
- whitespace treated the same as letters
- all letters to lowercase, but add a special character for the next letter being capitalized.

e.g., BERT's SentencePiece tokenizer:

Byte Pair Encoding

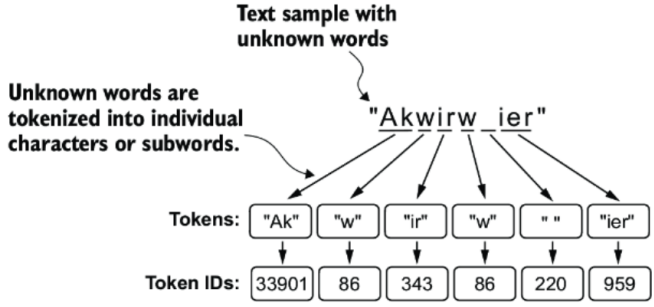


Figure 2.11 BPE tokenizers break down unknown words into subwords and individual characters. This way, a BPE tokenizer can parse any word and doesn't need to replace unknown words with special tokens, such as `<|unk|>`.

- character-level byte-pair encoder, learns character n-grams and treats spaces and punctuation the same as other characters.
- GPT tokenizer: `https://platform.openai.com/tokenizer`

Limitations of word/subword tokenization

- The tokenizers we have looked at so far are based only on data in the corpus.
  - bag-of-words tokenizers break down text into counts over the most relevant features.
  - subword tokenizers try to preserve the text as is.

Limitations of word/subword tokenization

- The tokenizers we have looked at so far are based only on data in the corpus.
  - bag-of-words tokenizers break down text into counts over the most relevant features.
  - subword tokenizers try to preserve the text as is.
- These tokenizers leave out a lot of information that we have from sophisticated and powerful conceptual models of language – that is, linguistics.

Segmenting paragraphs/sentences

- Many tasks should be done on sentences, rather than corpora as a whole.
  - spaCy is a good (but not perfect) job of splitting sentences, while accounting for periods on abbreviations, etc.
- There isn't a grammar-based paragraph tokenizer.
  - most corpora have new paragraphs annotated.
  - or use line breaks.

Parts of speech tags

- Parts of speech (POS) tags provide useful word categories corresponding to their functions in sentences:
  - Eight main parts of speech: verb (VB), noun (NN), pronoun (PR), adjective (JJ), adverb (RB), determinant (DT), preposition (IN), conjunction (CC).
  - The Penn TreeBank POS tag set (used in many applications) has 36 tags: `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`

Parts of speech tags

- Parts of speech (POS) tags provide useful word categories corresponding to their functions in sentences:
  - Eight main parts of speech: verb (VB), noun (NN), pronoun (PR), adjective (JJ), adverb (RB), determinant (DT), preposition (IN), conjunction (CC).
  - The Penn TreeBank POS tag set (used in many applications) has 36 tags: `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`
- Parts of speech vary in their informativeness for various functions:
  - For categorizing topics, nouns are usually most important
  - For sentiment, adjectives are usually most important.
- In particular, noun phrases are often informative features – spaCy can do fast noun phrase chunking.

Parts of speech tags

- Parts of speech (POS) tags provide useful word categories corresponding to their functions in sentences:
  - Eight main parts of speech: verb (VB), noun (NN), pronoun (PR), adjective (JJ), adverb (RB), determinant (DT), preposition (IN), conjunction (CC).
  - The Penn TreeBank POS tag set (used in many applications) has 36 tags: `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`
- Parts of speech vary in their informativeness for various functions:
  - For categorizing topics, nouns are usually most important
  - For sentiment, adjectives are usually most important.
- In particular, noun phrases are often informative features – spaCy can do fast noun phrase chunking.
- Can count POS tags as features – e.g., using more adjectives, or using more passive verbs.
  - provides style features, e.g. for authorship detection.

Named Entity Recognition

- Named entities such as “ETH Zurich” and “Marie Curie” are a special set of annotations, tagged by named entity recognizers (NER).
- usually identified by proper nouns; most pre-trained NER systems, like spACy, also give an entity category:

> [PER John Smith] , president of [ORG McCormik Industries] visited his niece [PER Paris] in [LOC Milan], reporters say .

Application: POS tags Predict Loan Repayment

Netzer, Lemaire, and Herzenstein (2019), "When Words Sweat"

Application: POS tags Predict Loan Repayment

Netzer, Lemaire, and Herzenstein (2019), "When Words Sweat"

- Imagine you consider lending $2,000 to one of two borrowers on a crowdfunding website. The borrowers are identical in terms of demographic and financial characteristics. However, the text they provided when applying for a loan differs:
  - Borrower #1: "*I am a hard working person, married for 25 years, and have two wonderful boys. Please let me explain why I need help. I would use the $2,000 loan to fix our roof. Thank you, god bless you, and I promise to pay you back.*"
  - Borrower #2: "*While the past year in our new place has been more than great, the roof is now leaking and I need to borrow $2,000 to cover the cost of the repair. I pay all bills (e.g., car loans, cable, utilities) on time.*"
- Which borrower is more likely to default?

Application: POS tags Predict Loan Repayment

Netzer, Lemaire, and Herzenstein (2019), "When Words Sweat"

- Imagine you consider lending $2,000 to one of two borrowers on a crowdfunding website. The borrowers are identical in terms of demographic and financial characteristics. However, the text they provided when applying for a loan differs:
  - Borrower #1: "*I am a hard working person, married for 25 years, and have two wonderful boys. Please let me explain why I need help. I would use the $2,000 loan to fix our roof. Thank you, god bless you, and I promise to pay you back.*"
  - Borrower #2: "*While the past year in our new place has been more than great, the roof is now leaking and I need to borrow $2,000 to cover the cost of the repair. I pay all bills (e.g., car loans, cable, utilities) on time.*"
- Which borrower is more likely to default?

"Loan requests written by defaulting borrowers are more likely to include words (or themes) related to the borrower's family, financial and general hardship, mentions of god, and the near future, as well as pleading lenders for help, and using verbs in present and future tenses."

Loan Application Words Predicting Repayment (Netzer, Lemaire, and Herzenstein 2019)

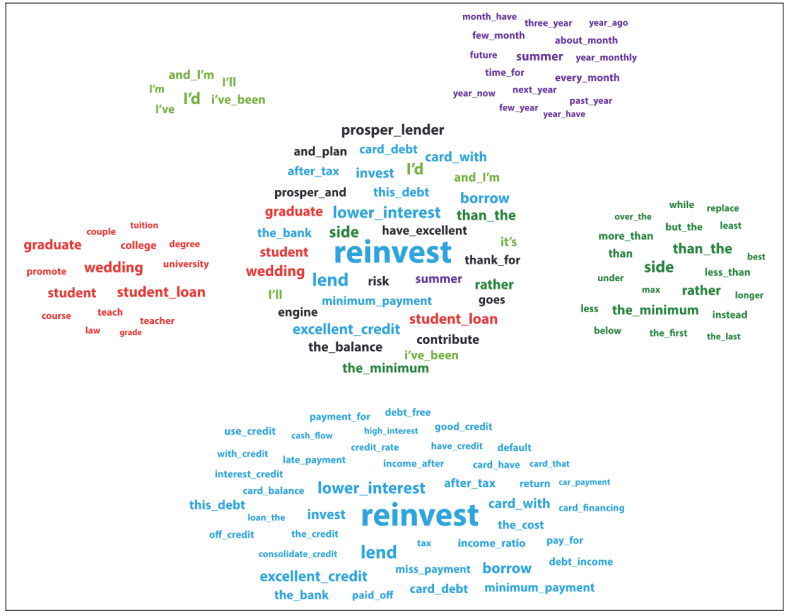


Figure 2. Words indicative of loan repayment.
Notes: The most common words appear in the middle cloud (cutoff = 1:1.5) and are then organized by themes. Starting on the right and moving clockwise: relative words, financial literacy words, words related to a brighter financial future, "I" words, and time-related words.

Loan Application Words Predicting Default (Netzer, Lemaire, and Herzenstein 2019)

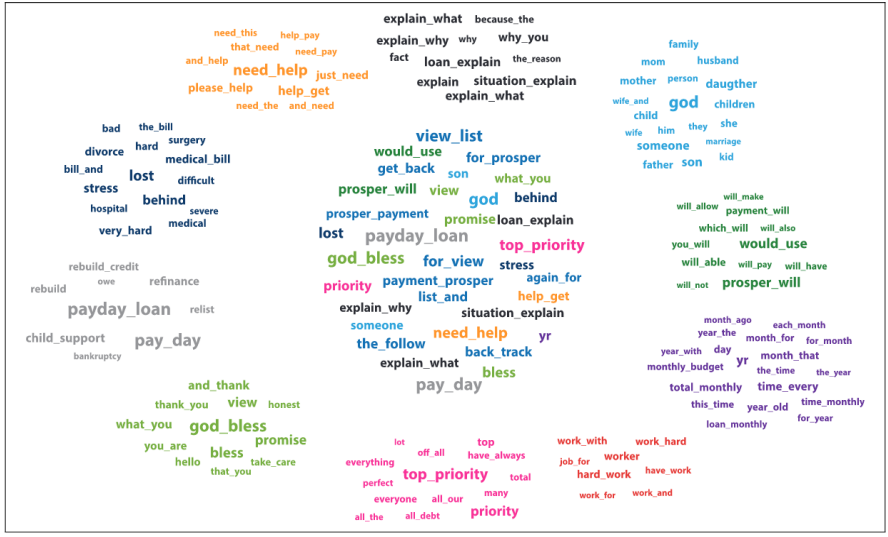


Figure 3. Words indicative of loan default.

Notes: The most common words appear in the middle cloud (cutoff = 1:1.5) and are then organized by themes. Starting on the top and moving clockwise: words related to explanations, external influence words and others, future-tense words, time-related words, work-related words, extremity words, words appealing to lenders, words relating to financial hardship, words relating to general hardship, and desperation/plea words.

Using Grammar: Constituency

Using Grammar: Constituency

- The idea of constituency is that groups of words behave as singular functional units in a sentence.
- Some example noun phrases:

Harry the Horse	a high-class spot such as Mindy's
the Broadway coppers	the reason he comes into the Hot Box
they	three parties from Brooklyn

  - these phrases consist of many POS's but function as nouns

Using Grammar: Constituency

- The idea of constituency is that groups of words behave as singular functional units in a sentence.
- Some example noun phrases:

Harry the Horse	a high-class spot such as Mindy's
the Broadway coppers	the reason he comes into the Hot Box
they	three parties from Brooklyn

  - these phrases consist of many POS's but function as nouns
- In English, constituents can be moved around in a sentence (e.g. these prepositional phrases):
  - John talked [to the students] [about linguistics].
  - John talked [about linguistics] [to the students] .

Using Grammar: Syntactic Dependencies

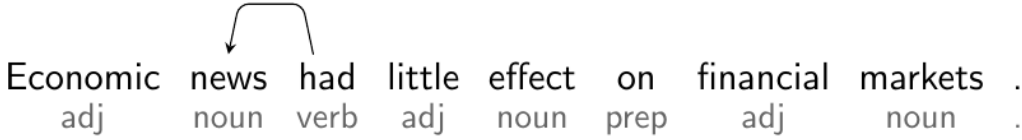
- The basic idea:
  - **Syntactic structure** consists of **words**, linked by binary directed relations called **dependencies**.
  - Dependencies identify the grammatical relations between words.

Dependencies: Binary Directed Relations Between Words (Head and Dependent)

Economic	news	had	little	effect	on	financial	markets	.
adj	noun	verb	adj	noun	prep	adj	noun	.

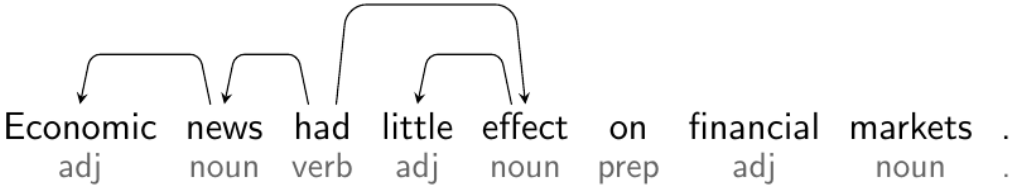
- dependency trees are mostly determined by the ordering of POS tags.

Dependencies: Binary Directed Relations Between Words (Head and Dependent)



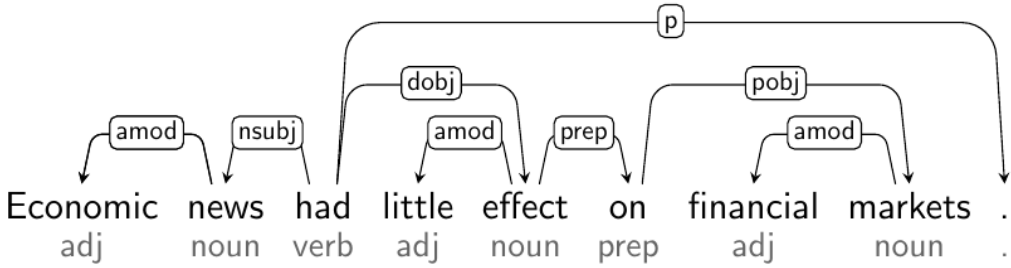
- the "root" of a sentence is the main verb (for compound sentences, the first verb).

Dependencies: Binary Directed Relations Between Words (Head and Dependent)



- directed arcs indicate dependencies: a one-way link from a "head" token to a "dependent" token.
- A word can be "head" multiple times, but "dependent" only once.

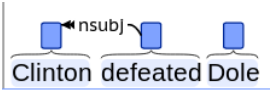
Dependencies: Binary Directed Relations Between Words (Head and Dependent)



- arc labels indicate functional relations, e.g.:
  - nsubj: verb → subject doing the verb
  - dobj: verb → object targeted by the verb
  - amod: noun → attribute of the noun
- spaCy dependency visualizer: `https://explosion.ai/demos/displacy`

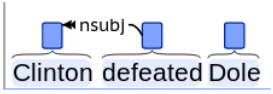
Who does What to Whom: Subjects

- **nsubj: nominal subject**
  - points from the active verb to the agent subject.

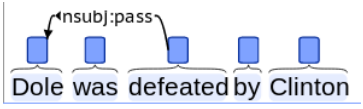


Who does What to Whom: Subjects

- **nsubj: nominal subject**
  - points from the active verb to the agent subject.



- **nsubjpass: passive nominal subject**
  - points from a passive verb to the patient subject



Who does What to Whom: Objects

dobj: direct object

- points from an active verb to the the (accusative) object noun phrase.

"She **gave** me a **raise**"

$$gave \overset{dobj}{\rightarrow} raise$$

Who does What to Whom: Objects

dobj: direct object

- points from an active verb to the the (accusative) object noun phrase.

"She **gave** me a **raise**"

$$gave \overset{dobj}{\rightarrow} raise$$

dative: dative or indirect object

- points from an active verb to the the (dative) object noun phrase.

"She **gave me** a raise"

$$gave \overset{dative}{\rightarrow} me$$

Who does What to Whom: Objects

dobj: direct object

- points from an active verb to the the (accusative) object noun phrase.

"She **gave** me a **raise**"

$$gave \overset{dobj}{\rightarrow} raise$$

dative: dative or indirect object

- points from an active verb to the the (dative) object noun phrase.

"She **gave me** a raise"

$$gave \overset{dative}{\rightarrow} me$$

pobj: object of a preposition

- noun phrase following a preposition

"I sat **on** the **chair**"

$$on \overset{pobj}{\rightarrow} chair$$

What Attributes do Entities Have?

acomp: adjectival complement

- points from verb to adjectival phrase functioning as object

"Bill **is honest** ": acomp(is $\rightarrow$ honest)

What Attributes do Entities Have?

acomp: adjectival complement

- points from verb to adjectival phrase functioning as object

"Bill **is honest** ": acomp(is $\rightarrow$ honest)

attr: attribute

- points from copula verb to an attribute noun phrase.

"Bill **is** a **saint**": attr(is $\rightarrow$ saint)

What Attributes do Entities Have?

acomp: adjectival complement

- points from verb to adjectival phrase functioning as object

"Bill **is honest** ": acomp(is $\rightarrow$ honest)

attr: attribute

- points from copula verb to an attribute noun phrase.

"Bill **is** a **saint**": attr(is $\rightarrow$ saint)

amod: adjectival modifier

- points from a noun to an adjective modifying it

"Sam eats **red meat**": amod(meat $\rightarrow$ red)

Verb phrases

- **aux: auxiliary**
  - points from a main verb to a helping verb, including modals.

"Reagan **has died**": aux(died → has)

"He **should leave**": aux(leave → should)

Verb phrases

- **aux: auxiliary**
  - points from a main verb to a helping verb, including modals.

"Reagan **has died**": aux(died → has)

"He **should leave**": aux(leave → should)

- **auxpass: passive auxiliary**
  - points from a main verb to a helping verb indicative passive voice.

"Laws have **been broken**": auxpass(broken → been)

Verb phrases

- **aux: auxiliary**
  - points from a main verb to a helping verb, including modals.

    "Reagan **has died**": aux(died $\rightarrow$ has)

    "He **should leave**": aux(leave $\rightarrow$ should)
- **auxpass: passive auxiliary**
  - points from a main verb to a helping verb indicative passive voice.

    "Laws have **been broken**": auxpass(broken $\rightarrow$ been)
- **neg: negation modifier**
  - points from a verb to a negation indicator
  - "Bill **is not** a scientist": neg(is $\rightarrow$ not)

Verb phrases

- **aux: auxiliary**
  - points from a main verb to a helping verb, including modals.

    "Reagan **has died**": aux(died → has)

    "He **should leave**": aux(leave → should)
- **auxpass: passive auxiliary**
  - points from a main verb to a helping verb indicative passive voice.

    "Laws have **been broken**": auxpass(broken → been)
- **neg: negation modifier**
  - points from a verb to a negation indicator
  - "Bill **is not** a scientist": neg(is → not)
- **prt: phrasal verb particle**
  - points from a verb to its particle, linking phrasal verbs.

    "They **shut down** the station": prt(shut → down)

Verb phrases

- **aux: auxiliary**
  - points from a main verb to a helping verb, including modals.

    "Reagan **has died**": aux(died → has)

    "He **should leave**": aux(leave → should)
- **auxpass: passive auxiliary**
  - points from a main verb to a helping verb indicative passive voice.

    "Laws have **been broken**": auxpass(broken → been)
- **neg: negation modifier**
  - points from a verb to a negation indicator
  - "Bill **is not** a scientist": neg(is → not)
- **prt: phrasal verb particle**
  - points from a verb to its particle, linking phrasal verbs.

    "They **shut down** the station": prt(shut → down)
- and more...

Semantic Role Labeling (PropBank Labels)

Ex1: [$_{\text{Arg0}}$ The group] *agreed* [$_{\text{Arg1}}$ it wouldn't make an offer].

Ex2: [$_{\text{ArgM-TMP}}$ Usually] [$_{\text{Arg0}}$ John] *agrees* [$_{\text{Arg2}}$ with Mary] [$_{\text{Arg1}}$ on everything].

ARG0	agent	ARG3	starting point, benefactive, attribute
ARG1	patient	ARG4	ending point
ARG2	instrument, benefactive, attribute	ARGM	modifier

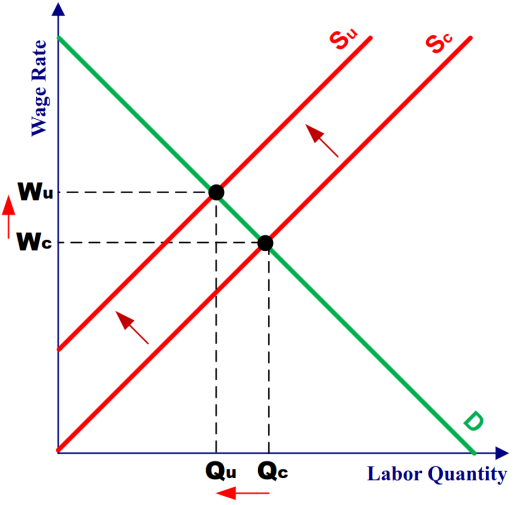
Table 1.1: List of arguments in PropBank

- Agent (ARG0)
  - Volitional/sentient involvement in event or state
  - Causes an event or change of state in another participant
- Patient (ARG1)
  - Causally affected by an agent/action
  - Undergoes change of state
- ARG2 has three functions:
  - instrument for an action ("Pat opened the door with a crowbar.")
  - attribute assigned to a patient ("Pat is an agent").
  - benefactive: the dative/indirect object ("Sasha gave the crowbar to Pat.")

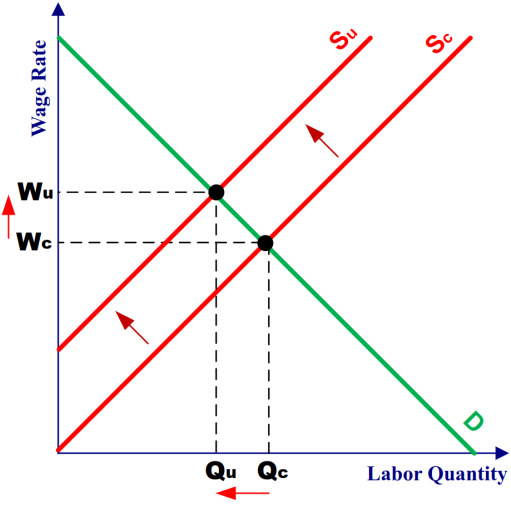
ARG-M: Modifiers

ArgM-TMP	when?	yesterday evening, now
LOC	where?	at the museum, in San Francisco
DIR	where to/from?	down, to Bangkok
MNR	how?	clearly, with much enthusiasm
PRP/CAU	why?	because ... , in response to the ruling
REC		themselves, each other
ADV	miscellaneous	
PRD	secondary predication	...ate the meat raw

Collective Bargaining, for Economists

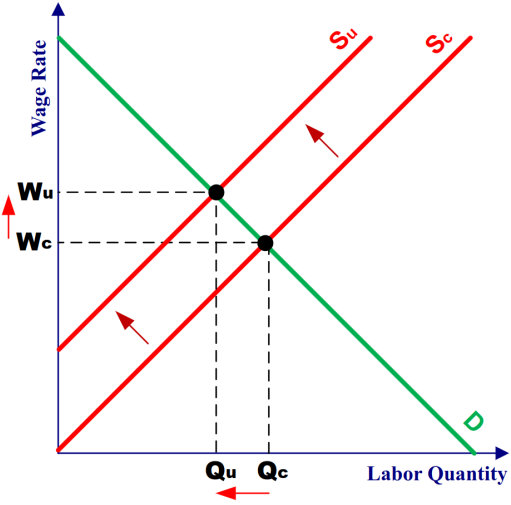


Collective Bargaining, for Economists



Collective Bargaining, for Lawyers

Collective Bargaining, for Economists



Collective Bargaining, for Lawyers

Does the text really matter?

"The difference between a one-page teaching contract [South Carolina] and a fifty-page teaching contract [New York] is that one of them has forty-nine extra pages of things that are good for teachers."

– Hamilton Nolan, *The Hammer* (2024, p. 47).

What a Union Contract Looks Like (first 3 pages)

TABLE OF CONTENTS

2005 – 2006 calendar

AGREEMENT

This Agreement ratified December 16,2005 is made and entered into between ST. CLAIR TECHNOLOGIES INC., Wallaceburg, Ontario (hereinafter called "the Company"), and the International Union, united Automobile, Aerospace and Agricultural Implement Workers of America (UAW-CLC) and its Local No. 251, (hereinafter called "the Union").

ARTICLE 1
RECOGNITION

1. The provisions of this Agreement shall apply to all employees covered by this Agreement without discrimination on account of race, creed, colour, sex, marital status, nationality, ancestry or place of origin.

2. Wherever the male noun or pronoun is used, it shall also mean the female.

3. The Company recognizes the Union as the sole bargaining agent of all its employees at Wallaceburg, Ontario, save and except supervisor, those above the rank of supervisor, office and sales staff, students for not more than twenty-four hours per week and students employed during the school vacation period (May 1st-September 15th). In case of reduction in force, students would be laid off first. Students will be paid at a rate to be determined by the Company, but will not be less than the Employment Standards Act.

4. The word "employee" or "employees" wherever used in this Agreement shall mean only the employees in the bargaining unit defined above unless the context otherwise provides.

5. The Company will negotiate with the Union for the purpose of adjusting any disputes which may arise concerning sickness and accident, wages, hours and working conditions.

ARTICLE 2

UNION SECURITY

1. All employees covered by this Agreement who are members of the Union at the signing date of this Agreement or who after become members thereof during the term of this Agreement, must retain their membership in the Union for the duration of the Agreement by paying the regular monthly dues levied against all members, as a condition of employment. All employees covered by this Agreement who are not members of the Union shall pay regular monthly dues, the same as the dues that are levied against those who are members of the union as a condition of employment.

2. All new employees, upon completion of thirty (30) days employment shall become members thereof in good standing in accordance with the constitution and bylaws of the Union for the life of this Agreement.

3. The Company will during the term of the Agreement, deduct initiation fees, monthly dues and assessments on a monthly basis from the pay cheque of all seniority employees, probationary employees and full-time students who have worked or been compensated for forty (40) hours in any one (1) month, or as required by the U.A.W. constitution, (full-time student being a student who works all or any time between May 1st and September 15th of the same year). Such deductions shall be credited to the Secretary-Treasurer of Local 251, not later than the tenth (10th) day of the calendar month next following the month in which such deductions are made. The Company and the Union will work out a mutually satisfactory arrangement by which the Company will furnish monthly records to the Financial Secretary of Local 251 of those from whom deductions were made, together with the amount of such deductions.

ARTICLE 3
MANAGEMENT RIGHTS

The Union recognizes and acknowledges that the management of the plant and direction of the working force are fixed exclusively in the Company and, without restricting the generality of the foregoing, the Union acknowledges that it is the exclusive function of the Company to:

1. Maintain order and efficiency

Text Pre-Processing Steps

- Contracts arrived as PDFs, along with matched metadata.
- Convert PDFs to machine-readable text (best was ABBBY FineReader)

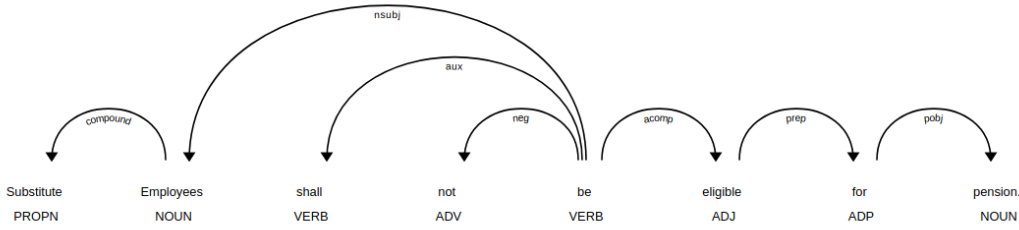
Text Pre-Processing Steps

- Contracts arrived as PDFs, along with matched metadata.
- Convert PDFs to machine-readable text (best was ABBBY FineReader)
- Exclude text for wage schedules, exhibits, appendices, etc.

Text Pre-Processing Steps

- Contracts arrived as PDFs, along with matched metadata.
- Convert PDFs to machine-readable text (best was ABBBY FineReader)
- Exclude text for wage schedules, exhibits, appendices, etc.
- Split the contracts into sections (RegEx) and sentences (spaCy):
  - 980,909 contract sections (33 per contract), 10.8 million sentences (11 per section)
- Co-reference resolution by section: replace pronouns with referent entity

Syntactic Parse for Contract Statements



- Dependency parsing (spaCy):
  - Output: Parse tree, giving functional relations between words in a sentence.
  - Identify syntactic subjects, and form statements around each subject
- Pipeline extracts clauses of the form: **Subject, Verb, Object**

Grammar Parse based on Legal Theory of Contracts

"Classic" legal linguistic indicators: e.g. must/shall indicate obligations, may/can indicate permissions (e.g. Hohfeld 1913, Balkin 1990).

Grammar Parse based on Legal Theory of Contracts

"Classic" legal linguistic indicators: e.g. must/shall indicate obligations, may/can indicate permissions (e.g. Hohfeld 1913, Balkin 1990).

- Subject categories:
  - worker, firm, union, manager
- Deontic modal verbs (deontic indicating "duty") capture necessity/possibility in social freedoms to act:
  - strict modals *(shall, will, must)* express necessity
  - permissive modals *(may, can)* express possibility
- Parser indicates negation ("shall **not**") and active/passive ("shall provide" vs "shall be provided")
- Special verbs:
  - *Obligation* Verbs (have to, ought to, be required, be expected, be compelled, be obliged, be obligated)
  - *Prohibition* Verbs (be prohibited, be forbidden, be banned, be barred, be restricted, be proscribed)
  - *Permission* Verbs (be allowed, be permitted, be authorized)
  - *Rights* Verbs (have, receive, retain)

Contract Statement Typology (Simplified)

Based on human (lawyer) annotation, machine assignments have precision of 91-99% (Ash et al, 2020).

Categorization Logic	Examples
Obligations	
Positive & Strict Modal & Active Verb	shall provide, shall include, shall notify, shall continue
Positive & Strict Modal & Obligation Verb	shall be required, shall be expected, shall be obliged
Positive & Non-Modal & Obligation Verb	is required, is expected
Prohibitions	
Negative & Any Modal & Active Verb	shall not exceed, shall not use, shall not apply
Negative & Permission Verb	shall not be allowed, is not permitted
Positive & Strict Modal & Constraint Verb	shall be prohibited, shall be restricted
Permissions	
Positive & Non-Modal & Permission Verb	is allowed, is permitted, is authorized
Positive & Strict Modal & Permission Verb	shall be allowed, shall be permitted
Positive & Permissive Modal & Active Verb	may be, may request, may use, may require, may apply
Negative & Any Modal & Constraint Verb	shall not be restricted, shall not be prohibited
Rights	
Strict Modal & Passive Verb	**shall be paid, shall be given, shall not be discharged**
Positive & Any Modal & Rights Verb	**shall have, shall receive, shall retain**
Negative & Any Modal & Obligation Verb	**shall not be required**

Summary Stats: Statement Type Shares

Subject	Clause Type				
	Obligation (%)	Prohibition (%)	Permission (%)	Right (%)	*Total (%)*
Worker	20.9	3.1	8.4	22.9	*55.3*
Firm	24.7	1.5	3.4	0.9	*30.5*
Union	7.0	0.6	2.0	2.1	*11.7*
Manager	1.7	0.1	0.4	0.2	*2.5*
Total	*54.4*	*5.3*	*14.1*	*26.2*	*100.0*

- Contracts consist mostly of worker rights (22.9%), worker obligations (20.9%) and firm obligations (24.7%)
- Firm rights are rare (0.9%); makes sense as management reserves rights.

"Worker Rights" Examples

1. Employees who retire as well as current retirees and survivors **will be provided with Life Insurance** in the amount of $6,000.
2. Where the Company schedules an employee to work in excess of seventy-seven (77) hours in one pay period, the **employee will be paid for the excess hours at the applicable overtime rate**.
3. Where an employee is prevented by circumstances beyond his control from returning to work on time, he **shall be paid for the holidays**.
4. However, where practicable, **senior employees in each job shall be given the opportunity to perform any available work** in that job, on their shift, within their Department.
5. An employee terminated during his probationary period would be **entitled to review under the grievance procedure** up to and including Step 3.

▸ Worker / Firm Obligations Examples

Getting Worker-Rights Topics

- Topic method:
  - Vectorize each worker rights clause with transformer-based context-sensitive sentence encoder (Reimers and Gurevych 2019): Pretrained S(entence)-BERT encoder to represent clauses as 768-dim vectors.
  - Uses context of sentences tuned to capture similar meanings (rather than word counts like LDA).
  - Construct topics using k-means clustering applied to the sentence embeddings.
- Three advantages over LDA:
  - allows word meanings to be interdependent, rather than independent. For example, our method registers that "employee" and "worker" are synonyms, whereas LDA treats those words as independent.
  - method learns context-sensitive representations. For example, the word "bank" can have a different meaning for bank tellers than for dock workers. LDA does not make such a distinction.
  - assigns each individual clause to a single topic, rather than a distribution across topics. That results in a simpler dataset, and makes more sense for short documents (single sentences) rather than long documents.

Back (Descriptives) Back (Tax) Back (Employment)

What do Worker-Rights Clauses Consist of?

Label	Frequency
Scheduling	0.26
Vacation	0.17
Health & Wellness	0.14
Seniority	0.12
Payments	0.11
Family Issues	0.10
Termination	0.10

Note: Clause topics constructed from embedding worker-rights clauses using MPNet, applying k-means clustering with $k = 32$, and aggregating up to 7 more interpretable topics. Other/miscellaneous topic (11%) omitted.

▸ Assigning Clauses to Topics

All Worker-Rights Topics ($k = 32$)

Topic Label	Frequency
Work Hours	0.058
Workplace Safety	0.054
Payment Rules	0.051
Vacations	0.045
Leaves of Absence	0.039
Seniority-Based Benefits	0.039
Seniority-Based Vacation	0.038
Holiday Pay	0.037
Position Classification	0.036
Recall	0.032
Grievance & Discipline	0.031
Job Security	0.03
Seniority & Promotion/Transfer	0.03
Scheduling	0.03
Bereavement Leave	0.028

Topic Label	Frequency
Notice Requirements	0.027
Parental Leave	0.027
Termination	0.026
Overtime	0.026
Holiday Work Pay	0.026
Shift Premiums	0.025
Sick Leave	0.025
Personnel Records	0.024
Workplace Injuries	0.024
Part-Time Employment	0.023
Reimbursements	0.022
Probation Period	0.015
Meals	0.015
Breaks	0.013
Jury Duty	0.009

Interpreting the Topics

- Summarization:
  - Sample 10 clauses from a topic; generate summaries with GPT-4 ("I will give you a list of clauses sampled from collective bargaining agreements. These clauses are on a related topic in terms of giving similar rights to workers. Summarize in one sentence what types of rights the clauses in this topic are giving to workers.").

Interpreting the Topics

- Summarization:
  - Sample 10 clauses from a topic; generate summaries with GPT-4 ("I will give you a list of clauses sampled from collective bargaining agreements. These clauses are on a related topic in terms of giving similar rights to workers. Summarize in one sentence what types of rights the clauses in this topic are giving to workers.").
- Distinguishing wage-like amenities from control rights:
  - U.S. BLS National Compensation Survey provides a list of benefits and amenities that impose a quantifiable cost to employers.
  - Extract list as plain-text phrases that would appear in contracts (incentive-based pay, a commission, a production bonus, a piece rate, a cost-of-living allowance, hazard pay, a uniform allowance, a tool allowance, free room and board, subsidized room and board, paid vacation leave, paid holiday leave, paid sick leave, paid personal leave, overtime pay, shift differential pay life insurance, health insurance, disability insurance, retirement benefits).
  - Construct simulated priced-amenity clauses as "Employees shall have ... [amenity phrase]", apply transformer sentence encoder to get embedding.
  - Compute cosine similiarity of topic embeddings to priced-amenity embeddings $\rightarrow$ topic clusters that are closest to the list of priced-amenity clauses on average are the most substitutable with wages, and vice versa.

Topic Label	Sim to Wages	Frequency	Topic Summary
Grievance & Discipline	**0.1590**	0.031	The clauses provide workers with rights related to disciplinary actions, grievance procedures, and representation, ensuring transparency, due process, and the ability to challenge or appeal employer decisions.
Recall	**0.1768**	0.032	The clauses provide rights related to job security and recall for workers who have been laid off, including options to accept vacant positions, refuse temporary recalls without penalty, and priority for rehiring in their former or equivalent positions if they become available.
Seniority & Promotion	**0.1792**	0.03	The clauses are granting workers rights related to job preference, promotion, and transfer based on seniority, qualifications, and experience.
Leaves of Absence	**0.2100**	0.039	The clauses provide workers with the right to take leaves of absence for union activities, public service, education, retraining, and other approved reasons, with varying conditions regarding pay and benefits.
Position Classification	**0.2115**	0.036	The clauses are providing workers with rights to receive pay adjustments or increases when they take on duties in higher paying classifications, substitute in higher paying roles, transfer to new positions with higher salary scales, or are temporarily appointed to positions of a higher pay grade.
Workplace Safety	**0.2118**	0.054	The clauses provide workers with rights related to workplace safety, health protection, and compensation in case of job loss due to technological changes, as well as opportunities for union engagement and training on safety procedures.
Scheduling	**0.2146**	0.03	The clauses provide workers with rights related to scheduling flexibility, compensation for working during non-standard hours or days off, and benefits during absences or layoffs.
Seniority-Based Benefits	**0.2194**	0.039	The clauses provide workers with rights related to pro-rated benefits, eligibility for allowances based on employment duration, credit for service and seniority during leaves, cost-sharing for benefits, and entitlements based on continuous service, including adjustments in pay and long-term disability plans.
Vacations	**0.2405**	0.045	The clauses are granting workers rights related to vacation entitlements, including the timing, duration, and pay during their vacation periods.
Payment Rules	**0.2414**	0.051	The clauses are giving workers rights related to the timing, frequency, and accuracy of their wage payments.

Validation of Worker-Rights Clauses

> If the parties mutually agree, the Company may hire temporary employees for short term periods not longer than 30 work days for non-routine work or special projects.

- Quite difficult: Scoring a given clause as "pro-worker" or not.

Validation of Worker-Rights Clauses

1. Employees who retire as well as current retirees and survivors will be provided with Life Insurance in the amount of $6,000.	2. If the parties mutually agree, the Company may hire temporary employees for short term periods not longer than 30 work days for non-routine work or special projects.

- Quite difficult: Scoring a given clause as "pro-worker" or not.
  - much easier: compare two clauses and say which one is more favorable to workers.

LLM Validation of Worker-Rights Clauses

1. Employees who retire as well as current retirees and survivors will be provided with Life Insurance in the amount of $6,000.	2. If the parties mutually agree, the Company may hire temporary employees for short term periods not longer than 30 work days for non-routine work or special projects.

- LLM coding (gpt-4o-mini):
  - Prompt: "Which of these sentences from a union collective bargaining agreement is more likely to be interpreted as an entitlement, benefit, or amenity for workers? Answer 'Definitely 1', 'Probably 1', 'Probably 2', 'Definitely 2', or 'Neither'. 1. [sentence 1]. 2. [sentence 2.]"

LLM Validation of Worker-Rights Clauses

1. Employees who retire as well as current retirees and survivors will be provided with Life Insurance in the amount of $6,000.	2. If the parties mutually agree, the Company may hire temporary employees for short term periods not longer than 30 work days for non-routine work or special projects.

- LLM coding (gpt-4o-mini):
  - Prompt: "Which of these sentences from a union collective bargaining agreement is more likely to be interpreted as an entitlement, benefit, or amenity for workers? Answer 'Definitely 1', 'Probably 1', 'Probably 2', 'Definitely 2', or 'Neither'. 1. [sentence 1]. 2. [sentence 2.]"
- Dataset:
  - 100 randomly sampled sentences for each of 16 clause types
    4 agents (worker, firm, union, manager) $\times$ 4 provisions (rights, obligations, prohibitions, permissions)
  - form across-clause-type pairs: $16 \times 15 \times 100$ clauses $= 24{,}000$ pairs
- **For each clause type (e.g. worker rights), compute % probability of being more pro-worker than other clause types.**

Ranking of Clause Types by Pair-Wise Pro-Worker Frequency

Clause Type	Clause Frequency (%)	Pro-Worker Frequency (%)
Worker Right	**22.9**	**80.9**
Union Right	2.1	67.8
Worker Permission	8.4	63.08
Manager Right	0.2	59.85
Firm Obligation	24.7	55.63
Worker Prohibition	3.1	55.51
Worker Obligation	20.9	55.33
Union Permission	2	46.33
Manager Prohibition	0.1	44.36
Firm Right	0.9	39.0
Union Obligation	7	38.74
Union Prohibition	0.6	38.73
Manager Obligation	1.7	38.5
Manager Permission	0.4	37.43
Firm Prohibition	1.5	36.17
Firm Permission	3.4	35.56

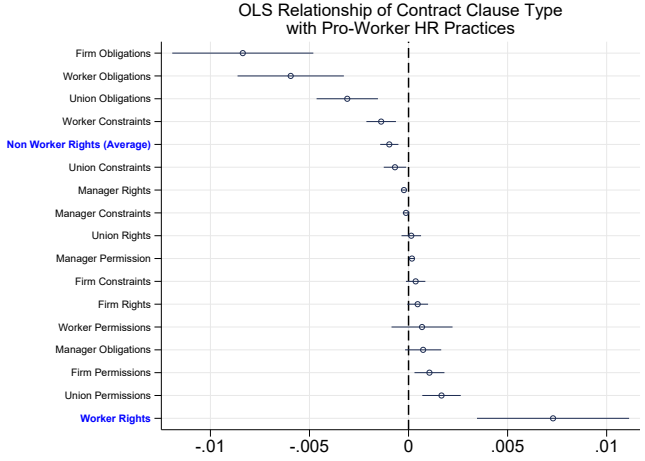
Note: Statistics from pairwise comparisons of clause types with GPT-3.5, as described in the text. Rows indicate clause types. Second column gives the frequency of that clause in the corpus; third column gives the proportion of pairwise comparisons where that category's clause is annotated as more beneficial to workers than the paired clause from another category. Sorted by third column.

Validation Against Pro-Worker HR Index

- Pro-Worker HR Index based on World Management Survey (Bloom et al, 2012)
  - Increases in "managers care about workers", "promotes good workers", "employees are valued"; decreases in "focus on top talent", "incentives", "fire poor performers"
  - Matched to 127 contracts by firm name and time.

Validation Against Pro-Worker HR Index

- Pro-Worker HR Index based on World Management Survey (Bloom et al, 2012)
  - Increases in "managers care about workers", "promotes good workers", "employees are valued"; decreases in "focus on top talent", "incentives", "fire poor performers"
  - Matched to 127 contracts by firm name and time.



Note: Figure presents coefficients and 95% confidence intervals of regression of contract clause types on index for Pro-Worker HR Practices. Outcome: Clause type, defined as share of clauses of given type (number of clauses of type in question over the number of all clauses) . Treatment: Standardized index of Pro-Worker HR Practices, defined as sum of approval rates to six statements about worker practices". Controls: None. Heteroscedasticity-robust standard errors.

Effect of 2000's Concession Bargaining on Auto Workers

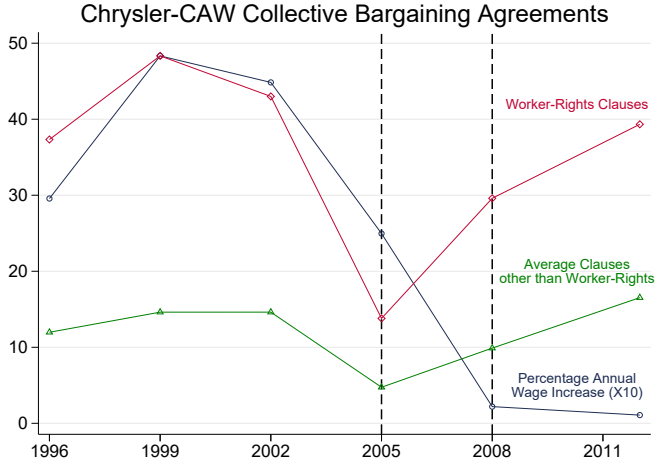
Canadian Auto Workers president Buzz Hargrove on 2005 agreement:

- "totally unprecedented....there was 'no business as usual' in this round of bargaining."
- "The companies started bargaining by demanding big concessions: like replacing wage increases with lump sums, abandoning COLA (even for pensioners), 10% co-pays on prescriptions, and giving up a week of paid time off per year."

Effect of 2000's Concession Bargaining on Auto Workers

Canadian Auto Workers president Buzz Hargrove on 2005 agreement:

- "totally unprecedented....there was 'no business as usual' in this round of bargaining."
- "The companies started bargaining by demanding big concessions: like replacing wage increases with lump sums, abandoning COLA (even for pensioners), 10% co-pays on prescriptions, and giving up a week of paid time off per year."



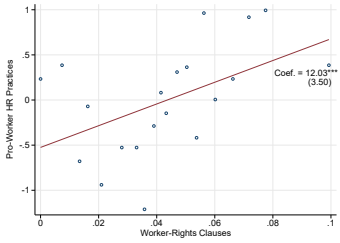
Comparison to Supervised-Learning-Based Method

- Lagos (2020) introduces a text-based measure of amenities in collective bargaining agreements based on "poaching":
  1. Vectorize contract clauses (e.g. using LDA topic shares) $\rightarrow \vec{L}$
  2. Fit a regression model to predict higher firm employment N with contract vectors, conditional on wages and FE.
  3. Higher amenities = higher $\hat{N}(\vec{L})$

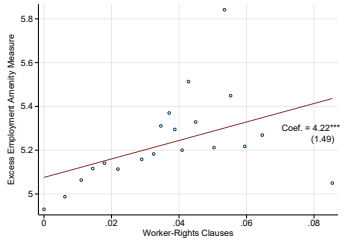
Comparison to Supervised-Learning-Based Method

- Lagos (2020) introduces a text-based measure of amenities in collective bargaining agreements based on "poaching":
  1. Vectorize contract clauses (e.g. using LDA topic shares) $\rightarrow \vec{L}$
  2. Fit a regression model to predict higher firm employment N with contract vectors, conditional on wages and FE.
  3. Higher amenities = higher $\hat{N}(\vec{L})$
- Following Lagos (2020), we train an LDA model on our contracts to get $\vec{L}$ and predict $\hat{N}$ using firm level employment, conditioning on province-sector-year wages and province-sector and sector-year FE.

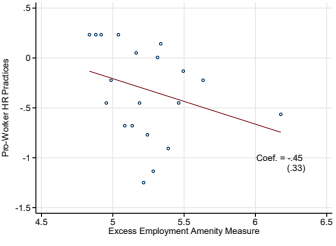
A. Pro-Worker HR Practices vs. Worker-Rights Clauses



B. Poaching Index vs. Worker-Rights Clauses



C. Poaching Index vs. Pro-Worker HR Practices



Note: Panel A: Binscatter plot of worker rights clauses (horizontal axis) and index for Pro-Worker HR Practices (vertical axis). Panel B: Binscatter plot of worker rights clauses (horizontal axis) and poaching index (vertical axis). Panel C: Binscatter plot of poaching index (horizontal axis) and index for Pro-Worker HR Practices (vertical axis). Worker rights clauses is defined as number of worker rights over the number of all clauses. Index of Pro-Worker HR Practices is defined as standardized sum of approval rates to six statements about worker practices; it increases in "managers care about workers", "promotes good workers", and "employees are valued," and decreases in "focus on top talent", "incentives", and "fire poor performers". Poaching index (text-predicted firm size) from Lagos (2020). Worker right measure significantly positively correlated with Pro-Worker HR Practices (Panel A). Poaching index significantly positively correlated with worker rights measure (Panel B), but not correlated with the pro-worker HR practices (Panel C). Data sources: Employment and Social Development Canada, World Management Survey (Bloom et al 2012).

Why Use Linguistic Features of Contracts?

- Why not measure wages and maybe some well-defined benefits like health insurance?

Why Use Linguistic Features of Contracts?

- Why not measure wages and maybe some well-defined benefits like health insurance?
  - For one: Many things that workers care about are not wages or pecuniary benefits – e.g. flexible scheduling, dignity at work (Dube et al 2022).

Why Use Linguistic Features of Contracts?

- Why not measure wages and maybe some well-defined benefits like health insurance?
  - For one: Many things that workers care about are not wages or pecuniary benefits – e.g. flexible scheduling, dignity at work (Dube et al 2022).
  - Contract terms matter to the parties: Firms and unions are spending a lot of money on contract drafting / labour lawyer services.

Why Use Linguistic Features of Contracts?

- Why not measure wages and maybe some well-defined benefits like health insurance?
  - For one: Many things that workers care about are not wages or pecuniary benefits – e.g. flexible scheduling, dignity at work (Dube et al 2022).
  - Contract terms matter to the parties: Firms and unions are spending a lot of money on contract drafting / labour lawyer services.
- Why not measure behavioral responses like hiring, strikes, and litigation?

Why Use Linguistic Features of Contracts?

- Why not measure wages and maybe some well-defined benefits like health insurance?
  - For one: Many things that workers care about are not wages or pecuniary benefits – e.g. flexible scheduling, dignity at work (Dube et al 2022).
  - Contract terms matter to the parties: Firms and unions are spending a lot of money on contract drafting / labour lawyer services.
- Why not measure behavioral responses like hiring, strikes, and litigation?
  - Litigation is rare – occurring only out of equilibrium. Well-designed contracts have no behavioral outputs, hence litigated contracts are selected sample.
  - hard to look at impacts of contract terms on strikes/litigation/etc.
  - still can look at language as outcome – how changes in incentives (e.g. tax rates, outside options) affect the language.

Options for Measuring "Pro-Worker" Clauses

1. Dictionaries: count pro-worker terms, e.g. "health benefits", "management rights", etc.
   - difficult to construct exhaustive list of pro-worker words/phrases
   - misses important context: directionality, negation

Options for Measuring "Pro-Worker" Clauses

1. Dictionaries: count pro-worker terms, e.g. "health benefits", "management rights", etc.
   - difficult to construct exhaustive list of pro-worker words/phrases
   - misses important context: directionality, negation
2. Supervised learning 1: hand-code clauses as pro-worker or pro-firm, train a classifier to extrapolate to whole dataset.
   - difficult to designate observed clauses as pro-worker or not – even labour lawyers are hesitant to do that except for a few special types of clauses.

Options for Measuring "Pro-Worker" Clauses

1. Dictionaries: count pro-worker terms, e.g. "health benefits", "management rights", etc.
   - difficult to construct exhaustive list of pro-worker words/phrases
   - misses important context: directionality, negation
2. Supervised learning 1: hand-code clauses as pro-worker or pro-firm, train a classifier to extrapolate to whole dataset.
   - difficult to designate observed clauses as pro-worker or not – even labour lawyers are hesitant to do that except for a few special types of clauses.
3. Supervised learning 2: predict a metadata variable, e.g. firm size – as in Lagos's (2020) text-based "poaching index".
   - relies on strong structural assumptions, especially given that observed text is equilibrium outcome.
   - classifier doesn't observe legal rules, brings in other confounding variation.

Options for Measuring "Pro-Worker" Clauses

1. Dictionaries: count pro-worker terms, e.g. "health benefits", "management rights", etc.
   - difficult to construct exhaustive list of pro-worker words/phrases
   - misses important context: directionality, negation
2. Supervised learning 1: hand-code clauses as pro-worker or pro-firm, train a classifier to extrapolate to whole dataset.
   - difficult to designate observed clauses as pro-worker or not – even labour lawyers are hesitant to do that except for a few special types of clauses.
3. Supervised learning 2: predict a metadata variable, e.g. firm size – as in Lagos's (2020) text-based "poaching index".
   - relies on strong structural assumptions, especially given that observed text is equilibrium outcome.
   - classifier doesn't observe legal rules, brings in other confounding variation.
4. Parser-based approach: combine legal knowledge and grammatical structure.
   - based on "classic" legal linguistic indicators (Hohfeld 1913, Balkin 1990): e.g. must/shall indicate obligations, may/can indicate permissions.
   - use syntactic parsers to attach rights/duties to regulated agents (e.g. worker, manager).

Why use grammar? Why not transformers all the way?

- Transformer-based NLP methods like BERT/GPT learn language by predicting masked tokens.
  - not informed by real-world inputs or outputs (outside the text).

Why use grammar? Why not transformers all the way?

- Transformer-based NLP methods like BERT/GPT learn language by predicting masked tokens.
  - not informed by real-world inputs or outputs (outside the text).
  - $\rightarrow$ associated language representations are not optimized for specific task of contract design.
  - $\rightarrow$ associated predictions come out of black box, may not be informative about worker rights/duties.

Why use grammar? Why not transformers all the way?

- Transformer-based NLP methods like BERT/GPT learn language by predicting masked tokens.
  - not informed by real-world inputs or outputs (outside the text).
  - $\rightarrow$ associated language representations are not optimized for specific task of contract design.
  - $\rightarrow$ associated predictions come out of black box, may not be informative about worker rights/duties.
- Our approach: Use legal knowledge on what the syntax means.
  - in principle, transformer models could be given such information during the training process.

Fixed-Effects Specification

$$Y_{psit} = \rho Z_{pst} + \alpha_{ps} + \alpha_{st} + \mathbf{X}'_{psit}\beta + \epsilon_{psit}, \tag{1}$$

- Y_{psit} = Text outcome (i.e. share of worker rights clauses) of contract adopted in province p, sector s, firm i, year t.

Fixed-Effects Specification

$$Y_{psit} = \rho Z_{pst} + \alpha_{ps} + \alpha_{st} + \mathbf{X}'_{psit}\beta + \epsilon_{psit}, \tag{1}$$

- Y_{psit} = Text outcome (i.e. share of worker rights clauses) of contract adopted in province p, sector s, firm i, year t.
- Z_{pst} = Economic treatment variable of interest:
  - log labor income tax rate τ at average income $\overline{y}_{p,t}$ in province p at year t:

$$Z_{pt} = \log(\tau_{p,t}(\overline{y}_{p,t}) + \tau_{\text{federal},t}(\overline{y}_{p,t}))$$

  - Z_{pst}= log employment rate in sector s of province p at year t

Fixed-Effects Specification

$$Y_{psit} = \rho Z_{pst} + \alpha_{ps} + \alpha_{st} + \mathbf{X}'_{psit}\beta + \epsilon_{psit}, \tag{1}$$

- Y_{psit} = Text outcome (i.e. share of worker rights clauses) of contract adopted in province p, sector s, firm i, year t.
- Z_{pst} = Economic treatment variable of interest:
  - log labor income tax rate τ at average income $\overline{y}_{p,t}$ in province p at year t:

$$Z_{pt} = \log(\tau_{p,t}(\overline{y}_{p,t}) + \tau_{\text{federal},t}(\overline{y}_{p,t}))$$

  - Z_{pst}= log employment rate in sector s of province p at year t
- α_{ps},α_{st} = province×sector and sector×year fixed effects.

Fixed-Effects Specification

$$Y_{psit} = \rho Z_{pst} + \alpha_{ps} + \alpha_{st} + \mathbf{X}'_{psit}\beta + \epsilon_{psit}, \tag{1}$$

- Y_{psit} = Text outcome (i.e. share of worker rights clauses) of contract adopted in province p, sector s, firm i, year t.
- Z_{pst} = Economic treatment variable of interest:
  - log labor income tax rate τ at average income $\overline{y}_{p,t}$ in province p at year t:

$$Z_{pt} = \log(\tau_{p,t}(\overline{y}_{p,t}) + \tau_{\text{federal},t}(\overline{y}_{p,t}))$$

  - Z_{pst}= log employment rate in sector s of province p at year t
- α_{ps},α_{st} = province×sector and sector×year fixed effects.
- $\mathbf{X}_{psit}$: Time-varying controls and additional FE, for robustness checks.

Fixed-Effects Specification

$$Y_{psit} = \rho Z_{pst} + \alpha_{ps} + \alpha_{st} + \mathbf{X}'_{psit}\beta + \epsilon_{psit}, \quad (1)$$

- Y_{psit} = Text outcome (i.e. share of worker rights clauses) of contract adopted in province p, sector s, firm i, year t.
- Z_{pst} = Economic treatment variable of interest:
  - log labor income tax rate τ at average income $\overline{y}_{p,t}$ in province p at year t:

$$Z_{pt} = \log(\tau_{p,t}(\overline{y}_{p,t}) + \tau_{\text{federal},t}(\overline{y}_{p,t}))$$

  - Z_{pst}= log employment rate in sector s of province p at year t
- α_{ps},α_{st} = province×sector and sector×year fixed effects.
- $\mathbf{X}_{psit}$: Time-varying controls and additional FE, for robustness checks.

→ Identification assumption: No time-varying province×sector-level confounders affecting both economic treatment and contract outcome.

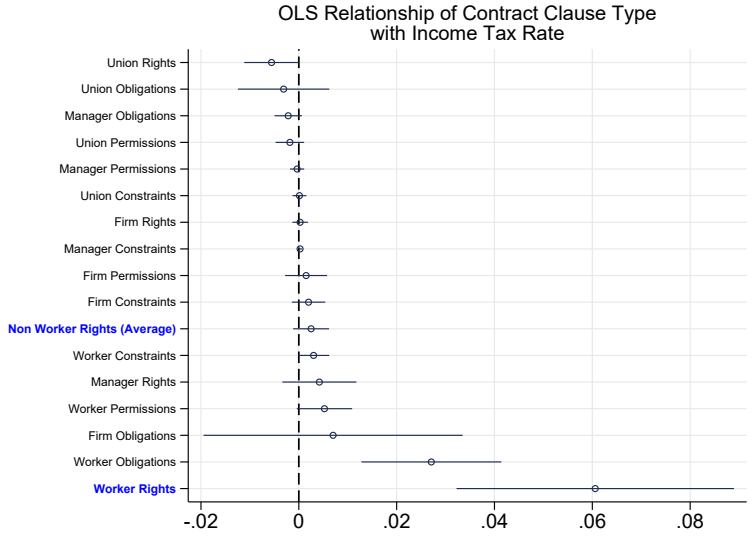
- exogenous timing motivated by pre-determined contract negotiation schedule.
- (in the data: treatment variables are unrelated to firm exits, the number of employees, and whether the employees have a COLA clause).

Effect of Income Tax Rate Change on Worker Rights

	Worker-Rights Clauses											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Log Income Tax Rate	0.060***	0.037***	0.041***	0.049***	0.060***	0.058***	0.060***	0.059***	0.060***	0.046***	0.035***	0.041***
	(0.014)	(0.011)	(0.014)	(0.018)	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)	(0.014)	(0.011)	(0.012)
R-Squared	0.15	0.16	0.55	0.34	0.15	0.15	0.15	0.15	0.15	0.30	0.47	0.16
Number of Observations	24,826	24,826	22,554	10,841	24,826	24,826	24,826	24,826	24,549	24,826	24,826	23,043
Province-Sector FEs	X	X	X	X	X	X	X	X	X	X	X	X
Sector-Year FEs	X	X	X	X	X	X	X	X	X	X	X	X
Province Trends		X										
Firm Fixed Effects			X									
Union Fixed Effects				X								
Cluster by Province					X							
Pro-Union Law Controls						X						
Anti-Union Law Controls							X					
NDP Party Control								X				
Employment Control									X			
Worker and Firm Obligation Control										X		
Share Parsed Clauses Control											X	
Drop Zero-Worker-Rights Clauses												X

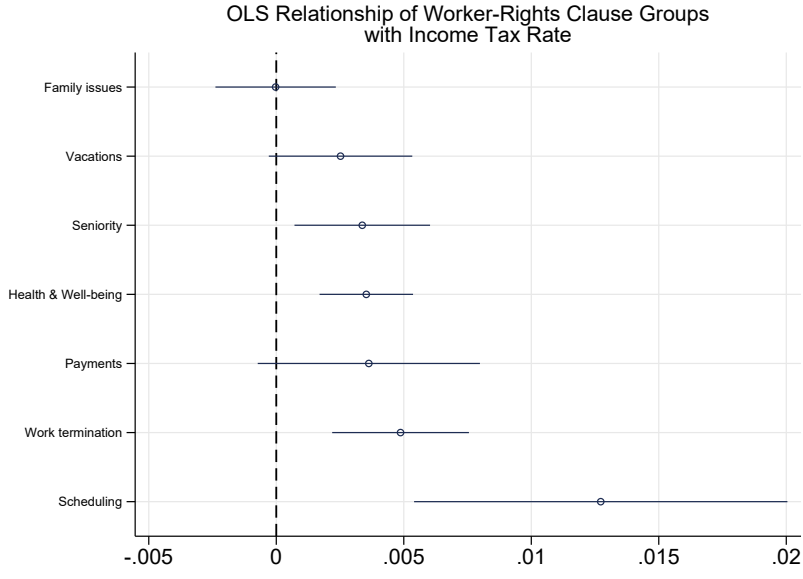
Note: Coefficients and standard errors of effect of labour tax rate on worker rights clauses, for different specifications as indicated in table footer. Outcome: Share of worker rights clauses, defined as number of worker rights clauses over the number of all clauses. Treatment: Labour tax rate is defined as logarithmized implicit personal income tax rate. Controls: Pro-Union (Anti-Union) Law Controls includes set of separate indicator variables for whether a given law favorable (unfavorable) to unions is in place. Inference: Standard errors clustered at the province-by-sector level, unless noted otherwise. Single, double, and triple asterisks indicate statistical significance at the 10%, 5%,and 1% levels, respectively.

Tax Effect is Specific to Worker Rights



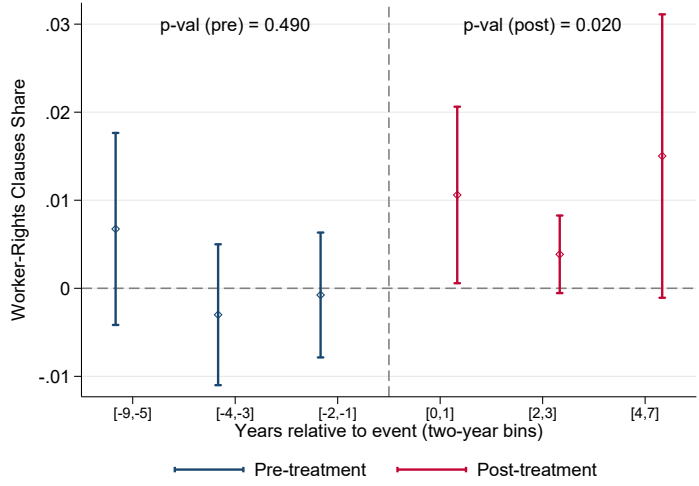
Note: Figure presents coefficients and 95% confidence intervals of effect of labour tax rate on contract clause types. Each coefficient is from a separate OLS regression. Outcome: Clause type, defined as the share of clauses of given type (number of clauses of type in question over the number of all clauses). Treatment: Labour tax rate, defined as logarithmized implicit personal income tax rate. Controls: Province-by-sector fixed effects and year-by-sector fixed effects. Inference: Standard errors clustered at the province-by-sector level. Data sources: Employment and Social Development Canada, Center for the Study of Living Standards.

Effect of Income Taxes on Worker Rights, by Topic



▸ Assigning Clauses to Topics

Event Study: Largest Discrete Increase in Tax Rates by Province



Note: Figure presents coefficients and 95% confidence intervals for time indicators before and after labour tax rate increase on share of worker-rights clauses. Callaway & Sant'Anna (2021) estimator, accounting for heterogeneous treatment effects and staggered treatment timing. Dynamic aggregation/event study effects, using doubly robust inverse probability weighting. Outcome: Worker rights share, defined number of worker rights over the number of all clauses. Controls: Not-yet-treated observations. Numbers on horizontal axis refer to final year of respective two-year bins; i.e., -1 = last two years prior to event. Event is defined as the largest labour tax increase in a given province in the 1990s, where labour tax rate is defined as implicit personal income tax rate. Inference: Standard errors clustered at the province-by-sector level.

▸ Event-Study: Tax Decrease

Instrumental Variables Strategy

- Idea for instrument: Exogenous changes in *province* income tax due to changes in *federal* tax rates, driven by associated deductions/credits (see Gruber and Saez, JPubE 2002; Akcigit, Grigsby, Nicolas, Stantcheva, QJE 2021).

Instrumental Variables Strategy

- Idea for instrument: Exogenous changes in *province* income tax due to changes in *federal* tax rates, driven by associated deductions/credits (see Gruber and Saez, JPubE 2002; Akcigit, Grigsby, Nicolas, Stantcheva, QJE 2021).
- Recall first Z_{pt} variable in fixed effects specification:

$$Z_{pt} = \log(\tau_{p,t}(\overline{y}_{p,t}) + \tau_{\text{federal},t}(\overline{y}_{p,t}))$$

Instrumental Variables Strategy

- Idea for instrument: Exogenous changes in *province* income tax due to changes in *federal* tax rates, driven by associated deductions/credits (see Gruber and Saez, JPubE 2002; Akcigit, Grigsby, Nicolas, Stantcheva, QJE 2021).
- Recall first Z_{pt} variable in fixed effects specification:

$$Z_{pt} = \log(\tau_{p,t}(\overline{y}_{p,t}) + \tau_{\text{federal},t}(\overline{y}_{p,t}))$$

- Instrument with variable constructed from Kevin Milligan's CTaCS tax calculator:

$$Z_{pt}^{\text{IV}} = \log(\tau_{p,t-k}(\overline{y}_{p,t-k}) + \tau_{\text{federal},t}(\overline{y}_{p,t-k}))$$

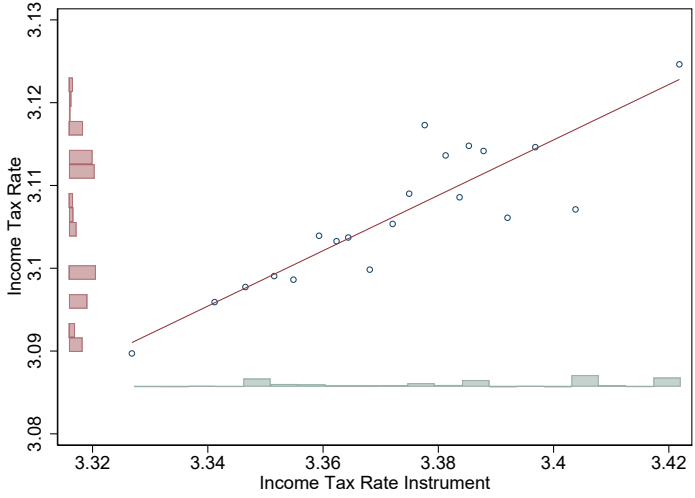
  where $k \in \{1,3\}$ is the lag in rates/income.
  - income is lagged $\rightarrow$ no endogenous income responses to taxes.
  - province rates are lagged $\rightarrow$ no endogenous tax responses to province-level confounders.
  - federal tax rates are not lagged $\rightarrow$ instrument isolates province tax changes due to changes in *federal* rates.

Tax IV First Stage

$$Z_{pt} = \gamma Z_{pt}^{IV} + \alpha_{ps} + \alpha_{st} + \mathbf{X}'_{psit}\beta + \eta_{psit}$$

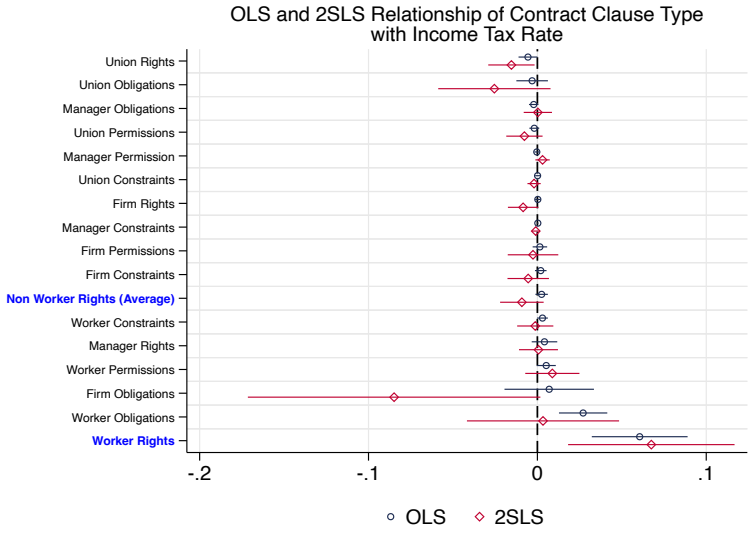
Tax IV First Stage

$$Z_{pt} = \gamma Z_{pt}^{IV} + \alpha_{ps} + \alpha_{st} + \mathbf{X}'_{psit}\beta + \eta_{psit}$$



Note: Figure presents first stage binscatter of labour income tax rate (vertical axis) and predicted income tax rate based on lagged rates and exemptions (horizontal axis). **Kleibergen-Paap First Stage F-Statistic = 65.08**. Controls: Province-by-sector fixed effects and year-by-sector fixed effects. Data sources: ESDC, Center for the Study of Living Standards, Statistics Canada.

Higher Income Taxes → More Worker Rights (OLS & IV)



Note: Figure presents OLS and 2SLS coefficients and 95% confidence intervals of effect of instrumented labour tax rate on contract clause types. Outcome: Clause type, defined as the share of clauses of given type (number of clauses of type in question over the number of all clauses). Treatment: Log Labour tax rate in province s and year t, for 2SLS instrumented as the log of the sum of federal income tax rate of year t, calculated for the average income of province s and year t-k, and the province income tax rate of province s and year t-k, calculated for the average income of province s and year t-k, for k=1. Each coefficient is from a separate regression. Controls: Province-by-sector fixed effects and year-by-sector fixed effects. Inference: Standard errors clustered at the province-by-sector level. Data sources: Employment and Social Development Canada, Center for the Study of Living Standards, Statistics Canada.

Labour Demand Shock

- Employment rate in sector $\times$ province $\times$ year is a measure of workers' outside option:
  - costliness of strike to employers – more difficult to hire replacements.
  - also a measure of labour demand.

Labour Demand Shock

- Employment rate in sector $\times$ province $\times$ year is a measure of workers' outside option:
  - costliness of strike to employers – more difficult to hire replacements.
  - also a measure of labour demand.
- Use leave-one-out sectoral employment rate ($\times$ province by year) as Bartik shifter:

$$Z_{pst} = \frac{1}{n_s - 1} \sum_{k \neq s} \log(\text{Emp}_{pkt})$$

  - i.e., average log employment rate in sectors besides s in province p at year t
  - helps to isolate outside-option component.
- Positive labour demand shock improves bargaining position of unions relative to firms: We expect an increase in worker rights.

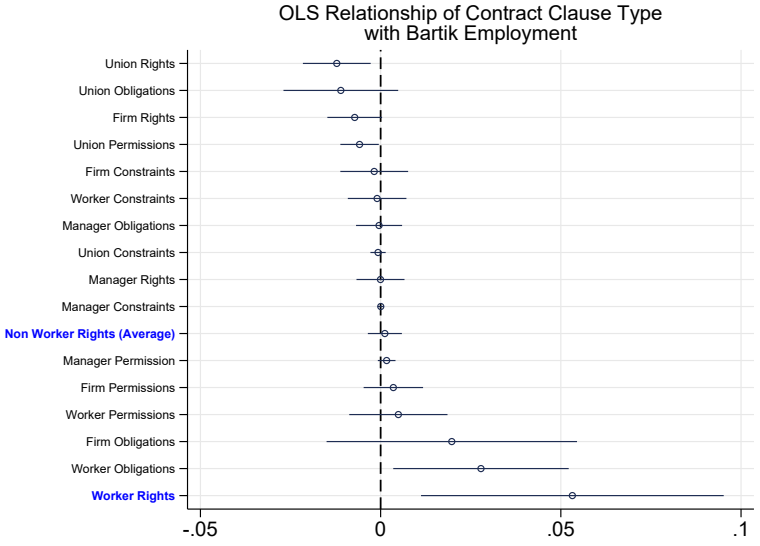
Better Outside Option → More Worker Rights

	Worker-Rights Clauses											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Log Emp. Rate	0.053**	0.050***	0.040**	0.055**	0.053*	0.078***	0.056***	0.050**	0.049**	0.037**	0.035**	0.052***
	(0.021)	(0.019)	(0.018)	(0.027)	(0.024)	(0.019)	(0.017)	(0.020)	(0.021)	(0.018)	(0.014)	(0.017)
R-Squared	0.15	0.16	0.56	0.36	0.15	0.15	0.15	0.15	0.15	0.31	0.47	0.16
Number of Observations	29,157	29,157	26,669	13,735	29,157	27,603	27,603	29,157	29,157	29,157	29,157	27,108
Province-Sector FEs	X	X	X	X	X	X	X	X	X	X	X	X
Sector-Year FEs	X	X	X	X	X	X	X	X	X	X	X	X
Province Trends		X										
Firm Fixed Effects			X									
Union Fixed Effects				X								
Cluster by Province					X							
Pro-Union Law Controls						X						
Anti-Union Law Controls							X					
NDP Party Control								X				
Employment Control									X			
Worker and Firm Obligation Control										X		
Share Parsed Clauses Control											X	
Drop Zero-Worker-Rights Clauses												X

Note: Coefficients and standard errors of effect of Bartik-style leave-one-out employment rate on worker rights clauses, for different specifications as indicated in table footer. Outcome: Share of worker rights clauses, defined as number of worker rights clauses over the number of all clauses. Treatment: Bartik-style leave-one-out employment rate in a given sector, defined as the logarithmized average over the employment rates in other sectors. Controls: Pro-Union (Anti-Union) Law Controls includes set of separate indicator variables for whether a given law favorable (unfavorable) to unions is in place. Employment control controls for logarithmized employment rate (own sector). Inference: Standard errors clustered at the province-by-sector level, unless noted otherwise. Single, double, and triple asterisks indicate statistical significance at the 10%, 5%,and 1% levels, respectively.

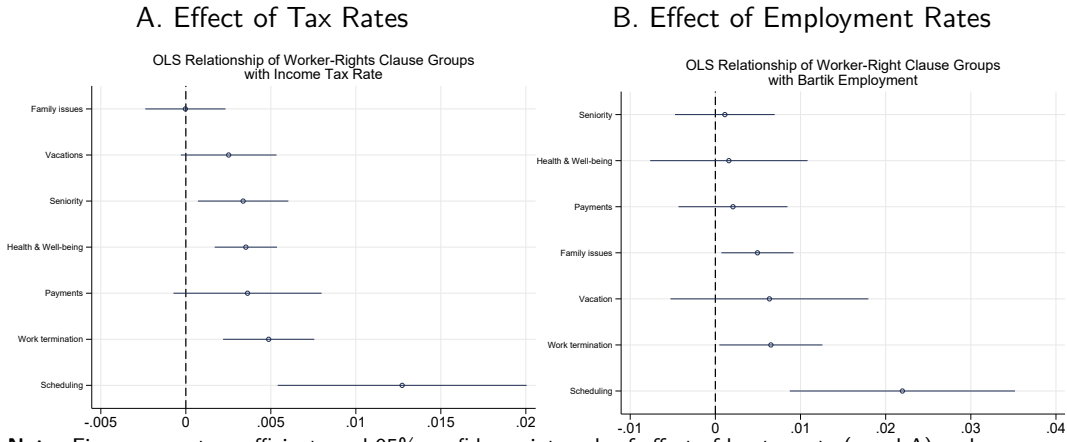
▸ Event Study ▸ Wage Regression

Better Outside Option → More Worker Rights



Note: Figure presents coefficients and 95% confidence intervals of effect of Bartik-style leave-one-out employment rate on contract clause types. Outcome: Clause type, defined as share of clauses of given type (number of clauses of type in question over the number of all clauses). Treatment: Bartik-style leave-one-out employment rate in a given sector, defined as the logarithmized average over the employment rates in other sectors. Controls: Province-by-sector fixed effects and year-by-sector fixed effects. Inference: Standard errors clustered at the province-by-sector level.

Taxes and Employment Shift the Same Topics



Note: Figure presents coefficients and 95% confidence intervals of effect of log tax rate (panel A) and Bartik-style leave-one-out log employment rate (Panel B) on worker right topics. Outcome: Worker-rights topic, defined as share of worker rights clauses that belong to given topic (number of clauses of topic in question over the number of all clauses). Controls: Province-by-sector fixed effects and year-by-sector fixed effects. Standard errors clustered at the province-by-sector level.

▸ Assigning Clauses to Topics

Valuing Worker Rights in Terms of Wages

	(1) Share Worker Rights (S.D.)	(2) Log Wages
Log Tax Rate	**2.34*** **	**0.23*** **
	(0.55)	**(0.06)**
Union		1.34***
		(0.50)
Log Tax Rate * Union		**-0.36** **
		(0.16)
R-Squared	0.15	0.31
Number of Obs	24,826	4,877,128
Province-Sector FEs	X	X
Sector-Year FEs	X	X
Dataset:	Union Contracts	Labour Force Survey

- In response to a 10% increase in income taxes:
  - share of worker-rights clauses increases by 0.23 standard deviations ($\frac{10}{100} \times 2.34$)
  - union wages fall by 1.3% ($\frac{10}{100} \times (0.23 - 0.36)$)

→ **One std deviation increase in share of worker-rights clauses is worth about** $1.3\% \times \frac{1}{0.23} =$ **5.7% of wages**.

Amenity Value of Worker Rights

One standard deviation increase in share of worker-rights clauses is worth about 5.7% of wages.

- Compare to:
  - Mas and Pallais (2017): option for remote work worth 8% of wages.
  - Lagos (2020): CBA employment protection worth 4% of wages.
  - Dube, Naidu, Reich (2021): one s.d. of "workplace dignity" worth 6% of wages.
  - Anelli and Koenig (2023): reducing workplace fatality risk by 1 in 100,000 is worth 9% of wages.
  - Roussille and Scuderi (2023): a one S.D. increase in amenities (in job posts) worth about 12% of wages.

Summary

- We demonstrate that the value of collective bargaining agreements is in worker rights clauses:
  - Personal Income Tax $\nearrow$ or Outside Option $\nearrow$: Increase in worker-rights clauses.
  - Substitution of wage and non-wage compensation.

Summary

- We demonstrate that the value of collective bargaining agreements is in worker rights clauses:
  - Personal Income Tax $\nearrow$ or Outside Option $\nearrow$: Increase in worker-rights clauses.
  - Substitution of wage and non-wage compensation.
- Evidence in support of Simon (1951): employment is an authority contract with constraints on employer power.
  - contracts allow employers to commit to providing protections and amenities that they otherwise would not provide ex post.
  - these constraints on employer discretion have real value for employees.

Summary

- We demonstrate that the value of collective bargaining agreements is in worker rights clauses:
  - Personal Income Tax $\nearrow$ or Outside Option $\nearrow$: Increase in worker-rights clauses.
  - Substitution of wage and non-wage compensation.
- Evidence in support of Simon (1951): employment is an authority contract with constraints on employer power.
  - contracts allow employers to commit to providing protections and amenities that they otherwise would not provide ex post.
  - these constraints on employer discretion have real value for employees.
- Structural text analysis is useful for studying legal texts and their economic value.
  - grammar-based approach to worker rights still simplifies a lot, could likely be improved further combining AI with legal expertise.