The Extent of Repetition in Contract Language

Dan Simonson, Daniel Broderick & Jonathan Herr

BLACKBOILER

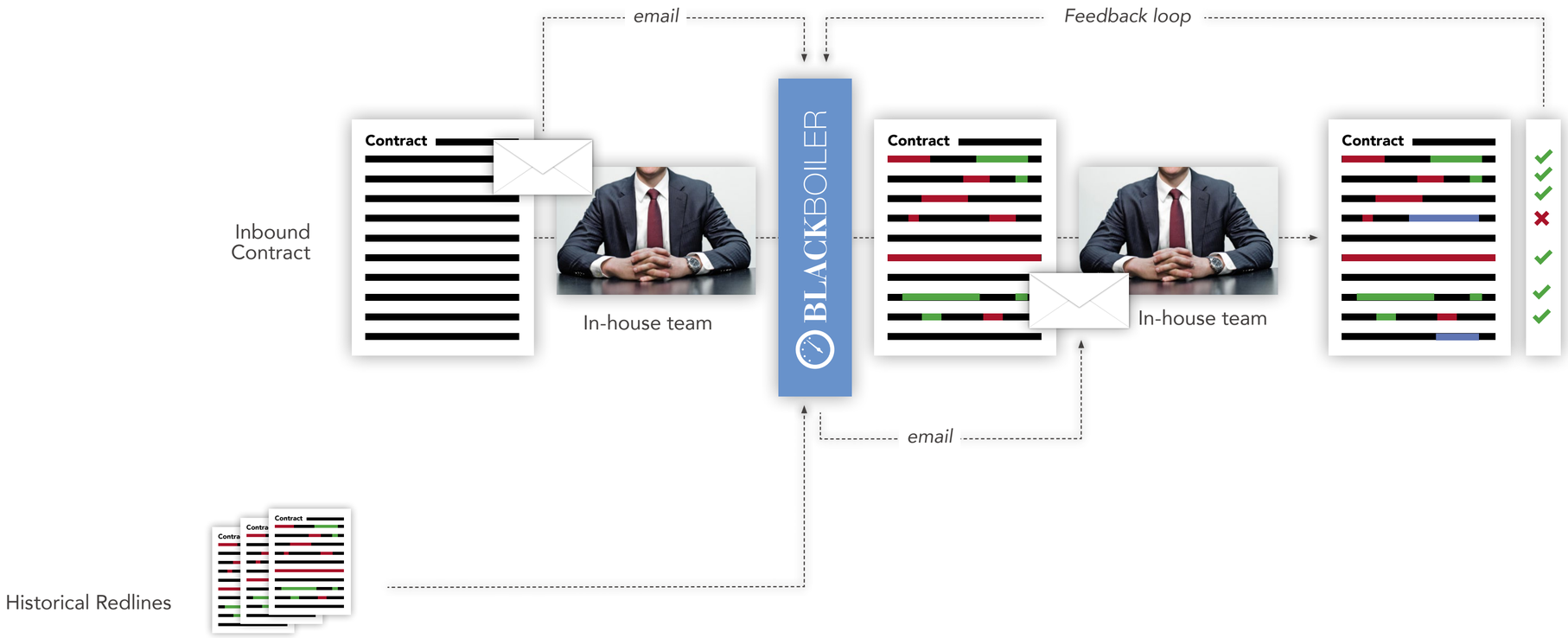
The First Natural Legal Language Processing (NLLP) Workshop,
June 7th, 2019

NAACL

dan.simonson@blackboiler.com

@thedansimonson

What is BlackBoiler?



Prior Work

Work on properties of conventionalized language:

- Halliday (1988), Argamon et al. (2005), Degaetano-Ortlieb and Teich (2017).

Many studies on contract corpora:

- Most were small (Faber and Lauridsen 1991, Anesa 2007, Carvalho 2008, Mohammed et al 2010, Curtotti and McCreath 2011)
- Anderson and Manns (2017) looked at a large corpus, but primarily with edit distance, looking at the graph of relationships between contracts.

The Gap:

- Corpus studies, but small, with no direct comparison to typical NLP data.
- Anderson and Manns (2017) big, but did not document the sort of metrics we wanted to know about.

Core Questions

- Can we demonstrate that contracts are different from “typical” natural language text (newswire, etc.)?
  1. Does that statement hold up quantitatively?
  2. To what extent is it true?
- Are we looking at something that is fundamentally language?
- To many of us the obvious answer is “yes,” but...

Another Point of View

- “For the purpose of AI training, [technical legal] language cannot be considered a natural language. For contract review and approval, Natural Language Processing (NLP) and off-the-shelf solutions do not work.”

(https://images.law.com/contrib/content/uploads/documents/397/5408/lawgeex.pdf)

Introduction

- We demonstrate some linguistic properties of contract language (English) and directly compare these with other genres of text (English).
- Contract language is more repetitive than other genres – in very particular ways.
- It’s parametrically different from other types of language, but it’s still language.

What should I get out of this?

Target audiences:

1. Computational linguists working in the legal domain.
2. Other computational linguists (to convince them to work in the legal domain!)
3. People interested automating their own contracting process.

Introduction

- Introduction
- Data
- Rank Counts Analysis (token scope)
  1. Hapax Legomena
  2. Pronoun Use
- Nearest Neighbors (sentential scope)
- Discussion

Data

- Contract Corpus
- Baseline Corpora

Data: Contract Corpus

Our subject matter expert acquired documents through search engines and through EDGAR.

Data: Contract Corpus

Non-Disclosure Agreements: 2,472 toks / doc

8. NOTICE OF UNAUTHORIZED USE OR DISCLOSURE. Recipient shall notify Accuride immediately upon
of any unauthorized use or disclosure of Confidential Information or any other breach of this Agreement by Rec
or any third party, and will cooperate with Accuride in every reasonable way to help Accuride regain possession
Confidential Information and prevent its further unauthorized use or disclosure.

9. OWNERSHIP AND RETURN OF CONFIDENTIAL INFORMATION. All Confidential Information disclos
Recipient shall be and remain the property of Accuride. Upon Accuride's written request,Recipient shall prompt
Confidential Information (including all originals, copies, reproduction sand summaries of such Confidential Inf
or certify its destruction in writing, and keep the same confidential and secret in accordance with this Agreemer

10. NO LICENSE. Nothing contained in this Agreement shall be construed as granting or conferring to Recipier
or license or otherwise, either expressly or by implication, in or to any Confidential Information disclosed by Ac
Recipient as a result of this Agreement, including, without limitation, rights or license under any present or futu
patent application, copyright, trademark,service mark, trade secret or other proprietary information owned, lice

Data: Contract Corpus

Purchase Orders: 8,443 toks / doc

8. INDEMNITY CLAUSE: The Contractor will release, protect, indemnify and hold the STATE and the respecti
subdivisions and their officers, agencies, employees, harmless from and against any damage, cost or liability, incl
reasonable attorney's fees for any or all injuries to persons, property or claims for money damages arising from a
omissions of the Contractor, his employees or subcontractors or volunteers.

9. EMPLOYMENT PRACTICES CLAUSE: The Contractor agrees to abide by the provisions of Title VI and VII
Civil Rights Act of1964 (42USC 2000e) which prohibits discrimination against any employee or applicant for em
or any applicant or recipient of services, on the basis of race, religion, color, or national origin; and further agree
by Executive Order No. 11246, as amended,which prohibits discrimination on the basis of sex; 45 CFR 90 which
discrimination on the basis of age; and Section 504 of the Rehabilitation Act of 1973, or the Americans with Dis
Act of 1990 which prohibits discrimination on the basis of disabilities. The Contractor agrees to abide by Utah's
Order, dated March 17, 1993, which prohibits sexual harassment in the work place.

10. SEVERABILITY: If any provision of this contract is declared by a court to be illegal or in conflict with any la

Data: Contract Corpus

Service Agreements: 8,881 toks / doc

8. REGULATORY, LEGAL AND SUPPLIER CHANGES. Customer acknowledges that the Service may be subje
or in part, to one or more provisions of state or federal tariffs filed by Pulsar360 or its suppliers and carriers. In t
of any conflict between any provision of the Agreement and any provision of such tariff, the provision of such ta
control. The Agreement and the Services shall be subject to such modifications as may be required or authorized
regulatory agency in the exercise of its lawful jurisdiction.

Customer acknowledges that certain of Pulsar360's suppliers and carriers establish prices charged to Pulsar360 a
terms on which such suppliers and carriers sell services to Pulsar360 based on governmental laws, rules, regulat
and decisions. If any of the prices charged to Pulsar360 by any of its suppliers increase or if any of the terms of s
change as a result of changes to governmental rules, laws or regulations or pursuant to new decisions or orders i
applicable regulatory or judicial bodies, or by unilateral action by suppliers, Pulsar360 reserves the right to incre
price charged to Customer and change the terms of Service hereunder, effective 30 days following notice to Cust
Customer does not agree to accept new pricing and revised terms, Customer may terminate the affected Service
penalty within 30 days of the date of such notice. Any continued use of the Services 30 days after the notice date

Data: Contract Corpus

Prime Contracts: 31,138 toks / doc

F. Education, Counseling, and Training Programs. All educational, counseling and vocational guidance program
apprenticeship and on-the-job training programs, under this Contract, shall be open to all qualified persons, wi
regard to race, sex, color, religion, national origin or ancestry. Such programs shall be conducted to encourage th
development of the interests, skills, aptitudes, and capacities of all students and trainees, with special attention t
problems of culturally deprived, educationally handicapped, or economically disadvantaged persons. Expansion
opportunities under these programs shall also be encouraged with a view toward involving larger numbers of pa
from these segments of the labor force where the need for upgrading levels of skills is the greatest.

G. Occupational Safety and Health. The Contractor shall comply with all the provisions of the Federal Occupati
and Health Act of 1970 (29 U.S.C. Section 651 et seq.) and all rules, regulations, and orders adopted pursuant th
Contractor shall comply with all the provisions of the California Occupational Safety and Health Act of1973 (La
Section 6300 et seq.) and all rules, regulations and orders adopted pursuant thereto. These laws provide for job s
health protection for workers. The Contractor shall obtain copies of such safety orders as are applicable to the ty
to be performed and shall be governed by their requirements in all construction operations. The Contractor sha

Data: Contract Corpus

Subcontracts: 12,388 toks / doc

8. SUBCONTRACTOR EMPLOYER.
Subcontractor has the status of an employer as defined by the Industrial Insurance, Workmen's Compensation A
Security, and other similar acts of the federal, state and local Governments. Subcontractor will withhold from its
applicable Social Security taxes, Workmen's Compensation, Unemployment Compensation contributions and w
taxes and pay the same. The Contractor shall in no way be liable as an employer to or on account of any of the e
of the Subcontractor. Before final payment is made upon this Subcontract,Subcontractor shall furnish satisfacto
to the Contractor that he has conformed to the laws, rules and regulations, and the Subcontractor agrees to inde
Contractor for any and all liability under such laws arising from the work performed under this Subcontract Ag

9. PERMITS, TAXES, TEMPORARY FUNCTIONS.
Subcontractor shall secure and pay for all permits, fees and licenses necessary for the performance of the Subco
Subcontractor shall pay any and all federal, state and municipal taxes, including sales taxes, if any,for which the
Subcontractor may be liable in carrying out the Subcontract. Subcontractor shall be responsible for all temporar
associated with its work, including but not limited to, lighting, wiring, protection, hoisting,scaffolding, rigging, f

Data: Contract Corpus

- Documents with OCR issues.

1. Removed (to the extent possible).
2. Contracts, without a standard for interchange, will likely include such errors as part of what we see exchanged.

Data: Contract Corpus

# Tokens	15,553,213
# Docs	1,737
Tokens / Doc	8,954

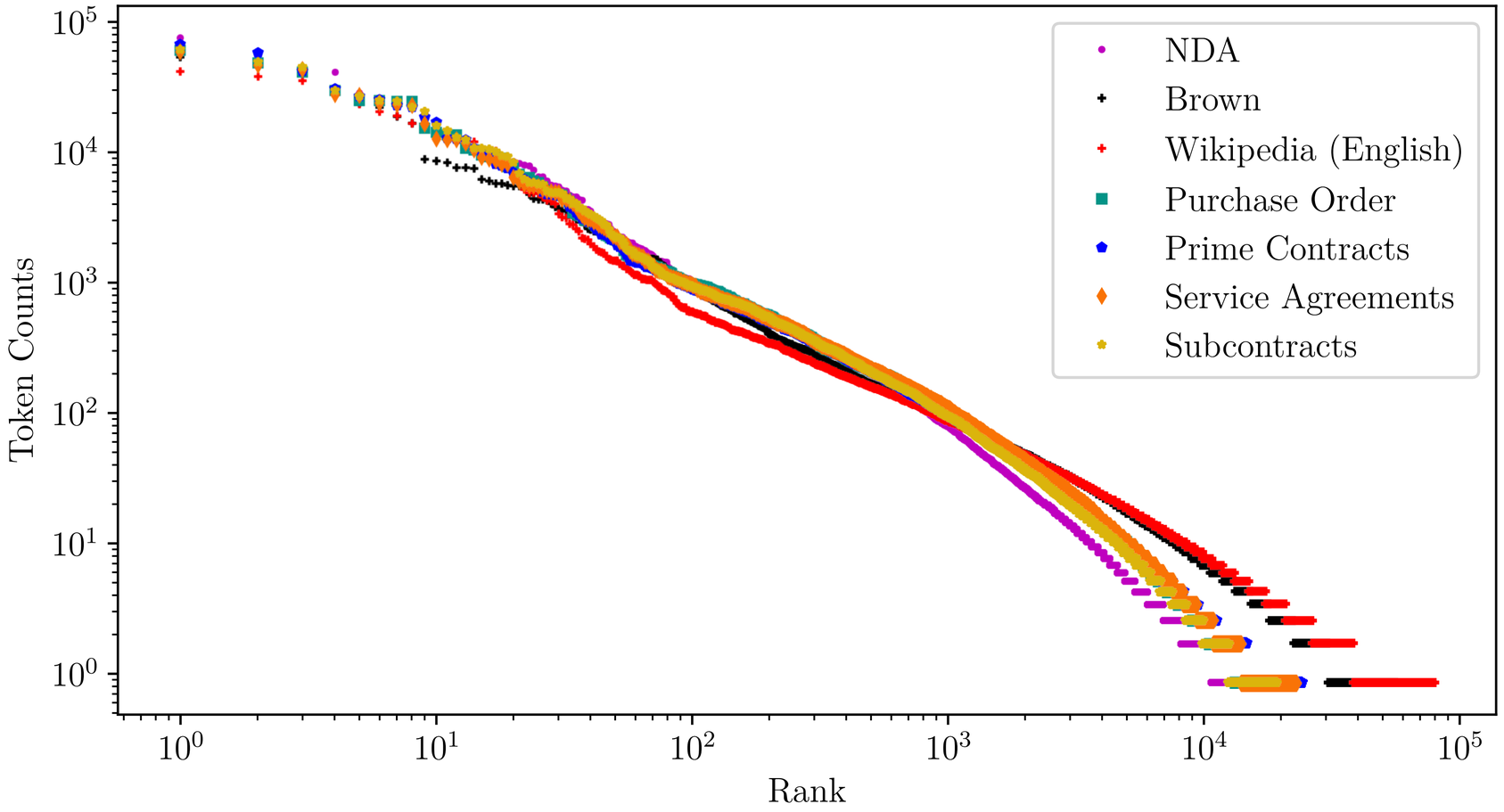
Data: Baseline Corpora

- Brown Corpus (Francis and Kučera 1964, 1971, 1979)
  1. Been studied for sixty years, hard to get more “typically studied” than that.
- English Wikipedia (King 2018)

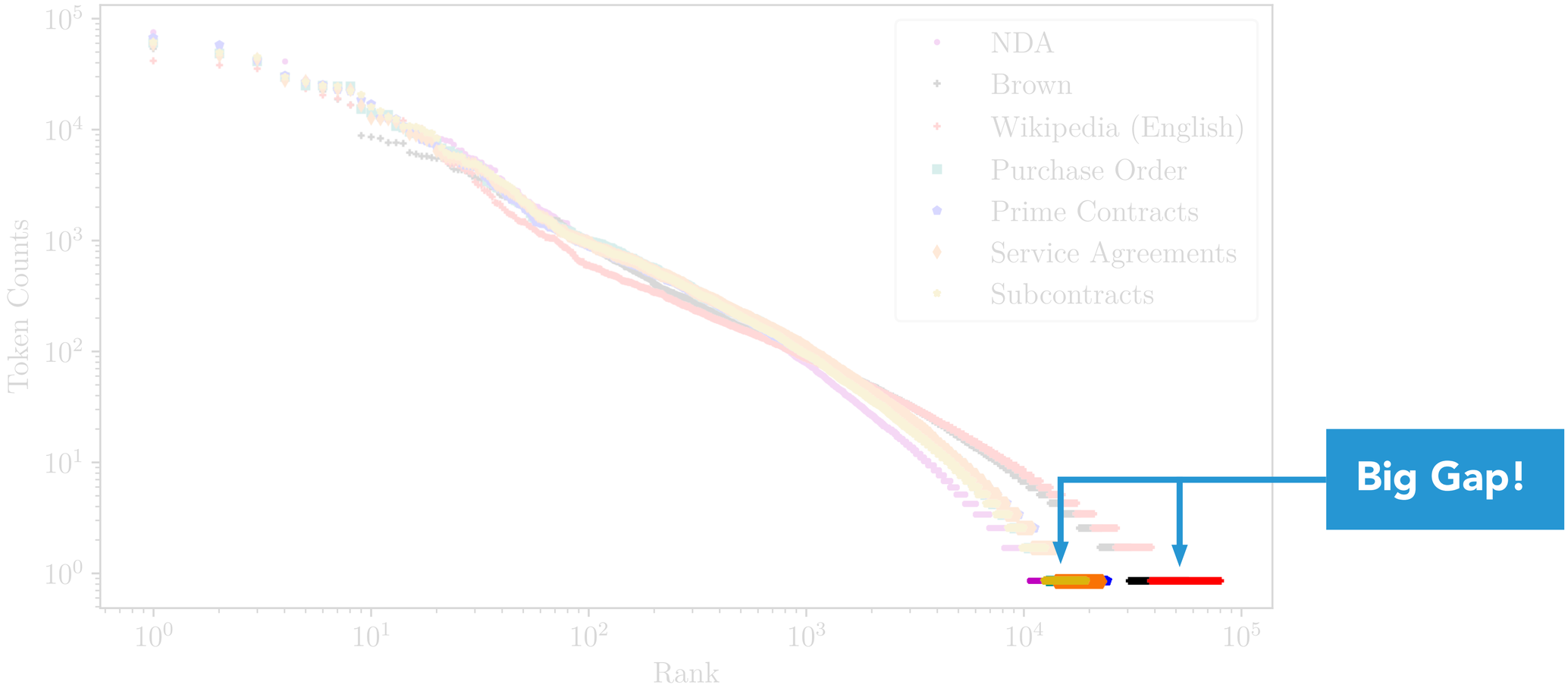
Rank–Counts Analysis

- Also known as a “Zipfian” analysis (Zipf 1949).
- Subsampled each subcorpus down to the size of the smallest one.
  1. Brown: 1.16 million tokens.
- Looking at raw tokens here.
  1. As given by SpaCy (Honnibal and Johnson 2015).
  2. Lemma: “-PRON-”.
  3. Case Sensitive defined terms.

Rank–Counts Analysis



Rank–Counts Analysis



Rank–Counts Analysis

- Hapax Legomena:
  1. Words that appear once in a corpus.
  2. Every corpus has some.
  3. Often problematic: get “UNKed,” etc.

Rank–Counts Analysis

Series	Hapax	H/Tok
Brown	25,559	2.20%
Wikipedia EN	40,820	3.52%
NDA	6,837	0.59%
Purchase Order	8,404	0.72%
Prime Contracts	9,461	0.81%
Services Agreements	8,915	0.77%
Subcontracts	6,670	0.57%

Rank–Counts Analysis

Series	Hapax	H/Tok
Brown	25,559	2.20%
Wikipedia EN	40,820	3.52%
NDA	6,837	0.59%
Purchase Order	8,404	0.72%
Prime Contracts	9,461	0.81%
Services Agreements	8,915	0.77%
Subcontracts	6,670	0.57%

5 hapax in Wikipedia for every 1 in Prime Contracts

Rank–Counts Analysis: Inspection

Series	Sample of Hapax Legomena Tokens
Brown	‘ARF’, ‘Piraeus’, ‘flint’, ‘Volta’, ‘paterollers’, ‘Schmalma’, ‘melanderi’, ‘bongo’, ‘hard-to-get’, ‘Beloved’, ‘miniscule’, ‘Tower’, ‘temerity’, ‘Fay’, ‘avidly’, ...
Wikipedia EN	‘appropriates’, ‘Puschmann’, ‘Muin’, ‘AC.7’, ‘sensing’, ‘Ambas’, ‘Kalutara’,‘Arnott’, ‘Ogrskem’, ‘48/73’, ‘Jayan’, ‘MK2020’, ‘beauticians’, ...
NDA	‘disapprove’, ‘mostly’, ‘wri+en’, ‘15260’, ‘48104’, ‘Loving’, ‘EXCLUSIVE’,‘Culver’, ‘Chih’, ‘Hwa’, ‘inch’, ‘Behalf’, ‘Opinions’, ‘HD8’, ‘appropriated’, ...
Purchase Order	‘ASNs’, ‘FRED’, ‘Party(i)wherethereceivingPartyistheSupplier’ ‘overturn’,‘Navigation’, ‘work.(iii’, ‘PLU’, ‘CDI’, ‘DFFRUGDQFH’, ‘INFRINGE’, ...
Prime Contracts	‘executers’, ‘Quote’, ‘derrick’, ‘FP-1’, ‘FP-3’, ‘FP-2’, ‘00:00’, ‘Ceiling’,‘EQUITABLE’, ‘OBTAIN’, ‘fan’, ‘ticket’, ‘prolonged’, ‘Macao’, ‘19.2.2(c’, ...
Services Agreements	‘Sophia(R’, ‘sacrifice’, ‘adhesives’, ‘transloader’, ‘totheChangeinwriting’,‘salient’, ‘simulate’, ‘KG’, ‘15/29’, ‘divert’, ‘ownedbyCorsearch’, ‘biologic’, ...
Subcontracts	‘ENCOURAGED’, ‘closer’, ‘INTRODUCTION’, ‘projecting’, ‘14607’, ‘CUT’,‘Higher’, ‘interfaces’, ‘percipient’, ‘takeover’, ‘postponement’, ‘timesheet’, ...

Rank–Counts Analysis: Inspection

- Some noise present.
  1. Just part of the reality of contract text.
  2. But even cleaning this noise would amplify the effect observed.

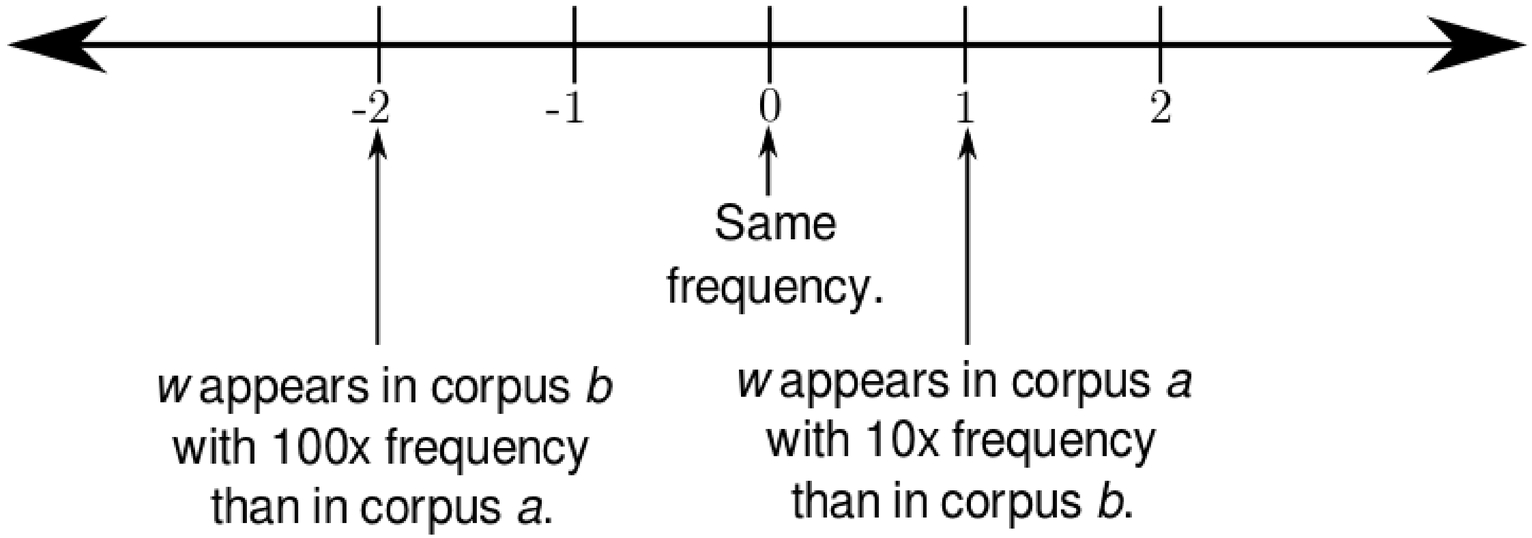
Rank–Counts Analysis: Inspection

- Pronouns
  1. Reverse of hapax.
  2. Even these pattern differently.

Rank–Counts Analysis: Inspection

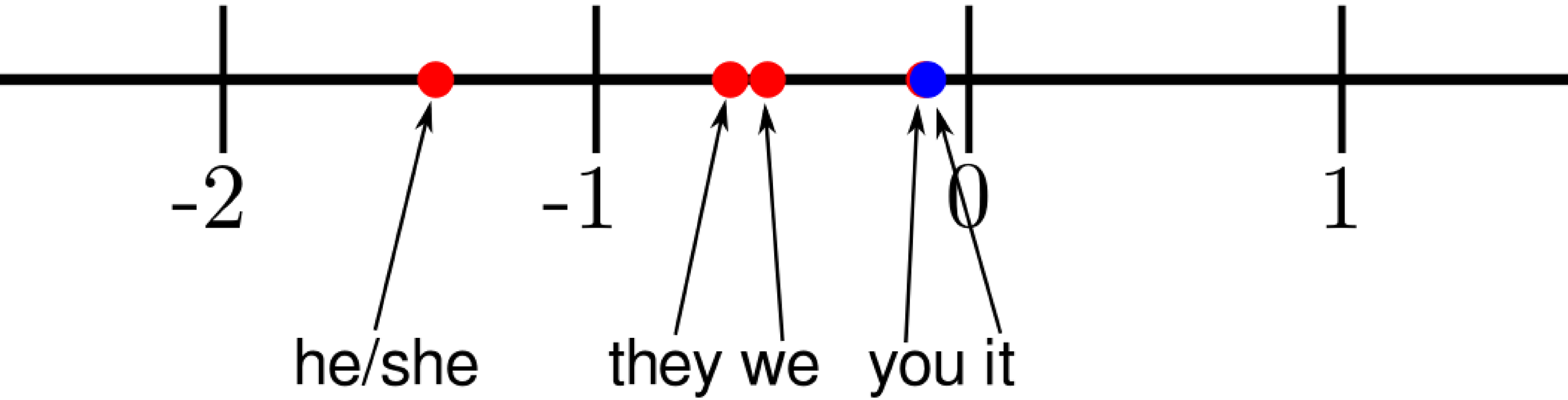
- Looked at log frequency ratio:

$$LF_{a,b}(w) = \log_{10} \frac{f_a(w)}{f_b(w)}$$



Rank–Counts Analysis: Inspection

$$LF_{a,b}(w) = \log_{10} \frac{f_a(w)}{f_b(w)}$$



Rank–Counts Analysis: Inspection

Anaphora:

- “Employee agrees that all information communicated to him/her concerning the work…”

Cataphora:

- “… it is the intention of the Recipient to give the Information Provider the broadest possible protection…”

Deictic:

- “…the terms ‘you and your’ are used in this Agreement, the same shall be construed as including…”

Nearest Neighbors

- So far we've discussed tokens broadly, seeing less repetition that typically expected.
- Does this hold up for the sentential level?

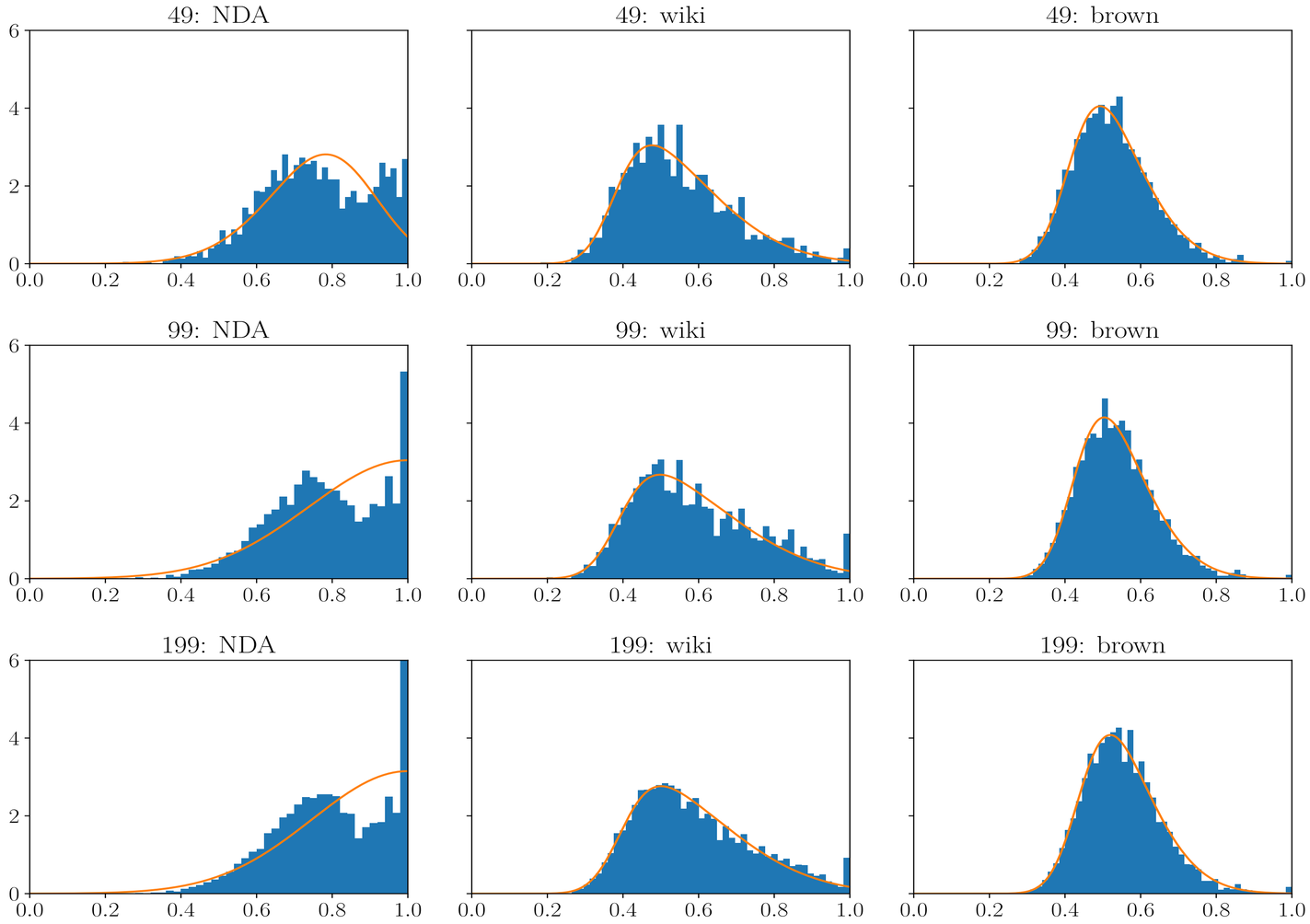
Nearest Neighbors: Model

- In a given corpus of documents, how similar is the next most similar sentence in the corpus?
- Old-fashioned unigram vector model.
- One vector per sentence, weighted by term frequency, normalized by sentence length.

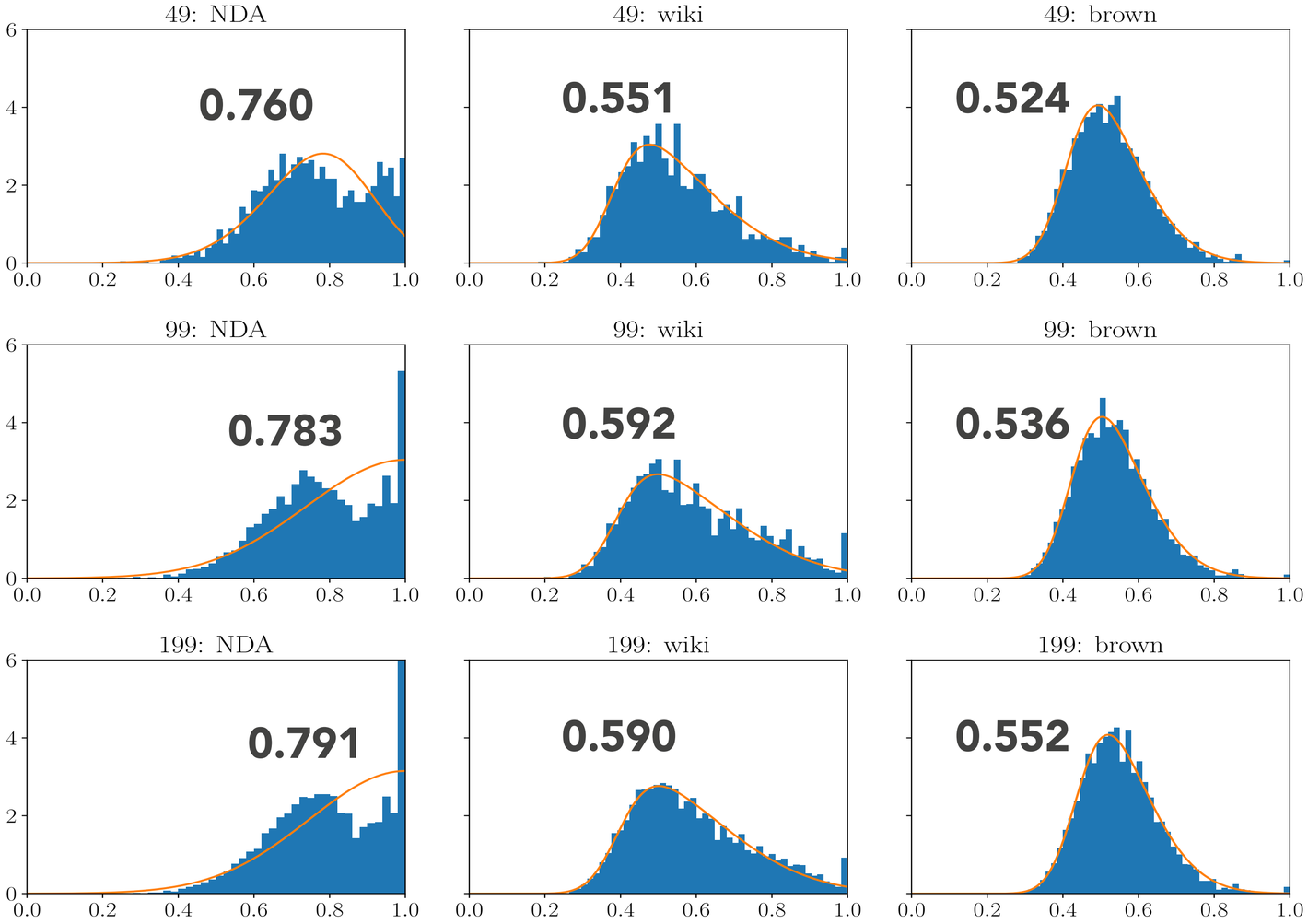
Nearest Neighbors: Model

- Sentences shorter than five tokens, removed.
- Take the dot product of every sentence in the corpus against every other.
- Save the highest similarity for each sentence.
- Discussing here 50, 100, 200 documents.

Nearest Neighbors: Results



Nearest Neighbors: Results



Nearest Neighbors: Model

- How do things look at the average similarity?

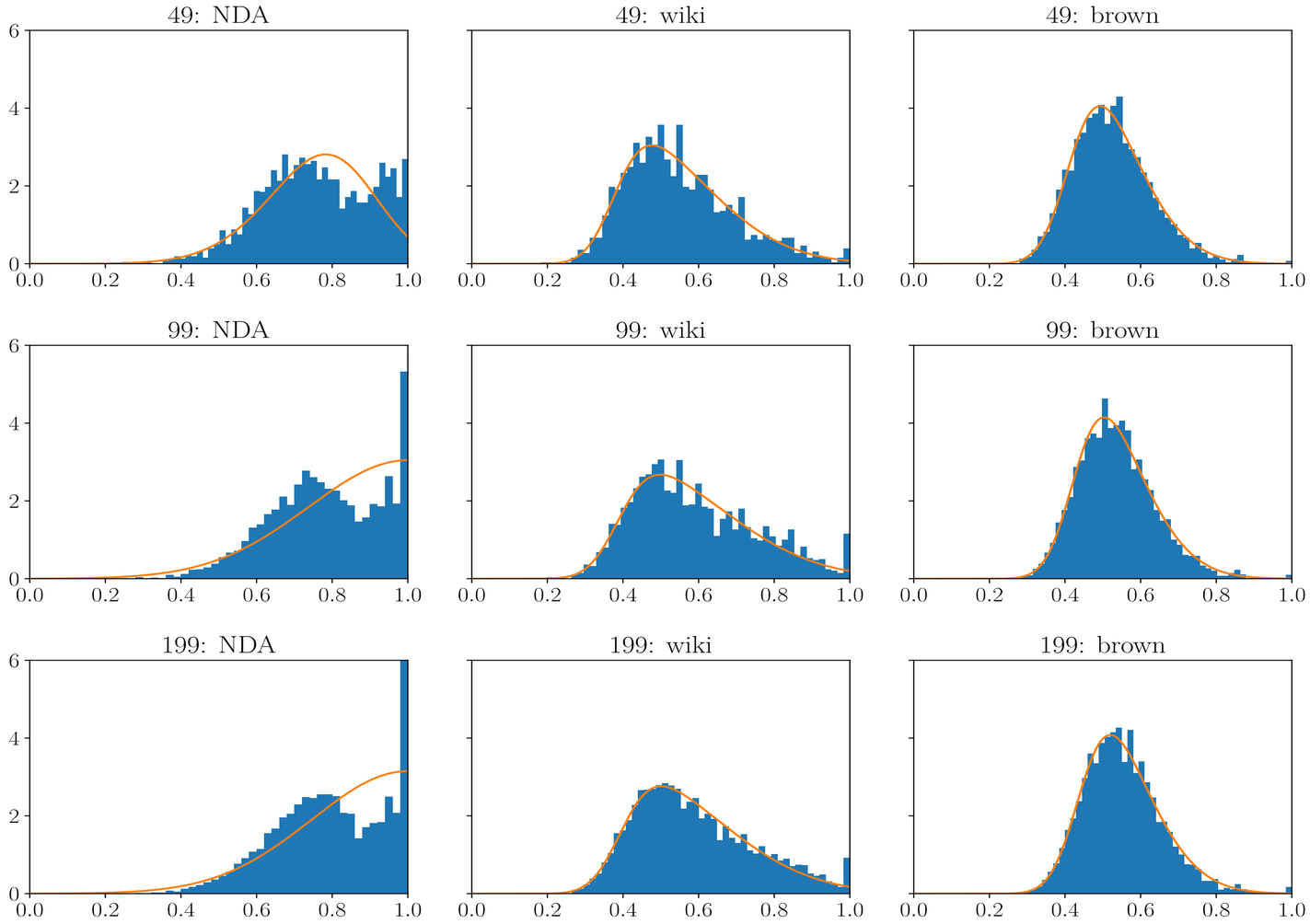
1a) Any assignment without such written consent shall be null and void and of no force or effect.

1b) Any such attempted assignment shall be void and of no effect.

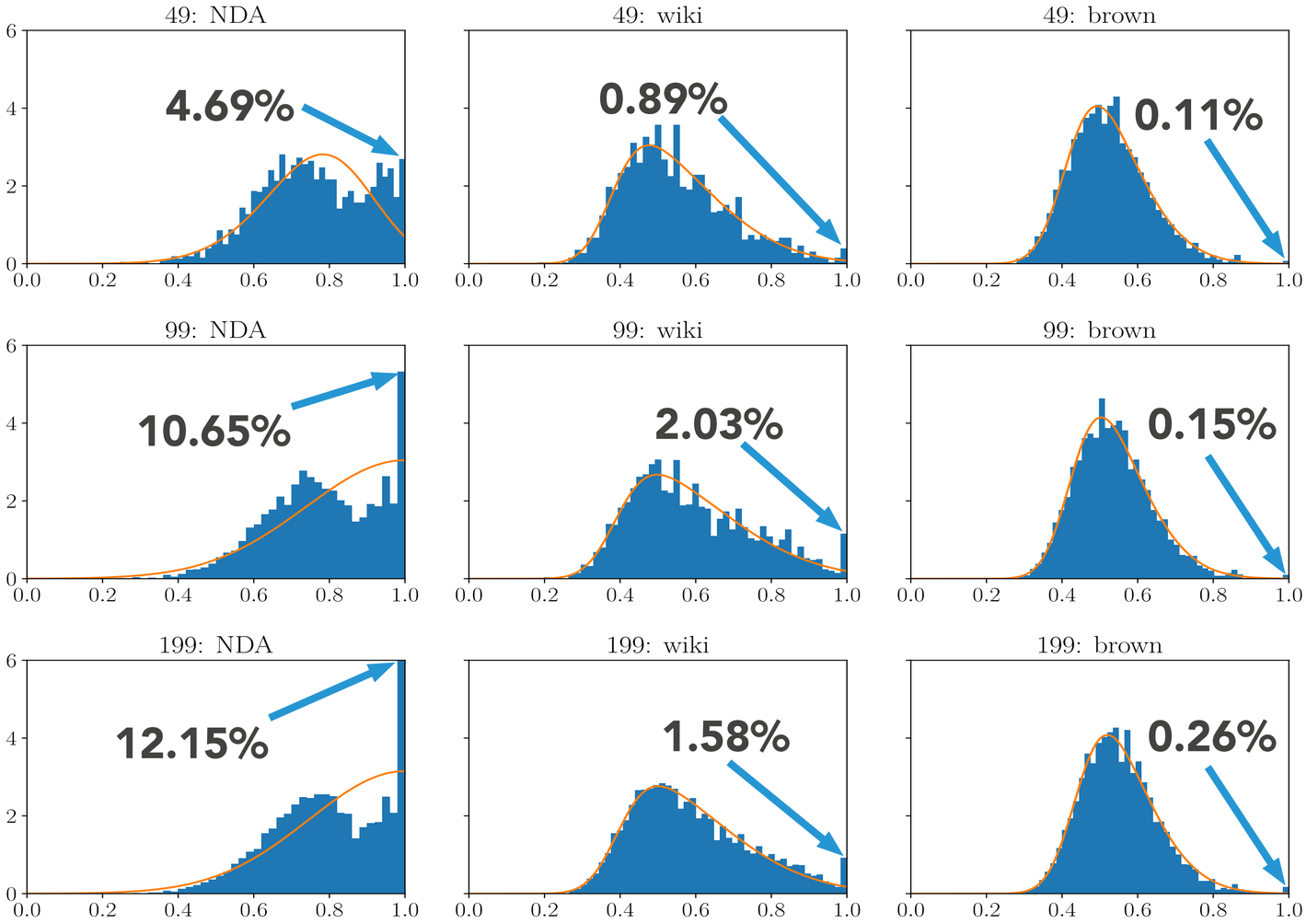
2a) Notwithstanding any other provision of this Agreement to the contrary, this Agreement shall be effective as of the date first above written and shall remain in full force and effect thereafter for a period of two (2) years, whereupon the Agreement shall automatically terminate, unless otherwise terminated by the mutual written agreement of the Parties.

2b) This Agreement shall be effective as of the Effective Date and continue for a period of five (5) years, or until termination of the Relationship, unless this Agreement is earlier terminated by mutual written agreement of the parties.

Nearest Neighbors: Results



Nearest Neighbors: Results



Nearest Neighbors: Results

- Exact Matches
  1. The more of these, the better for us!
  2. Exact matches in baseline corpora:
     - “Miami, Fla., March 17 –”
     - Wikipedia Infoboxes

Discussion

- We see differences that indicate repetition:
  1. Fewer hapax
  2. Fewer (and different) Pronouns
  3. Greater sentence similarity
- But it’s definitely language.
  1. “I pronounce you husband and wife” is too.
  2. Originality is not a prerequisite for language.
  3. We’ve had access to a lot of genres of text for a long time where originality is the reason for the communication.
  4. Contracts are fundamentally a different speech act from newswire, encyclopedias, etc.

Thank you!

dan.simonson@blackboiler.com

@thedansimonson

References

Robert Anderson and Jeffrey Manns. 2017. Engineering greater efficiency in mergers and acquisitions. The Business Lawyer, 72:657–678.

Patrizia Anesa. 2007. Vagueness and precision in contracts: a close relationship. Linguistica e filologia, 24:7–38.

Shlomo Argamon, Paul Chase, and Jeff Dodick. 2005. The languages of science: A corpus-based study of experimental and historical science articles. In Proceedings of the Annual Meeting of the Cognitive Science Society, volume 27.

J.L. Austin. 1962. How To Do Things With Words. Harvard University Press.

Adelchi Azzalini and Antonella Capitanio. 1999. Statistical applications of the multivariate skew normal distribution. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61(3):579–602.

Dorrit Faber and Karen Lauridsen. 1991. The compilation of a Danish-English-French corpus in contract law. English computer corpora. Selected papers and research guide, pages 235–43.

W. Nelson Francis and Henry Kučera. 1964, 1971, 1979. A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown). Brown University, Providence, Rhode Island, USA.

S. Bird, E. Loper, and E. Klein. 2009. Natural Language Processing with Python. O'Reilly Media Inc.

Michael AK Halliday. 1988. On the language of physical science. Registers of Written English: Situational Factors and Linguistic Features, pages 162–178.

Bjarne Blom and Anna Trosborg. 1992. An Analysis of Regulative Speech Acts in English Contracts - Qualitative and Quantitative Methods. Quantitative and quantitative methods. Hermes (Aarhus), 82:83.

Zellig S Harris. 2002. The structure of science information. Journal of biomedical informatics, 35(4):215–221.

Daniel P. Broderick, Jonathan Herr, and Daniel E. Simonson. 2016. Method and System for Suggesting Revisions to an Electronic Document. U.S. Patent and Trademark Office, US20170039176A1/US10216715B2.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Luciana Carvalho. 2008. Translating contracts and agreements: a Corpus Linguistics perspective. Avanĉos da linguística de Corpus no Brasil, page 333.

Matthew W. Crocker, Vera Demberg, and Elke Teich. 2016. Information density and linguistic encoding (ideal). KI - Künstliche Intelligenz, 30(1):77–81.

Michael Curtotti and Eric C. McCreath. 2011. A Corpus of Australian Contract Language: Description, Profiling and Analysis. In Proceedings of the 13th International Conference on Artificial Intelligence and Law, ICAIL '11, pages 199–208, New York, NY, USA. ACM.

Mark Davies. 2010. The Corpus of Contemporary American English as the first reliable monitor corpus of English. Literary and linguistic computing, 25(4):447–464.

Stefania Degaetano-Ortlieb and Elke Teich. 2017. Modeling intra-textual variation with entropy and surprisal: topical vs. stylistic patterns. SIGHUM, Pages 68–77, Vancouver, Canada. Association for Computational Linguistics.

Stefania Degaetano-Ortlieb and Elke Teich. 2018. Using relative entropy for detection and analysis of periods of diachronic linguistic change. SIGHUM Workshop, pages 22–33, Santa Fe, New Mexico. Association for Computational Linguistics.

Jason King. 2018. English Wikipedia Articles 2017- 08-20 SQLite. Kaggle.

Julia Kristeva. 1980. Desire in language: A semiotic approach to literature and art. Columbia University Press.

Christian Mair. 1997. The spread of the going-to-future in written English: A corpus-based investigation into language change in progress. Trends in Linguistics Studies and Monographs, 101:1537–1544.

Christopher D Manning and Hinrich Schütze. 1999. Foundations of Statistical Natural Language Processing. MIT Press.

Abdel Karim Mohammad, Nabil Alawi, and Maram Fakhouri. 2010. Translating contracts between english and arabic: Towards a more pragmatic outcome. Jordan Journal of Modern Languages and Literature.

Jane Norre Nielsen and Anne Wichmann. 1994. A frequency analysis of selected modal expressions in German and English legal texts. HERMES-Journal of Language and Communication in Business, 7(13):145–155.

Claude Elwood Shannon. 1948. A mathematical theory of communication. Bell system technical journal, 27(3):379–423.

George K. Zipf. 1949. Human Behavior and the Principle of Least Effort. Addison-Wesley.

Nearest Neighbors: Results

Corpus	Statistic	@ 50	@ 100	@ 200
NDA	Average	0.760	0.783	0.791
Wikipedia EN	Average	0.551	0.592	0.590
Brown	Average	0.524	0.536	0.552
NDA	Frac Max	4.69%	10.65%	12.15%
Wikipedia EN	Frac Max	0.89%	2.03%	1.58%
Brown	Frac Max	0.11%	0.15%	0.26%

Data: Contract Corpus

Category	# Docs	# Tokens	Toks/doc
NDAs	791	1,955,522	2,472
Prime Contract	174	5,417,987	31,138
Purchase Order	229	1,933,547	8,443
Services Agreement	137	1,216,724	8,881
Subcontract	406	5,029,433	12,388
Total	**1,737**	**15,553,213**	**8,954**

Rank–Counts Analysis

Series	| C |		# of Tokens	# Types	TTR	Hapax	H/Ty	H/Tok
Brown	500	500	1,161,192	56,057	4.83%	25,559	45.59%	2.20%
Wikipedia EN	4.9M	1,559	1,161,264	78,973	6.80%	40,820	51.69%	3.52%
NDA	791	484	1,164,051	17,454	1.50%	6,837	39.17%	0.59%
Purchase Order	229	132	1,164,421	21,670	1.86%	8,404	38.79%	0.72%
Prime Contracts	174	36	1,162,939	23,971	2.06%	9,461	39.47%	0.81%
Services Agreements	137	131	1,164,687	22,854	1.96%	8,915	39.01%	0.77%
Subcontracts	406	96	1,163,421	19,052	1.64%	6,670	35.01%	0.57%