NYT紙の量的テキスト分析を通じた150年間の地政学的脅威の測定

渡辺耕平

インスブルック大学（早稲田大学，LSE）

自己紹介

- 所属大学
  - デジタル科学センター／政治学部，University of Innsbruck (オーストリア)
- その他の所属
  - 招聘研究員，早稲田大学
  - 訪問研究員，US Centre, London School of Economics (イギリス)
- 学歴
  - 2007-2009，人文科学研究科，社会学専攻，武蔵大学
  - 2010-2011，修士課程，政治学部，Central European University (ハンガリー)
  - 2013-2017，博士課程，社会科学研究法，London School of Economics (イギリス)
- 研究分野
  - 政治コミュニケーション（ニュースのバイアス，メディアによる議題設定）
  - 国際コミュニケーション（プロパガンダ，グローバリゼーション）
  - 量的テキスト分析用ソフトウェアの開発（quanteda, newsmap, LSXなど）

量的テキスト分析とは？

- 政治学では文書が重要な研究資料
  - 質的テキスト分析
    - 例：スピーチ，選挙マニフェスト，政府報告書，新聞記事など.
  - 量的テキスト分析（2000年代以降）
    - 自然言語処理のツールを用いて多数の文書を分析する
    - 文書データを数値データに変換し，統計的な分析を行う.
- 政治学者は文書そのものに関心があるわけではない
  - 文書を通じて政治的行為者に関する推定を行う.
    - 例：イデオロギー，精神状態，偏見，行動など.
  - 文書を直接的には測定できないものを推計する
    - 例：政治的な影響力，経済的の不確実性，地政学的な脅威認識など.

発表の流れ

1. 量的テキスト分析の応用
   - “The Geopolitical Threat Index: A New Text-based Computational Approach to Identifying Foreign Threats”
     - “地政学的脅威指数：外的脅威を特定するためのコンピューターを用いた新しいアプローチ”
     - Peter Trubowitz (US Centre, LSE) との2017年ごろからの共同研究で、最近論文が完成した.
2. 量的テキスト分析の方法
   - Latent Semantic Scaling (LSS)
     - 準教師あり機械学習による脅威の測定
     - 種語を学習に用いるため柔軟で効率的

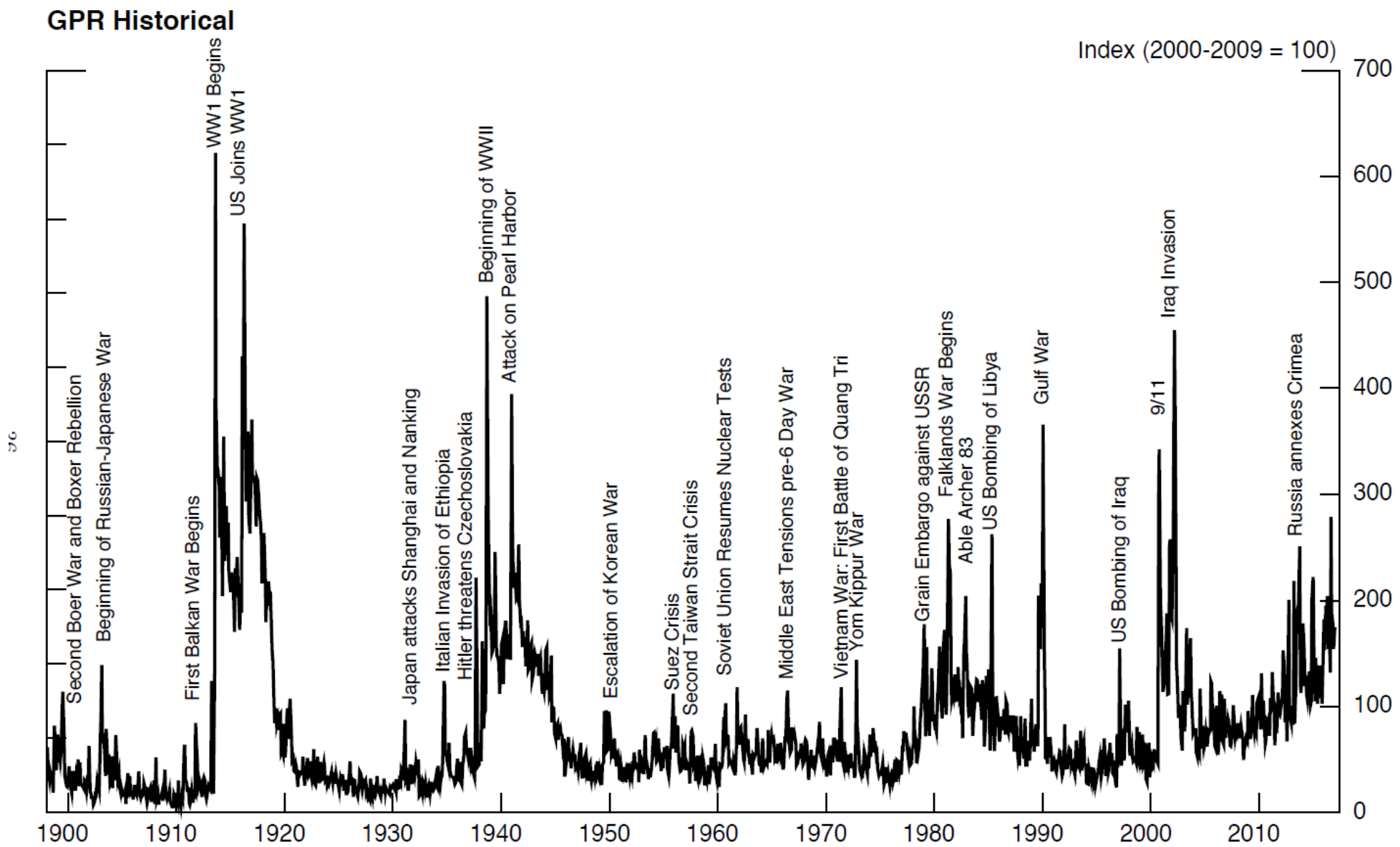
量的テキスト分析の応用

NYT紙を通じた地政学的脅威の歴史的な測定

研究の目的

- 地政学的脅威指数 (Geopolitical Threat Index) を作成する
  - 当指数を用いると，政治学者がアメリカの外交政策を統計的なモデルによって説明できるようになる.
  - 1861年から2017年にかけて発行されたNYT紙の記事を分析する.
  - 国際関係論の研究で使われている同様な指数は,
    - Composite Index of National Capability
      - 軍事支出，兵員数，エネルギー消費，鉄鋼生産量，都市人口，総人口などから構成
    - Militarized Interstate Dispute
      - 国家間の軍事的な衝突のイベントを記録
    - Geopolitical Risk Index (Caldara & Iacoviello 2018)
      - アメリカの主要な新聞に掲載された地政学的リスクに関する記事の頻度
      - データベースをキーワードで検索することで頻度を取得

Search Category	Search Terms
1. Geopolitical Threats	Geopolitical AND (risk* OR concern* OR tension* OR uncertaint*) “United States” AND tensions AND (military OR war OR geopolitical OR coup OR guerrilla OR warfare) AND (“Latin America” OR “Central America” OR “South America” OR Europe OR Africa OR “Middle East” OR “Far East” OR Asia)
2. Nuclear Threats	(“nuclear war” OR “atomic war” OR “nuclear conflict” OR “atomic conflict” OR “nuclear missile*”) AND (fear* OR threat* OR risk* OR peril* OR menace*)
3. War Threats	“war risk*” OR “risk* of war” OR “fear of war” OR “war fear*” OR “ military threat*” OR “war threat*” OR “threat of war” (“military action” OR “military operation” OR “military force”) AND (risk* OR threat*)


GPR Historical
Index (2000-2009 = 100)
700
600
500
400
300
200
100
0
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
Second Boer War and Boxer Rebellion
Beginning of Russian-Japanese War
First Balkan War Begins
WW1 Begins
US Joins WW1
Japan attacks Shanghai and Nanking
Italian Invasion of Ethiopia
Hitler threatens Czechoslovakia
Beginning of WWII
Attack on Pearl Harbor
Escalation of Korean War
Suez Crisis
Second Taiwan Strait Crisis
Soviet Union Resumes Nuclear Tests
Middle East Tensions pre-6 Day War
Vietnam War: First Battle of Quang Tri
Yom Kippur War
Grain Embargo against USSR
Falklands War Begins
Able Archer 83
US Bombing of Libya
Gulf War
US Bombing of Iraq
9/11
Iraq Invasion
Russia annexes Crimea

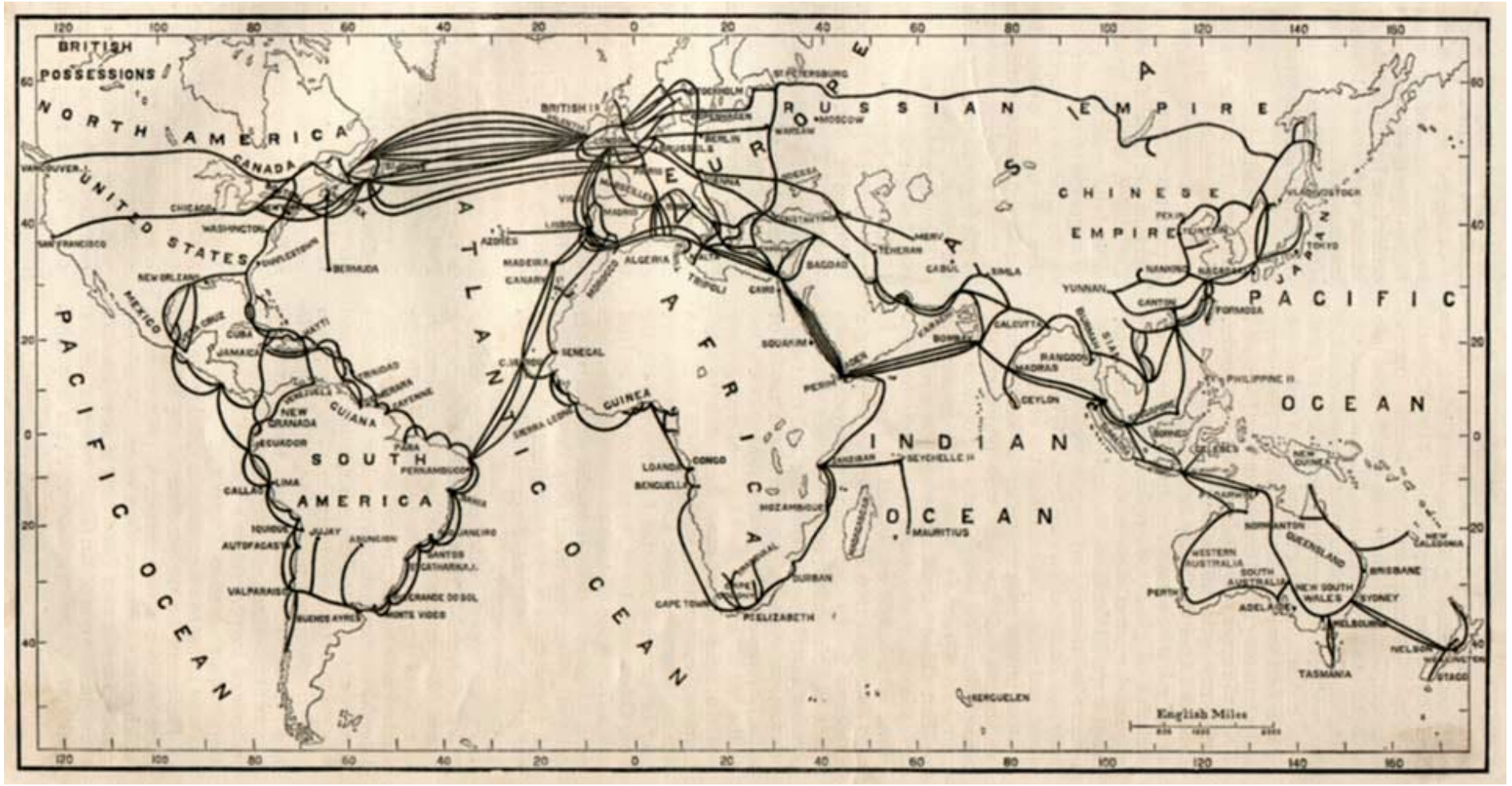
New York Times API

- NYT紙のアーカイブをAPIを通じて検索
  - NYT APIから，記事の最初の節を無料でダウンロードできる.
  - 軍事的脅威に関係したキーワードでおおまかにデータを収集
    - (military OR soldier* OR air force OR navy OR army) AND (threat* OR danger* OR fear* OR risk*)
  - 全文検索だが，記事の最初の数節しかダウンロードできない.
  - 1861 年から2017年の間に38万件の記事を収集.
- 合計で 387,896件の記事の一部を取得
  - “The two great German Powers have consented to pause in their career of conquest. After invading and taking possession of the Duchies, and just on the borders of Jutland,...” (1864年3月9日)

ニュース収集網の歴史

- 電信はイギリス人(William Cooke and Charles Wheatstone)とアメリカ人 (Finley Morse)のによって1837年に同時に発明された.
  - 電信網は1861年にアメリカの東から西海岸に到達.
  - 大西洋を渡る海底ケーブルが1866年に完成.
- フランスの電信機器が1869年に日本に輸入された.
  - 1871年に電信網がロンドン，ウラジオストク，アモイを経由して長崎に到達.

1895年の電信網



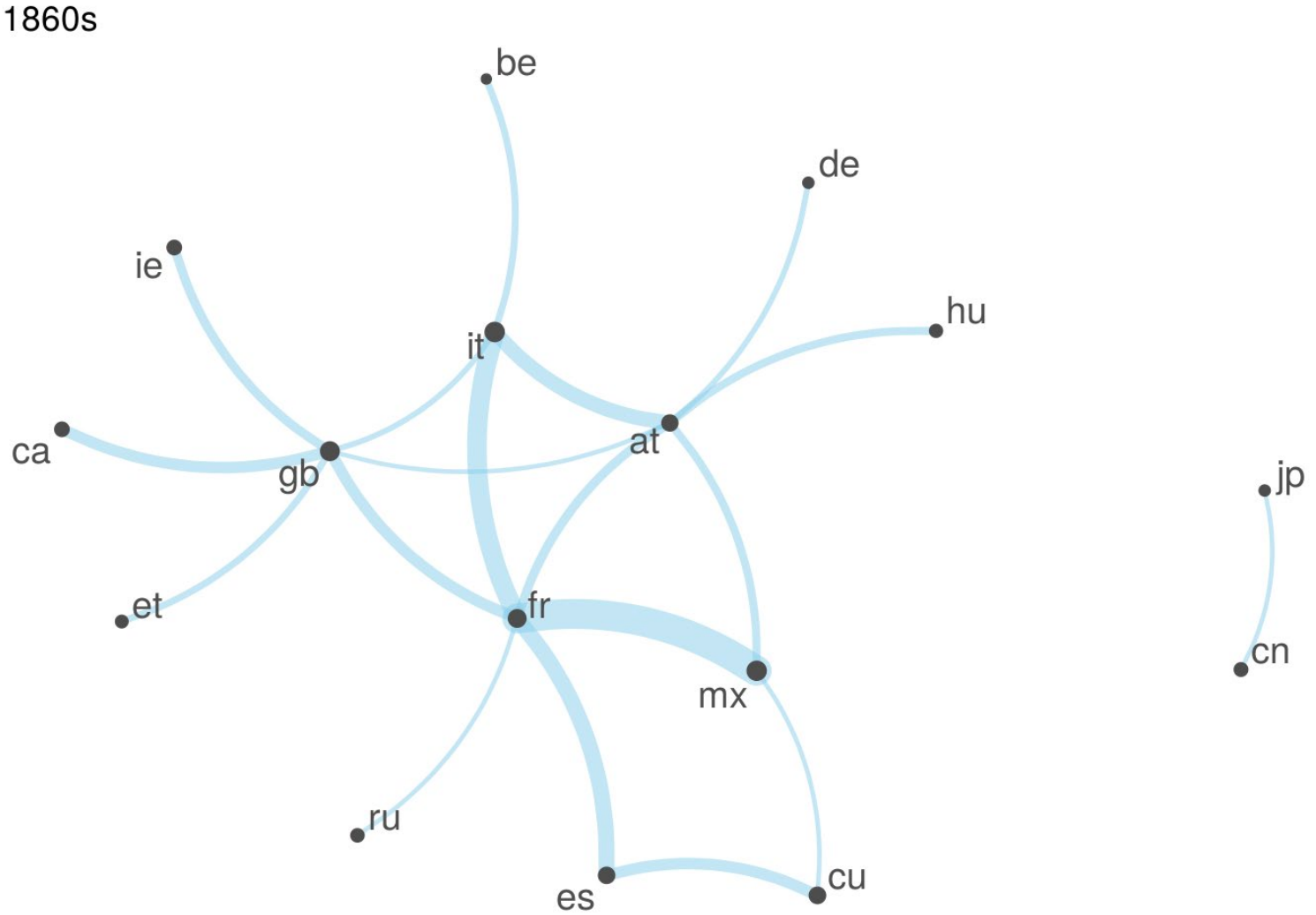
http://atlantic-cable.com/Article/1895MunroNerves/

国際的なニュースメディア

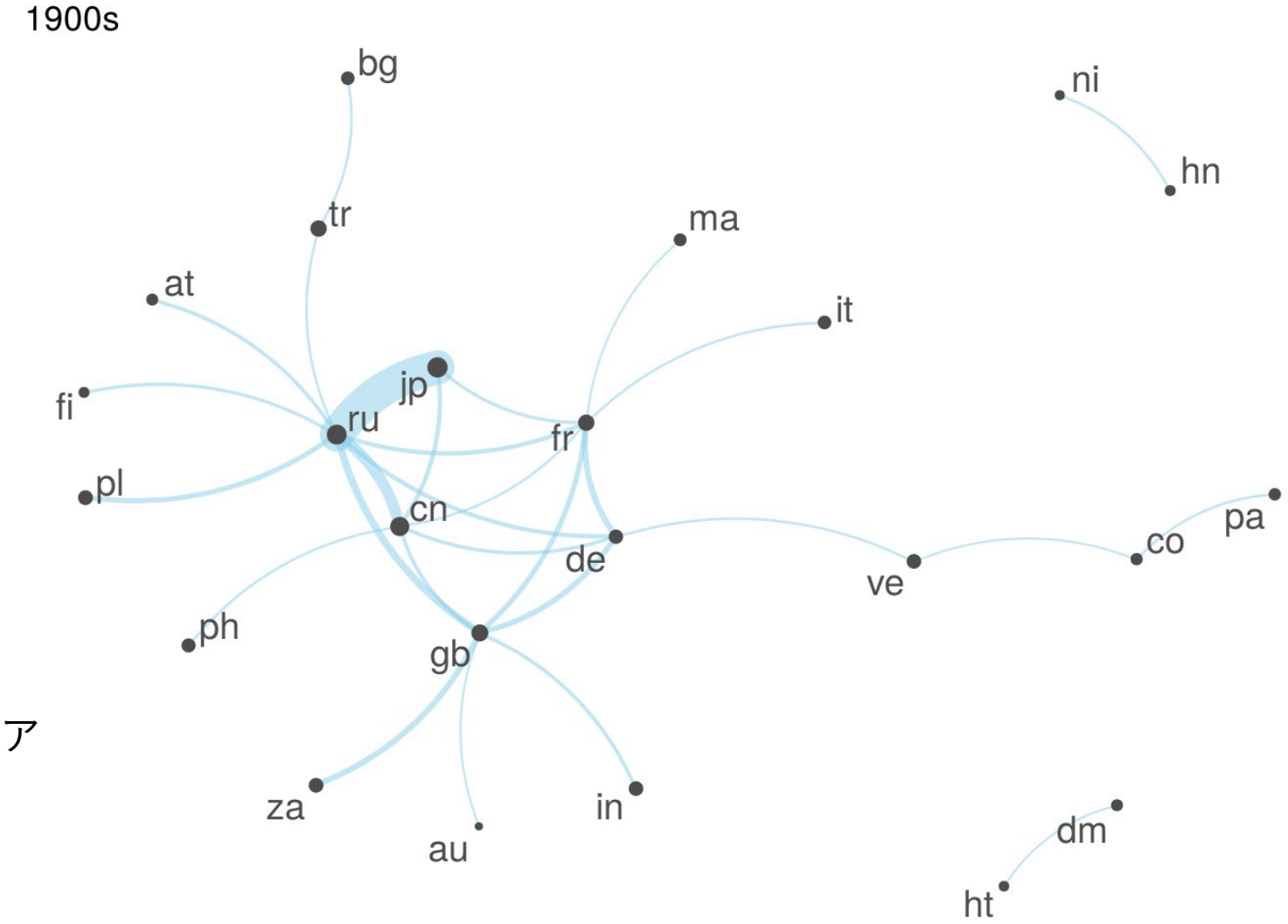
- 通信社
  - Agence France-Presse (1835)
  - Associated Press (1848)
  - Reuters (1851)
- 新聞
  - The Times of London (1785)
  - The New York Times (1851)

1860年代

at　オーストリア
cn　中国
cu　キューバ
es　スペイン
fr　フランス
gb　イギリス
it　イタリア
jp　日本
mx　メキシコ
ru　ロシア

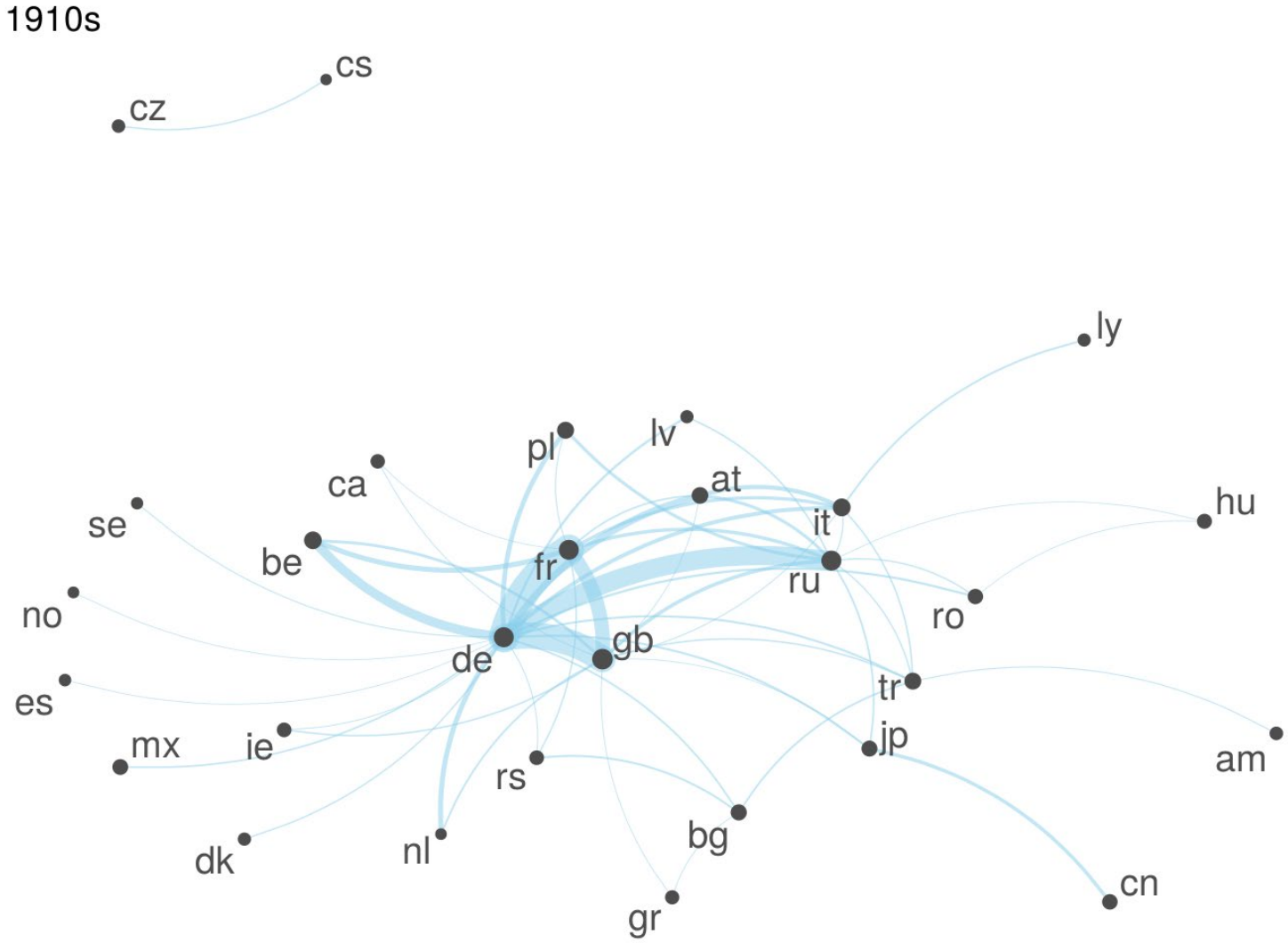


1900年代



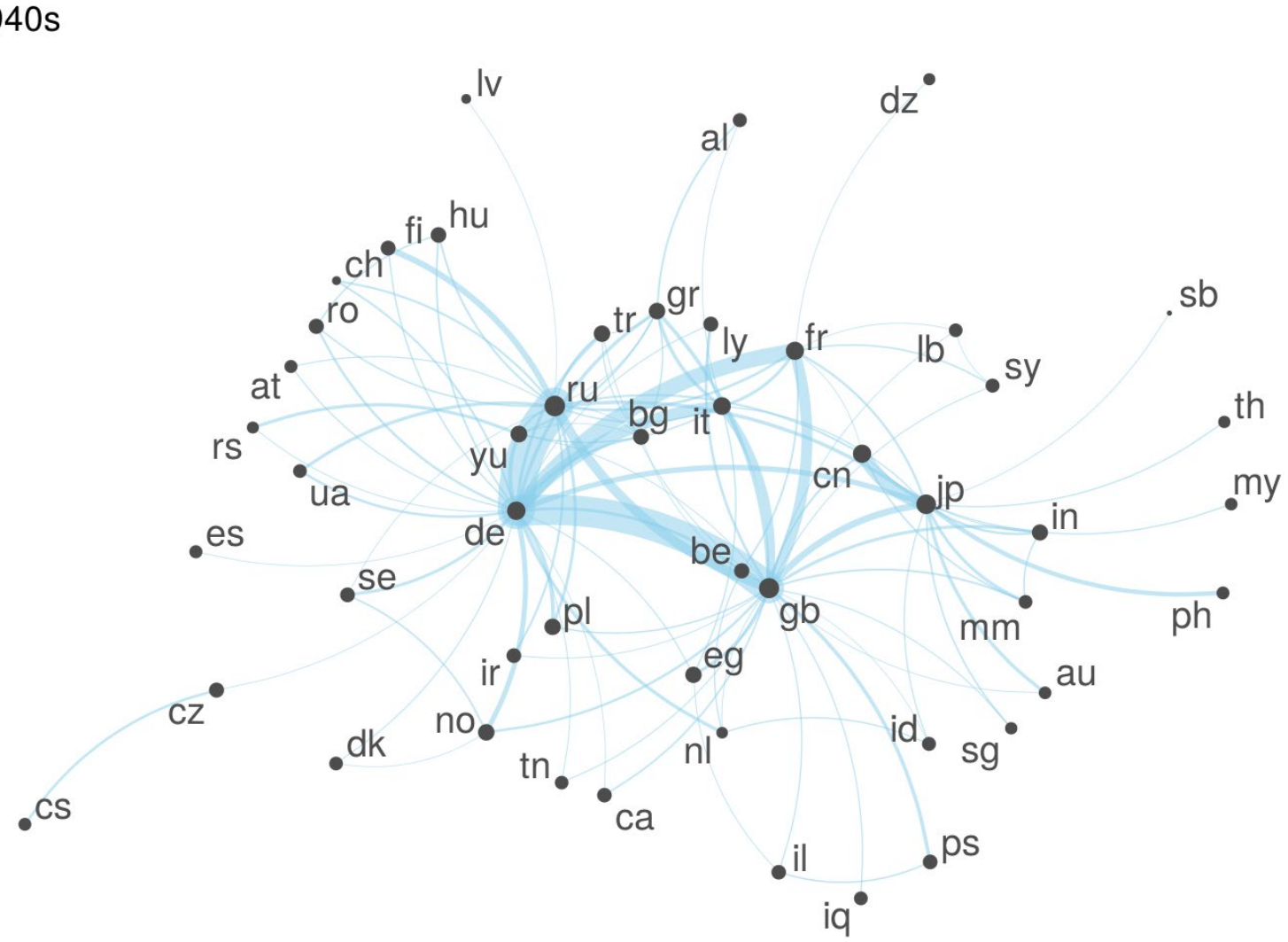
au　オーストラリア
de　ドイツ
in　インド
za　南アフリカ

1910年代



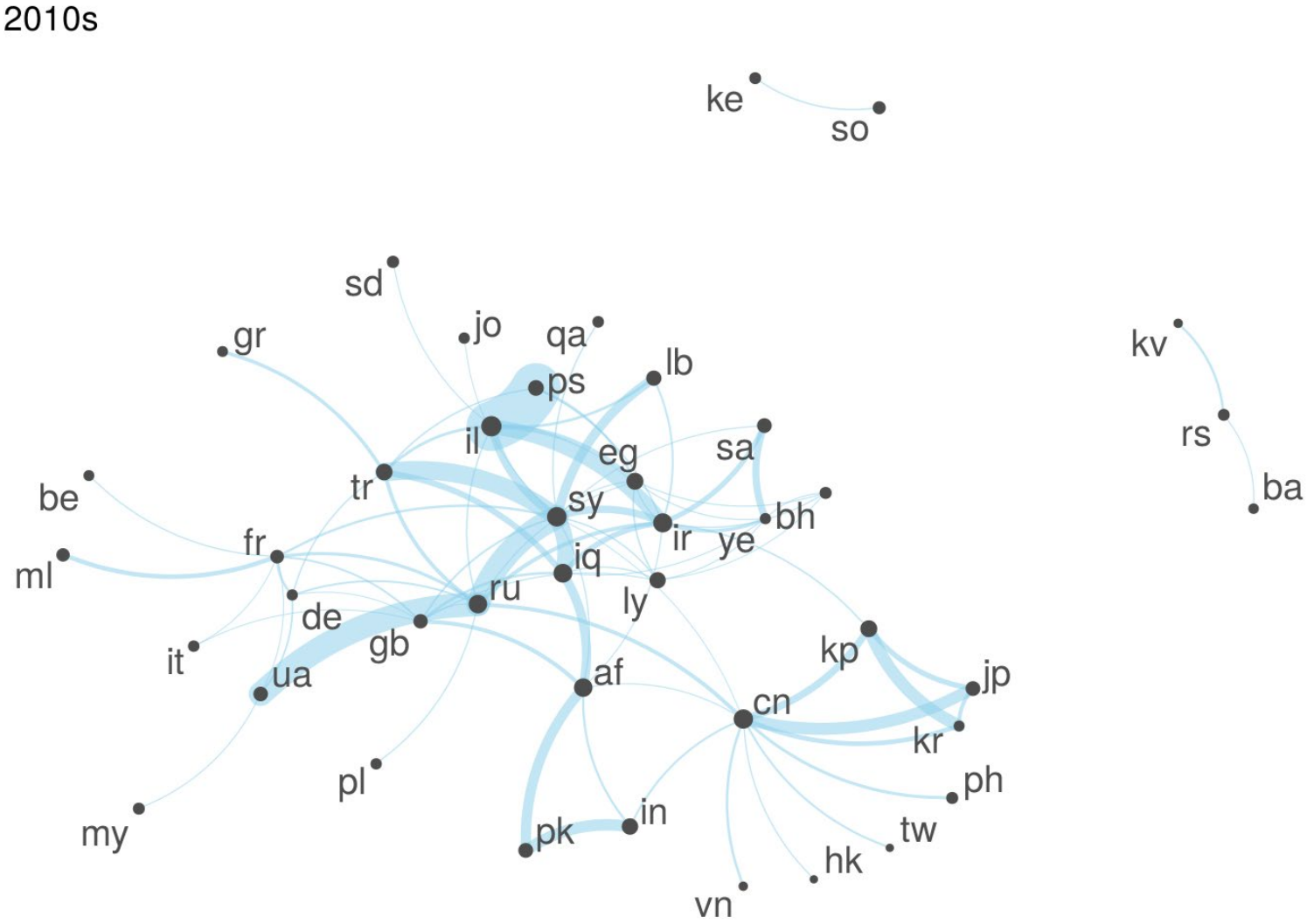
be　ベルギー
nl　オランダ
pl　ポーランド

1940年代



gr　ギリシャ
hu　ハンガリー
lb　レバノン
my　マレーシア
ph　フィリピン
sy　シリア
th　タイ
tr　トルコ

2010年代



il　イスラエル
iq　イラク
kp　北朝鮮
kr　韓国
ps　パレスチナ
sa　サウジアラビア
ua　ウクライナ

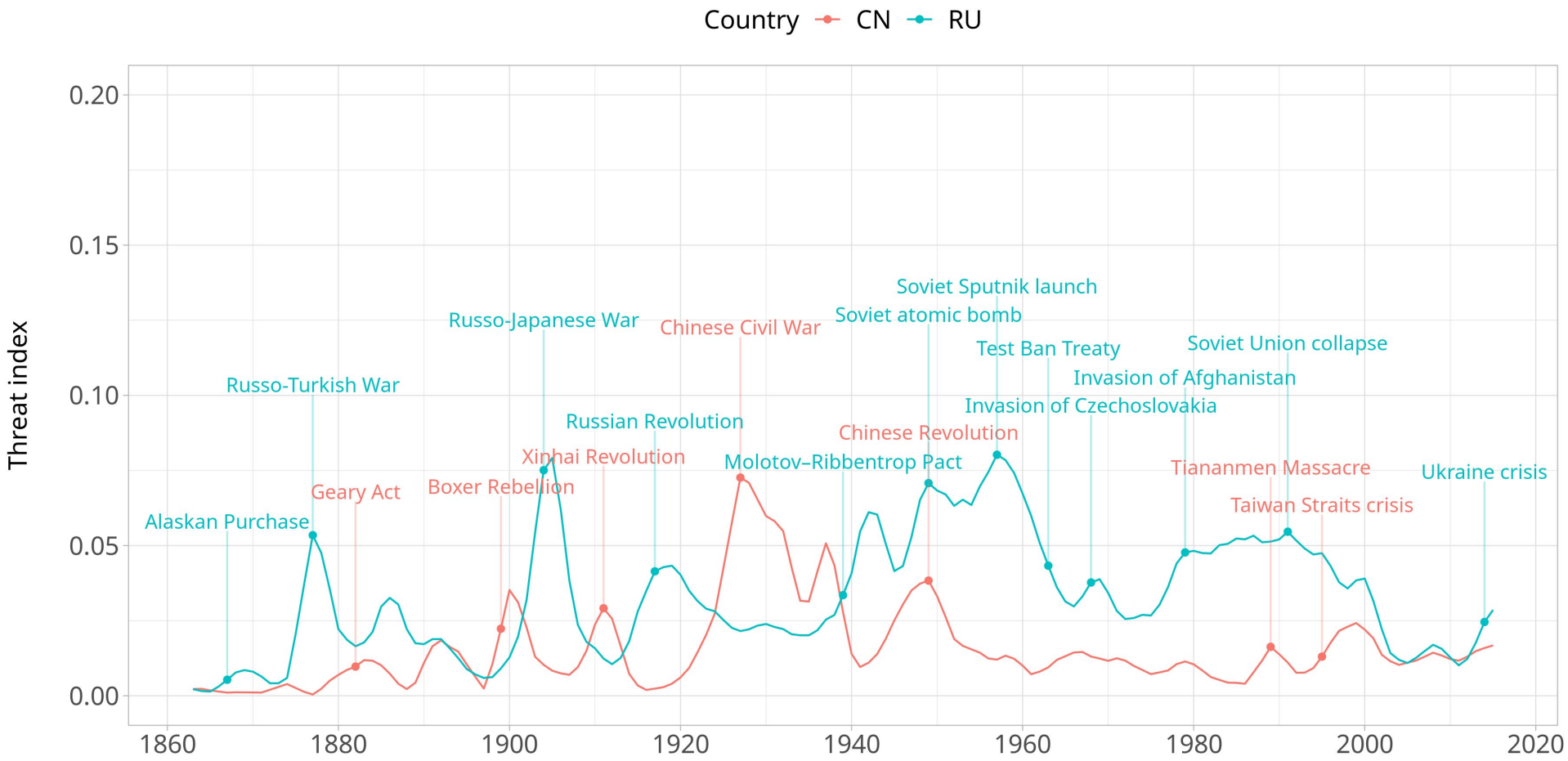
予備的分析の解釈

- NYT紙の記事を分析すると，国家間の緊張関係を測定できる可能性がある.
  - 記事を軍事的なキーワードで選択しているため，国名の共起は多くの場合、戦争などの敵対的な行動を表している.
  - しかし、NYT紙の報道アメリカの読者の興味関心に基づくため，同国のエリートの世界観を反映する.
- ネットワーク分析は視覚的でわかりやすいが，脅威指数の作成には不適切だろう.
  - 脅威指数のためには，主要な国の脅威の度合いを，一つの値に要約しなくてはいけない.

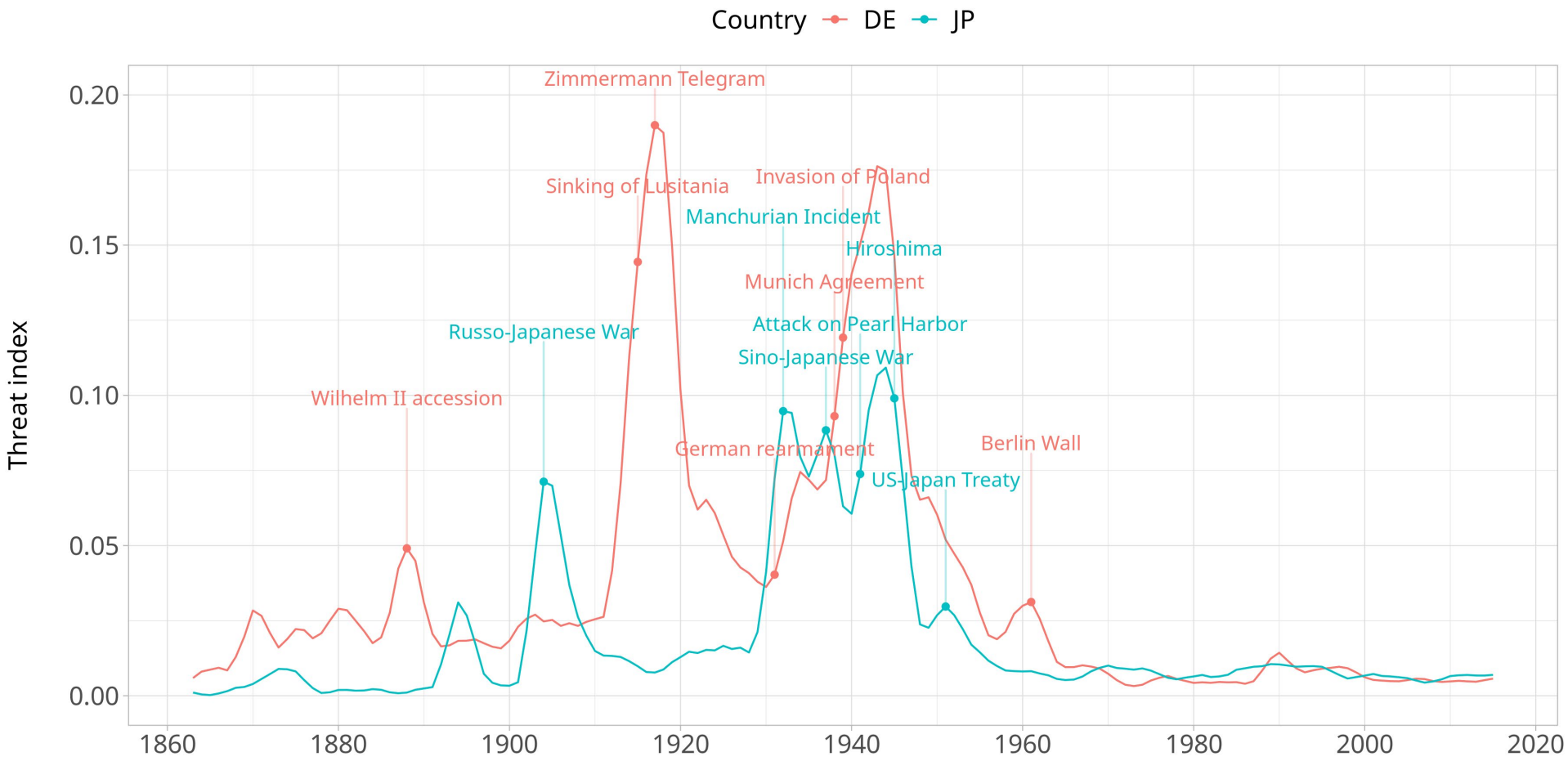
準教師あり機械学習による脅威の測定

- 地理的な分類
  - 記事が最も注目する国を特定する.
  - Newsmapを用いて辞書に含まれていない地理的語を特定
    - 150年間にはあまりにも多くの地名，人名，組織名が現れる.
- 敵対性による分類
  - 記事が報じる出来事の敵対性を判定する.
  - Latent Semantic Scaling (LSS) を用いて，敵対－友好の尺度を作成
    - 軍事に関する記事であっても，友好的な出来事は地政学的な脅威と関係ない.

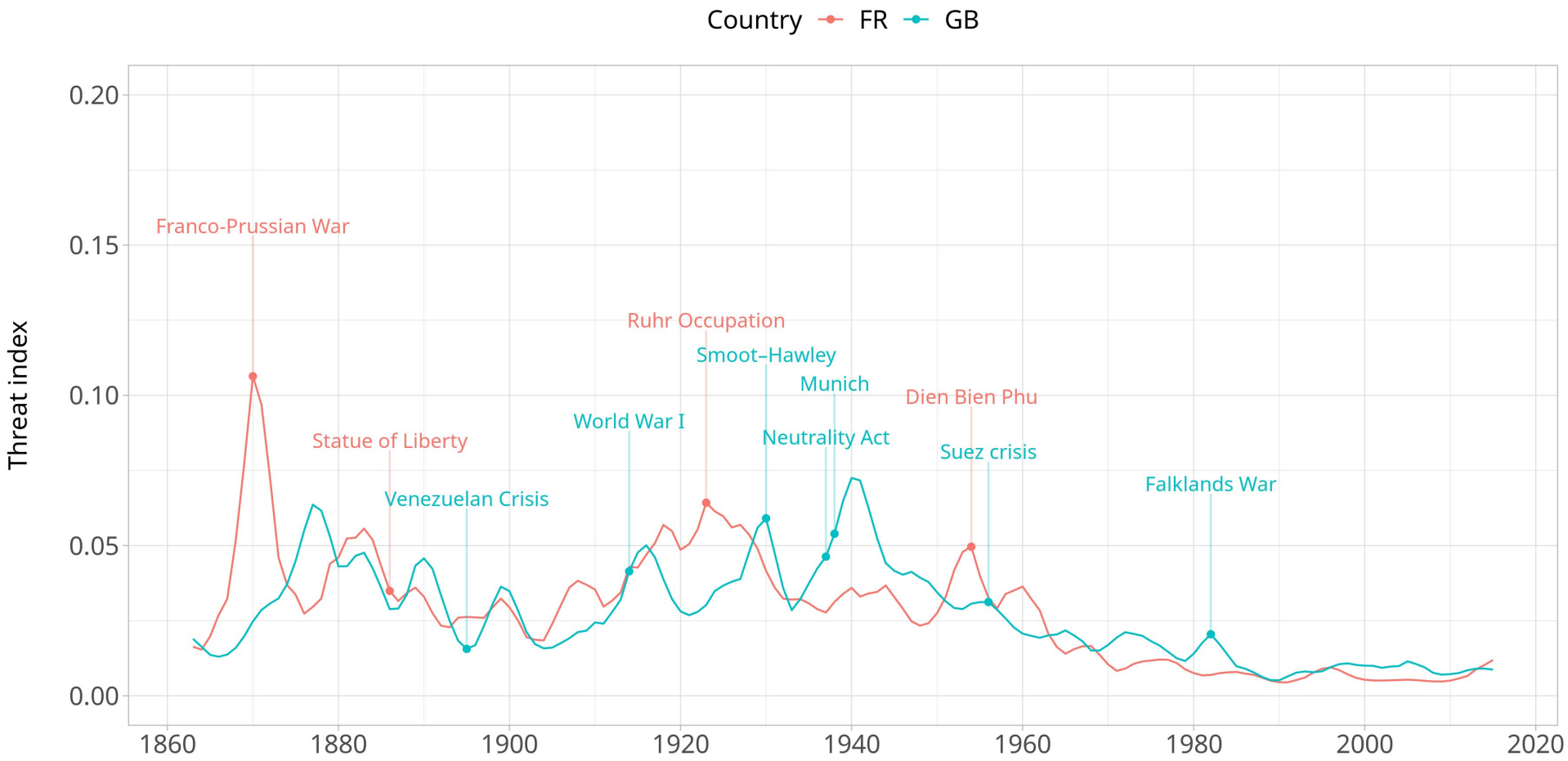
中国，ロシア


Country
CN
RU
Threat index
0.20
0.15
0.10
0.05
0.00
1860
1880
1900
1920
1940
1960
1980
2000
2020
Russo-Turkish War
Alaskan Purchase
Geary Act
Russo-Japanese War
Boxer Rebellion
Xinhai Revolution
Russian Revolution
Chinese Civil War
Molotov–Ribbentrop Pact
Soviet Sputnik launch
Soviet atomic bomb
Chinese Revolution
Test Ban Treaty
Invasion of Czechoslovakia
Invasion of Afghanistan
Soviet Union collapse
Tiananmen Massacre
Taiwan Straits crisis
Ukraine crisis

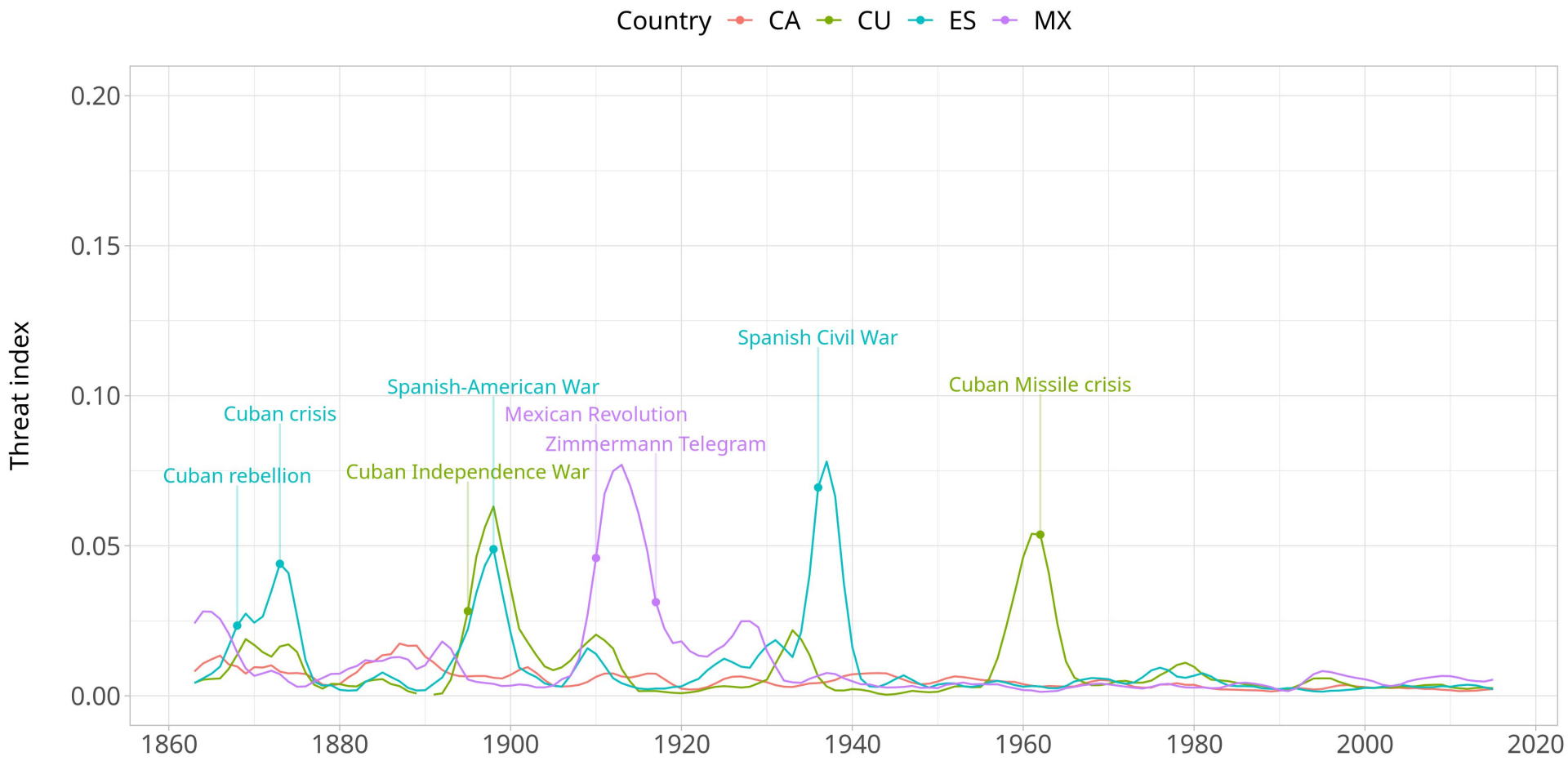
ドイツ，日本


Country DE JP
Threat index
0.20
0.15
0.10
0.05
0.00
1860
1880
1900
1920
1940
1960
1980
2000
2020
Zimmermann Telegram
Sinking of Lusitania
Invasion of Poland
Manchurian Incident
Hiroshima
Munich Agreement
Attack on Pearl Harbor
Russo-Japanese War
Sino-Japanese War
Wilhelm II accession
German rearmament
Berlin Wall
US-Japan Treaty

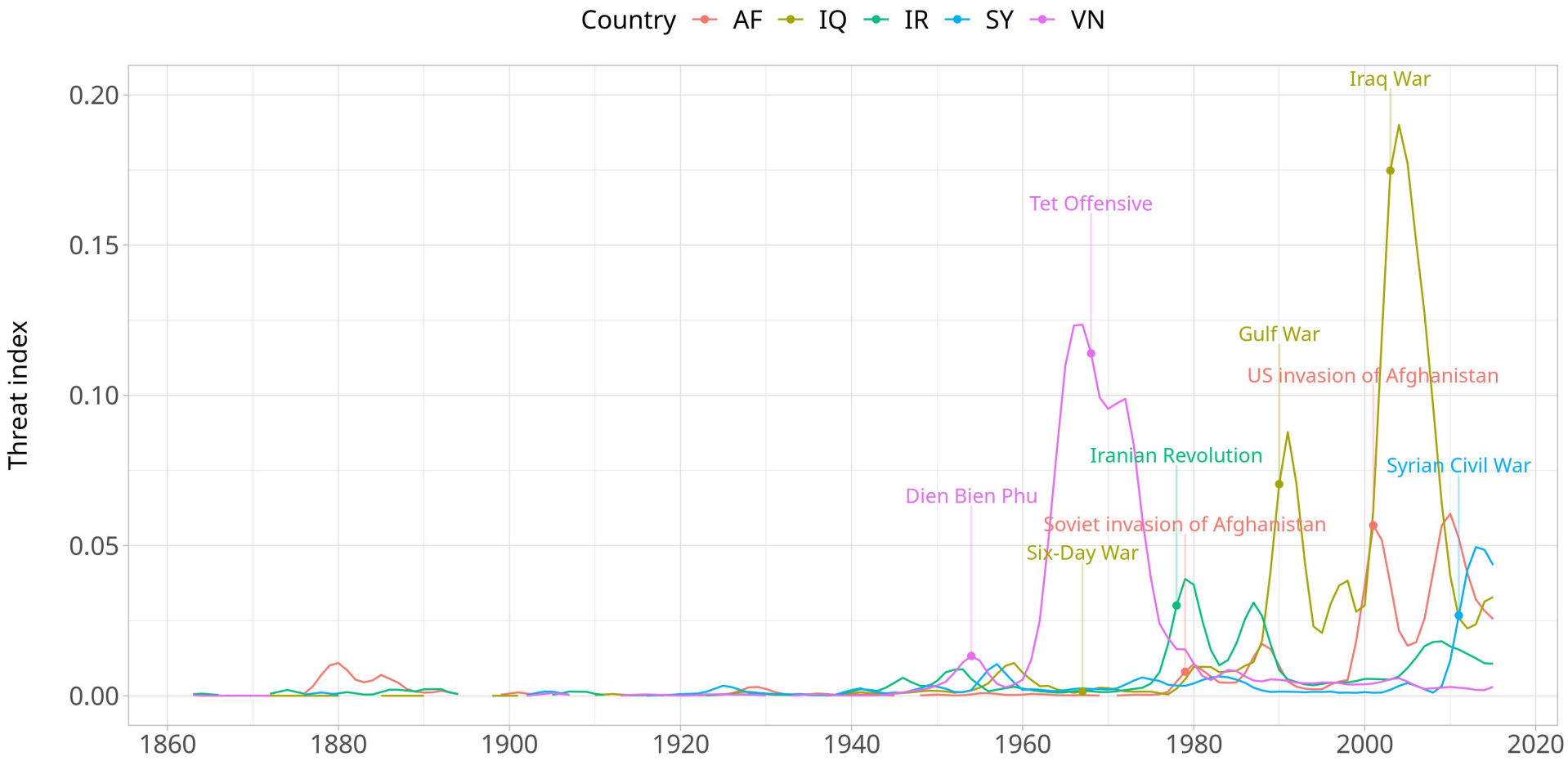
フランス，イギリス


Country FR GB
Threat index
0.20
0.15
0.10
0.05
0.00
1860
1880
1900
1920
1940
1960
1980
2000
2020
Franco-Prussian War
Statue of Liberty
Venezuelan Crisis
World War I
Ruhr Occupation
Smoot–Hawley
Munich
Neutrality Act
Dien Bien Phu
Suez crisis
Falklands War

カナダ，キューバ，スペイン，メキシコ



アフガニスタン，イラク，イランなど



すべての地政学的脅威


Threat index
0.5
0.4
0.3
0.2
0.1
0.0
1860
1880
1900
1920
1940
1960
1980
2000
2020
Pearl Harbor
Zimmermann Telegram
Sinking of Lusitania
Munich
Invasion of Manchuria
Chinese Civil War
Berlin Wall
Russo-Japanese War
Soviet bomb
Tet offensive
Santo Domingo Treaty
Mexican Revolution
Cuban Missile crisis
September 11
Spanish-American War
Gulf War
Ukraine Crisis
Invasion of Afghanistan
Sputnik

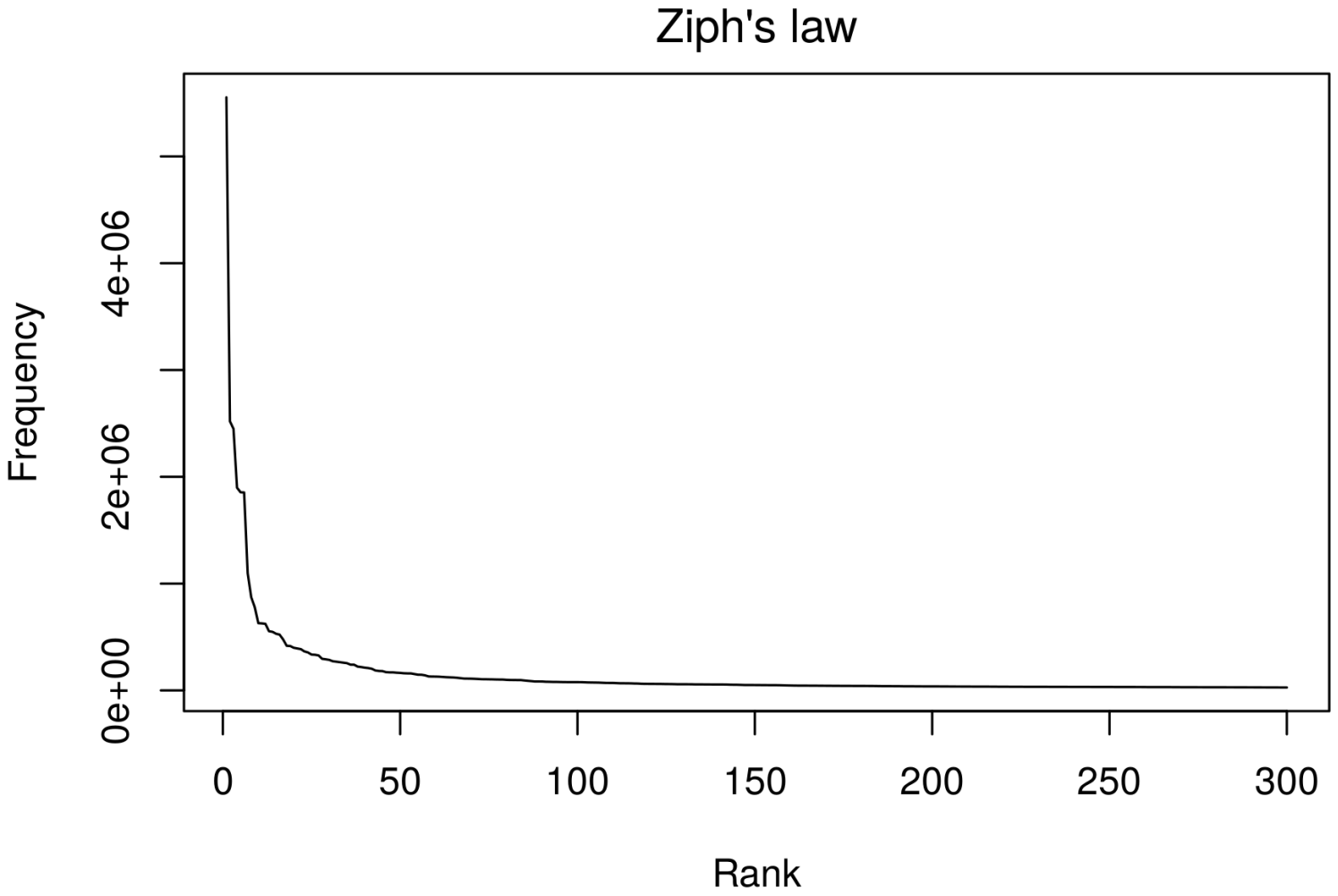
量的テキスト分析の方法

準教師あり学習モデルによる文書の分類

文書データの特徴

- 多次元性
  - ひとつひとつの語が変数となり，統計分析や機械学習では数万個の変数が分析の対象となる.
  - 記事の内容が多様であると，さらに多次元性が高くないり，分析が難しくなる.
- データ疎性
  - 文書には意味が乏しい文法的な語が多いが，政治学的に興味深い語は少ない.
  - データの疎性は，コーパスの内容が多様で、文書が短いほど高くなる.
  - データ疎性は統計的分析を難しくする.

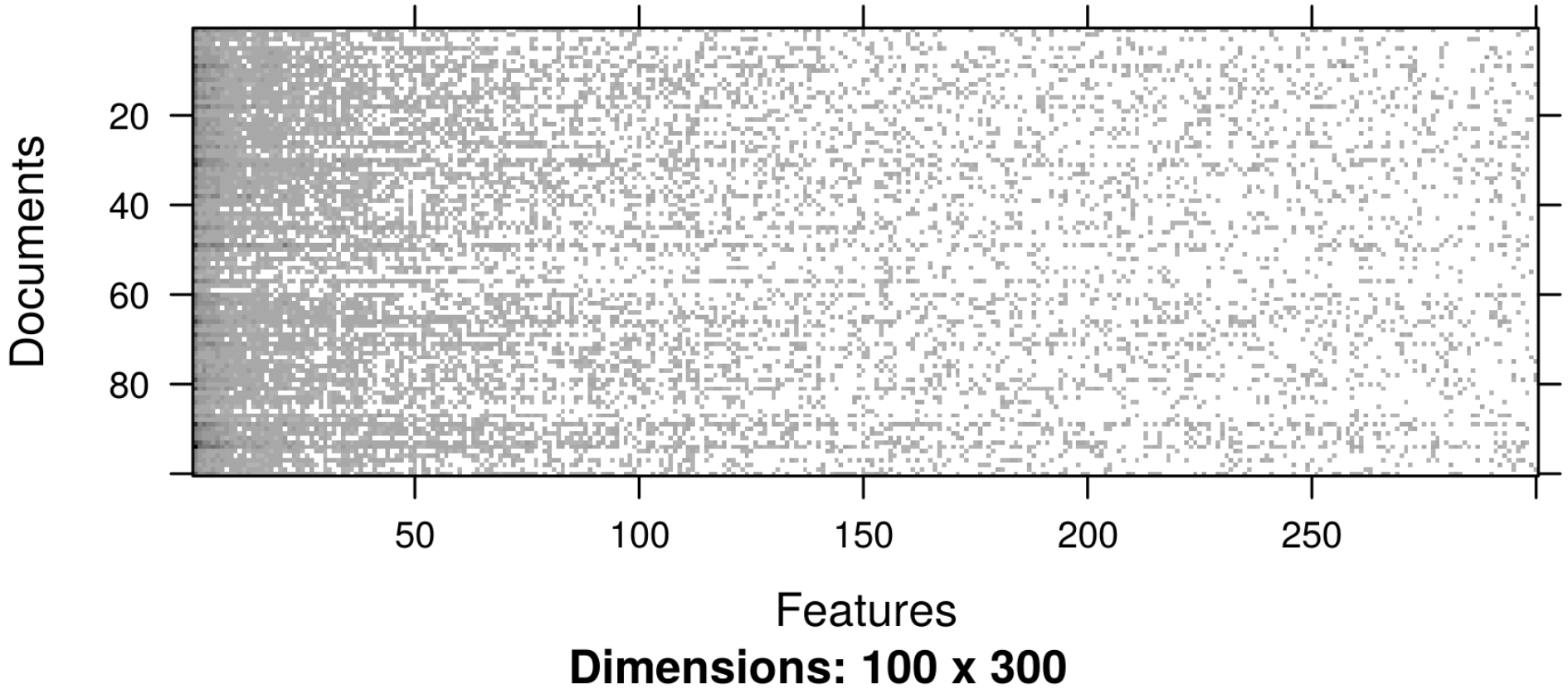
語頻度の分布


Ziph's law
Frequency
0e+00
2e+06
4e+06
0
50
100
150
200
250
300
Rank

最も頻度が高い100語

```
the, of, to, a, in, and, that, is, for, on, by, was,
at, as, with, has, from, today, it, an, his, be,
this, have, are, new, united, he, president, been,
states, which, its, war, not, who, will, last, but,
their, were, had, american, one, said, military,
they, government, more, yesterday, after, two, here,
when, about, there, or, would, over, all, into,
than, first, years, army, i, state, against, week,
most, no, out, york, world, now, some, lead, soviet,
what, up, city, may, made, time, between, if,
before, other, people, foreign, we, mr, national,
general, so, house, many, washington, her, officials
```

データ疎性


Documents
20
40
60
80
50
100
150
200
250
Features
Dimensions: 100 x 300

一般的な量的テキスト分析の流れ

- データ収集
  - NYT APIからRでダウンロード
- 文書の前処理
  - クリーニング（記者名，日付などの削除）
  - トークン化（記事を単語に分割）
  - 数字，記号，文法的語の削除
- 統計的分析
  - ネットワーク分析，辞書分析，相対頻度分析，機械学習など
- 結果の解釈

機械学習の種類

- 教師あり学習
  - Support Vector Machine，ナイーブベイズ，Random Forestsなど
    - 訓練データを通じてユーザーが分析結果を制御できる.
    - 複雑なモデルを訓練するためのコストが高い.
- 教師なし学習
  - Latent Dirichlet Allocation，対応分析，Multi-dimensional Scalingなど
    - ユーザーが分析結果を制御できない.
    - 訓練をするための費用が全くかからない.
- 準教師あり
  - Seeded LDA，Newsmap，Latent Semantic Scaling
    - 種語を通じてユーザーが分析結果を制御できる.
    - 訓練するためのコストが小さい.

Latent Semantic Scaling (LSS)

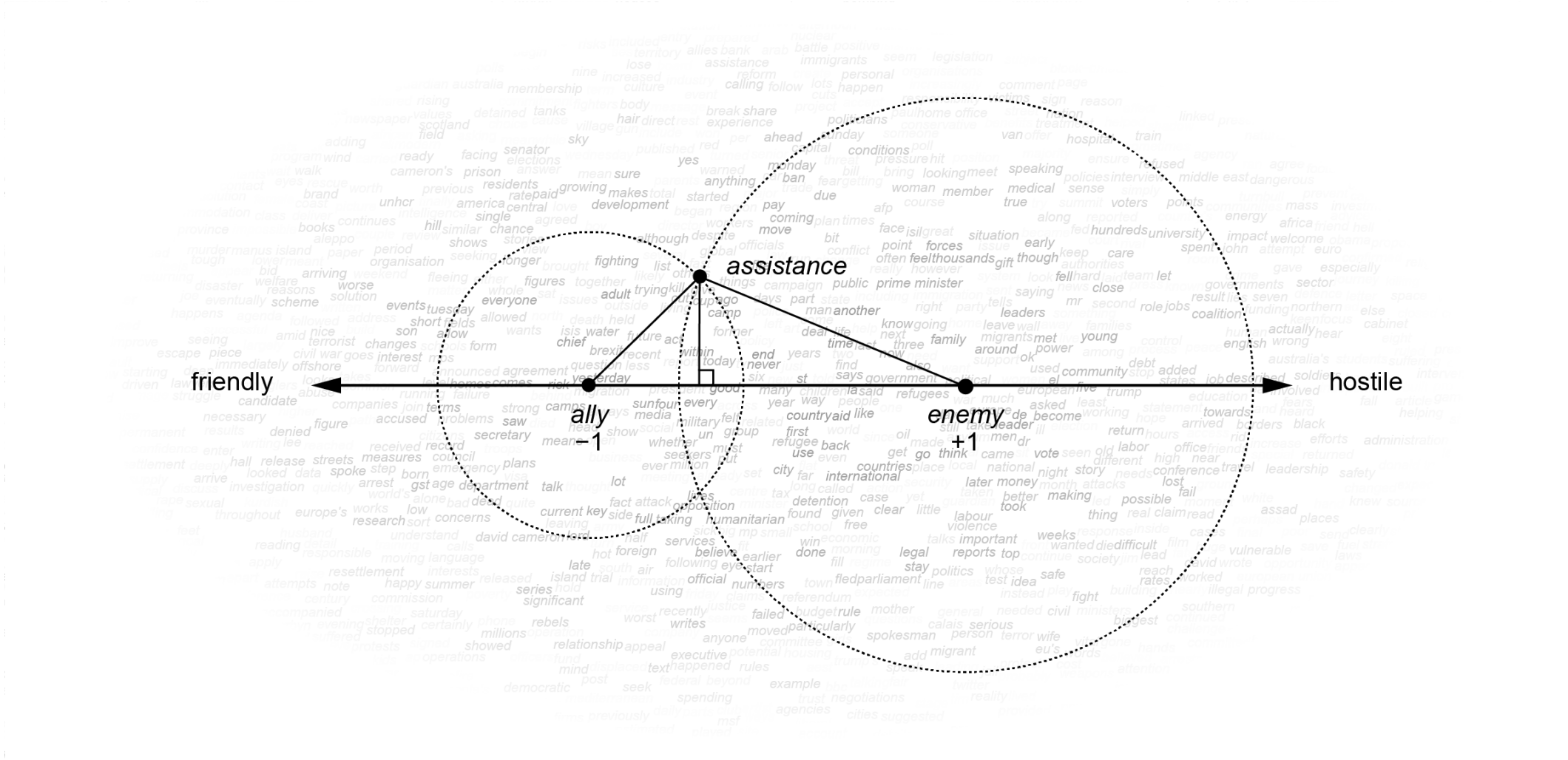
- Word-embeddingの手法を準教師あり学習モデルとして文書の計測のために応用した
  - 少ない数の種語から測定する尺度を学習する.
  - およそ70%程度の精度で分類を行える.
- LSEでの博士課程の間に開発した
  - 2014年のウクライナ危機の最中のロシアの国営通信社の国際プロパガンダの分析を行った (Watanabe 2017).
  - 特定された分野の分析や多言語での分析で非常に有用性が高い (Watanabe 2020).

敵対性を測るための種語

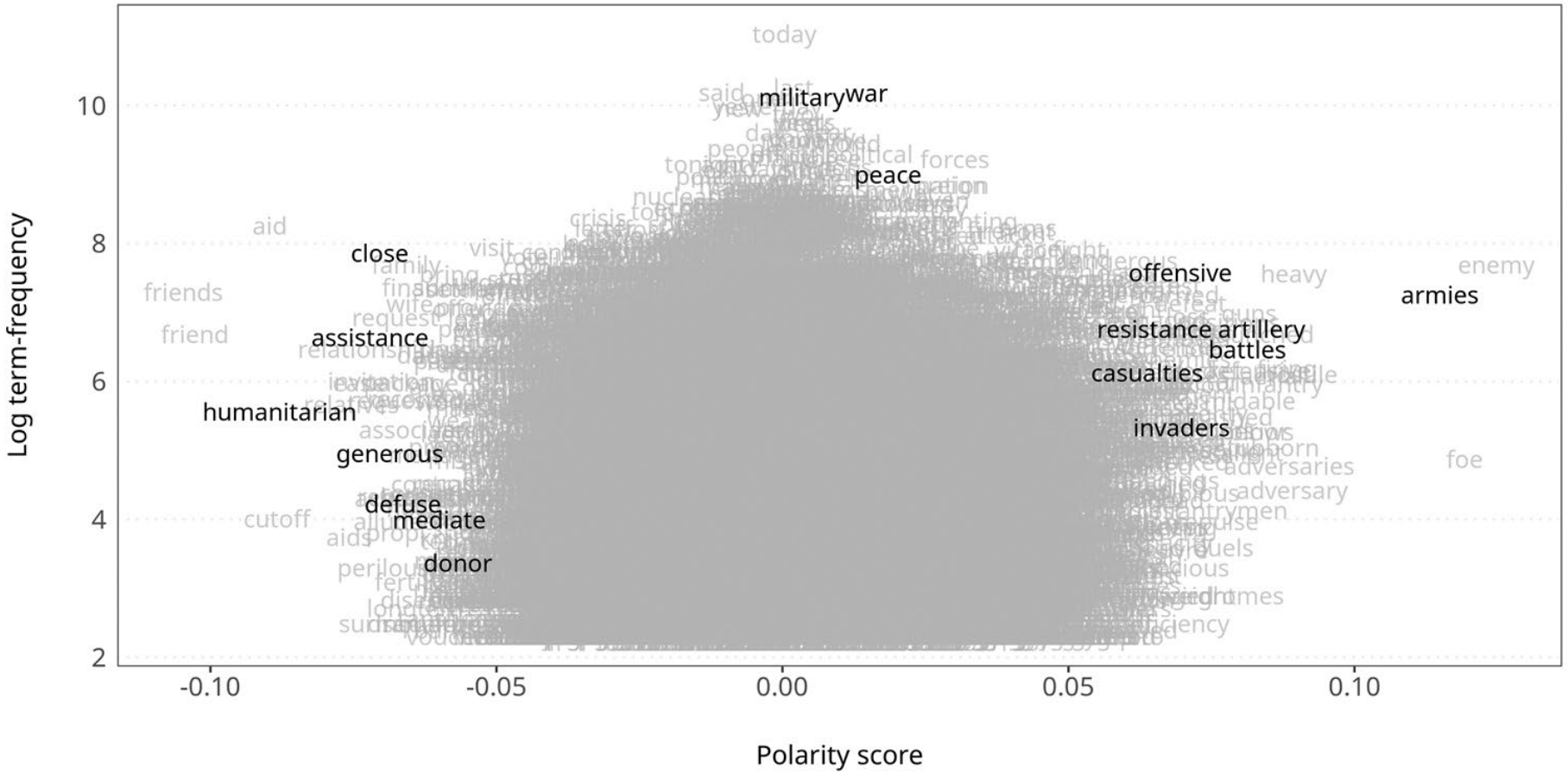
- 種語は測定しようとする概念を定義する
  - 種語は「弱い教師」としてモデルを訓練する.
  - 内容分析用のキーワード辞書分析に似ている.
- ユーザーは種語を通じて機械に図るべき尺度を教える

概念	種語
敵対的 (hostile)	adversary, adversaries, enemy, enemies, foe, foes
友好的 (friendly)	aid, ally, friend, peaceful

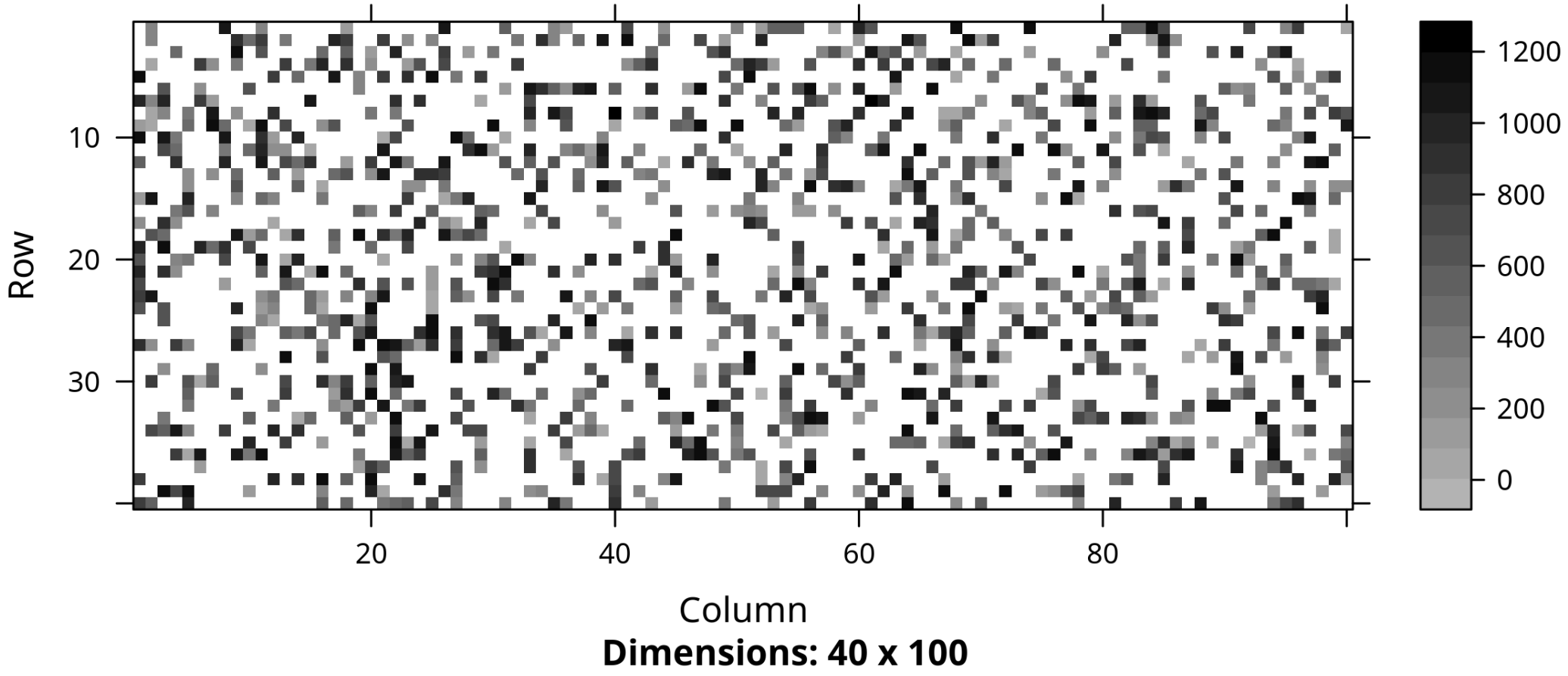
種語による重みづけの概念



敵対性で重みづけられた語

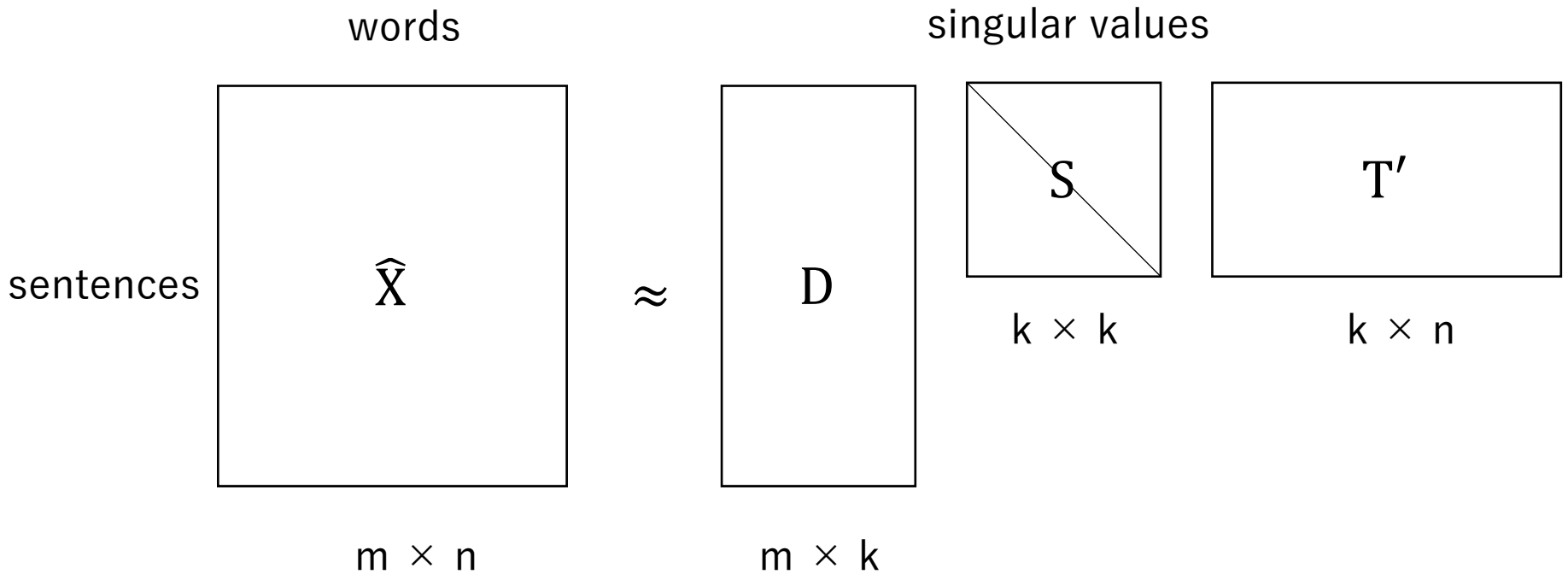


文書特徴行列

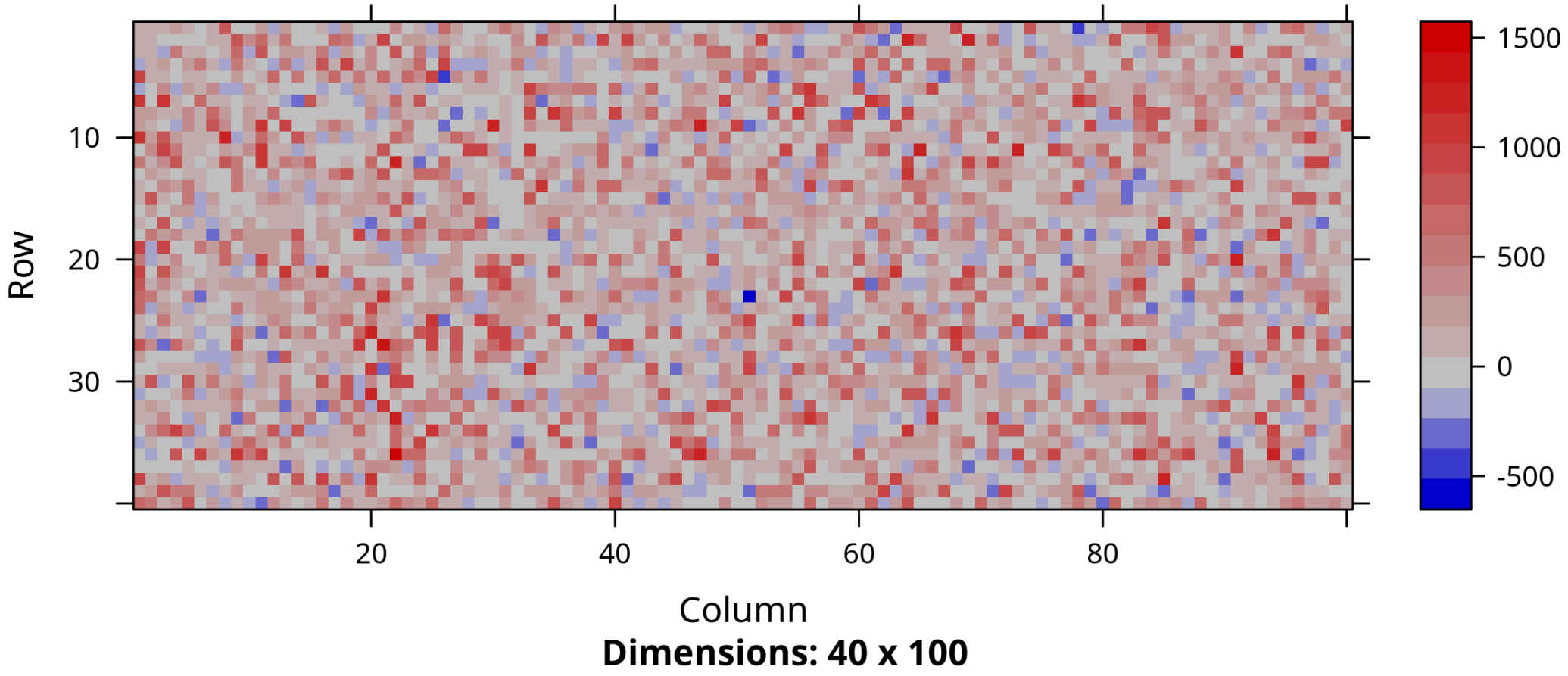

Row
10
20
30
20
40
60
80
Column
1200
1000
800
600
400
200
0
Dimensions: 40 x 100

特異値分解 (SVD)

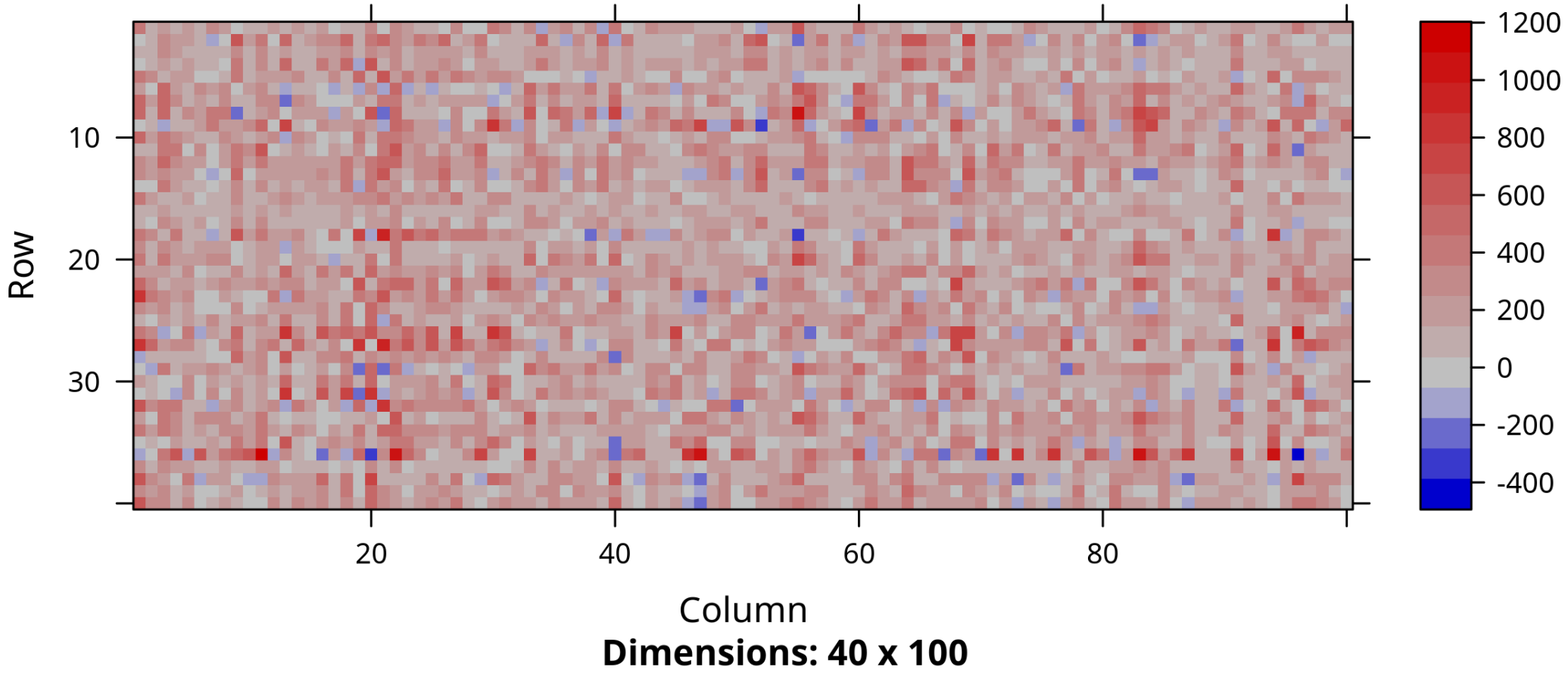
$$X \approx \hat{X} = DST'$$



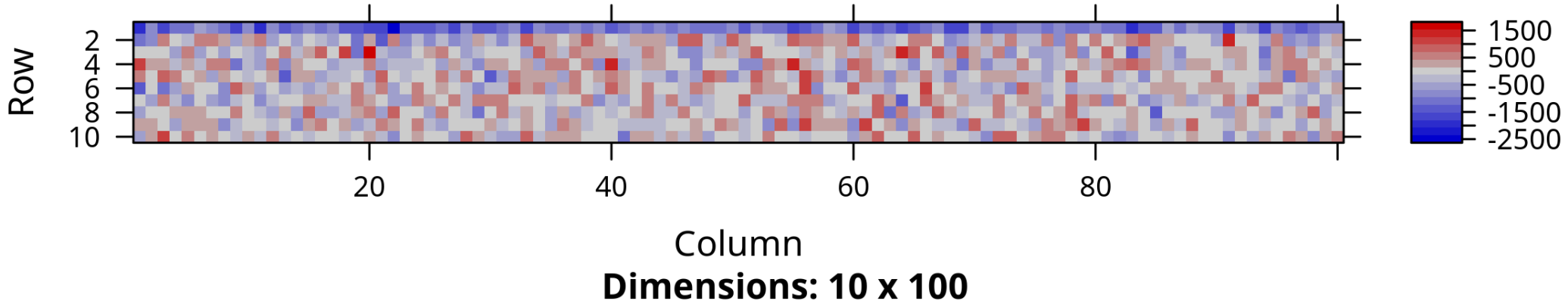
スムーズ化された文書特徴行列 (k = 10)



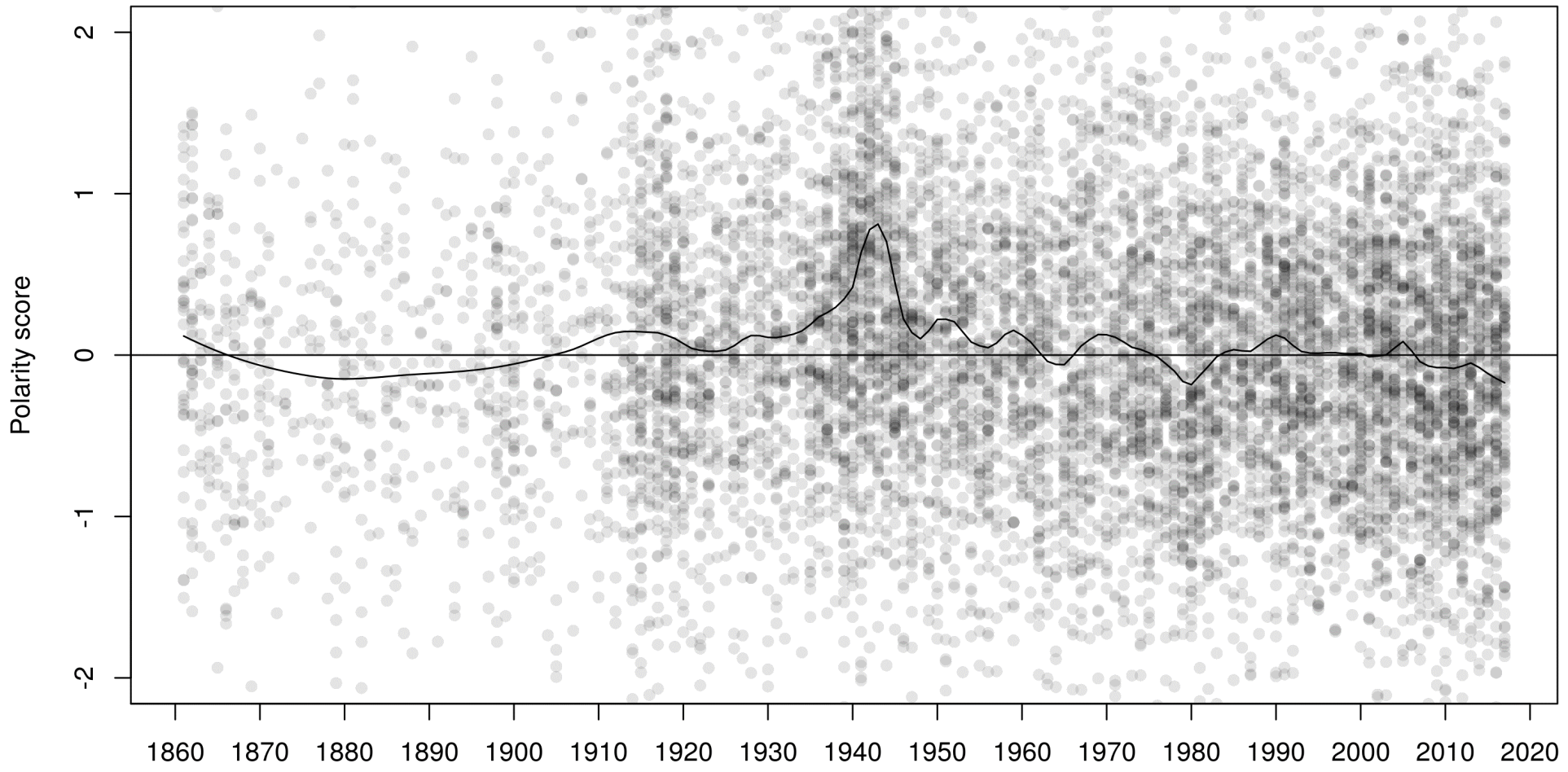
スムーズ化された文書特徴行列 (k = 5)

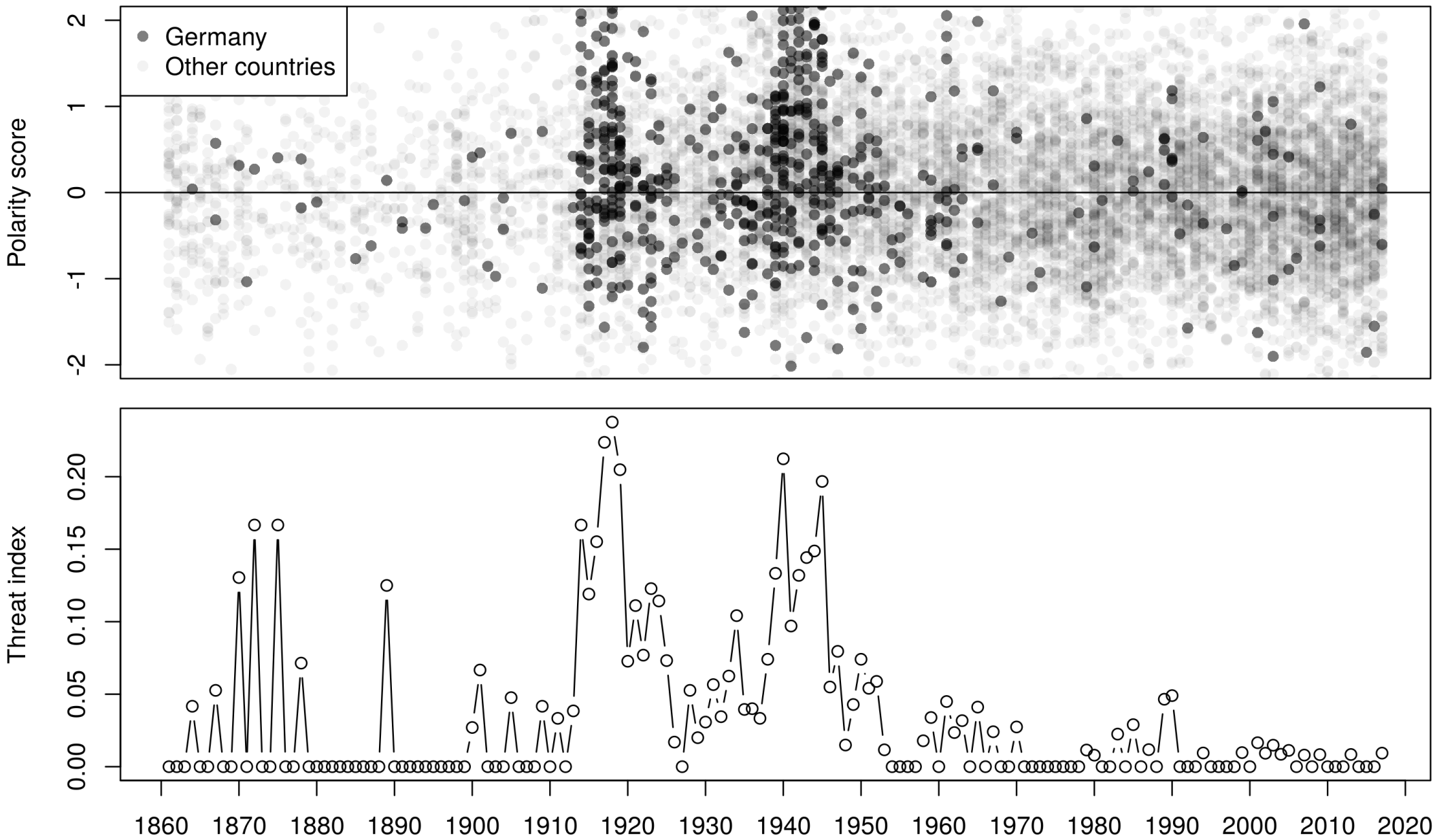


語ベクトル


Row
2
4
6
8
10
20
40
60
80
Column
Dimensions: 10 x 100
1500
500
-500
-1500
-2500

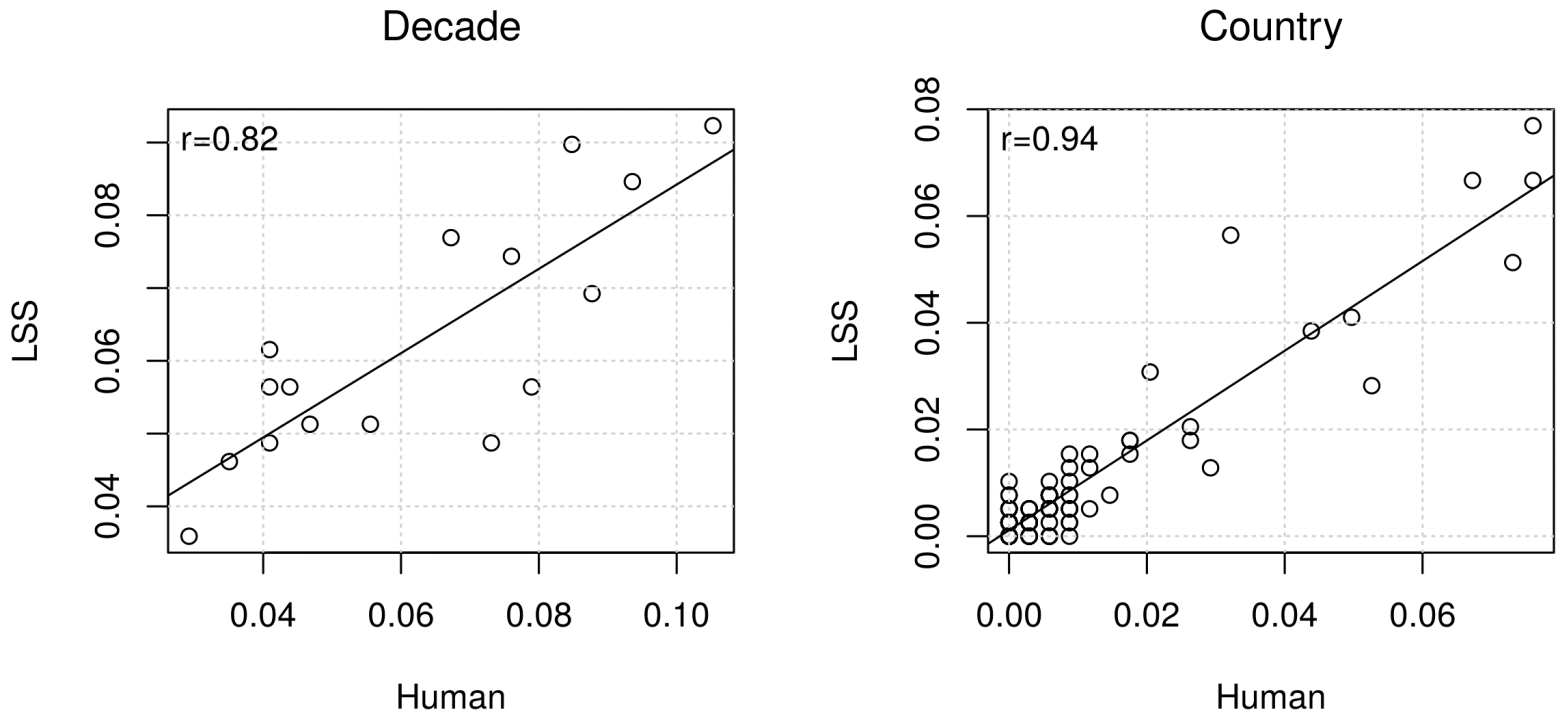
LSSによる記事の分類




Germany
Other countries
Polarity score
2
1
0
-1
-2
Threat index
0.00
0.05
0.10
0.15
0.20
1860
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020

測定精度の検証

- 人による分類との比較



まとめ

- 量的テキスト分析は社会科学においてすぐに活用できる
  - 準教師あり学習は，低いコストで多数の文書を分析できる.
    - 日本語やアラビア語を含む文書も同様に分析できることが確認されている.
    - 特殊なキーワード辞書が存在しなくても，種語だけで任意の尺度を測定できる.
  - Rのパッケージだけで，データ収集から統計的分析までを一貫して行うことができる.
    - Quanteda，Newsmap，LSSはすべてCRANで公開されているオープンソースのソフトウェア.
    - 同じツールで日本語、中国語などのアジア言語の分析もできる.
  - 多くの文書データを有料及び無料のAPIを通じてダウンロードできるようになってきた.
    - NYT APIやTwitter APIはテキストは誰でも無料で利用できる.
    - NYTやFactiva，Nexisなどが商用の全文APIの販売を始めている.

追加情報

- ブログ
  - Watanabe Kohei (https://blog.koheiw.net)
- Rパッケージ
  - Quanteda (CRAN, https://quanteda.io)
  - Quanteda Tutorials (https://tutorials.quanteda.io)
  - LSX (CRAN)
- 論文
  - Watanabe, K, 2017, Measuring news bias: Russia's official news agency ITAR-TASS' coverage of the Ukraine crisis, *European Journal of Communication,* doi:10.1177/0267323117695735.
  - Watanabe, K, 2020, Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages, *Communication Methods and Measures,* doi:10.1080/19312458.2020.1832976*.*