大規模言語データと人工知能が切り開く新しい社会意識分析の可能性

「社会調査データの・デジタルデータ公共的な利活用に向けて」日本学術会議学術フォーラム, 2023.09.24

瀧川裕貴（東京大学）

自己紹介

- 東京大学大学院人文社会系研究科社会学研究室所属
- 専門は数理社会学、計算社会科学
- ビッグデータや大規模テクストデータを用いた新しい社会学方法論の確立を研究テーマに

・今日は大規模言語データの人工知能による分析が社会学の主題の一つである社会意識の分析にどのようなインパクトを与えるかについて、自分の研究もまじえて展望したいと思います。

社会科学の主要課題としての社会意識分析

- 社会意識（集合意識・集合表象）
  - 一定の集団において共有された思想や態度、意見、ものの見方
  - 性別役割意識、民族ステレオタイプ、労働観、家族観、健康や医療に対する意識、犯罪に対する意識、幸福感etc…
  - 政策決定の基礎的情報としても重要→公共的な利活用

従来の社会意識の研究手法

- アンケート調査
  - 代表的、量的
  - 情報に乏しい（発見が少ない）
  - 反応的（社会的望ましさバイアスなど）
  - 意識的（意識して言語化できることが主）

- インタビュー調査
  - 少数、質的
  - 情報は豊か（発見が多い）
  - 反応的（社会的望ましさバイアスなど）
  - 意識的（意識して言語化できることが主）

第３の方法としての大規模言語データと人工知能を用いた新しい社会意識分析の可能性

- 社会意識
- 大規模言語データ（テキスト）
- 人工知能

社会意識の測定

- 社会意識とは？
  - 一定の集団において共有された思想や態度、意見、ものの見方
  - 行動や言葉と違って見えないもの→潜在的
  - 何らかの「メディア」から何らかの「ツール」を使って測定する必要がある。

- 従来のメディア
  - 質問紙（サーベイ）
  - 会話（インタビュー）

  →反応的で意識的、意図的に収集

テキストデータ：書籍、雑誌、新聞、SNSデータetc.

- 書き手の（無意識を含めた）社会意識を反映
- 読者に一定の社会意識を植え付ける

「見出されたデータ」としてのテキストの特徴	
非反応的	自然な状況で書き手が表現したもの。現場性
潜在意識的	書き手、読み手が必ずしも意識していない考え方、バイアスや偏見など。言葉の選び方、連想の仕方
歴史的	テキストさえあればアンケートやインタビュー記録が存在しない過去の社会意識にアクセス可能

人工知能

- 認識や推論、判断を人間の代わりに機械により実行する方法
- データから学習することにより能力を獲得

➢テキストを「読む」ことで人間のものに似た意味やカテゴリを獲得可能

人工知能・バイアス・社会意識

- データのバイアスや差別的態度を学習してしまう可能性

  例）MicrosoftのTay

→社会意識の研究として考えれば、人工知能のふるまいから逆算することでデータの、つまりは人間のバイアスや差別意識、社会意識を特定可能

人工知能によるテキスト分析まとめ

- 人工知能によるテキスト分析を通して、自然な状況での、非反応的で、潜在意識を含む様々な思考や態度を、歴史をさかのぼって情報量豊かに測定可能
- さまざまな人工知能のモデルがあるが、ここではテキスト分析のモデルとして単語埋め込みモデルに焦点

単語埋め込みモデルによるテキスト分析

- 単語埋め込みモデルとは？
  - 単語の分散表現（ベクトル表現）を得るためのモデル
- 単語の表すカテゴリや概念を定量化することで、カテゴリや概念の意味や関係性を量的に分析可能

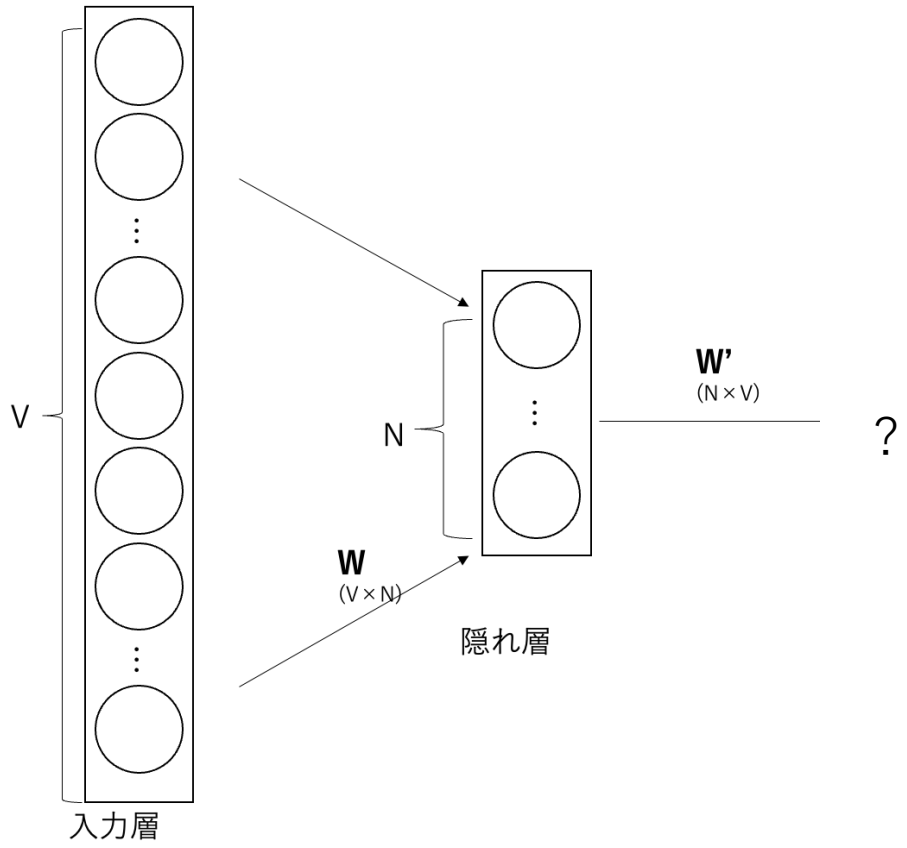
女性 =
```
[0.1776111326084173 2,
0.919838573152116 3,
0.42507285078693 8,
0.29866627399661 04,
0.08203703765477 743,
0.73336927779059 5,
0.68586246835541 54,
0.76047015677415 36,
0.08438666176841 88,
0.38995440782073 53,
0.12289144186867 14]
```

男性 =
```
[0.9794849463480628,
0.731621981532410 2,
0.76509773580411 2,
0.12997237380831 084,
0.97568582306651 66,
0.37260071109950 11,
0.41543818563064 67,
0.07709026961802 501,
0.09562466916991 641,
0.72171093156163 05,
0.64744430497635 24]
```

単語埋め込みモデル

- ここでは人工ニューラルネットワーク（ANN）モデルに限定
- ANNモデルでは、ある課題を解決するために、単語の意味を表すようなベクトル表現を獲得する。

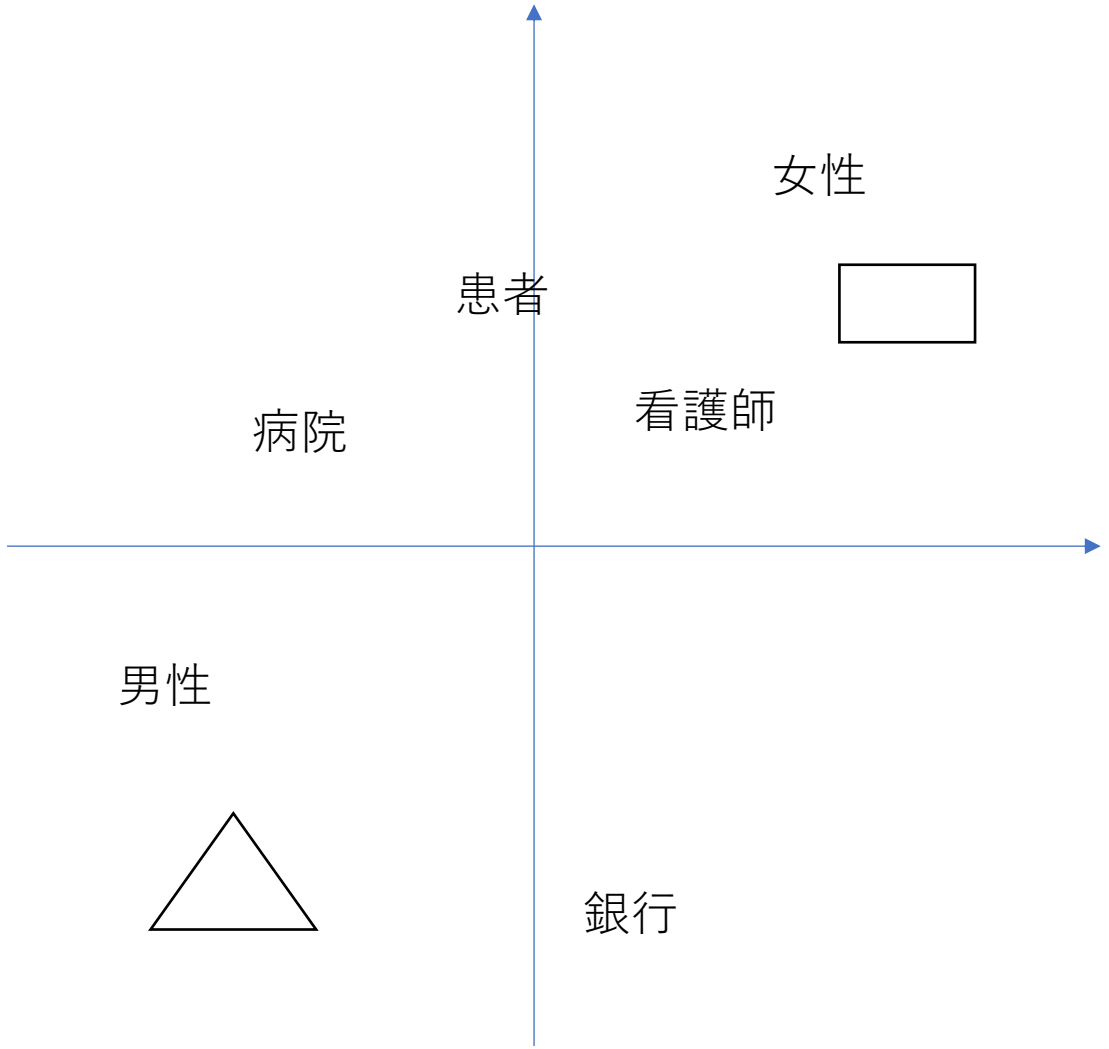
彼女は都内の病院で□□としてはたらいて
患者から好かれている



単語のベクトル表現

- 女性、病院、患者などの個々の単語を数百次元のベクトルとして表現

→語の意味が空間的に配置される意味空間を獲得
例えば、単語間の距離（類似性や差異）を定量的に評価できるように



ステレオタイプの学習

- 学習データのステレオタイプ
  - データにおいて、彼女＋病院と共起するのが、医師より看護師が多いとしたら…

「彼女は都内の病院で□□としてはたらいて患者から好かれている」

「看護師」と答える

→女性から医師より看護師を連想するステレオタイプを学習

単語埋め込みモデルを用いた社会意識分析

- ステレオタイプの分析

  「女性は□である」

  「中国人は□である」

  →カテゴリと特定の事物・概念（職業、性格、ふるまい等）を結びつける仕方が問題

- 概念やカテゴリの意味分析

  「階級にはどんな文化的意味、イメージが結びつけられているか」

  「人は幸福をどのように意味づけているのか。どのようなことに幸福を見出すのか」

  →カテゴリのもつ複数のイメージや意味づけが問題

ステレオタイプの研究

- 心理学のステレオタイプ研究
  - IAT（潜在連合テスト）：人が意識できない潜在的態度を測定するテスト
  - 「男性：仕事」「女性：家庭」などの語の結びつきの強さを測定することで潜在的偏見を明るみに出す
- 単語埋め込みモデルによるテキスト分析
  - 語をベクトル化して、距離の測定が可能
  - 「男性：仕事」の距離と「女性：仕事」の距離、「男性：家庭」の距離と「女性：家庭」の距離等を比較してIATに類似した測定が可能
    →WEAT

IATとWEAT

Table1 「花VS虫」と「快いVS不快な」の連合強度の測定における刺激の例

カテゴリー	刺激
花	バラ、さくら、チューリップ、ユリ、ライラック
虫	あり、ごきぶり、はち、くわがた、はえ、くも
快い	やさしい、うれしい、すばらしい、かわいい、うつくしい
不快な	きたない、くさい、いやしい、みにくい、やかましい

森尾（2007）

WEAT
A)「花」と「快い」のカテゴリに属する語同士の距離と「花」と「不快な」のカテゴリーに属する語同士の距離の差
B)「虫」と「快い」のカテゴリに属する語同士の距離と「虫」と「不快な」のカテゴリーに属する語同士の距離の差

AとBを比較。A-Bとして差をとる。A>Bなら「花」は「虫」よりも「快い」に結びついている。

ステレオタイプの研究

- Caliskan et al. (2017)はIATとWEATの結果を比較して、両者が一致することを示している（データはCommon Crawl）。

  「花vs.虫」「快vs.不快」
  「楽器vs.武器」「快vs.不快」
  「男性名vs.女性名」「仕事vs.家庭」
  etc..

- Chalesworth et al. (2021)はさまざまなテキストデータにおいて男性-女性に関するステレオタイプがみられることをWEATで測定している。

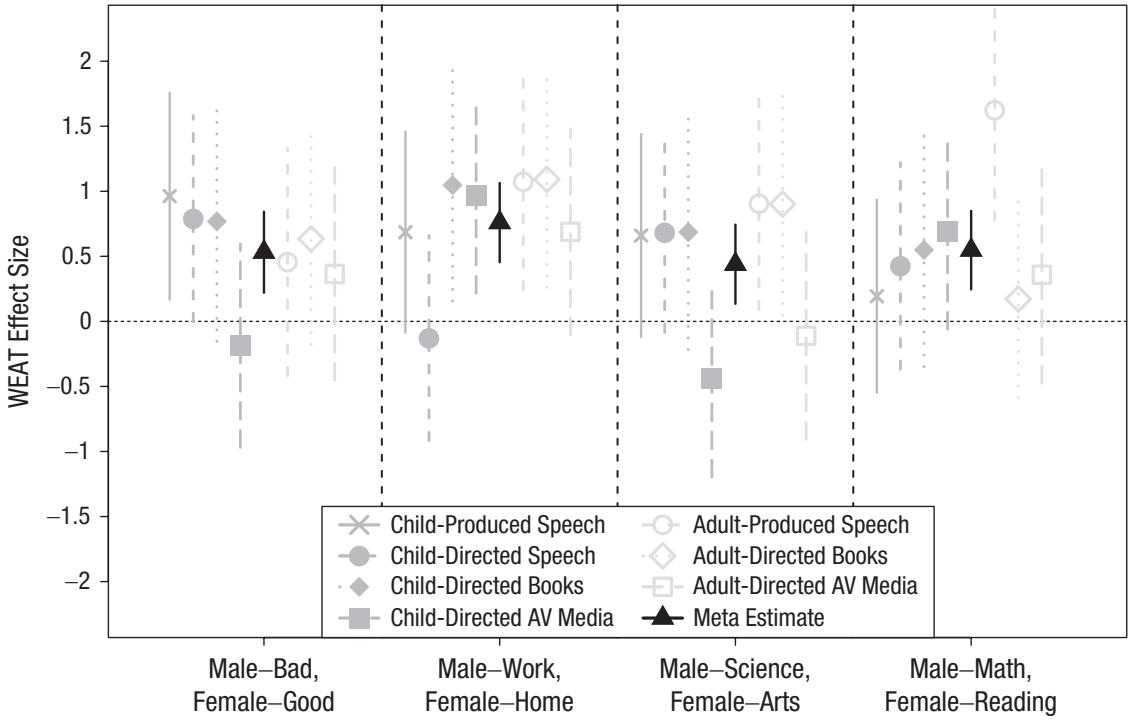
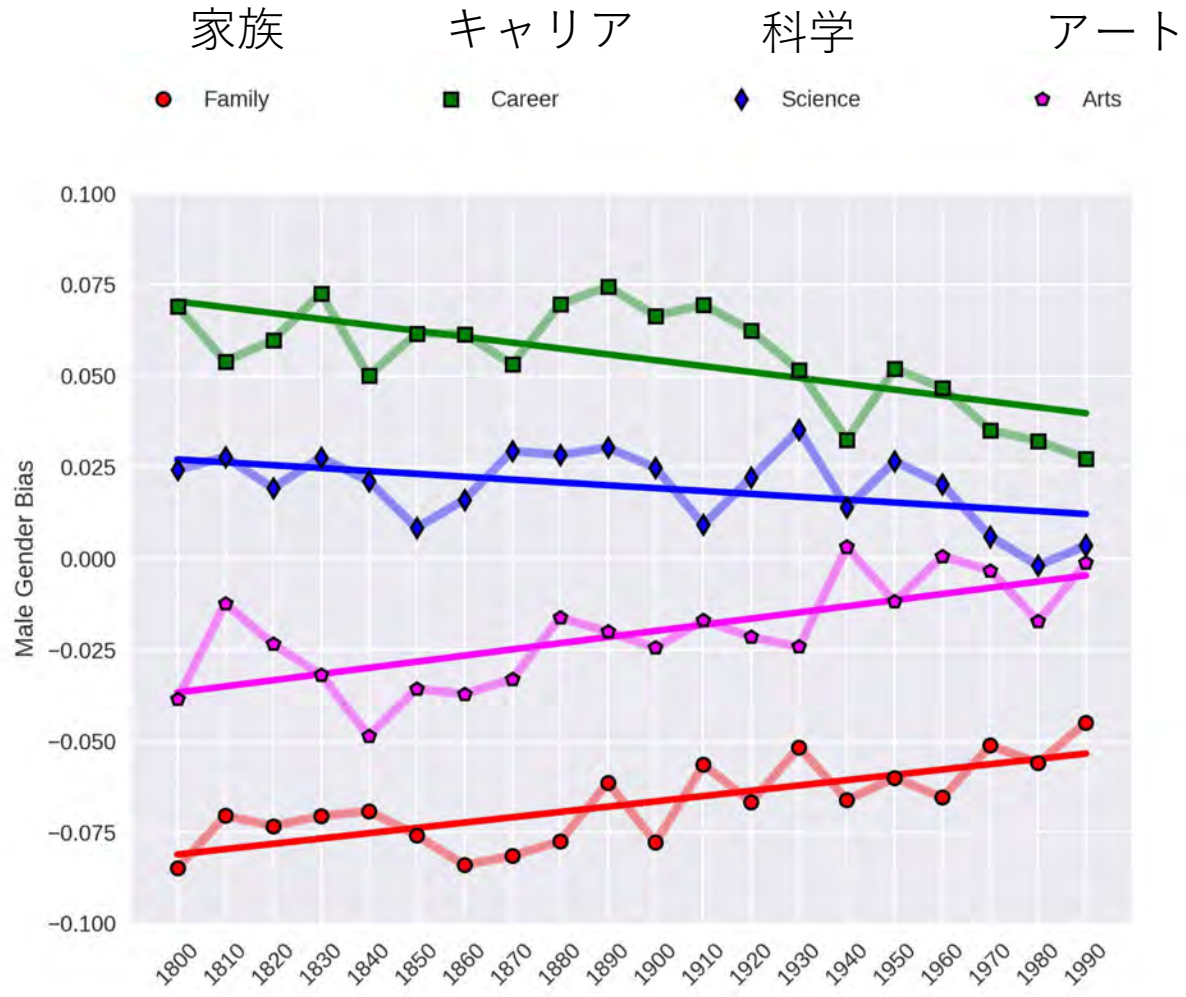


Fig. 1. Gender associations in child and adult language (Study 1). Word-Embedding Association Test (WEAT) *D*-score effect sizes are shown as a function of gender association (stereotypes and attitudes), separately for each type of child-directed/child-produced and adult-directed/adult-produced speech, books, and audiovisual (AV) media. Also shown is the meta-analytic estimate, which was computed from a fixed-effects meta-analysis across all sources. Error bars represent 95% confidence intervals computed from the standard error (i.e., the standard deviation of the permutation distribution of WEAT effect scores).

Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021).

ステレオタイプの研究

- Jones et al. (2019)はGoogle Ngramというテキストデータを用いて、英語圏の1800-2000までのジェンダーステレオタイプをWEATと同様の手法で分析している

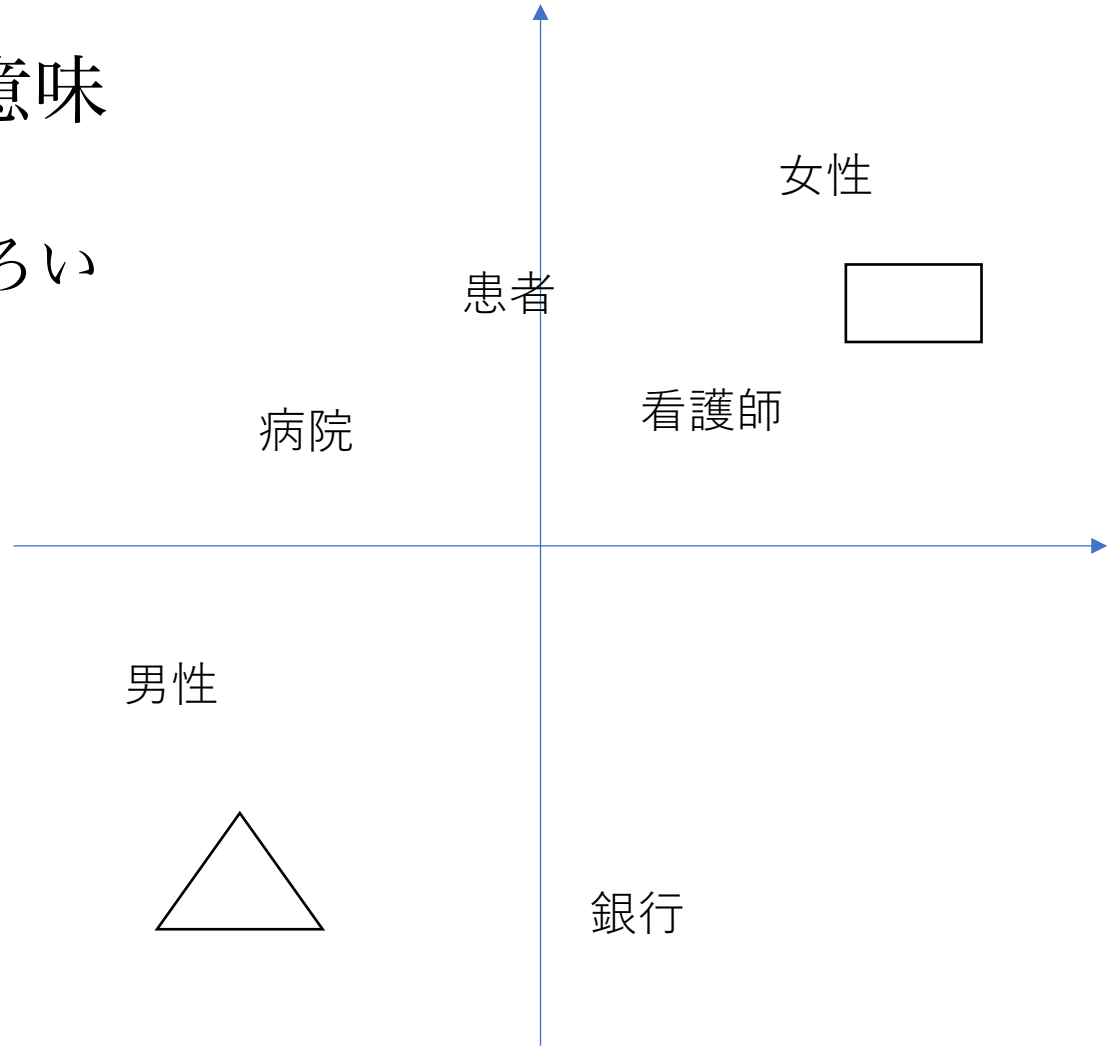


Joens et al. (2017).

概念の意味分析：意味空間アプローチ

- 埋め込みモデルの与える意味空間の次元は解釈不能

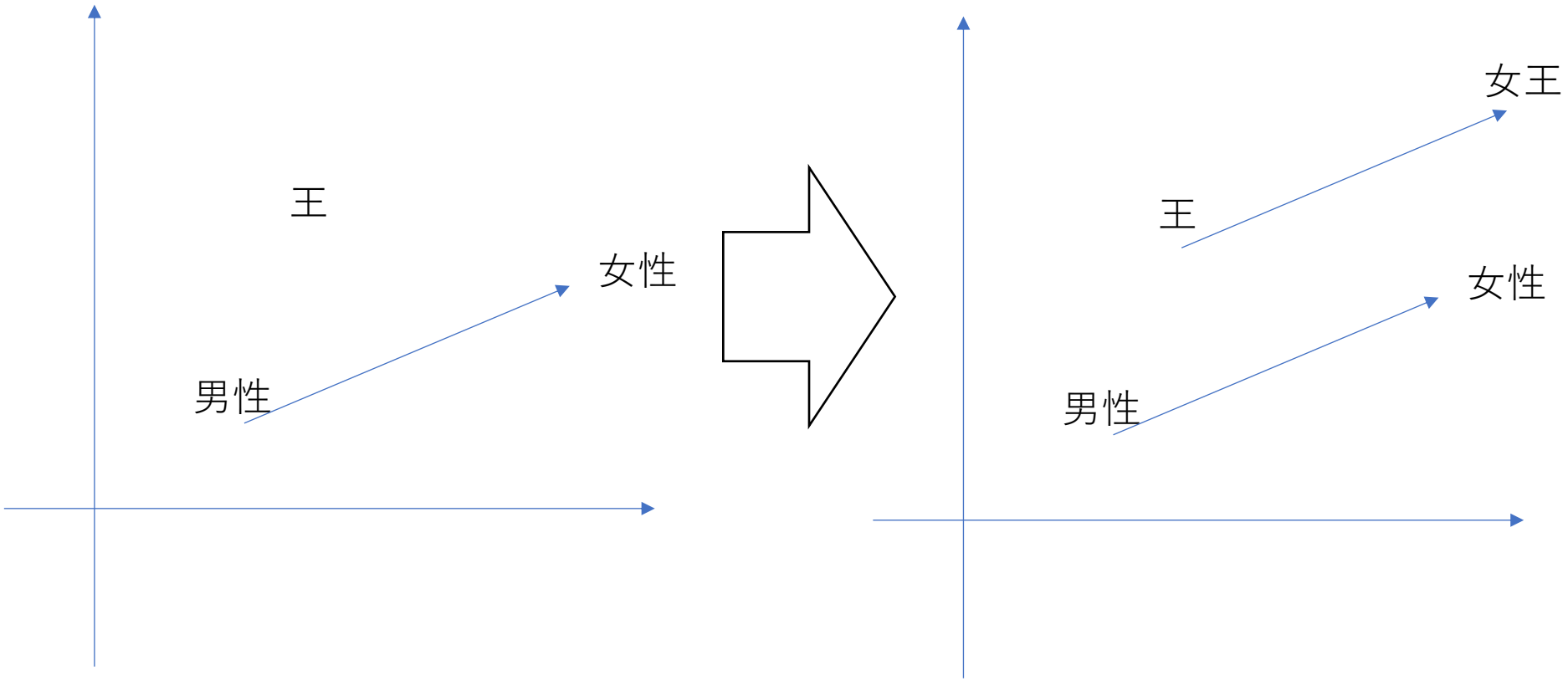
→解釈可能な次元が作れればいろいろ分析できる



概念の意味分析：意味空間アプローチ

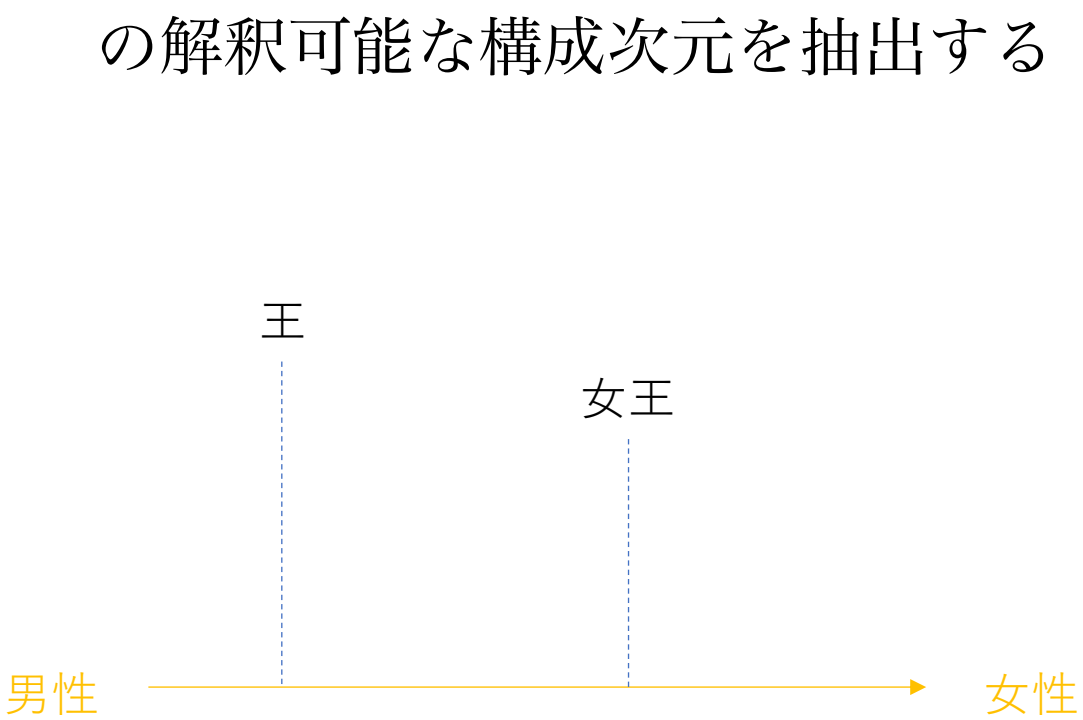
- 単語埋め込みモデルとアナロジー計算

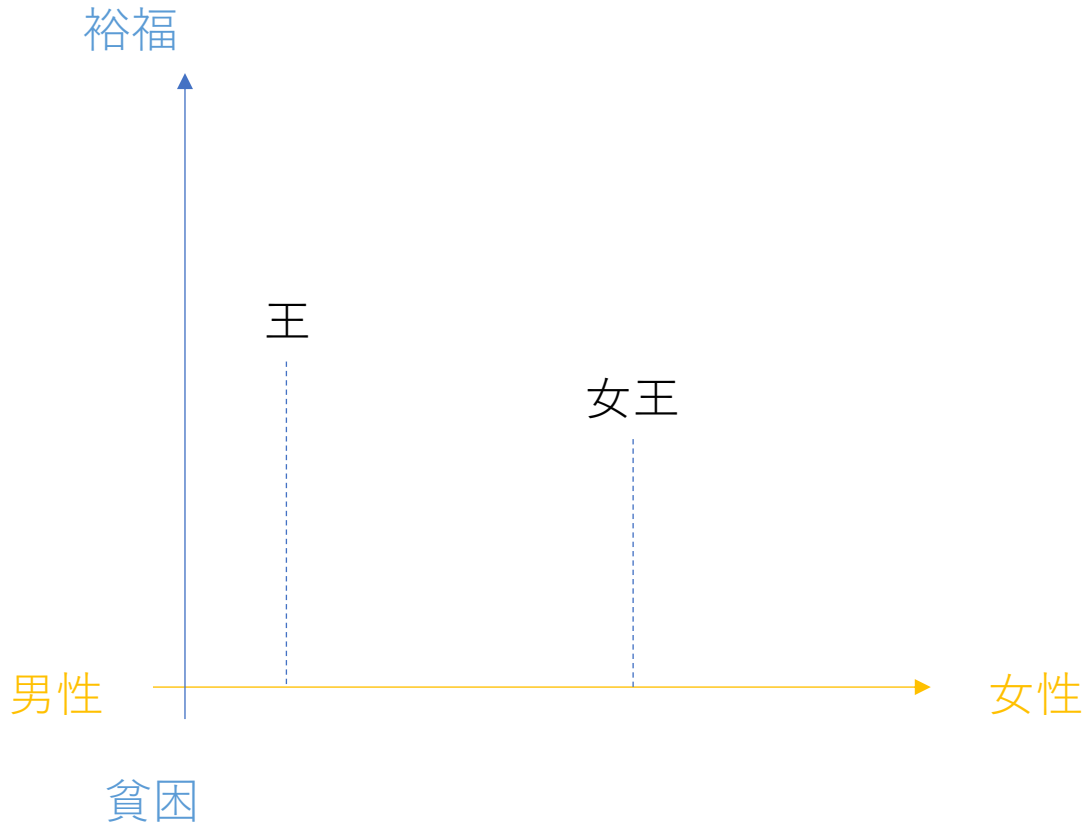
王　－　男性 + 女性　＝　女王



概念の意味分析：意味空間アプローチ

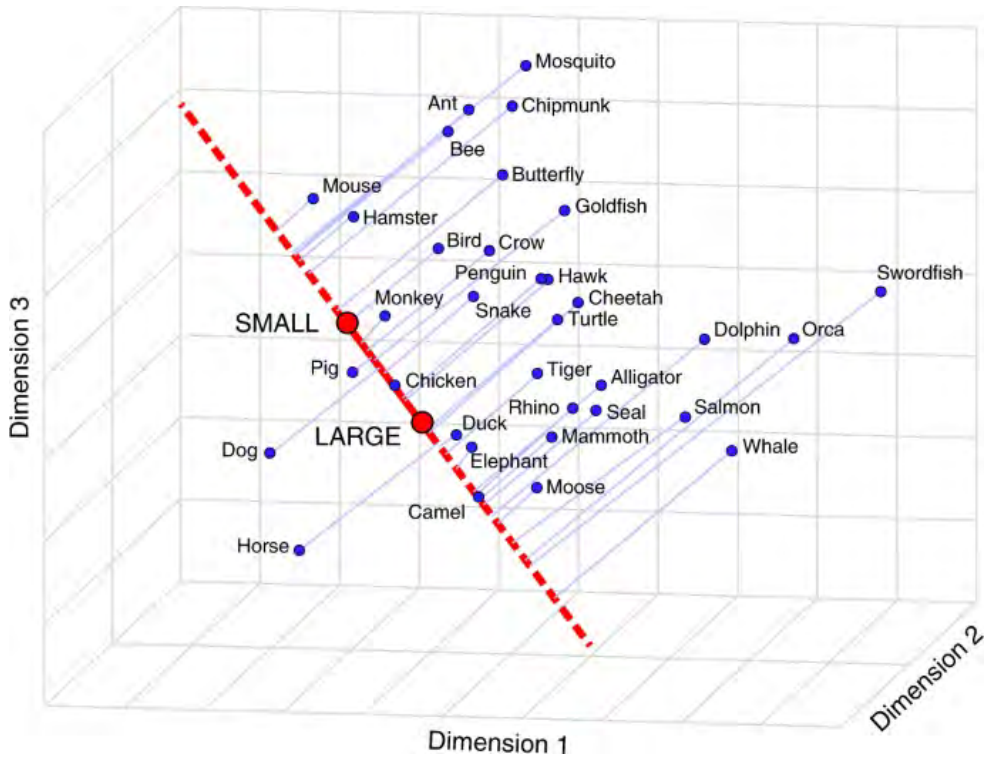
- 対義語ベクトルのペア集合から意味空間の解釈可能な構成次元を抽出する

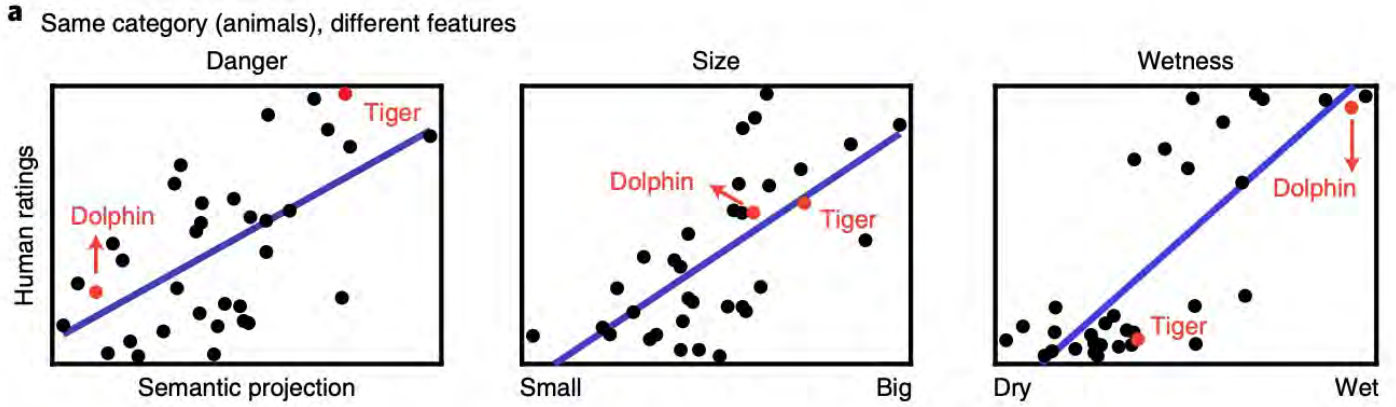




概念の意味分析：意味空間アプローチ

- 意味空間は概念（の意味）の複合的情報を潜在的に含む
→適切な操作で意味次元を特定可能





- 動物（イルカ、とら）という複合的概念を特定の意味次元（危険性、サイズ、湿り気）に分解し、位置づける

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022).

概念の意味分析：意味空間アプローチ

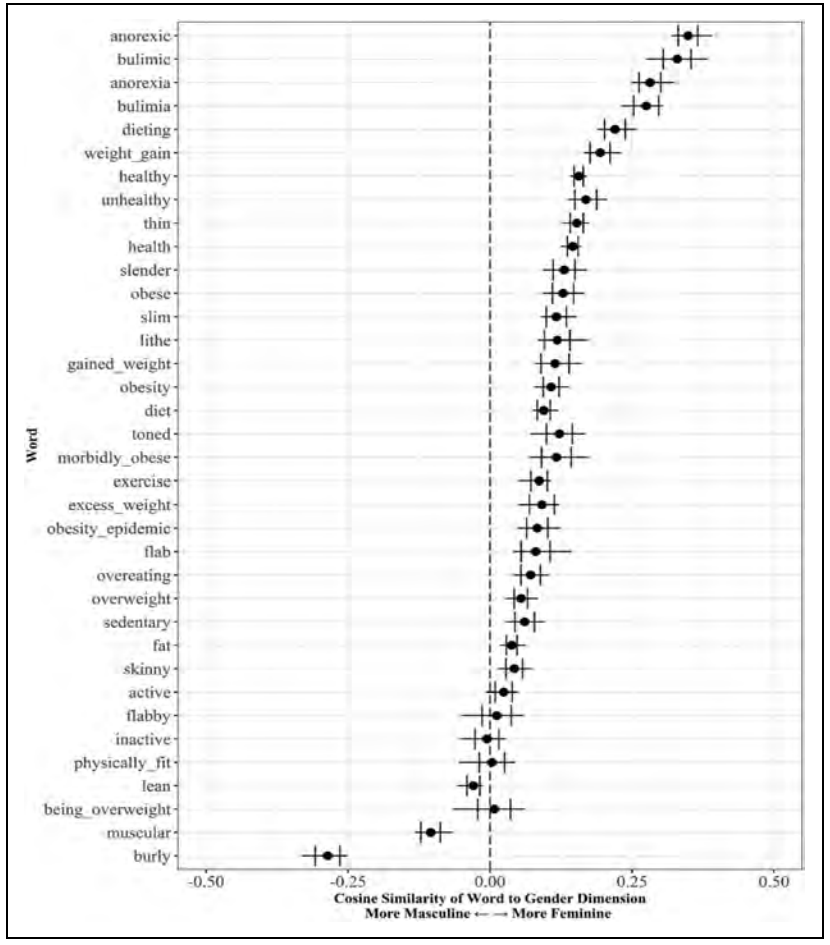


Figure 2. Gendering of obesity-related words.

Arseniev-Koehler, A., & Foster, J. G. (2022)

- anorexic: 拒食症の
- bulimic: 過食症の
- anorexia: 拒食症
- bulimia: 過食症
- dieting: ダイエット
- weight gain: 体重増加
- healthy: 健康的な
- unhealthy: 不健康な
- thin: 痩せている
- health: 健康
- slender: 細身の
- obese: 肥満の

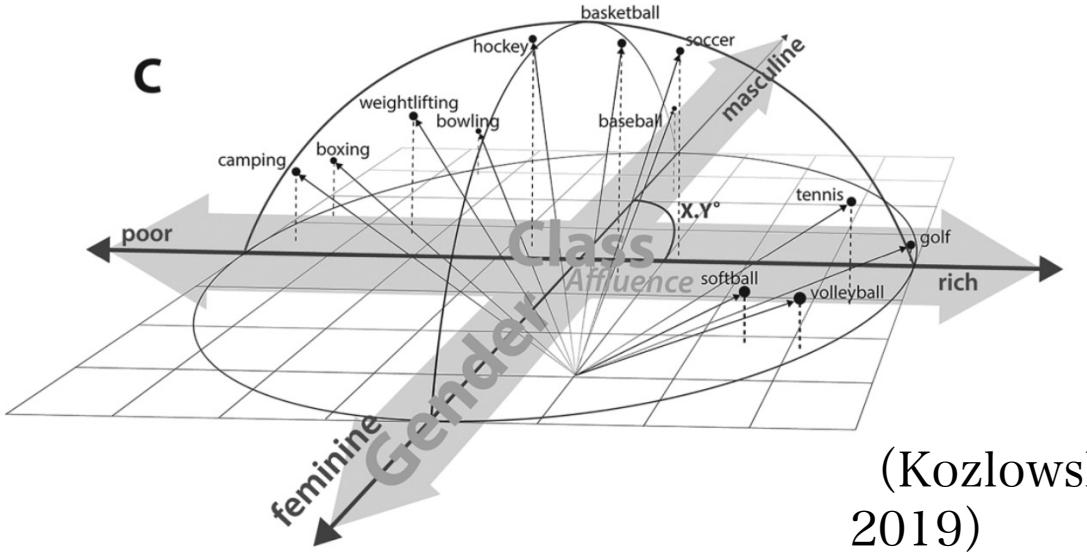
- 構成次元に語を投射して語が表す事象の文化的意味を検討
- Arseniev-KoehlerとFosterは肥満に関わる語が女性的イメージと深く関係づけられていることを明らかに

概念の意味分析：意味空間アプローチ

- さらに進めて、意味構成次元間の関係を分析することで、抽象的な概念間の関係性を分析することができる（cf. Kozlowski et al. 2019）

→社会階級に対する意識のような複雑な社会意識を分析可能

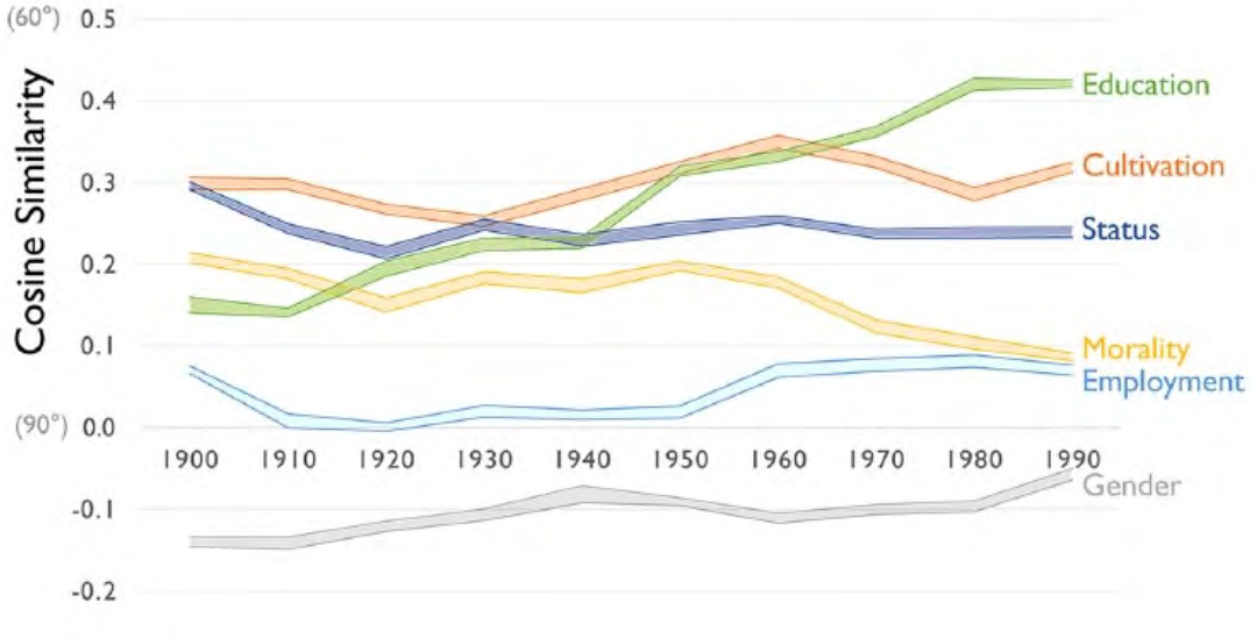
- 例えば、（経済的）階級とジェンダーの意味次元を構成して、階級概念とジェンダー概念の関係性を分析すること。



(Kozlowski et al. 2019)

概念の意味分析：意味空間アプローチ

- Kozlowskiらは社会階級の複合的な意味と歴史的変遷の分析を試みている。
- 社会階級の意味を「経済的」「教育」「文化教養」「ステータス」「道徳性」「従業上の地位」「ジェンダー」の6つの意味次元に分析し、「経済的」階級次元が他の次元とどう関係しているかの歴史的トレンドを分析した。



(Kozlowski et al. 2019)

意味空間アプローチの応用：幸福の多次元的意味の分析

- アンケートでは明らかにすることが難しい幸福をめぐる複合的な意識を明らかにする。
- アンケートの存在しない戦前の社会意識を明らかにし、歴史的に分析する

この研究は、ムーンショット型研究開発事業：目標９研究開発プロジェクト「脳指標の個人間比較に基づく福祉と主体性の最大化」の一環として行われたものです。

意味空間アプローチの応用：幸福の多次元的意味の分析

- 以後、ウェブ上ではスライド不掲載

テイクホームメッセージ：大規模言語データと人工知能を用いた新しい社会意識分析の可能性

情報量豊かで、無意識の反応も含めた自然な状況での社会意識を、歴史をさかのぼって検討することが可能