日本音響学会第150回(2023年秋季)研究発表会

深層学習で獲得される音声シンボルは自然言語シンボルと同様に Zipf 則に従うか？

前田 紘希，○高道 慎之介，朴 浚鎔，猿渡 洋
（東京大学）

MOONSHOT RESEARCH & DEVELOPMENT PROGRAM
AVATAR SYMBIOTIC SOCIETY

Zipf則（ジップ則，Zipf's law）とは？

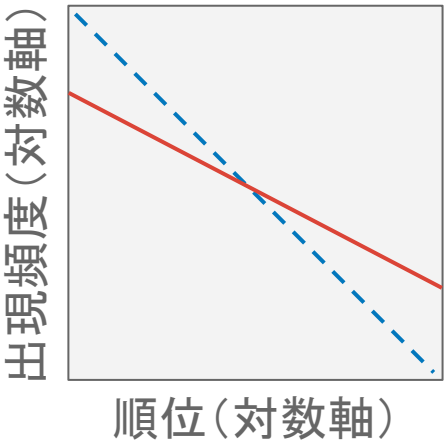
- **Zipf則** [Zipf49]
  - データ集合に含まれる要素の出現頻度に関する経験則
  - “ある要素の出現頻度が k 番目に大きいとき，その頻度は 1 番目に大きい要素の頻度の $1/k$ である”

パラメータ

$$f_r = ar^{-\eta}$$

出現頻度　順位

- **両対数グラフで図示すると直線状になる**



- グラフが直線状である場合：
  - $\eta \fallingdotseq 1$：”当該データ集合は Zipf 則に従う”
  - $\eta > 0$：”当該データ集合は冪乗則に従う”
- すなわち，Zipf 則は冪乗則の特殊例

自然言語シンボルは Zipf 則に従う

- **具体的には，データ集合 ＝ ある文章における単語とする場合**
  - この場合，自然言語シンボルは単語を指す (文字 n-gram でも同様)
  - 例えば “the”, “and” がそれぞれ 1, 3 位のとき，”and” の出現回数は ”the” の 1/3

- **Zipf 則に従う事実とそこからの逸脱は自然言語の解析手段** [田中19]
  - 子供の言語発達：2 ～ 4 歳の発話した単語は直線にならず上に凸. すなわち子供の発話は頻度の大きい単語に偏る.
  - 言語間差異：異なる表記システムを持つ言語の間では分布が異なる

音声の話：深層学習で獲得される音声シンボル

- **音(声)合成の機械学習では，離散シンボル (音声シンボル)による音声表現が主流**
  - 例えば，GSLM [Lakhotia21]，SoundStorm [Borsos23]
  - 離散シンボルを介した，完全なデータ駆動型の音声表現獲得
  - 離散シンボルは，音価や言語的意味の表現と見做すことも可能

- **自然言語シンボルと音声シンボルを対比すると**
  - 自然言語シンボル：音韻や意味を表すために人間が定義したもの
  - 音声シンボル：音韻や意味をデータ駆動で表したもの

自然言語シンボルにおいて成り立つ Zipf 則は
音声シンボルにおいても成り立つか？

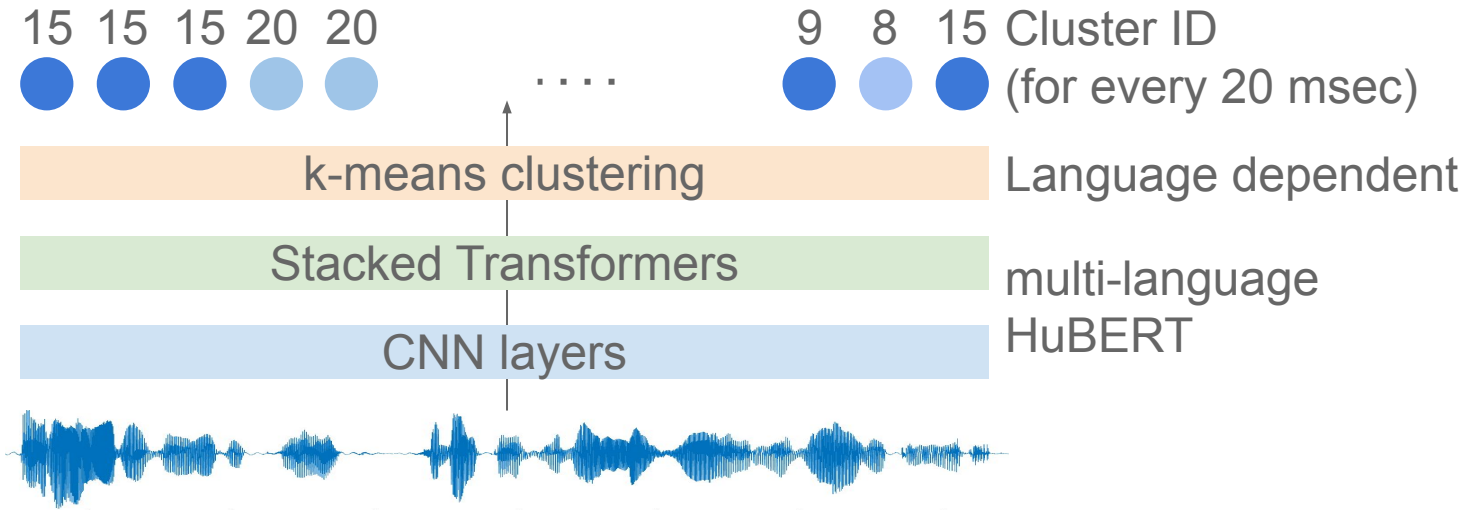
本発表の概要

- **深層学習で獲得される音声シンボルは自然言語シンボルと同様にZipf 則に従うか？（本発表の題目）**
  - これを調べて何の意味があるかについては，最後に述べます．

- **検証方法**
  - GSLM を用いて日本語と英語について調査
  - 文と音声の対データを用いた検証方法を提案
    - 自然言語シンボルと対比させた検証

Generative spoken language model (GSLM) [Lakhotia21]

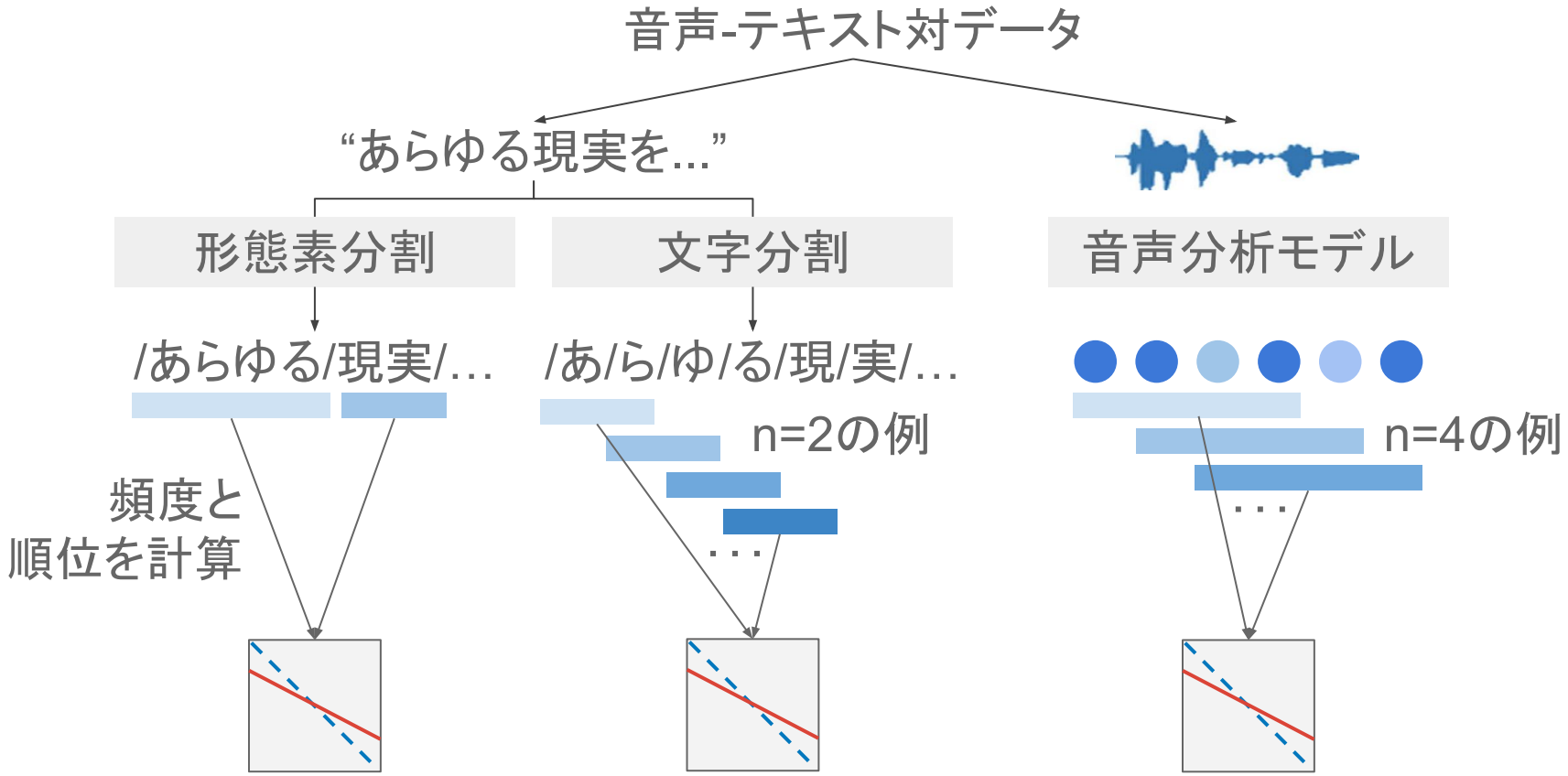
- **音声シンボルを介した音声分析合成系 (符号復号)**



- **本研究における使用法**
  - 連続する同一シンボルは，単一のシンボルに置き換える
    - 上記の例では，”15 15 15 20 20 …” → “15 20 …”
    - 継続長に対応する要素を除外するため
  - 言語依存の学習済みモデルを使用する
    - 現時点で日英の各モデルが存在 [朴23]
    - モデルに強い言語依存性があるため

自然言語シンボルと音声シンボルを対照させた検証

- **Zipf 則を検証する単位**
  - 単語 (日本語の場合は単語原型)
  - 文字 n-gram (n は1単語に対応する数)
  - 音声シンボル n-gram (n は1文字/1単語に対応する数)



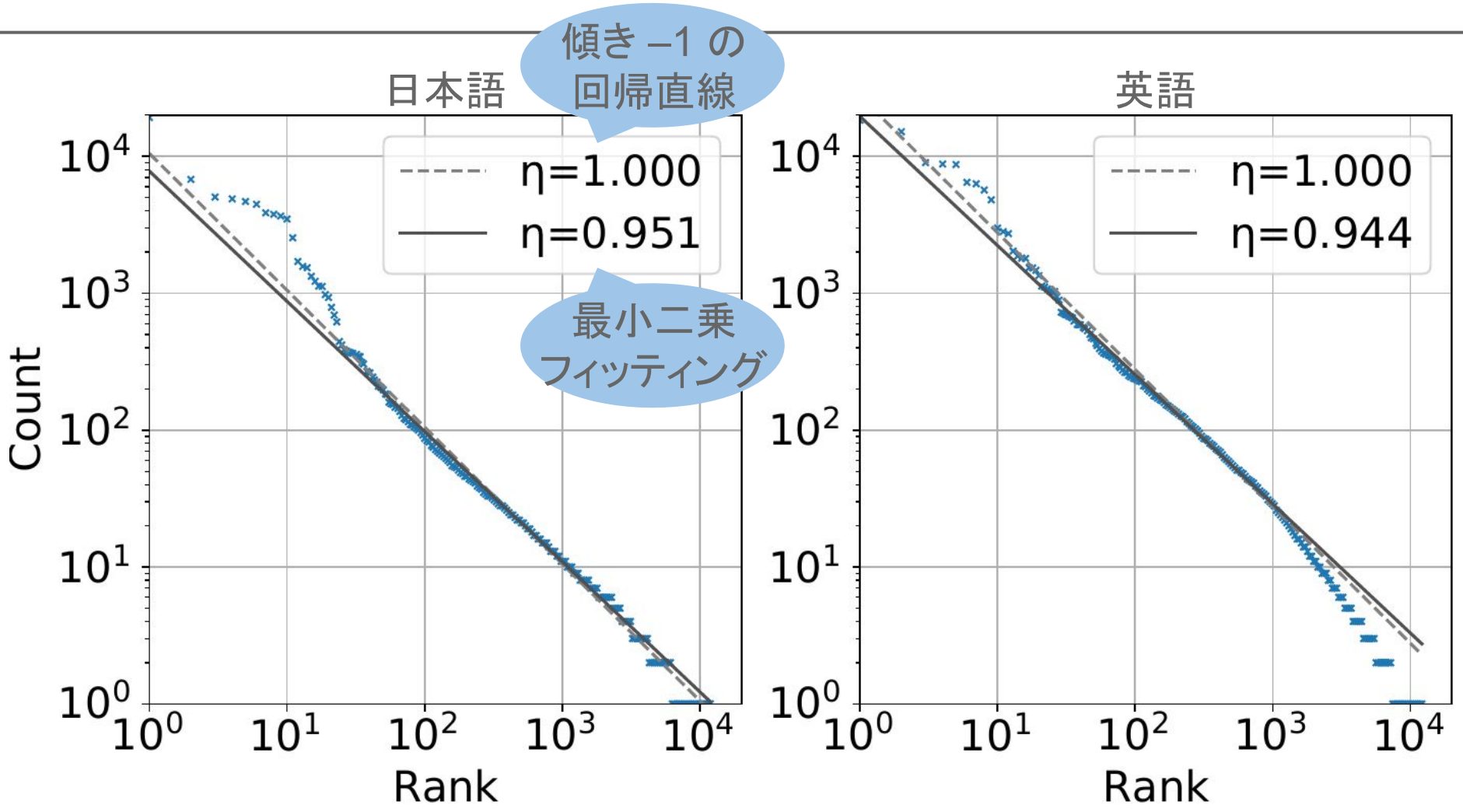
実験的評価

実験条件

要素	日本語	英語
学習済みGSLM	JSUT などで学習 [朴23] (我々が公開中)	LibriSpeech [Panayotov15] で学習 (Meta が公開中)
形態素解析	Mecab [Kudo04]	NLTK
コーパス	JSUT [Takamichi20] 約 7600 文	LJspeech [Ito17] 約13,000文
平均文字数 / 単語	1.6 (→ 文字 2-gram)	5.1 (→ 文字 6-gram)
平均シンボル数 / 文字	5.7 (→ シンボル 6-gram)	1.9 (→ シンボル 2-gram)
平均シンボル数 / 単語	8.9 (→ シンボル 9-gram)	9.0 (→ シンボル 9-gram)
回帰直線を計算する順位*	上位 0.1 ~ 10 %	上位 0.1 ~ 10 %

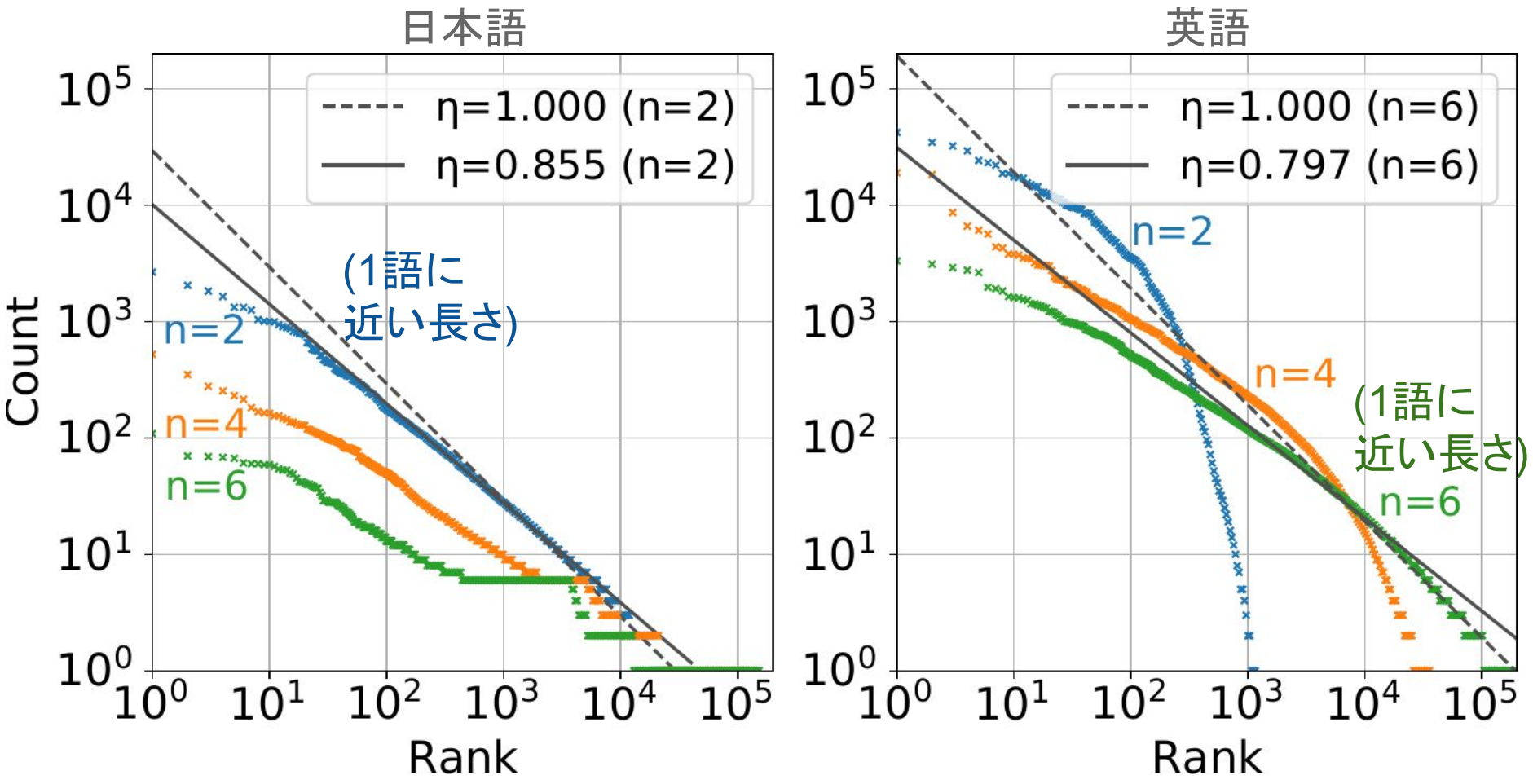
* 高頻度要素と低頻度要素は直線から乖離することが知られているため.

実験結果（単語）



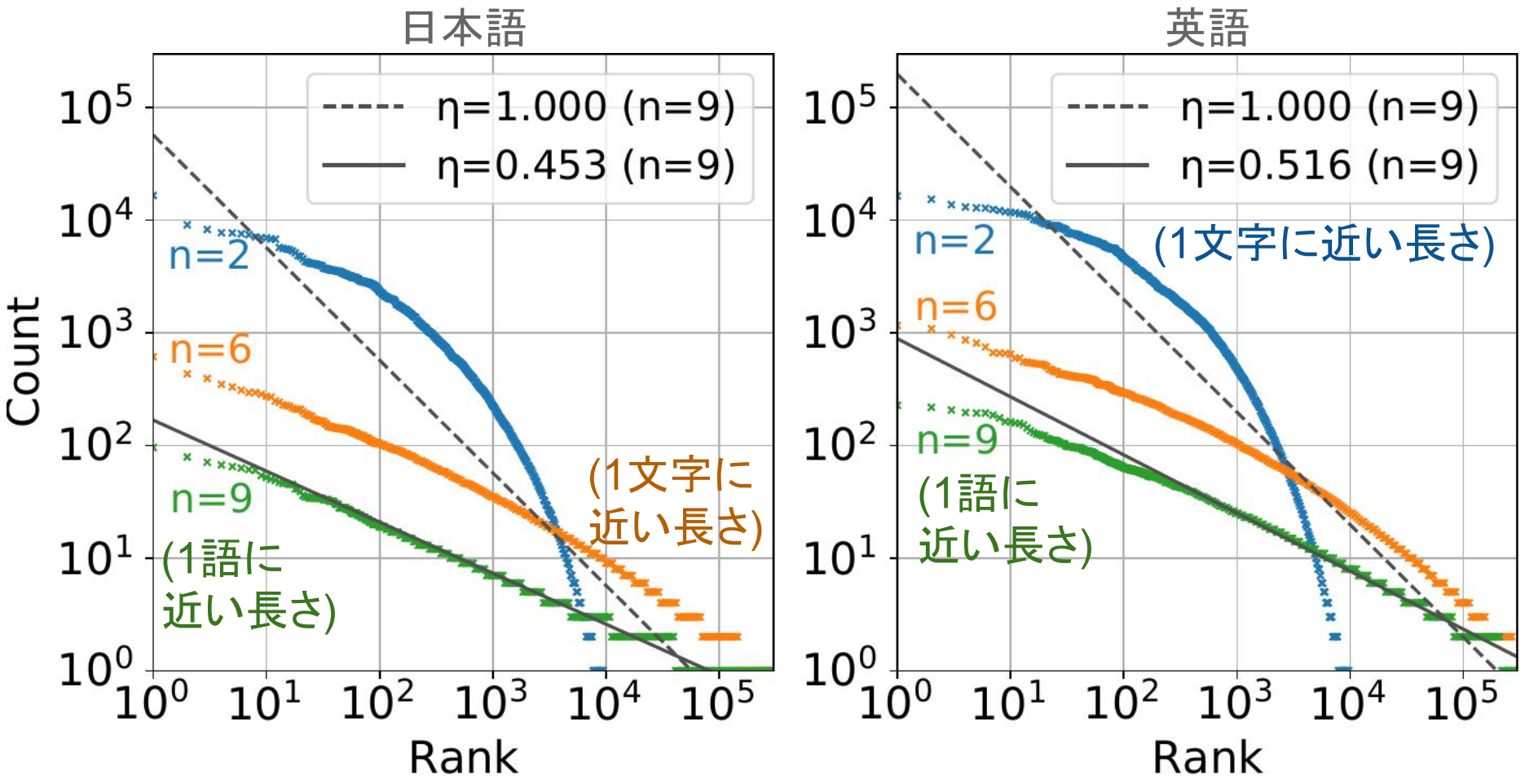
日本語・英語ともに $\eta \fallingdotseq 1.0$ であるため，
本実験のコーパスは単語の観点で Zipf 則に従う．
（既存研究と同様の性質を持つ，不偏のコーパスと言って良い）

実験結果（文字 *n*-gram）



日本語 2-gram (≒1語の文字数) の分布は直線上であり，冪乗則に従う.
英語 n-gram の分布は，n に依らず上に凸 (頻出文字列がより使われる).

実験結果（音声シンボル *n*-gram）



日本語 9-gram (≒1語のシンボル数) の分布は直線上であり，冪乗則に従う.
英語 n-gram の分布は，n に依らず上に凸 (頻出シンボルがより使われる).

まとめ

- **まとめ**
  - 音声シンボル n-gram が冪乗則に従う場合がある
  - 言語によって音声シンボルの分布が異なる

- **これを調べて何の意味があったのか？**
  - 自然言語を介さず，言語音を解析する新たな手段になるかも？
    - 例えば，子供の発話の発達，言語間差異，合成音声．
    - → 言語習熟度については arXiv 論文を参照 (“takamichi zipf”で検索)
  - 言語音以外の解析にも使えるようになるか？
    - 自然言語で記述できない音声（非言語音声）
    - 非音声（例えば環境音，楽音）