自然言語処理
―準備、形態素解析―

https://satoyoshiharu.github.io/nlp/

形態素解析および100本ノック第4章の位置づけ

- 日本語を処理する場合、処理単位へ分割する「語分ち」が必須となります（語分かちした結果を分かち書きといいます）。形態素解析は、その語分ちを提供してくれます。

- 以下で、形態素解析とニューラルネットとの関係を説明します。スライドだけだとわかりにくいので動画をリンクしています。動画のしゃべり原稿はpptのノートに入れてあります。

- 形態素解析はすでに優秀なエンジンがいくつもフリーで利用でき、プログラミング的にはそれらを呼ぶだけです。100本ノックの第4章は、形態素解析エンジンの出力をいろいろ加工してみるという課題になっています。

自然言語処理
形態素解析とは？

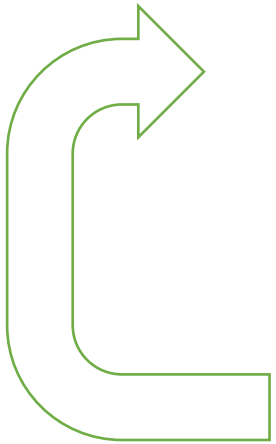
解説動画

https://yo-sato.com/

形態素とは？

表層の音：からすがきた

形態素列　　表層のテキスト：カラスが来た。



形態素

単語は表層で多様な形態をとる

からす
カラス
烏

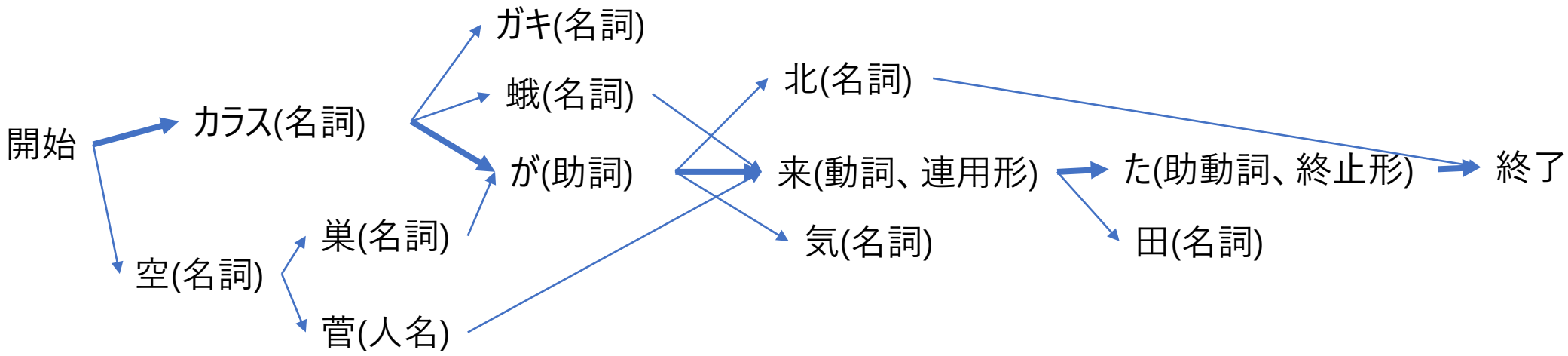
来[ク]る
来[コ]
来[クレ]
来[キ]
来[コ]い

単語　　“カラス”　　“来る”

形態素解析とは？

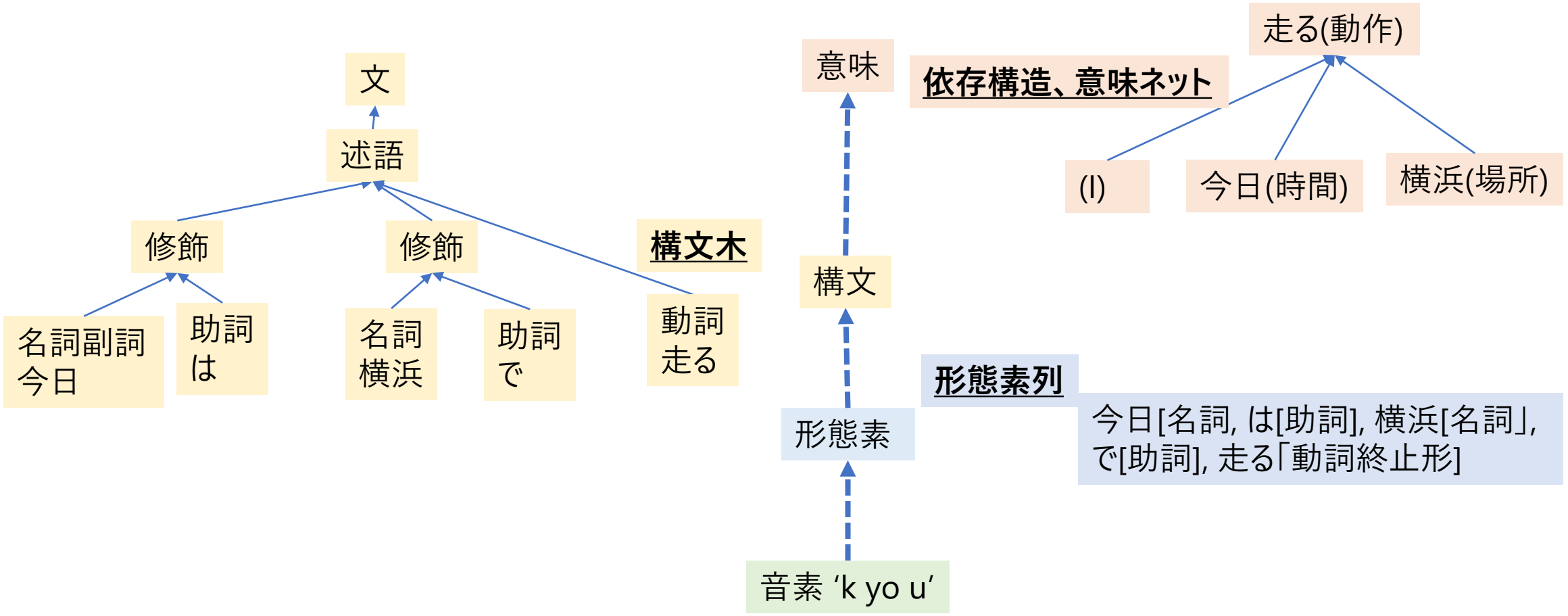
形態素解析は、表層の形態素の列から、単語の活用、表記、送り仮名などの、単語の派生タイプを明らかにしつつ、最も確からしい単語列を推定すること。

表層の音：からすがきた



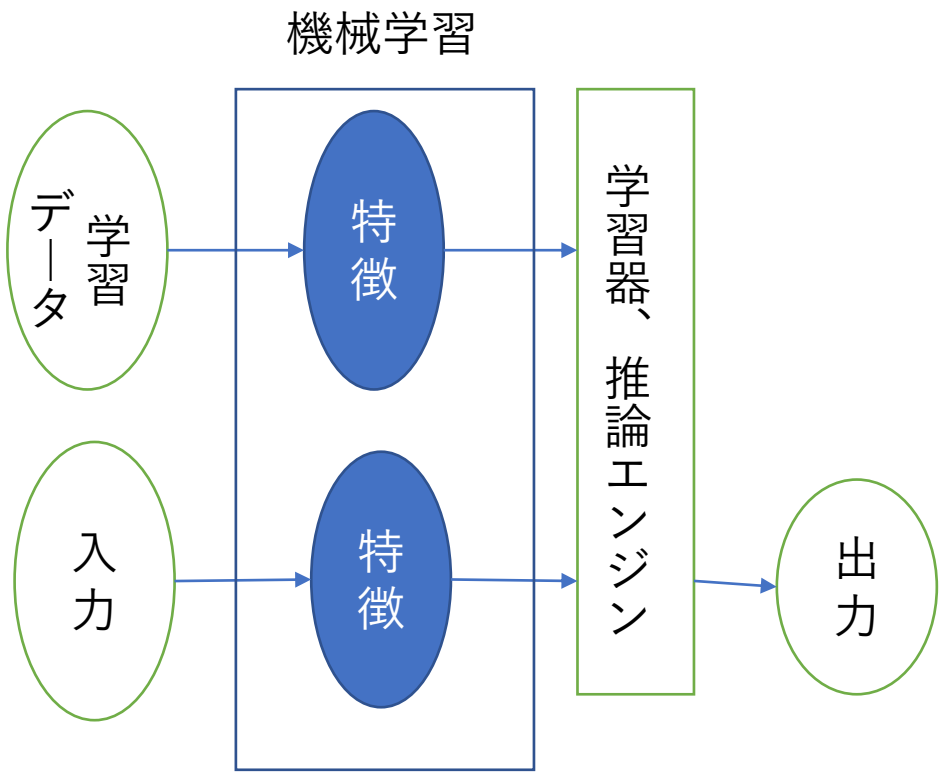
形態素解析は、自然言語処理の中で、どう利用されているか？

伝統的な自然言語処理

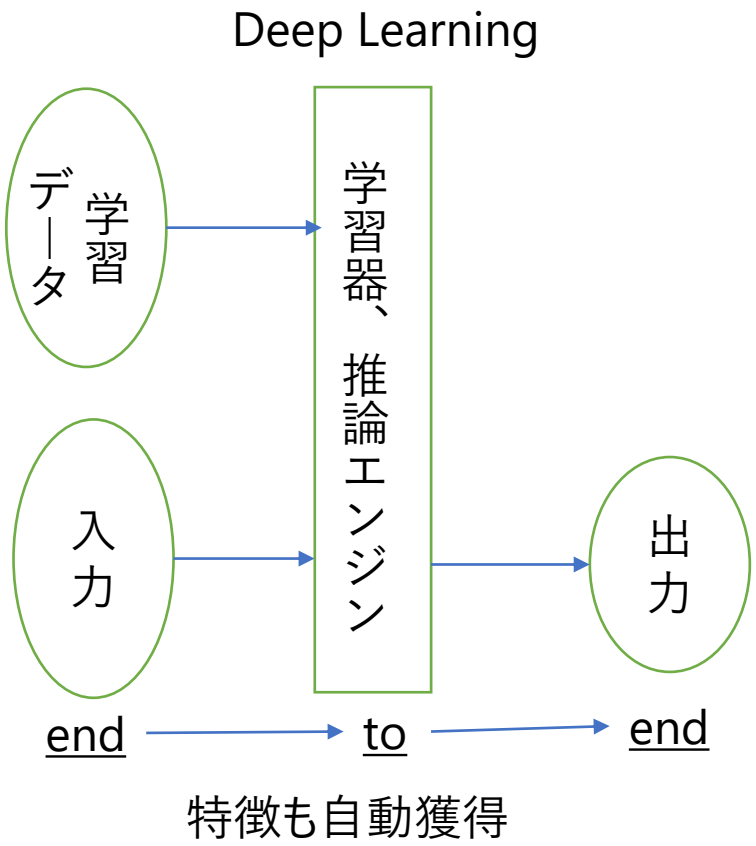


特徴(Feature)の階層

ニューラルネットで、人が特徴を設定するのでなく、それも学習するEnd-to-Endへ

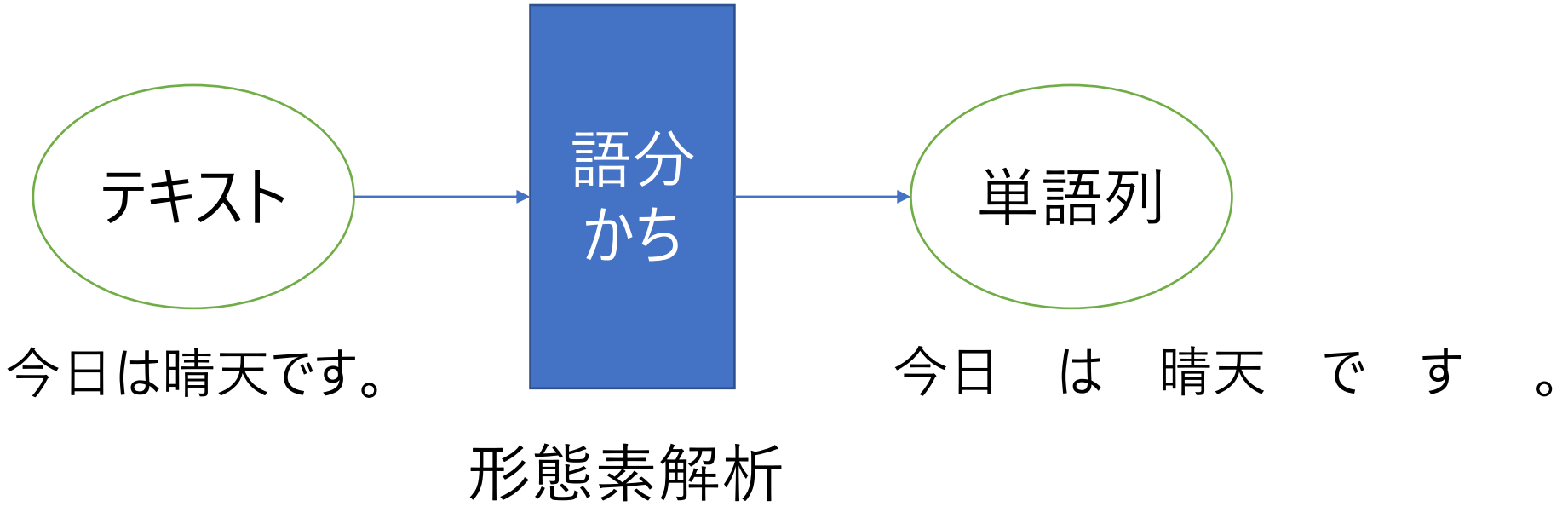


人手で設計
（伝統的な自然言語処理の
階層は特徴の一種）



特徴も自動獲得

形態素解析の結果、単語の区切りが得られる



語分ちは、End-to-end のニューラルネットでも利用する

入力は分割された単語列

今日　は　晴天　で　す　。

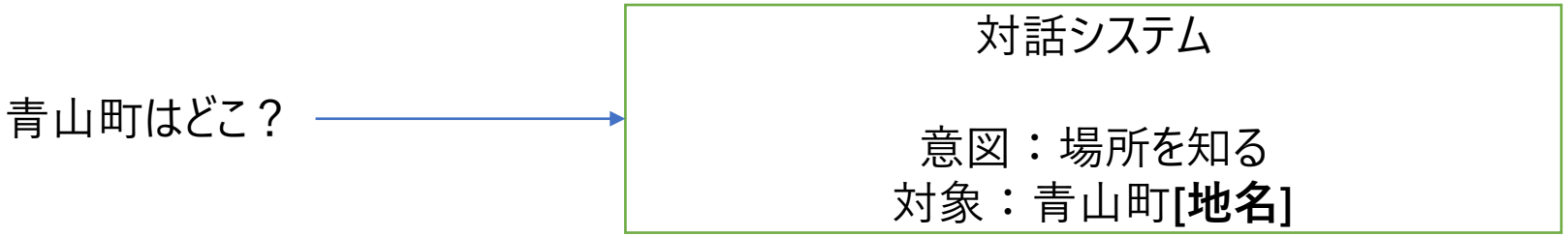
End-to-End ニューラルネットワーク

応用の出力

(分類、応答文、訳文,...)

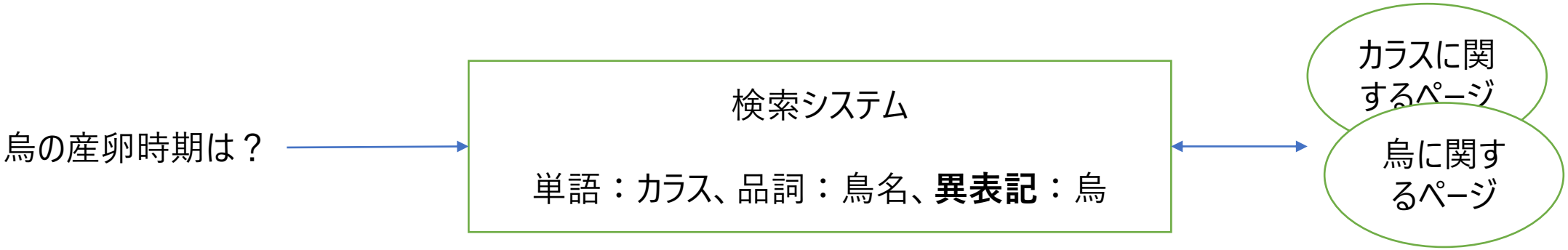
形態素解析のほかの用途

エンティティの抽出



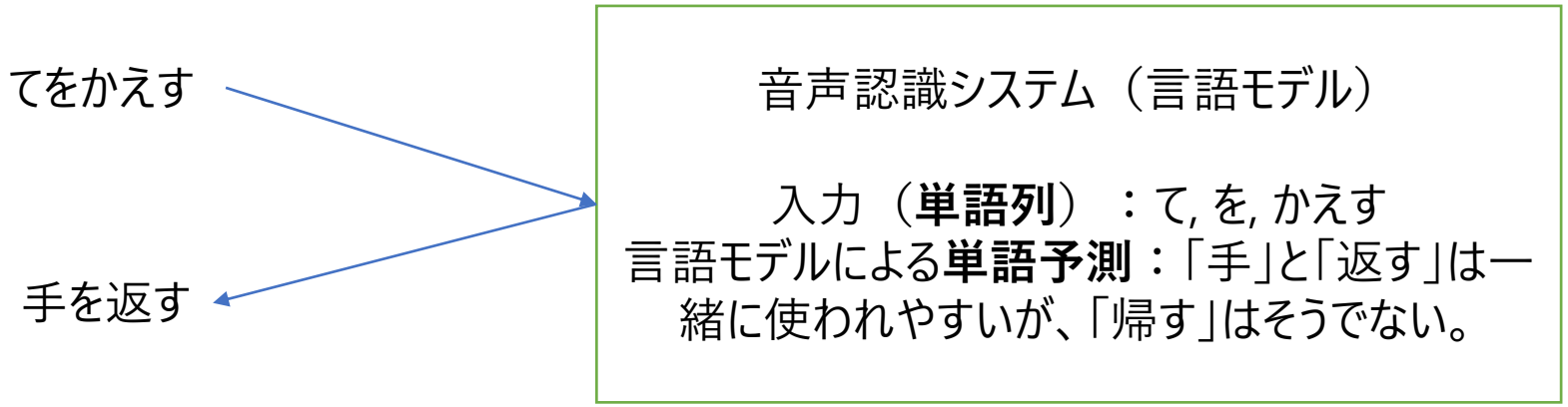
形態素解析の結果、品詞という構文的・意味的な情報を得られる。

異表記の認識



形態素解析の結果、辞書に異表記情報があれば、異表記単語の処理結果を統合できる。

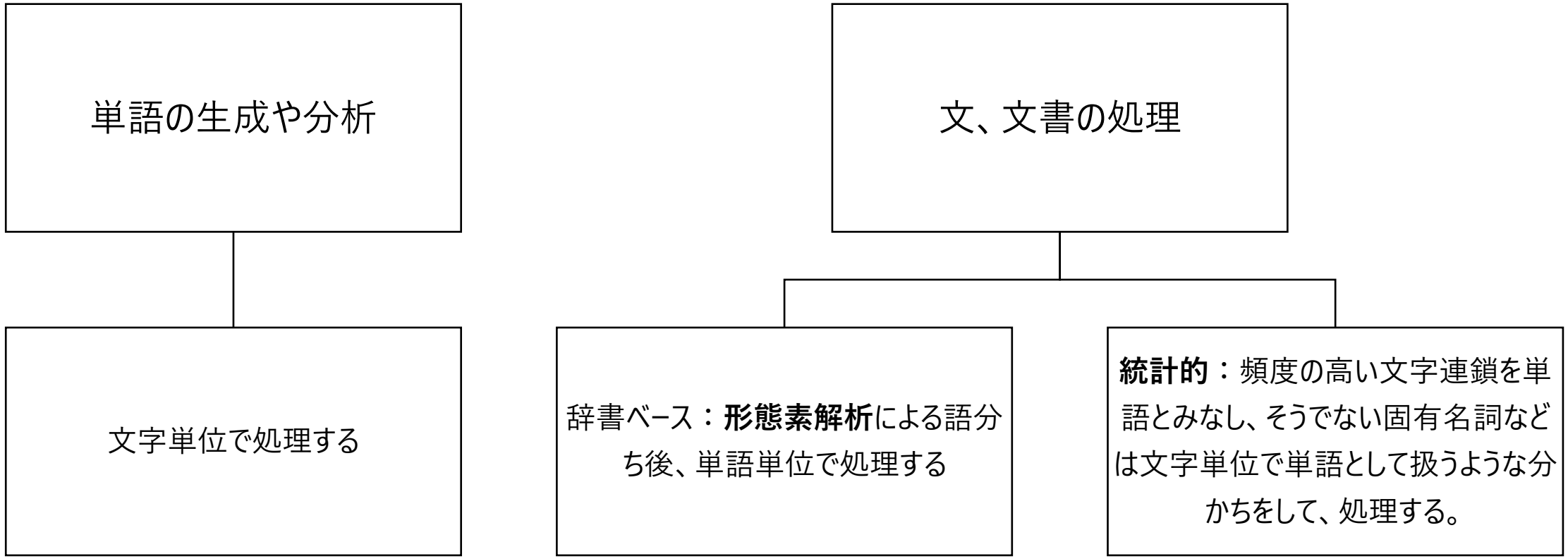
言語モデルによる単語予測



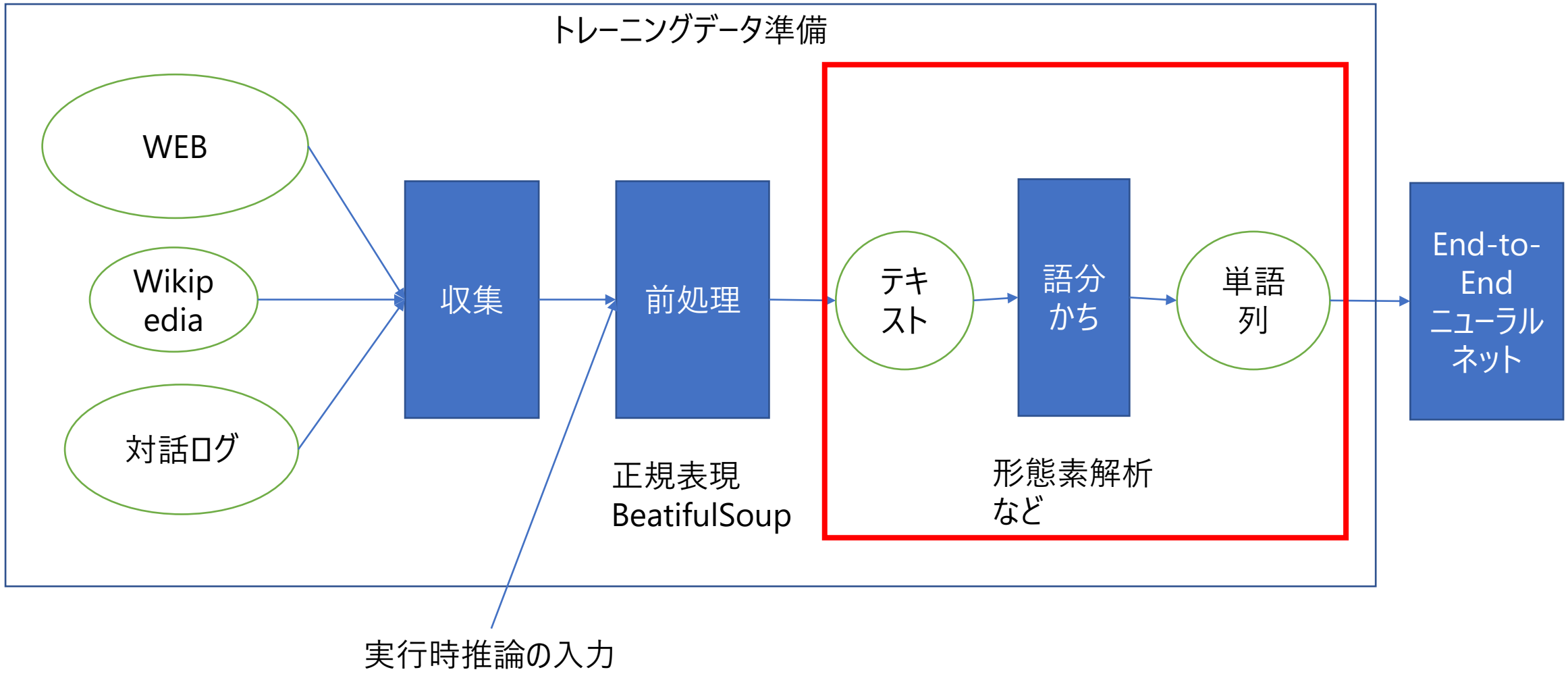
分かち書きをした結果、単語列をニューラルネットに覚えさせれば、コンテキストから単語予測ができるようになる。

多様な処理単位

形態素解析は、必須ではなく、応用による



通常のニューラルネット開発プロセス内での形態素解析の位置づけ



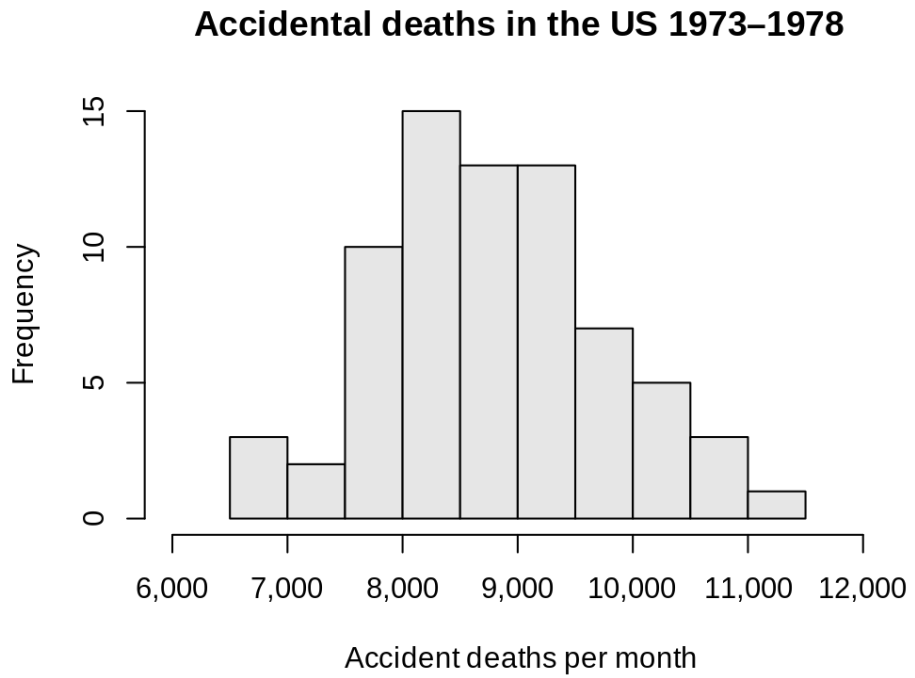
100本ノック第4章課題30～39

- 「100本ノック」の4章の課題を解いてみましょう。
- 第4章の課題は、形態素解析エンジンがすでにあってそれが利用できるという前提で、その出力加工を Python でどう書くかという課題です。
- 「NLP準備、形態素解析.ipynb」というノートをコピーし、冒頭の準備、基本事項をやった後、各課題のセクション下のコードセルに解答コードを完成させ、実行ログを残してください。
- データを扱う際、pyplotというグラフ描画パッケージを重宝します。ついでに基本的な使い方をマスターします。
- 以下に、課題を解く際に参考となることを説明します。

課題を解くための参考情報

ヒストグラム

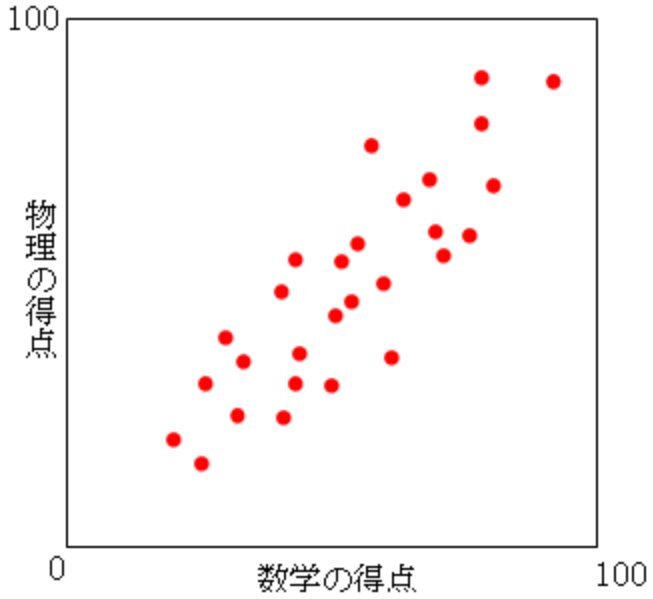
- 縦軸に度数、横軸に階級をとった統計グラフ



https://ja.wikipedia.org/wiki/%E3%83%92%E3%82%B9%E3%83%88%E3%82%B0%E3%83%A9%E3%83%A0

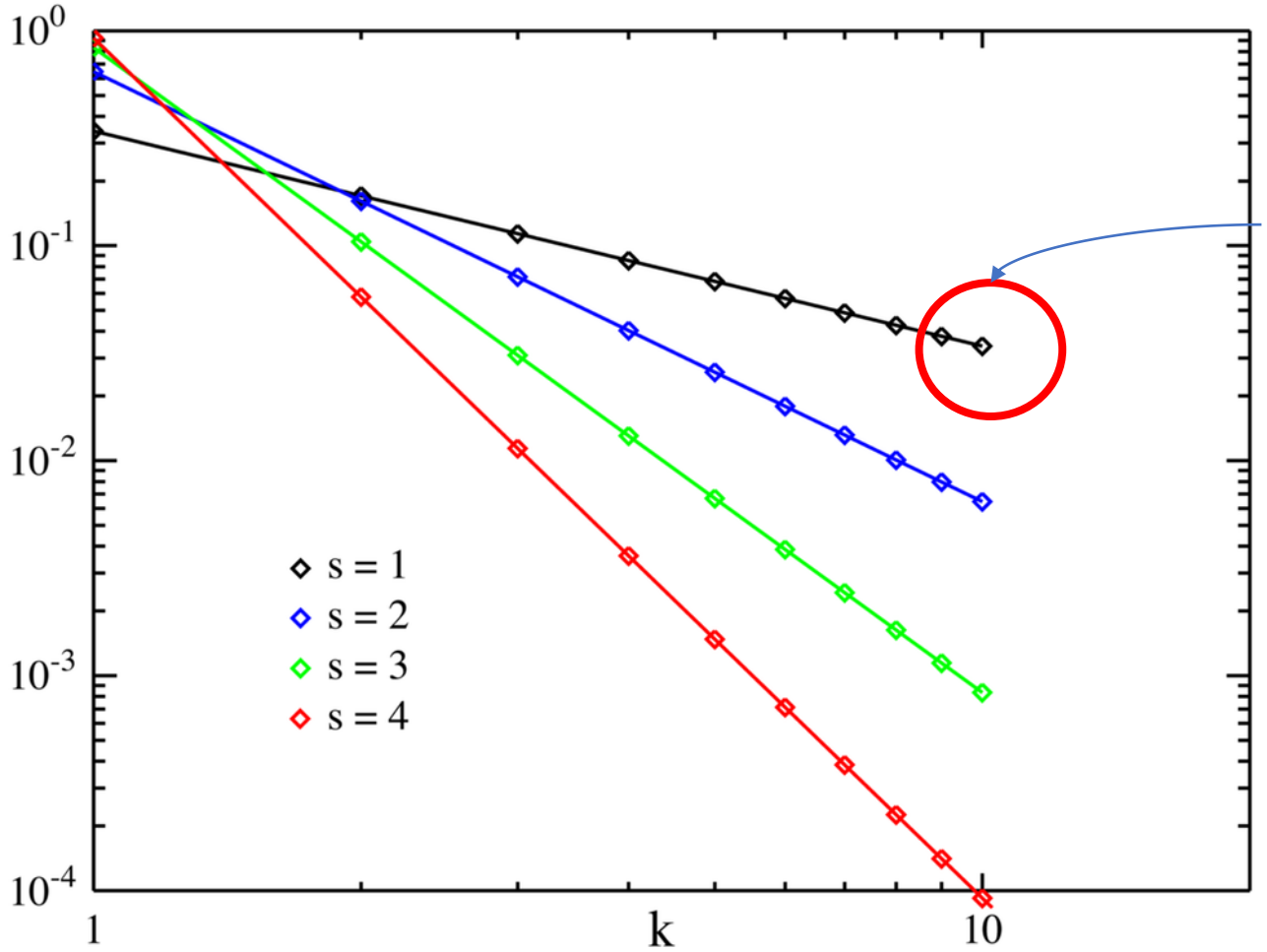
散布図

- 二つの特性を横軸と縦軸とし，観測値を打点して作るグラフ表示



https://ja.wikipedia.org/wiki/%E6%95%A3%E5%B8%83%E5%9B%B3

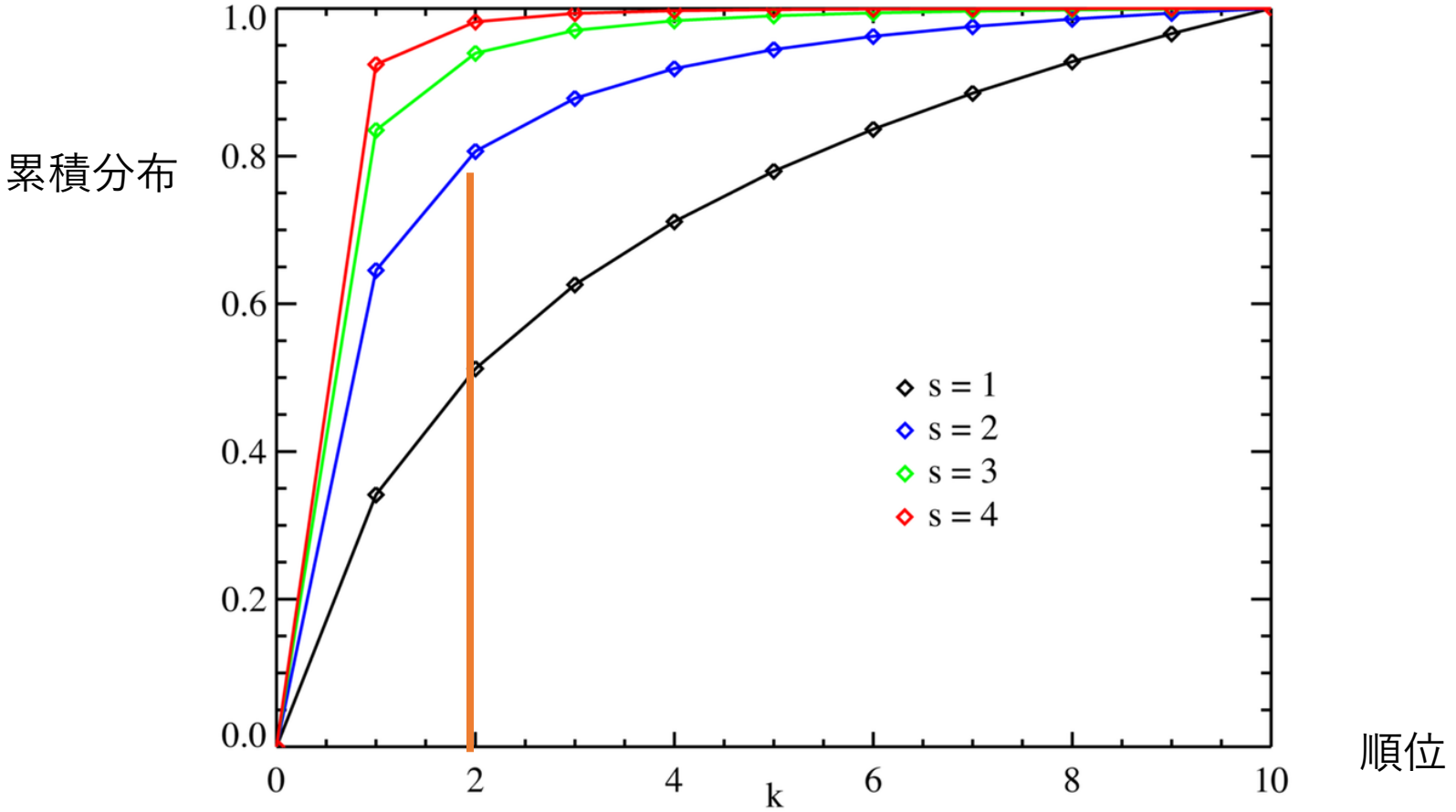
Zipf(ジップ)の法則：。言語や人文科学の分野で広範囲にみられる現象です。



10番目の項目の頻度は、1番目の項目の頻度の1/10。

https://ja.wikipedia.org/wiki/%E3%82%B8%E3%83%83%E3%83%97%E3%81%AE%E6%B3%95%E5%89%87

Zipfの法則≒パレート(80/20)の法則：上位20%の原因が、80%の問題を引き起こす。上位20%の営業員が80%の利益を稼ぐ



https://ja.wikipedia.org/wiki/%E3%82%B8%E3%83%83%E3%83%97%E3%81%AE%E6%B3%95%E5%89%87
https://effectiviology.com/80-20-rule-pareto-principle/#Related_concept_Zipf%E2%80%99s_law

確認クイズ

- 形態素解析の確認クイズをやってください。