2023.03.25 14:00-14:30 オンライン

『BTSJ日本語自然会話コーパス(トランスクリプト・音声) 2022年3月NCRB連動版』の形態素解析について

山崎　誠(やまざき まこと)(国立国語研究所)

『BTSJ1000人日本語自然会話コーパス』と『自然会話リソースバンク（NCRB）』の新展開（―その多様な活用方法―）

目次

1. コーパスの規模

1. コーパスの規模

- 『BTSJ日本語自然会話コーパス(トランスクリプト・音声)2022年3月NCRB連動版』(以下、「BTSJ2023」と略す)

- コーパスに同梱されているファイル「1. はじめにお読みください.docx」に以下の記述がある。

- このフォルダには, 本コーパスに収録されている514会話(約127時間10分, 延べ人数1028人)のトランスクリプトが収められています。

1. コーパスの規模

- 話し言葉のコーパスは「時間数」が一つの目安。

- 『名大会話コーパス』:129会話、合計約100時間
- https://mmsrv.ninjal.ac.jp/nucc/

- 『日本語日常会話コーパス』:200時間、577会話、約240万語
- https://www2.ninjal.ac.jp/conversation/cejc/design.html

1. コーパスの規模

- 書き言葉のコーパスでは「時間」がないので、文字数や語数が規模の目安となる場合が多い。
- 話し言葉のコーパスでも、語数をひとつの属性として取り上げると、分析の幅が広がる。
- 文節や文も共通の量的な指標となりうるが、語よりも認定に難しさがある。

1. コーパスの規模

- BTSJ2023の語数はどのくらいか？

- データの語数を把握するには、現在、形態素解析という技術を用いるのが一般的である。

- 以下、16枚目のスライドまで山崎・宇佐美（2019）をアレンジして使用しています。

2. 形態素解析

2.1　形態素解析とは

- 「形態素は(morpheme)は, 言語の意味や文法機能を担う最小の単位と定義される. 日本語の形態素解析(morphological analysis)とは, 与えられた文を形態素の単位に分割し, その文法機能(一般には品詞および活用情報)を同定する処理を言う. 日本語の形態素解析は次の3つを行う処理と考えるのが普通である.

  (1)形態素の同定(tokenization)

  (2)活用形の処理(stemming, lemmatization)

  (3)品詞タグ付け(POS(Part Of Speech) tagging)」

- 松本裕治(2009: 138)「2.2　品詞と形態素解析」『言語処理学事典』共立出版, 2009刊　より

2.1 形態素解析とは

- Morphological analysis is the analysis of morphology in various fields:
  - Analysis of morphology (linguistics), the internal structure of words. Morphological parsing is conducted by computers to extract morphological information from a given wordform.
- https://en.wikipedia.org/wiki/Morphological_analysis

2.1 形態素解析とは

- Morphilogical parsing

- Morphological parsing, in natural language processing, is the process of deter mining the morphemes from which a given word is constructed. It must be able to distinguish between orthographic rules and morphological rules.
- For example, the word 'foxes' can be decomposed into 'fox' (the stem), and 'es' (a suffix indicating plurality).

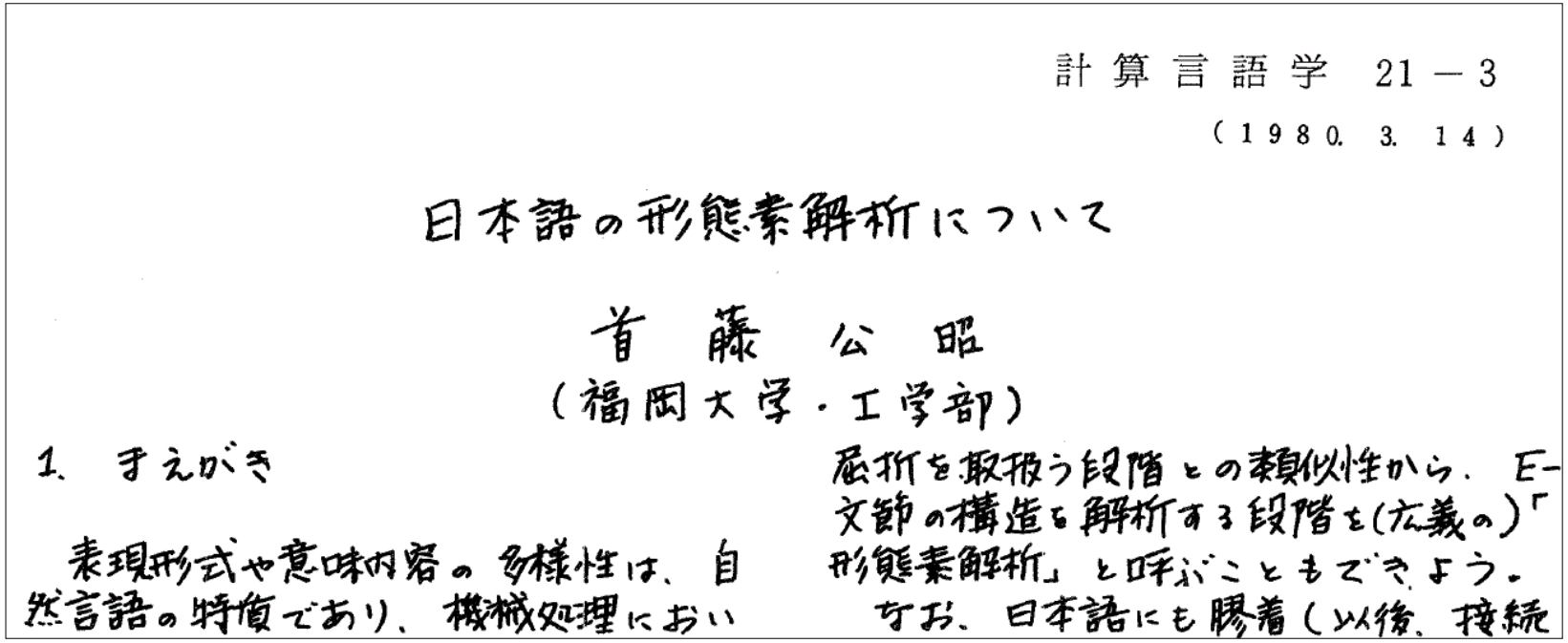
- https://en.wikipedia.org/wiki/Morphological_parsing

2.1 形態素解析とは

- 「形態素」解析は、言語学的には誤解を招く使い方である。
- 実際にどのような言語単位に分割されるかは、形態素解析器が利用する「辞書」に収録されている語彙項目（＝語）に依存する。
- 現在、一般的に使われている形態素解析用の辞書は、ipadic、UniDic、NEOlogd、Sudachi（の辞書）など。

2.1 形態素解析とは

- 「形態素解析」という用語は、1980年頃から自然言語処理の分野で使われはじめたようである。



計 算 言 語 学 21－3

（1980. 3. 14）

日本語の形態素解析について

首藤 公昭

（福岡大学・工学部）

1. まえがき

表現形式や意味内容の多様性は、自然言語の特徴であり、機械処理における解析を取扱う段階との類似性から、E-文節の構造を解析する段階を(広義の)「形態素解析」と呼ぶこともできよう。

なお、日本語にも膠着(以後、接続

- 首藤公昭「日本語の形態素解析について」「情報処理学会研究報告自然言語処理(NL)」1979(46(1979-NL-021)), 1-6, 1980-03-14

2.1 形態素解析とは

- 初期のころは、「形態素分析」という言い方もあった。

- 「本論文では, 言語処理の第一段階である形態素分析の日本語文に対する適用について考察した. 」(長尾他1978: 520)
- 長尾真, 辻井潤一, 山上明, 建部周二(1978)「国語辞書の記憶と日本語文の自動分割」「情報処理」 19(6), pp.514-521, 1978-06-15

- 「日本語の形態素分析」
- 中野洋, 野村雅昭(1979)「日本語情報処理:日本語の形態素分析」「情報処理」 20(10), pp.857-864, 1979-10-15

2.2 形態論情報

- 語彙素読み　:コオリ　ハシル　アジサイ　ビーナス
- 語彙素　:氷　走る　紫陽花　ビーナス-Venus
- 書字形出現形　:氷　走れ　あぢさゐ　ヴィーナス
- 書字形(基本形):氷　走る　あぢさゐ　ヴィーナス
- 発音形出現形　:コーリ　ハシレ　アジサイ　ビーナス
- 発音形(基本形):コーリ　ハシル　アジサイ　ビーナス
- 仮名形出現形　:コオリ　ハシレ　アヂサヰ　ヴィーナス
- 仮名形(基本形):コオリ　ハシル　アヂサヰ　ヴィーナス
- 語形出現形　:コオリ　ハシレ　アジサイ　ビーナス
- 語形(基本形)　:コオリ　ハシル　アジサイ　ビーナス

2.2 形態論情報

- 「表記形(word-form)(表記語[graphic word]とも言う)とは, テクストの中で実際に表記される形のことを言う。表記形は綴りに着目した語の単位であるため, 綴りが異なれば別語となり, たとえば, study, studies, studied, studyingは4語と数えられる。」

- 「語彙素(lexeme)とは, 各種の活用形に共通する基底形(base form)のことで, テクストの中でそれぞれの表記形を生み出す「語の素」となる。表記形は実際に書いたり, 読んだりすることができるが, 語彙素は語のプロトタイプであり, 抽象的な概念である。」

- 石川慎一郎(2008: 77-78)『英語コーパスと言語教育』

2.2 形態論情報

- 「レマ(lemma)とは，活用形や綴り字の違いを問わず，語幹と語類を同じする各種の表記形を包含する基準形(canonical form)のことである。レマもまた実際に書いたり読んだりできない抽象的概念であるため，コーパス研究では，通常の表記形と区別するために，全大文字で記載するのが通例である。また，レマが辞書の見出し単位になりやすいことから，ほぼ同じ意味で見出し語(headword)という用語が用いられることもある。」

- 石川慎一郎(2008: 77-78)『英語コーパスと言語教育』

2.3　形態素解析の問題点

- 形態素解析の問題点
- 解析精度が低い場合がある。とくに話し言葉において。

  - [001,60,JM001] ちが、ちが、ちが。
  - (会話番号,ライン番号,話者] 発話内容)

  - を MeCab 0.996 + UniDic-csj-2.2.0(現代話し言葉用)で解析すると,「ち」(記号)+「が」(格助詞)+読点　となる。

  - 語断片、言いよどみなどは前もってタグ付けをしておく必要がある。
  - 臼田他(2015: 182)も参照

2.3 形態素解析の問題点

- 「また, 転記テキストを対象に形態論情報が付与されるが, 全体の約9割が自動解析の範囲で行なわれる。そのためフィラーや語断片などは, テキストの可読性だけでなく, 自動形態素解析の精度を落とす要因ともなる。そこでこのようなフィラーや語断片などに対し, 以下の例に示すように, (F) や (D) といったタグを付与することにした。」

- 「(F えー) この点につきましては後程 (F えー)(D わ) 今後 (D ちゃん) ちゃんと (D ふや) 被験者を増やしていこうと」

- 国立国語研究所(2006: 27)

3. BTSJ2023の形態素解析

3.1 準備

- 形態素解析を行うための準備
  - プログラムのインストール：Python（パイソン）がおすすめ。さまざまなライブラリも充実。解説本やネットの情報も多い。

- 注意点
  - フリーのアプリケーションは、突然仕様が変更になって、うまく動作しなくなったり、開発が止まって、最新の機械環境で使えなくなったりする。
  - ネットの情報は自分の機械環境に合ったものでないとうまくいかないことがある。

3.2 ファイル一覧の取得

- ファイルの抽出
- BTSJ2023のファイルは、発話を書き起こしたエクセルファイルと、その音声ファイルがひとつのフォルダに入っているものが多い。

- 形態素解析のためには、エクセルファイルだけあればよいので、エクセルファイルだけを取り出す必要がある。

3.2 ファイル一覧の取得

- files=glob.glob(データのあるフォルダ+’ /**’ , recursive=True)

- これでサブフォルダも含めてすべてのファイルの一覧が取得できるので、その中から拡張子が xlsx のものだけを抜き出せばよい。

- ただし、「3. 本コーパスに収録されている会話データの情報一覧.xlsx」ちという解説のファイルも対象になってしまうので、それは最初から違う場所に移動しておくなどの措置が必要。また、発話データ以外のエクセルファイルを同じフォルダに作成しない。

- 対象ファイル数：544

3.3 エクセルの操作

- 対象となるファイルの一覧が取得できたら、それを開いて、内容を取り出す。

- Pythonでエクセルを扱えるライブラリはいくつかあるが、今回はopenpyxlを使う。

- ライブラリの仕様変更が多いので要注意。2020年のときに使っていた、xlrdは xlsxの扱いをやめてしまった。

3.3 エクセルの操作

- 対象となるファイルの一覧が取得できたら、それらのファイルを開いて、内容を取り出す。

- セルを指定してその値を取り出すことができるので、操作自体は比較的簡単。

- どの範囲を取り出すか。

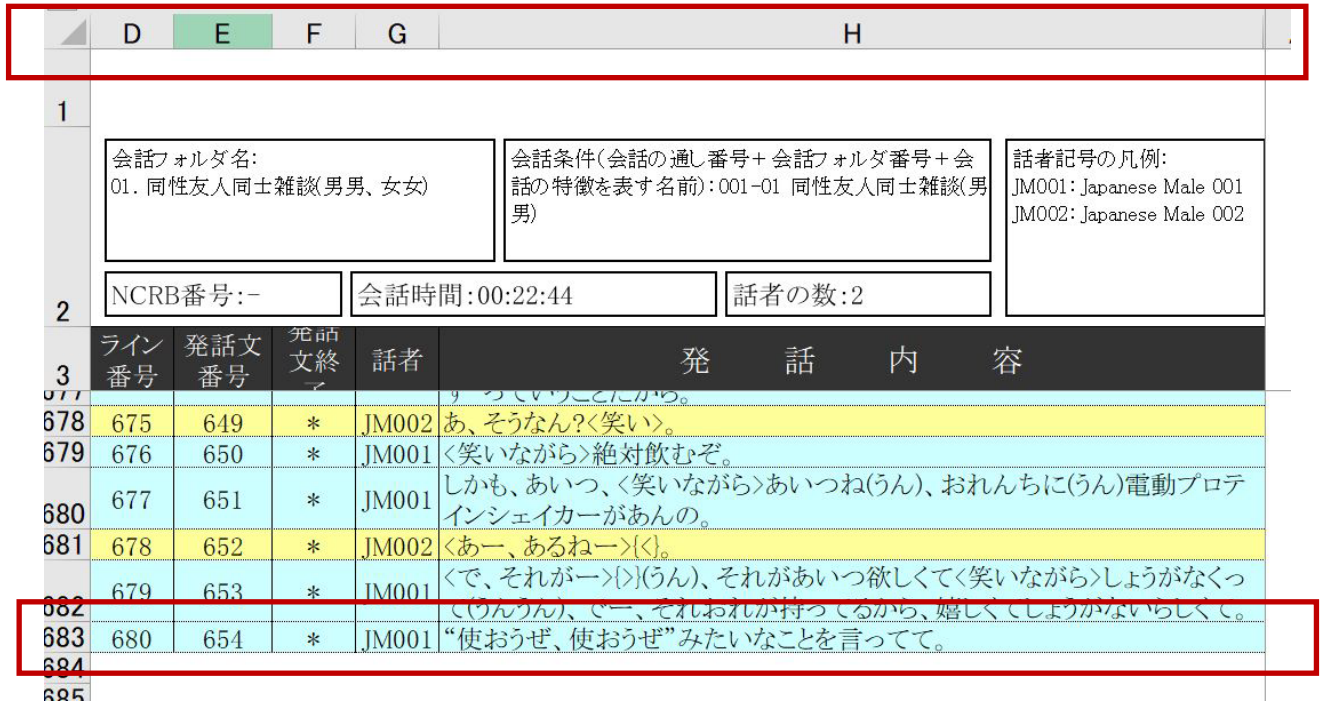
3.3 エクセルの操作

- 開始点　D4～H4（あるいはD3～H3）。発話内容はH列。

	D	E	F	G	H
1					
2	会話フォルダ名: 01. 同性友人同士雑談(男男、女女)		会話条件(会話の通し番号＋会話フォルダ番号＋会話の特徴を表す名前):001-01 同性友人同士雑談(男男)		話者記号の凡例: JM001: Japanese Male 001 JM002: Japanese Male 002
	NCRB番号:-		会話時間:00:22:44	話者の数:2	
3	ライン番号	発話文番号	発話文終了	話者	発　話　内　容
4	1	1	*	JM001	<おれ>{<}【【。
5	2	2	*	JM002	】】<じゃ>{>}、さっそく見てみようか。
6	3	3	*	JM001	ちがう<笑いながら><2人で笑い>。
7	4	4	*	JM001	まだよくない?<笑いながら>。
8	5	5	*	JM002	<早い?>{<}<笑いながら>。
9	6	6	*	JM001	<ちょ>{>}、まだいいよ<笑いながら>。

3.3 エクセルの操作

- 終了点　583行目(D583～H583)



	D	E	F	G	H
1					
2	会話フォルダ名: 01. 同性友人同士雑談(男男、女女)		会話条件(会話の通し番号＋会話フォルダ番号＋会話の特徴を表す名前):001-01 同性友人同士雑談(男男)		話者記号の凡例: JM001: Japanese Male 001 JM002: Japanese Male 002
	NCRB番号:-		会話時間:00:22:44		話者の数:2
3	ライン番号	発話文番号	発話文終了	話者	発話内容
678	675	649	*	JM002	あ、そうなん?<笑い>。
679	676	650	*	JM001	<笑いながら>絶対飲むぞ。
680	677	651	*	JM001	しかも、あいつ、<笑いながら>あいつね(うん)、おれんちに(うん)電動プロテインシェイカーがあんの。
681	678	652	*	JM002	<あー、あるねー>{<}。
682	679	653	*	JM001	<で、それがー>{>}(うん)、それがあいつ欲しくて<笑いながら>しょうがなくって(うんうん)、で一、それおれが持ってるから、嬉しくてしょうがないらしくて。
683	680	654	*	JM001	“使おうぜ、使おうぜ”みたいなことを言ってて。
684					
685					

max_rows（エクセルの最終行を得る関数）を使うときは要注意。

この例では、max_rowsの値は995になっている。

データがなくても、書式が設定されているとそのセルもデータがあると見なされる。

https://www.relief.jp/docs/openpyxl-worksheet-max-rom-column.html

3.3 エクセルの操作

- 終了点　1行ずつ読み込んでいって、データのない行が出てきたら、その1つ前で終了。

- データがない＝セルの値がない。
- 書き出したファイル上は、Noneと示されるものと、空文字(長さ0の文字列)がある。それら以外であればデータがあると見なす。

- 取り出した行数:193588行(見出し行を含む)、19304行(見出し行を含まない場合)

3.4 解析

- ライブラリ mecab-python3を使う。

- import MeCab
- m = MeCab.Tagger(”-d c:/unidic-cwj-3.1.0”)

使用する形態素解析辞書のフォルダ

- result=m.parse(解析対象文字列)

3.4 解析

- 「日本語の本」の解析結果

- 日本,28455,日本,ニッポン,名詞,名詞-固有名詞-地名-国,固
- 語,13334,語,ゴ,名詞,名詞-普通名詞-一般,漢
- の,28989,の,ノ,助詞,助詞-格助詞,和
- 本,34867,本,ホン,名詞,名詞-普通名詞-一般,漢

- 各形態論情報の順序や種類はカスタマイズ可能。

3.5 結果

- 単純集計(発話内容に何も手を加えていないもの)

- 全体
  - 延べ語数　:2,799,231語
  - 異なり語数:　　18,761語

- 空白・補助記号等を除く場合
  - 延べ語数　:1,701,555語
  - 異なり語数:　　18,691語

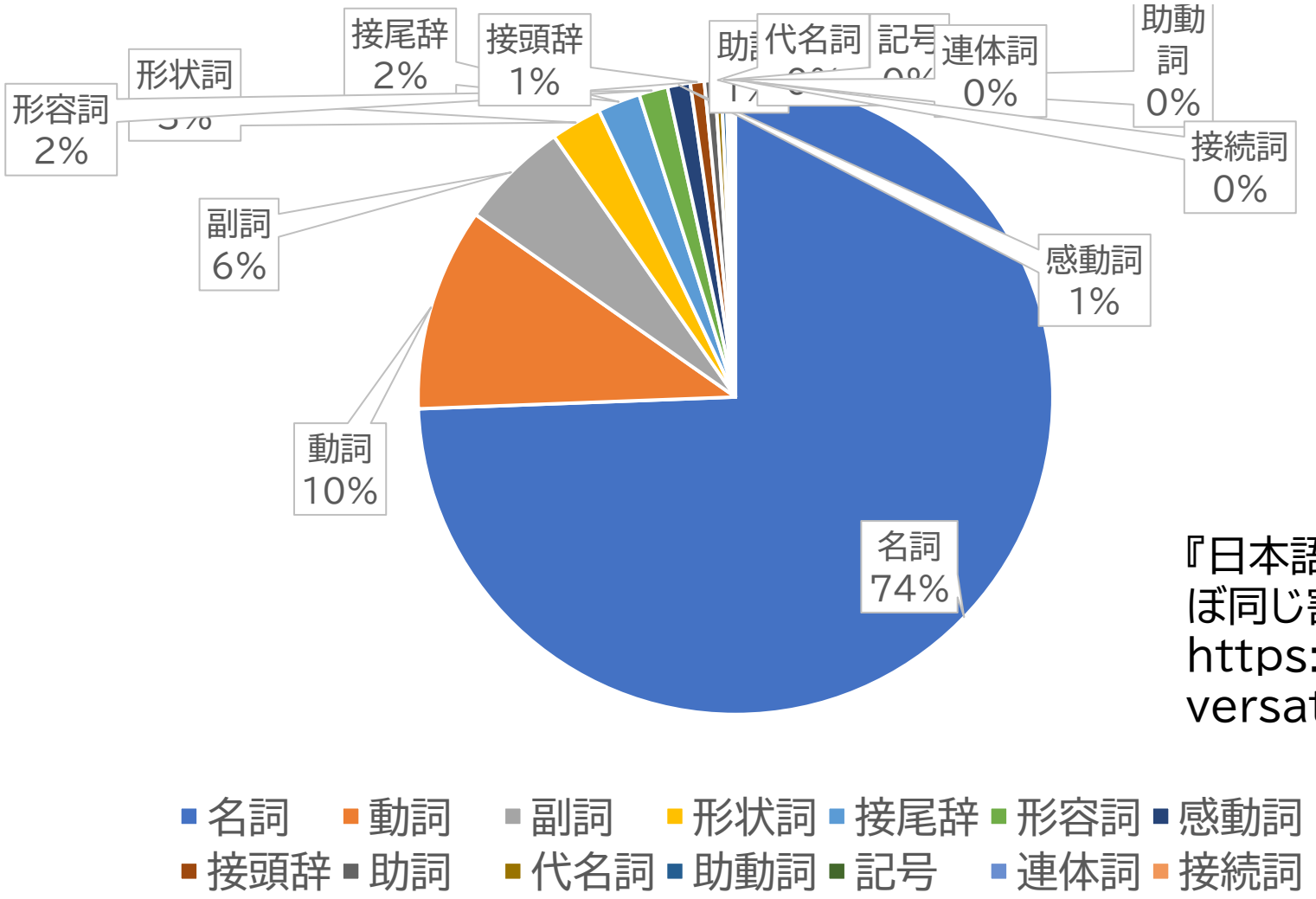
3.5 結果

- 上位語

語彙素ID	語彙素	読み	品詞	品詞2	語種	頻度
22916	だ	ダ	助動詞	助動詞	和	75592
3568	うん	ウン	感動詞	感動詞-一般	和	42972
24874	て	テ	助詞	助詞-接続助詞	和	37307
28754	ね	ネ	助詞	助詞-終助詞	和	37022
28990	の	ノ	助詞	助詞-準体助詞	和	35918
25653	です	デス	助動詞	助動詞	和	34266
21642	た	タ	助動詞	助動詞	和	33753
5568	か	カ	助詞	助詞-副助詞	和	32927
25826	と	ト	助詞	助詞-格助詞	和	31413
25518	で	デ	助詞	助詞-格助詞	和	30693

『日本語日常会話コーパス』の場合と似ている
https://www2.ninjal.ac.jp/conversation/report/report04.pdf

3.5 結果:品詞の割合(異なり)



『日本語日常会話コーパス』の場合とほぼ同じ割合
https://www2.ninjal.ac.jp/conversation/report/report04.pdf

3.5　結果:今後の展開

- 今回は準備不足で間に合わなかったが、会話の属性、話者の属性と結びつけると、既存のコーパス言語学の手法を用いて、さらに多角的な分析ができるようになるだろう。

4. 課題

4. 課題

- 解析の前に行っておくべきこと

- Excelファイルから情報の抽出の際
- 発話内容の半角文字を全角文字に変更
- 記号の除去
  - [001,70,JM001] <らしくない>{>}=。
  - [会話番号,ライン番号,話者] 発話内容
- パラ言語情報の除去
  - [001,57.JM002] うん<笑い>。
  - 発話の重複の場合との記述法の違いを要確認。

4. 課題

- 伏せ字の登録
  - [001,74,JM001] ＝あれ、「大学名1」。
  - このままだと「大学」「名」「1」に分割される。もし，他の会話に「大学名1」が出てきた場合，同一の語とするかどうかの判断も必要（元の会話に戻って確認するという主張ではない）。

- 相づちの分離
  - [001,49,JM001] ヨーグルト買って（うん），，
  - 「うん」は会話の相手（JM002）の発話。

4. 課題

- 伏せ字の登録
  - [001,74,JM001] ＝あれ、「大学名1」。
  - このままだと「大学」「名」「1」に分割される。もし，他の会話に「大学名1」が出てきた場合，同一の語とするかどうかの判断も必要（元の会話に戻って確認するという主張ではない）。

- 相づちの分離
  - [001,49,JM001] ヨーグルト買って（うん），，
  - 「うん」は会話の相手（JM002）の発話。

おわりに

- 量的な観察は常に質的な観察とともに行うべきであり、単に数えてみましたというのは、ブレーンストーミングとしてはあり得ても、研究の段階では意味を持たない。素朴な疑問を研究の文脈に乗せる必要がある。

- また、単純な数だけを見ていると思わぬ誤判断を招く可能性がある。以前、集計の中で、女性の話者が人称代名詞の「俺」を使う例がわずかに確認されたが、データを見ると、いずれも他者(男性)の発話の引用だったことがあり、それを知らなければ謝った判断をしていたかもしれない(宇佐美・山崎2018)。

参考文献

- 宇佐美まゆみ, 山崎誠(2018)『BTSJ日本語自然会話コーパス2018年版』における一人称・二人称代名詞の使用実態, 日本語学会2018年度秋季大会, 2018.10.14
- 臼田泰如, 川端良子, 西川賢哉, 石本祐一, 小磯花絵(2015)「『日本語日常会話コーパス』における転記の基準と作成手法」『国立国語研究所論集』15, pp.177-193
- 国立国語研究所(2006)『日本語話し言葉コーパスの構築法』 https://pj.ninjal.ac.jp/corpus_center/csj/k-report-f/CSJ_rep.pdf
- 首藤公昭「日本語の形態素解析について」「情報処理学会研究報告自然言語処理(NL)」 1979(46(1979-NL-021)), 1-6, 1980-03-14
- 長尾真, 辻井潤一, 山上明, 建部周二(1978)「国語辞書の記憶と日本語文の自動分割」「情報処理」 19(6), pp.514-521, 1978-06-15
- 中野洋, 野村雅昭(1979)「日本語情報処理:日本語の形態素分析」「情報処理」 20(10), pp.857-864, 1979-10-15
- 松本裕治(2009)「2.2　品詞と形態素解析」『言語処理学事典』共立出版
- 山崎誠, 宇佐美まゆみ(2019)「『BTSJ自然会話コーパス』の形態素解析のための補助ツールの開発について」, 第1回語用論コーパス科研成果発表会, 2019.06.29

- ご清聴ありがとうございました。

次の発表

「JFL環境における日本語学習者の「日本語らしさ」
ーインドの授業のデータ分析よりー」

大塚 容子（岐阜聖徳学園大学）・重光由加（東京工芸大学）