2021/08/26

第12回コーパス利用講習会

オンライン検索システム「中納言」講習会

対象：
『日本語日常会話コーパス モニター公開版』

国立国語研究所
言語変化研究領域　　山崎　誠
音声言語研究領域　　柏野　和佳子

今日の講習内容

10:00-10:10	導入	山崎
10:10-10:25	『日本語日常会話コーパス モニター公開版』概説	山崎
10:30-11:30	『中納言』での活用	柏野
11:40-12:00	質疑応答	山崎・柏野

（以下、資料の紹介。時間があれば簡単な説明）

オープンハウスで公開したポスターによる概説	柏野
『ひまわり』用 CEJCの紹介	山崎

今日の目標

* Webインターフェース『中納言』を使って『日本語日常会話コーパス』（以降, CEJC）の形態論情報を検索できるようになることが目標
* そのために必要なCEJCに関する基礎知識を知る
  * とりわけ形態論情報（短単位）や頻度の知識
* ダウンロードした検索結果のExcelによる処理

今日の講習についての注意

* 『中納言』で検索できるのは形態論情報だけ
  * 性別・年齢などの話者情報は，『中納言』で直接検索することはできないが，ダウンロード後に分析することは可能。
  * 統語情報，複合語の内部構造などは，CEJCには付与されていないため，そのような検索は出来ない。
  * CEJCは話者の性別，年齢層のバランスをとって設計しているが，多少の偏りはある。

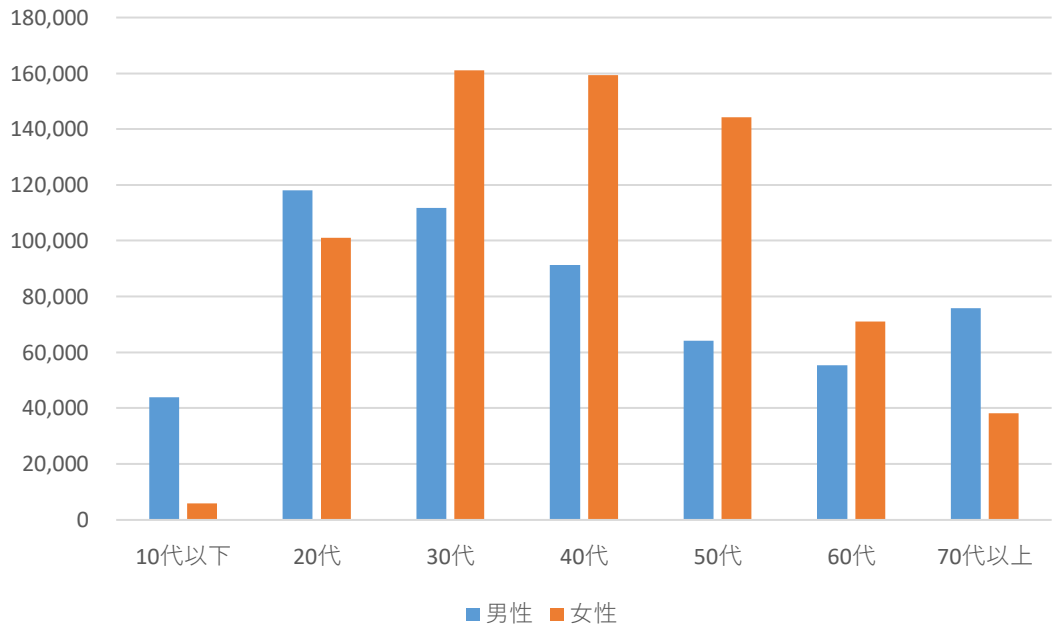

図　性別・年代別の収録語数

コーパスにおける
形態論情報の意義

* 日本語形態素自動解析技術は1990年頃から実用の域に達しはじめ，Juman, ChaSen, MeCabなど優秀な解析器が無償公開されている。
* しかし解析用辞書における「語」のあつかいが問題。

Yahoo!	IPA	JUMAN	UniDic
国立国会図書館	国立	国立	国立
	国会図書館	国会	国会
		図書	図書
		館	館
国立	国立	国立	国立
公文書館	公文書	公文書	公
			文書
			館
国立			国立
歴史	歴史	歴史	歴史
民俗	民俗	民俗	民俗
博物館	博物館	博物	博物
		館	館

CEJC等の短単位用解析辞書であるUniDicは，言語学的には首尾一貫しているが，やや短めに分割される。

短単位と長単位による二重解析

短単位	短単位品詞	長単位	長単位品詞
公害	名詞-普通名詞-一般		名詞-普通名詞-一般
紛争			
処理			
法			
に	助詞-格助詞	における	助詞-格助詞
おけ	動詞-一般		
る	助動詞		
公害	名詞-普通名詞-一般	公害紛争処理	名詞-普通名詞-一般
紛争	名詞-普通名詞-サ変可能		
処理	名詞-普通名詞-サ変可能		
の	助詞-格助詞	の	助詞-格助詞
手続	名詞-普通名詞-サ変可能	手続	名詞-普通名詞-一般
は	助詞-係助詞	は	助詞-係助詞

CEJCは，現時点では
『中納言』で検索できるのは短単位のみ。