# 全文検索システム『ひまわり』における言語分析支援機能の拡張

山口昌也（国立国語研究所）

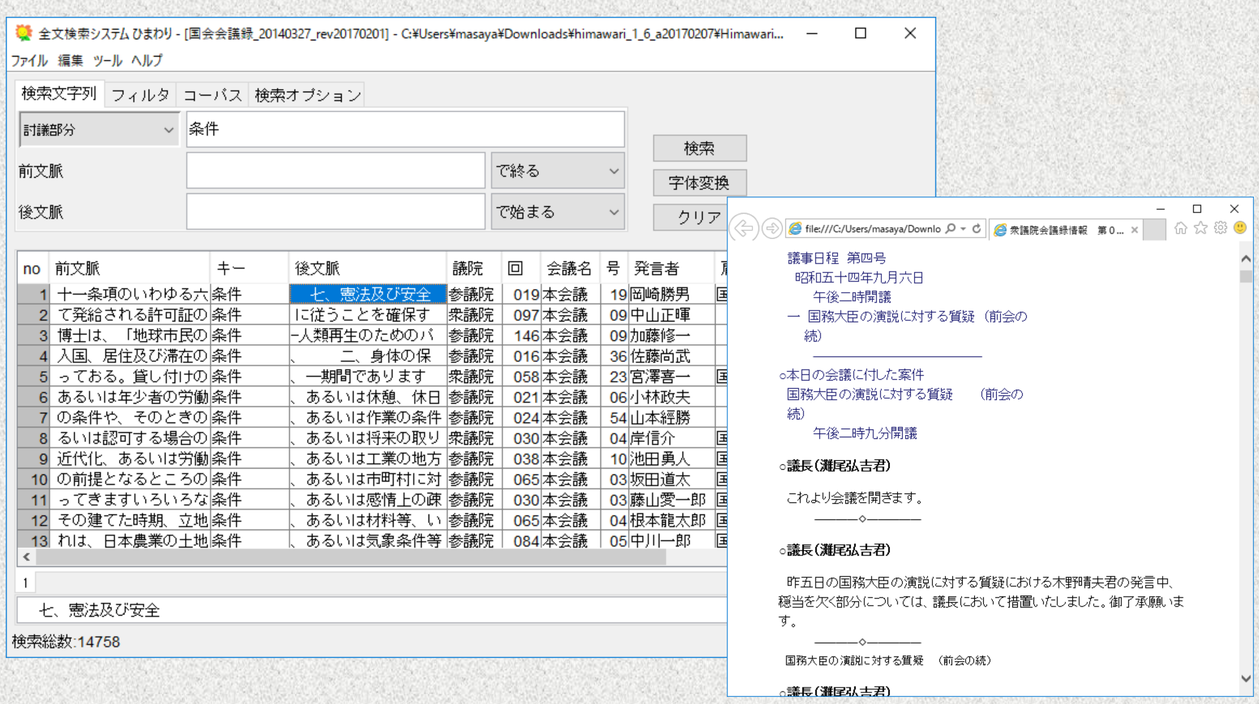
## 背景

- 従来の『ひまわり』の主な言語分析支援機能
  - XML文書に対するコンコーダンサ
  - 言語資料に適した表示

利用者自身が用例を「目で見る」タイプの分析方法

- 問題点
  - コーパスの巨大化への対応
    - 大量の用例（表示しきれない場合も）
    - 統計的分析に必要な情報の取得が困難（例：総文字数）
    - 外部アノテーションDB（形態素解析結果など）の肥大化

検索結果・コーパスの内容を集約する機能の必要性




## 拡張機能の設計

- 基本的な方針
  - 統計的分析の支援ができるようにする（≠ 統計分析機能）
    - アノテーション結果の集計
    - 外部アノテーション機能の改善
  - 検索結果を集約できるようにする
    - 検索中の集約
    - 一覧作成機能の改善

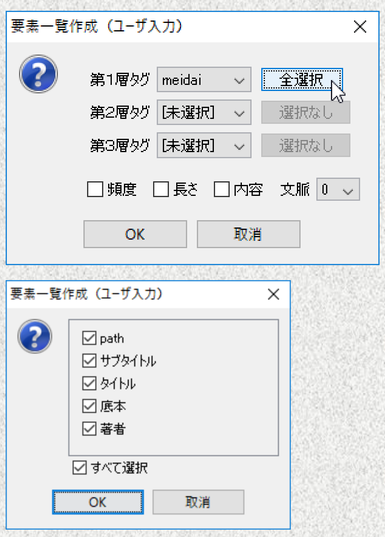
- 既存システムの機能分類と『ひまわり』の位置づけ
  - コンコーダンサ（少納言など）
  - 検索結果の集約機能（AntConc, NINJAL-LWPなど）
  - 統計分析機能（KHCoder など）

- XML文書に対するコンコーダンサ
- 検索結果の集計機能（検索結果をすばやく目で概観する）
- 統計分析支援機能（外部プログラムで統計分析しやすくする）

新しい『ひまわり』

## 拡張機能の実現

- アノテーション結果の集計
  - アノテーションした要素と属性の一覧を作成
  - 一覧作成対象の要素は，3階層まで指定可能
  - 要素の各種情報の集計・抽出
    - 要素の出現頻度
    - 要素中のテキスト長
    - 要素中のテキスト
    - 前後要素の情報




- 会話ごとの文字数（名大会話コーパス）

| meidai/@データ名 | meidai%文字数 |
|---|---|
| data001 | 15610 |
| data002 | 25820 |
| data003 | 12682 |
| data004 | 14225 |
| data005 | 22389 |
| data006 | 20725 |
| data007 | 15998 |
| data008 | 36379 |
| data009 | 34869 |
| data010 | 18410 |
| data011 | 18148 |
| data012 | 15508 |
| data013 | 14741 |

  - meidai タグ（会話）
  - 文字数オプション

- 話者ごとの発話数（名大会話コーパス）

| meidai/@データ名 | u/@話者 | 頻度 |
|---|---|---|
| data001 | M023 | 133 |
| data001 | F128 | 61 |
| data001 | F107 | 347 |
| data001 | F023 | 307 |
| data001 | × | 10 |
| data001 | unknown | 62 |
| data002 | F023 | 384 |
| data002 | F107 | 614 |
| data002 | F128 | 447 |
| data002 | unknown | 143 |
| data002 | × | 3 |
| data003 | F033 | 574 |
| data003 | F056 | 410 |

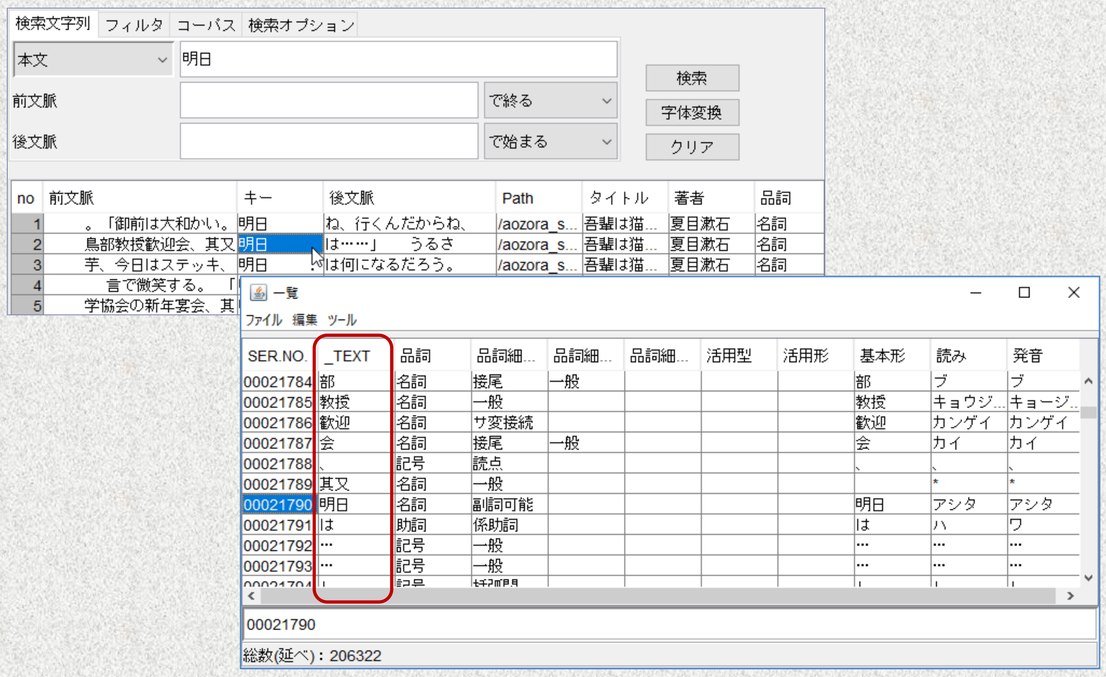
  - meidai タグ（会話）
  - u タグ（発話）
  - 頻度オプション

- bigram（名大会話コーパス）

| s/@品詞 | s/@{t} | s[1]/@品詞 | s[1]/@{t} | 頻度 |
|---|---|---|---|---|
| 助詞-終助詞 | ね | 補助記号-句点 | 。 | 13150 |
| 助詞-格助詞 | と | 助詞-副助詞 | か | 9588 |
| 助詞-終助詞 | ね | 補助記号-読点 | 、 | 9168 |
| 助詞-準体助詞 | ん | 助動詞 | だ | 8012 |
| 感動詞-一般 | うん | 補助記号-句点 | 。 | 7687 |
| 助詞-格助詞 | で | 助詞-係助詞 | も | 7186 |
| 代名詞 | なん | 助詞-副助詞 | か | 6213 |
| 助詞-接続助詞 | て | 補助記号-読点 | 、 | 6051 |
| 補助記号-一般 | ＊ | 補助記号-一般 | ＊ | 6021 |
| 助詞-終助詞 | よ | 補助記号-句点 | 。 | 5524 |
| 助動詞 | だ | 助詞-接続助詞 | から | 5227 |
| 感動詞-一般 | うん | 補助記号-読点 | 、 | 4737 |
| 助詞-副助詞 | か | 補助記号-読点 | 、 | 4716 |

  - s タグ（短単位）
  - 文脈オプション（長さ1）
  - 頻度オプション

- 外部アノテーション用データベースの改善
  - コーパスファイル(XML文書)に直接アノテーションしない方式
  - 検索結果からデータベースを閲覧可能
  - RDBから独自データベースに変更
    - サブコーパスごとに作成可能
    - データサイズの削減（約1/4）
    - 高速化（2～3倍程度）




- 検索結果の集約（国会会議録における発言者の年齢分布作成を例に）

**1** 会議の開催年と発話者年齢一覧の作成

| minutes/@開催日 | utteran... | 頻度 |
|---|---|---|
| 1947-08-12 | 1892 | 4 |
| 1947-08-12 | | 3 |
| 1947-08-12 | 1888 | 2 |
| 1947-08-12 | 1895 | 1 |
| 1947-08-13 | 1893 | 27 |
| 1947-08-13 | 1892 | 25 |
| 1947-08-13 | 1888 | 21 |
| 1947-08-13 | | 17 |
| 1947-08-13 | 1903 | 6 |
| 1947-08-13 | 1896 | 4 |
| 1947-08-13 | 1884 | 3 |
| 1947-08-13 | 1900 | 3 |
| 1947-08-13 | 1897 | 2 |

1949-10-20
総数(延べ)：429153, 異なり：25343

**2** 年月日を年に置換

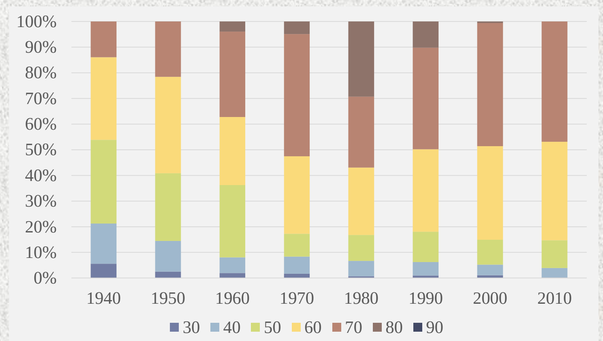
| minutes/@開催日 | utteran... | 頻度 |
|---|---|---|
| 1947 | 1892 | 4 |
| 1947 | | 3 |
| 1947 | 1888 | 2 |
| 1947 | 1895 | 1 |
| 1947 | | |
| 1947 | | |
| 1947 | | |
| 1947 | | |
| 1947 | | |

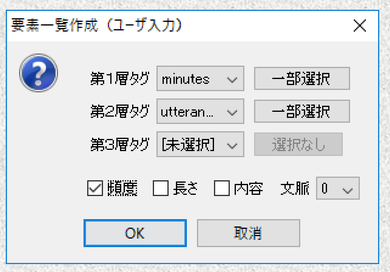
置換（正規表現）
検索する文字列 -*
置換後の文字列
OK Cancel

**3** 再集計

- 頻度を再集計
- Excel などの外部プログラムにエクスポート

ピボットテーブルでクロス集計し，グラフ表示した例




※『ひまわり』用『国会会議録』パッケージ（衆議院・予算委）を使用