音声言語コーパス目録検索システムCatalog-Search の構築および応用の検討*

沈睿†
早稲田大学人間科学研究科†
raymondshenrui@gmail.com†

菊池英明✞
早稲田大学人間科学学術院✞
kikuchi@waseda.jp✞

概要

音声言語コーパスの増加に伴い、利用者の目的に適したコーパス検索のニーズが高まっている。著者らのグループが開発した音声言語コーパスの類似性可視化システムCorpus Search(http://corpus-search.nii.ac.jp/)によって、コーパス利用者に選択の基準を提供することができたが、当該システムのアンケート評価により、類似性可視化だけでは所望のコーパスの検索が困難であるケースがあることがわかった。本稿では、コーパスの属性をフィルタリングすることによって検索を実現するシステムCatalog-Searchを実装し、Corpus Searchとの併用によってより容易なコーパス検索方法を提案する。

キーワード：音声コーパス、情報検索、言語資源

1 はじめに

近年、音声言語データの需要に伴い、言語資源を扱う機構も増えてきた。その中一番規模の大きいのはアメリカのLDC(Linguistic Data Consortium)[1]とヨーロッパのELRA(European Language Resources Association)[2]である。しかし、両センターがそれぞれコーパスの規格を持ち、提供しているコーパス検索システムも違うので、コーパス利用者が膨大なコーパスから、自らの目的に適したコーパスを選び出すことが困難である。日本においても国立情報学研究所音声資源コンソーシアム（NII－SRC）において、音声言語コーパスの管理・運用が行われており、取り扱い件数が増えるにしたがって検索機能のニーズが高まっている。

この問題に対処するために、我々はコーパスの内容を表すいくつかの特徴(コーパス特徴属性)[3]を表現し、その属性をフィルタリングすることにより、利用者が絞った条件に合うコーパスの検索システムCatalog-Searchを構築した。そして我々が開発した音声言語コーパスの類似性可視化システムCorpus Search[4]との併用によってより容易なコーパス検索方法を提案する。

本稿では、まず音声言語コーパスの類似性可視化システムCorpus Searchについて概説し、本研究が提案する音声言語コーパス目録検索システムCatalog Searchについて説明する。

2 先行研究

Corpus Searchシステムは複数のコーパスについて、コーパスの内容を表すいくつかの特徴(コーパス特徴属性)で表現し、その類似性を多次元尺度構成法により2次元に圧縮して表示するシステムである[5]。

多次元尺度構成法を適用してコーパス

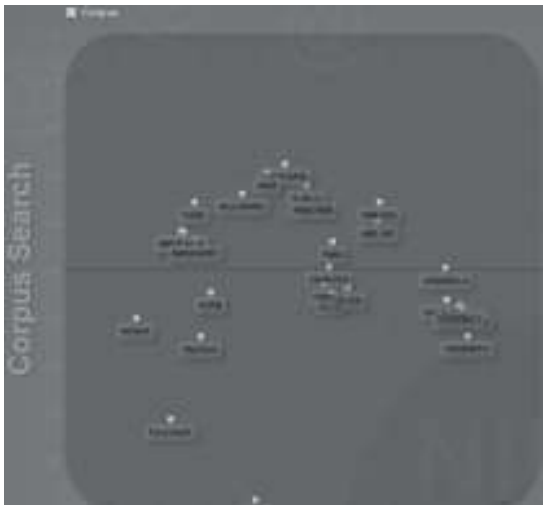

Figure 1 Corpus map [5]

の空間配置を導出し、Corpus Searchでは上位2次元のマップを表示する(図1参照)。

注目したい特徴の重みを調節することで、目的にかなった特徴を持つコーパスがマップ上にまとまって表示され、位置関係のわかるようなシステムである。しかし、コーパスに馴染みのない利用者の視点からみると、コーパスの類似度よりまず「コーパスを探したい」ということは先になるかもしれない。なので、コーパスの検索システムが必要になってくる。

3 Catalog-Search

3.1 概要

コーパスの検索を実現するため、まずコーパス利用の実態を調査し、コーパス特徴属性を決めた。さらにそれぞれの属性の下にいくつかの特徴項目を設定している。[4]で導入された属性と項目に基づいて、寄せられた意見に参考し修正を加え、合計１０属性(表１参照)６６項目を扱う。

Table 1 Corpus attributes

Attributes	Items	Contents
Sources	7	Recording devices
Environment	4	Recording environment
Speakers	12	Numbers of speakers
Quantity	7	Quantity of data
Style	4	Speech style
Mode	5	Speech mode
Sampling Rate	3	Sampling rate
Data	9	Miscellaneous data
Languages	4	Languages of data
Purpose	11	Purpose of construction

現在、国立情報学研究所音声資源コンソーシアム（NII－SRC）［6］で扱われている 76 コーパスを対象とした。今後もさらに国内外のコーパス情報を追加していく予定である。

3.2 システム動作

システム全体の流れを以下に示す(図 2 参照)。

ユーザによる操作は全てウェブブラウザーにて行われる。ユーザが注目したい属性と項目をチェックすると、サーバ上にあるコーパス情報との照合が行われる。その照合結果がファイルとしてまとめられ、サーバからウェブブラウザーに送られ、さらにテーブル形式でユーザに提供される（図 3 参照）。

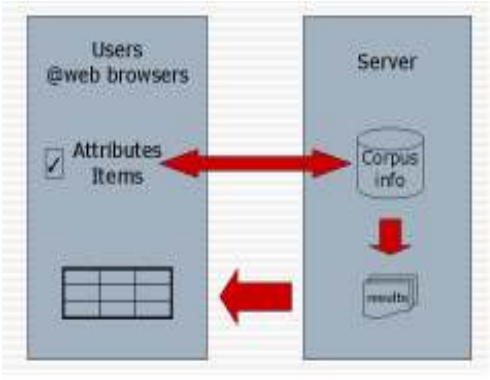

Figure 2: flowchart of Catalog-Search

Figure 3: An example of items check

結果出力では、検索条件に合致したコーパスの詳細が表示され、コーパスの名前がクリックされると、そのコーパスの詳細ページを表示する（図 4 参照）。

Figure 4: An example of output results

ユーザが行われた検索の結果を見て、さらに検索条件を修正しようとする場合、同ページ内で再検索することも可能になっている。

3.3 システム応用

第 1 章でも述べたように、本システムは[6]の音声言語コーパスの類似性可視化システム Corpus Search との併用を前提としている。本研究が提案するシステム上では、属性や項目のフィルタリングによりコーパスをヒットすることが可能だが、指定した条件における、登録された全コーパスの相互の位置関係を表現することはできない。そのことは可視化システム Corpus Search では実現できる。

つまり、両システムの併用によって、ユーザの目的に応じたコーパス情報が得られ、さらにそのコーパス間の関係性を知ることもできる。ユーザが自分のニーズと嗜好で自由に使用することが望まれる。

可視化システム Corpus Search の評価アンケートで「使いやすさ」の調査が行われ、知っているコーパスの多いユーザの評価が高いという結果が見られた[7]。そうでないユーザにとって、Catalog-Search システムがより容易に利用できるかもしれない。

4 システムの予備評価

構築したシステム Catalog-Search の一般公開に先立って予備的な評価を行った。数名の大学学生から自由記述で意見を求めた。

検索について、属性の分類や名称に関する指摘が数件見られた。「“Languages”の類には mixed-language があるべきのでは」や「“Data”の類には、もっとある気はする」などの意見があった。やはり

ユーザの利用目的や使い方によって、コーパスに対して期待する内容がそれぞれ違い、［8］で述べられたように、コーパス内容を表す特徴属性には、標準化の必要がある。

ユーザインタフェースについて、画面デザインに関して、「もっと属性間の境を分かるように」のような有益な指摘が寄せられた。さらに、「検索ボタンをもっと大きく」のようなインターネット検索などの馴染みのある表現形式の提案も複数寄せられた。

5 おわりに

[6]が提案する音声言語コーパスの特徴属性を改良して、フィルタリングする手法によって検索を実現するシステムに実装して予備評価を行った。

今後は、予備評価に際して寄せられた意見を参考にして、(1)特徴属性標準化の検討、(2)インターネット検索表現や形式の見直し、(3)可視化システム Corpus Search との連携の再考を行い、一般公開し、システム評価を行う予定である。

謝辞　本研究を行うにあたり、音声資源コンソーシアムより、共同研究の一環として音声言語コーパスの属性データを利用させていただいた。記して感謝する。また、日頃議論していただく共同研究のメンバーに感謝する。

参考文献

［1］ ［Linguistic Data Consortium］-LDC http://www.ldc.upenn.edu/

［2］ ［European Language Resources Association］ - ELRA http://www.elra.info/

［3］ 山川，松井，板橋，“多次元尺度化構成法を用いた複数音声コーパスの可視化法の検証”，音講論(春)，3-Q-10，395-396，2008.

［4］ 山川，松井，菊池，板橋，“複数音声コーパスの可視化における音響特徴量の利用”，音講論(春)，2-P-6，457-458，2009.

［5］ 菊池，沈，山川，板橋，松井，“音声言語コーパスの類似性可視化システムの構築”，音講論(秋)，3-P-33，441-442，2009.

［6］ ［音声資源コンソーシアム］ http://research.nii.ac.jp/src/

［7］ 石本，板橋，山川，沈，菊池，松井，“音声コーパスの類似性可視化システムの改良”，音講論(春)，2011(予定)

［8］ Itahashi，Yamakawa，Matsui，Ishimoto，“A proposal for standardizing catalogue specifications of speech corpora”，Proc. Oriental COCOSDA Workshop 2010，2010.

［9］ 山川，松井，板橋，“多次元尺度化構成法を用いた複数音声コーパスの可視化”，音講論(秋)，1-P-20，447-448，2007.

＊ Construction and application of the search system of speech corpora - Catalog-Search，by KIKUCHI Hideaki，Raymond SHEN (Waseda University)