統合データベース開発：

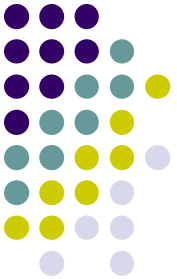
専門用語辞書管理システムと専門用語解析技術の開発

奈良先端大
松本裕治，新保仁，浅原正幸，原一夫

- 専門用語解析技術
  - 専門用語辞書システムの開発
  - 専門用語解析技術の開発
- 専門用語抽出ツールの設計と開発
  - 専門用語辞書拡張支援ツールの設計と開発

今年度設定した成果目標

- 専門用語辞書システムの開発
  - 次のような機能をもつ辞書（用語管理）システムを開発
    - 専門用語の内部構造や同義語識別子を用いた検索
    - 専門用語の意味情報を管理（シソーラスコード等）
- 専門用語解析技術の開発
  - 専門用語の内部構造解析済みデータの拡大（800語程度のデータを2000語程度にまで拡大）
  - 専門用語の内部構造の自動解析手法の開発
  - 文書内の並列表現の言語解析技術の高性能化
- 専門用語抽出ツールの設計と開発
  - 文書中の専門用語の意味クラス分類手法の高性能化
  - 新規の用語をシソーラスへ登録するユーザインタフェースの設計



専門用語解析システム
1．専門用語辞書システム

辞書管理システムCradle

- 形態素解析用辞書の管理ツール
  - 現在登録している辞書，用語集
    - これまでライフサイエンス辞書（京大金子研究室）
    - 今年度，標準病名マスターv2.80を格納
    - 先日，仲里さんよりいただいた専門語候補(18万語)を登録
  - 複合語に対する内部構造付与
    - 現在約1800語について人手により内部構造付与
  - 見出し，読み，品詞，シソーラスコード，英訳．．．の項目に加えて，同義語検索，内部構造に基づく検索を実装
  - 表示項目のカスタマイズ機能を実装

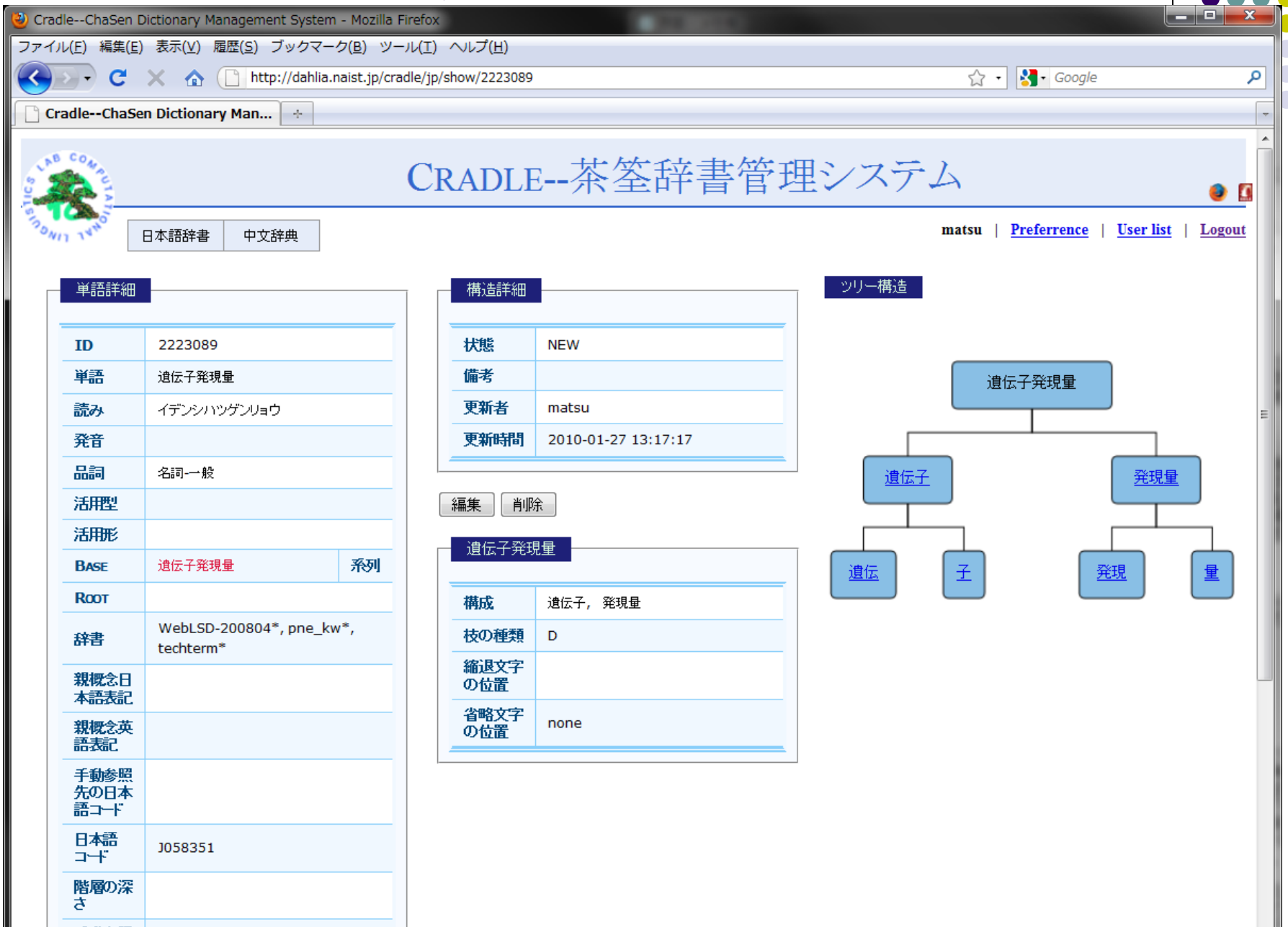
検索画面


Cradle--ChaSen Dictionary Management System - Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://dahlia.naist.jp/cradle/
Google
Cradle--ChaSen Dictionary Man...
CRADLE--茶筌辞書管理システム
日本語辞書
中文辞典
matsu | Preferrence | User list | Logout
単語属性
リセット
検索
ID
単語
読み
発音
品詞
活用型
活用形
Base
辞書
NAIST-jdic-20080707
WebLSD-200804*
標準病名マスター-V2.80*
pne_kw*
techterm*
文字数
状態
新規者
更新時間
更新者
親概念日本語表記
親概念英語表記
手動参照先の日本語コード
日本語コード
階層の深さ
手動参照先の日本語表記
自動参照先ID
ツリー番地
自動参照先表記
ツリー日本語
親概念ID
ツリー英語
ICD10
頻度(文数)
確信度
頻度(文献数)
複合語属性
リセット
検索
内部表記
include
内部読み
include
内部POS
include
状態

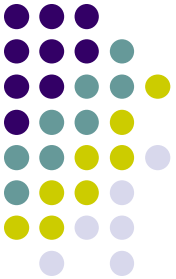
検索例(遺伝子)

	ID	単語	読み	発音	Base	Root	辞書	品詞	活用型	活用形	構造
詳細	2221835	融合遺伝子	ユウゴウイデンシ		融合遺伝子			名詞一般			
詳細	2222191	ターゲット遺伝子	ターゲットイデンシ		ターゲット遺伝子			名詞一般			
詳細	2223089	遺伝子発現量	イデンシハツゲンリョウ		遺伝子発現量			名詞一般			
詳細	2223353	感受性遺伝子	カンジュセイイデンシ		感受性遺伝子			名詞一般			
詳細	2223586	候補遺伝子	コウホイデンシ		候補遺伝子			名詞一般			
詳細	2224194	がん遺伝子	ガンイデンシ		がん遺伝子			名詞一般			
詳細	2224256	癌抑制遺伝子	ガンヨクセイイデンシ		癌抑制遺伝子			名詞一般			
詳細	2224278	遺伝的多型性	イデンテキタケイセイ		遺伝的多型性			名詞一般			
詳細	2224289	逆遺伝学	ギャクイデンガク		逆遺伝学			名詞一般			
詳細	2224628	致死遺伝子	チシイデンシ		致死遺伝子			名詞一般			
詳細	2224854	遺伝子工学	イデンシコウガク		遺伝子工学			名詞一般			
詳細	2227511	遺伝性腫瘍	イデンセイシュヨウ		遺伝性腫瘍			名詞一般			
詳細	2228263	遺伝子ファミリー	イデンシファミリー		遺伝子ファミリー			名詞一般			
詳細	2230000	myc遺伝子	ミックイデンシ		myc遺伝子			名詞一般			

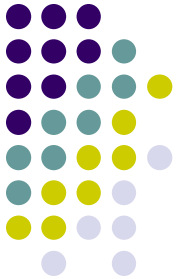
単語情報の表示


Cradle--ChaSen Dictionary Management System - Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://dahlia.naist.jp/cradle/jp/show/2223089
Google
Cradle--ChaSen Dictionary Man...
CRADLE--茶筌辞書管理システム
日本語辞書
中文辞典
matsu | Preferrence | User list | Logout
単語詳細
ID 2223089
単語 遺伝子発現量
読み イデンシハツゲンリョウ
発音
品詞 名詞一般
活用型
活用形
BASE 遺伝子発現量 系列
ROOT
辞書 WebLSD-200804*, pne_kw*, techterm*
親概念日本語表記
親概念英語表記
手動参照先の日本語コード
日本語コード J058351
階層の深さ
構造詳細
状態 NEW
備考
更新者 matsu
更新時間 2010-01-27 13:17:17
編集
削除
遺伝子発現量
構成 遺伝子, 発現量
枝の種類 D
縮退文字の位置
省略文字の位置 none
ツリー構造
遺伝子発現量
遺伝子
発現量
遺伝
子
発現
量

専門用語辞書システムの今後の予定

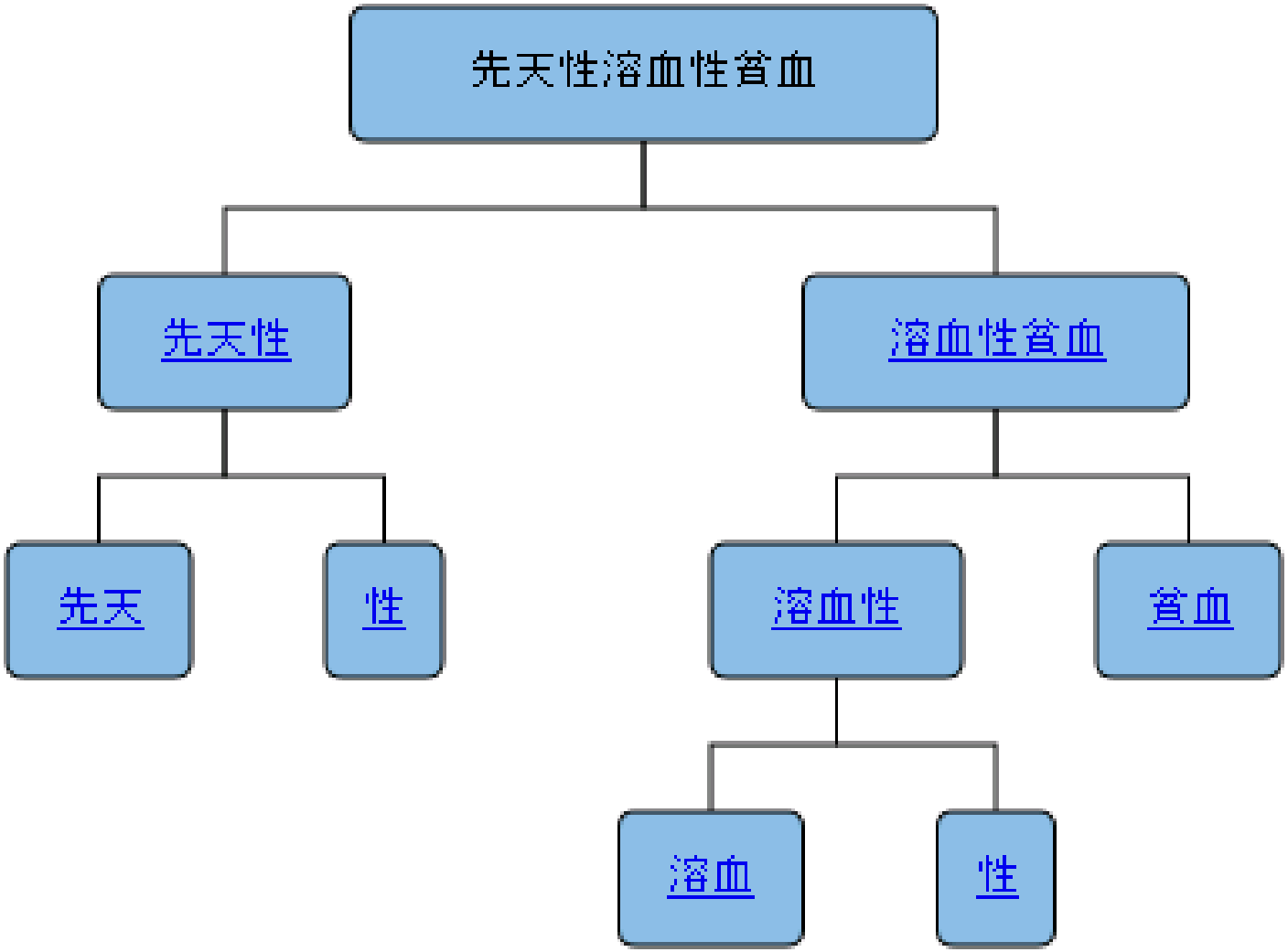


- 検索・表示機能の拡張
  - 意味情報に基づく検索
  - 意味関係（上位・下位関係，類似度など）の表示
- 内部構造解析データの拡張
  - 現状の1800語を2000語以上に拡張（できれば3000語以上）
  - 内部構造の自動解析の精度評価
- 他の機能拡張
  - 汎用のWebシステム化（現在は，Firebox対応のみ）
  - 表示項目のカスタマイズ機能：数値属性，順序属性

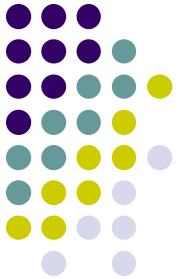


専門用語解析技術
２．専門用語解析技術の開発

複合語の内部構造解析



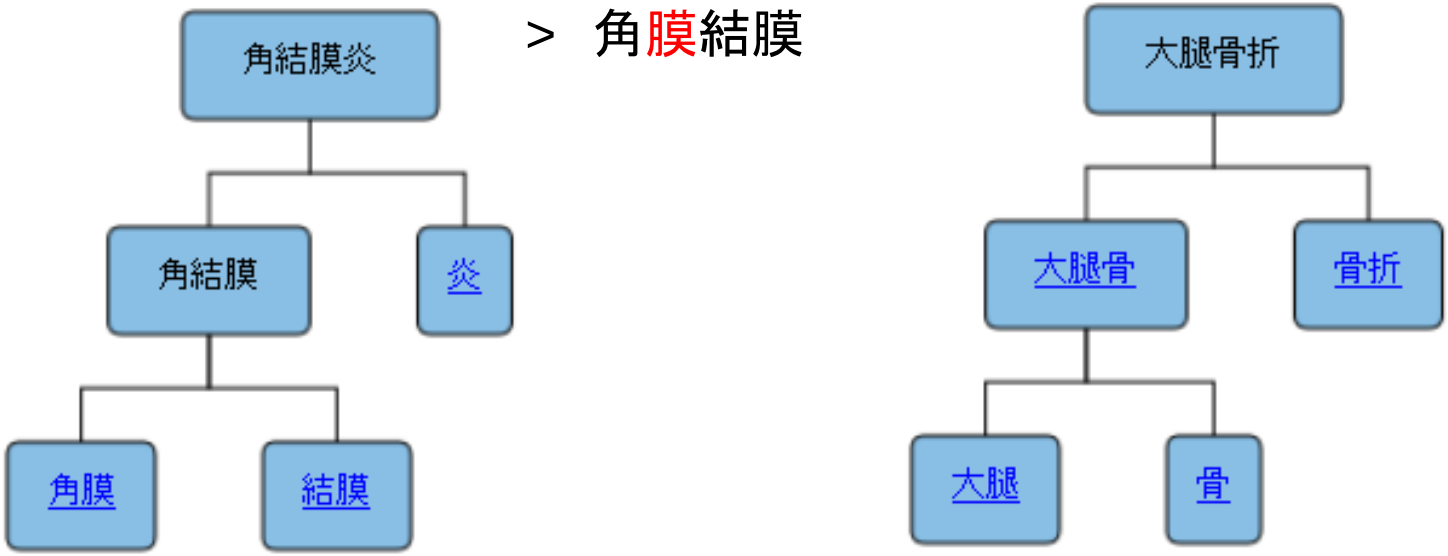
内部構造の関係の分類



- 構造間の関係の分類
  - 通常の係り受け(D)
    - 急性　=>　肺炎
  - 逆向きの係り受け(R)
    - 糖尿病　<=　I型
  - 並列(P)
    - 脊髄　=>　小脳
  - その他・方向無し(U)
    - B　=>　1　=>　6　(B16メラノーマ細胞)

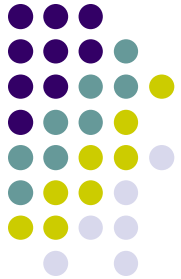
内部構造における縮退現象

- 縮退
  - 縮退する文字の位置
    - End + Begin
      - 大腿骨 + 骨折 => 大腿骨骨折
    - End + End



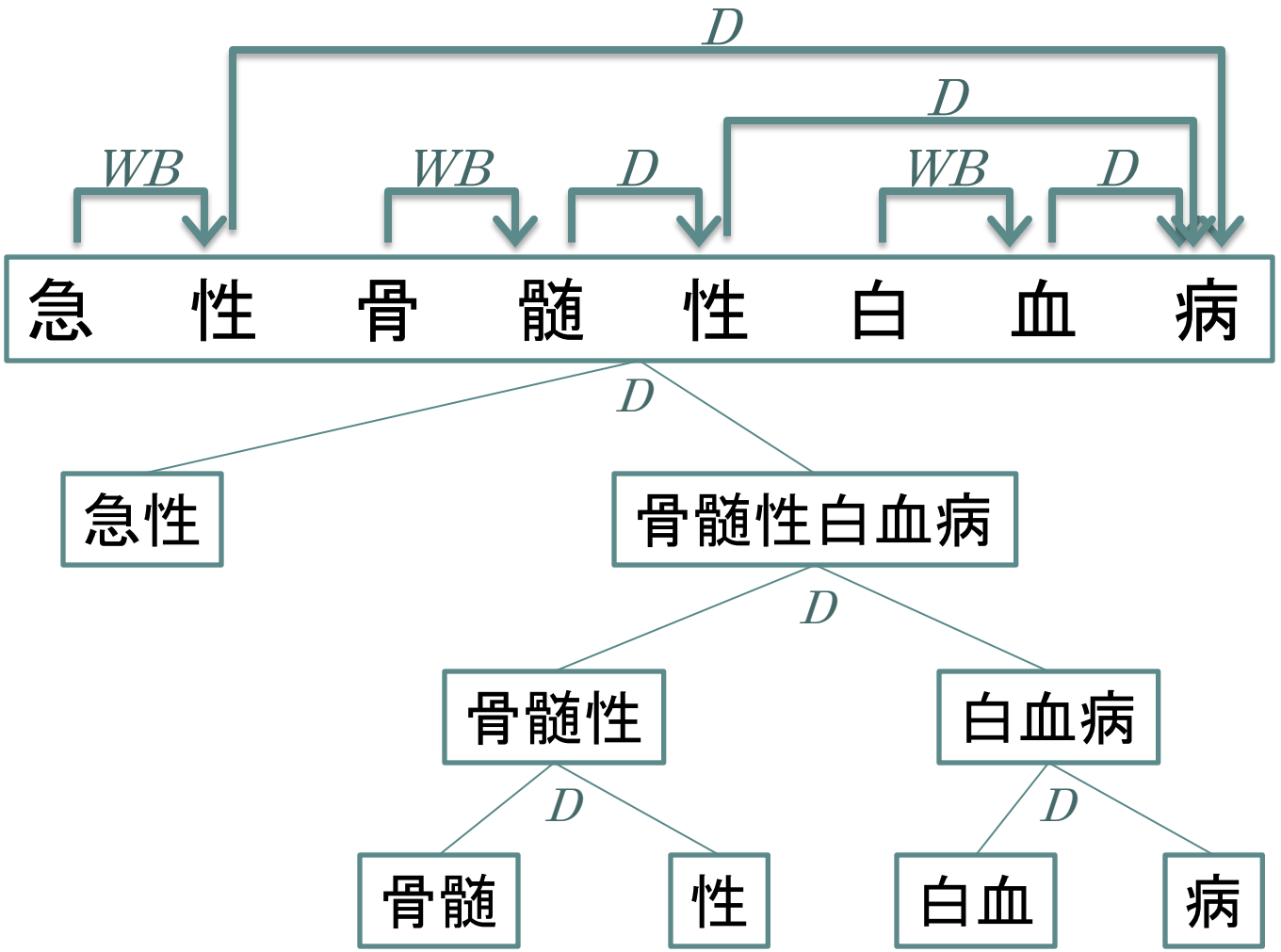
> 角膜結膜

文字単位の係り受けによって記述

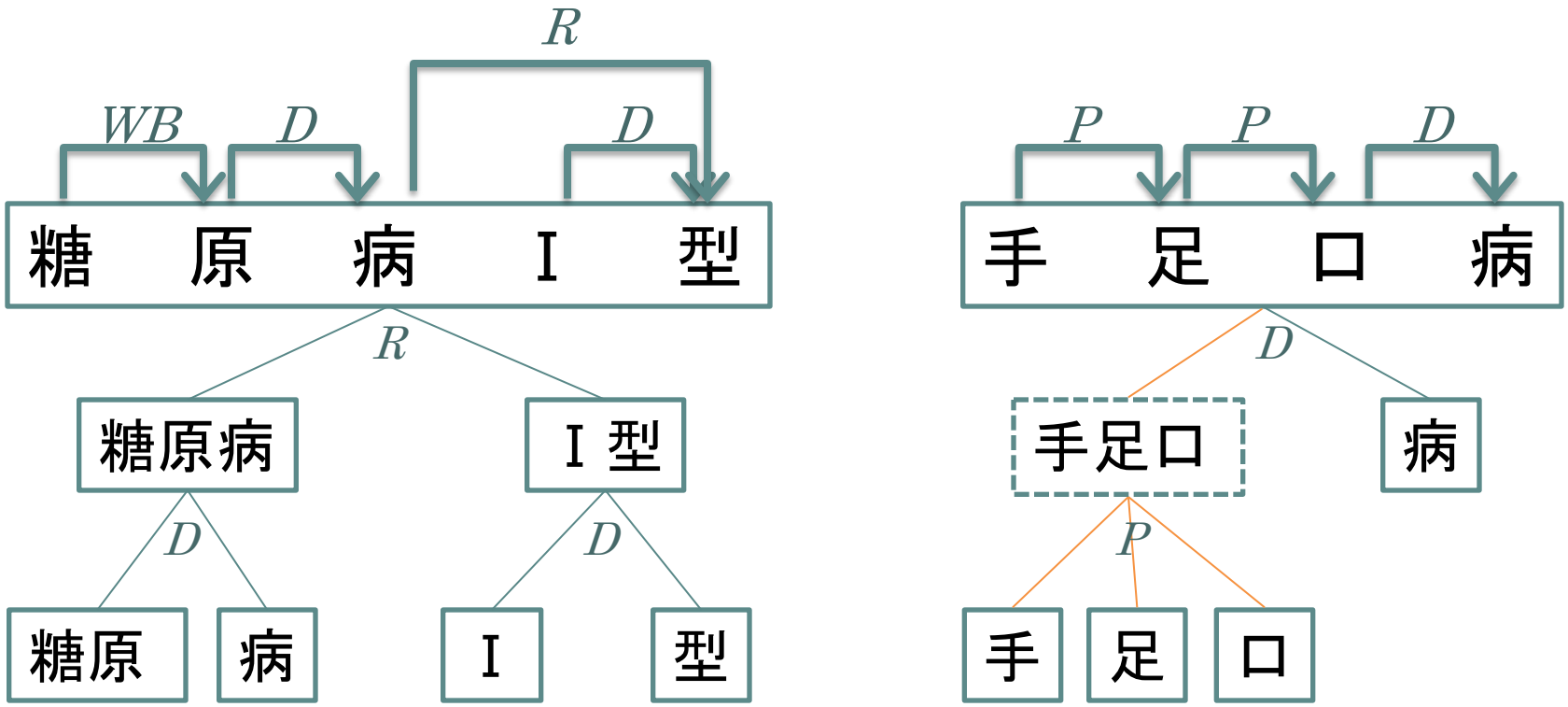


- 係り受けの種類（合計６種類）
  - 形態素を構成する係り受け
    - 形態素の先頭（WB）、中間（WI）
  - 形態素間の係り受け（前述の４種類）

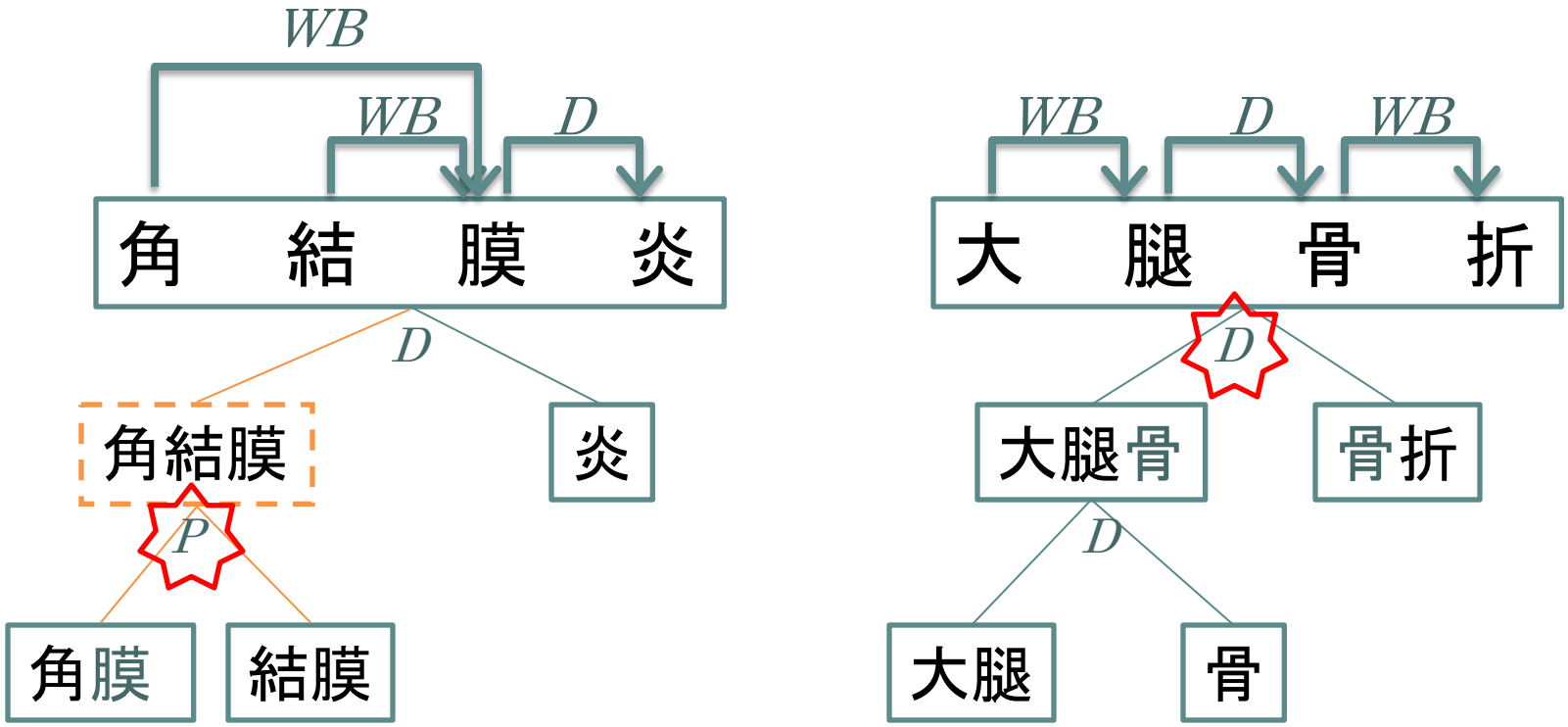
文字単位の係り受けによる表現



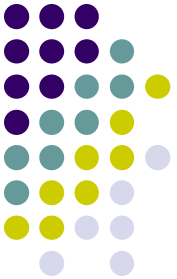
文字単位の係り受けによる表現



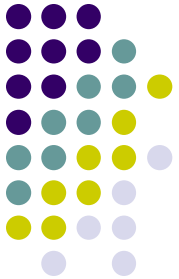
文字単位の係り受けによる表現



内部構造解析実験:現状と予定



- ライフサイエンス辞書中の804語に対して人手で解析を行い，5分割交差検定
- 文字単位の係り受け解析
  - 解析アルゴリズム:Nivreによるshift-reduce解析
  - 係り受け関係の学習:Support-Vector-Machine
- 解析精度
  - 文字単位精度: 96.9% (ラベル無しでは97.8%)
  - 用語単位精度: 87.6% (ラベル無しでは89.9%)
- 今後:学習事例の追加，解析アルゴリズムの改良，機械学習に有効な素性を選択



（2）専門用語抽出ツールの設計と開発
専門用語辞書拡張支援ツール

本研究の目的＝シソーラス拡張

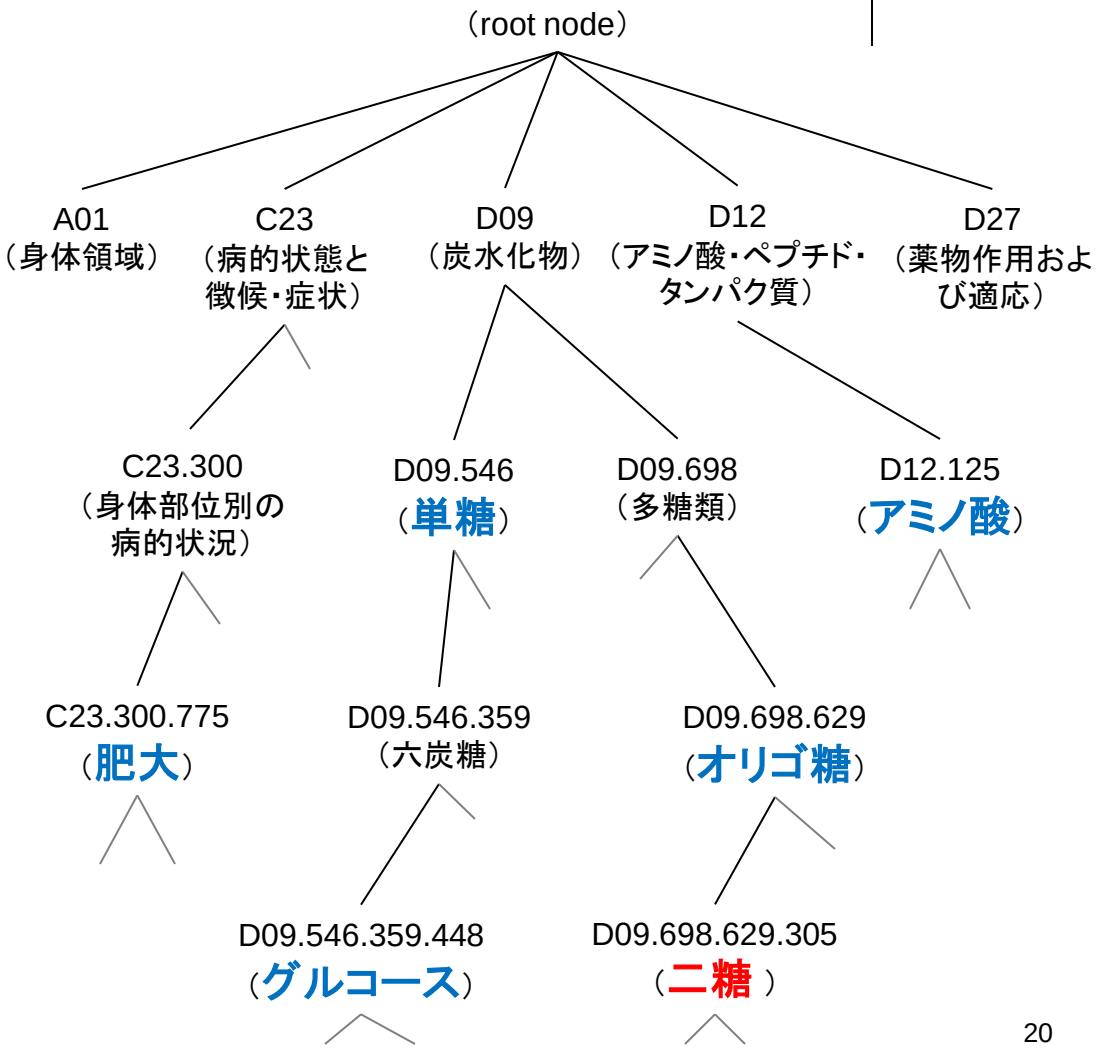
- 新規登録対象の専門用語（クエリ）に対して，類似度が高い順に登録済の専門用語をランク付けし，提示するシステムの構築
- シソーラス辞書の編集者は，システムが提示するランキング上位語を参考に，新しい専門用語をシソーラス辞書にマッピングする

シソーラス拡張の例:指定した語の類義語を検索して表示したい

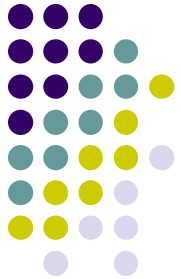
クエリ: 二糖

類義の上位語:

類似度ランキング	
1位	オリゴ糖
2位	単糖
3位	アミノ酸
4位	肥大
5位	グルコース

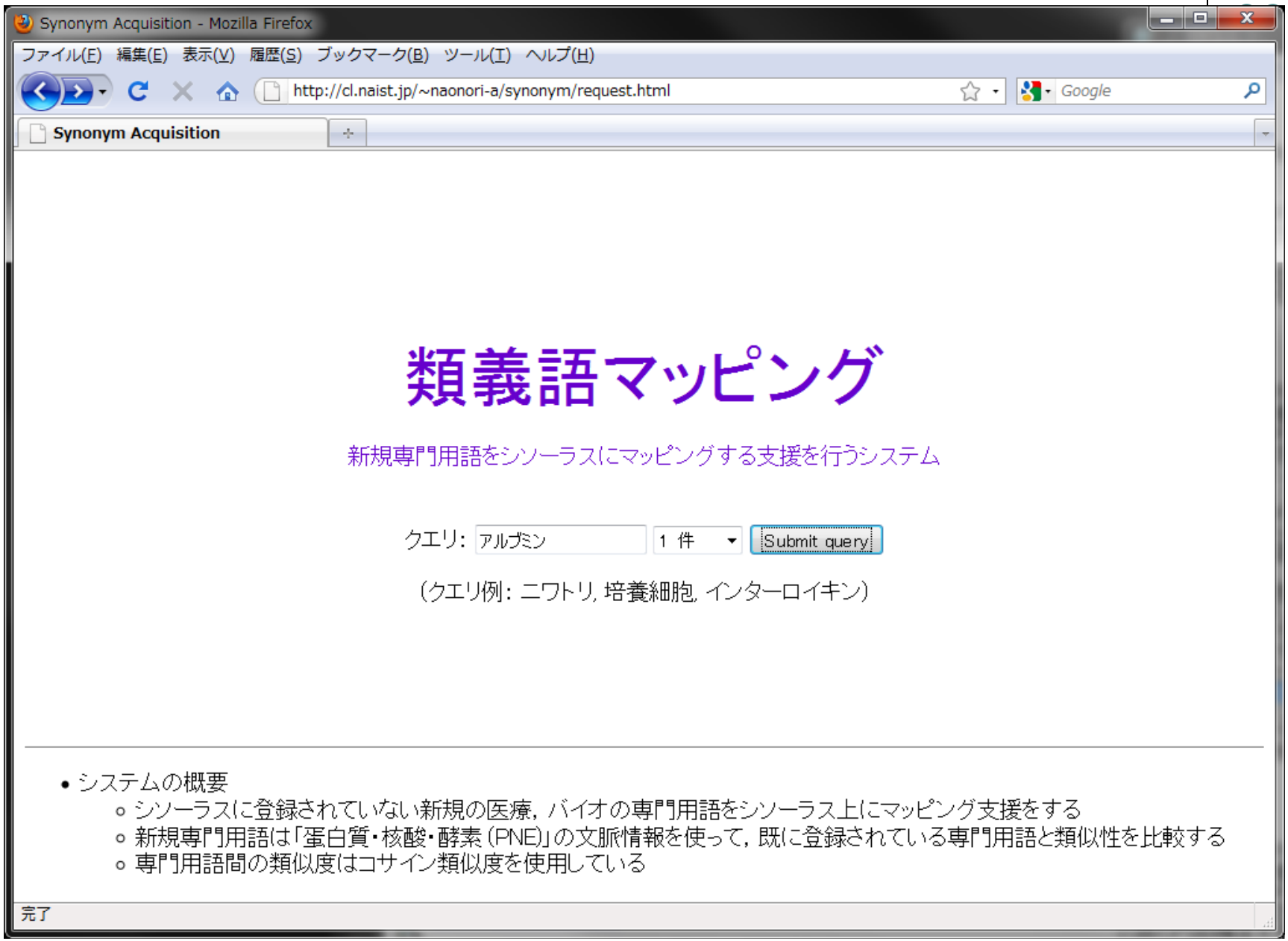


専門用語辞書拡張支援ツールの現状



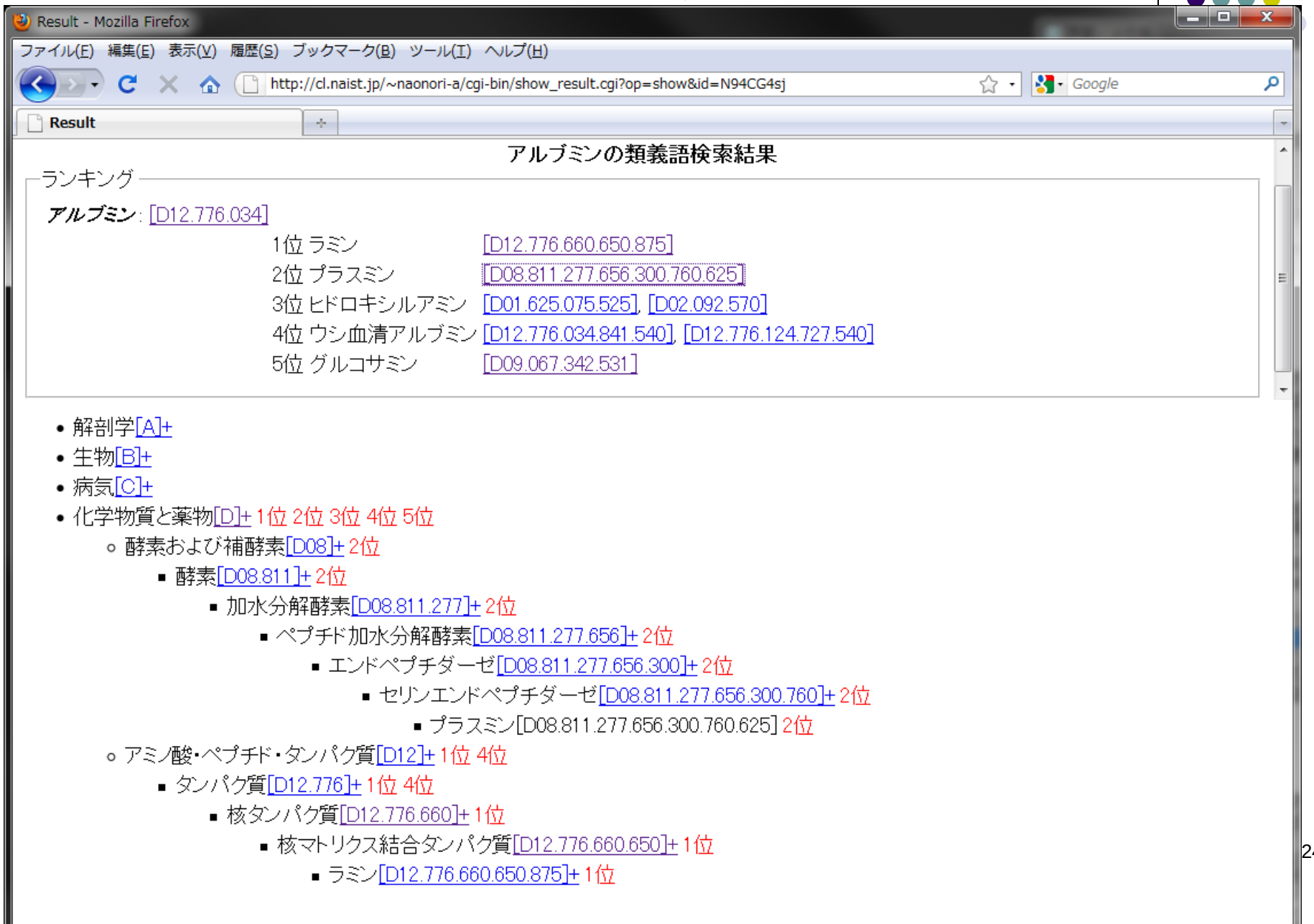
- 専門文書から対象とする語の文脈情報を抽出して用語の隣接グラフを作成し，グラフ構造を用いて用語間の類似度を算出する手法を提案
- 雑誌「蛋白質・核酸・酵素」を実験データとして用い，そこに登場する新規の専門用語と類似度の高い語をライフサイエンス辞書から検索するツールを開発した
- 指定した用語に対して，類似度上位の語を検索し，シソーラス内の位置を表示するインタフェースを構築

類義語検索ツールの初期画面

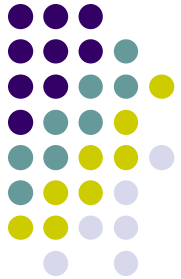


検索結果の表示（上位5位の場合）

シソーラスの部分的表示



専門用語辞書拡張支援ツールの今後の予定



- 新規語に対する類義語を動的に検索する機能の実装
- 新規語をシソーラスに登録する機能の実装
  - MeSHのシソーラスコードを付与する機能
  - 専門用語辞書システムCradleとの連携
- 類義語の表示インタフェースの改良

来年度計画のまとめ

- 専門用語辞書システムの開発
  - 用語間の意味関係（上位/下位など）や類似度を表示する機能の実現
  - 種々のWebブラウザ上で辞書の検索，編集ができる辞書システムとして汎用化
- 専門用語解析技術の開発
  - 内部構造解析済みデータの拡大（2000語以上に拡大）
  - 一般的な統語解析アルゴリズムを拡張し，解析精度の向上
    - 用語単位の解析精度90%以上を目指す
- 専門用語抽出ツールの設計と開発
  - 文書中の専門用語の意味クラス分類手法の高性能化
  - 新規の用語をシソーラスへ登録する機能の実装
  - Cradleとの連携