資料　6

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

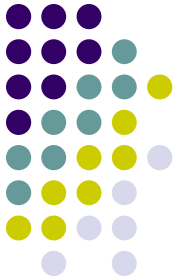
- 専門用語解析技術
  - 専門用語辞書システムの開発
  - 専門用語解析技術の開発
- 専門用語抽出ツールの設計と開発
  - 専門用語辞書拡張支援ツールの設計と開発

研究目標

- 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
    - 用語の内部構造の記述
    - 用語の意味クラス等の詳細情報の記述
    - 同義語・同一語の異表記の情報の記述
- 専門用語解析技術の開発
  - 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同定と意味クラス推定
  - 新規の専門用語をシソーラスへ登録する支援ツールの構築

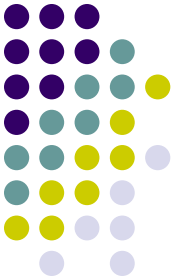
今年度の進捗状況

- 専門用語辞書システムの設計
  - 辞書(用語管理)システムCradleを試作
    - 品詞等の文法情報以外に，シソーラスコード等の意味情報も記述
    - 複合語の内部構造の記述と編集機能
- 専門用語解析技術の開発
  - 専門用語(複合語)の内部構造解析法を試作
  - 専門文書中の並列表現の解析手法を提案
- 専門用語タグ付け手法の設計
  - 専門文書に出現する新規語の意味クラスの推定法を試作



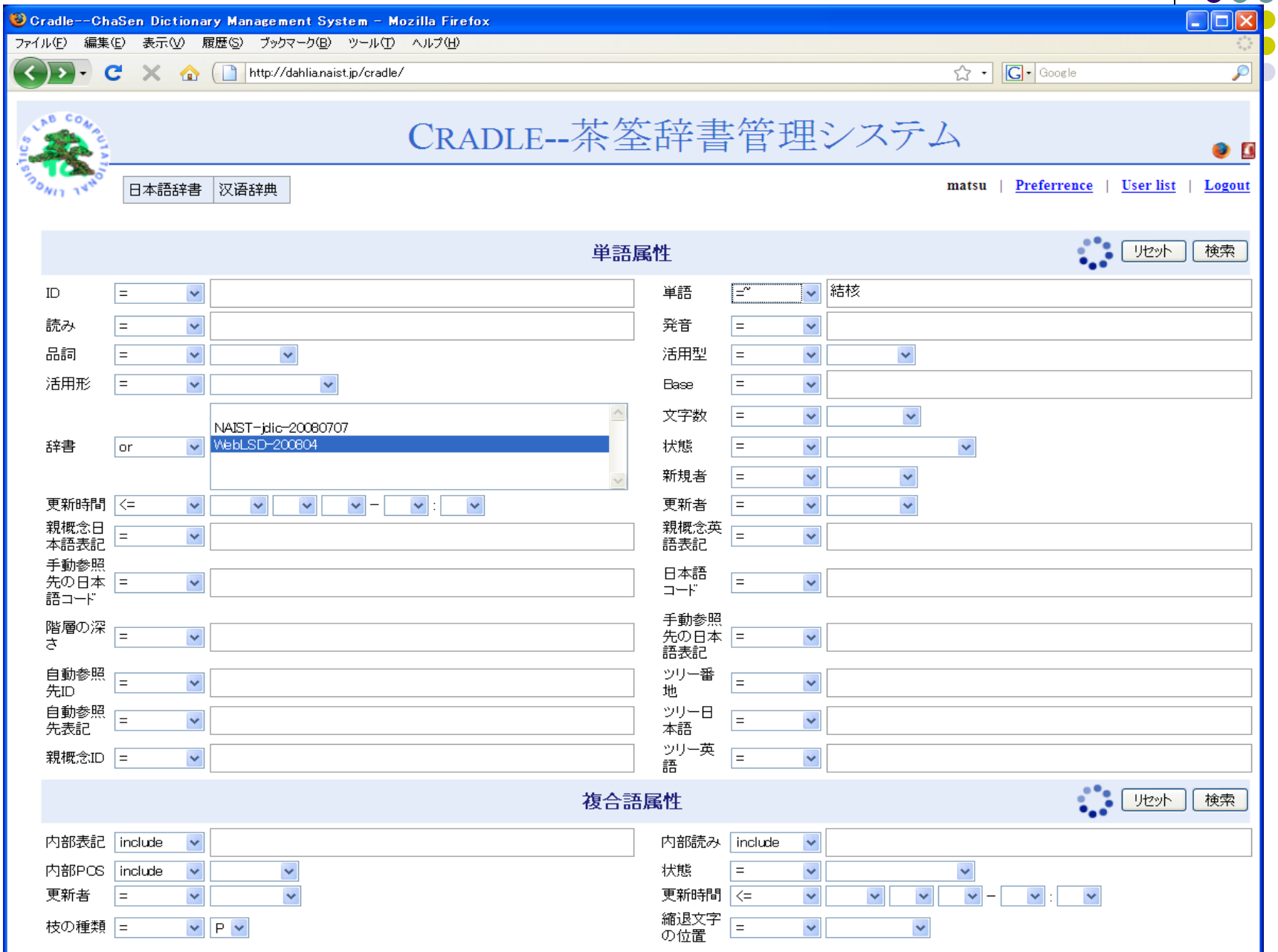
辞書（専門用語管理）システム

辞書管理システムCradle



- 形態素解析用辞書の管理ツール
  - ライフサイエンス辞書（京大金子研究室）を格納
  - 見出し，読み，品詞などの基礎情報
  - 約2万語に対してMeSHのシソーラスコード，および，英訳の情報
  - 複合語に対する内部構造付与機能
    - 約800語について人手により内部構造付与
  - 様々な検索機能
  - 同義語情報の付与機能

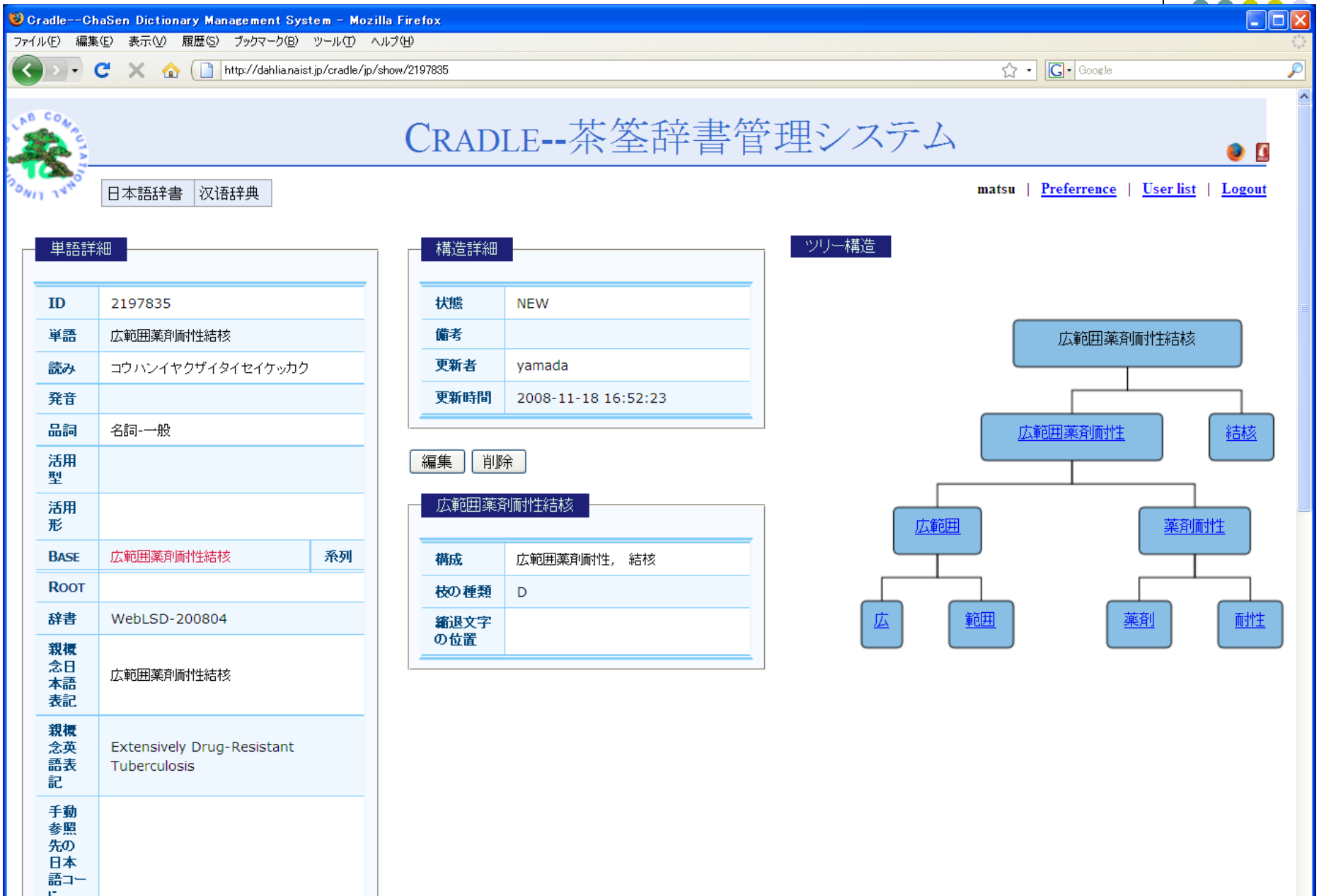
検索画面


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--茶筌辞書管理システム
日本語辞書
汉语辞典
matsu | Preferrence | User list | Logout
単語属性
リセット
検索
ID
=
読み
=
品詞
=
活用形
=
辞書
or
NAIST-jdic-20080707
WebLSD-200804
更新時間
<=
親概念日本語表記
=
手動参照先の日本語コード
=
階層の深さ
=
自動参照先ID
=
自動参照先表記
=
親概念ID
=
単語
=~
結核
発音
=
活用型
=
Base
=
文字数
=
状態
=
新規者
=
更新者
=
親概念英語表記
=
日本語コード
=
手動参照先の日本語表記
=
ツリー番地
=
ツリー日本語
=
ツリー英語
=
複合語属性
リセット
検索
内部表記
include
内部POS
include
更新者
=
枝の種類
=
P
内部読み
include
状態
=
更新時間
<=
縮退文字の位置
=

検索結果(結核)

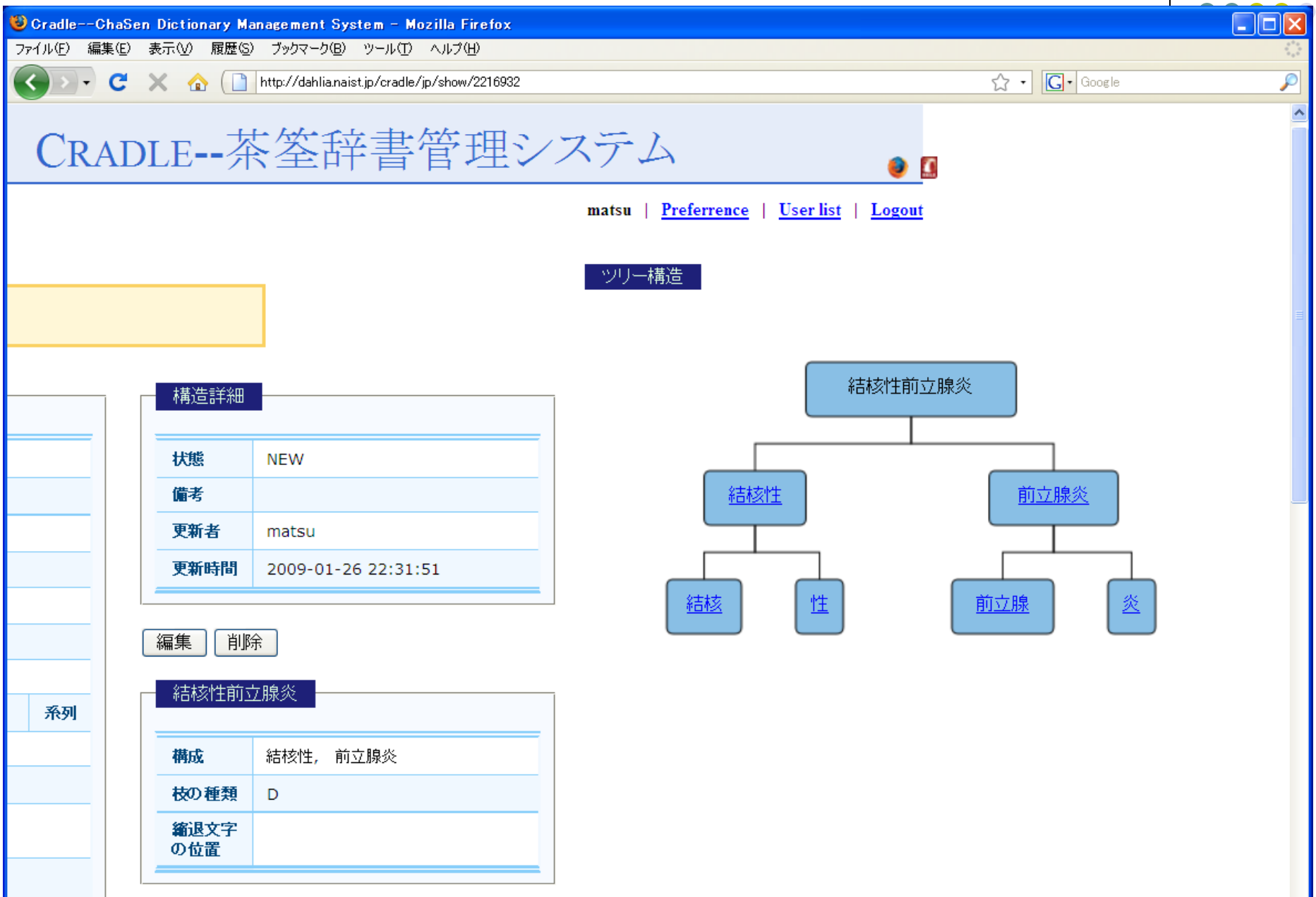

Cradle--ChaSen Dictionary Management System - Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://dahlia.naist.jp/cradle/jp/list?dynamic_lexeme_condition=&dynamic_synthetic_condition=&show_conditions=単語%3D~結核%2C
Google
CRADLE--茶筌辞書管理システム
日本語辞書
汉语辞典
matsu | Preferrence | User list | Logout
表示件数: 30
Review
辞書<=>色
表示フィールド
結果出力
« Previous 1 2 3 4 Next »
条件: 単語=~結核, 辞書:(WebLSD-200804)
93 Hits
ID 単語 読み 発音 Base Root 辞書 品詞 活用型 活用形 構造
詳細 80772 結核 ケッカク ケッカク 結核 名詞-一般
詳細 198848 肺結核 ハイケッカク ハイケッカク 肺結核 名詞-一般
詳細 2197835 広範囲薬剤耐性結核 コウハンイヤクザイタイセイケッカク 広範囲薬剤耐性結核 名詞-一般
詳細 2198017 結核性脊椎炎 ケッカクセイセキツイエン 結核性脊椎炎 名詞-一般
詳細 2198710 極度薬剤耐性結核菌 キョクドヤクザイタイセイケッカクキン 極度薬剤耐性結核菌 名詞-一般
詳細 2198905 結核性 ケッカクセイ 結核性 名詞-一般
詳細 2200540 凝結核 ギョウケッカク 凝結核 名詞-一般

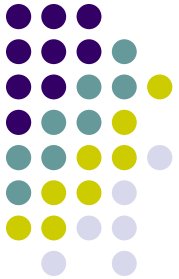
単語情報の表示


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/2197835
CRADLE--茶筌辞書管理システム
日本語辞書 汉语辞典
matsu | Preferrence | User list | Logout
単語詳細
ID 2197835
単語 広範囲薬剤耐性結核
読み コウハンイヤクザイタイセイケッカク
発音
品詞 名詞-一般
活用型
活用形
BASE 広範囲薬剤耐性結核 系列
ROOT
辞書 WebLSD-200804
親概念日本語表記 広範囲薬剤耐性結核
親概念英語表記 Extensively Drug-Resistant Tuberculosis
手動参照先の日本語コー
構造詳細
状態 NEW
備考
更新者 yamada
更新時間 2008-11-18 16:52:23
編集 削除
広範囲薬剤耐性結核
構成 広範囲薬剤耐性, 結核
枝の種類 D
縮退文字の位置
ツリー構造
広範囲薬剤耐性結核
広範囲薬剤耐性
結核
広範囲
薬剤耐性
広
範囲
薬剤
耐性

複合語内部構造のアノテーション

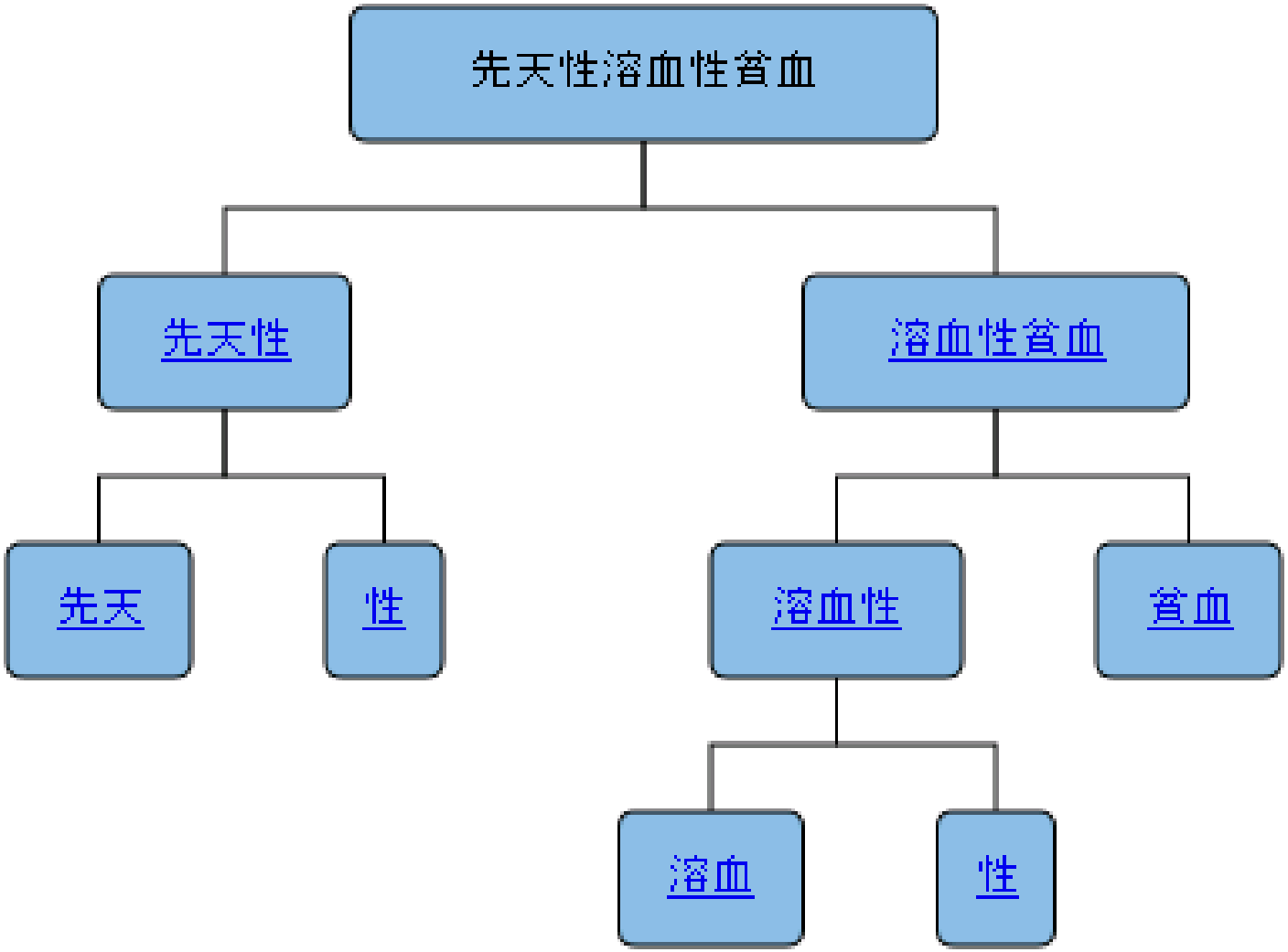
複合語の内部構造情報付与結果



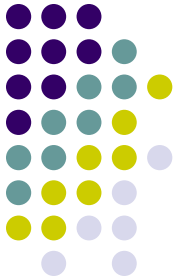


複合語の内部構造解析

複合語の内部構造



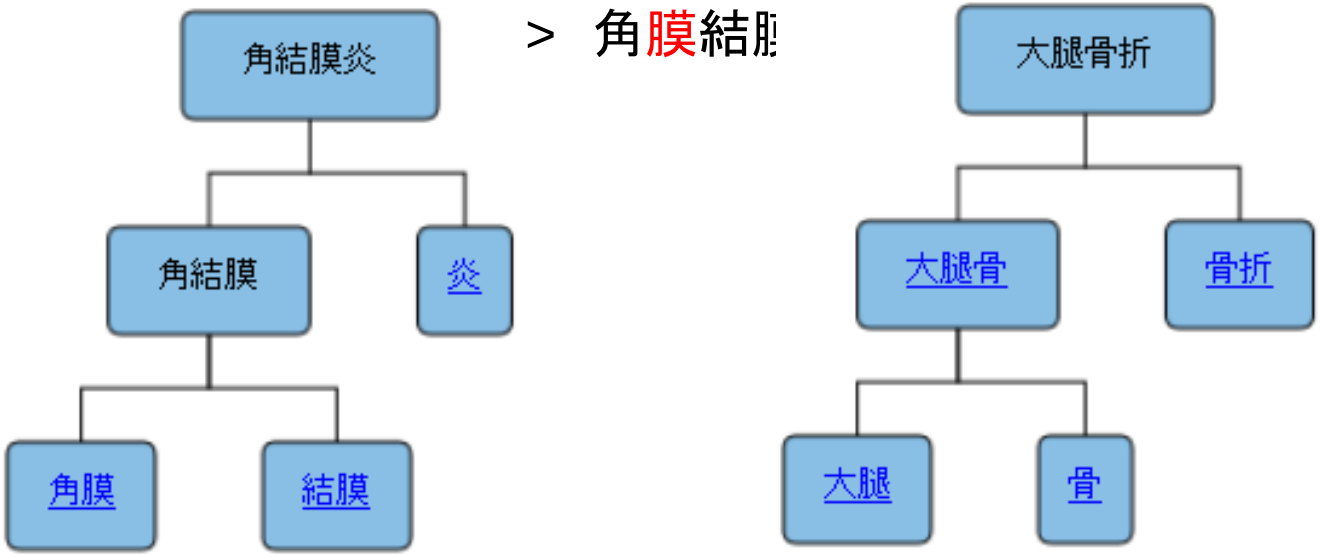
内部構造の関係の分類



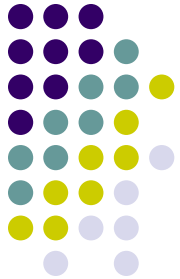
- 構造間の関係の分類
  - 通常の係り受け(D)
    - 急性　=>　肺炎
  - 逆向きの係り受け(R)
    - 糖尿病　<=　I型
  - 並列(P)
    - 脊髄　=>　小脳
  - その他・方向無し(U)
    - B　=>　1　=>　6　(B16メラノーマ細胞)

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



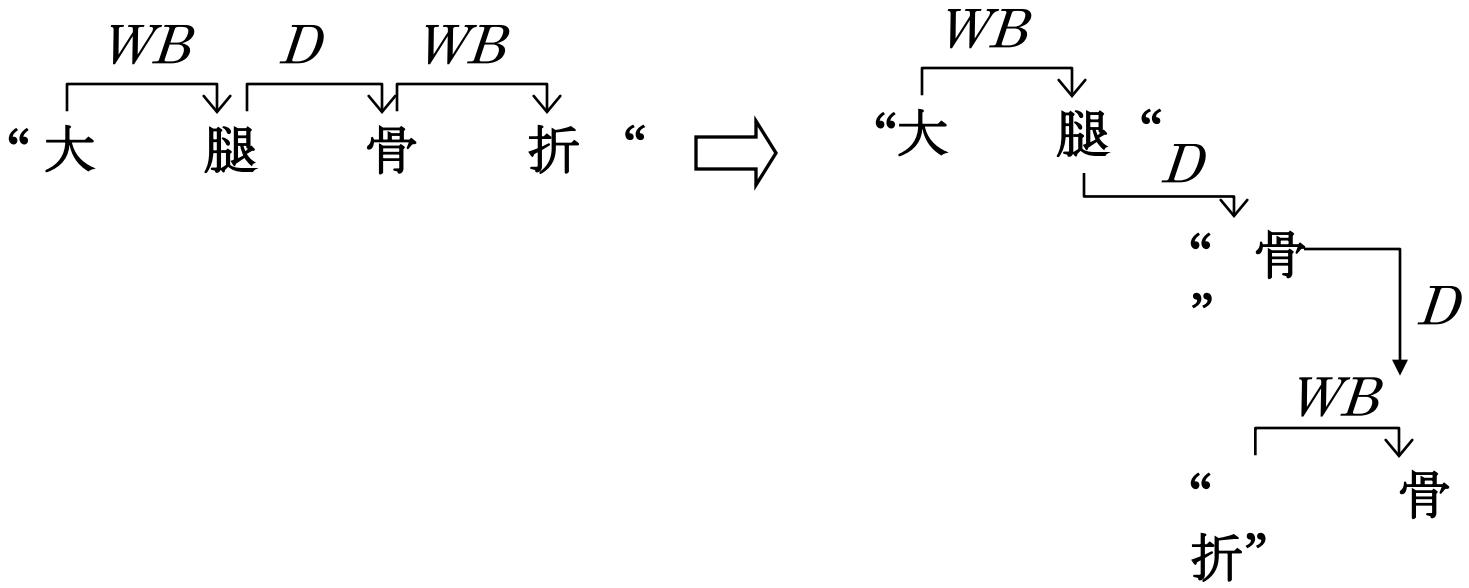
文字単位の係り受けによって記述



- 係り受けの種類
  - 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）

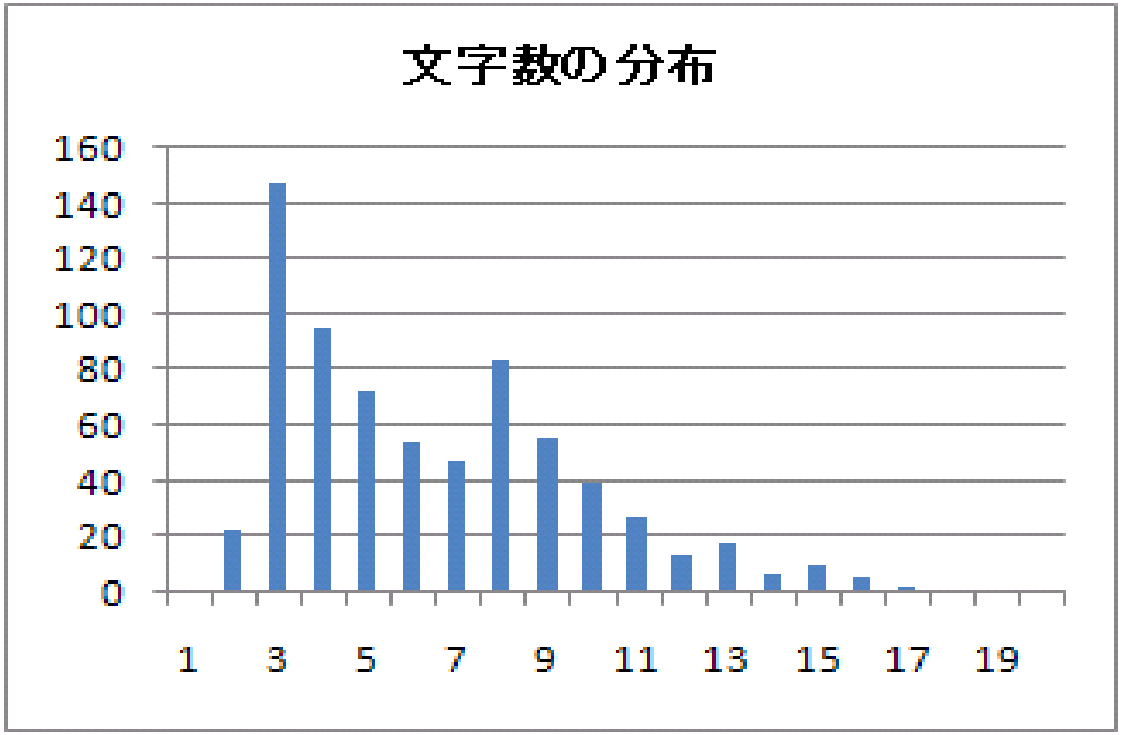
仮定

1. 縮退で消える文字は1文字
2. 縮退したときの形態素間の係り受け関係



自動解析実験のためのデータ作成

- ライフサイエンス辞書
- 病名・解剖部位など794語をタグ付け
  - 8文字以上の病名251語



実験

- 文字単位の係り受け
- SVMを用いて係り受け判定
  - 係る・係らないの二値分類（ラベルは無視）
  - TinySVM、線形カーネル

  - 素性（例：溶**血**性**貧**血）
    - 漢字（血、貧）
    - 文字種（漢字、漢字）
    - 係り先の1文字後がカタカナ（no）
    - 距離（2）
    - 元の文字列中での位置（中間、中間）
    - その二文字から成る語が存在する（血貧：no）
    - その文字で終わる語が文字列に含まれている
      （溶血：yes、性貧、血性貧、溶血性貧：no）

精度(文字対)

Data	Accuracy	Precision	Recall
1	93.7%	87.7%	85.2%
2	94.6%	89.2%	87.6%
3	93.9%	89.3%	83.8%
4	94.7%	89.3%	87.5%
5	95.1%	90.1%	88.5%
6	94.9%	90.9%	86.3%
7	92.4%	84.7%	81.1%
8	93.0%	87.4%	82.4%
9	92.1%	83.2%	83.7%
10	92.4%	85.3%	82.2%
Average	93.7%	87.7%	84.8%

精度（語）

Data	Accuracy
1	52.1%
2	59.2%
3	63.4%
4	54.9%
5	63.4%
6	59.2%
7	56.3%
8	52.9%
9	45.7%
10	48.9%
Average	55.6%

平均文字長 = 6.5

<文字対の精度>の6.5乗 = 0.655

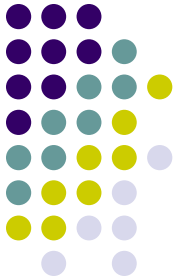
原因：データの偏り？

「～ / 症候 / 群」が56語

専門用語と並列句

- しばしば部分的な省略が起こるため，正確な解析・インデクシングのためには辞書の整備だけでは不十分

- erythroid, myeloid and lymphoid cell types
  - = erythroid cell type, myeloid cell type and lymphoid cell type
- human T and B cells
  - = human T cells and human B cells
- recombinant human nm23-H1, -H2, mouse nm23-M1, and -M2 proteins
  - = ???

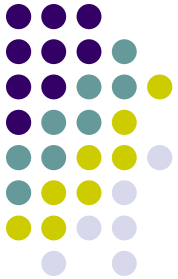


並列構造解析

並列句同定

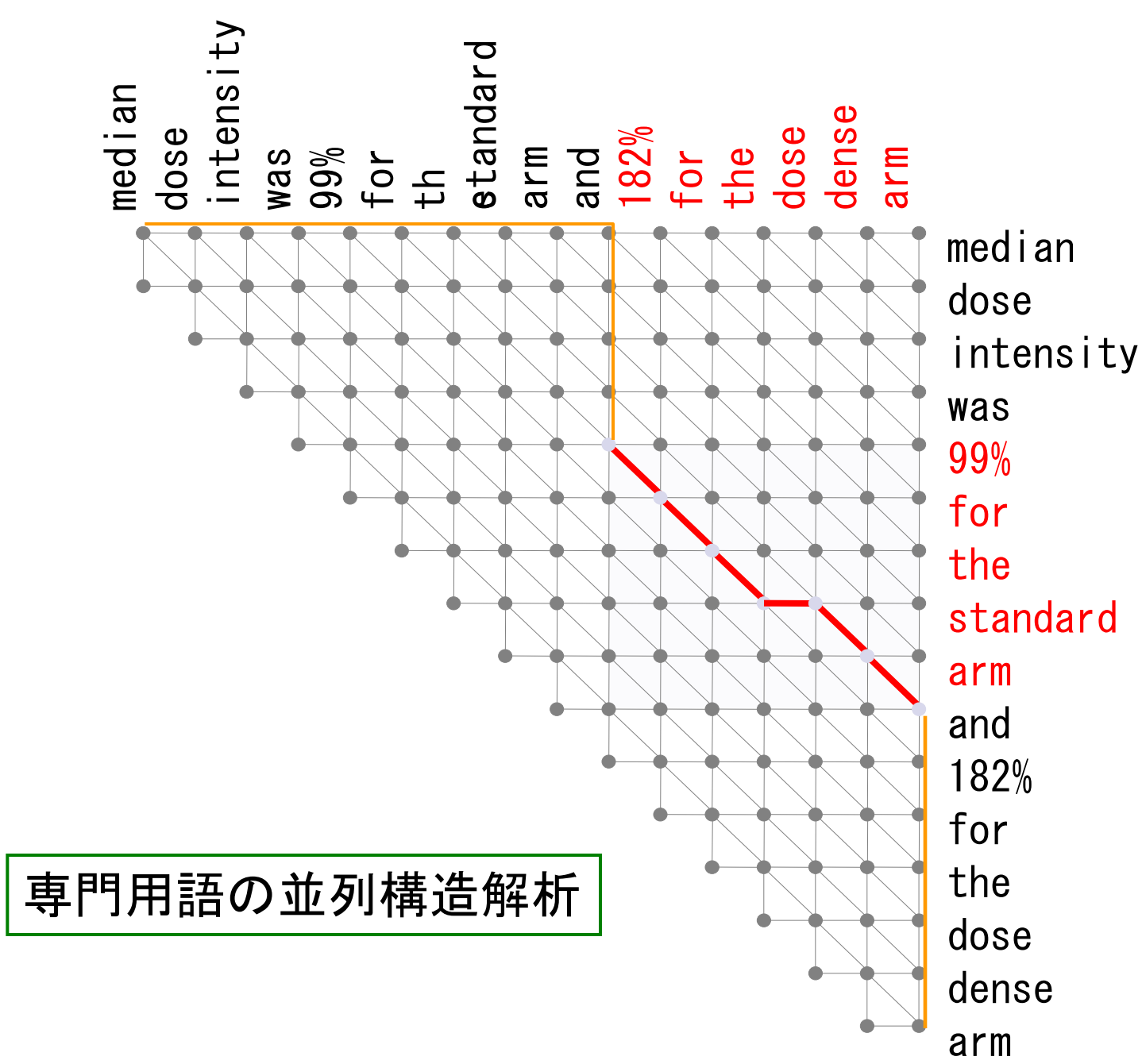
- 並列句の構成要素には構文的な類似性が高いことが多い
  - ➔文中の類似部分を検出

文中の類似部分検出による並列句同定



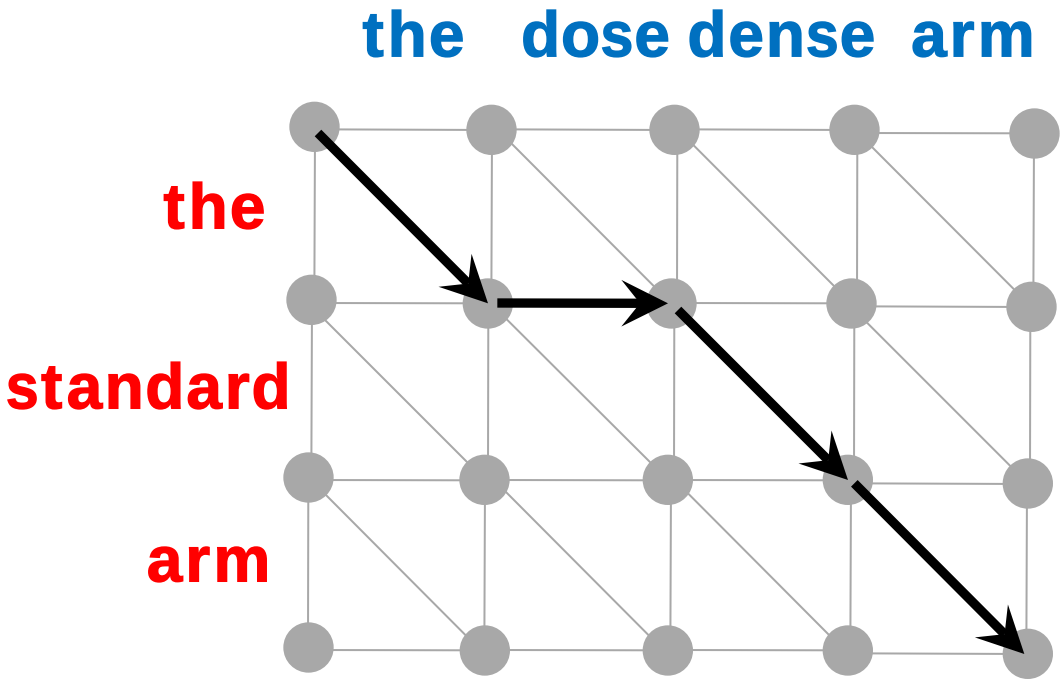
Median dose intensity was 99% for the standard arm and 182% for the dose dense arm.

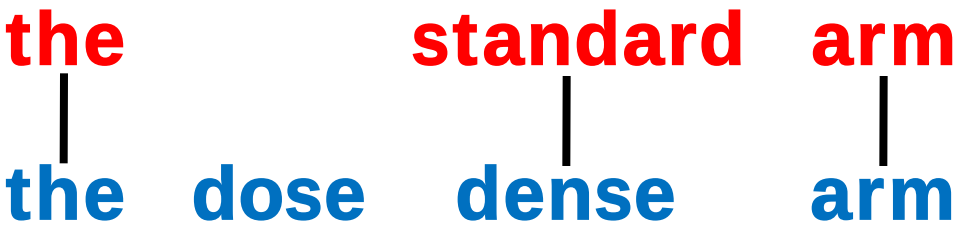
99%	for	the		standard	arm
182%	for	the	dose	dense	arm



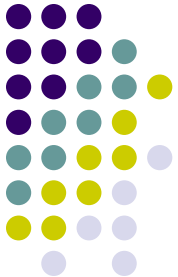
専門用語の並列構造解析

DPマッチングによる並列句のアラインメント





埋め込みを伴う並列句の解析

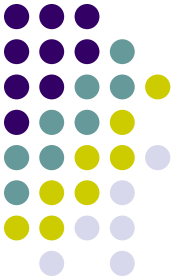


- DPマッチングに基づく手法では，複数の並列構造が入れ子になった表現（並列句の埋め込み）の解析が困難
- 入れ子になった並列句同士は交差してはならない（文法的制約）

“**((Median times to progression) and (median survival times))**
were
((((6.1 months) and (8.9 months)) in group A) and
(((7.2 months) and (9.5 months)) in group B))”

文法制約とDPマッチング法の同時適用

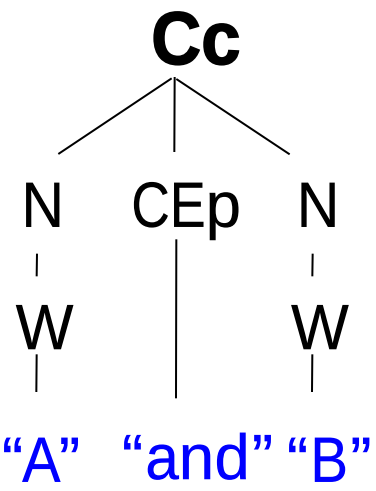
[Hara, Shimbo, Matsumoto 08]

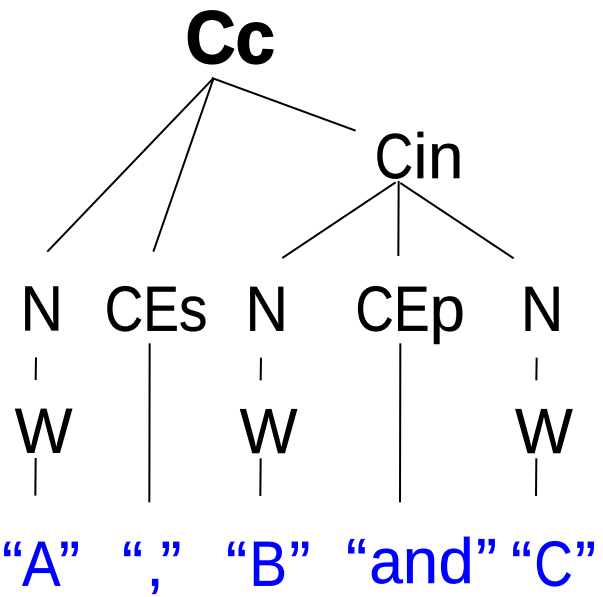


1. アラインメント用スコアの自動学習
   - CRFに基づく素性学習 [Shimbo et al 07]
2. 並列表現を解析するための文脈自由文法規則の記述（文法制約の記述）
3. 文法制約を保持したアラインメントスコアの自動学習
   - 文脈自由文法の構文解析アルゴリズムとアラインメントスコア学習の同時適用

並列構造のための文法規則 (1)

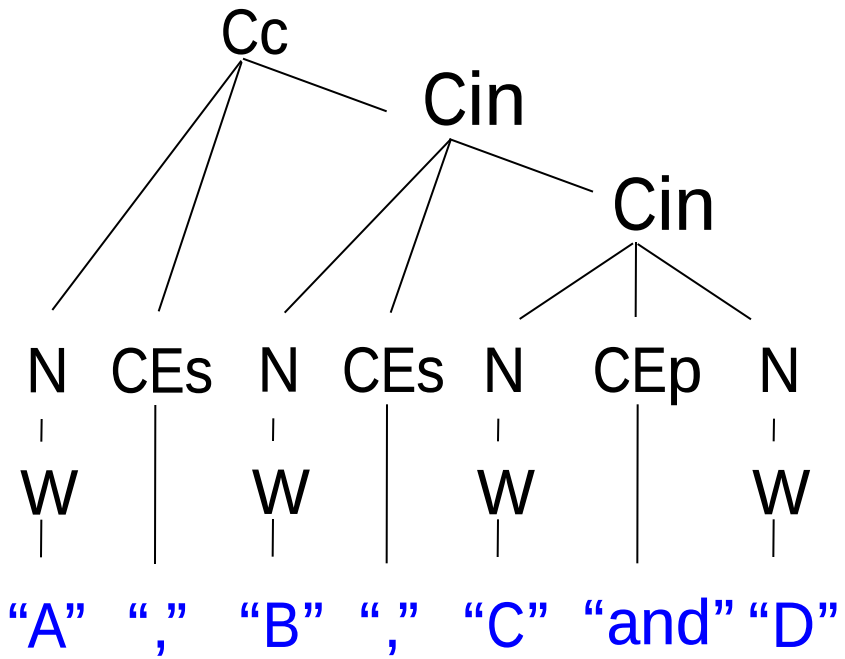
- Cc → {N, Cc} CEp {N, Cc}
- Cc → {N, Cc} CEs Cin





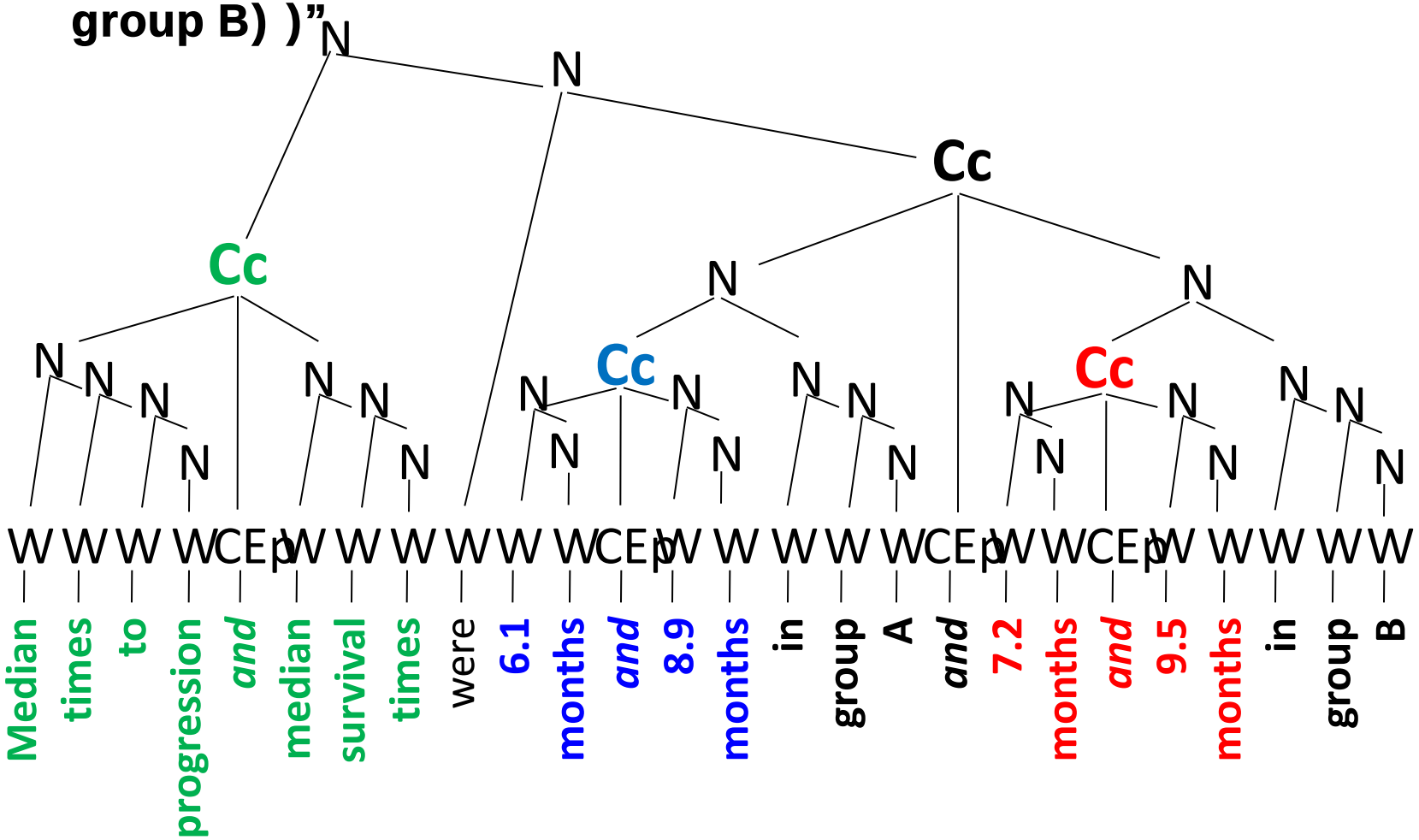
並列構造のための文法規則 (2)

- Cin → {N, Cc} CEp {N, Cc}
- Cin → {N, Cc} CEs Cin



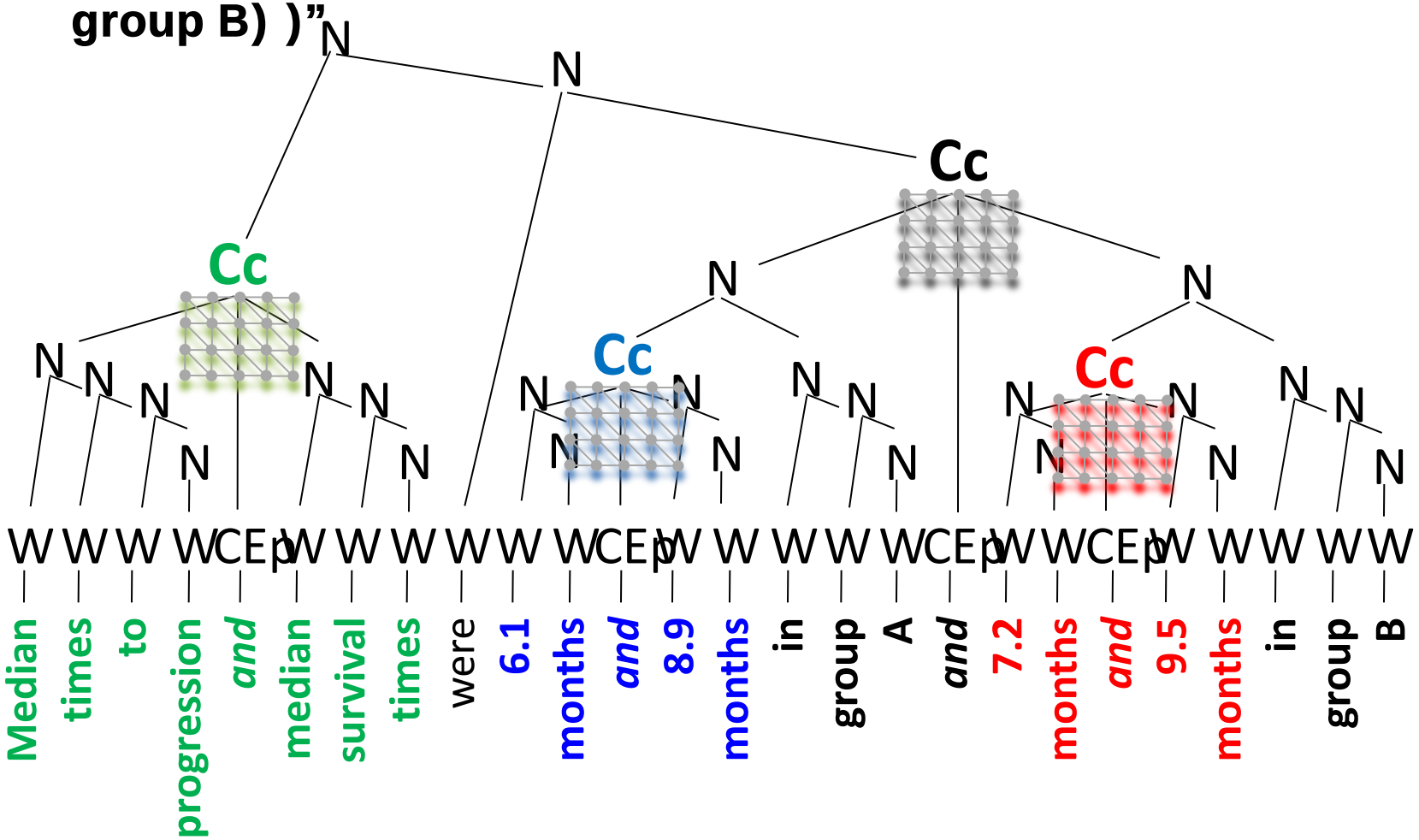
埋め込まれた並列構造は句構造木によって表現できる

“((Median times to progression) and (median survival times)) were ((((6.1 months) and (8.9 months)) in group A) and (((7.2 months) and (9.5 months)) in group B))”



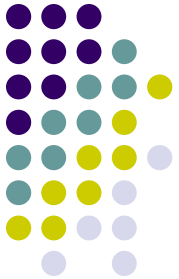
木の総合スコアは，文法制約を満たすという条件のもとで各並列構造のスコアの和

“((Median times to progression) and (median survival times)) were ((((6.1 months) and (8.9 months)) in group A) and (((7.2 months) and (9.5 months)) in group B))”



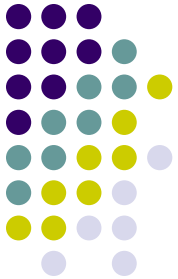
他手法との比較実験

	精度	再現率	F-値
Proposed	**56.8%**	**54.5%**	**55.6**
[Charniak & Johnson 05]	45.6%	42.9%	44.2
[Bikel 04]	43.9%	44.4%	44.1



専門用語の意味クラス推定

本研究の目的＝シソーラス拡張



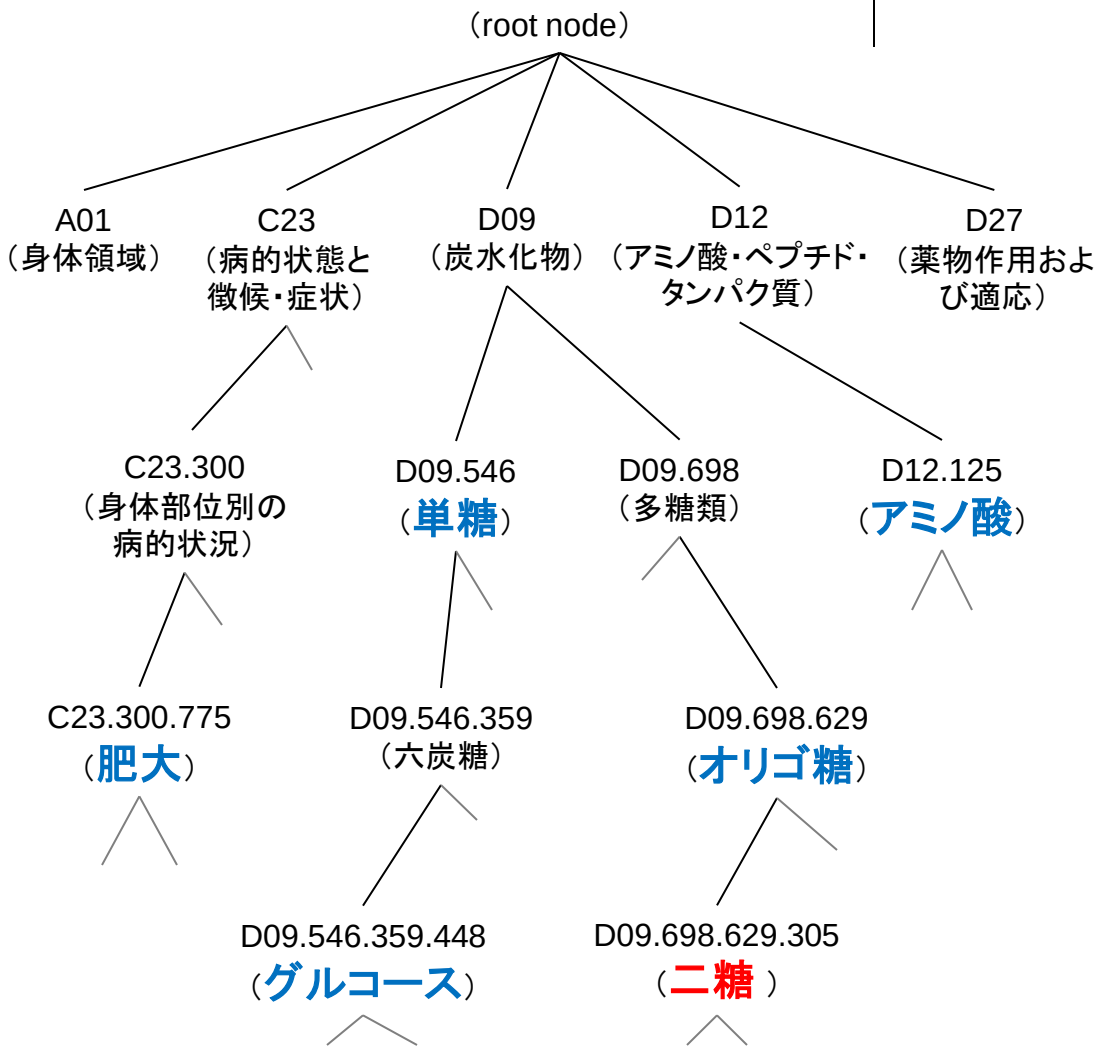
- 新規登録対象の専門用語（クエリ）に対して，類似度が高い順に登録済の専門用語をランク付けし，提示するシステムの構築
- シソーラス辞書の編集者は，ランキング結果を参考に，新しい専門用語をシソーラス辞書にマッピングすることができる

シソーラス拡張の例:専門用語「二糖」をシソーラスに登録する.

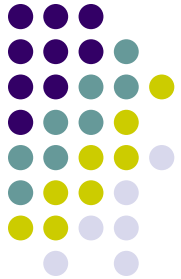
クエリ: 二糖

システム出力:

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



実験とその結果

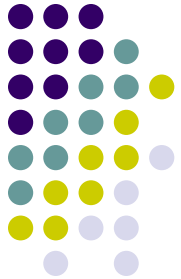


- 専門文書から対象とする語の文脈情報を抽出して用語の隣接グラフを作成し，グラフ構造を用いて用語間の類似度を算出する手法を提案
- 雑誌「蛋白質・核酸・酵素」を実験データとして用い，そこに登場する専門用語をライフサイエンス辞書のシソーラスにマッピングすることを目的とする実験を行った.
- 文書での出現頻度が少ない専門用語の場合には，提案するラプラシアン拡散カーネル行列を用いた手法が高い精度を示すことがわかった

来年度計画

- 専門用語辞書システムの開発
  - 機能と内部情報の拡張
    - 用語の内部構造の記述（約2000語）
    - 同義語・同一語の異表記の情報のタグ付け機能の簡素化，一覧表示
- 専門用語解析技術の開発
  - 専門用語の内部構造の自動解析手法の高性能化
  - 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同定と意味クラス推定の高性能化
  - 新規の専門用語をシソーラスへ登録する支援ツールの構築．意味クラス推定機能との連携

今年度の成果



- Mamoru Komachi, Taku Kudo, Masashi Shimbo, Yuji Matsumoto, “Graph-based Analysis of Semantic Drift in Espresso-like Bootstrapping Algorithms,” Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008), pp.1011-1020, October 2008.
- Kazuo Hara, “Classifying Narrative Patient Records without Any External Resources,” Proceedings of the Second i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, November 2008.
- 原 一夫, 新保 仁, 大熊 秀治, 松本 裕治, “GENIAコーパスからのネスト並列句同定,” 情報処理学会研究報告, 自然言語処理研究会, 2008-NL-187, pp.53-58, September 2008.
- 大熊秀治, 原一夫, 新保仁, 松本裕治, “機械学習と系列アラインメントを応用した日本語並列句解析,” 人工知能学会全国大会（第22回）論文集, 1H1-03, June 2008.
- 鈴木郁美, 原一夫, 新保仁, 松本裕治, “グラフを用いたバイオ医療専門用語の類義語推定,”情報処理学会研究報告, 自然言語処理研究会, 2008-NL-189, pp.65-70, January 2009.