

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

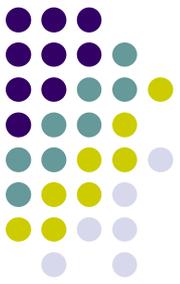
松本裕治, 新保仁, 浅原正幸, 原一夫

- 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



今年度の成果目標

- 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(現在の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



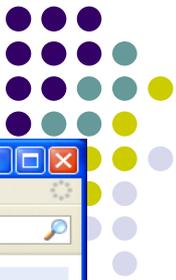
辞書(専門用語管理)システム



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, 品詞などの用語の基本情報
 - 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(E) 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://dahlianaist.jp/cradle/ Google

 **CRADLE--茶釜辞書管理システム** 

日本語辞書 汉语辞典 matsu | [Preference](#) | [User list](#) | [Logout](#)

単語属性



ID	=	<input type="text"/>	単語	=	<input type="text"/>	結核	<input type="text"/>
読み	=	<input type="text"/>	発音	=	<input type="text"/>		
品詞	=	<input type="text"/>	活用型	=	<input type="text"/>		
活用形	=	<input type="text"/>	Base	=	<input type="text"/>		
辞書	or	<input type="text" value="NAIST-jdic-20080707"/> <input type="text" value="WebLSD-200804"/>	文字数	=	<input type="text"/>		
更新時間	<=	<input type="text"/>	状態	=	<input type="text"/>		
親概念日本語表記	=	<input type="text"/>	新規者	=	<input type="text"/>		
自動参照先の日本語コード	=	<input type="text"/>	更新者	=	<input type="text"/>		
階層の深さ	=	<input type="text"/>	親概念英語表記	=	<input type="text"/>		
自動参照先ID	=	<input type="text"/>	日本語コード	=	<input type="text"/>		
自動参照先表記	=	<input type="text"/>	自動参照先の日本語表記	=	<input type="text"/>		
親概念ID	=	<input type="text"/>	ツリー番地	=	<input type="text"/>		
			ツリー日本語	=	<input type="text"/>		
			ツリー英語	=	<input type="text"/>		

複合語属性



内部表記	include	<input type="text"/>	内部読み	include	<input type="text"/>
内部POS	include	<input type="text"/>	状態	=	<input type="text"/>
更新者	=	<input type="text"/>	更新時間	<=	<input type="text"/>
枝の種類	=	<input type="text" value="P"/>	縮退文字の位置	=	<input type="text"/>

検索結果(結核)



CRADLE--茶釜辞書管理システム

日本語辞書 汉语辞典

matsu | [Preference](#) | [User list](#) | [Logout](#)

表示件数: 30

<< Previous **1** 2 3 4 Next >>

条件:	単語=~結核, 辞書:(WebLSD-200804)										93 Htrs
	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	キョウ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(E) 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://dahlianaist.jp/cradle/jp/show/2197835

CRADLE--茶釜辞書管理システム

matsu | [Preference](#) | [User list](#) | [Logout](#)

日本語辞書 汉语辞典

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ド		

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

```

graph TD
    A[広範囲薬剤耐性結核] --> B[広範囲薬剤耐性]
    A --> C[結核]
    B --> D[広範囲]
    B --> E[薬剤耐性]
    D --> F[広]
    D --> G[範囲]
    E --> H[薬剤]
    E --> I[耐性]
    
```

7

複合語内部構造のアノテーション



Cradle--ChaSen Dictionary Management System - Mozilla Firefox

ファイル(E) 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://dahlianaist.jp/cradle/jp/show/2216932

CRADLE--茶筌辞書管理システム

日本語辞書 | 汉语辞典

matsu | [Preference](#) | [User list](#) | [Logout](#)

内部構造: 結 核 性 前 立 腺 炎

Top: 結核性前立腺炎(2216932) ==> [結核性前立腺炎0]

結核性 前立腺炎

選択	更新	ID	単語	読み	品詞	構造	辞書	状態
<input checked="" type="radio"/>	<input type="radio"/>	2198905	結核性	ケツカクセイ	名詞-一般	<input type="button" value="show"/>	<input type="checkbox"/>	NEW
DUMMY	更新	ID	単語	読み	品詞	構造	辞書	状態

選択	更新	ID	単語	読み	品詞	構造	辞書	状態
<input checked="" type="radio"/>	<input type="radio"/>	2240106	前立腺炎	ゼンリツセンエン	名詞-一般		<input type="checkbox"/>	NEW
DUMMY	更新	ID	単語	読み	品詞	構造	辞書	状態

複合語の内部構造情報付与結果



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://dahlianaist.jp/cradle/jp/show/2216932

CRADLE--茶筌辞書管理システム

matsu | [Preference](#) | [User list](#) | [Logout](#)

ツリー構造

構造詳細

状態	NEW
備考	
更新者	matsu
更新时间	2009-01-26 22:31:51

[編集](#) [削除](#)

系列

構成	結核性, 前立腺炎
枝の種類	D
縮退文字 の位置	

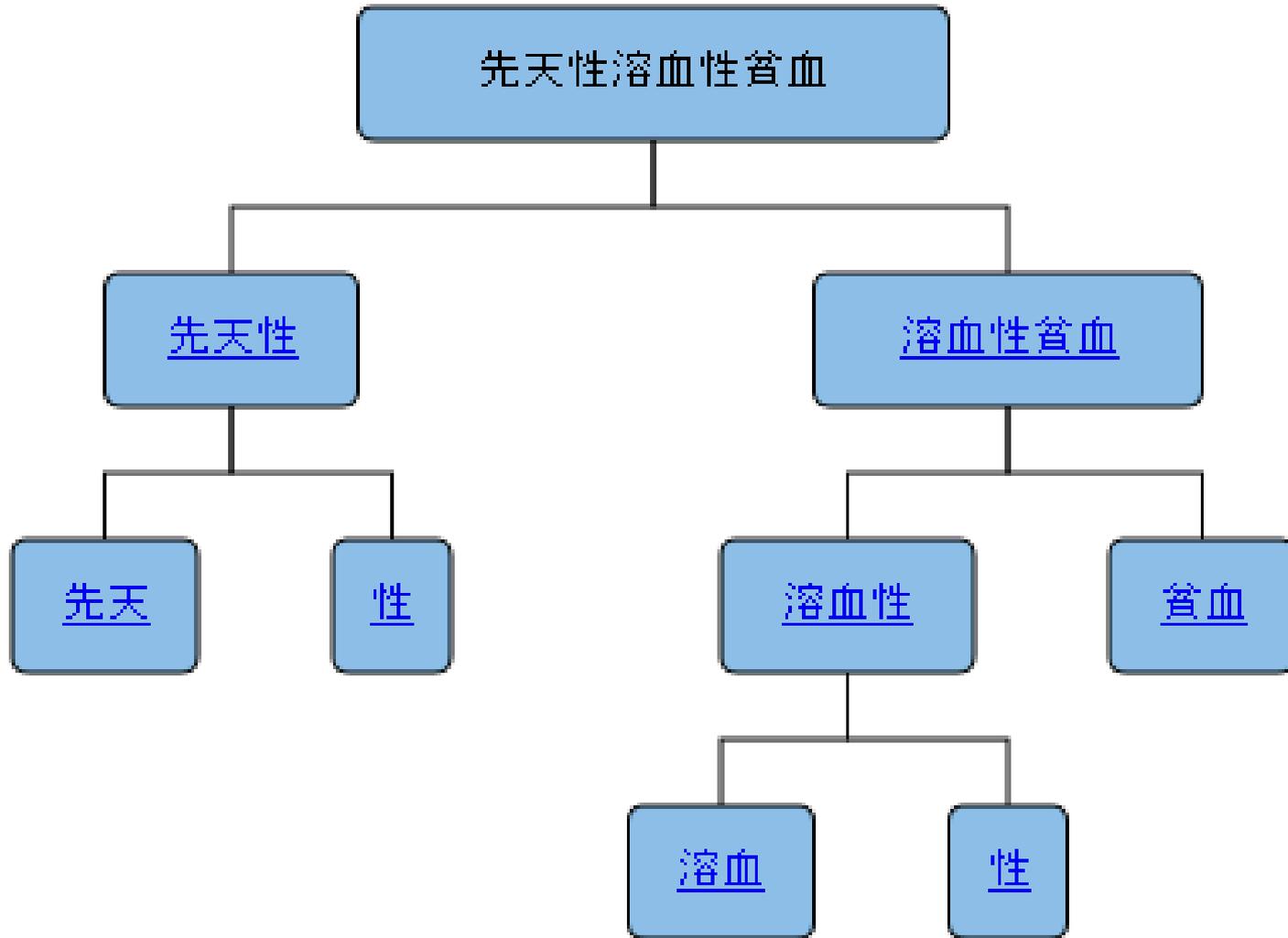
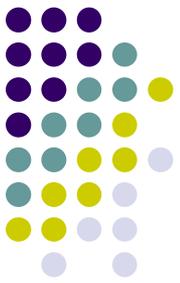
```
graph TD; A[結核性前立腺炎] --> B[結核性]; A --> C[前立腺炎]; B --> D[結核]; B --> E[性]; C --> F[前立腺]; C --> G[炎];
```

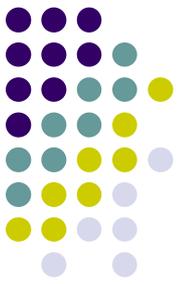


専門用語解析技術

1. 複合語の内部構造解析

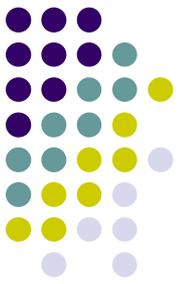
複合語の内部構造





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 <= 1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

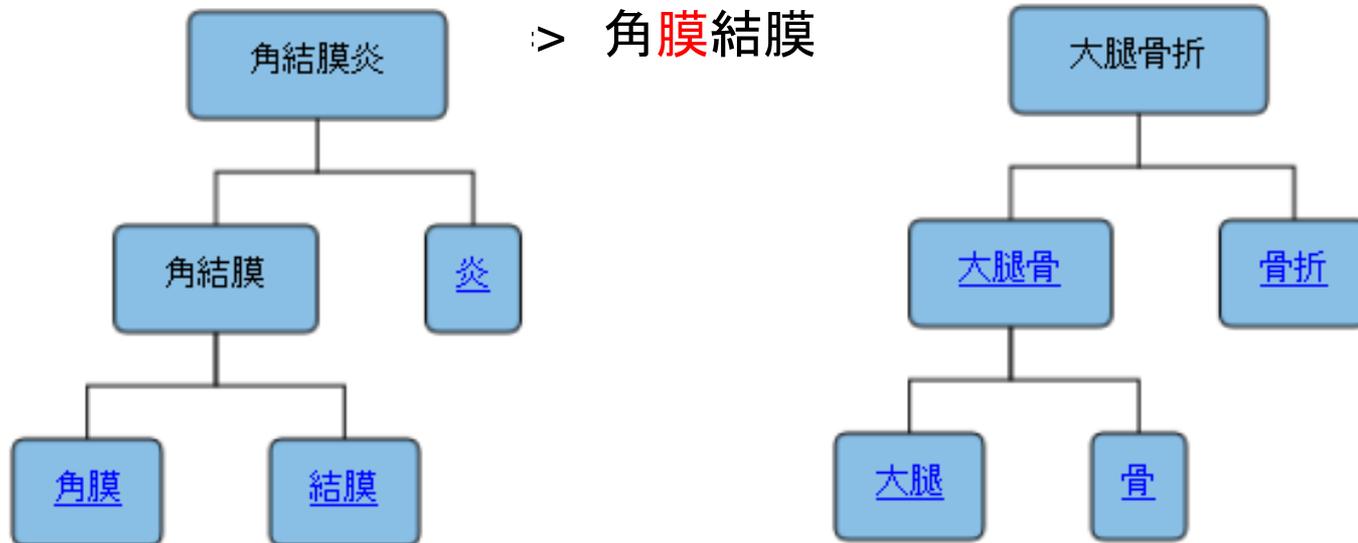
- 縮退

- 縮退する文字の位置

- End + Begin

- 大腿骨 + 骨折 => 大腿骨骨折

- End + End

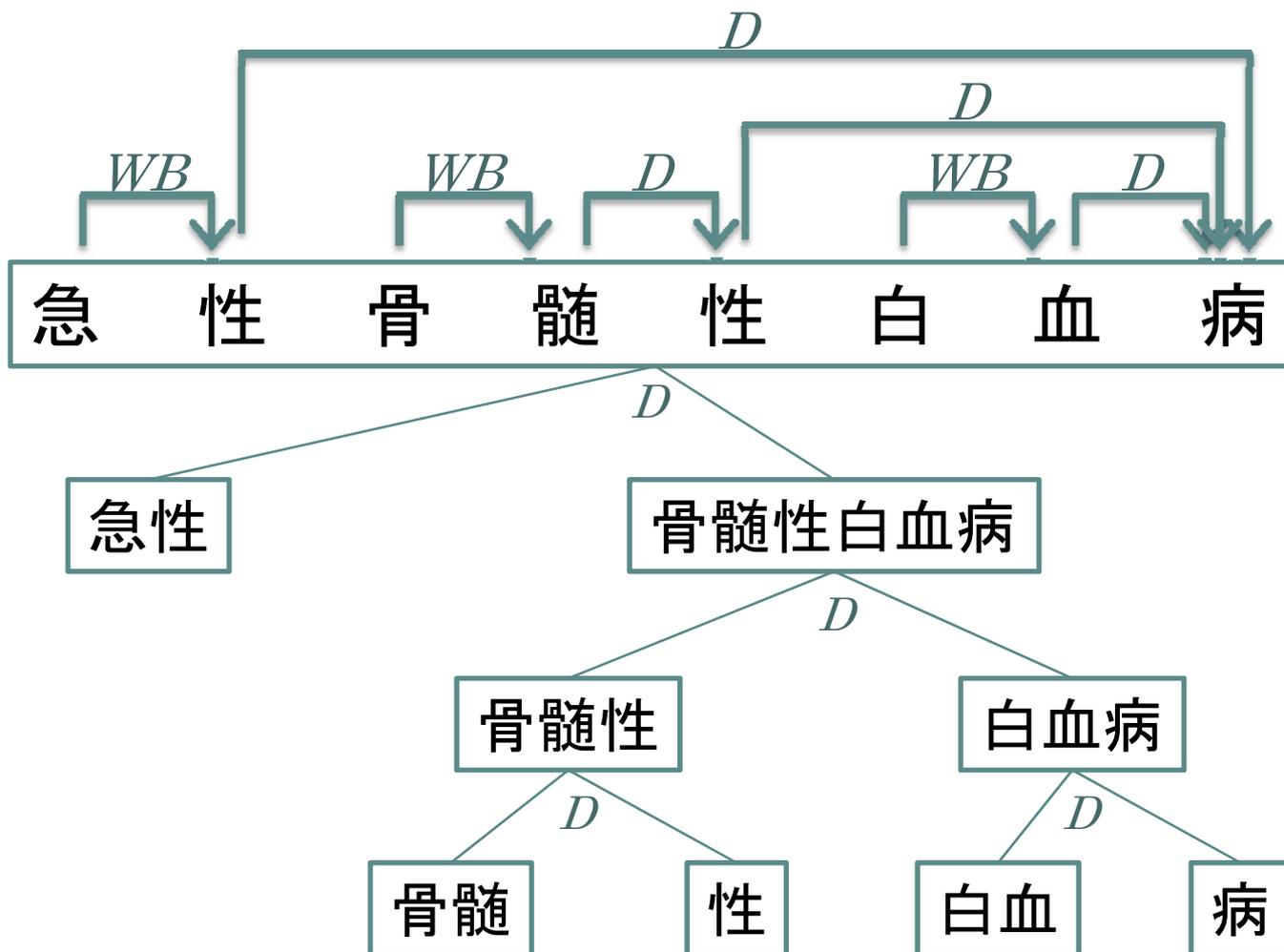


文字単位の係り受けによって記述

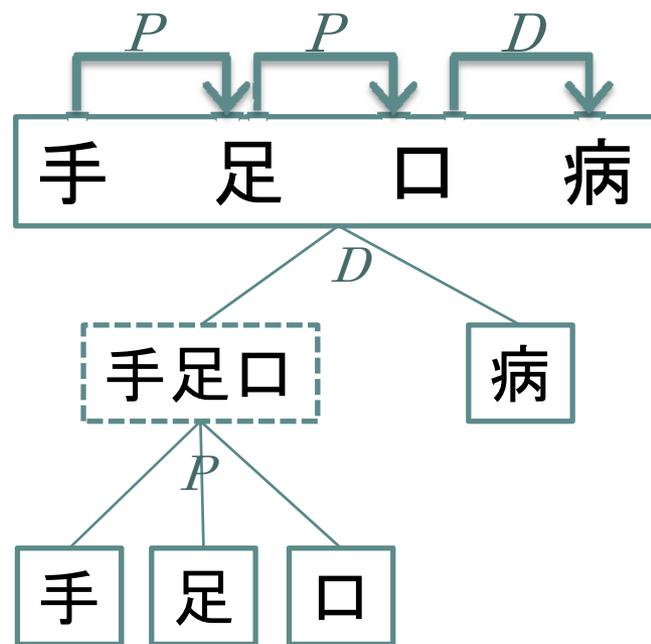
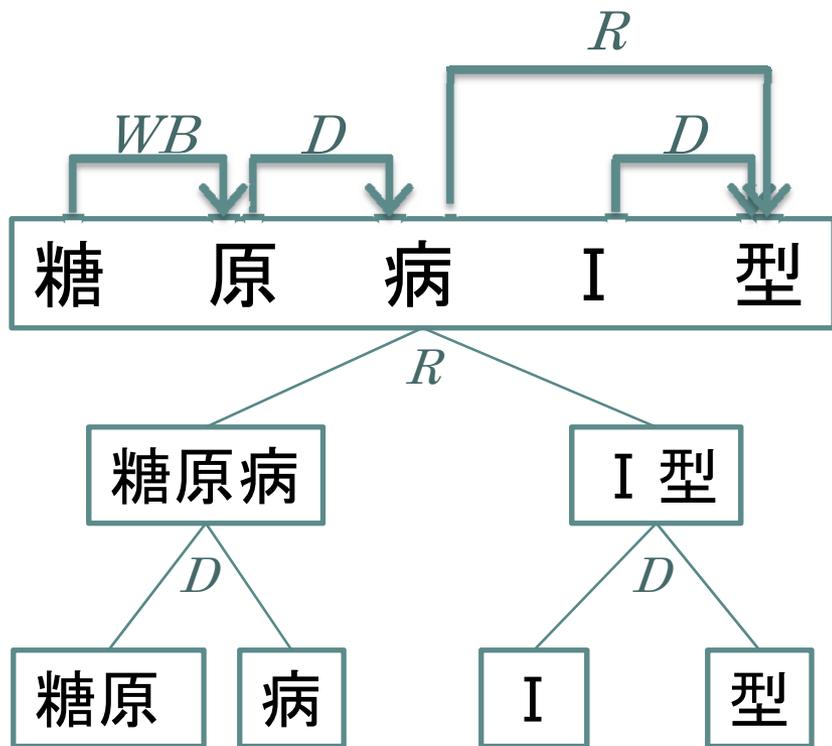


- 係り受けの種類(合計6種類)
 - 形態素を構成する係り受け
 - 形態素の先頭(WB)、中間(WI)
 - 形態素間の係り受け(前述の4種類)

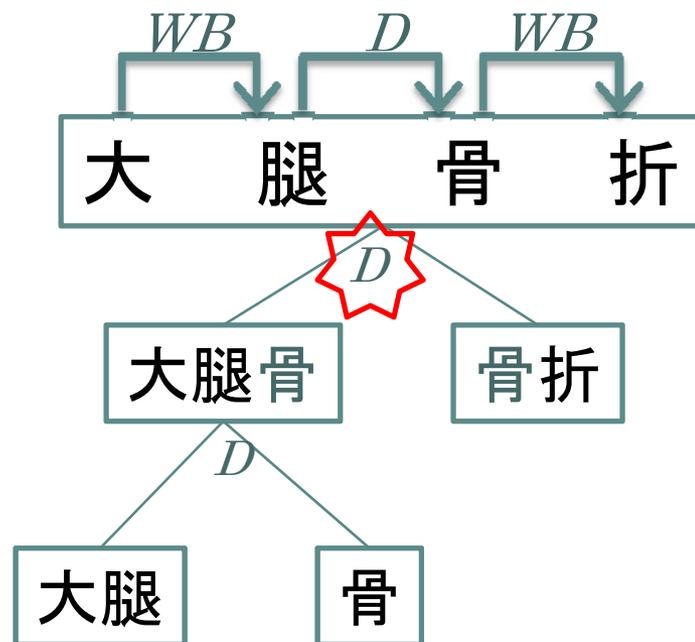
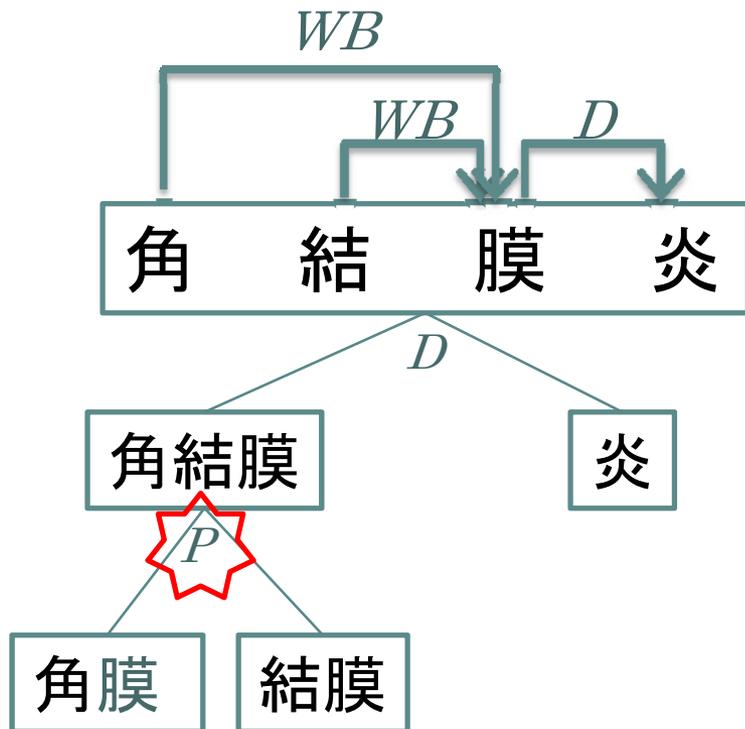
文字単位の係り受けによる表現



文字単位の係り受けによる表現



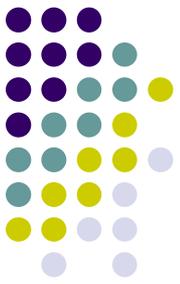
文字単位の係り受けによる表現





專門用語解析技術

2. 並列構造解析



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
 - = ???

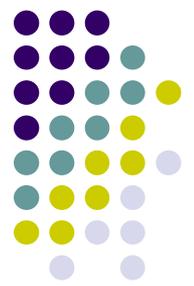
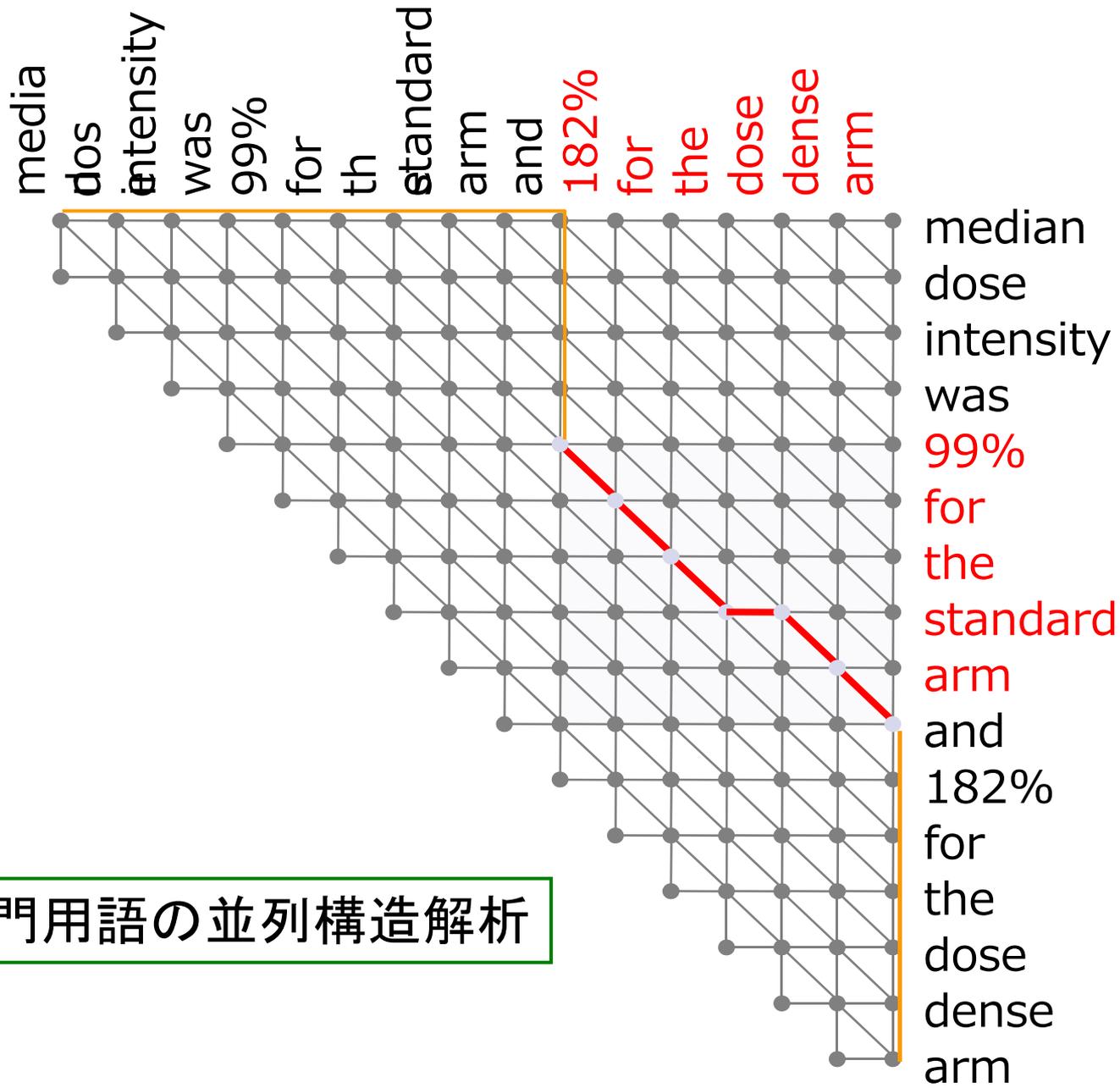
文中の類似部分検出による 並列句同定

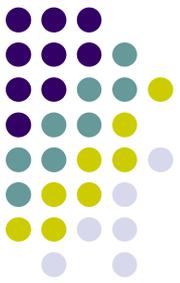


Median dose intensity was 99% for the standard arm and 182% for the dose dense arm.

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

専門用語の並列構造解析





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))”

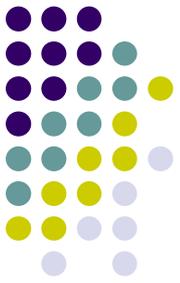
文法制約とDPマッチング法の同時適用

[Hara, Shimbo, Okuma, Matsumoto 09]

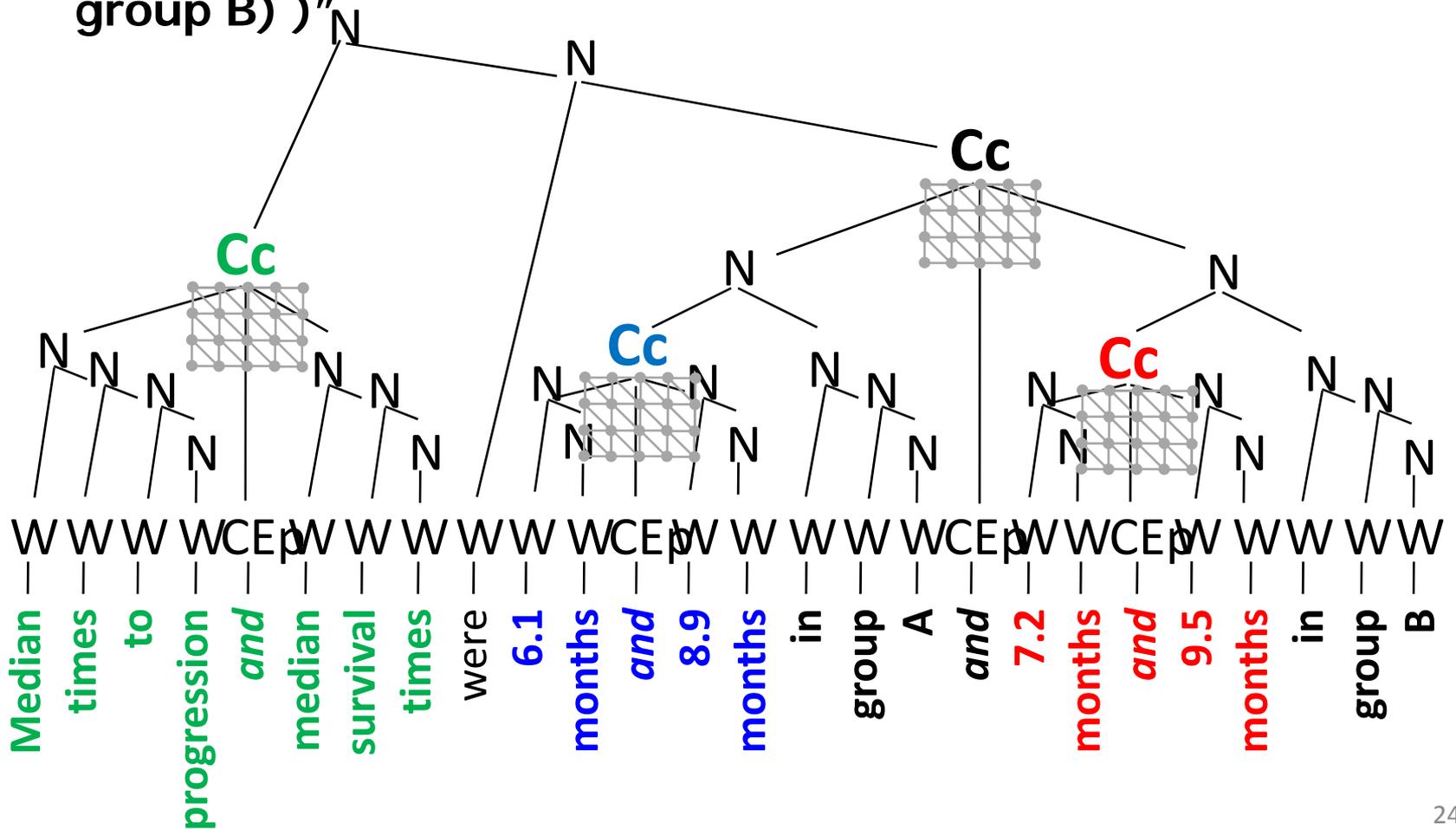


1. アラインメント用スコアの自動学習
 - CRFに基づく素性学習 [Shimbo et al 07]
2. 並列表現を解析するための文脈自由文法規則の記述 (文法制約の記述)
3. 文法制約を保持したアラインメントスコアの自動学習
 - 文脈自由文法の構文解析アルゴリズムとアラインメントスコア学習の同時適用

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	60.2%	55.6%	57.8
[Charniak & Johnson 05]	49.0%	46.8%	47.9
[Bikel 04]	45.6%	46.1%	45.8



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、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

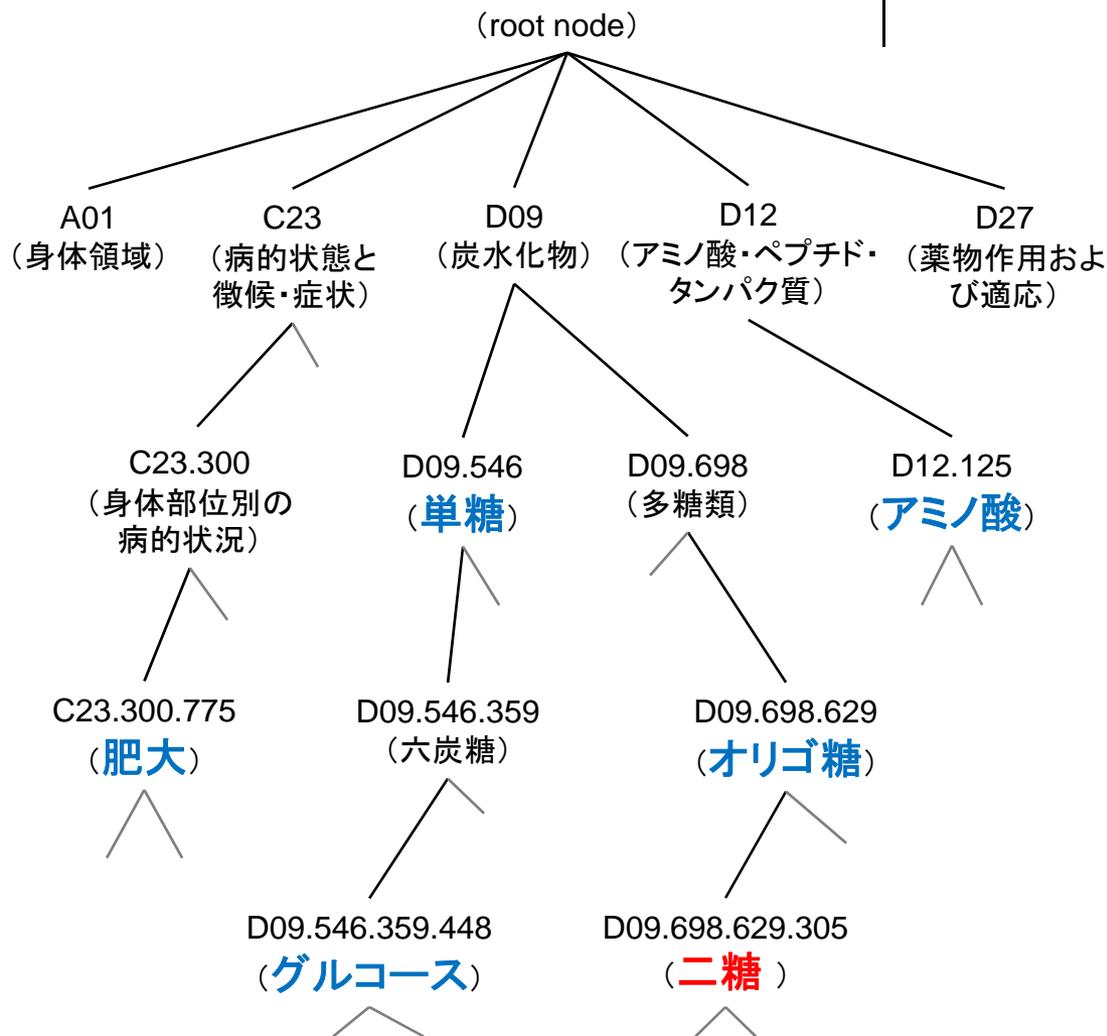
シソーラス拡張の例: 専門用語「二糖」と類似の語を表示

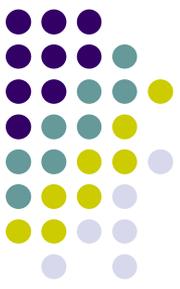


クエリ: 二糖

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





進捗

- 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実験を行った
- 文書での出現頻度が少ない(出現頻度2回以下)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従来のコサイン尺度を用いた類似度尺度よりも良好な結果を示すことがわかった
- 現在、類似度上位語のシソーラス内の位置を表示するインタフェースを構築中



今年度の成果目標(再掲)

- 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(現在の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