不均一な情報の再構成による事典コンテンツの構築

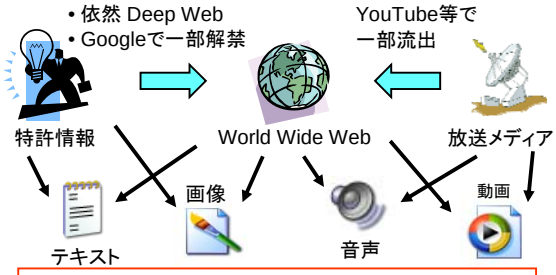
～新たな価値の創造を目指して～

藤井　敦
筑波大学
大学院図書館情報メディア研究科
2007.12.20

1

研究の動機

多種多様な情報が爆発的に増えている



不均一な情報に一定の秩序を与えたい

2

Web等における情報の多様性・不均一性

本講演の中心的な話題

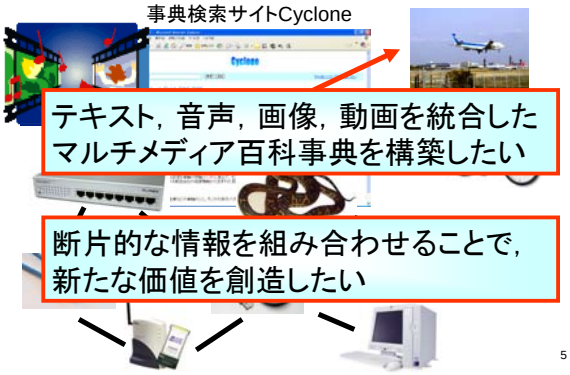
メディア	テキスト ／ 音声 ／ 画像 ／ 動画
専門性	専門的（論文，特許）／ 大衆的（CGM）
言語	日本語 ／ 外国語（英中韓蒙）
目的	情報伝達（新聞，論文）／ 娯楽・芸術（小説）
感情	客観的（新聞，論文）／ 主観的（意見，批評）
玉石	有益 ／ 無益（有害）

3

研究の構想

4

研究の構想



5

事典検索サイトCycloneの概要

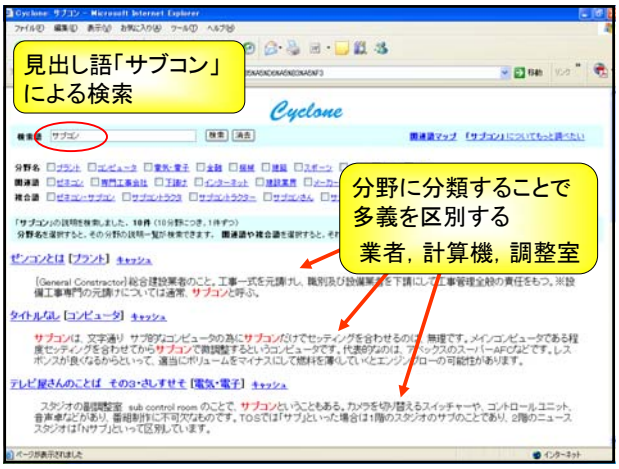
6

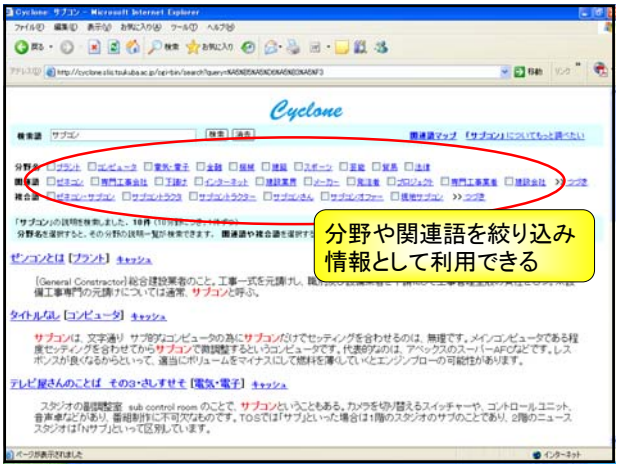
事典検索サイト Cyclone

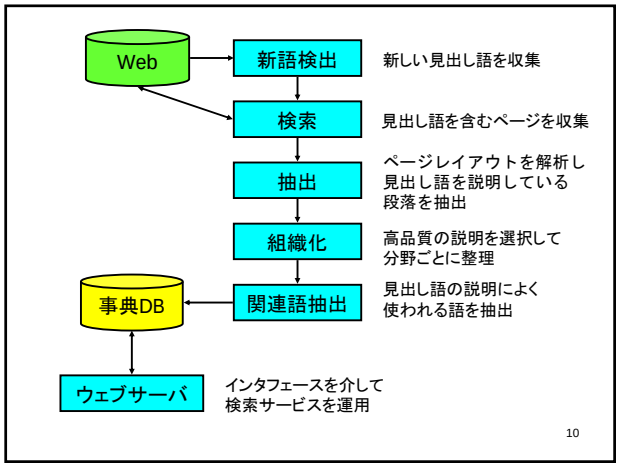
http://cyclone.slis.tsukuba.ac.jp/

- コンテンツの構築　→コンテンツとしての価値を追求
  - Webから見出し語と説明を収集し，体系化する
  - 現在の見出し語数：約75万語
- 多様な検索機能　→ サービスとしての価値を追求
  - 見出し語，同義語，関連語による検索
  - 質問文による検索
  - 鳥瞰による検索：関連語グラフの可視化

7







10

分野の分類

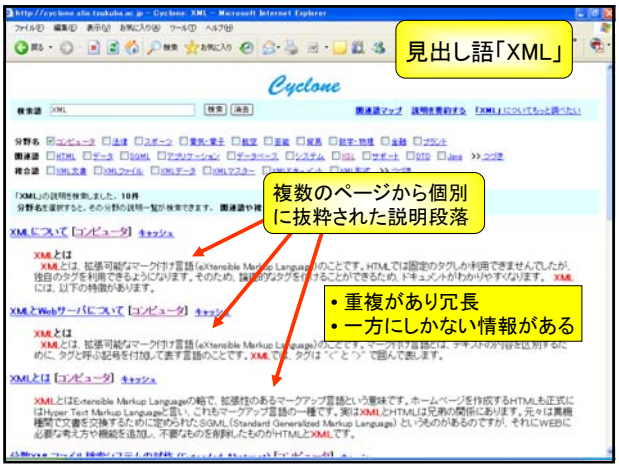
- 説明段落を分野に分類する
  - 分野（カテゴリ）は事前に決めておく
- 以下の言語資源から抽出した語の頻度分布を用いて統計モデルを学習
  - 機械翻訳用の専門用語辞書（20分野）
    - コンピュータ，バイオ，機械工学，法律など
  - 新聞記事テキスト
    - 芸能面，スポーツ面

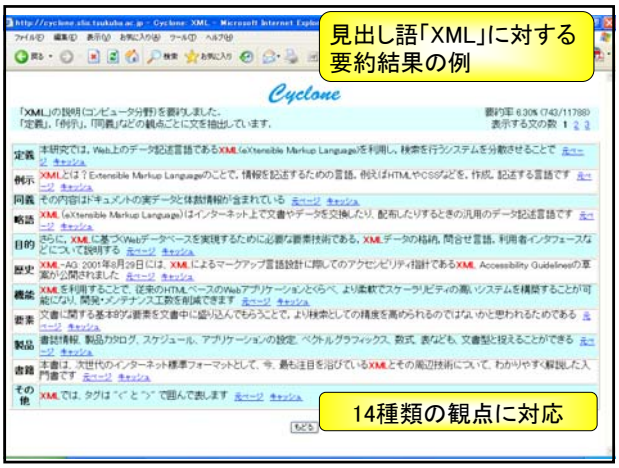
11

組織化によって区別できる曖昧性

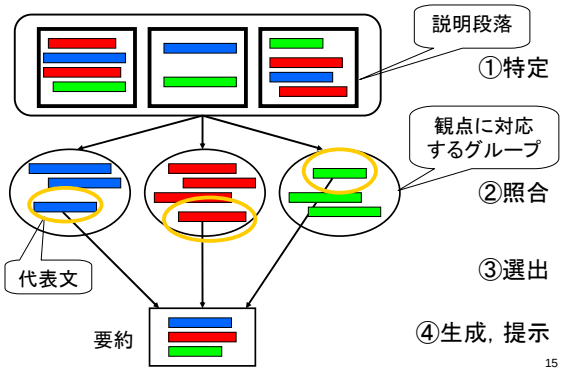
- 多義性：基本的な意味は同じ
  例：「プロービング」（検査）
  - データ（コンピュータ分野），計測行為（医療）
- 同音意義：異なる語が偶然同じ発音をもつ
  例：「ハブ」（装置，ヘビ，中心）
- 頭字語：略語が偶然同じ形をもつ
  例：「NLP」（自然言語処理，夜間離着陸訓練）
- 観点
  例：「クレオソート」（薬品，防腐剤）

12





要約処理の概要



15

多様な検索手法

- 見出し語による検索（通常の検索）
- 関連語による検索（絞り込み，芋づる検索）
- 文字列一致による検索
- 同義語による検索
- 本文一致による検索
- 質問応答

ユーザの入力が見出し語にない場合に，代替案を提示するための手段

- 関連語グラフによる検索

16

文字列一致による検索

- ユーザ入力と表層的に（文字列のレベルで）似ている見出し語を検索する

- 前方・後方一致
- 部分一致
  - 異表記（インターフェイス，インタフェース）
  - 入力誤り（人口知能，人工知能）
- 略語
  - 住民基本台帳ネットワークシステム，住基ネット

17

部分一致検索

1. 全見出し語を文字Nグラムで索引付け
   例：インタフェース（N=2，バイグラム）
   イン，ンタ，タフ，フェ，ェー，ース
2. ユーザ入力もNグラムに分割
3. 多くのNグラムを共有する見出し語を検索
4. 長さが近い見出し語を優先

※ DPマッチングでは実時間応答に耐えない

18

同義語による検索

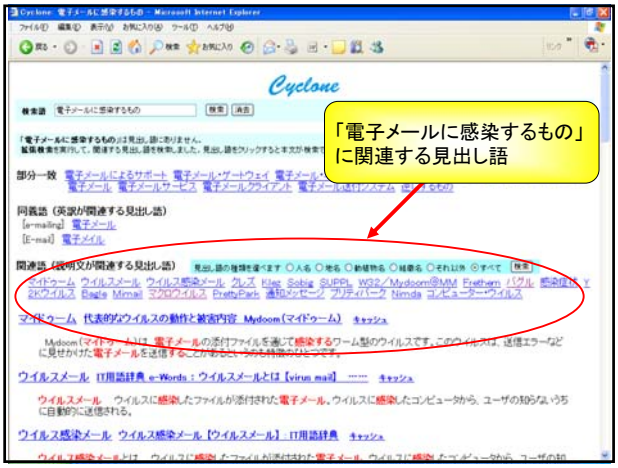
- （表層的に似ていなくても）意味が似ている見出し語を検索
  - 同義語を厳密に定義することは困難
  - 同じ英訳に対応する語を同義語と見なす
- 複数の翻訳手法を併用
  - 対訳辞書による翻訳
  - カタカナ語を元の英訳に対応付ける（翻字）
    - レゾナンス → resonance
    - マイニング → mining

19

本文一致による検索

- コンテンツの全文検索によって事典を逆引き
  - ユーザが入力した語を本文に多く含む見出し語を検索
- 2つの効果がある
  - ユーザが入力した言葉が他の見出し語に対する段落で説明されている場合に有効
  - 自然言語による質問文の入力が可能
    - 電子メールに感染するソフト → マクロウイルス

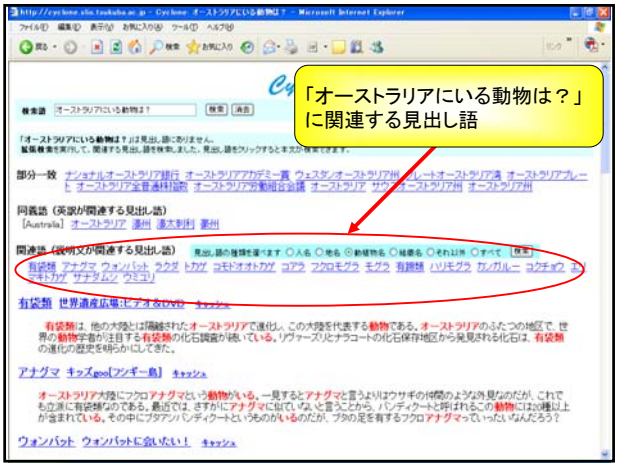
20



質問応答

- 「本文一致による検索」を拡張
- 質問表現によって回答すべき言葉の「種類」を特定する
  - ～は誰ですか？ → 人名
  - ～はどこですか？ → 地名
  - ～の動物は？ → 動物名
- 人名辞書，地名辞書などを用いて見出し語に種類を付与しておく

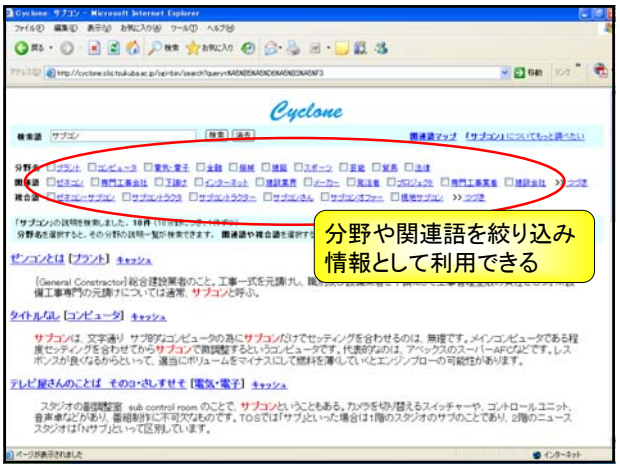
22

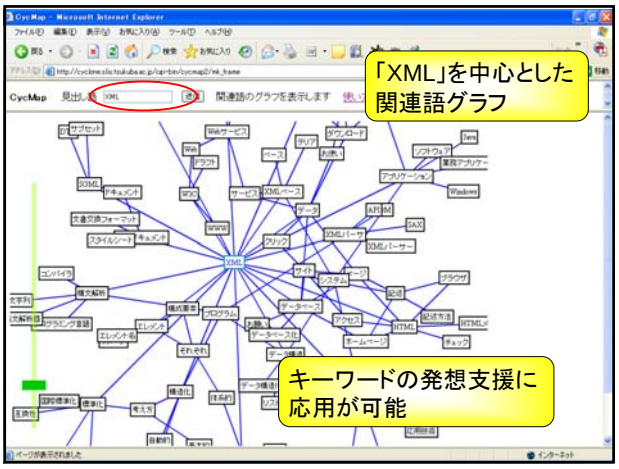


関連語グラフによる検索

- 関連語抽出によってグラフが構成される
  - 言葉 → ノード
  - 言葉の関係 → エッジ
- グラフを可視化することで，ある語を手がかりにして周辺の語を探索することができる
- 「テキスト」と「グラフ」を横断的に検索可能

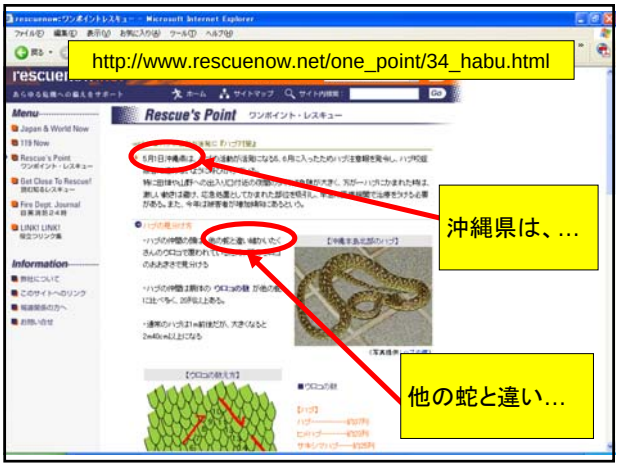
24

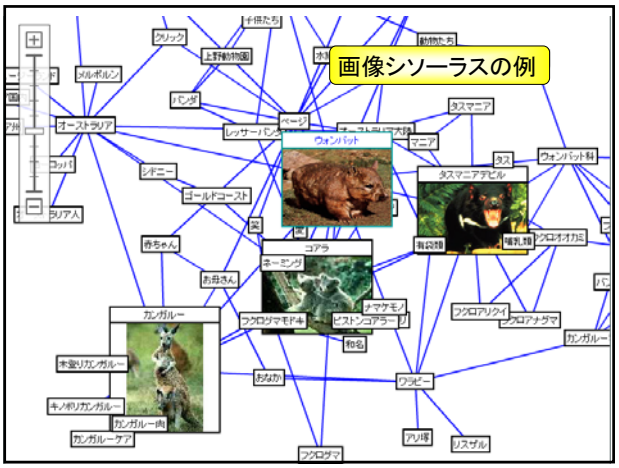




テキストと画像の対応付け

27





テキストと音声・動画の対応付け

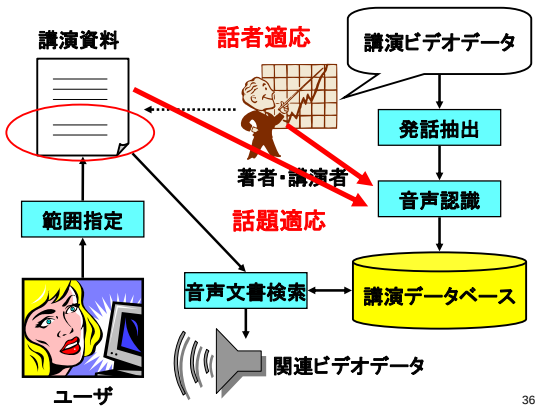
オンデマンド講演システム

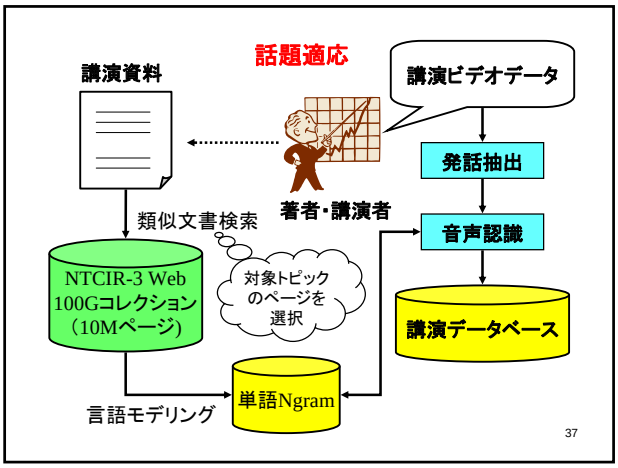
- 講演ビデオデータを対象にして，要求に応じた内容を視聴するオンデマンドシステムを実現
- 講演資料（テキスト）を閲覧しながら，特定のビデオ内容（音声，動画像）を選択的に視聴することが可能
- 音声認識と情報検索技術を統合し，高精度の音声文書検索を実現

資料と講演の相違点

資料（書き言葉）	講演（話し言葉）
ランダムアクセス可能 ・構造（章立て），表層（字種） ・重要箇所の特定が容易	逐次アクセスが原則 ・早送り，巻き戻しでは不十分 ・書き起こしは読みにくい
比較的簡潔 ・ページ数の制限	**分かりやすい** ・説明が詳しい ・適度に冗長

書き言葉と話し言葉の長所を統合する





音声認識の確率モデル

単語列W　入力音声X　ベイズの定理

$$\arg\max_W P(W \mid X) = \arg\max_W \frac{P(X \mid W) \cdot P(W)}{P(X)}$$

$$= \arg\max_W P(X \mid W) \cdot P(W)$$

デコーダ（確率最大のWを探索）

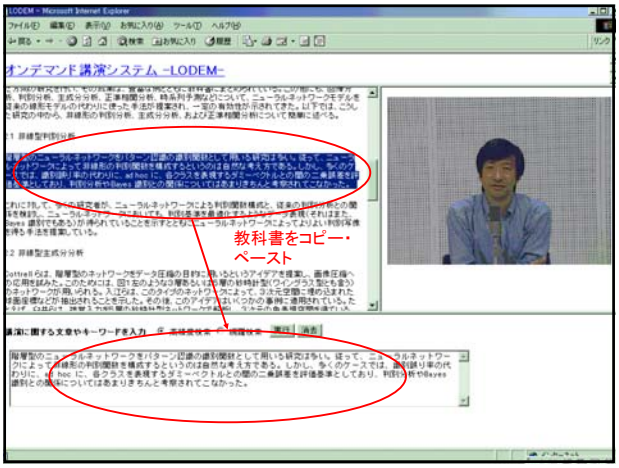
言語モデル（単語列の出現確率）

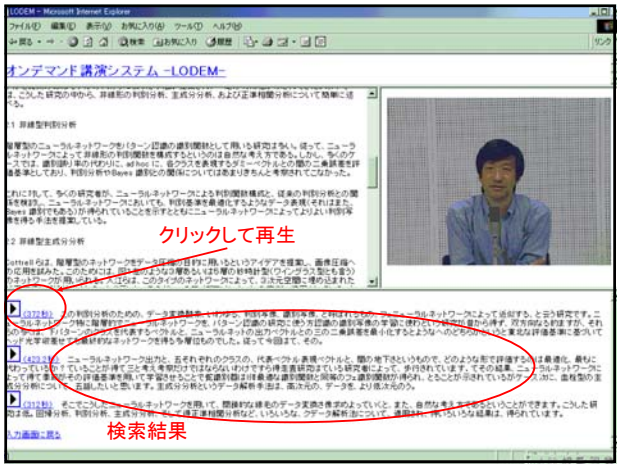
音響モデル（単語列が話者によってどのような音声になるか）

話題適応

話者適応

38





オンデマンド講演システムの評価

- データ：放送大学 45分×5講義，検索質問 127件
- 音声認識の語彙サイズ：10万語

	適応なし	音響	言語	両方
単語誤り率	57.5	47.3	47.6	39.2
再現率	36.9	41.3	44.0	48.3
精度	38.2	39.8	39.2	42.8
F値	37.6	40.5	41.5	45.4

41

Cycloneとの統合

- ある見出し語に関する講義ビデオのシーンを検索する
  - 放送大学の講義，話し言葉コーパスなど
  - 著作権の関係で一般公開はできない

42

特許情報からの辞典構築

43

Web以外の情報源はないのか？

- 特許情報の利用
  - 特許には高度な発明に関する言葉や説明がある
  - Webでは（ほとんど）見つからない用語がある
    - 公開特許公報 1993～2005年発行分
- NEDO「産業技術研究助成事業」H17～20
  - 用語辞典・シソーラスの構築
  - 特許検索インタフェースの開発
    - 辞典 → 専門用語だけ（事件などは含まない）

44

Webには（ほとんど）ないけど，特許にはある用語の例

ジルコニウムジクロリド，重合体成分，感光性平版印刷版，焼付定盤，絶縁基体，沃臭化銀乳剤，ハロゲン化銀乳剤，スルファモイル基，プラテンドラム，塩基プレカーサー，エチレン性不飽和単量体，トラッキング誤差信号，静電潜像保持体，スロットル弁開度，マゼンタカプラー

大抵はWebに掲載された特許が検索される

45

定義文抽出へのこだわり

- 定義文の典型
  - 「○○とは，～である」という文型
- しかし，定義文でない「とは」が多数存在する
  - ○○とは全く関係ない
  - ○○とは△△で連結している
  - これが○○とは恐れ入った
- こうした非定義文の「とは」を区別するための規則を体系化
  - 精度評価： 規則 82%，SVM 93%

46

Webでは引けない用語定義の例

- 感光性平版印刷版とは、一般に、適当な表面処理を施したアルミニウム、紙あるいはプラスチックなどの支持体の表面に、感光性化合物を含有する感光層を設けたものである。
- トラッキング誤差信号とは、対物レンズが光ディスクの半径方向に移動する場合、記録されたピットの中心で0となり、ピットの中心からずれるに従って、値が大きくなる信号である。
- 塩基プレカーサーとは、加熱下で塩基を遊離する化合物をいい、塩基と有機酸の塩等が挙げられる。塩基プレカーサーを構成している塩基としては、前記塩基で例示したものが好ましい。
- マゼンタカプラーとは、Ｎ－エチル－Ｎ－（β－メタンスルホンアミドエチル）－３－メチル－４－アミノアニリン硫酸塩（ＣＤ－３）との酸化的カップリングによって生成する色素が、メタノール中での極大吸収波長が500～600nmの範囲にあるカプラーをいい、

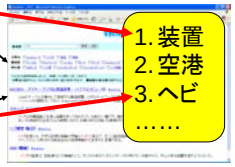
47

個人化への対応
「芋づる検索」における文脈の利用

LAN → ケーブル → ハブ
（装置）

毒蛇 → まむし → ハブ
（ヘビ）

現状のCycloneでは，いつ誰が到達しても提示される説明文章の順番が同じ



文脈（たどってきた語の連鎖）によって
提示する説明文章の順番を変更する

51