



# 人工知能特論II 第1回

二宮 崇<sub>1</sub>

# 講義の目的



# この講義の目的 (1/2)

- 最先端の自然言語構文解析技術

- 自然言語

- 社会において自然に発生して用いられている言語。  
[デジタル大辞泉]

- 主に人間同士で意志疎通を行うことを目的として、人間が日常、読み、書き、聞き、喋り、思考するための言語。

- 例：日本語、英語、ドイツ語、フランス語、...

- ⇔人工言語

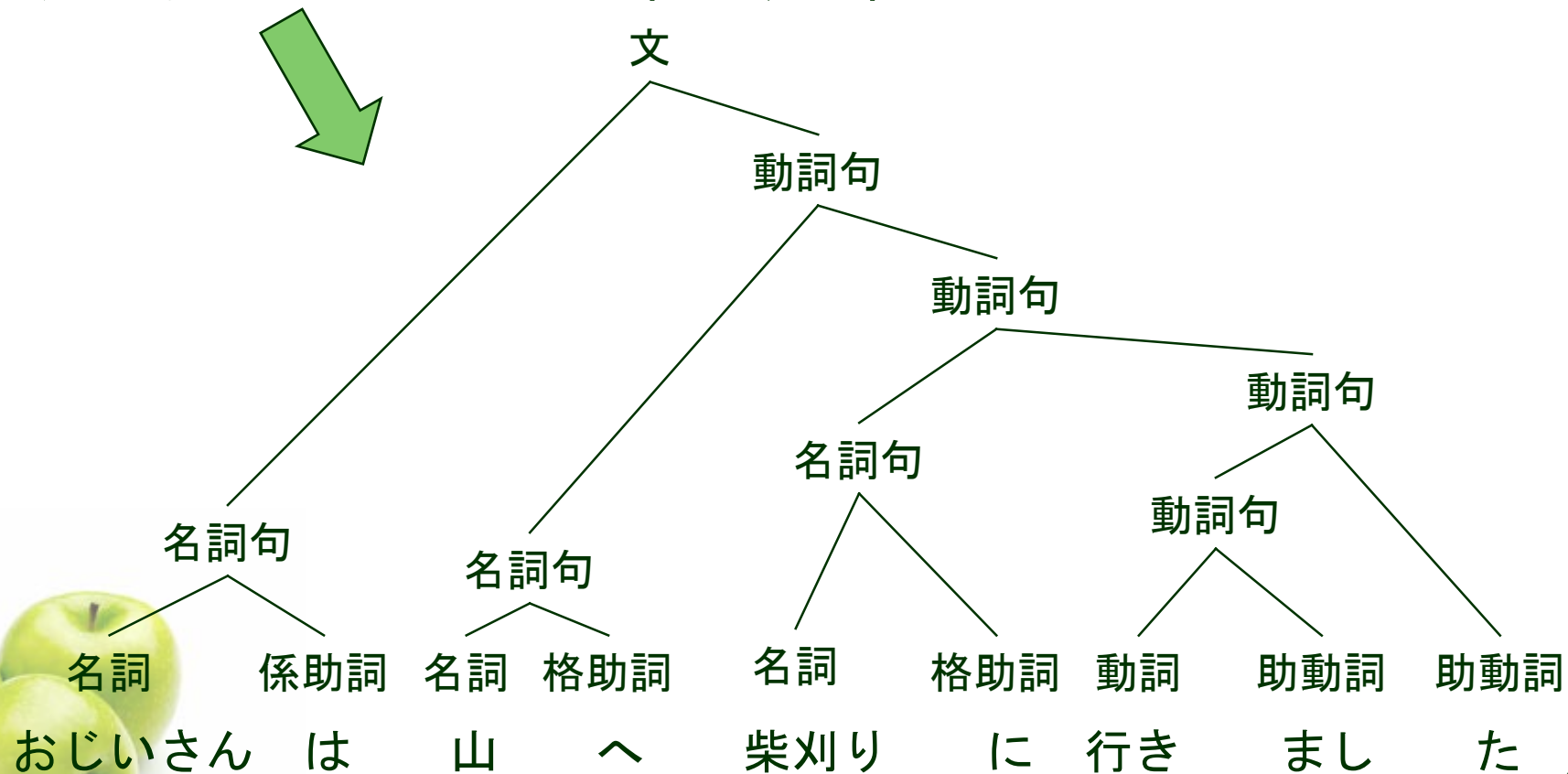
- エスペラントなど人工的につくられた言語)
- 形式言語（記号論理学やコンピュータープログラミングのためにつくられた記号や式によって作られた人工言語)



# この講義の目的 (2/2)

## ● 構文解析

文：おじいさんは山へ柴刈りに行きました

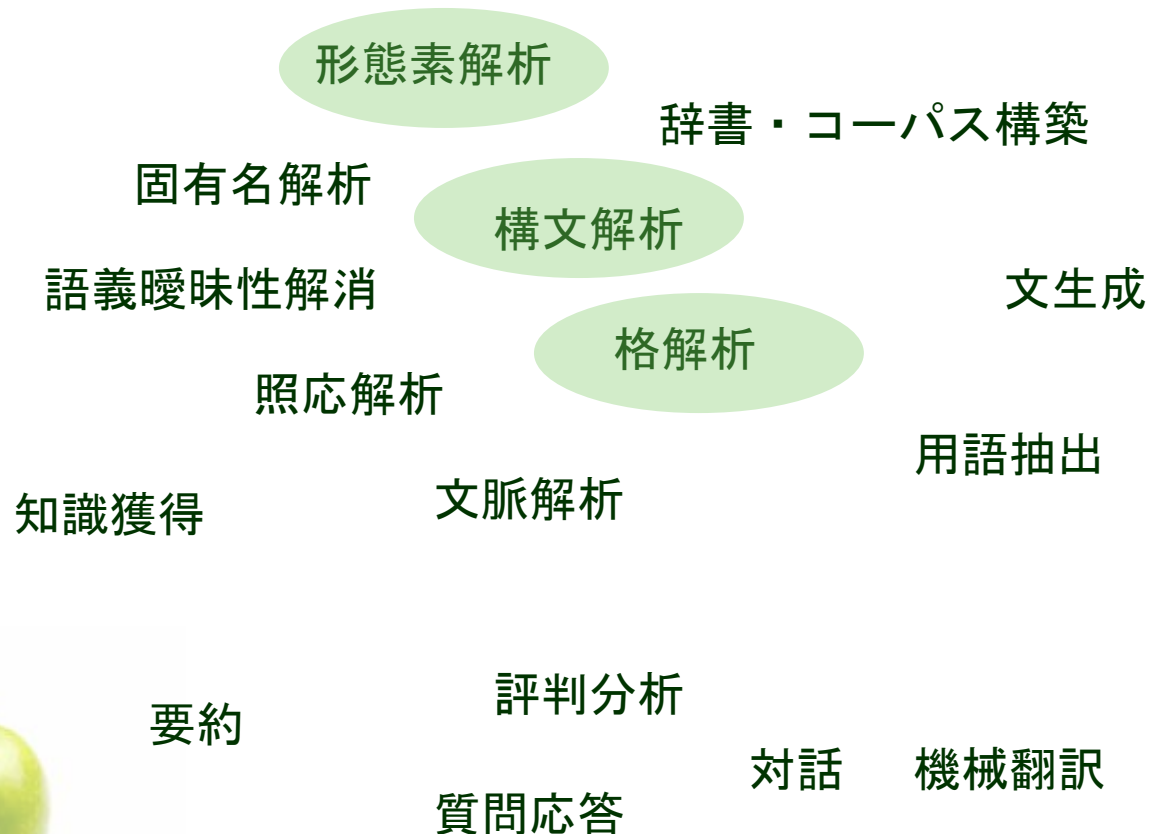


# 構文解析の背景



# 自然言語処理全般と構文解析

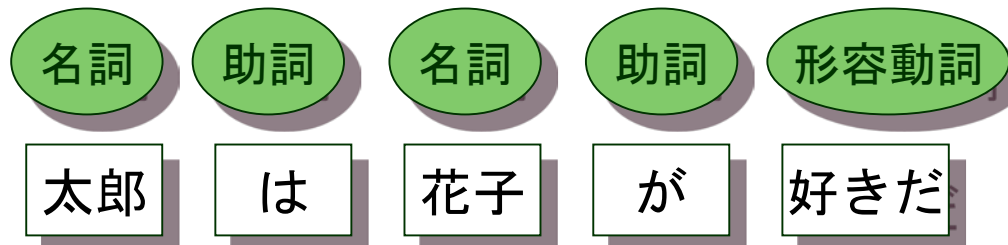
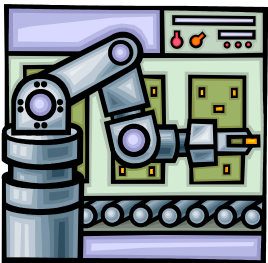
## 自然言語処理全般



# いわゆる“自然言語処理”

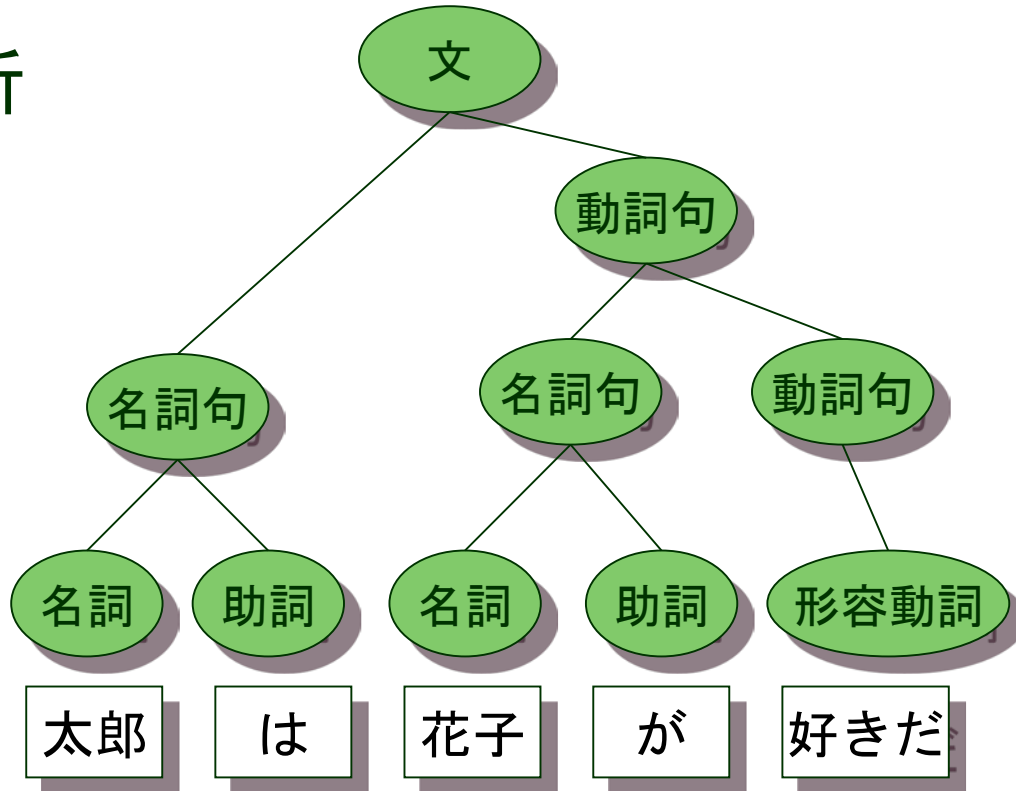
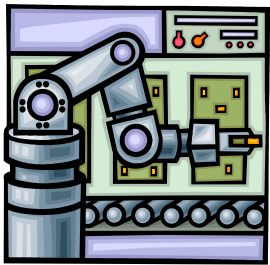
## 形態素解析

“太郎は花子が好きだ”



# いわゆる“自然言語処理”

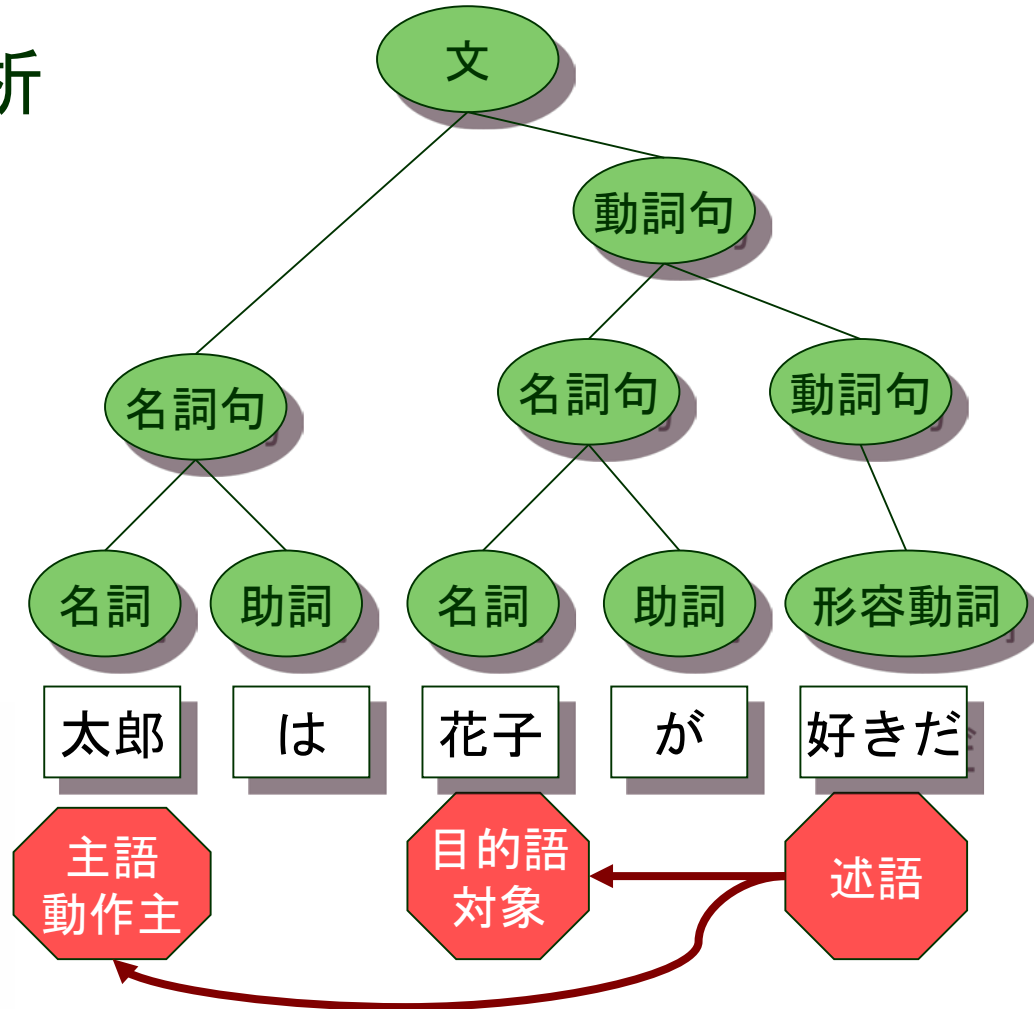
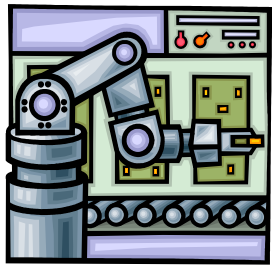
## 構文解析



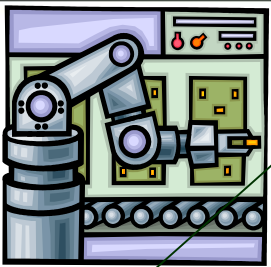


# いわゆる“自然言語処理”

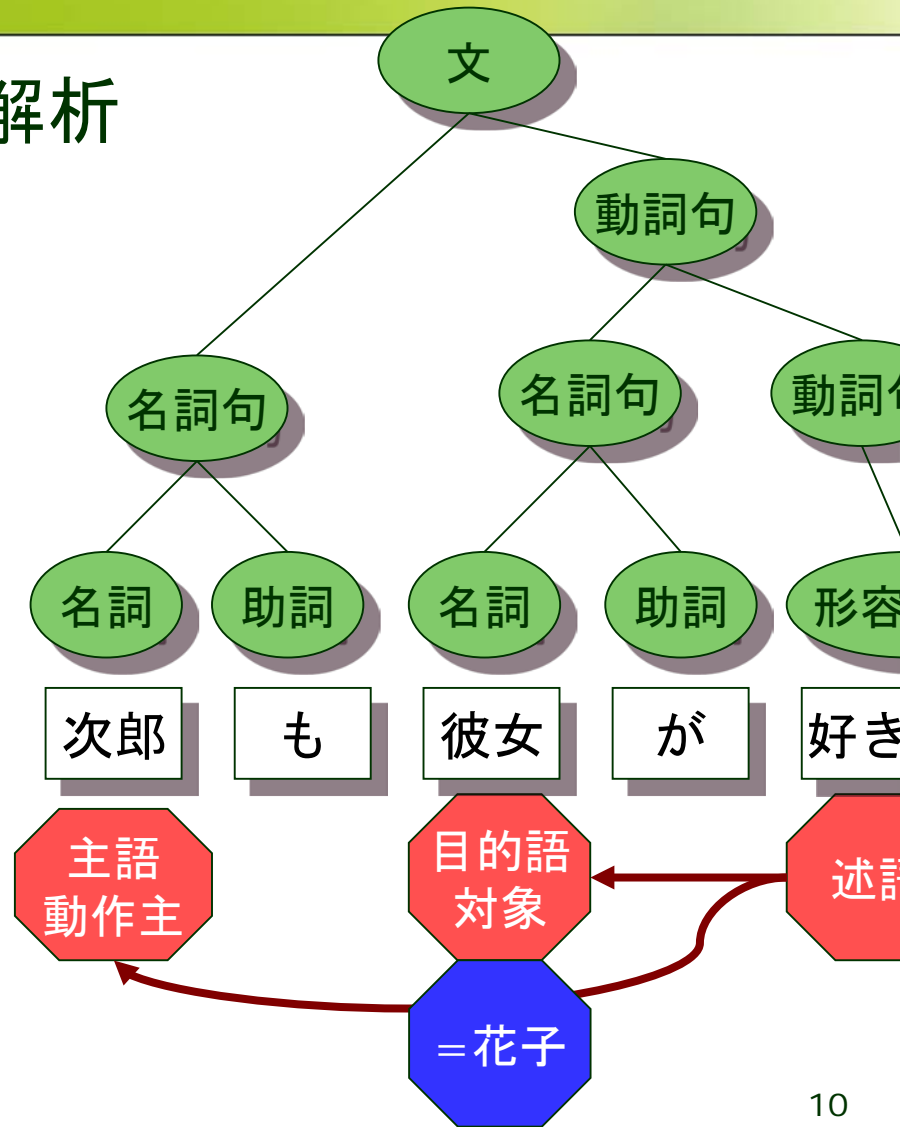
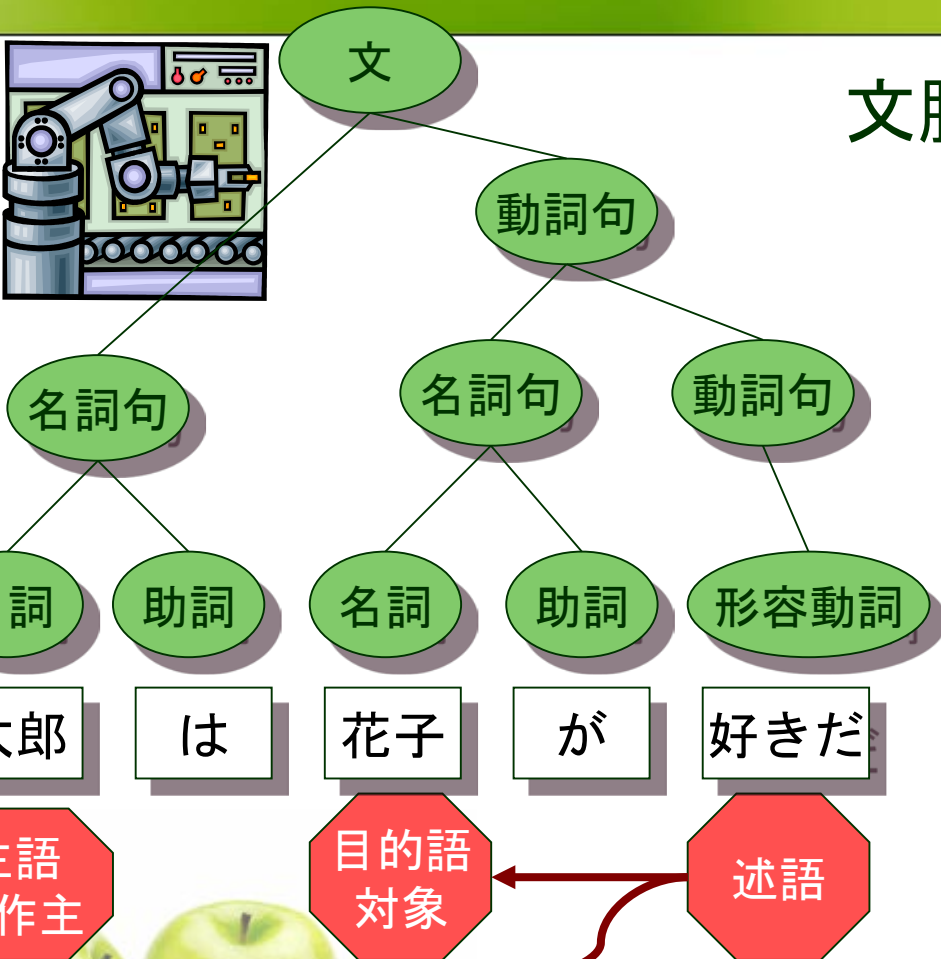
## 意味解析



# いわゆる“自然言語処理”

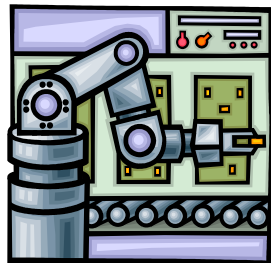


## 文脈解析



# いわゆる“自然言語処理”

“太郎は花子が好きだ”

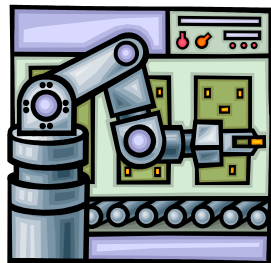
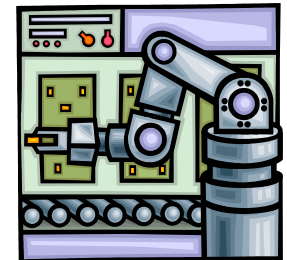


形態素解析

太郎/名詞 は/助詞 花子/名詞 が/助詞 好きだ/形容動詞

構文解析

(文 (名詞句-主語太郎/名詞 は/助詞) (名詞句-目的語  
花子/名詞 が/助詞) (動詞句 好きだ/形容動詞))

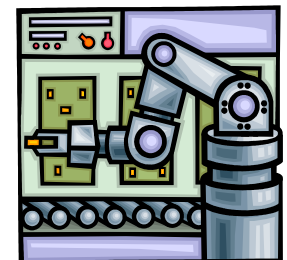


意味解析

太郎/名詞/主語/動作主/人物  
花子/名詞/目的語/対象/人物  
好きだ/動詞/動作主-太郎/対象-花子

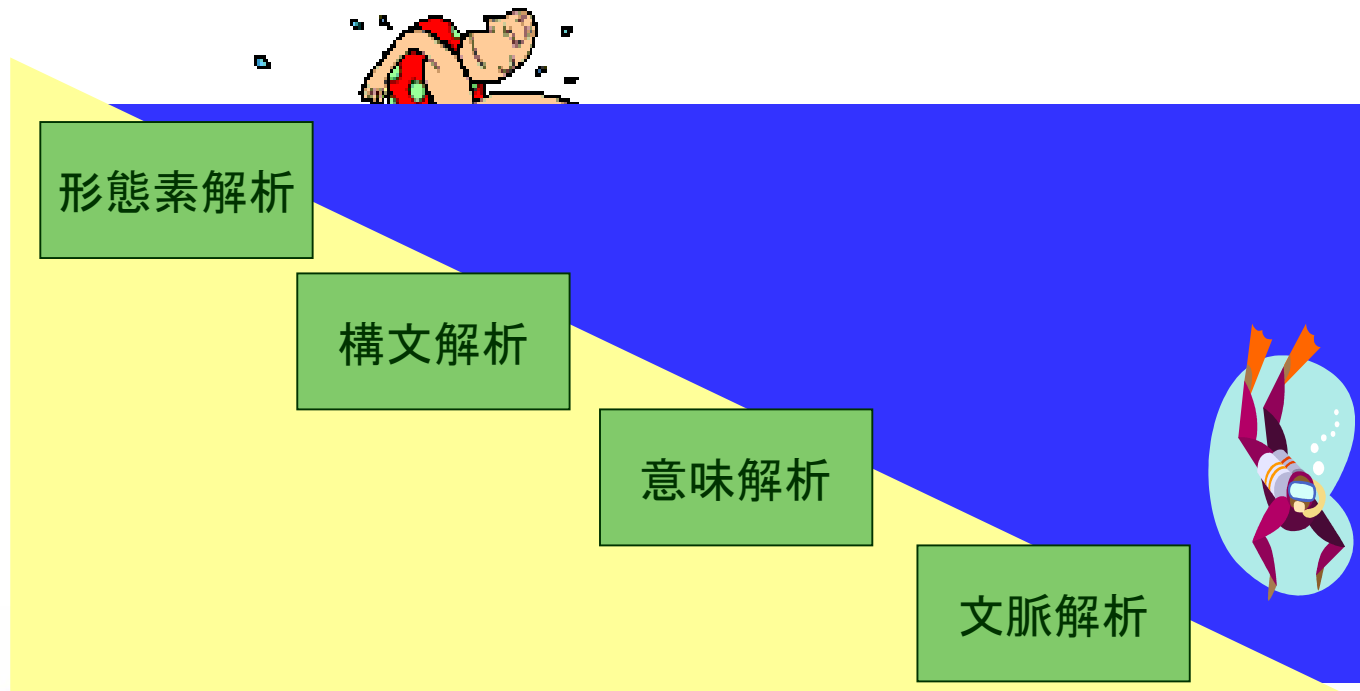
文脈解析

...



# “浅いところ”から“深いところ”へ

理想



# “浅いところ”から“深いところ”へ

現実

形態素解析

構文解析

意味解析

文脈解析

思いの他  
深い！

TAG, LFG, HPSGなど複雑精巧な文法が提案・研究されてきたにも関わらず、実テキストを解析することはなかなかできなかった

# なぜ構文解析に注目するのか？

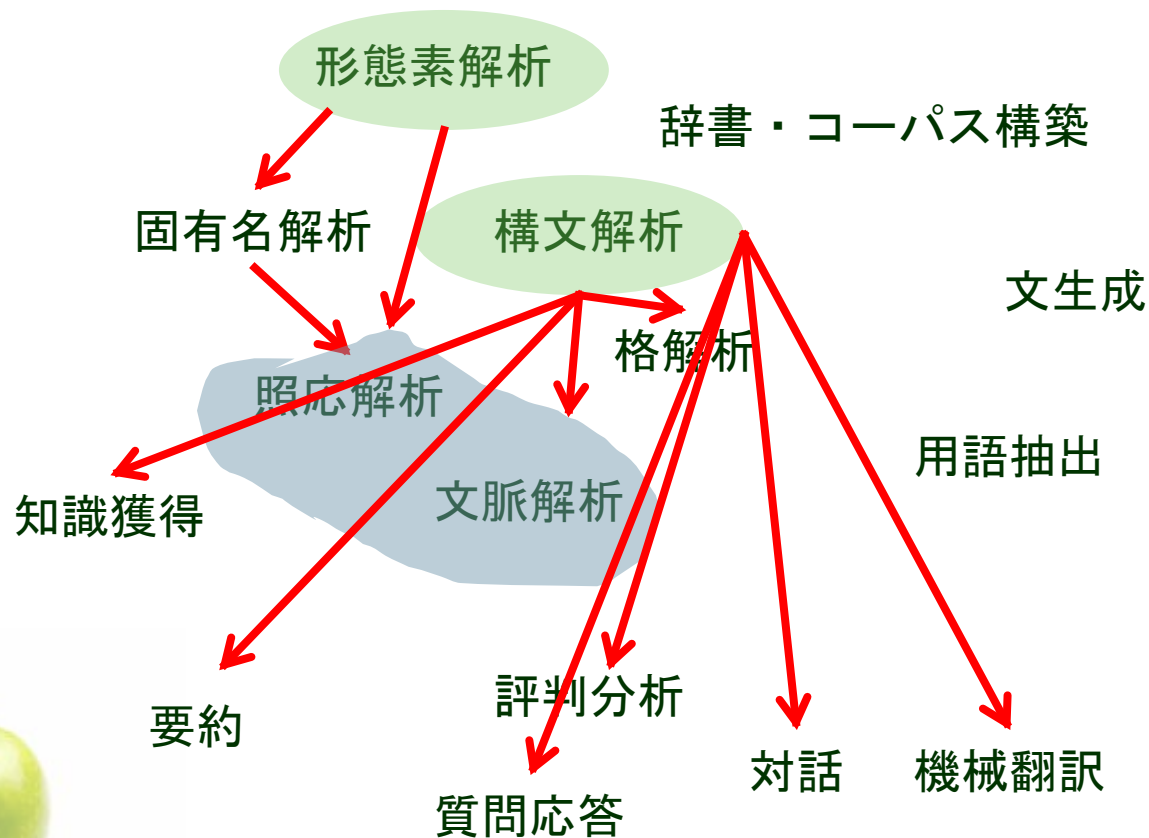
## (1/2)

- 構文解析の基礎技術が他の多くの自然言語処理の基礎技術を包含
- 長く使える（であろう）技術
- 意味処理、文脈処理へとつながる重要な技術
  - 格解析、照応解析は構文解析結果を利用
- 自然言語処理アプリケーションでの利用
  - 機械翻訳、質問応答、自動対話システム



# 自然言語処理全般と構文解析

## 自然言語処理全般



# なぜ構文解析に注目するのか？

## (2/2)

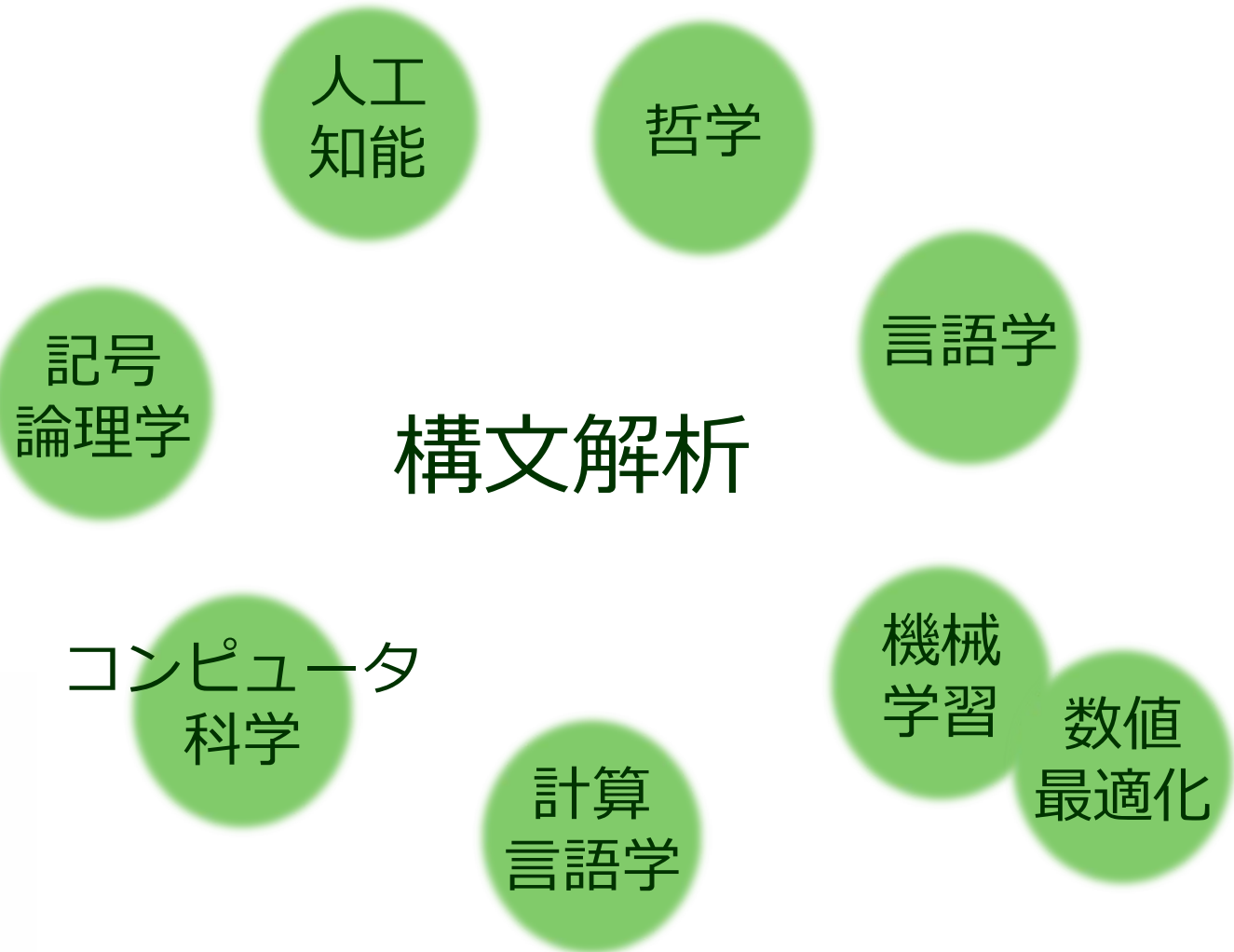
- 歴史は長いが、この十年で大きな技術革新
  - 計算機の高速化
  - メモリの大容量化
  - 言語リソースの出現
  - 実テキストの解析が可能に！
- 句構造解析は世界的にはよく研究されているが、日本ではあまり研究されていない！





# 構文解析は面白いのか？

- 複合的



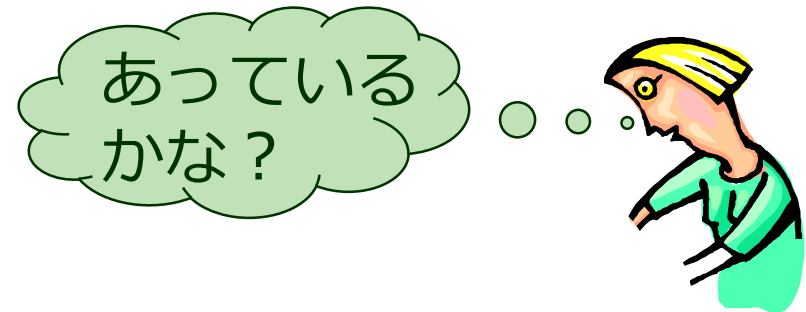
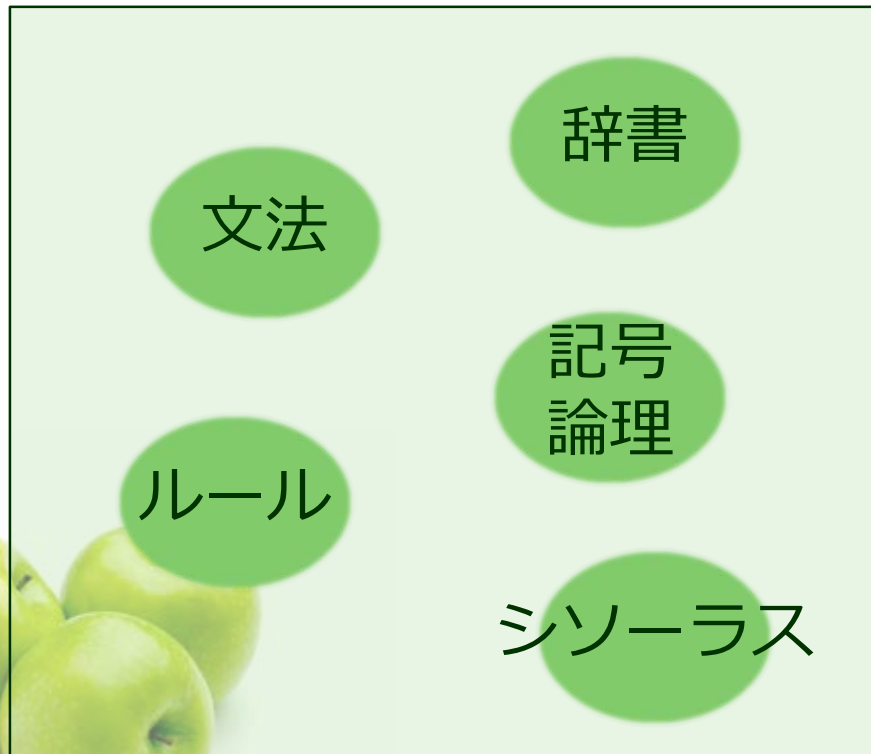
# 構文解析が今面白い理由

- おもちゃから実用的なシステムへ
  - 合理的(rational)から、経験的(empirical)へ
  - 80年代後半に構文木や係り受け木を人手で付与した正解データの登場
    - Penn Treebank [Marcus+1993] ← 5万文ぐらい
    - SUSANNE [Sampson 1995]
    - TIGER Treebank [Brants+ 2002]
    - Prague Dependency Treebank [Hajic 1998]
    - Verbmobil [Hinrichs+ 2000]
    - EDRコーパス [EDR 1995]
    - 京都大学テキストコーパス [黒橋+ 1997] ← 4万文ぐらい
    - 日本語話し言葉コーパス [前川+ 2000]
  - 90年代後半には、高精度構文解析が可能に
  - 00年代は、高速化や機械学習によるさらなる高精度化

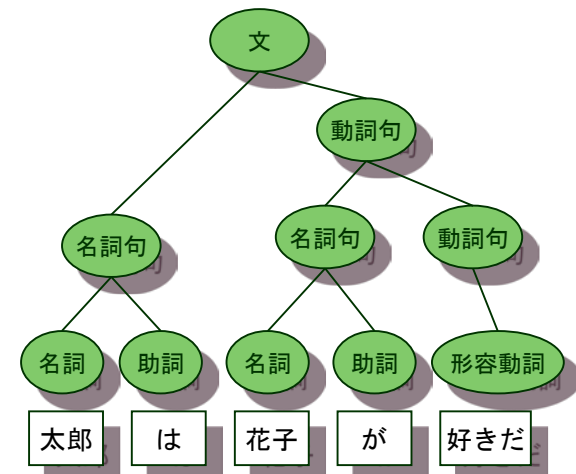
# 合理的から経験的

- 合理的（～90年代後半）

入力：“太郎は花子が好きだ”



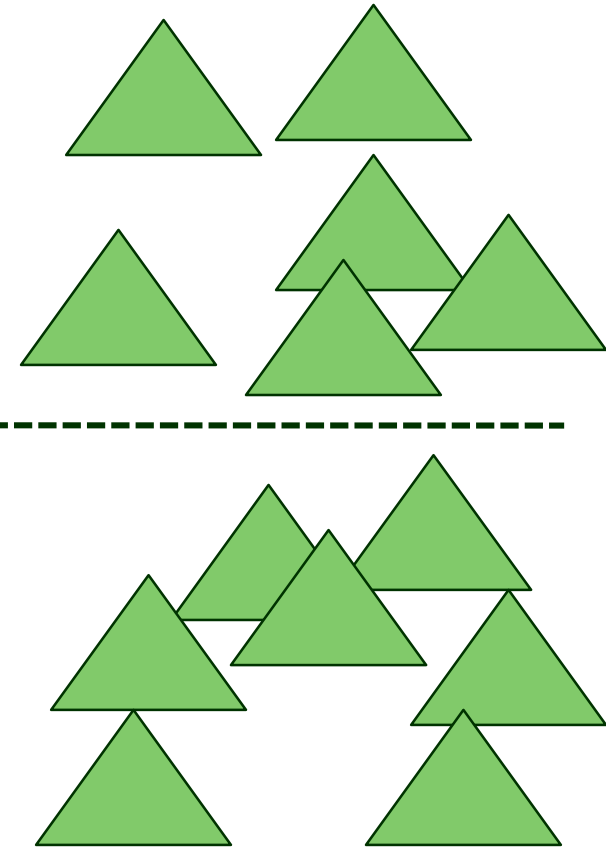
出力：構文木



# 合理的から経験的

- 経験的（90年代後半～）

データ（構文木の集合）



より良い  
モデル

改良

訓練データ

より良い  
モデル

統計モデル  
・  
識別器

一般化

一般化  
モデル

より良い  
モデル

評価データ



# 合理的研究の黄昏

- 合理的モデルから、実データに対して有効な数理モデルへ
  - 正解データに対する精度の評価が中心に
  - 複雑精巧なモデルよりも単純なモデル
    - 同じ精度がでるのならより単純な方法の方が良い！
- 合理的方法の衰退
  - 論理に基づく研究の衰退
    - 複雑精巧なのに実用に結びつかない
  - 文法開発の失敗
  - (参考) I. カント「純粹理性批判」1781、黒崎政男「カント『純粹理性批判』入門」2000
- 理論的計算量よりも実測値による計測

# 講義内容の射程



# ここではすでに処理済みと考える問題

文：おじいさんは山へ柴刈りに行きました



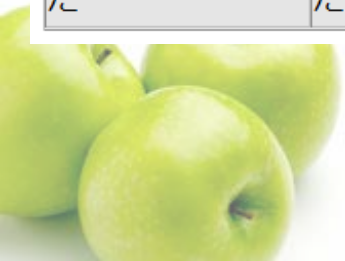
## 形態素解析

### 分析の結果

単字	読法	原形	品詞	語詞情報	詞尾変化
おじいさん	おじいさん	おじいさん	名詞-一般		
は	は	は	助詞-係助詞		
山	やま	山	名詞-一般		
へ	へ	へ	助詞-格助詞-一般		
柴刈り	しばかり	柴刈り	名詞-一般		
に	に	に	助詞-格助詞-一般		
行き	いき	行く	動詞-自立	五段・力行促音便	連用形
まし	まし	ます	助動詞	特殊・マス	連用形
た	た	た	助動詞	特殊・タ	基本形

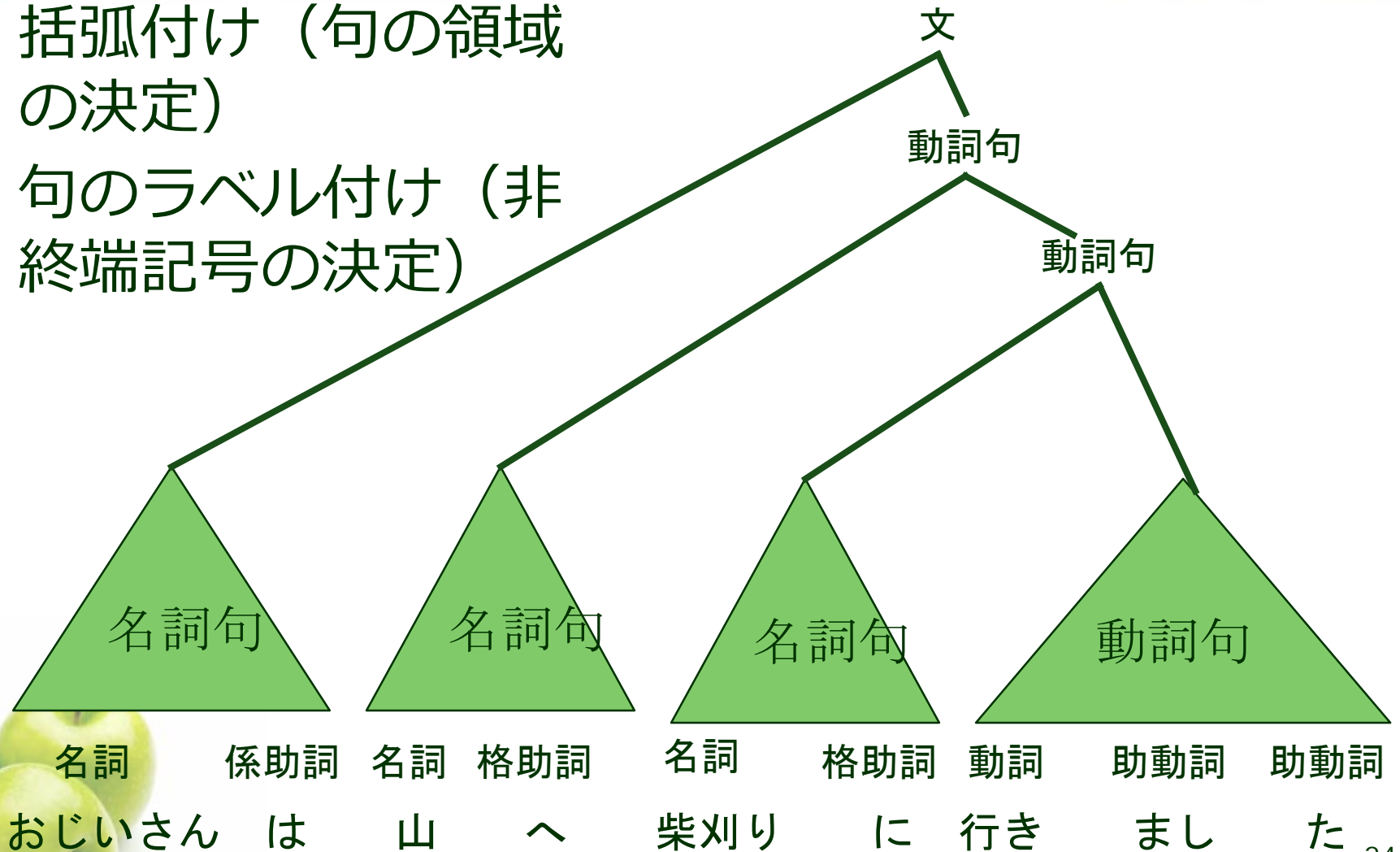


## 構文解析



# 構文解析で行うこと

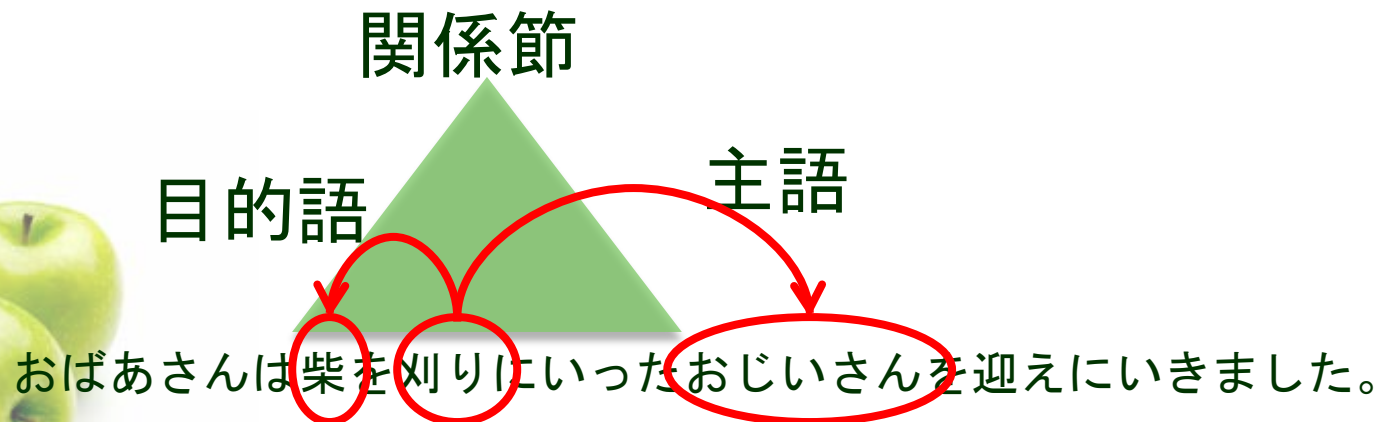
- 括弧付け（句の領域の決定）
- 句のラベル付け（非終端記号の決定）





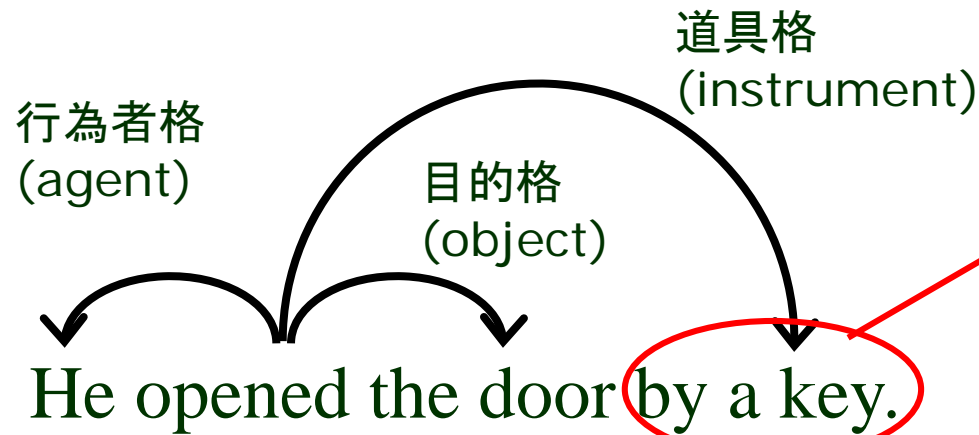
# 構文解析で行うこと

- 言語学的現象の解析（関係節、Control verb, 補文、受動態、疑問文、命令文、テンス、アスペクト、動名詞、It-Cleft、タフ構文など）
- 述語と項の関係（主語、述語など）

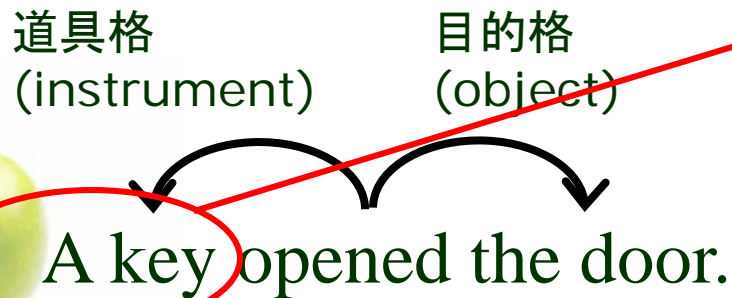


# 構文解析に含まれていると考えられがちな問題 (1/3)

- 格解析, Semantic Role Labeling



構文解析では前置詞句としか解析しない



構文解析では主語としか解析しない



# 構文解析に含まれていると考えられがちな問題 (2/3)

- 固有表現解析 (Named Entity Tagging)
  - 人名
  - 地名
  - 組織名
  - 日付
  - 時間
  - 金額
  - 割合

太郎は、10月1日、朝9時に根津で花子とうどんを食べた。



人名

日付

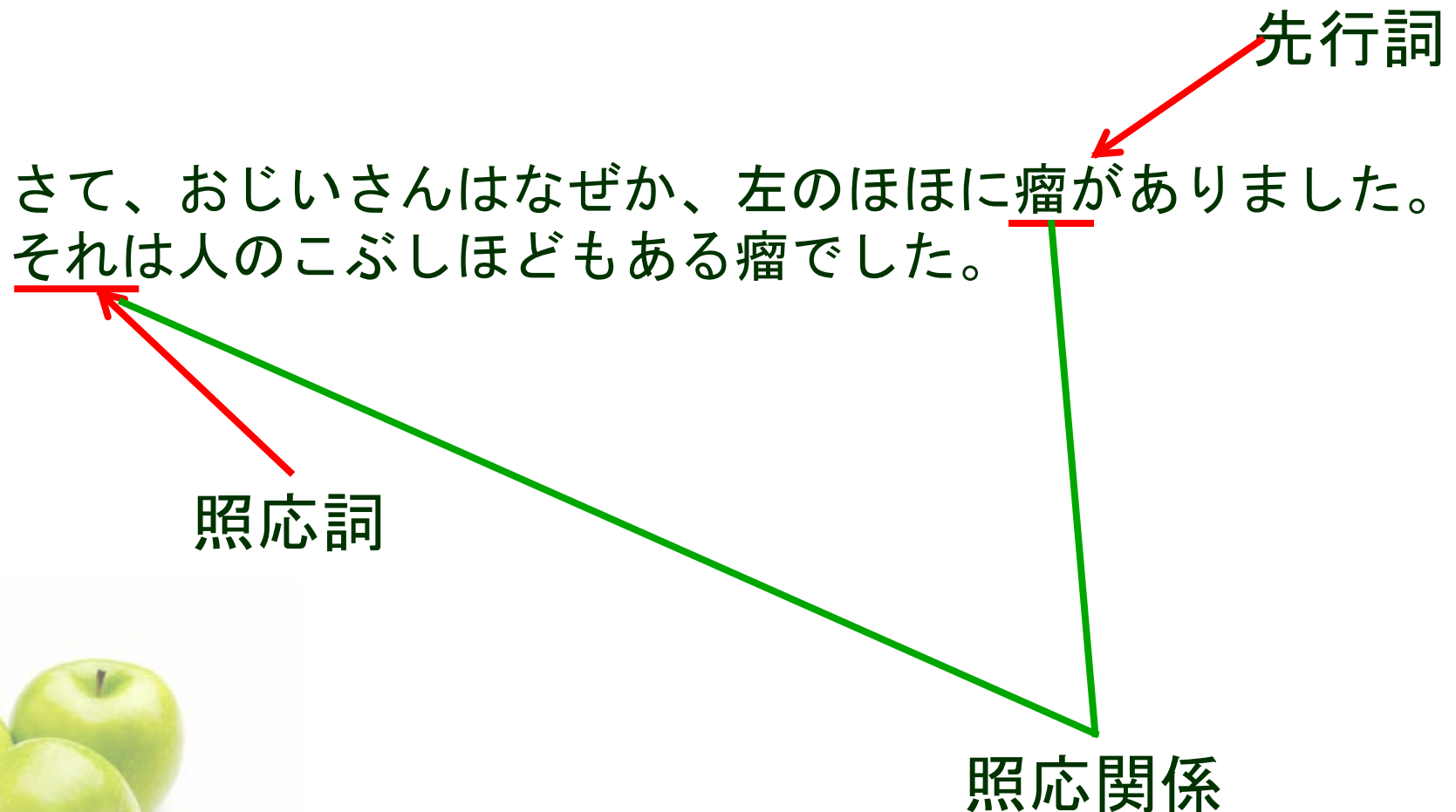
時間

場所

人名

# 構文解析に含まれていると考えられがちな問題 (3/3)

## ● 照応解析



# 講義內容 (1/2)

- 文法柢組
  - CFG (Context Free Grammar, 文脈自由文法)
  - Dependency Grammar (依存文法)
  - Categorical Grammar (CG, 範疇文法)
- 確率的文法柢組
  - 確率的文脈自由文法 (Probabilistic CFG, PCFG)



# 講義内容 (2/2)

- パラメータ推定
  - 最尤推定とEMアルゴリズム
  - 内側外側アルゴリズム
  - 条件付確率場
- 解析アルゴリズム
  - 系列ラベリング
  - フルパーズィング
    - CKYアルゴリズム
  - 確率的（スコア付）文法のための解析アルゴリズム
    - ビタビアルゴリズム
    - ビームサーチ
  - 依存構造解析
    - MSTアルゴリズム



# 連絡

- 評価は2回のレポートで行います。内容と時期はまだ未定です。

- 連絡・資料

<http://aiweb.cs.ehime-u.ac.jp/~ninomiya/ai2/>



# 参考書

- 英語

- D. Jurafsky, J. H. Martin, A. Kehler, K.V. Linden & N. Ward  
“Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition” Prentice Hall Series in Artificial Intelligence, 2000
- C. D. Manning & Hinrich Schütze “FOUNDATIONS OF STATISTICAL NATURAL LANGUAGE PROCESSING” MIT Press, 1999

- 日本語

- 高村大也「言語処理のための機械学習入門」奥村学監修、コロナ社、2010年
- 北研二「言語と計算 4 確率的言語モデル」辻井潤一編、東京大学出版会、1999年
- 長尾真、佐藤理史、黒橋禎夫、角田達彦「岩波講座 ソフトウェア科学 15 自然言語処理」長尾真編、岩波書店、1996年
- 長尾真、中川裕志、松本裕治、橋田浩一、John Bateman、乾健太郎「岩波講座 言語の科学 8 言語の数理」岩波書店、1999年

