共生社会特論
第1回　自然言語処理基礎

小川 泰弘

共生

- 計算機と人間の共生

- 言語が異なる人々の間での共生

機械翻訳

講義計画

11月29日　自然言語処理基礎
12月　6日　ルールベース機械翻訳
12月13日　統計的機械翻訳
12月18日　レポート締切
12月20日　機械翻訳における諸問題
1月17日　評価方法、レポート採点締切
1月24日　派生文法とウイグル語機械翻訳
1月31日　法令翻訳、レポート返却

レポート課題

書評

- 課題図書から1冊選ぶ
  - 早い者勝ち
- PDF形式1ページ
- 名前ありと名前なしの2ファイルを提出
  - 名前なしは採点用
- メールで提出：詳細は講義用サイト参照
- 書評の内容と、他人への採点・添削を評価

課題の補足説明

- 書評であり、概要でも感想でもない
  - 概要や感想を書いていけない訳ではない
- 講義用サイトにリストを掲載
  - 書評が提出された図書はリストから削除
- 後日、他人の書評を採点する
  - 一人あたり五人分を予定
  - 採点方法は後日説明

自然言語

- 自然言語 (natural language)
  - 我々人間が使用する言語
    - 日本語
    - 英語
- 人工言語 (artificial language)
  - 人工的に作られた言語
    - プログラミング言語
    - エスペラント
    - 手話

形式言語 (formal language)

形式言語

- アルファベット: Σ
- 語 (word) : $w \in \Sigma^*$
  - アルファベットの要素の列
- 言語 (language): $L \subseteq \Sigma^*$
  - 語の集合

自然言語をモデル化

by Noam Chomsky

言語学

- 音韻論
- 形態論
- 構文論/統語論
- 意味論
- 語用論

音韻論 (phonology)

- 言語の音素を扱う
- 生成音韻論
  - 生成文法の観点からの音韻論

音声認識や音声合成などの音声言語処理は自然言語処理とは別の分野とされる

形態論 (morphology)

- 語(word)の構造などを扱う
- 形態素(morpheme)
  - 言語において意味を有する最小の単位
  - 文法解析における最小単位

形態素

形態素と形態を区別しないことも多い

- 形態素(morpheme): 抽象的→2種類に分類
  - 意味概念を持つもの（語根）e.g. *door*, *blue*, *take*
  - 抽象的素性 e.g. 過去や複数
- 形態(morph): 形態素が具現化したもの
  e.g. *doors* → *door*, *-s*：それぞれが形態
  e.g. *take*, *took*：同じ形態素の異形態(allomorph)
  - 自由形態: それ自体で語になれる形態
    e.g. *door*（語としての *door* は monomorphemic)
  - 拘束形態: 他の形態と一緒にのみ出現
    e.g. 接辞, 複数形の *-s*, *exclude* の *ex-*と*-clude*

語/単語 (Word)

- 直接的な定義はない
  e.g. *database* or *data-base* or *data base*
  e.g. *言語処理学会、 言の葉、 東西南北*
- 統語論的定義 : 文を構成する単位
  - 形態論は語の内部構造を扱う
  - 構文論的文脈により形が変化
    e.g. *degrade, degrades, degrading, degraded*

- 音韻論的定義 : 音韻過程のある範囲

基本形/基底形 (Base Form/Lemma)

基本形：語の意味を持つ部分

- 英語：単数形、非三人称単数現在能動態
- 日本語：終止形もしくは語幹(stem)

基本形が複数の形態から成る場合もある

- 語根(root) ：中心的意味を担う形態

e.g. *degrade = de + grade* の *grade*

e.g.「集まる」「集める」の「集(m)」

語彙素と語彙目録

- 語彙素(lexeme)
  - 形態は異なるが同じ語と考えられる集合

  e.g. *go - went - gone - going*

- 語彙項目(lexical entry)
  - 語の形式・意味・文法的特性

- 語彙目録(lexicon)
  - 語彙項目の集まり、辞書
  - 「語彙」とも言うが、vocabulary とは異なる

品詞 (Part-of-Speech; POS)

- 品詞の分類
  - 意味的: 物を指すか、動作を指すか
  - 文法的: 日本語における形容詞と形容動詞
- 基本的な品詞
  - 名詞
  - 動詞
  - 形容詞
  - その他

単語と品詞の関係

- 英語：　同じ単語が複数の品詞を持つ
  - 特に多くの名詞が動詞になる
    - *hand, eye, line, circle, bird, cat, ship*
    - *E-mail me.*
    - *Google it.*
    - *Man the boat!*
- 日本語：　単語の品詞は一つ
  - 「*大きい*」と「*大きな*」は別品詞
  - 例外は「*今日*」「*実際*」「*結果*」など副詞的名詞

屈折 (Inflection)

- 品詞はそのままで文法的機能を変更
  - 変化形の集合→語形変化表(paradigm)
  - 屈折は「完全」
- 屈折による語の分類
  - 不変化詞：前置詞、副詞、接続詞、冠詞
  - 活用語：動詞
  - 名詞語句：名詞、形容詞、代名詞

屈折の分類

- 活用 (conjugation):
  動詞が時制・態・人称や数の一致により変化
- 曲用 (declension):
  名詞や形容詞が、各種の一致により変化
  - 数・格・性・比較などの一致

新語の生成方法　1

派生(derivation)

computationally = compute + tion + al + ly

- 基本形＋拘束形態
- 品詞が異なる語も生成
- 「完全」ではない
  - 特定の派生が適用できるサブクラスが存在
- 再帰的適用も可
- 派生接尾辞は通常、一つの意味を表す

新語の生成方法　２

合成(compounding)

- 基本形＋基本形 e.g. *bedtime, red wine*
- 連結形態が入る場合もある e.g. *bull's eye*
- 意味解釈は複雑
  e.g. *外国、愛国、帰国、米国*
- 派生との区別は曖昧 e.g. *-ful*
- 句との区別も曖昧 e.g. *red wíne* vs. *réd wine*
  roter Wein vs. *Rotwein*

語の構成方法

- 基本は語根に接辞（affix）が連結
  - 膠着語はもっぱらこの方法
- それ以外にも多くの種類がある
  - 母音交替　　e.g. *swim - swam - swum*
  - 韻律変化　　e.g. *increase*
  - 補充法　　e.g. *went* は元々 *wend* の過去形
  - ゼロ形態　　e.g. *deer* の複数形
  - 品詞転換　　例は既出

接辞添加 (Affixation)

接辞：拘束形態　音素の列として具現化

- 接頭辞：語幹の前に接続 e.g. *un*common
- 接尾辞：語幹の後に接続 e.g. *shoe*s
- 接周辞（両面接辞）：前後にペアで出現
  e.g. *お*疲れ*さま*、*お*待ち遠*さま*
- 接中辞（挿入辞）：語幹の中に出現
  e.g. ボントク語（フィリピンのボントク族の言語）
  /fikas/ (strong)　*/f*um*ikas/* (to be strong)
  /kilad/ (red)　*/k*um*ilad/* (to be red)
  /fusul/ (enemy)　*/f*um*usul/* (to be an enemy)

構文論/統語論(Syntax)

- 文の構造を扱う
  - 句構造文法
  - 依存文法

句構造文法(Phrase Structure Grammar)

- 生成規則

CFG:

S → NP VP

NP → ADJ N

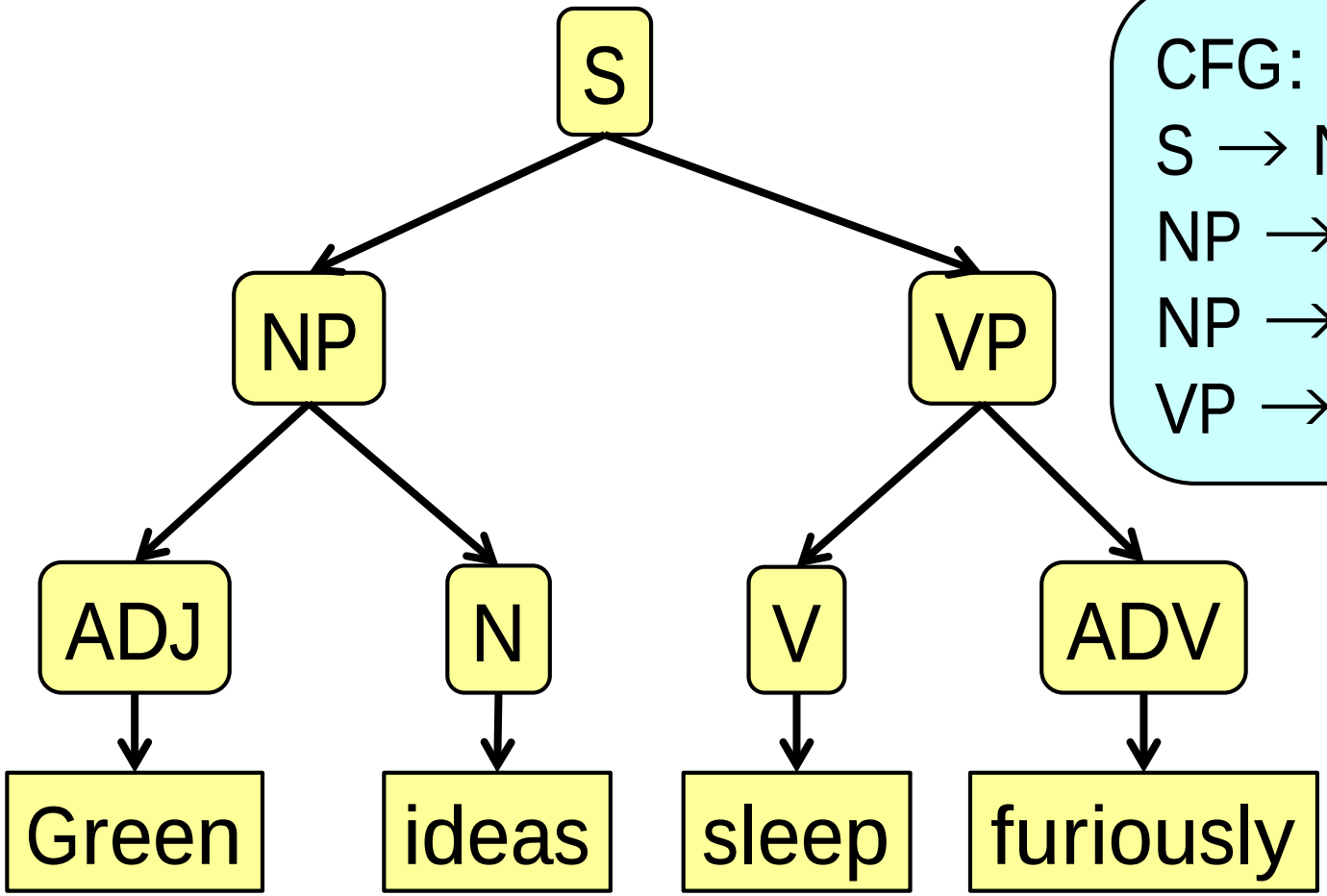
NP → ADJ ADJ N

VP → V ADV

VP → V NP

名詞句や動詞句といった句を考える

構文木(syntax tree / parse tree)



CFG:

S → NP VP

NP → ADJ N

NP → ADJ ADJ N

VP → V ADV

(S: (NP: (ADJ: Green) (N: ideas)) (VP: (V: sleep) (ADV: furiously)))

構文木ができない例

Furiously sleep ideas green colorless.

ADV	V	N	ADJ	ADJ
ADV	V	N	N	ADJ
ADV	N	N	ADJ	ADJ
ADV	N	N	N	ADJ

非文法的：構文木を構成できる文法規則なし

チョムスキー階層

- タイプ0
  - チューリングマシン
- タイプ1　文脈依存文法
  - 線形拘束オートマトン
- タイプ2　文脈自由文法 (context-free grammar)
  - プッシュダウン・オートマトン
- タイプ3　正規文法 (regular grammar)
  - 有限オートマトン

依存文法(dependency grammar)

- 単語と単語の間の関係; 係り受け
  - 主語と述語
  - 動詞と目的語
  - 修飾・被修飾

生成文法は主辞のマークをつけることで対応

意味論

- 文の意味を扱う
  - 語や句のタイプ　　例：固有名詞
  - 文中での役割　　例：主語と目的語
  - その他様々な情報
    - 生物か無生物か
    - 会社・組織・場所・日付・金額

意味的な役割と構文的な役割

構文は異なるが同じ意味の文

The Federal Court chastised Microsoft.

Microsoft was chastised by the Federal Court.

- 文法的な主語は異なる
- 動作主(agent)と受容者(recipient)は同じ

照応 (anaphora)

- 代名詞などによる語の指示

  e.g. *太郎の兄に会った。彼は大学生だ。*

- ゼロ代名詞の照応

  e.g. *本を買った。*

語用論(pragmatics)

- 環境に依存する文の意味を扱う

  e.g. *ソースある？*

  e.g. *Do you have a watch?*

曖昧性 (Ambiguity)

自然言語につきもの

- 単語の多義性

  crane　*take*

- 品詞の曖昧性

  Time files like an arrow.

- 構文的曖昧性

  John saw a girl with a telescope.

曖昧性解消

人間は文脈・常識などで解消している

e.g. 無礼なのは誰？

ジョンは横柄で無礼な音楽家の子供の世話をした

ジョンは横柄で無礼な音楽家の子供が嫌いだった

言語と方言

明確な定義はない

- 旧ユーゴスラビア
  - 分裂前：セルボクロアチア語
  - 現在：　セルビア語、クロアチア語、ボスニア語
- ドイツ語の方言とオランダ語
  - 低地ドイツ語と高地ドイツ語はドイツ語の方言
    - ✧意思疎通は困難
  - 低地ドイツ語とオランダ語は別言語
    - ✧意思疎通は容易

ピジン言語とクレオール言語

- ピジン(pidgin)
  - 現地人と貿易商人の間の意思疎通などのために生まれた混成言語
  - 自然言語と人工言語の中間
- クレオール(creole)
  - 母語化したピジン

  e.g. 小笠原語(Bonin English)
    - 英語と日本語の混成

    「ユーは何のティーチャーかい?」

言語の数

世界に何言語あるかは不明

- エスノローグ第19版には7,097言語登録
  - 日本で使用されているのは
    - 日本語(jpn, ja)
    - アイヌ語(ain)
    - 中央沖縄語(ryu), 南奄美語(ams), 喜界語(kzg), 宮古語(mvi), 沖永良部語(okn), 北奄美語(ryn), 八重山語(rys), 徳之島語(tkn), 国頭語(xug), 与那国語(yoi), 与論語(yox)

言語の分類

- 孤立語（e.g. 中国語）
  - 拘束形態がない（＝接辞がない）言語
- 膠着語（e.g. チュルク諸語、フィン・ウゴル諸語）
  - すべての拘束形態が接辞である言語
  - それぞれの接辞が異なる一つの素性を表現
- 屈折語（e.g. インド・ヨーロッパ語族）
  - 複数の素性を一つの拘束形態が表現
  - 同じ素性を異なる形で表現
- 抱合語（e.g. イヌイト語、アイヌ語）
  - 目的語などが動詞に付着して一語文を形成

オマケ：音韻論での問題

その前に

「よったり」という単語（名詞）を知っていますか？

あたくしと橘さんと、それからリリちゃんに牧さん。

そのよったりでしたわ。

鮎川哲也 『りら荘事件』
1956年9月 - 1957年12月

発音と表記の一致

ghoti どう発音するか？

- laugh
- women
- nation

fish と同じ

- though
- people
- ballet
- business

発声なし

日本語の発音と表記

必ずしも一致してない

- 「は」：ha，wa
- 「へ」：he，e
- 「ぢ」と「じ」
- 「づ」と「ず」
- 「を」と「お」
- 「遠い(とおい)」「父さん(とうさん)」
- 「映画(えいが)」「姉さん(ねえさん)」
- 「はんのう」「はんぱ」「はんこ」
  「はんを」「はん」

正書法 (orthography)

言語を文字で表記するルール

- 文字（アルファベット）
- 綴り
- 仮名遣い

発音は時代とともに変化するので
正書法とずれが生じる
e.g. 英語の大母音推移

日本語の正書法

- 平仮名の統一（変体仮名の廃止）
- 常用漢字表(昭和56年内閣告示第1号)
  - 最新版は（平成22年内閣告示第2号）
- 公用文作成の要領(昭和27年内閣閣甲第16号)

「正しい表記」が存在

片仮名で表記される外来語などでは、
定まっていない語が多い（発音が異なる場合も）
e.g.「メール」「メイル」
e.g.「プリンター」「プリンタ」
e.g.「スパゲティー」「スパゲッティ」「スパゲチー」

変体仮名

- 平仮名のバリエーション

e.g.

そ
は
な
か

画像は以下および Wikipediaから引用
http://www10.plala.or.jp/koin/koinhentaigana.html

発音の変化

- 定着したもの
  - 「新た（あらた）」→「新しい（あたらしい）」
  - 「山茶花（さんさか）」→「さんざか」→「さざんか」

- 進行中？
  - 「体育（たいいく）」→「たいく」
  - 「雰囲気（ふんいき）」→「ふいんき」
  - 「自転車（じてんしゃ）」→「じでんしゃ」

現在は間違いとされる
定着するかどうかは不明

変音現象（連声・音便）

一（イチ）＋ 本（ホン）＝ 一本（イッポン）

- 一本（イッポン）
- 二本（ニホン）
- 三本（サンボン）
- 四本（よんホン）

- 一歩（イッポ）
- 二歩（ニホ）
- 三歩（サンポ）
- 四歩（よんホ）

人数の数え方

ひとり	一人	イチニン
ふたり	二人	ニニン
みたり	三人	サンニン
よったり	四人	よニン ~~シニン~~
いったり	五人	ゴニン
むたり	六人	ロクニン
ななたり	七人	シチニン ななニン
やたり	八人	ハチニン
ここのたり	九人	キュウニン
とたり	十人	ジュウニン

一人前

二人三脚

よったり

あたくしと橘さんと、それからリリちゃんに牧さん。

そのよったりでしたわ。

鮎川哲也 『りら荘事件』
1956年9月 - 1957年12月