

N I I 市民講座 9月8日
(木)



言語情報処理

コンピュータはコトバをどこま
で理解できるようになるのか

国立情報学研究所
相澤彰子

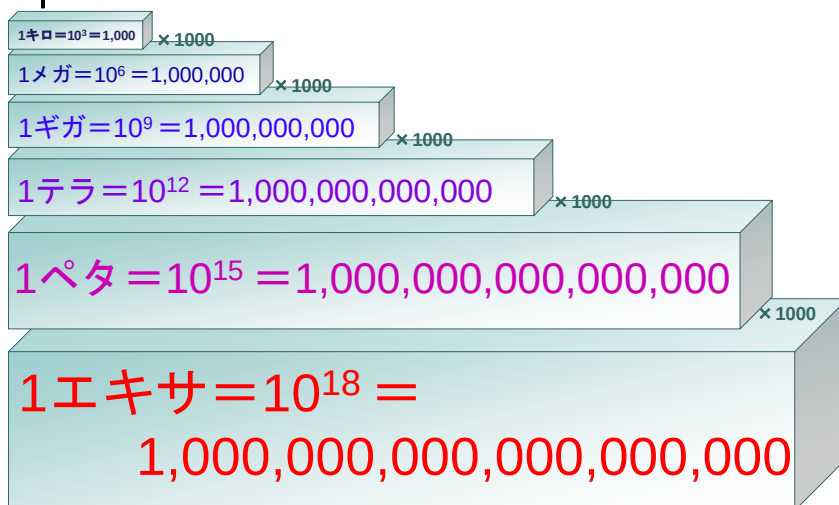


「ペタ」「エキサ」
「ゼタ」、
それぞれ何を表す言
葉？



大きな数を表す単位

N
I
I
市
民
講
座



2005.8.25

3



1文字 = 1バイトとして考えると...

N
I
I
市
民
講
座



2005.8.25

4



人類と情報の記録量

我々が今後2年間で見る情報の量は
人類が過去に見た情報の量よりも大
きい

2005.8.25

<http://www.sims.berkeley.edu/research/projects/how-much-info-2003/execsum.htm#summary>
http://www.emc.com/images/new_bc_images/information_generation/digital_watch/57billion_slide.tif

5



「情報爆発」時代の苦難



2005.8.25

資料提供
文部科学省特定領域研究「情報爆発！丁基盤」

6



ヒトの見方、コン ピュータの見方

「コンピュータが空気になる時代」 2005.06
ユビキタス社会のキーテクノロジー（丸善ライブラリー）
東倉、山本、上野、三浦、p.17



次の文章を読んでください。

N
「」
市民講座

コンピュータが空気と同じように、世界中に
充満している状況を実現する手段の一つは、身
に着けて持ち歩くことでしょう。これがもう一
歩進むと、持ち歩こうと意識しなくても身につ
いている、すなわち、パラサイト（寄生虫）の
ような存在になることです。

さらに将来に向けた究極の形は、身体の中に
コンピュータが埋め込まれている状態で、こ
うなればコンピュータは身体の一部となります。
コンピュータが空気になれば、空気と同じよう
に、なくなれば窒息することを意味します。

2005.8.25

8



そこで質問

- 筆者の主張を表すタイトルをつけてください。
- 「これ」とは何を指すでしょう？
- 「パラサイトのような存在になる」とは、コンピュータがどのようなことになることですか？
- 「コンピュータがなくなると窒息する」理由を説明しなさい。



もっと質問

- 何文字あったでしょう？
- 何種類の単語があったでしょう？
- 一番長い文は何だったでしょう？
- 一番多くでてきた単語は何だったでしょう？
- 接続詞は何個あったでしょう？



遠い道のり

- 筆者の主張を表すタイトルをつけてください。
- 「これ」とは何を指すでしょう？
- 「パラサイトのような存在になる」とは、コンピュータがどのようなことになるのですか？
- 「コンピュータがなくなると窒息する」理由を説明しなさい。

- 何文字あったでしょう？
- 何種類の単語があったでしょう？
- 一番長い文は何だったでしょう？
- 一番多くでてきた単語は何だったでしょう？
- 接続詞は何個あったでしょう？

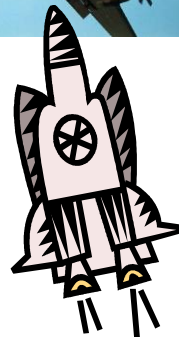
2005.8.25

11



「鳥と飛行機」

N
「」
市民
講



2005.8.25

12

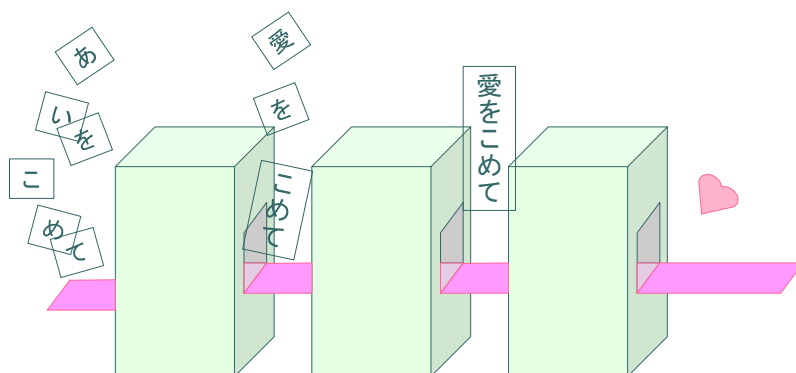


言語処理の階層 ～「もうすぐ8月が終わります。」



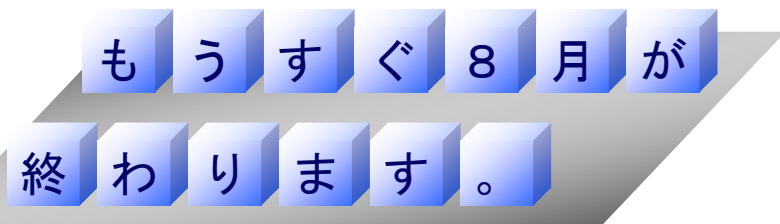
言語処理の階層

N
「
」
市
民
講
座





「もうすぐ8月が終わります。」 文字の切り分け



2005.8.25

15



「もうすぐ8月が終わります。」 形態素の解析

<http://chasen.naist.jp/hiki/ChaSen/>

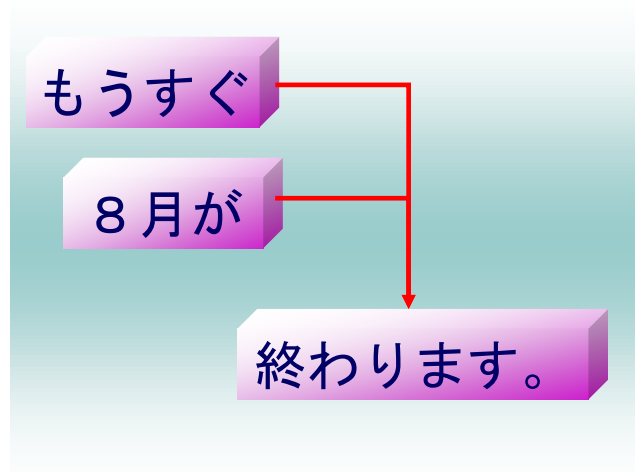
もうすぐ	モウス グ	もうす ぐ	副詞-一般
8月	ハチガ ツ	8月	名詞-副詞可能
が	ガ	が	助詞-格助詞-一般
終わり	オワリ	終わる	動詞-自立 五段・ラ行 連用形
ます	マス	ます	助動詞 特殊・マス 基本形
。	。	。	記号-句点
。			

2005.8.25

16



「もうすぐ8月が終わります。」 文構造の解析



2005.8.25

17



「もうすぐ8月が終わります。」 す。I

8月 **意味解析**

はち - がっ [- グワツ] 【八月】
一年の八番目の月。葉月(はづき)。
《季 秋》「一の太白低し海の上／子規」
(大辞泉より)

終わる

[1] 続いていたものがそこで切れる。
N 1 ガ> 一年が あっという間に 終わった。

[2] 続けていた事をそこで止める。
N 1 ガ N 2 ヲ> 彼は 調べを 終わった。

[3] していた事やしようとしていた事などが思わしくない結果になる。
N 1 ガ N 2 ニ> 計画は 失敗に 終わった。
(IPA基本動詞辞書より抜粋)

2005.8.25

18



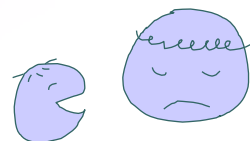
「もうすぐ8月が終わります。」

談話（状況）の解析

「お母さん、遊びに行っていていい？」

「もうすぐ8月が終わります。」

「宿題は明日やるって。」

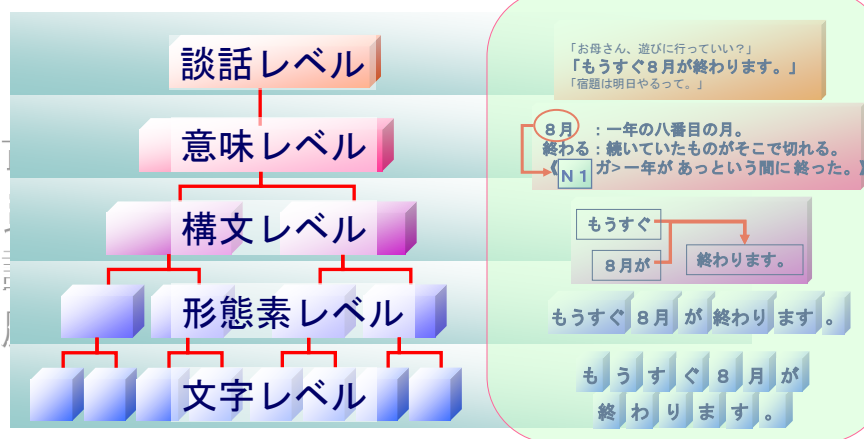


2005.8.25

19



もう一度、言語処理の階層



2005.8.25

20

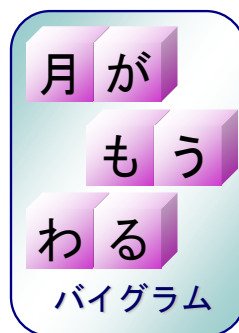


縁の下の力持ち：文字レベルの処理



「Nグラム」

N
E
I
G
H
T
L
E
S
S
O
N
S



半無限長Nグラム

コンピュータが空気と同じように、世界中に充滿している状況を実現する手段の一つは、身に着けて持ち歩くことでしょう。これがもう一歩進むと、持ち歩こうと意識しなくても身についている、すなわち、パラサイト（寄生虫）のような存在になることです。

さらに将来に向けた究極の形は、身体の中にコンピュータが埋め込まれている状態で、こうなればコンピュータは身体の一部となります。コンピュータが空気になれば、空気と同じように、なくなれば窒息することを意味します。



Nグラムの数え上げ ～「サフィックスアレイ」の仕 組み

0 1 2 3 4 5 6 7 8 9 10 11 12
も う す ぐ 8 月 が 終 わ り ま す 。

1
6
3
2
11

サフィックスアレイ

2005.8.25

23



Nグラムの数え上げ ～「サフィックスアレイ」の仕 組み

0 1 2 3 4 5 6 7 8 9 10 11 12
も う す ぐ 8 月 が 終 わ り ま す 。

ユバダイグラム

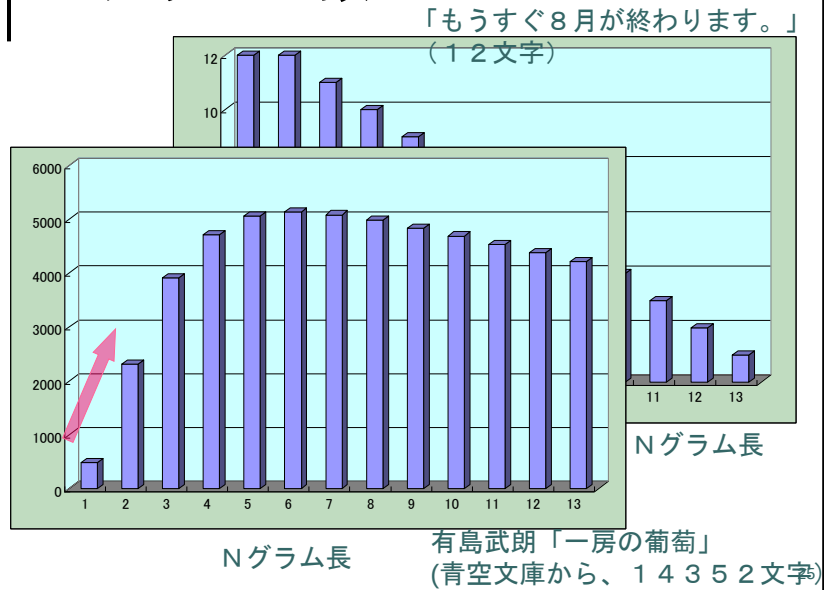
1 う す ぐ 8 月 が 終 わ り ま す 。
6 が 終 わ り ま す 。
3 ぐ 8 月 が 終 わ り ま す 。
2 す ぐ 8 月 が 終 わ り ま す 。
11 す 。

2005.8.25

24



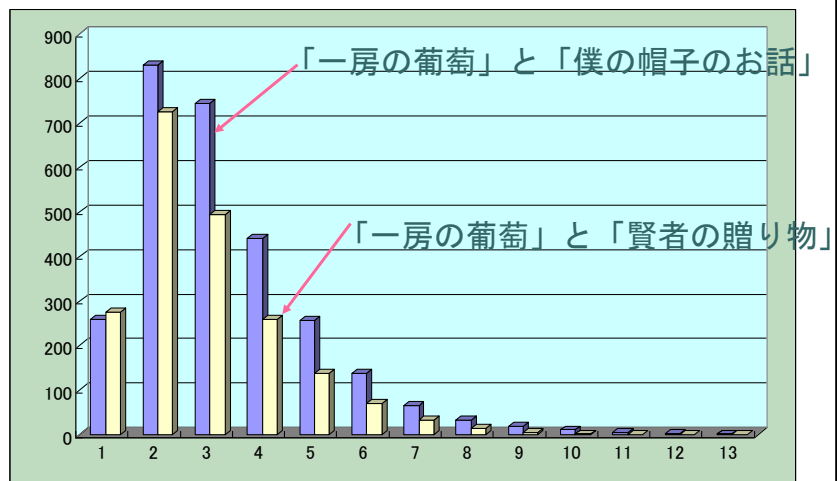
N グラムの数



2005.8.25



共通Nグラムのカウント

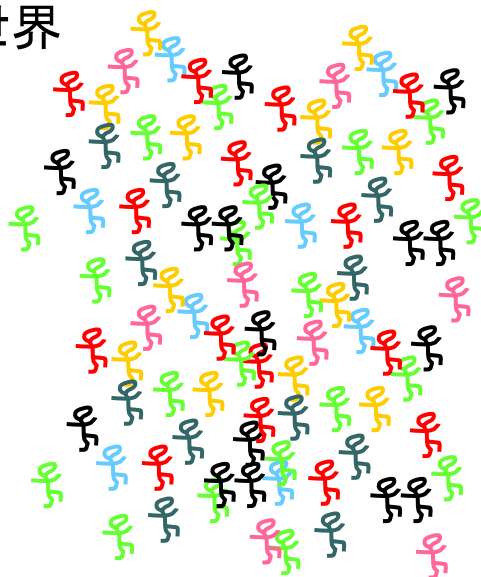


2005.8.25

26



高速な並べ替えを追求するアル ゴリズムの世界

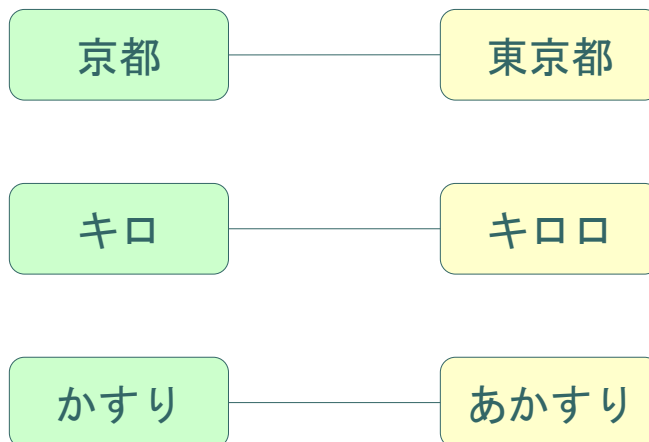


2005.8.25

27



N グラムの落とし穴



2005.8.25

28



「言語処理問題」を解 く コンピュータ



言語処理タスクの設定

N
I
I
市
民
講
座

入力テキスト



出力テキスト

簡単なキーワード抽出プログラム

テキスト入力インターフェース [[2画面表示へ](#)]

テキスト:

入力テキスト

☒ キーワード(最長)
 ☐ キーワード(最短)
 ☐ キーワード(すべて)

出力テキスト

2005.8.25

31

<http://chasen.naist.jp/hiki/ChaSen/>

コンピュータが空気と同じように、世界中に充満している状況を実現する手段の一つは、身に着けて持ち歩くことでしょう。これがもう一歩進むと、持ち歩こうと意識しなくても身についている、すなわち、パラサイト（寄生虫）のような存在になることです。

☑ 形態素解析

コンピュータ	コンピュータ	コンピュータ
が	が	が
空気	空気	空気
と	と	と

助詞-格助詞-一般	助詞-格助詞-一般	助詞-格助詞-一般
名詞-一般	名詞-一般	名詞-一般
助詞-格助詞-一般	助詞-格助詞-一般	助詞-格助詞-一般

2005.8.25

32



コンピュータが空気と同じように、世界中に充満している状況を実現する手段の一つは、身に着けて持ち歩くことでしょう。これがもう一歩進むと、持ち歩こうと意識しなくても身についている、すなわち、パラサイト（寄生虫）のような存在になることです。

☒ 構文解析

コンピュータが---D
空気と-D
同じように、-----D
世界中に-D |

2005.8.25

33



コンピュータが空気と同じように、世界中に充満している状況を実現する手段の一つは、身に着けて持ち歩くことでしょう。これがもう一歩進むと、持ち歩こうと意識しなくても身についている、すなわち、パラサイト（寄生虫）のような存在になることです。

☒ 用語抽出

コンピュータ (4), 空気 (3), 身 (2), 身体 (2), 実現 (1), 状態 (1), 究極 (1), 一つ (1), 手段 (1), 存在 (1), 形 (1), パラサイト (1), 意味 (1), 窒息 (1), 状況 (1), 意識 (1)

2005.8.25

34



コンピュータが空気と同じように、世界中に充満している状況を実現する手段の一つは、身に着けて持ち歩くことでしょう。これがもう一歩進むと、持ち歩こうと意識しなくても身についている、すなわち、パラサイト（寄生虫）のような存在になることです。

☒ 要約

例

コンピュータが空気になる



コンピュータが空気と同じように、世界中に充満している状況を実現する手段の一つは、身に着けて持ち歩くことでしょう。これがもう一歩進むと、持ち歩こうと意識しなくても身についている、すなわち、パラサイト（寄生虫）のような存在になることです。

☒ 自動翻訳

It will be one of the means to realize the situation full of a computer in the same way as air all over the world can touch the body, and to carry it. When this already goes step further, stick to the body even if is not conscious to carry it || in other words is to come to seem to be a parasite (a parasitic worm).



コンピュータが空気と同じように、世界中に充満している状況を実現する手段の一つは、身に着けて持ち歩くことでしょう。これがもう一歩進むと、持ち歩こうと意識しなくても身についている、すなわち、パラサイト（寄生虫）のような存在になることです。

☑ 関連書籍の検索

モバイルコンピューティング 塚本昌彦著 -- 岩波書店, 2000.12, v, 114p. -- (岩波科学ライブラリー; 77)

2005.8.25

37



最適化問題としての言語処理

入力テキスト

知らない言葉

規則外の文法

意外な言い回し

書き間違い

出力テキスト

2005.8.25

38



「Aに似たB」

「Aに似たB」という表現を捜せば、AとBが類似関係にあるという知識が得られる。

...に違いない。

ファミコンのコントローラに似たゲームパッド

あなたに似た人

太陽－木星に似た惑星系

水に似た感情

ペットに似た飼い主

下手の考え休むに似たる

2005.8.25

39



言語処理の中身 ～形態素解析の場合

http://chaSen.naist.jp/hiki/ChaSen/

「もうすぐ8月が終わりま す。」 形態素の解析

● ● ●

N
I
I
市民講座

もうすぐ	モウス グ	もうす ぐ	副詞-一般
8月	ハチガ ツ	8月	名詞-副詞可能
が	ガ	が	助詞-格助詞-一般
終わり	オワリ	終わる	動詞-自立 五段・ラ行 連用形
ます	マス	ます	助動詞 特殊・マス 基本形
。	。	。	記号-句点

2005.8.25

1

コンピュータの「言語知識」

● ● ●

N
I
I
市民講座

単語辞書

表記と品詞

もうすぐ	副詞
夏	名詞
が	助詞
終わり	動詞
ます	助動詞

品詞接続表

AにBが続くかどうか

A	B	
名詞	名詞	○
名詞	助詞	○
名詞	副詞	×

2005.8.25

42



あいまいなテキスト

N
「
」
市民講座

全国都道府県議会議長会

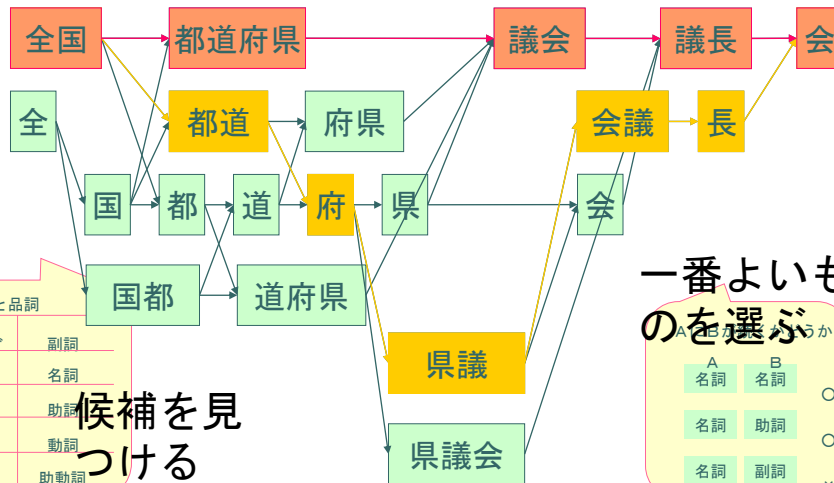
2005.8.25

43



最適パスを求める

N
「
」
市民講座



20



辞書と形態素解析

- 東（名詞-一般）
- 東北（名詞-固有名詞-地域-一般）
- 北東北
 - 北（名詞-固有名詞-地域-一般）
 - 東北（名詞-固有名詞-地域-一般）
- 北北東
 - 北（名詞-固有名詞-地域-一般）
 - 北東（名詞-一般）
- 東北東（名詞-一般）

2005.8.25

45



辞書資源の構築

人手による言語資源の構築

WordNet Online Search で調べた “school” の意味

Noun

- S: (n) school (an educational institution)
- S: (n) school, schoolhouse (a building where young people receive education)
- S: (n) school, schooling (the process of being formally educated at a school)
- S: (n) school (an educational institution's faculty and students)
- S: (n) school, schooltime, school day (the period of instruction in a school; the time period when schools is in session)
- S: (n) school (a body of creative artists or writers or thinkers linked by a similar style or by similar teachers)
- S: (n) school, shoal (a large group of fish)

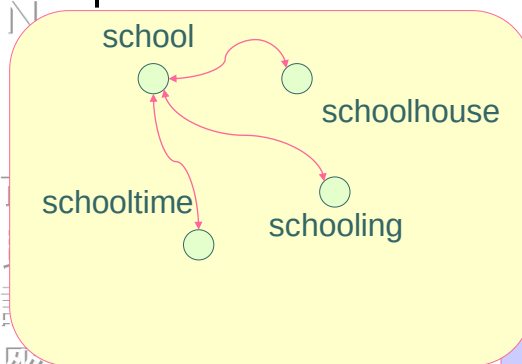
EDR の概念識別子の例

- 3bc262 学校という教育を行う組織
- 3cf8b1 学校という教育施設

2005.8.25

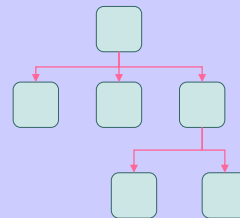
47

「意味」の記述方法



WordNet : 語のネットワーク

EDR : 概念階層

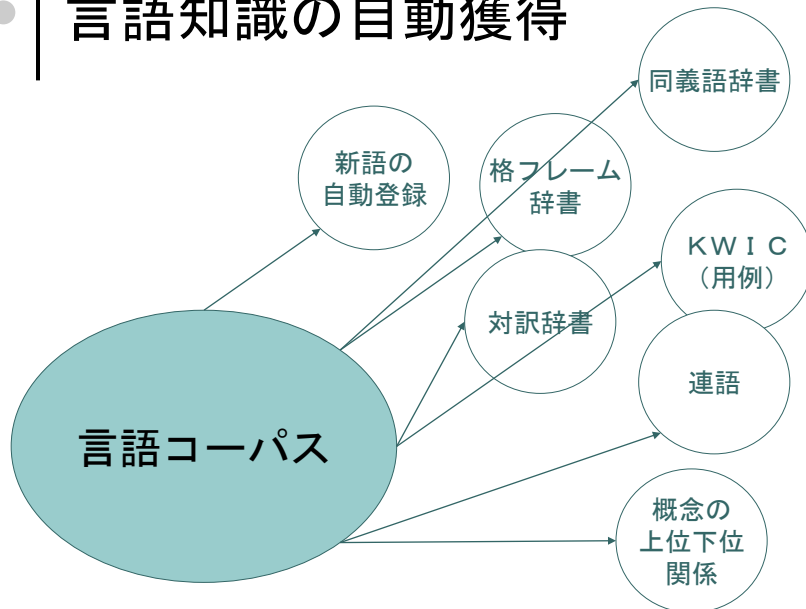


2005.8.25

48



言語知識の自動獲得



2005.8.25

49



KW I C (KeyWord In Context)

KWICFinder
<http://www.kwicfinder.com/KWiCFinder.html>

OP-ED: NURTURING EARTH: ' **MOTTAINAI**' AND THE FOURTH
 esnt have knob. from spirit of " **MOTTAINAI**" means ecology.. Po
 e sense and the word itself of " **MOTTAINAI**" come to get lost. T
 " come to get lost. This is rally **MOTTAINAI**. . Nice composition
 t) tripflyin says: thanx kaz-1,, **MOTTAINAI** is japanese soul to
 article title list IMPERIAL OIL & **MOTTAINAI** 1. Oil, Geopolitics,
 the Earth. Nobelist Marthai's " **MOTTAINAI**(wasteful)" 〇 〇 〇 〇
 OSAMASU wake up ATARASHI **MOTTAINAI** waste WOKASHI JO

2005.8.25

50



自動構築した学術用語オンライン辞書の例

学術用語対訳／類語オンライン辞書

検索したい学術用語(和or英)

和 テレビ会議 (17.8) **英** TV-conference (8.6)

和類 ▶ 通信会議 (11.8), 電子会議 (10.6), TV会議 (5), 遠隔会議 (4.8), ビデオ会議 (3.4), 画像会議 (3), テレコンファレンス (3)

英類 ▶ TELECONFERENCE (16.4), Video-conference (6.6), Tele conference (4.2), Electronic meeting (2.8), Electronic Conference (2.8), Teleconferencing (2.6), Videoconferencing (2), Videoconference (1.6), Video Conferencing (1.6), Electric Meeting (1), Computer Conference (1), Remote Conference (0.8)

2005.8.25

<http://mic.ex.nii.ac.jp/dict/> 51



新語との遭遇



テキストを支配するZipfの法則

- 語の頻度
- 科学論文の引用数
- Webのヒットカウント
- 書籍の販売数
- 電話の通話回数
- 地震の規模
- 月のクレーターの直径
- 太陽フレアのガンマ線ピーク値
- 戦争による死者
- お金持ちの人の財産
- 苗字の頻度
- 都市の人口

語の出現頻度
語の出現頻度
書のクラス間分布
イトあたり文書数
照リンク数

Mark Newman, "Power Laws, Pareto
Distributions and Zipf's Law" Condensed
Matter, cond-mat/0412004

2005.8.25

53



80-20ルール

パレートの法則：
社会全体の富の80%は20%の
人間に集中している。

rich-get-richer

乗算過程：変化量が現
在量に比例する過程。

2005.8.25

54



新語の割合の予測

- 未知の語が観察される割合は頻度 1 の語の割合に等しい。



複合語

固有名詞

外来語

...



規則にあわない新語

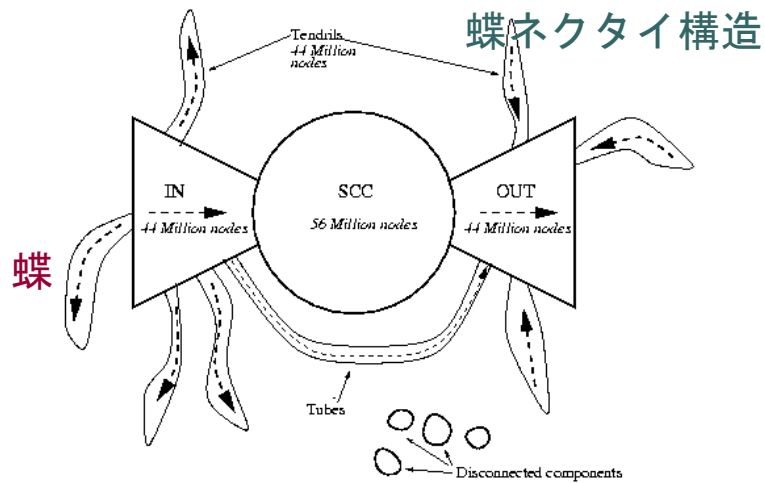
- 「乗り上がり脱線」
- 「想定地震」
- 「わだち掘れ」
- 「モーニング娘。」
- 「ペ・ヨンジュン」
- 「むない」
- 「e気持ち」
- 「王里イヒ学」



NI市民講座

ドラマチックな新語

A. Broder, R. Kumar et.al. : "Graph structure in the Web : experiments and models", Proc. of the 9th WWW Conf. (2000)



終わりに

● ● ●

身近に

小学2年生の国語の問題を解く
<http://www.languagecraft.net/dennou/>

ブログで社会の動き
<http://blogwatcher.pi.jp/>

電脳優子2年生
 国語

ブログで社会の動きをチェック!!

なんでもRSS メタブログ
 延べ11,833,069日のエン

専用ツールバー Mozilla/Firefox/Chrome

Copyright © 2003-2005 OKURA Lab, Precision and Intelligence Laboratory, Tokyo Institute of Technology. All rights reserved.

2005.8.25

59

● ● ●

市
民
講
座

Q & A

2005.8.25

60