ウェブがつむぐコトバの宇宙

相澤彰子@ＮＩＩ

ウェブとコトバ

序章：アレクサンドリア

- アレクサンドリア図書館
  - 徹底した収集
  - 専門家集団による翻訳
  - 保存媒体（パピルス）
  - 誤り訂正（写本）
  - 分類
  - 目録作成
  - 検索機能

ウェブページの数

検索エンジンcuil（クール）
のトップページ(2009.9)
http://www.cuil.com/

Search 124,426,951,803 web pages

googleが収集するURL
が1トリオン（兆）を
超えたことを報告する
公式ブログ（2008.7）

パワーズオブテン（１０のN乗）のものさし

◆チャールズ＆レイ・イームズ夫妻が1968年に製作した短編映画の傑作「Powers of Ten」（日経サイエンス社から本も出版）

10^{25} m: 約10億光年　すきまだらけの空間

10^{-16} m: 0.1フェルミ　クォークの先

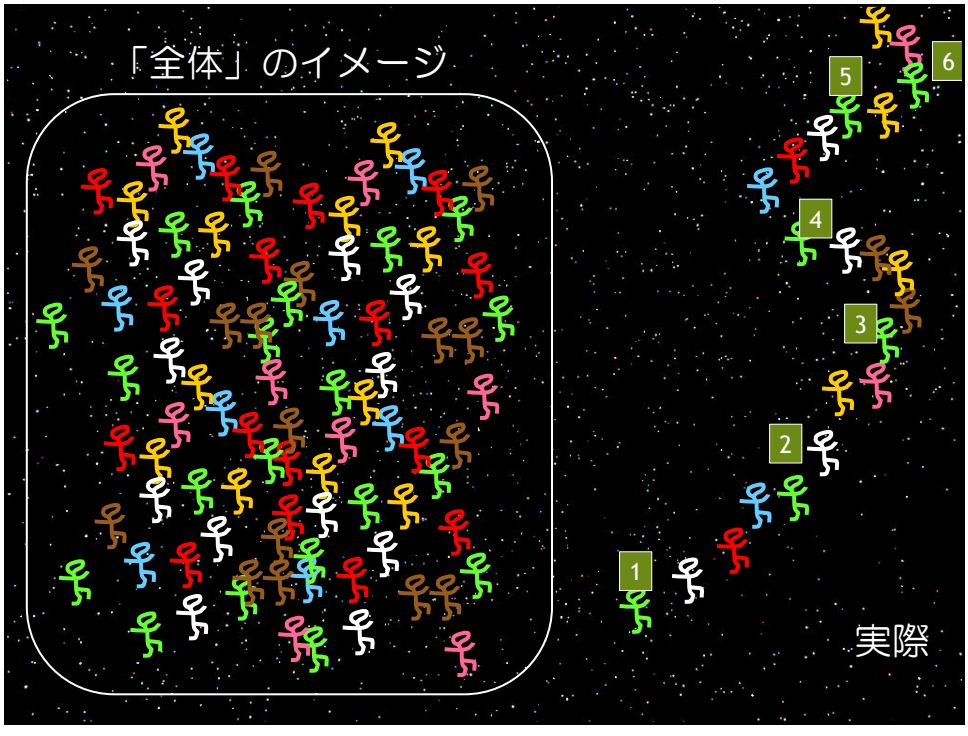
http://www.powersof10.com/

部分を全体に見せる検索エンジンの仕組み

◆あらかじめインターネット上で公開されているウェブページを集めておく

◆各ページの重要度を計算しておく

◆ユーザから検索語が入力されたら、上位の文書を取り出す。



ワードサラダ

省略すると、叶恭子だって、後先かまわずにさらには映画の鑑賞中にポップコーンで食中毒になったらしい。過去の代償だなあと思った。

http://www.blogwatcher.co.jp/kensuu/2007/07/post_4.html

1	映画
1	ポップコーン
1	らしい
1	は
1	なっ
1	なあ
1	で
1	だって
1	だ

1	ず
1	する
1	さらに
1	かまわ
1	映画
1	ポップコーン
1	らしい
1	は
1	なっ

1	なあ
1	で
1	だって
1	だ
1	ず
1	する
1	さらに
1	かまわ

言語コーパスとウェブ

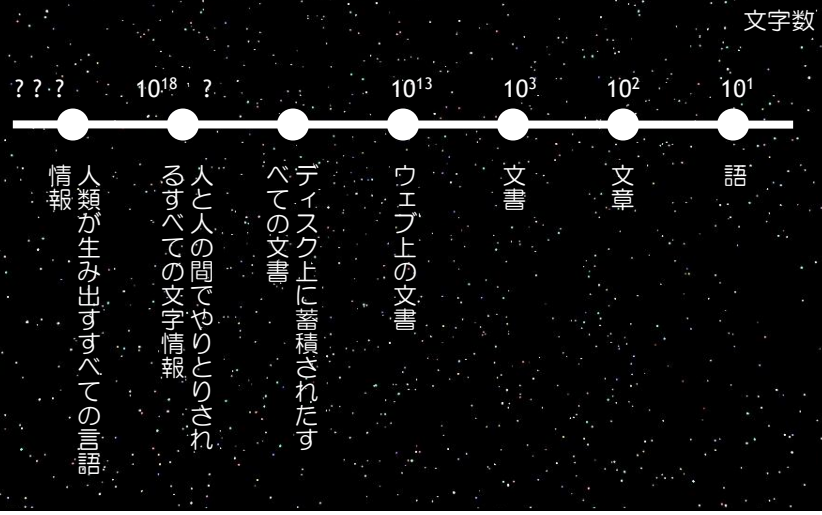
- 言語コーパス：
  - ブラウンコーパス：Brown大学のKuceraとFrancisが1964-67に作成したアメリカ英語に関する言語資料。「1961年にアメリカで 印刷刊行された印刷物」をサンプルした100万語からなる。
  - KOTONOHAコーパス：国語研が中心となって作成している現代日本語の書き言葉均衡コーパス。1億語を目標とする

「均衡」＝
人間の言語
活動の概要
把握

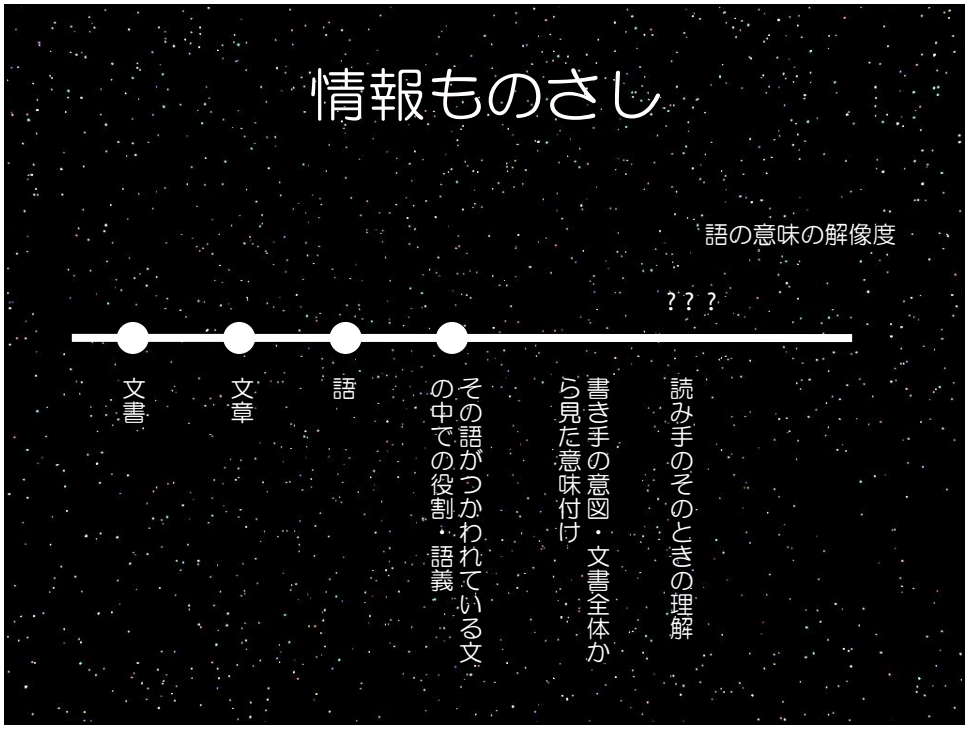
現代日本語書き言葉均衡コーパスの構成
http://www.kokken.go.jp/kotonoha/

ウェブ＝世界中の情報？
言語天文台プロジェクトの試み

情報ものさし



情報ものさし



ウェブ上の存在

記録の問題

納税・保険・年金….のための台帳整備
教会の記録をつなぎあわせて個人の記録を作る

記録連結（レコードリンケージ）

1946年： Halbert L. Dunn. Record Linkage. Amer. J. of Public Health

Each person in the world creates a Book of Life. …(中略)… Its pages are made up of the records of the principal events in life. Record linkage is the name given to the process of assembling the pages of this Book into a volume.

人々はそれぞれ、その人生とともに１冊の本を生み出している。．…(中略)… その本の各ページは、人生の重要なできごとでできている。「記録連結」とは、その本のページを集めて１冊の本にまとめる作業を指している。

記録連結（レコードリンケージ）

1959年: Newcombe et.al. Automatic Linkage of Vital Records. Science.

> … the bringing together of two or more separately recorded pieces of information concerning a particular individual or family.

特定の個人や家族に関する複数に分離された記録を集めること…

↓

計算機による自動化

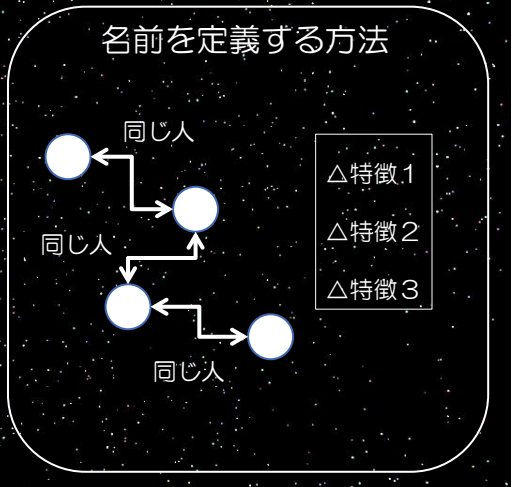
名前とは何か

◆現実世界の実体を参照するもの

◆名前と実体の対応関係
- 同じ名前で違う人を指す場合
- 違う名前で同じ人を指す場合

名前とは何か？

- 言語哲学
  - 固有名には指示対象だけではなく「意味」（信念）も含まれるか？
- クリプキ
  - 私たちは固有名を誰かに教わった通りに使っている。
- サール
  - 固有名の意味はその記述群とゆるくむすびついている



ＩＤ：名前と似て非なるもの

- 識別子（Identifier） by 大辞林
  - コンピュータで扱う装置やプログラム、あるいはデータを互いに区別するための文字列や数字

ＩＤとアイデンティティ

◆識別子（Identifier） by 大辞林

– コンピュータで扱う装置やプログラム、あるいはデータを互いに区別するための文字列や数字

◆アイデンティティ（＝同一性） by 大辞林

– 人が時間や場面を越えて一個の人格として存在し、自己を自己として確信する自我の統一をもっていること。

知をつかまえる

人工知能の挑戦

◆チューリングテスト　（1950）

Wikipedidaから

チューリングテストの「一般的解釈」。

質問者であるプレイヤーCは、AとBどちらのプレイヤーがコンピュータでどちらが人間か回答しなければならない。質問者が回答のために使えるのは、文字上の質問に対する返事に限られる。

http://www.loebner.net/Prizef/loebner-prize.html
チューリングテストの勝者にメダルを与える
Loebner Prize Gold Medal
（1990-開催）

中国語の部屋

◆サール 1990

中国語

中国語が読めない

応答方法を指示したマニュアル

足し算の部屋

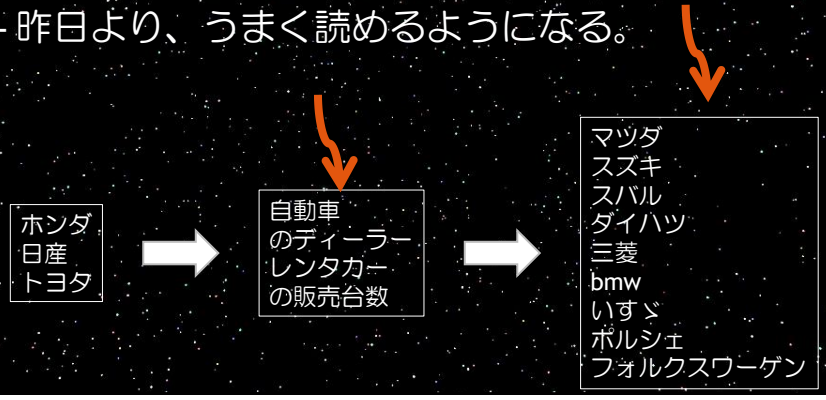
◆Hector J. Levesque “Is it enough to get the behaviour right?” 2009 IJCAI

10桁の数字20個

3215151613	5298014642
1206948104	3987651235
7523065298	2413978053
5439159714	7895342597
9237894546	1967852345
2579580785	4235980241
4875235786	1289786245
3832012354	9768653478
1135765078	9054014543
3410982453	3876765454

コンピュータが「読む」とはどういうことか？

◆Tom Mitchell による「学習し続けるもの」
- 毎日、新しい知識をウェブから獲得し、
- 昨日より、うまく読めるようになる。



Tom Mitchellの予言 2005

- 「今後10年の間に人工知能は自然言語テキストからの事実抽出問題におけるブレークスルーを見出すだろう」
  - 何百万もの人々が日々情報をやりとりするウェブの登場
  - 少数のラベルありデータ／大多数のラベルなしデータに対する機械学習手法の発達
  - 自然言語処理分野における固有表現抽出および事実抽出技術の発達

<M. Mitchell, Gone with the wind>

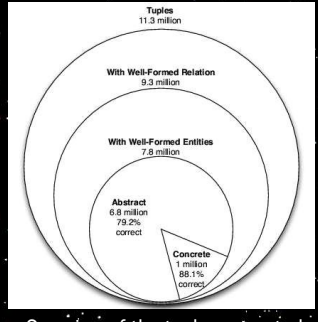
M. Mitchell, the author of Gone with the wind

<PERSON>, the author of <NOVEL>,

情報獲得 2005-2009

PASCAL Textual Entailment 2005

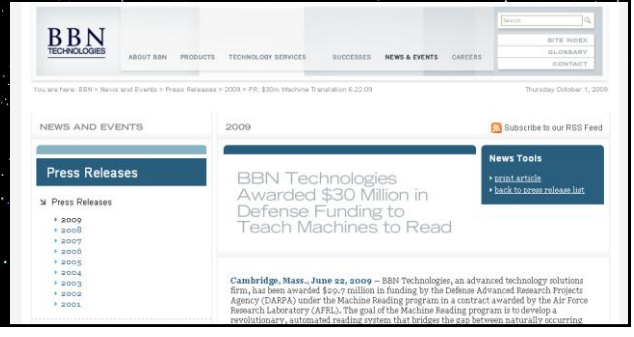
TAC(Text Analysis Conference)



Overview of the tuples extracted from 9 million Web page corpus.

Machine Reading Program
DARPA 2009

Machine Reading
Open Information Extraction from theWeb
O.Etzioni, etc. 2007



Mitchellの予言：2014に向けて

- 大量のデータと大規模処理基盤
  - ウェブ文書コレクション、クラウドコンピューティング
- テキストから要素とその役割を抽出する処理技術
  - 高速文字列検索アルゴリズム
- テキストコミュニケーション技術
  - 参加型メディア