テキスト世界の歪み：文章から現実世界を測るということ

荒牧英治
東京大学 知の構造化センター
JST さきがけ「情報環境と人」

自己紹介

- 学部: 京大 総合人間学部（山梨研）
- 修士: 〃　言語メディア研（黒橋研）
- 博士: 東京大学 情報理工（黒橋研）

自然言語処理

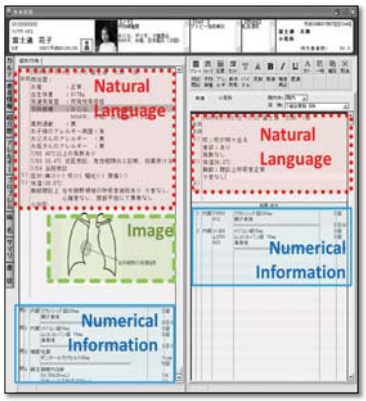
- 助教（2006-2008）
  - 東大 医学部附属病院
- 講師 (2008-)
  - 東大 知の構造化センター

医療分野での言語処理研究に従事

電子化 ≠ 標準化

電子カルテは Natural Language（自然言語）を含んでいる

病院ごとに，各診療科ごとに，
（場合によっては）医師ごとに
表現が異なる



様々な表現にバリエーションを扱うために
言語処理（Natural Language Processing; NLP）
が必要

マルケサーニ症候群
マルケサニー症候群
マルケサニ症候群
マルケザーニ症候群
マルケザニー症候群
マルケザニ症候群

marchesani症候群の同義語テーブル（標準病名マスター）

誤って筋弛緩剤投与、患者死亡　徳島・鳴門の病院

2008年11月19日（水）23:24　asahi.com

徳島県鳴門市撫養（むや）町黒崎の健康保険鳴門病院（増田和彦病院長）は１９日、誤って抗炎症剤ではなく筋弛緩（きんしかん）剤を点滴で投与された７０代の男性患者が死亡した、と発表した。当直医が電子カルテに薬剤の名称を記入した際に誤表示され、そのまま薬剤師が用意してしまったのが原因で、蘇生を試みたが意識が戻らなかったという。病院から届け出を受けた県警は業務上過失致死の疑いもあるとみて、医師ら関係者から事情を聴いている。

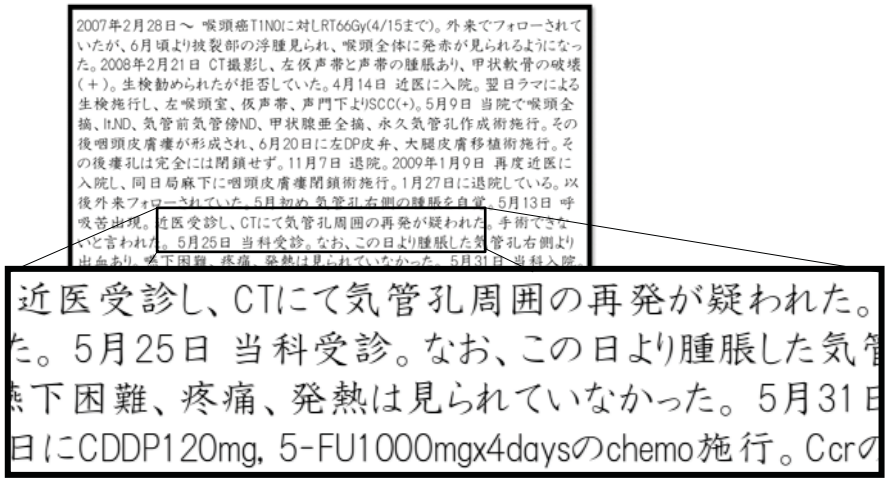
同病院によると、この男性患者は肺炎と胸膜炎で入院していた。１７日午後９時ごろ、３９度を超える発熱があったため、看護師が当直医に連絡。当直の３０代の女性医師は、患者のアレルギー体質を考慮して抗炎症剤の副腎皮質ホルモン「サクシゾン」の投与を決め、電子カルテのパソコン端末に記入。その際、最初の３文字（サクシ）だけを入力して薬剤名を検索したが、同病院でサクシゾンは扱っていなかったため、画面には筋弛緩剤の「サクシン」だけが表示された。

>>　続きは asahi.comへ

サクシゾン ⟷ サクシン

(スキサメトニウム)

医療テキスト（入力）
患者の現病歴記述から


2007年2月28日～ 喉頭癌T1N0に対しLRT66Gy(4/15まで)。外来でフォローされていたが、6月頃より披裂部の浮腫見られ、喉頭全体に発赤が見られるようになった。2008年2月21日 CT撮影し、左仮声帯と声帯の腫脹あり、甲状軟骨の破壊（＋）。生検勧められたが拒否していた。4月14日 近医に入院。翌日ラマによる生検施行し、左喉頭室、仮声帯、声門下よりSCC(+)。5月9日 当院で喉頭全摘、ltND、気管前気管傍ND、甲状腺亜全摘、永久気管孔作成術施行。その後咽頭皮膚瘻が形成され、6月20日に左DP皮弁、大腿皮膚移植術施行。その後瘻孔は完全には閉鎖せず。11月7日 退院。2009年1月9日 再度近医に入院し、同日局麻下に咽頭皮膚瘻閉鎖術施行。1月27日に退院している。以後外来フォローされていた。5月初め 気管孔右側の腫脹を自覚。5月13日 呼吸苦出現。近医受診し、CTにて気管孔周囲の再発が疑われた。手術できないと言われた。5月25日 当科受診。なお、この日より腫脹した気管孔右側より出血あり。嚥下困難、疼痛、発熱は見られていなかった。5月31日

近医受診し、CTにて気管孔周囲の再発が疑われた。
た。5月25日 当科受診。なお、この日より腫脹した気管
嚥下困難、疼痛、発熱は見られていなかった。5月31日
日にCDDP120mg, 5-FU1000mgx4daysのchemo施行。Ccrの

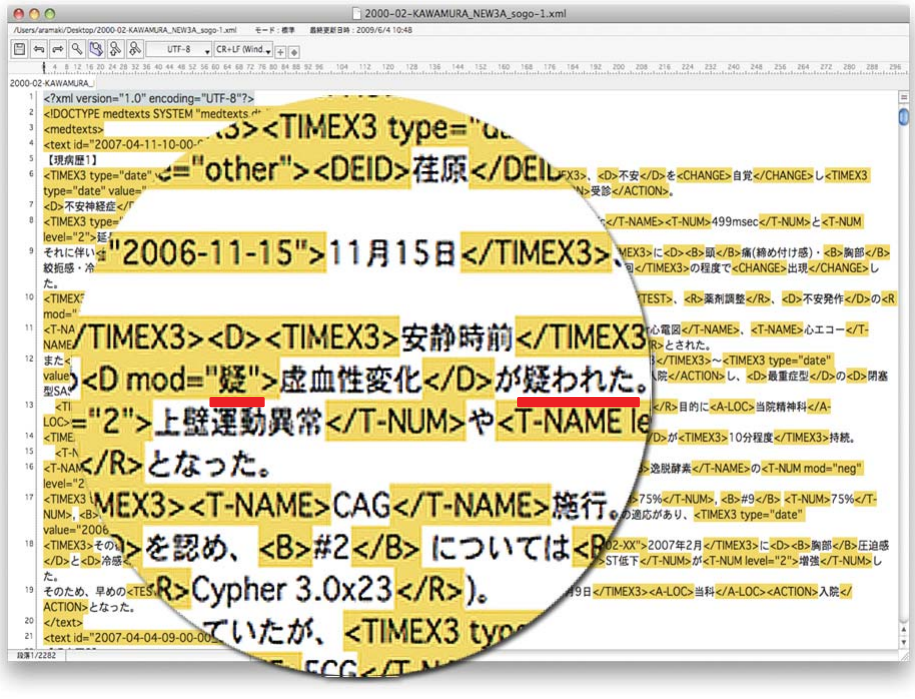
（医師により作成された）ダミーのカルテ文章

言語処理結果（出力）
いつ/何があったのかを抽出

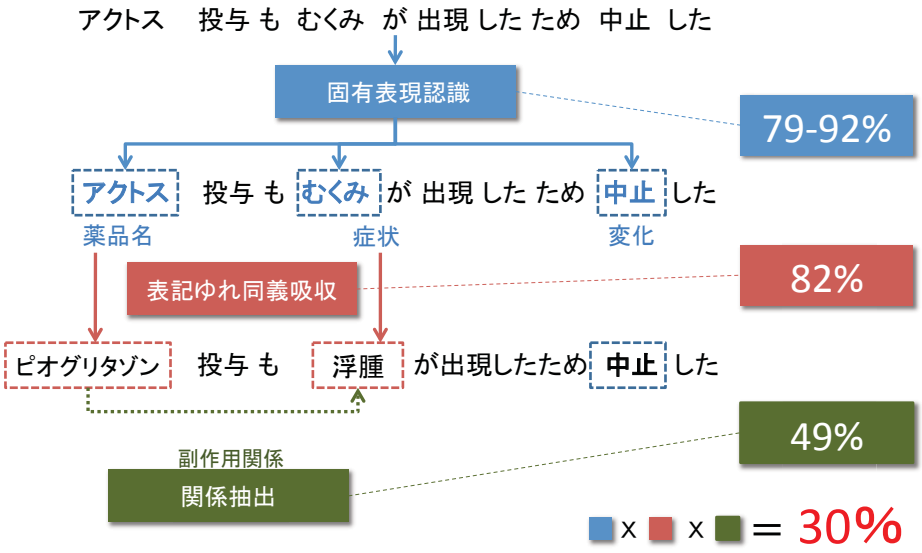
日付	施設名[科]	入退院	疾患	変化	治療
2009-01-09（2009年1月9日）					
	近医	入院			
			咽頭皮膚瘻		咽頭皮膚瘻閉鎖術
2009-01-27（1月27日）		退院			
		外来フォロー			
2009-05-XX（5月初め）			腫脹		
2009-05-13（5月13日）	近医		呼吸苦		
		受診			
			再発（気管孔周囲）		
					手術
2009-05-14（翌日）					

いつ　どこで　患者の行動　医師の診断　医師の処置

事実　否定　疑い ▪▪▪　適応なし　希望　予定

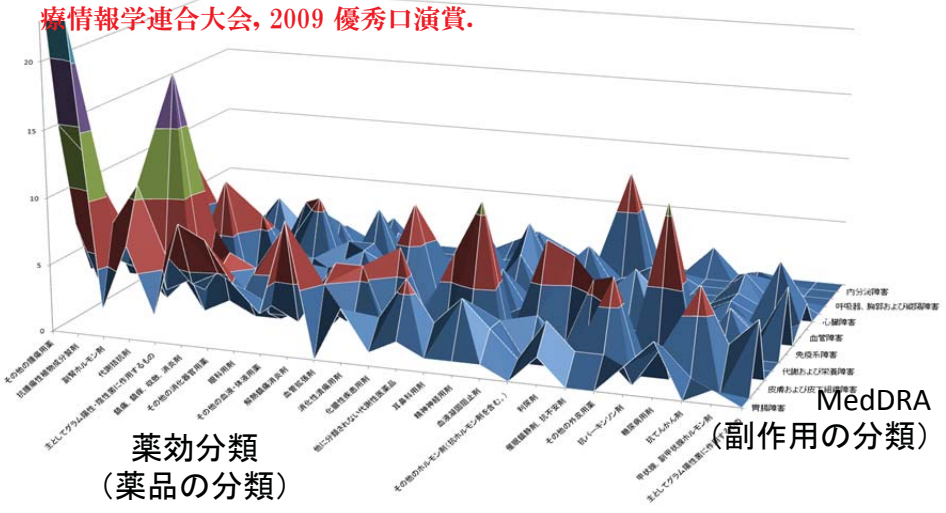


副作用関係の自動抽出



東大病院での副作用情報抽出

外池昌嗣, 大熊智子, 荒牧英治, 三浦康秀, 増市博, 大江和彦: 自然言語表現の現病歴情報を時系列表形式で表示するシステムとその評価, 第29回医療情報学連合大会, 2009 優秀口演賞.



【内科学会;循環器学会】症例報告の有効活用に関する研究

[荒牧, 大江2009]

- 二つの学会に同義／表記ゆれ検索機能をもった症例検索システムを提供

社団法人 日本内科学会 The Japanese Society of Internal Medicine　社団法人 日本循環器学会

高速 症例くん

現在データベースには、すでに約25,000件の演題が登録されており、毎年、新たに約3,100件の演題が追加登録

http://member.naika.or.jp/member/content/ninsho1/search.html

自動匿名化に関する研究[荒牧+; 2006]

- カルテは個人情報のかたまり
  → 研究利用する前に個人情報を除く必要がある
- 個人情報とは？

<DATE>9月12日</DATE>, <HOS>東大病院</HOS>紹介受診. . .

AGE	HOSPITAL
DATE	ID
DOCTOR	PHONE
PATIENT	LOCATION

HIPPAの定義による

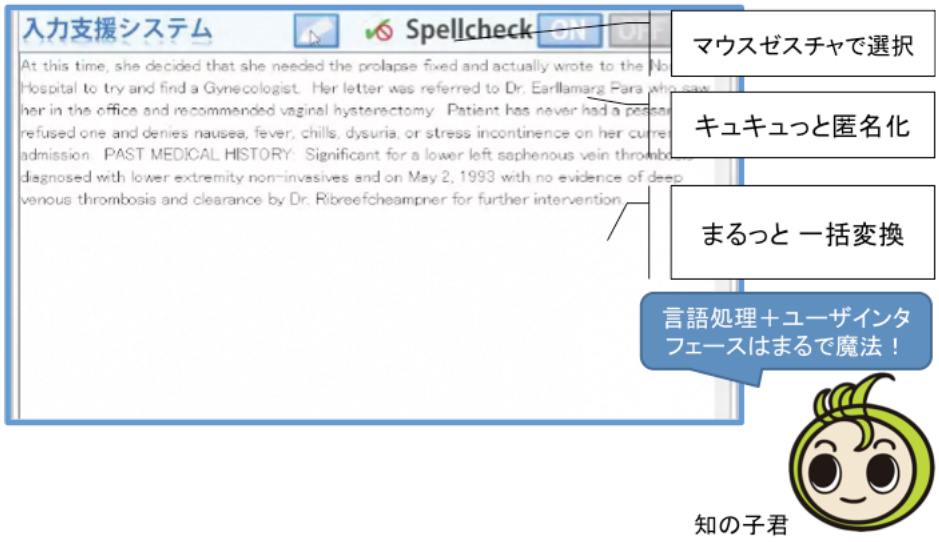
	適合率	再現率	F値
人間	99.6	**95.9**	**97.7**
提案手法	98.3	**96.4**	**97.3**

人間の精度は [Dorr2006]による

匿名化ツール

より使いやすく！

[宮部+2011]



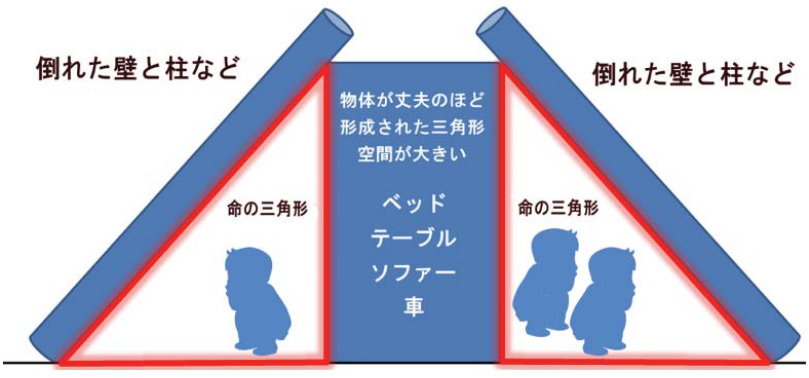
日々蓄積される大量の医学知識（上）と臨床知識（下）

流言の広がるパターン

甲南大灘本先生との共同研究

拡散する「命の三角形」情報

命の三角形：地震が起こったときに身を守る方法



流言の訂正

RT @atomfe: .@dfnt [命の三角形]は地震で建物が倒壊した場合も有効ではありません。「揺れがなく倒壊」した時のみ有効(かもしれない)です。揺れによってベッドや車が移動し押しつぶされる可能性のほうが高いためです。http://bit.ly/bxs2vI

命の三角形

尾田栄一郎氏の寄付(左)と社内サーバ(右)

ワンピースの作者　尾田栄一郎氏、地震の被害者救済に15億円を寄付 「自分が幸せになったということは、世の中から受けたひとつの借りだ」

地震が起きた時、社内サーバールームにいたのだが、ラックが倒壊した。腹部を潰され、血が流れている。

関西電力節電呼びかけ

- 関西電力で働いている友達からのお願いなのですが、本日18時以降関東の電気の備蓄が底をつくらしく、中部電力や関西電力からも送電を行うらしいです。一人が少しの節電をするだけで、関東の方の携帯が充電を出来て情報を得たり、病院にいる方が医療機器を使えるようになり救われます！

より大規模に
日本中の人々から疾患情報を得る

医薬基盤研究所 森田先生との共同研究
エスエス製薬への技術提供

2011, CLIO Healthcare Awards 2011, Gold Awards (最高賞) 2011, Spikes Asia Advertising festival 2011, デジタル部門 Gold Awards (第二位)（技術提供サイト「カゼミル」による）

msn 産経ニュース

トップ　速報　事件　政治　経済・IT　国際　スポーツ　エンタメ
暮らし・トレンド　からだ　教育　皇室　学術・アート　ブックス

【松田聖子さんのCMで話題】ぷるぷる肌ジェリーお試し５００円

［ライフ］ニュース

インフルエンザ、来週にも流行入り　ピーク
旬～２月上旬　国立感染症研究所

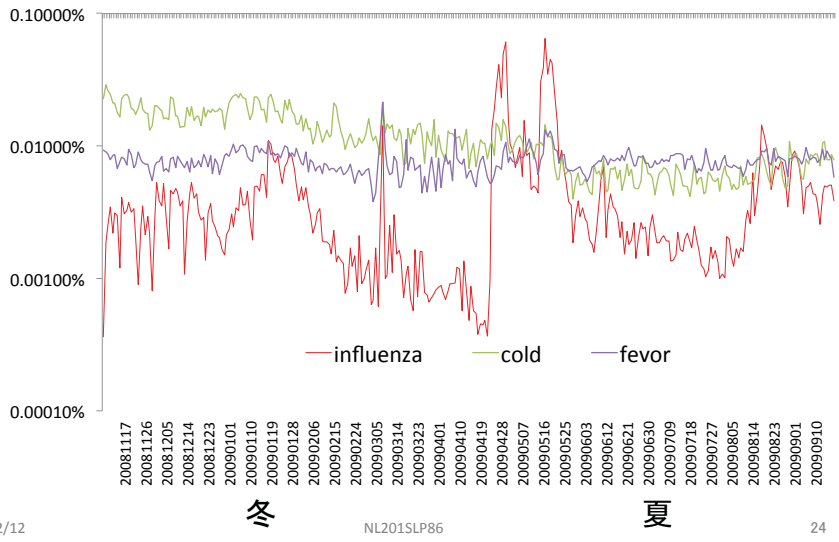
2010.12.17 21:32

インフルエンザの患者報告が、１２日までの１週間で１医療機関当た
り、８週連続で増加しているいることが１７日、国立感染症研究所の調
染研では１医療機関当たりの患者数が１人を超えると「流行入り」と判
まま感染者が増え続ければ、来週にも全国的な流行入りが発表される見

調査の集計には時間がかかり、１週間前の流行状況が発表されるため、実際はす
行入りしてい
染研感染症
かる。今な
る。

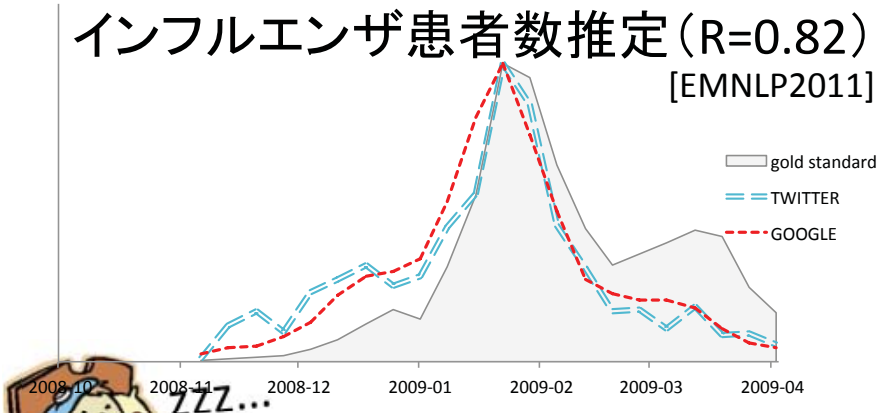
調査の集計には時間がかかり、１週間前の流行状況が発表されるため、実際はすでに流行入りしている可能性もある。

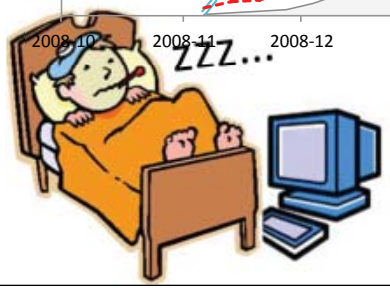
単語の頻度
「風邪」「熱」「インフルエンザ」



Webを用いた インフルエンザ患者数推定（R=0.82）

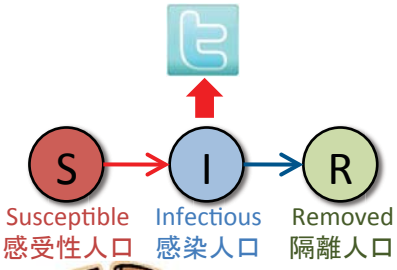
[EMNLP2011]



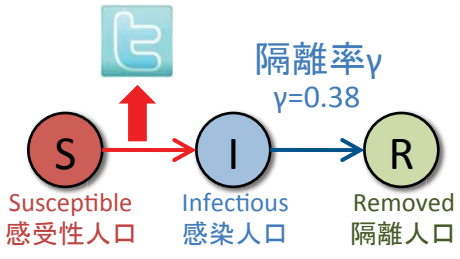


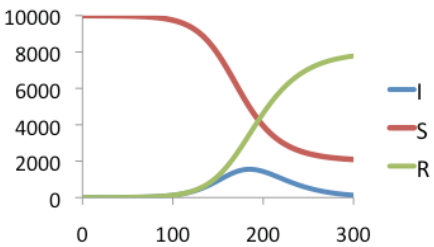
病人はネットを使わない
a Sick Person is Invisible in the Web

従来モデルとSIRモデル



r=0.821





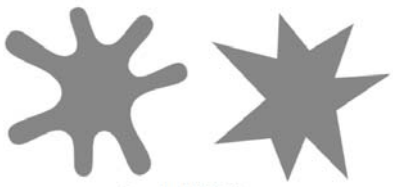
r=0.834

現実世界と知覚世界
現実世界とテキスト世界

最強のポケモン

名前によってモンスターの＜強さ＞がどの程度伝わるのか？

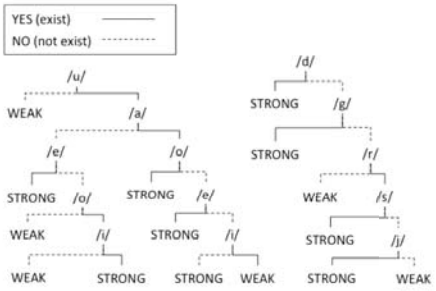
http://mednlp.jp/~baba/pokemon/



***Bouba-Kiki* Effect**

The 95% of examinee calls the shape on the RIGHT "*Kiki*" and the one on the LEFT "*Bouba*". This phenomenon is called "*Bouba-kiki*" effect (Ramachandran & Hubbard 2001).

```
Which is strong (j)"pippi" or (k)"poppo"?
Please a key [j] or [k].
>j
```



This rule shows the high agreement 0.704, which is the almost same agreement with inter-human agreement (0.692).

No.	アイアンマン	Score	バットマン
Human 1		-0.177023	
Human 2		2.58551	
Human 3		-1.35475	
Human 4		0.48017	
Human 5		1.56725	
Human 6		1.54221	
Human 7		0.443805	
Human 8		1.13444	
	2	Points	6

Name2 バットマン win

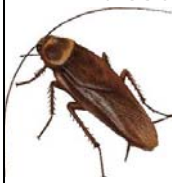
名前の強さを機械学習し，名前が強くなる原因を調査する．
①音素が重要 ②位置は無関係 ③長いと強い

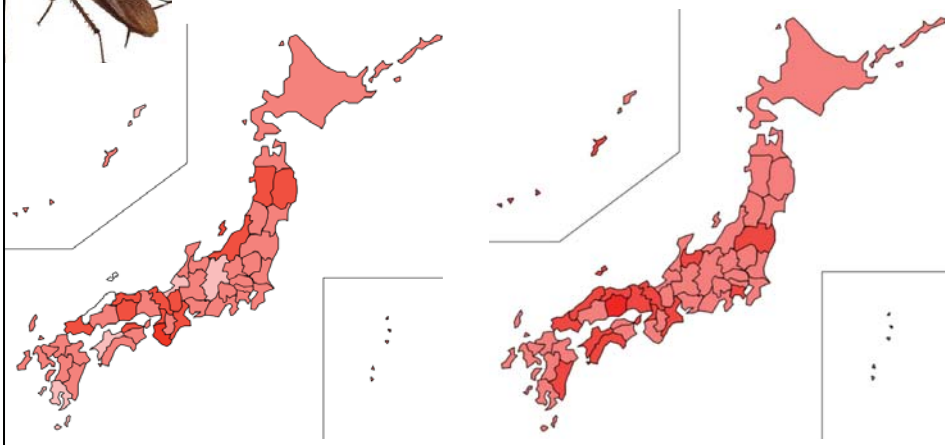
Twitterによる動物生態地図の作成

http://mednlp.jp/~baba/animalByPref/

003	鶫	708	499	20	0	0
004	鵠の鳥	0	0	0	0	0
005	トキ	22852	18011	187	22	27
006	朱鷺	18359	10879	147	6	79
007	鴇	6346	5289	21	0	0
008	ハヤブサ	11323	6874	165	34	44
009	隼	21564	14941	240	27	34
010	鵠	4	0	0	0	0
011	イリオモテヤマネコ	4702	2338	155	21	0
012	西表山猫	422	230	0	0	0
013	アマミノクロウサギ	740	407	0	0	0
014	奄美野黒兎	0	0	0	0	0
015	ゲンゴロウ	2123	1514	9	9	3
016	源五郎	3190	2439	102	19	0
017	トンボ	23460	15038	283	27	75
018	蜻蛉	20382	15849	180	15	27
019	とんぼ	20110	12040	256	31	58
020	クワガタムシ	1611	1227	0	0	0

Twitterによる動物生態地図の作成
ゴキブリ





夏　　冬

テキスト世界と現実世界の差異

―動物の部位分布における3つのプロトタイプ効果―

保田祥・岡本雅史・荒牧英治

「認知言語学論考11」

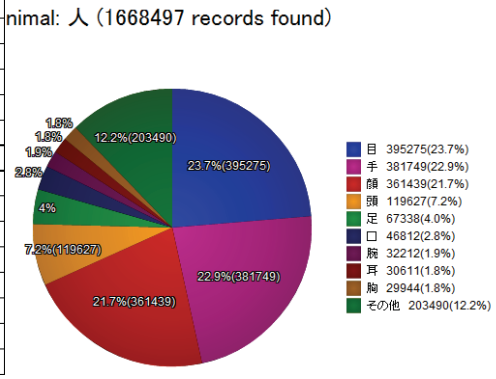
(本年度刊行予定)

テキスト世界のホムンクルス

[保田，岡本，荒牧;2012]

- Google n-gramデータにて「人の＜人体部位＞」を調査
- 例：人の腕，人間の腿，etc

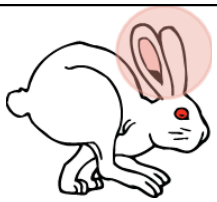
身体部位	体表面積	テキスト出現率	補正面積	補正倍率
腿	15.85%	0.01%	0.04%	0.003
腕	13.75%	3.01%	12.43%	0.904
脚	13.10%	0.66%	2.58%	0.197
背	12.25%	2.12%	7.80%	0.636
尻	9.70%	0.26%	0.76%	0.079
足	7.20%	3.99%	8.62%	1.198
胸	6.80%	1.55%	3.17%	0.467
臍	5.85%	0.32%	0.56%	0.096
手	5.20%	22.98%	35.90%	6.903
頭	4.25%	7.34%	9.37%	2.206
顔	2.80%	20.44%	17.19%	6.139
頸(首)	2.55%	0.99%	0.76%	0.296
耳	0.50%	2.63%	0.39%	0.789
鼻	0.086%	0.51%	0.01%	0.154
目	0.049%	25.14%	0.37%	7.540
口	0.041%	2.70%	0.03%	0.802



＜動物＞の＜部位＞にて調査

- 動物
  - ＜犬＞＜猫＞など200種
- 身体部位
  - ＜足＞＜手＞など100部位
- 組み合わせの用例
  - 「犬の尾」、「猫の手」など
- Google n-gramコーパスにて頻度調査
- 13,907,774用例（1動物1表記あたり，平均32,034の用例）が収集された
- http://mednlp.jp/~baba/animal
  →現実との違いは3つの現象に分類可能

兎の耳（Rabbit-Ear）現象



- 兎の耳（30%）
- 白鳥
  - 5%以上の出現率の部位を見ると、翼・羽（30%）・首（27%）・足（11%）・頭（7%）・嘴（6%）があり、用例の8割を超える
- ライオン
  - 顔（16%）、たてがみ（14%）、口（14%）、歯（9%）、頭（8%）
- その動物のもっとも特徴的な部分があたる
- 従来「プロトタイプ効果 (Lakoff1987)として説明されてきた現象

兎の耳現象と Positive Prototype

- 兎の耳現象は，典型的な属性がテキストに現れやすく，最も言及される頻度が高いことを示している
- なぜ，これらがよく言及されるのかは，たとえば兎の耳は、上位カテゴリである草食系哺乳網の他の成員（例えば，羊，リス，など，本稿ではこれらを**兄弟メンバ**と呼ぶ）との差異となり，メタファーなどに多用される

吾一は眼をこすった。向うの空が兎の耳のように、薄く色づいてきた 。（山本有三「路傍の石」）

兎の角（Rabbit-horn）現象

- 兎の角（1%）
- 現実には在るはずのない部位が、テキストでのみ言及される
- 蛇
  - 皮（22%）・頭（17%）・尾（14%）・舌（9%）・鱗（6%）・口（5%）・目（5%）に続き、足（4%）
- カバの角（1%）
- 幽霊の足（11%）
- 慣用化「兎に角（←兎角蛇足）」
- 「もしカバに角があればサイ」

兎の角現象とNegative Prototype

- 兄弟メンバと比べた際，その動物だけが持っていないような属性もまた，差異として知覚され，言及される頻度が増す(Cf。Levinson 2000)
- たとえば、蛇は足を持たないが，その兄弟メンバであるワニ，カメ，トカゲなどすべて足を持っており，それらとの顕著な（典型的な）差異が蛇に足がないことである
- よって，ないことが知覚され言及される

兎のヒゲ（Rabbit-whisker）現象

- 兎のヒゲ（0.07%）
- 虎
  - 目（27%）・尾（26%）・牙（10%）・皮（10%）
  - 現実的には当然有しているはずの足や耳（2%）が出現しにくい
- ミミズク
  - 目（44%）・耳（33%）・頭（12%）・顔（11%）のみ
- あるはずの部位が無視される現象

兎のヒゲ現象と Common Prototype

- 兄弟メンバとの差異が言語化されるという現象
- 逆に，差異でない属性は，めったに言語化されない
- 例えば，ヒゲのような身体部位は、兎の兄弟メンバ，例えば，鹿やヤギに一般的であり，差異となっていない

テキスト世界と現実世界の分類

	テキスト		
現実		記述される	記述されない
	存在する	兎の耳	兎のヒゲ
	存在しない	兎の角	N/A

現実世界ー認知世界＝テキストとして算出

おわりに

まとめ

- 「事実というものは存在しない．存在するのは解釈のみである」ニーチェ
- 「多くの人は、見たいと欲する現実しか見ていない」カエサル
- 「今よりも幸せな未来を想像できないからこそ、現在の幸福感と不安が両立するのだ」古市憲寿 → 発言と事実が反転する可能性

→書かれた結果（テキスト）のみを扱う危険性