

KOTONOHA

『現代日本語書き言葉均衡コーパス』
へのブログデータの追加

大学共同利用機関法人人間文化研究機構

国立国語研究所 言語資源研究系長

前川 喜久雄

KOTONOHA

書き言葉

書籍
新聞
雑誌
WEB

言文一致完成期の
総合雑誌700万語

「太陽」コーパス

「近代女性雑誌」
コーパス

2011年以降の
開発候補

現代日本語書き
言葉均衡コーパス

1875

1900

1925

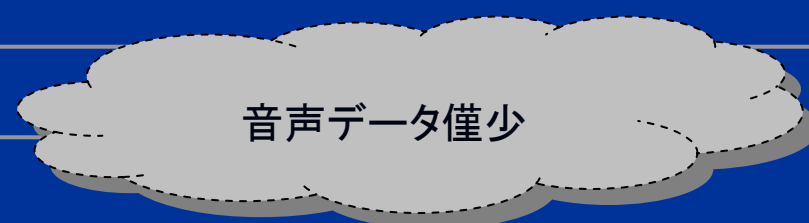
1950

1975

2000

話し言葉

独話
対話
雑談



現代の独話
752万語

日本語話し言葉
コーパス

2011年以降
の開発候補

日本語史上の
できごと

言文一致運動

現代かなづかい
当用漢字

JISコード

常用漢字

『現代日本語書き言葉均衡コーパス』

Balanced Corpus of Contemporary Written Japanese (BCCWJ)

出版(生産実態)サブコーパス

2001～2005年に出版された
書籍、雑誌、新聞

3500万語

図書館(流通実態)サブコーパス

東京都の13自治体以上の図書館に
収蔵されている書籍

対象期間: 1986-2005

3000万語

特定目的(非母集団)サブコーパス

白書、教科書、国会会議録、ベストセラー、[Yahoo!知恵袋](#)、等
対象期間はさまざま、最長30年。 3500万語

- 全体で1億語以上、著作権処理を施して公開する
- 3種のサブコーパスから構成
- ウェブ上のテキストとして「Yahoo!知恵袋」(500万語)を含む

『現代日本語書き言葉均衡コーパス』の開発

- 開発期間

2006～2010年度の5年間

- 開発費

国語研運営費交付金

文科省科学研究費補助金特定領域研究「日本語コーパス」

- 進捗状況(2010年3月の実績)

テキストデータ約9000万語以上を作成し、XML化

そのうち約5000万語の著作権処理を終了

著作権処理済の約4600万語を検索デモサイトで一般公開

検索デモンストレーションサイト

http://www.kotonoha.gr.jp/demo/

KOTONOHA「現代日本語書き言葉均衡コーパス」 検索デモンストレーション - 検索結果 - Windows Internet Explorer

http://www.kotonoha.gr.jp/cgi-bin/test/search_result.cgi?word=NE3%8D%82%8E3%81%A8%8E3%81%84%8E3%81%86%8E3%81%8B&genre_all=%E7%99%BD%E6%9B%B8&ger

Google

KOTONOHA「現代日本語書き言葉均衡コーパス」 検...

独立行政法人 国立国語研究所

KOTONOHA

「現代日本語書き言葉均衡コーパス」

検索デモンストレーション

検索結果

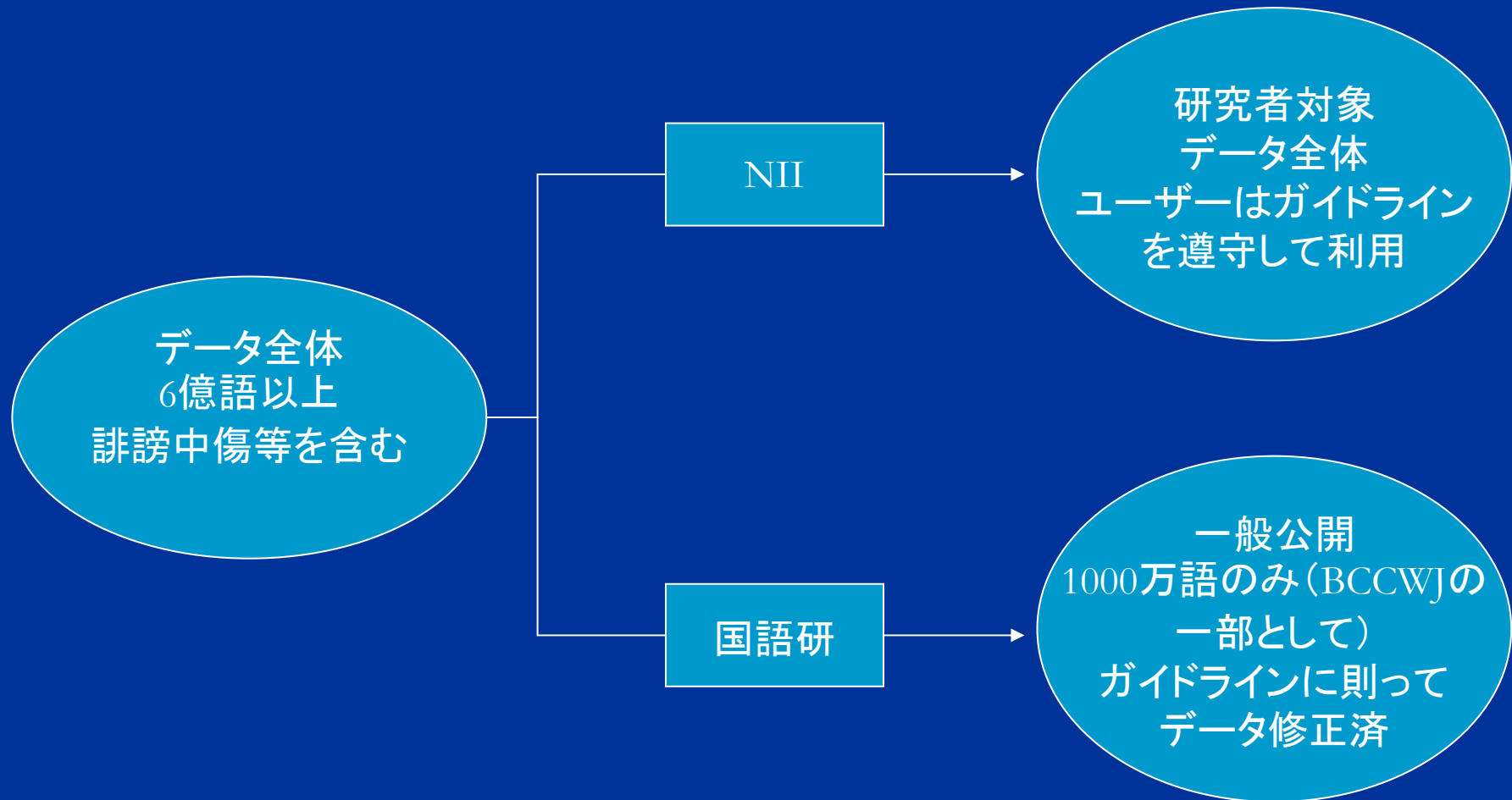
37 件の結果が見つかりました。そのうち 37 件を表示しています。

前文版	検索文字列	後文版	出典	種別	著者	出版元
いけど、お金の話をクリアすればあとはラク	。というか	じつはココが、らくらく持ち家生活を実現す	ゼットイ失敗しないマイホーム 購入大満足ガイドブック	3 社会科学	らくらく持ち家委員会 著;造事務所 編著	情報センター出版
?！ 金はなぜかあるので、働いてません	。というか	、働けるほど集中力も技能もありません。	Yahoo!知恵袋/暮らしと生活ガイド	暮らしと生活ガイド		Yahoo!
です。それが破局したとは公表できません	。というか	実際にしていません。現実のアイドルの結	Yahoo!知恵袋/エンターテインメントと趣味	エンターテインメントと趣味		Yahoo!
）生気論の結合への、パラダイム転換である	。というか	、日本語版の「Anyhow」に見られる通	建築と時間/対論	5 技術・工学	磯崎新, 土居義岳 著	岩波書店
対して -iオプション(-aiはこれを含む	。というか	-rlptg oD を全部指定したのと同	Yahoo!知恵袋/インターネット、PCと家電	インターネット、PCと家電		Yahoo!
す。それで終わりっぽい人なら怒りますね	。というか	「あたしまだなのに・・・」って拗ねてみ	Yahoo!知恵袋/健康、美容とファッション	健康、美容とファッション		Yahoo!
えが戻ってきそうで、いかにも気の毒な話だ	。というか	、この中国人は本当に「馬鹿を見た」のだ。	日中ことばの澳ちがい	8 言語	張麟声 著	くろしお出版
ってば、飛行機のことば、どうでもよかった	。というか	、完璧な二日酔いの中で、身体も起こせずに	ステップ	9 文学	黒田武一郎 著	熊本日日新聞情報文化センター
っこを断念したわけではまったくなかった	。というか	、かれこれとしては、それは必ずしもただのご	消えた女官/マルガ麗宮殺人事件	9 文学	栗本薫 著	早川書房
上委員大変よいことを教えていただきました	。というか	、今の質疑のやりとりを聞いていて、確かに	第140回国会衆議院その他 予算委員会第六分科会	衆議院	佐藤国務大臣	
に。よく見たらカラーの液が垂れていました	。というか	流し台までの移動の間頭はタオルは巻いても	Yahoo!知恵袋/健康、美容とファッション	健康、美容とファッション		Yahoo!
いちゃったか」何もい空中から声でした	。というか	、声というよりも、心話のひびき、であった	ヤーンの朝	9 文学	栗本薫 著	早川書房

BCCWJでのブログデータ公開

- 2008年4月から2009年4月までの1年間
- 抽出時に1000記事以上の投稿があるもの
- 約350万記事から半角文字のみからなる記事などを除外
 - ⇒ 推定6.5億語
 - ⇒ ランダムに1000万語(1.6%、約5万記事)を抽出
 - ⇒ 伏字処理(手作業)
- 今回500万語分をBCCWJデモ検索サイトに追加

ブログデータの公開:NIIと国語研



削除の実例：部分的削除

■例1：個人情報

お問い合わせ：

ファンタステ・ネオ代表 <private type="name">田中夏尾利</private>さんまで
〒141-0022

<private type="address">東京都品川区東五反田3-16-46-201</private>



お問い合わせ：

ファンタステ・ネオ代表 =====さんまで
〒141-0022

=====

■例2：容疑者・被告などの氏名

記事によると、同校野球部の<criminal>木戸夏雄</criminal>監督(47歳)が、2006年12月下旬から、同校を担当していた毎日新聞のA記者(20代女性)にセクハラ行為をしたというのだ。



記事によると、同校野球部の=====監督(47歳)が、2006年12月下旬から、同校を担当していた毎日新聞のA記者(20代女性)にセクハラ行為をしたというのだ。

削除の実例：記事全体の削除

■例3 誹謗中傷

「<pejorative>基地外</pejorative>に殺されたら殺され損」って事ですよ。



φ

■例4 風俗広告

【名古屋東山人妻援護会】

◆待ち合わせの魅力◆

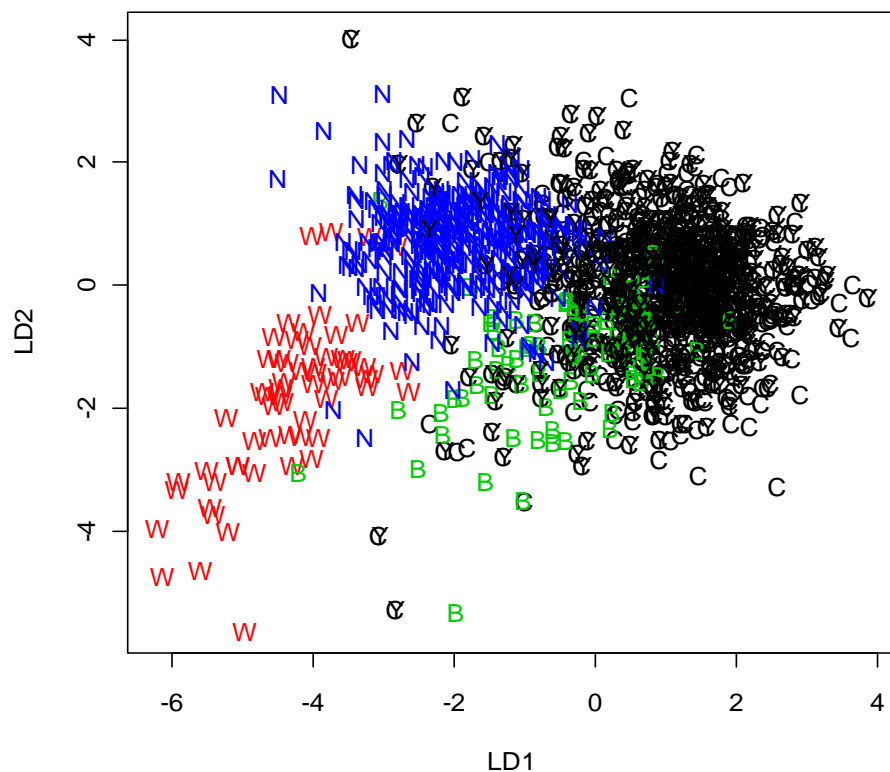
『名古屋東山人妻援護会』です。いつもご利用ありがとうございます
当店は名古屋の東方に位置する待ち合わせ型のデリバリーヘルスです。
東山という高級住宅街のハイクラスな環境下で、ビジネスエリアからも程よく
距離があり



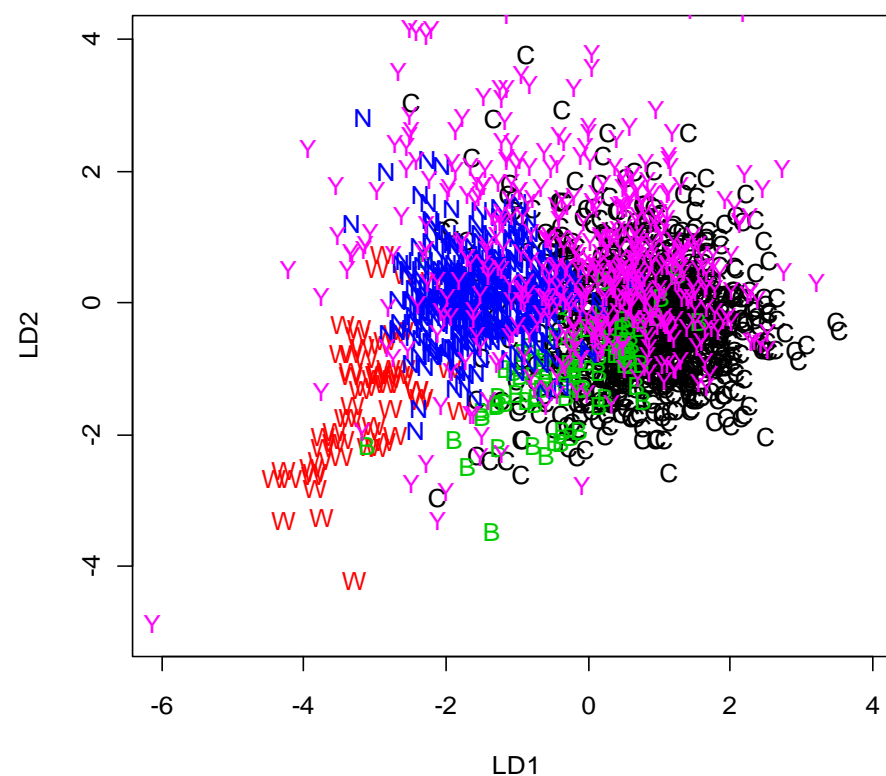
φ

ブログデータの言語的特性

品詞と文長データを用いたジャンルの判別分析

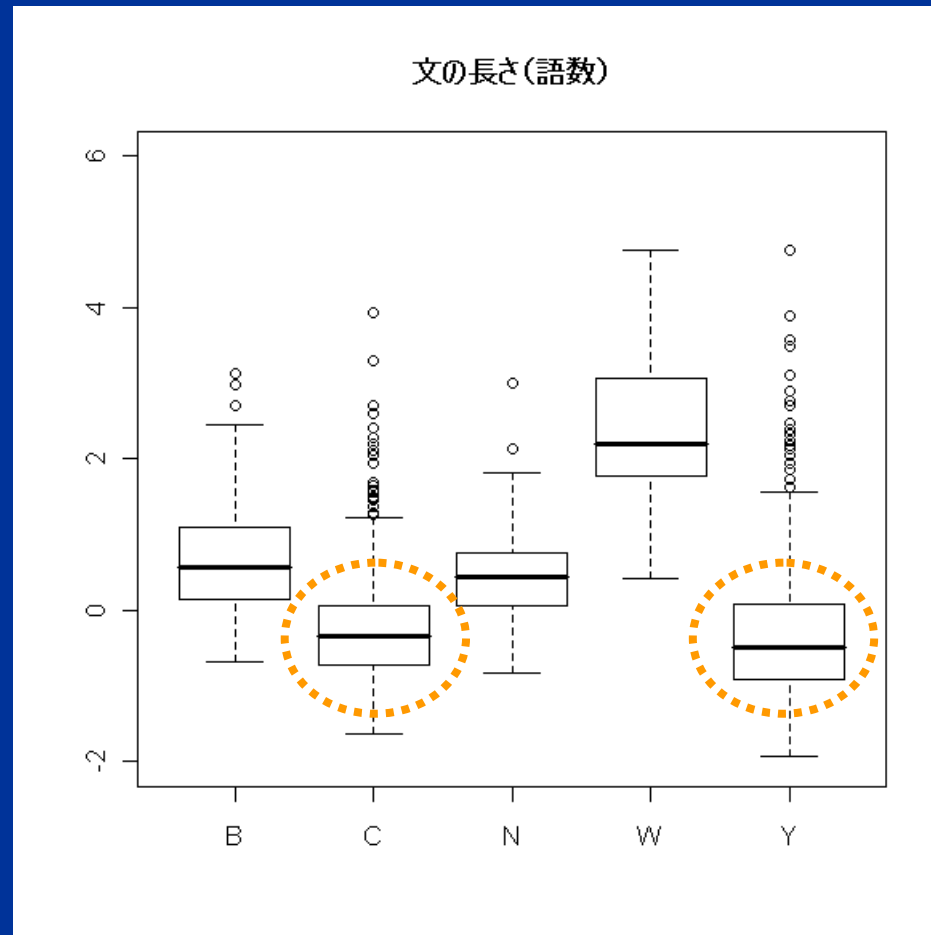


B:書籍 **W:**白書
N:新聞 **C:**知恵袋



B:書籍 **W:**白書
N:新聞 **C:**知恵袋 **Y:**ブログ

文の長さ(語数)

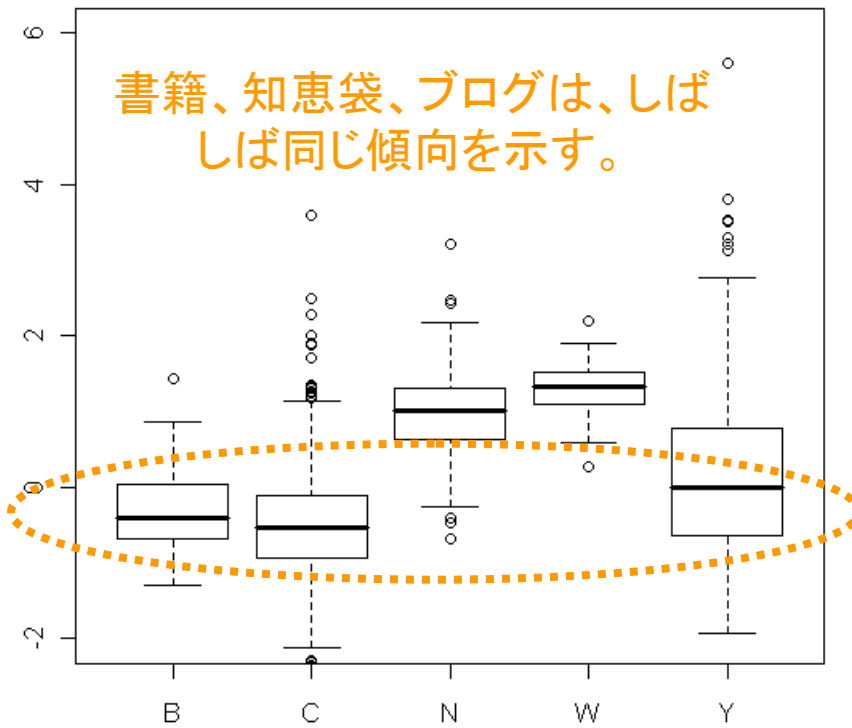


ウェブのテキスト
は文が短い

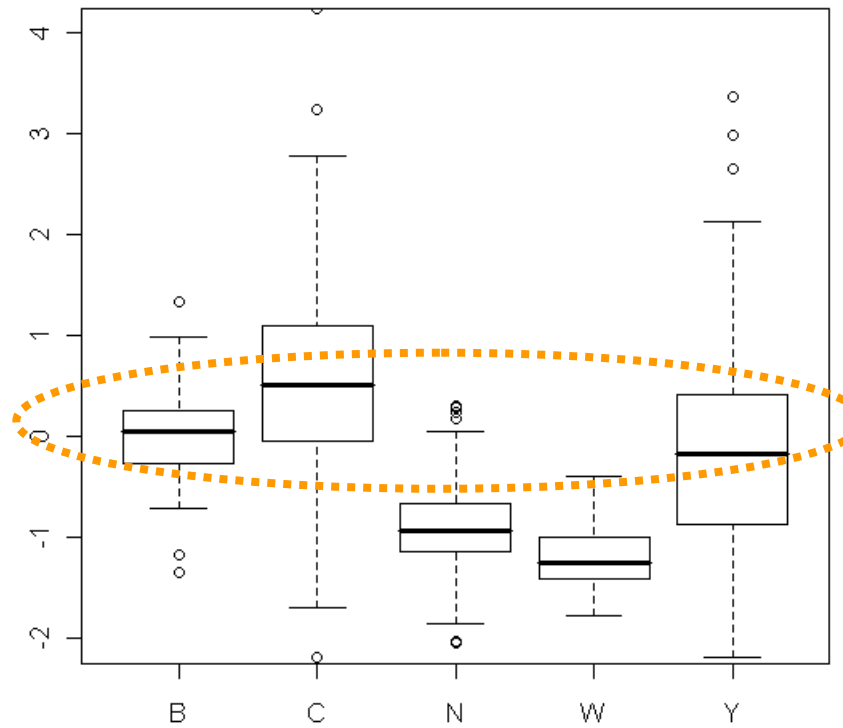
B:書籍 C:知恵袋 N:新聞 W:白書 Y:ブログ

名詞、助動詞の率

名詞の率



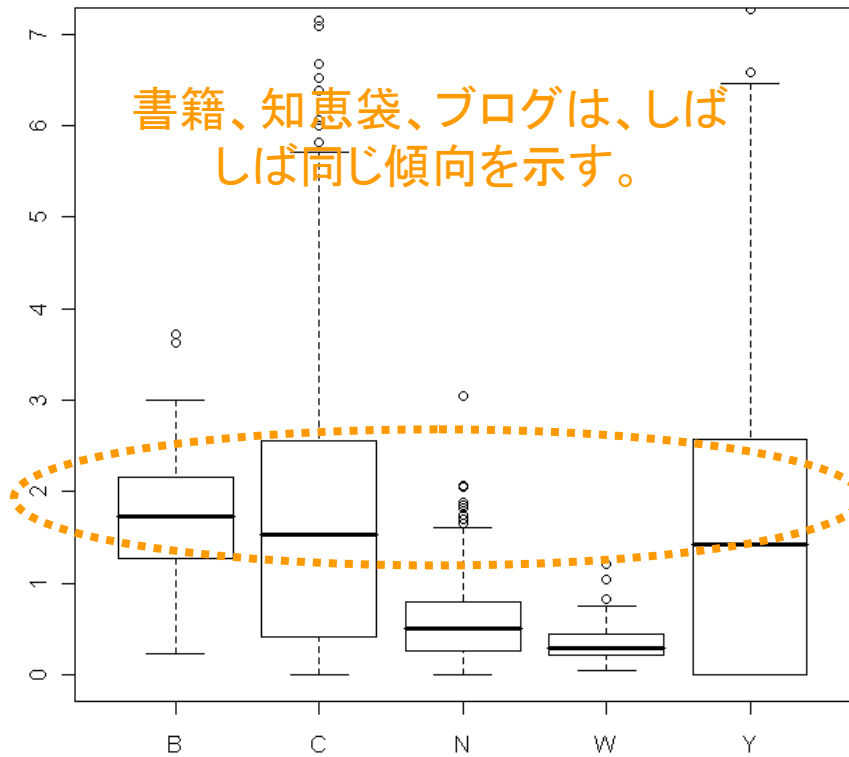
助動詞の率



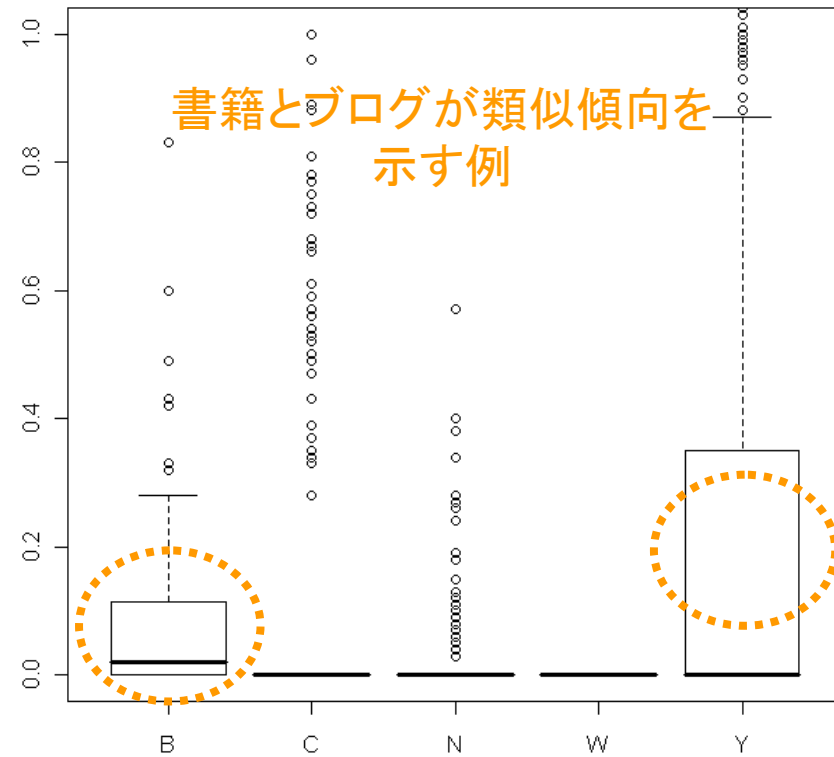
B:書籍 C:知恵袋 N:新聞 W:白書 Y:ブログ

副詞、感動詞の生起率

副詞の率

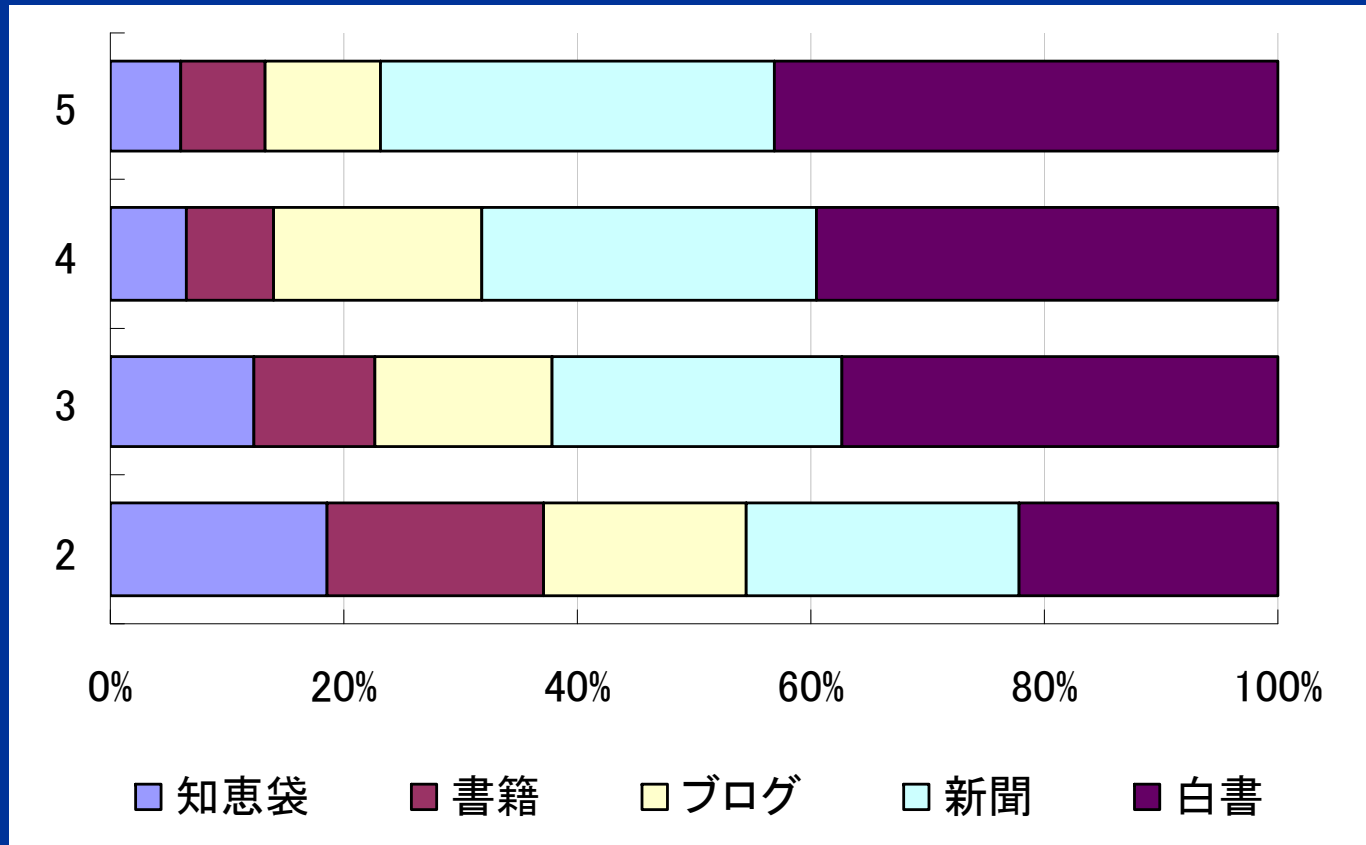


感動詞の率



B:書籍 C:知恵袋 N:新聞 W:白書 Y:ブログ

複合名詞にみるジャンルの比



複合名詞が名詞全体に占める比率の相対値

縦軸は複合名詞に含まれる単純名詞の数

ブログの言語的性格

- 今回分析した5つのジャンルのテキストは、知恵袋に代表されるグループ(文が短く、名詞が少なく、助動詞、副詞、感動詞が多い⇒意見・感想を述べるタイプ。)と、白書・新聞に代表されるグループ(文が長く、名詞が多く、助動詞、副詞、感動詞が少ない⇒事実を述べるタイプ。)の二つに大きく分かれる。
- ブログは書籍とともに知恵袋のグループを形成しているが、新聞・白書グループとオーバーラップするサンプルも多い。
- データの分散が非常に大きく、例外となるサンプルを多数含んでいる。

新宿御苑にて

C|a|n|o|n| |E|O|S| |K|i|s|s| |D|i|g|i|t|a|l| |X| |(|S|I
G|M|A| |十|八| - |二|百|mm| |F|3|. |5| - |6|. |3|
|D|C|)

フラッシュ|使用|:| |未|使用

レンズ|の|焦点|距離|:| |二|百|. |0|mm

CCD|の|幅|:| |6|. |1|1|mm

露出|時間|:| |0|. |0|0|4|0| |秒| |(|1| / |二|
百|五十|)

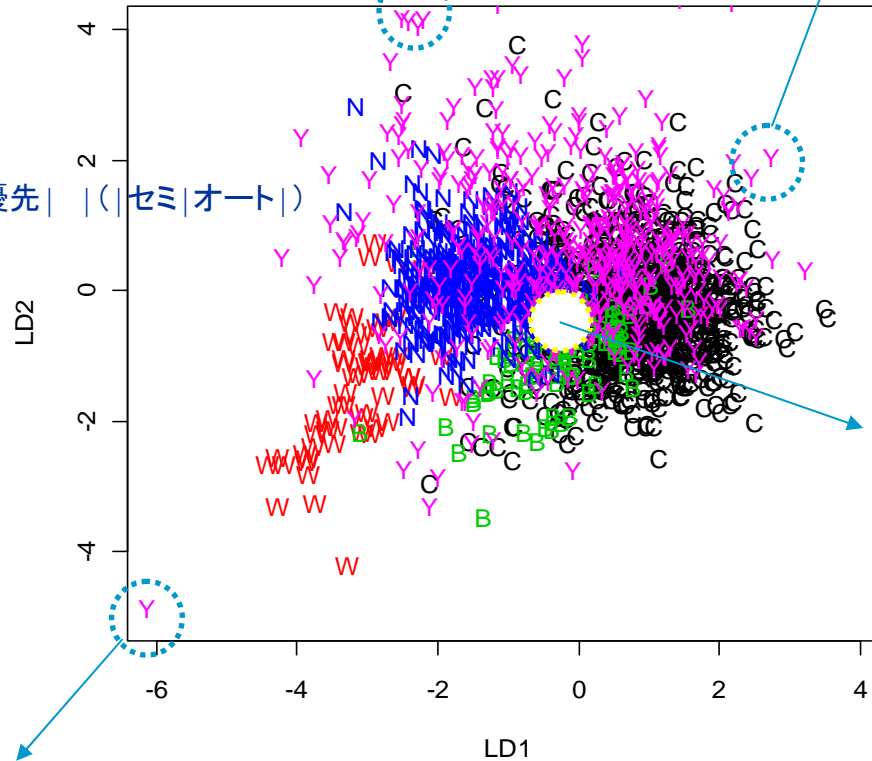
絞り|値|:| |f| / |9|. |0

ISO| |感度|:| |四|百

ホワイト|バランス|:| |オート

測光|方式|:| |マトリックス

露出|プログラム|:| |絞り|優先| |(|セ|ミ|オ|オ|ト|)



まだまだ|写真|が|沢山|あり|まし|た|が|ぼち
ぼち|打ち止め|と|し|ます|。
笑い|。
(|^|^|♪|)。

今日|は|「|株式|投資|これ|だけ|心得|帖|」の|二十|八|日|目|「|売買|
|注文|の|板画|面|を|どう|みる|か|
|」|です|。

【|内容|】

Q| |売買|注文|の|板画|面|を|ど
う|みる|か

A| |あなた|が|「|銘柄|A|を|二|百|
五|十|円|で|五|千|株|売|ろ|う|」と|考
え|て|い|た|時|に|下|記|の|よ|う|な
|板画|面|が|現|れ|た|。

どう|行動|する|か|?

「|買い|板|が|厚|い|の|で|少|し|様子|
|を|見る|」と|答|え|た|人|は|、株
式|投資|に|あ|ま|り|向|い|て|い|ない|
|。

今|なら|希望|通|り|に|二|百|五|十|
円|で|五|千|株|売|れる|。

し|か|し|十|秒|後|に|も|二|百|五|十|
円|に|買|い|板|が|五|千|株|以|上|あ
る|保|証|な|ど|全|く|ない|。

直|ち|に|売|り|注|文|を|出|す|べ|き|で
あ|る|。

「|迷|わ|ず|二|百|五|十|円|で|五|千|
株|売|る|」と|答|え|た|人|は|投資
家|と|し|て|正|しい|。

http://tradings.jp/0977.html | Author: 株山株次郎 |
|割安株で1億円をつつてみる | 指数提供: 株式投資情報サ
イト | バグース | . | . | . | | 二|千|八| - |0|4| - |二|十|一|大|引|け|時|の|割|安|
株 | (|0|4| / |二|十|一|) | . | . | . | | デ|タ|提|供|: |割|安|株|の|バ|グ|ース|
|【|二|千|七| / |0|5| / |0|2| | . | . | . | kabutarou3 | b|log|九|十|三| | f|c
|2|. com | / |b|log| - |entry| - |六|百|七|十|二|. |html| | 株

まとめ

- BCCW検索デモサイトでガイドライン適用済ブログデータ500万語を公開。
- 2011年前半には、ブログと知恵袋のデータ各1000万語を含むBCCWJ(1億語)を公開。
- ブログのテキストは知恵袋と似たふるまいを示すが、多様性が高く、新聞や白書に類似したデータも含まれている。
- 現代日本語の「生きた」姿を幅広く把握するために好適。