文体指標を特徴づける係り受け部分木の抽出

浅原 正幸 (国立国語研究所) *
加藤 祥 (国立国語研究所)

Extraction of Dependency Subtree Features for Writing Style Indexing

Masayuki Asahara (National Institute for Japanese Language and Linguistics)
Sachi Kato (National Institute for Japanese Language and Linguistics)

要旨

柏野 (2013), 柏野・奥村 (2012b) は文体を計量する指標として，専門度，客観度，硬度，くだけ度，語りかけ性の 5 種の分類指標を提案し，現代 日本語書き言葉均衡コーパス (BCCWJ) の図書館サブコーパス 10,551 サンプルに対して悉皆的に付与を行った。浅原ほか (2014) では，この分類指標に対して語彙素を特徴量とした制約付き主成分分析を行い，各指標と特徴的な語彙分 布の対応を品詞ごとに定量的に評価した。浅原ほか (2015b) では語彙素を語彙の系列 (n-gram, p-mer) に拡張し対応分析を行い，既存の定性的な分析結果との比較を行った。今回は，係り受け解析結果の部分木を特徴量とした決定株とブースティングに基づく分類器を用い，文体指標に対して代表的な係り受け部分木の評価を行った。

1. はじめに

柏野 (2013), 柏野・奥村 (2012b) は文体を計量する指標として，専門度，客観度，硬度，くだけ度，語りかけ性度の 5 種の分類指標を提案し，『現代日本語書き言葉均衡コーパス』(BCCWJ) の図書館サブコーパス (LB サンプル)10,551 サンプルに対して悉皆的に付与を行った。このデータに対して，硬度・語りかけ性度を中心に，定量的・定性的な分析が進められてきた (柏野ほか (2012a), 保田ほか (2012b,a,c, 2013d,a,c,b), 加藤ほか (2014))。

また，浅原ほか (2014) では，この分類指標に対して語彙素を特徴量とした制約付き主成分分析を行い，各指標と特徴的な語彙分布との対応を品詞ごとに定量的に評価した。さらに，浅原ほか (2015b) ではこの手法を拡張し，各指標と語彙系列 (語彙素の連続・非連続列) との制約付き主成分分析を行った。本予稿集の別の発表 (浅原・森田 (2015a)) では，主成分分析に必要な「文書-単語行列」・「文書-部分構造行列」をプログラミングすることなしに生成する方法について紹介している。

本稿では，文体を評価する特徴量として係り受け解析結果に基づく単語単位係り受け部分木を用いた識別学習を行い，文体分析を行う。識別学習器として，与えられた木構造の部分木を特徴量とした決定株 (Decision Stumps) を弱学習器とした Boosting アルゴリズムによる機械学

* masayu-a@ninjal.ac.jp

習器 bact (Kudo and Matsumoto (2004)) を用いる。BCCWJ LB サンプル (10,511 サンプル) に対する識別分析のほか約 14 億文からなる Web コーパス (Asahara et al. (2014)) に適用し，識別性能を確認した。

2. 分析手法

2.1 文体指標

柏野 (2013) は文体指標として以下の 5 種類を規定した：

- 【専門度】: 1 専門家向き，2 やや専門的な一般向き，3 一般向き，4 中高生向き，5 小学生・幼児向きの 5 段階指標
- 【客観度】: 1 とても客観的，2 どちらかといえば客観的，3 どちらかといえば主観的，4 とても主観的の 4 段階指標
- 【硬度】: 1 とても硬い，2 どちらかといえば硬い，3 どちらかといえば軟らかい，4 とても軟らかいの 4 段階指標
- 【くだけ度】: 1 とてもくだけている，2 どちらかといえばくだけている，3 くだけていないの 3 段階指標
- 【語りかけ性度】: 1 とても語りかけ性がある，2 どちらかといえば語りかけ性がある，3 特に語りかけ性はないの 3 段階指標

対象は BCCWJ に収録されている図書館サブコーパス 10,551 サンプル（書籍サンプル）とし，20〜50 代女性作業者延べ 9 名に可変長サンプルを呈示して文体指標付与を行った。作業において，インタビューなどのテキスト構造が文体付与に適さないものや外国語や数式などが多いサンプルなど内容や表現が文体付与に適さないものなど 1,664 サンプルを，文体指標付与対象から除外している。本研究ではこれらのサンプルもラベルなしデータとして利用した。

2.2 識別学習器 bact

識別学習器として，ラベル付き順序木の部分木を特徴量とした決定株 (Decision Stumps) を弱学習器とした Boosting アルゴリズムによる機械学習器 bact (Kudo and Matsumoto (2004)) を用いる[(1)]。

決定株は深さ 1 の決定木と同一で単一の特徴量に基づく分類器である。bact では，ラベル付き順序木の部分木を特徴量として考慮した決定株を逐次生成し，これを弱学習器とした Boosting (重み付き多数決) を行う。最右拡張などに基づく部分木構造マイニングアルゴリズムを適応することで，効率的に最適な弱学習器（に対応するラベル付き部分木）を枚挙する実装になっている。

Support Vector Machines などの Large Margin Classifier が事例スパースの解（Support Vector となる事例）を導出するのに対し，Boosting が特徴量スパース（弱学習器に対応する特徴量＝ラベル付き部分木）を導出する。このことは人文系研究者にとって，「どのような特徴量を用いて分析しているのか」を陽に示すだけでなく，SVM と比べて解析速度が高速であるという

(1) http://chasen.org/~taku/software/bact/

利点がある。

2.3　係り受け解析結果に基づくラベル付き部分木の与え方

本研究ではラベル付き部分木として文節係り受け解析器 CaboCha-0.69 の UniDic 主辞規則[2]により生成したものを用いる。文節係り受け解析結果を単語単位係り受け解析に変換する手法として，(1) 文節内最右要素を主辞として残りの要素を鎖状にかける手法と，(2)CaboCha-0.69 の UniDic 主辞規則に基づく内容語主辞[3]を主辞として残り要素を左右から鎖状にかける手法があるが，事前の実験の結果，(1) より (2) の方が良い性能が得られたために (2) を用いる。

他の手法として，Mori et al. (2014) のような単語係り受け木 (上記 (1) に近い木) や Universal Dependencies (UD)(McDonald et al. (2013), Universal-Dependencies-contributors (2015)) のような単語係り受け木 (金山ほか (2015))(上記 (2) に近い木) が考えられる。さらに田中・永田 (2015) は Stanford typed dependency (SD)(Marneffe and Manning (2008)) に基づくラベル付き単語係り受け木において，3 種類の主辞決定規則[4]を定義している。

3.　BCCWJ によるモデリング実験

3.1　手法

BCCWJ LB サンプル (10,511 サンプル) を文単位 (1,651,084 文) に分割し，CaboCha-0.69 (UniDic 主辞規則) により文節係り受け解析を行う。文節係り受け解析結果は 2.3 節に述べた手法で単語係り受け解析結果に変換する。以下に変換事例を示す：

```
(˜EOS(居る (。(た (て (支える))(顎 (を))(両手 (で))(伸ばす (，(床 (に))(体 (を (しなやか (だ (スリム (だ))))))))(オクタビアン (は (　(˜BOS))))))))
(˜EOS(感ずる (。(其れ (を))(横顔 (に (˜BOS))))))
(˜EOS(分かる (。(ない (私 (は (に)))(行く (，(か (の (た (何処 (へ))(人 (が (メンネンカルト (˜BOS))))))))))))
```

文体指標は n 段階評価によりレーティングラベルである。一方，今回用いる識別学習器は二値分類器である。二値分類器をレーティングラベルに対して適用する手法として，順序ラベルのようにレーティングの上位下位に基づく手法[5]が考えられるが，指標によってはラベルが規定されていないものもあり，単純な one-vs-others 法を用いることとした。評価において，全ての二値分類器が負の値を返した場合には「ラベルなし」として認定することとした。

---

(2) `./configure --with-posset=unidic`

(3) 以下のような CaboCha の出力において，1 行目の `* 0 1D 2/4 0.000000` の `2/4` が主辞を表す。/左の `2` が内容語主辞で，`4` が機能語主辞：

```
* 0 1D 2/4 0.000000
”　　補助記号,括弧開,*,*,*,*,,” ,*,,*,*,*,*,*,*,*,*
警察　　名詞,普通名詞,一般,*,*,*,ケイサツ,警察,*,ケーサツ,*,*,*,*,*,*,*
メディア　　名詞,普通名詞,一般,*,*,*,メディア,メディア,*,メディア,*,*,*,*,*,*,*
”　　補助記号,括弧閉,*,*,*,*,,” ,*,,*,*,*,*,*,*,*,*
が　　助詞,格助詞,*,*,*,*,ガ,が,*,ガ,*,*,*,*,*,*,*
```

(4) 主辞後置型 1：内容語と格要素となる後置詞句の間で先に構造を作る句構造を作り，格構造で最右要素を主辞とする；
主辞後置型 2：接続助詞を除いた述部の文節相当の単位で先に構造を作る句構造を想定し，格構造で最右要素を主辞とする；
述部内容語主辞型：述部の文節相当の単位で先に構造を作る句構造を想定し，述部において最左要素を主辞とする。

(5) {1,2,3,4} というレーティングラベルに対して，{1}vs.{2,3,4}・{1,2}vs.{3,4}・{1,2,3}vs.{4} の 3 種類の二値分類器を構成し，これらの多数決により分類する手法。

本実験では bact の iteration 回数を 10,000 回として，BCCWJ LB サンプル全てでモデルを学習し，bact によって得られた連続部分木を分析する。

3.2 得られた規則

得られた規則数を本稿末尾の表 2 の「規則数」の列に示す。およそ，各ラベルごとに数百 (min. 157, max. 411)，文体指標ごとに千前後 (min. 558, max. 1683) 程度の特徴量に基づく規則が得られている。

得られた規則の一例として，表 1 に客観度に対して得られた特徴量（上位 10 位・下位 10 位まで）を示す。客観的なものの例として，引用表現（“言う 。”, “」。” など)，法律用語（“法”, “条” など）などが上位に来る傾向がある。一方，主観的なものの例として，一人称表現（“私”, “僕”, “俺” など）や感嘆符・疑問符などが上位に来る傾向がある。

客観度	1 とても客観的	2 どちらかといえば客観的	3 どちらかといえば主観的	4 とても主観的
デフォルト	−0.0024571855	−0.0006028928	−0.001862472	−0.0022119238
上位1位	0.0012419398 ，	0.0027332267 ユダヤ	0.0032610654 オウ	0.0019894564 ちゃう
上位2位	0.0010513154 権	0.0027150333 ［ ~BOS	0.0020614907 ~EOS 無い 。 か	0.0018879718 僕
上位3位	0.0010498933 」 。	0.0018783132 食品	0.0006848079 ，	0.0014030064 有る 。 だ
上位4位	0.0010301565 有る 。 だ	0.0015717303 寺	0.0006361754 。 だ	0.0013411624 私
上位5位	0.0010125919 ） （ ~BOS	0.0010845271 居る 。	0.0005789535 子供	0.0012610379 ！
上位6位	0.0008768302 図	0.0006354240 ~EOS 有る 。 だ	0.0005688861 作品	0.0012515885 ？
上位7位	0.0008472139 細胞	0.0005610703 ~EOS 有る 。 た だ	0.0005660776 私	0.0008426375 。
上位8位	0.0007825627 ）	0.0004399209 。 ます	0.0005103392 君	0.0007344837 よ
上位9位	0.0006846557 法	0.0003861222 言う 。	0.0005022981 ~EOS 有る 。 だ	0.0006952075 俺
上位10位	0.0005550320 条	0.0002879612 は	0.0004832747 、 、	0.0005798904 ね
下位10位	−0.0006896641 子供	−0.0005263574 思う	−0.0005475509 場合	−0.0006841687 因る
下位9位	−0.0006971665 自分	−0.0005590740 。 た 私	−0.0005766390 銀行	−0.0006940692 ）
下位8位	−0.0007551097 ね	−0.0005657794 ね	−0.0006162622 ．	−0.0007068332 化
下位7位	−0.0008653166 御	−0.0005927404 な	−0.0006934120 発生	−0.0007170490 千
下位6位	−0.0008783045 。 だ	−0.0008522566 ！	−0.0007463329 条	−0.0007193994 企業
下位5位	−0.0010509838 ！	−0.0009144568 ~EOS 居る 。	−0.0007609778 図 ~BOS	−0.0007560440 図
下位4位	−0.0010796309 ？	−0.0009766754 よ	−0.0008078806 ）	−0.0008436630 的
下位3位	−0.0012399752 さん	−0.0015972556 私	−0.0010902414 有る 。 だ	−0.0009392407 ．
下位2位	−0.0012867395 僕	−0.0020705533 」 ？	−0.0012004897 無い 。 か	−0.0017862343 ~EOS 有る 。 だ
下位1位	−0.0015122806 私	−0.0021344768 僕	−0.0020124672 ，	−0.0028774426 ，

表 1　分類指標「客観度」に対する規則（連続部分木パターン）

4. 交差検定による評価

次に，構成した識別学習器の性能を評価するために，BCCWJ LB データ (10,551 サンプル) 上での 5 分割交差検定を行う。ファイル名がサンプルの属性の情報を含んでいるために，乱数を発行することによりサンプル単位で LB データを 5 分割した。識別学習は文単位で行い，文単位評価 (4.1 節)・サンプル単位評価 (4.2 節)・サンプル全体における文の位置に対する正答率 (4.3 節) の 3 種類の評価を実施した。

4.1 文単位評価

文単位の評価結果を本稿末尾の表 2 の「文単位評価」の列に示す。OK は左に示すラベルをシステムが正答した件数，SYS は左に示すラベルをシステムが出力した件数（右に全体における割合を % で表示），GOLD は左に示すラベルを人手により付与された件数（右に全体における割合を % で表示），PREC が精度 (precision) で OK/SYS，REC が再現率 (recall) で OK/GOLD を意味する。

全体の傾向として，GOLD における分布の大きいものが，SYS において大量に生成されるように尖度が高くなる傾向にある。言い換えると，学習データにおいて多数のものの再現率が

高くなる傾向にあり，学習データにおいて少数のものの精度が高くなる傾向にある。

さらに，識別学習器の出力は必ずしも元のサンプルの分布を保存するようなものではなく，識別学習器の分布を用いて，コーパスの文体分布の計量的な調査を行うことは不適切であることがわかる。

一方，低頻度のラベルについて識別結果の精度の高いことは，稀な文体ラベルの事例に似た事例を大量のコーパスから抽出するのには適していると考える。

4.2　サンプル単位評価

サンプルを構成する文単位の評価の重みなし多数決を用いて，サンプル単位の評価を行った。サンプル単位の評価結果を本稿末尾の表 2 の「サンプル単位評価」の列に示す。各列の意味は「文単位評価」と同じである。

重みなし多数決を行う結果，より一層 GOLD における分布の大きいものが SYS において大量に生成される傾向が強くなり，分布の小さいものの判定の出現する確率が下がる傾向にある。例えば，専門度においては 98.3% が「3 一般向き」と出力される。客観度においては 94.3% が「ラベルなし」と出力される。

4.3　サンプル全体における文の位置に対する正答率

評価は 10 文以上のサンプルのみについて行った。表中 (n)-(n-1)% はサンプル全体における評価対象文の位置を表す。

どの指標も 80-90%, 90-100% で正答率の下がる傾向が見られた。サンプリングにおいて 90-100% に位置するデータが少なくなる傾向にある [(6)]にしても，有意に差がある。

これはサンプルの末尾にプロフィールやまとめなどの本文と異なる文体が出現しているためではないかと考える。

5.　超大規模コーパスへの適用

最後に現在国語研コーパス開発センターで開発している超大規模コーパス (Asahara et al. (2014))（2014 年 10-12 月収集分）全文（1,463,142,939 文=14.6 億文，23,836,100,595 語=238 億語，EOS・句点を含まず）に 3 節で構成した識別学習器を適用した結果を表 2 右の「超大規模コーパス分布」に示す。

基本的に SYS （システムが出力した件数）とその割合のみを示す。4 節に示した通り，学習元データにおける分布の大きいラベルがより多く出力される傾向にある。しかしながら，分布の小さいラベルでも出力されることがあり，これらの出力結果は精度の高いものであると考える。

[(6)] 例えば，サンプル中 19 文の場合，90-100% の文数が 1 に対し，他の箇所は文数 2 となる。

6. おわりに

6.1 本研究のまとめ

本研究では，BCCWJ LB サンプルに付与された文体指標を単語係り受け部分木を特徴量とした識別学習器によりモデル化し，分析を行った。

特徴量の抽出 (3 節) においては，単語・単語列の特徴的な表現を抽出しているが，単語係り受け木を使う有効性までは確認することができなかった。交差検定による性能評価 (4 節) においては，チャンスレベルと比較するとよい性能が得られたが，学習元データの分布をそのままシステム出力することがないことが確認された。

今回学習した識別学習器を Web から収集した 14 億文規模のテキストコーパスに適用した (5 節)。識別学習器の出力の分布を用いて文体指標の分布を分析することが困難である一方，少ないラベルの精度が高いことから，大量のテキストから似た文体の事例を高精度で収集することが可能であると考えられる。

6.2 今後の検討課題

今後検討すべき課題は以下のとおりである：

- 特徴量スパース vs. 事例スパース
  2.2 節に述べた通り，今回用いた識別学習器は特徴量を抽出することによる二値分類器である。計算量は大きくなるが Tree Kernel に基づく Large Margin Classifier などを用いることで事例スパースな解を与えて，境界事例を分析することを考えたい。
- bact に与える単語係り受け木
  日本語においては文節係り受け木に基づく自然言語処理の研究が進んでいる一方，文節係り受け木に基づいてどのような単語係り受け木を与えるかを検討する必要がある。2.3 節に示した通り，ここ数年日本語の単語係り受け木の規定について様々な提案がなされており，文の識別問題として定式化した文体指標分析に最適な単語係り受け木を調査する必要がある。
- 二値分類器のレーティングラベル適応
  本研究では浅原ほか (2014, 2015b) の対応分析の結果から，レーティングラベルが必ずしも線形上に分布していないとし，one-vs-others 法を用いた。その結果，分布の偏りを増長するような識別学習器を構成することになった。どのようにラベル空間を構造学習器に反映させるかを検討する必要がある。

謝辞

本研究の一部は国語研基幹型共同研究プロジェクト「コーパスアノテーションの基礎研究」および国語研「超大規模コーパス構築プロジェクト」によるものです。

参考文献

浅原正幸・加藤祥・立花幸子・柏野和佳子 (2014).「文体指標と語彙の対応分析」　第 6 回コーパス日本語学ワークショップ, pp. 11–20.

浅原正幸・森田敏生 (2015a).「コーパスコンコーダンサ『ChaKi.NET』の「文書-部分構造行列」出力機能」　第 8 回コーパス日本語学ワークショップ.

浅原正幸・加藤祥・立花幸子・柏野和佳子 (2015b).「文体指標と語彙系列の対応分析」　第 7 回コーパス日本語学ワークショップ, pp. 7–16.

Asahara, Masayuki, Kikuo Maekawa, Mizuho Imada, Sachi Kato, and Hikari Konishi (2014). “Archiving and analysing techniques of the ultra-large-scale web-based corpus project of ninjal, japan.” *Alexandria*, 25:1-2, pp. 129–148.

de Marneffe, Marie-Catherine, and Christopher D. Manning (2008). “The stanford typed dependencies representation.” *Prof. of COLING-2008: Workshop on Cross-framework and Cross-domain Parser Evaluation*.

金山博・宮尾祐介・田中貴秋・森信介・浅原正幸・植松すみれ (2015).「日本語 Universal Dependencies の試案」　言語処理学会第 21 回年次大会発表論文集, pp. 505–508.

柏野和佳子・立花幸子・保田祥・丸山岳彦・奥村学・佐藤理史・徳永健伸・大塚裕子・佐渡島紗織 (2012a).「テキストの硬さと軟らかさの考察–『現代日本語書き言葉均衡コーパス』の収録書籍を対象に–」　第 1 回コーパス日本語学ワークショップ, pp. 131–138.

柏野和佳子・奥村学 (2012b).「書籍テキストへの分類指標人手付与の試み–『現代日本語書き言葉均衡コーパス』の収録書籍を対象に–」　言語処理学会第 18 回年次大会, pp. 1260–1263.

柏野和佳子 (2013).「書籍サンプルの文体を分類する」　国語研プロジェクトレビュー, 4:1, pp. 43–53.

加藤祥・柏野和佳子・立花幸子・丸山岳彦 (2014).「語りかける書きことばの表現」　国立国語研究所論集, 8, pp. 85–108.

Kudo, Taku, and Yuji Matsumoto (2004). “A boosting algorithm for classification of semi-structured text.” *Proc. of EMNLP-2004*, pp. 301–308.

McDonald, Ryan T., Joakim Nivre, Yvonnne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, and Slav Petrov Hao Zhang Oscar Täckström Kuzman Täckström, Keith B. Hall (2013). “Universal dependency annotation for multilingual parsing.” *Prof. ACL-2013(2) 92-97*.

Mori, Shinsuke, Hideki Ogura, and Tetsuro Sasada (2014). “A japanese word dependency corpus.” *Proc. of LREC-2014*, pp. 1631–1636.

田中貴秋・永田昌明 (2015).「日本語のラベル付き依存構造解析の検討」　言語処理学会第 21 回年次大会発表論文集, pp. 1044–1047.

Universal-Dependencies-contributors (2015). *Universal Dependencies*. `https://universaldependenceis.github.io/docs/`.

保田祥・柏野和佳子・立花幸子・丸山岳彦 (2012a).「「語りかけ性」を有すると判断される書きことばの表現」　第 2 回コーパス日本語学ワークショップ, pp. 43–50.

保田祥・柏野和佳子・立花幸子・丸山岳彦 (2012b).「「語り性」を有する書きことばの典型例の分析」　第 1 回コーパス日本語学ワークショップ, pp. 139–146.

保田祥・柏野和佳子・立花幸子 (2012c).「総体として印象を与える表現：「語りかけ性」を有すると判断する根拠」　人工知能学会第 41 回ことば工学研究会.

保田祥・立花幸子・柏野和佳子・丸山岳彦 (2013a).「「ベテランは足を保護する」が語りかけるとき」　第 4 回コーパス日本語学ワークショップ, pp. 345–354.

保田祥・柏野和佳子・立花幸子・丸山岳彦 (2013b).「アノテーターコメントを用いた「語りかけ性」分析の試み–頻度情報から捉え難いテキスト性質の解明に向けて–」　言語処理学会第 19 回年次大会発表論文集, pp. 358–361.

保田祥・柏野和佳子・立花幸子・丸山岳彦 (2013c).「語りかけると判断される文体–大規模コーパスを用いた特徴的表現の分析–」　日本文体論学会第 104 回大会.

保田祥・柏野和佳子・立花幸子・丸山岳彦 (2013d).「書きことばにおける「語りかけ」は何のために用いられるのか」　第 3 回コーパス日本語学ワークショップ, pp. 143–152.

表 2　文体分析結果

専門度

ラベル	規則数	文単位評価 OK	SYS		GOLD		PREC	REC	サンプル単位評価 OK	SYS		GOLD		PREC	REC	超大規模コーパス分布 SYS	
ラベルなし		3,193	17,595	1.1%	284,947	17.3%	0.181	0.011	10	40	0.4%	1,730	16.4%	0.250	0.005	4,399,250	0.3%
1 専門家向き	391	28	367	0.0%	15,992	1.0%	0.076	0.001	0	0	0.0%	141	1.3%	0.000	0.000	13,833	0.0%
2 やや専門的な一般向き	411	7,961	21,998	1.3%	106,685	6.5%	0.361	0.074	81	138	1.3%	929	8.8%	0.586	0.087	2,042,510	0.1%
3 一般向き	238	1,076,748	1,602,737	97.1%	1,093,905	66.3%	0.671	0.984	7,036	10372	98.3%	7,065	67.0%	0.678	0.995	1,456,613,821	99.6%
4 中高生向き	395	280	621	0.0%	91,866	5.6%	0.450	0.003	0	0	0.0%	384	3.6%	0.000	0.000	9,238	0.0%
5 小学生・幼児向き	248	4,699	7,766	0.5%	57,689	3.5%	0.605	0.081	1	1	0.0%	302	2.9%	1.000	0.003	64,287	0.0%
計	1,683	1,092,909	1,651,084		1,651,084				7,128	10,551		10,551				1,463,142,939	

硬度

ラベル	規則数	文単位評価 OK	SYS		GOLD		PREC	REC	サンプル単位評価 OK	SYS		GOLD		PREC	REC	超大規模コーパス分布 SYS	
ラベルなし		53,846	276,337	16.7%	284,947	17.3%	0.194	0.188	282	1,548	14.7%	1,730	16.4%	0.182	0.163	526,050,369	36.0%
1 とても硬い	389	1,775	4,741	0.3%	68,194	4.1%	0.374	0.026	10	13	0.1%	619	5.9%	0.769	0.016	464,034	0.0%
2 どちらかといえば硬い	252	78,699	175,765	10.6%	420,734	25.5%	0.447	0.187	622	1,112	10.5%	3,065	29.0%	0.559	0.202	47,655,505	3.3%
3 どちらかといえば軟らかい	182	604,739	1,192,266	72.2%	753,383	45.6%	0.507	0.802	4,087	7,877	74.7%	4,440	42.1%	0.518	0.920	888,566,194	60.7%
4 とても軟らかい	214	819	1,975	0.1%	123,826	7.5%	0.414	0.006	0	1	0.0%	697	6.6%	0.000	0.000	406,837	0.0%
計	1,037	739,878	1,651,084		1,651,084				5,001	10,551		10,551				1,463,142,939	

くだけ度

ラベル	規則数	文単位評価 OK	SYS		GOLD		PREC	REC	サンプル単位評価 OK	SYS		GOLD		PREC	REC	超大規模コーパス分布 SYS	
ラベルなし		43,721	251,880	15.3%	284,947	17.3%	0.173	0.153	55	357	3.4%	1,730	16.4%	0.154	0.031	179,609,675	12.3%
1 とてもくだけている	157	328	1,052	0.1%	89,419	5.4%	0.311	0.003	0	0	0.0%	473	4.5%	0.000	0.000	347,197	0.0%
2 どちらかといえばくだけている	163	158,122	353,419	21.4%	511,680	31.0%	0.447	0.309	813	1,526	14.5%	2,696	25.6%	0.532	0.301	18,614,436	1.3%
3 くだけていない	287	550,230	1,044,733	63.3%	765,038	46.3%	0.526	0.719	5,188	8,668	82.2%	5,652	53.6%	0.598	0.917	1,264,571,631	86.4%
計	607	752,401	1,651,084		1,651,084				6,056	10,551		10,551				1,463,142,939	

語りかけ性度

ラベル	規則数	文単位評価 OK	SYS		GOLD		PREC	REC	サンプル単位評価 OK	SYS		GOLD		PREC	REC	超大規模コーパス分布 SYS	
ラベルなし		31,947	125,616	7.6%	284,947	17.3%	0.254	0.112	288	1,216	11.5%	1,730	16.4%	0.187	0.131	46,693,906	3.2%
1 とても語りかけ性がある	206	13,182	45,451	2.8%	112,441	6.8%	0.290	0.117	9	14	0.1%	833	7.9%	0.642	0.010	1,297,863	0.1%
2 どちらかといえば語りかけ性がある	179	57	416	0.0%	197,646	12.0%	0.137	0.000	0	1	0.0%	1,379	13.1%	0.000	0.000	805	0.0%
3 特に語りかけ性はない	173	1,003,963	1,479,601	89.6%	1,056,050	64.0%	0.678	0.950	6,409	9,320	88.3%	6,609	62.6%	0.687	0.969	1,415,150,365	96.7%
計	558	1,049,149	1,651,084		1,651,084				6,706	10,551		10,551				1,463,142,939	

客観度

ラベル	規則数	文単位評価 OK	SYS		GOLD		PREC	REC	サンプル単位評価 OK	SYS		GOLD		PREC	REC	超大規模コーパス分布 SYS	
ラベルなし		652,752	1,084,518	65.7%	937,252	56.8%	0.592	0.685	4,609	9,948	94.3%	4,650	44.1%	0.463	0.991	1,302,522,743	89.0%
1 とても客観的	340	7,039	18,627	1.1%	102,858	6.2%	0.377	0.068	73	112	1.1%	950	9.0%	0.651	0.076	3,761,009	0.3%
2 どちらかといえば客観的	198	99,538	374,577	22.7%	299,282	18.1%	0.265	0.332	241	487	4.6%	2,523	23.9%	0.494	0.095	142,993,449	9.8%
3 どちらかといえば主観的	276	8,332	60,169	3.6%	200,814	12.2%	0.138	0.041	1	3	0.0%	1,566	14.8%	0.333	0.000	516,827	0.0%
4 とても主観的	278	11,416	113,193	6.9%	110,878	6.7%	0.100	0.102	1	1	0.0%	862	8.2%	1.000	0.000	13,348,911	0.9%
計	1092	779,077	1,651,084		1,651,084				4,925	10,551		10,551				1,463,142,939	

全体における位置に対する正答率

(10文以上)	0–10%	10–20%	20–30%	30–40%	40–50%	50–60%	60–70%	70–80%	80–90%	90–100%	90–100%
専門度	0.675	0.677	0.675	0.675	0.675	0.675	0.675	0.673	0.627	0.401	0.401
硬度	0.463	0.467	0.469	0.467	0.470	0.468	0.467	0.465	0.442	0.331	0.331
くだけ度	0.548	0.550	0.550	0.551	0.549	0.548	0.544	0.546	0.510	0.365	0.365
語りかけ性度	0.640	0.641	0.641	0.640	0.640	0.639	0.639	0.638	0.607	0.450	0.450
客観度	0.493	0.495	0.494	0.492	0.491	0.495	0.491	0.493	0.477	0.358	0.358

サンプル単位5分割交差検定(ランダム順分割)

文単位評価:　サンプルを文単位に分割して、各文がどのラベルに属するかを推定。どれにも分類されないばあいには「ラベルなし」をつける。
サンプル単位評価:　文単位で評価したものの多数決。
超大規模コーパス分布:　超大規模コーパスを解析したもののシステム出力の分布。

OK は正当した数、SYSはシステムが出力した数で右に総数に対する割合(%)を付与、GOLDは人手で付与した数で右に総数に対する割合(%)を付与。
PREC は精度 (OK/SYS)、REC は再現率 (OK/GOLD)。

全体における位置に対する正答率:　サンプルが10文以上のものに対し、全体における位置に対する正答率。