

特定の構文グループにおける形容詞の分布：現代中国語大規模コーパスによる統計とクラスタリング

メタデータ	言語: Japanese 出版者: 公開日: 2017-10-03 キーワード (Ja): キーワード (En): 作成者: 林, 智 メールアドレス: 所属:
URL	http://hdl.handle.net/2297/7649

特定の構文グループにおける形容詞の分布

—現代中国語大規模コーパスによる統計とクラスタリング

林 智

1. Introduction

コーパス言語学においてコーパスの環境整備とアクセスのパフォーマンス改善、種々のクエリや統計処理の組み合わせを手早く行えるようにする事は、実験サイクルの高速化により研究のスピードの迅速化と研究の裾野を広げる事に直結する。

我が研究室・及び金沢大学 IT 教育推進プログラム内の活動においては XML を中心としたコーパスのデータベース化とアプリケーション開発に努め、これらの研究を援護する環境の構築を進めてきた。

本稿では、今年度までのコーパス作成プロジェクトの活動の概括とデータベースを利用した研究の一例として現代中国語の特定の構文における形容詞の振る舞いとクラスタリングについての論考を行う。

2. Chinese Corpora and XML-database

金沢大学における中国語コーパスの集積は、以前紹介したように独自で入力した現・当代文学コーパス、5 言語対照パラレルコーパス、及び国家語委現代汉语通用平衡语料库 (CYLK)¹などの外部のソースを中心としたコーパス群を中心に行っている²。近年ではこれ以上のコーパスを増やすよりは利用環境を中心とした整備を進めてきた。

現在では、コーパスのクオリティや提供されているベースの形式に応じて、クエリ用のアプリケーションを複数動作させているが、その中で

¹ [教育部语言文字应用研究所, 2005]

² [林, 2005]参照

も特に前掲の CYLK については精度の高い POS タグ（品詞タグ）がつけられている事もあって最も優先的に環境構築をすすめている。

この CYLK は、もともと種々の現代中国語のテキストに対しテキスト本体のメタ情報及び Paragraph と POS タグを付加したテキスト形式のデータとして提供されていたものであり、これを XML 形式に変換して頂いた版本をさらにパースして、POS タグのないテキストの除去と POS タグの属性ノード化を行い、特定の構文形式を目標とするクエリ（データベースへの問い合わせ要求）を容易にした XML 文書の集合として再構築したものを現在では利用している³。

この XML 化したコーパスは、ネットワークのサーバ上に構築したデータベースによって管理・提供されているわけであるが、現在ではネットワーク越しの利用はセキュリティ及び著作権処理の問題が発生することを防ぐために制限し、サーバ OS にログインすることによってのみ可能になっているため、事実上はローカルでの利用を行っているのが現状である。

データベース本体は、Oracle 10g、IBM DB9、eXist、及び Microsoft SQL Server 2005 という Native XML Database(NXDB)としての動作が可能な複数のデータベースを導入し、それぞれに全ての XML ベースのコーパスを収納しクエリ発行可能な状態にしているが、この中でも現在は主に DB9 での利用を中心としている。それぞれ XQuery⁴を発行する部分においても組み込み関数による動作の違いが若干あるものの、ラッピングによりほぼ透過的にアクセスする事ができるようにしている。従って、安定運用を行うまではどのデータベースを中心としても構わないのだが、データのメンテナンスやクエリの管理のしやすさから DB9 をメインのデータベースとして採用している。

リレーショナルデータベース(RDB)ではなく NXDB を採用している理

³ この作業の間に well-formed XML Document ではなかったソースが若干除去されているため CYLK に含まれる全てのテキストが収められているわけではない。

⁴ (The World Wide Web Consortium(W3C), 2007)

特定の構文グループにおける形容詞の分布

由は今さら語るまでもないことだが、メタ情報を含んだコーパス、特に複層階層を前提としているような言語コーパスの場合は、テーブル形式のデータよりもツリー形式でノードを構築できる XML の方がより適している上、昨年度あたりからまともに XQuery を発行でき、かつ商用データベースとして最も信頼性の高いデータベースをリリースしてきた上記の 3 社⁵が Native に XML に対応したデータベースを相次いでリリースし、クエリのパフォーマンス面についても RDB でのクエリより効率のよいアクセスが期待できるためである。

CYLN などの XML データベース化と XPath 及び XQuery によるクエリが発行できるようになった事によって、様々な条件を組み合わせた複雑なクエリを組んで、かなり目標を絞り込んだ特定の構文を対象とした検索、統計をすることが可能になっている。

以下の論考における統計は全てこの NXDB と XML に対するクエリによって行っており、クエリの組み方で同時に特定の構文を抽出・統計処理を加えたものである。

3. Statistic

中国語の形容詞については、その内包するいくつかの要素、例えば褒貶義（[±評価]）⁶や計量性など複数次元のカテゴリー類別によってその振る舞いや生起可能性が異なる事は、これまでの先行研究において明らかである。このうち、比較構文における計量形容詞及び褒貶義を有する属性形容詞の生起可能性についてはかつて分析を加えた事があるが⁷、本節では数量的に形容詞の特定構文における出現について分析する。

形容詞についての基本的な統計であるが、まずそれぞれの形容詞の出

⁵ eXist については早期から XQuery を発行できる数少ない Open Source の NXDB であったので前々から利用してきたが、上述の状況からいってこの先選択肢に残る事はほぼない。

⁶ [相原, 1976] [陆, 说量度形容词, 1989]等、形容詞の分類についていくつかの呼称があるが概念は同等である。

⁷ [吉田 林, 2001]

現頻度について概括すると、全ての形容詞の出現数 $S(A)$ (ただし、 $A=\{A_0, A_1, \dots, A_n\}$) は 1,369,391 であり、 $n=9728$ であった。出現数最大の形容詞は“大”で約 2.7%、次が“多”で約 1.6%である。また、出現数上位の 64 個の形容詞でおよそ 30%、269 個の形容詞で 50%、約 1300 個の形容詞で 70%を占める。以下の統計においては基本的に上位 300 個の形容詞、すなわち形容詞の出現数の 60%程度を占める事象を対象に結果を検討している。

次に、構文グループとして 特に特徴の出やすいと思われる比較構文を取り上げて比較する。比較構文については以前にも論じた“和(跟/同/像/与) ~一样 A”と“有(没有/只有) ~那么(这么/那样/这样) A”のように同じ程度か否かに言及する構文や⁸、それ以外にもっと一般的な“比”“不比”を用いる構文があげられる。⁹これらの構文について個々の形容詞が実際にはどの程度生起しているのか、その傾向がこれまで認識されていた通りかどうかを検討する。

下にクエリから得られた結果をまとめて図示する。(表の上部は構文の代表例)

	跟+一样 A	有+那么 A	比+X+A	比+X+A+的	不比+A
計量 (+)	+	+	+	+	+
計量 (-)	-	-	+	+	+
属性 (+)	+	+	+	+	+
属性 (-)	+	+	+	+	+
属性 (±)	-	-	+	+	+

上表は生起が見られたか見られないかを図示したものである。同程度

⁸ 同じ程度を表す比較構文における形容詞の生起については一般的に計量形容詞の場合は形容詞ペアのうちスケールがマイナス方向の形容詞、褒貶義ならば Negative である方の形容詞に制限が強くなる事を論じた。

⁹ [陆 马, “比” 字句新探, 1999]参照

特定の構文グループにおける形容詞の分布

に関する比較構文で制限が見られるがその他の構文については生起自体には制限がないようである。

さらに共起の傾向についていえば、“没有”の構文では“多”が特に多いが二音節形容詞が生起するパターンが目立っている。対して“有”の構文については生起自体が“没有”の半分程度であるが、計量形容詞のプラス方向の形容詞がよく生起しており、以前の論考を裏付ける結果になっている。

“跟～一样”についても、プラス方向マイナス方向双方の形容詞、及び色彩をあらわす形容詞などの中立的な形容詞も生起している。計量形容詞のマイナス方向の形容詞についての生起がみられないのは以前の仮説とは合わない結果である。統計対象の母集団が不足しているのでなければ、生起条件の制限がたとえ緩く生起自体はできるとしても、実際の言語環境では“跟～一样”とマイナス方向の計量形容詞は現在ではあまり使われない表現という事になる。¹⁰

“比”については各カテゴリーの形容詞が等しく生起できているが、共起という観点から見ると傾向が異なる。“比A”“比A的”については、前者が単音節の計量形容詞に特に多いのに対し、後者は計量以外の単音節形容詞や二音節形容詞も同じ程度に共起する。特に“重要”は共起頻度が高い。

全般的に言えば、計量形容詞のプラス方向の形容詞が最も振る舞いが自由で、計量形容詞のマイナス方向の形容詞が最も振る舞いに制限がある。

さらに一般的な構文である“副詞+A”と“副詞+A+的”について生起頻度と共起頻度について統計を行い、特に副詞“很”“最”“更”の場合について比較してみると、やはり形容詞の生起頻度に偏りがみられる。

まず突出を表現する“最A”と“最A的”に関しては、概して“最A”

¹⁰ 以前の調査では、かなり古い年代のソースも検索の対象に含んでいたのも、現代語のコーパスとはずれが生じる可能性がある。

の生起頻度が高いが、褒貶義を含む形容詞や一部の二音節形容詞に関しては、“最 A 的”の頻率が高い。

同じく“非常”については二音節形容詞との結びつきが強く単音節形容詞の生起頻度は概して低い。また“非常 A 的”の方が生起頻度が高いのも明らかである。

“很”は特定の形容詞との結びつきが強く、これらの形容詞は“很 A”の単位で形容詞として扱われている場合もある。これらのものを除いて考慮した場合は、褒貶義形容詞や二音節形容詞では“很 A 的”の生起頻度が比較的少ないが、プラス方向の計量形容詞は逆転現象もしくは漸近が起きている場合がある。

さらに、“更”についてふれると、計量形容詞、褒貶義形容詞の生起頻度の占有率が他と比べて高くなる。特に“大”“多”“高”“重要”“主要”などでは“更 A 的”の生起頻度が他と比べて高く、逆転している場合もある。全体的に“更 A(的)”自体の生起頻度が低いのは、比較の義が生起に制限をかけているためと推測できる。

全体的に、“A+的”の生起頻度の高い形容詞は共通している場合が多いが、“很 A”と“更 A”で逆転現象が起きている場合と起きていない場合があり、これらの複数の副詞及び“的”との共起性の強弱を特徴として総合的にとらえることで形容詞のカテゴリ化が可能にみえる。

そこで、これらの統計結果をもとに多変量解析を行う事によって、形容詞一般について共起に基づくカテゴリーの再構築を試みる。

特定の構文グループにおける形容詞の分布

4. Clustering

そこで、本節では前節に挙げてきたような特定構文における形容詞の出現頻度の占有率をもとに、Ward 法によるクラスタリングと Self Organization Map(SOM)¹¹による形容詞カテゴリーの数量的表現の可視化を試みる。

Ward 法は多次元ベクトルに基づくクラスタリング手法であり、クラスの融合の際に次にあらわされる距離関数 $D(p,q)$ に基づいて融合する手法である。

$$D(C_1, C_2) = E(C_1 \cup C_2) - E(C_1) - E(C_2)$$

$$E(C_i) = \sum_{x \in C_i} (D(x, C_i))^2$$

Ward 法は一般的にクラスタリングの感度が高いとされる手法であり、これを SOM と組み合わせることで情報が読み取りやすくなる。

SOM は教師なし学習を行うニューラルネットワーク (Neural Network) の一種であり、高次元の特徴ベクトル (Feature vector) で表わされるサンプルを位相を保ったまま低次元 (多くは 2 次元か 3 次元) に投影する解析である。主成分分析 (principal component analysis, PCA) と異なり、サンプルに対してネットワークが超曲面に展開して張り付く事になる。SOM を構成する各ノードは格子状もしくはハニカム状に配置されており、入力シグナルに対して距離が最も近いノードが唯一発火するネットワークである。

SOM の学習の基本的アルゴリズムは、入力シグナルに対して勝者ノード (Winner Node/Best Matching Unit) を決定し、近傍関数に基づき周囲のノードの重みを更新する事によって行われる。勝者ノードは特定の距離関数により求められる入力シグナルとの距離によって決定される。すなわち、ステップ $t+1$ における重み W の更新式は、

$$W_i(t+1) = W_i(t) + \theta(i, t)\alpha(t)(D(t) - W_i(t))$$

¹¹ [T.Kohonen, 1999]参照

となる。ただし、 $\theta(t)$ は近傍関数、 $\alpha(t)$ は学習係数を時間とともに減少させる関数、 $D(t)$ はステップ t における入力である。

本節では、以下に掲げる特定構文における形容詞 A の出現数の統計に基づき、それぞれの構文における占有率を n 次元ベクトルとして上述のクラスタリングを行う。

【対象構文】

- $D+A$. 及び $D+A$ 的.

D: 非常 也 最 更 很 还 较 又

- $X+C/V+Y+一様+A$.

C/V: 和 与 跟 同 像

- $X+V+Y+R+A$.

V: 有 没有 只有

R: 那么 这么 那样 这样

- $X+V+Y+A(+的)$.

V: 比 不比

学習の対象とする形容詞は、出現数の多い順に上位 300 種までで、入力ベクトルは上記に挙げた 29 次元、学習は Batch-Learning、距離関数は分散重み付きユークリッド距離、ノードの初期化ベクトルは PCA により第一主成分と第二主成分を XY 方向に展開したものである。

以下に SOM 上に Ward 法によるクラスタリング結果をセパレータとして加えた図を示す。

特定の構文グループにおける形容詞の分布

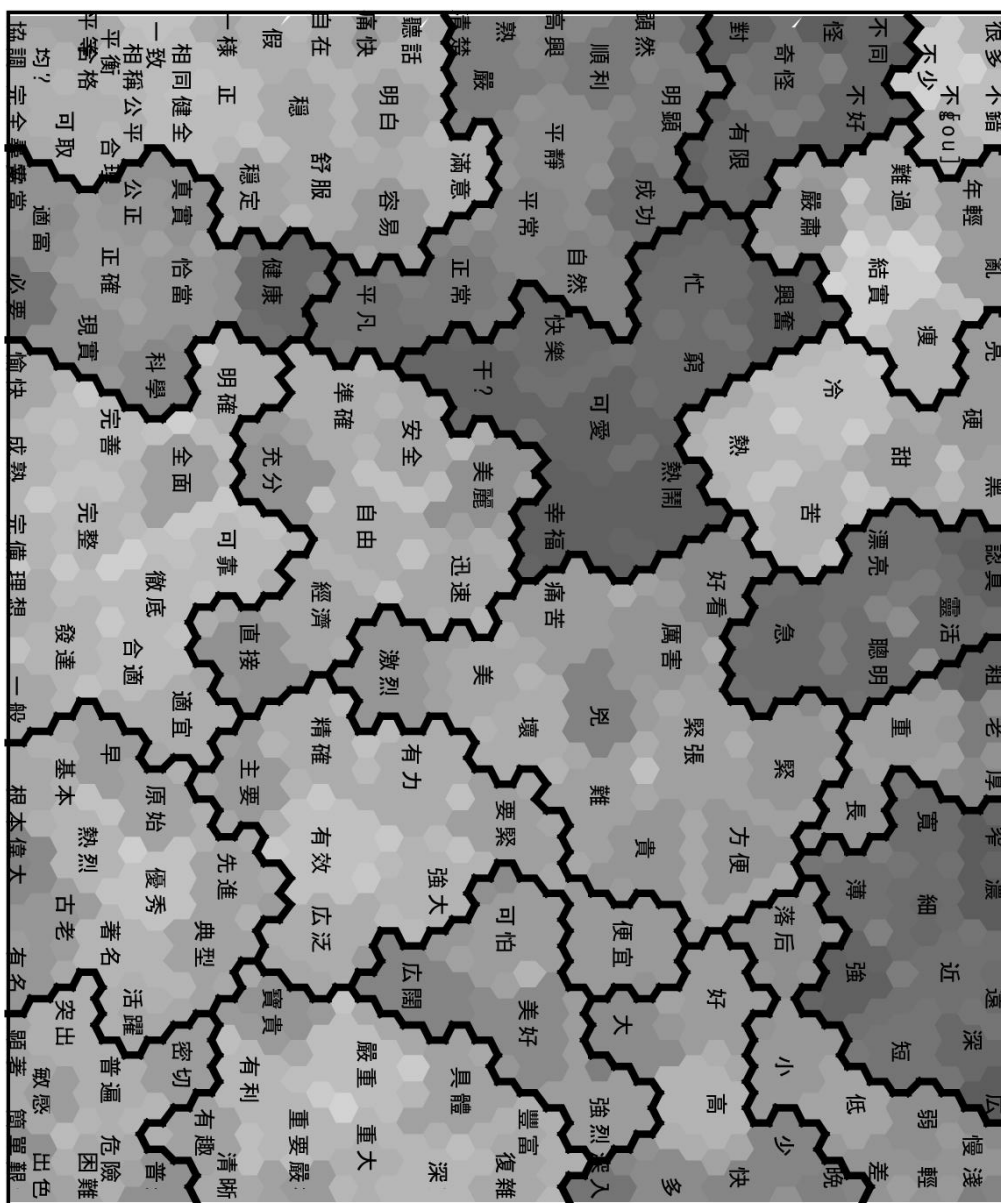


図 形容詞による Self Organization Map

この SOM 上から、ほぼ単音節形容詞と二音節形容詞でクラスの大分類が行われることと形容詞が内包する属性によって一定のクラスの傾向がみられる事が分かる。

マップの中央から右上方向に単音節形容詞が投射されているが、この中でも計量形容詞と属性形容詞とされる一群はそれぞれ別の領域に展開されている。右上のクラスが計量形容詞とされるものであり、中央部のクラスには褒貶義の強い形容詞、中央上のクラスには無属性、中立の形容詞が分類される。

対して、二音節形容詞については不+A などの合成語が左上、中央左から左側にかけて感情形容詞、中央から左下にかけては基本的に Positive な形容詞か中立の形容詞がそれぞれ語義のカテゴリーごとにクラスを形成し、右下に一部 Negative な形容詞が見られる。

総じていえば、これらの構文における共起傾向を基準に形容詞をカテゴライズするのであれば、およそ計量／属性の違い > 単音節／二音節 > プラスマイナス指向（量的・褒貶義）の順に形容詞の特徴を決定する条件の重要度が高いと推定できる。この基準は明らかに前節で挙げた形容詞の数量的な特徴とよく近似するものであり、さらにこの SOM に未知語や Phrase を投入する事により、該当後の内包する特徴が最も漸近するクラスに適合させることもできる。

5. Conclusion

以上に述べたとおり、コーパスによる統計分析は個々の事象を追求することもさる事ながら、全体の事象を見渡す際に特にそのメリットを発揮しやすい。本稿における形容詞のクラスタリング手法などは、どのような方法を採用するかはともあれ、ベースの POS タグ付きコーパスが十分な数に達すれば、さらに応用が可能なアプローチの方向性を示している。

また、インフォーマント調査に時に見られるような曖昧性¹²を排除し、数

¹² 前回比較構文と形容詞について論じた時はインフォーマント調査を一次資料

特定の構文グループにおける形容詞の分布

量的な根拠から既存の議論を再確認もしくは反証することが可能になる。この際、例外事象の取り扱いや母集団の不足による観察可能圏外の事象についての問題も残されるが¹³、一定の成果を期待する事はできよう。

文献目録

T. Kohonen. (1999). *Self-Organizing Maps*. Springer.

吉田清香, 林智. (2001). 同程度を表す 2 つの比較構文とその差異の考察. 金沢大学中国語学中国文学教室紀要第 5 輯, 31-62.

教育部语言文字应用研究所. (2005). 国家语委现代汉语通用平衡语料库.

林智. (2005). 中国語言語コーパスを用いた研究とソフトウェア環境. 金沢大学中国語学中国文学教室紀要第 8 輯, 23-47.

陆剑明, & 马真. (1999). “比”字句新探. 出处 陆剑明・马真, 现代汉语虚词散论 (页 179-203). 语文出版社.

とした。

¹³ 本稿においてもコーパス内の範囲では十分にサンプルを得られない事象については意図的に記述を削除している。