

## 日本語助詞「と」コーパスの構築

花岡 洋輝<sup>†</sup>増田 勝也<sup>‡</sup>植松 すみれ<sup>‡</sup>美馬 秀樹<sup>§</sup><sup>†</sup> 東京大学大学院情報理工学系研究科 <sup>‡</sup> 東京大学知の構造化センター <sup>§</sup> 東京大学大学院工学系研究科

hkhana@is.s.u-tokyo.ac.jp {masuda,uematsu}@cks.u-tokyo.ac.jp mima@t-adm.t.u-tokyo.ac.jp

## 1 はじめに

日本語の統語解析研究においては係り受け解析 [1] が大きな成功を収め、これを利用した基盤的・応用的な研究が広く行われている。一方で、生成文法のような文法理論に基づく構文解析器 [2] の研究も進められており、機械翻訳や知識検索など様々な応用が期待される。

実用的な統語解析器を構築する場合、現実 に即した統計モデルの学習や、語彙辞書の半自動的な獲得が重要であるため、今やコーパスは必要不可欠なものである。日本語の統語解析研究において広く利用されているコーパスの一つとして、京都大学テキストコーパス [3] が挙げられる。このコーパスは、1995 年度の毎日新聞の記事・社説の、約 40,000 文に対して形態素・構文情報を付与したものである。また、これと同じ文集合に対して述語項構造・照応関係を付与したものとして NAIST テキストコーパス [4] があり、これを利用した意味解析の研究にも期待がかかる。

日本語において助詞は、「が」「を」「に」のように格標識を表すもの、「から」「ので」のように複文を構成するもの、「か」のように単純疑問文を作るものなど、様々な統語的役割を担っている。これらの統語的役割を弁別せずして日本語の統語・意味解析は成り立たない。したがって、助詞についてのコーパス整備が不可欠である。

NAIST テキストコーパスでは、格解析・照応解析のための資源として、主要な格標識である「が」「を」「に」格を対象にアノテーションが行われている。これは助詞に限らず、表層格の関係が広くアノテーションされているものであるが、「が」「を」「に」あるいはそれらに相当する「は」などの助詞に関してアノテーションが行われた資源であると捉えることもできる。本研究は、更に細かな統語・意味解析に向けて、別の助詞についても情報を追加することを目的とするものであり、我々は、その頻度の高さと用法の多様さから、助詞「と」を標的にアノテーションを施した。

表 1: 京都大学テキストコーパスに含まれる高頻度助詞とその品詞細分類の頻度。

	格助詞	接続助詞	副助詞	終助詞	
の	2909	50328	1	0	53238
を	32949	0	0	0	32949
は	3	0	32231	0	32234
に	30562	437	0	0	30999
が	23812	3273	0	0	27085
と	21980	25	1	0	22006
で	13369	2	1	2	13374
も	0	8	9139	2	9149

## 2 日本語における助詞「と」

国立国語研究所 [5, pp. 99–119] によれば、助詞「と」には大きく分けて、格助詞、接続助詞、並立助詞の三つの用法がある。この内、格助詞用法には、『彼女はそこで、ディックと会う。』のような共格標識や、『八百長だったと知ったチェンバレンが〜』のような補文標識、あるいは『心臓はとつくとつくとはげしく脈をうっています。』にあるような擬音語・擬態語を受ける用法など、様々なものが含まれる。

京都大学テキストコーパスにおいて、助詞は、格助詞、接続助詞、副助詞、終助詞に分類される。頻度の高い助詞に対して、この分類に対する頻度は表 1 のようになる。最頻出の「の」については、統語的な曖昧性よりもむしろ意味的な曖昧性の解決が重要であると考えられ<sup>1</sup>、アノテーションに困難が予想されるので今後の課題として扱うこととし、次いで頻出する「を」「は」「に」「が」については、格標識として出現しているものに関しては NAIST テキストコーパスにより情報が付与されているので、本研究では、次いで頻出する助詞であって、かつ既存の言語資源では情報の少ない「と」に関してアノテーションを行うことを目指した。

<sup>1</sup> 統語的に異なる用法の代表的なものとして、名詞的な「の」や、断定等を表す「のだ」があるが、それぞれ形式名詞/助動詞としてアノテーションされているため、表 1 には含まれていない。

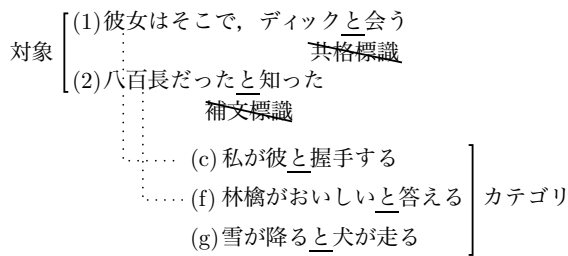


図 1: 用例に基づくアノテーション。ラベルの代わりに用法の近い文例を選択する。カテゴリは表 3 の一部。

表 2: アノテーション対象の助詞とその頻度。

	訓練	評価	全体
と	618	18835	19453
とは	14	556	570
とも	5	394	399
	637	19785	20422

表 7: Cohen の  $\kappa$  係数によるアノテータ間の一致度。

	評価	全体
と	0.8178	0.8234
とは	0.6938	0.7019
とも	0.6641	0.6683

表 1 が示すように、助詞「と」は出現頻度が高いにも関わらず、そのほとんどは格助詞としてアノテーションされているだけで、その多様な用法を周辺の状況から判断する以外にない。京都大学テキストコーパス中、約 5,000 文に対しては、格関係がアノテーションされており、そこには「と」に関する関係もアノテーションされているので、用法判断の困難が幾らか緩和されるものの、依然として残りの文に含まれる「と」については格助詞という手がかりしかない状況である。本研究で報告するコーパスは、この状況を打開し、更に細かな解析の可能な統語解析器を構築する一助となると考えられる。

### 3 助詞「と」コーパス

本稿は既存研究 [6] で提示されたコーパスの完成を報告するものである。既存研究では、用例に基づくアノテーションフレームワークにより、京都大学テキストコーパスに含まれる約半分の「と」について、用法のアノテーションを施している。本研究では、残りの「と」に対して既存研究と同様のフレームワークでアノテーションを施した。

#### 3.1 用例に基づくアノテーション

統語的なアノテーションを行う場合、アノテータにある程度の文法知識を要求することとなる。この要求を緩和するため、アノテーションカテゴリには文法用語を用いず、文の実例を用いることとした。具体的

表 3: アノテーションカテゴリ<sup>2</sup>。最右列は設計者の意図する用法。／で区切られた前者は「と」に前置される句を、後者は用法を表す。

(a)	私が林檎と桃を食べる。	体言並列
(b)	「話す」と「聞く」。	用言並列
(c)	私が彼と握手する。	体言／補語
(d)	山と積まれた桃を食べる。	体言／修飾
(e)	「林檎」と子供。	体言／述部省略
(f)	林檎がおいしいと答える。	用言／補語
(g)	雪が降ると犬が走る。	用言／接続
(h)	仕事が終わったと喜ぶ。	用言／修飾
(i)	「おいしい」と子供。	用言／述部省略
(j)	やっと終わった、と。	文末
(k)	彼は思った。おいしいと。	転置
(l)	わんわんと犬が吠える。	副詞／修飾
(m)	というのも、	文頭
「とは」に対する追加カテゴリ		
(n)	酵素とは触媒のことだ。	体言／命題
(o)	「大志を抱け」とは、識者の弁。	用言／命題
「とも」に対する追加カテゴリ		
(p)	両者とも一歩も引かない。	体言／接尾辞

は、図 1 が示すように、「共格標識」や「補文標識」というラベルを直接にアノテーションするのではなく、各々のカテゴリに対応する『私が彼と握手する』『林檎がおいしいと答える』という文例を提示して、アノテーション対象の用法が、どの文例での用法に最も近いかを判断させることでアノテーションを行った。このフレームワークであれば、アノテータに特別な文法知識を要求することがないので、アノテータは対象言語に堪能でありさえすれば良く、アノテータの文法知識によらずある程度一貫したアノテーションが可能であると期待される。

#### 3.2 実際のアノテーション

本稿で報告する第一版では、取り立て助詞を介する「とは」「とも」以外の複合助詞は対象外とし、それらを除いた 20422 個に対して (表 2)、二人のアノテータにより重複アノテーションを施した。一人目のアノテータには、簡単な口頭説明だけでアノテーションを施してもらい、二人目のアノテータについては、1 月 4 日分 (950104.KNP) のデータに対して一人目の結果を参照しながらアノテーションしてもらうことで簡単な訓練とした。本稿では便宜上、1 月 4 日分を訓練データ、それ以外を評価データと呼ぶことにする。実際の作業では、文脈情報を参照するために記事単位で、「と」「とは」「とも」に対して出現順にアノテーションを行った。

<sup>2</sup>紙幅の都合で実例は多少改変してある。

表 4: 「と」に対するアノテータ間の混同行列. (\*) は該当無しを表す.

(a)	2907.5	9	108.5	11	11	64	1	3	4	0	0	0	0	2
(b)	0	3	0	0	0	0	0	0	0	0	0	0	0	0
(c)	50.5	3	5032	35.5	58	957.5	9	12	17	0	0	3	0	0
(d)	7	0	121	32	14.5	51.5	8	5.5	2	0	0	0	0	1
(e)	0	0	0	0	1	0	3	0	2	0	0	0	0	0
(f)	2	0	66	1	2.5	7000	121.5	99	198	5	3.5	0	2	0
(g)	0	0	0	2	0.25	46.5	1213	15	7.25	1	0	0	0	0
(h)	0	8	2	0	0	33.5	47	18.5	2	1	0	1	0	0
(i)	0	0	0	0	0.25	1	0	0	39.75	0	0	0	0	0
(j)	0	0	1	0	0	0	3	0	3	18	21.5	0	0	0
(k)	0	0	0	0	0	0	0	0	0	0	0	0	0	1
(l)	1	0	38	73	1	15	0	0	0	0	0	137	0	0
(m)	0	0	0	1	0	1	0	0	0	0	0	0	24	0
(*)	2	0	2	0	0.5	1	0	0.5	0	0	0	0	1	0

表 5: 「とは」に対するアノテータ間の混同行列. 表 6: 「とも」に対するアノテータ間の混同行列.  
アノテーション数が 0 のカテゴリは省略.

(c)	238	13	25	3	0	0	0	34	0	0
(d)	2	0	0	0	0	0	0	0	0	0
(f)	6	0	127	7	0	0	0	0	3	0
(g)	0	0	2	1	0	0	0	0	0	0
(k)	0	0	0	0	4	0	0	0	0	1
(l)	1	0	1	0	0	0	0	0	0	0
(m)	0	0	0	0	0	0	2	0	0	0
(n)	2	0	2	2	0	0	0	72	4	0
(o)	0	0	1	0	0	0	0	0	2	0
(*)	0	0	0	0	0	0	0	1	0	0

(a)	2	0	7	0	0	0	0	0	0	1
(b)	0	0	0	0	1	0	0	0	0	0
(c)	2	0	120	2	12	0	0	0	0	35
(e)	0	0	0	0	0	0	0	0	0	0
(f)	0	0	3	0	104	10	0	0	0	0
(g)	0	0	0	0	2	8	0	0	0	0
(i)	0	0	0	0	0	0	0	0	0	0
(j)	0	0	0	0	0	0	2	0	0	0
(l)	0	0	4	0	0	7	0	0	2	1
(p)	0	0	5	0	0	0	0	0	0	64

アノテーションカテゴリとして表 3 に示すカテゴリを用意しているが、統語的な違いを重視したので、意味的に同じものが別のカテゴリに分かれているものもある。たとえば、(d) と (l) はいずれも修飾句を作る用法であるが、前置される句でカテゴリを分けている。アノテータはこの他に「該当無し」を選ぶこともでき、また単一の判断をできない場合には二つ以上のカテゴリを選ぶことも許されている。これらのカテゴリについて、評価データに対するアノテータ間の混同行列は表 4,5,6 のようになった<sup>3</sup>。これに対して Cohen の  $\kappa$  係数 [7] を計算すると表 7 のようになるので、アノテータ間の一致度は比較的高いと言えるが、複合助詞に対する一致度は相対的に低い。

## 4 不一致の分析

### 4.1 「と」の不一致

既存研究 [6] での報告と同じく、表 4 を見ると、(c) と (f) での不一致が最も多い。約 1024 個あると考え

<sup>3</sup>複数のカテゴリが選択されている場合には、カテゴリ数で割った数を頻度とみなしているため、各セルの数値は整数とは限らない。

られるこの種の不一致のうち<sup>4</sup>、「と」直前の語の品詞を調べると、名詞 756 個、名詞性接尾辞 83 個、形容詞語幹 73 個と、名詞的な用法の語が多いことが分かる。これは例えば、『「民営化の足下に爆弾」という見出し』のように、前置される句が体言止めされているのが原因であるものが多いと考えることができる。これについては、京都大学テキストコーパスの構文情報を用いて、「と」直前の句が用言的であることを知ることができればある程度は事後修正可能であると考えられる。

次いで頻出する不一致は、(f) と (i) での不一致である。199 個の不一致のうち、「と」直後の語の品詞が名詞のものが 184 個あり、『破壊したと発表』のように、後置される句が体言の場合に、(f) と (i) を混同してしまう場合があったと考えられる。不一致の仕方に偏りがあることから、一方のアノテータは (i) を述部の省略ではなく、述部が体言であるような用例と勘違いしてしまったと推察され、用例に基づくフレームワークを、本研究のように単純な形で用いることの難しさを示していると考えられる。

<sup>4</sup>複数カテゴリが付与されている場合に頻度が小さく見積もられている。

## 4.2 「とは」の不一致

「と」の場合と同じく、(c) と (f) での不一致が多いが、それ以上に (c) と (n) での不一致が多い。そのうち約半分の 19 個については、『文化とは一体何だろう』のように疑問詞の含まれるものであった。これはガイドラインの不足によるところが大きいと考えられるが、用例に基づくフレームワークにおいて、一つのカテゴリに複数の用例を用意し、設計者の意図するカテゴリに誘導するといった戦略も考えられる。残りの半分の中には、『「ナラダ」とは、光と熱情と喜びを天から地に運び、地上から経験を天に運ぶ意味だというくらい哲学的なのだ』のように、命題表現とそうでない用法が混ざってしまっているものや、『研究会とは名ばかりで』のように、慣用表現と関わる用法も含まれており、カテゴリの判別を難しくさせていたと考えられる。

## 4.3 「とも」の不一致

「とは」の場合と同じく、「と」のアノテーションでは使用されなかったカテゴリに対する不一致、すなわち (c) と (p) での不一致が高頻度である。一つの原因として、複合助詞は単体の「と」に比べると圧倒的に頻度が少ないため、訓練が有効に働いていないことが考えられる。また「と」「とは」「とも」を並行してアノテーションしたことにより、アノテータを混乱させ、結果に揺れが生じてしまったのではないかと予想される。

40 ある (c)(p) 間の不一致のうち、4 つは『日本はどの国ともマニュアルを共有していない』のように (c) の用法を含みながらも、対象が複数であり意味的には (p) に近いため、アノテーションの揺れ易い境界例であると考えられるが、残り 36 個については、『二人とも人気者ではない』のように (p) の用法に該当すると考えられるもので、両アノテータが揃って (p) を選択している『四党とも大差はない』のような文例との大きな違いを確認できない。アノテーションの揺れも疑って、より注意深くデータを精査する必要があると考えられる。

## 5 おわりに

多様な機能を持ち、かつ使用頻度の高い日本語助詞「と」と一部の複合助詞について、その用法を統語的に分類し、アノテーションを施した。単体の「と」については、アノテータ間の一致度は比較的高い値を記

録しており、実用に耐える資源になったと考えられる。今後は、一致度を下げている要因について、本稿で示した分析をもとに整理し直し、一般公開を目指していく。複合助詞「とは」「とも」に対するアノテーション結果については、未だアノテータ間の一致度が十分とは言えず、更なる分析と整理が必要である。幸いにし頻度はそれほど多くないため、「と」コーパス公開時に付随的な情報としてこれらを加えることも可能であると期待される。

## 謝辞

本研究は、東京大学知の構造化センターの助成を受けています。

## 参考文献

- [1] 工藤拓, 松本裕治. 相対的な係りやすさを考慮した日本語係り受け解析モデル. 情報処理学会論文誌, Vol. 46, No. 4, pp. 1082–1092, April 2005.
- [2] Melanie Siegel and Emily M. Bender. Efficient deep processing of Japanese. In *Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization*, pp. 1–8, 2002.
- [3] Sadao Kurohashi and Makoto Nagao. Building a japanese parsed corpus while improving the parsing system. In *Proceedings of the NLPRS-97*, pp. 451–456, 1997.
- [4] Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. Annotating a japanese text corpus with predicate-argument and coreference relations. In *Proceedings of the Linguistic Annotation Workshop*, pp. 132–139, June 2007.
- [5] 国立国語研究所. 現代語の助詞・助動詞 用法と実例. 秀英出版, 1951.
- [6] Hiroki Hanaoka, Hideki Mima, and Jun'ichi Tsujii. A japanese particle corpus built by example-based annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pp. 1876–1880, May 2010.
- [7] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, Vol. 20, No. 1, pp. 37–46, 1960.