決定木に基づく多義語分析：「明らか」を例に

李在鎬（筑波大学）・進藤三佳（京都大学）

1. 目的と狙い

本研究では，コーパスデータに対する語彙意味論の方法として，決定木を使った多義語分析の手法を紹介する。分析対象は，進藤（他）（2011）において記述分析を行なった「明らか」を取り上げる。分析においては，BCCWJ（現代日本語書き言葉均衡コーパス(Balanced Corpus of Contemporary Written Japanese)）から得たKWIC データを表層情報に基づいてコーディングした上，「明らか」の語義を従属変数に，「明らか」の生起文脈を独立変数にして実験を行った。調査の結果，「明らか」の語義の判別に関して，後続要素を独立変数にした場合，学習データと評価データのいずれにおいても高い正答率が得られたことを報告する。

2. 問題提起：「明らか」の意味における多様性

2010年世論をにぎわせた「海上保安庁ビデオ流出事件」のニュース報道の中で，次に示すように「明らか」が頻繁に使われている。

(1) 海上保安庁の情報管理のずさんさが要因の一つだったことは明らかで，守秘義務違反に問うほどの「秘密性」があるのかという疑問符も付く。
(http://www.yomiuri.co.jp/net/news 2010.11.16. 読売新聞)

(2) 馬淵国土交通大臣は閣議のあとの記者会見で，尖閣諸島沖で起きた中国漁船による衝突事件の映像が流出した事件について，海上保安庁に対して捜査への全面的な協力を指示したことを明らかにしました。
(http://www.nhk.or.jp/news/html 2010.11.8.)

(3) その（海上保安庁）関係者は「実際に私も見た。明らかに中国の船が当たりにきていたのが一目瞭然（りょうぜん）で誰でも分かると思った」と語る。
（http://www.nikkansports.com 社会ニュース 2010.11.11.08:17 紙面から）

例文（1）から例文（3）における「明らか」の意味同士の関連性については，以下のように考えることができる。例文（1）は「AはBだ」という命題を表し，「明白」であることを表している。その論証として，「ずさんさが要因の一つだったことは明白だ。」に置き換えられる。例文（2）は「指示したことを公表した。」に置き換えられることから，「公表する」ことを表している。例文（3）は話者の判断および評価を表すもので，連用形「明らかに」は「分かる」と修飾関係にある。さて，これら三種類の「明らか」以外に，もう一つよく使われる用法として例文(4)がある。なお，以下においては，出典を明記していないものは，すべてBCCWJのものである。

(4) 入るだけ損」とする未納者が増えています。しかし、それは明らかな間違いです。Q公的年金の制度は崩壊してしまうの？A　年金財政の収支は、このところ急速に

例文（4）は「はっきりとした間違い」というように，言い替えられる。従って，知覚できる（し

やすい），認識できる（しやすい）という意味を表す[1]。

以上の考察を踏まえ，進藤（他）（2011）では「明らか」に対して以下の４タイプが存在することを指摘している。

・タイプA.　「明白だ」型（明白型）
・タイプB.　「公表する」型（公表型）
・タイプC.　話者判断提示型（判断提示型）
・タイプD.　「はっきり」型（はっきり型）

本研究では，進藤（他）（2011）を踏まえ，次の３つの問題を検討する。１）各タイプのコーパスにおける分布を明らかにすること，２）個々の語義タイプの分岐がどのような言語的要素によってなされるのかを検証すること，３）４つのタイプに分類することの妥当性について検討する。

3. 分析方法とデータ

決定木は従属変数や独立変数に関して，量的変数と質的変数の両方を使うことができるため，コーパスの頻度情報を含め，人手によるコーディング情報（名義尺度）も扱うことができる（李 2008）。その意味で，コーパス日本語学における分析手法として汎用性が高いと言える。また，統計検定による有意水準を使用し，説明変数の値を評価する。その際，統計的に等質である場合は値を結合し，異質である場合は値を保持することで，従属変数の値に対する最適な分類を行う（玉岡 2008）。

調査においては，BCCWJ の中納言（https://chunagon.ninjal.ac.jp/）を使用し，「明らか」の 9844 個の KWIC データを作成した。全データをタイプ A の明白型からタイプ D のはっきり型まで人手で分類した。

表１．コーパス別の「明らか」タイプの頻度

コーパス	明白型	公表型	判断提示型	はっきり型	合計
生産・雑誌	46	125	17	96	284
生産・書籍	870	2165	249	888	4172
生産・新聞	20	250	5	16	291
非母集団・ブログ	41	346	30	197	614
非母集団・ベストセラー	72	103	31	94	300
非母集団・教科書	3	63	3	6	75
非母集団・知恵袋	28	59	34	393	514
流通・書籍	849	1549	212	984	3594
合計	**4660**	**581**	**1929**	**2674**	**9844**

次に個々の用例の表層の出現文脈を３つのタグセットでコーディングした。

独立変数１群：先行要素のコーディング
独立変数２群：直前格のコーディング
独立変数３群：後続要素のコーディング

決定木分析では，タイプ A からタイプ D までを従属変数に，生起文脈を独立変数に投入し，生起文脈から語義を予測するタスクを行った。決定木の成長方法は，CHAID 法を用いた。無作為抽出でデータの全体の 50%を学習データ，残り 50%を評価データとして使用した。実験は同じ条件で，

[1] 古語における「明らか」は，もともと知覚形容動詞で，知覚的な意味しかなかった（進藤 2009, Shindo 2010）。例文(5)は，視覚的な明るさを示し，例文(6)は，聴覚的な鮮明さを示す。

(5) 夜深き月の明らかにさし出でて，山の端近き心地するに，念誦いとあはれにしたまひて，昔物語したまふ。（1006年頃 源氏・椎本）

(6) それを何（なに）ぞといふに，音声（おんじやう）がいささか鼻声で，明らかにないと申すが，真や，この比（ごろ）は音声も明らかになって，歌はせらるる声も面白いと承る。（1593 エソポのハブラス，p.89 l.14-17）

独立変数のみをかえて５回行った。

実験１：独立変数１群のみで分析
実験２：独立変数２群のみで分析
実験３：独立変数３群のみで分析
実験４：独立変数１群＋２群で分析
実験５：独立変数１群＋２群＋３群で分析

各実験におけるタイプの予測の精度を比較検討した。

4. 結果

５つの実験における正答率は，表２の通りである。

表２．全タイプの正答率

区分	学習データ	評価データ
実験1	44.5%	45.2%
実験2	46.4%	45.3%
実験3	77.1%	76.9%
実験4	43.9%	44.2%
実験5	87.1%	86.2%

実験３と実験５の正答率の上昇から全体の多義の分岐を予測する上で，独立変数３群（後続要素）が大きく貢献していることが明らかになった。次に，各タイプにおける正答率を確認した。

表３．学習データのタイプ別正答率

区分	明白型	公表型	話者判断型提示	はっきり型
実験1	44.8%	61.6%	10.0%	61.7%
実験2	56.2%	97.1%	12.3%	20.2%
実験3	59.8%	84.2%	48.6%	100%
実験4	30.1%	97.2%	12.2%	36.1%
実験5	56.8%	96.7%	52.0%	100%

表４．評価データのタイプ別正答率

区分	明白型	公表型	話者判断型提示	はっきり型
実験1	46.8%	62.2%	11.2%	60.5%
実験2	53.8%	96.7%	11.6%	19.3%
実験3	58.2%	83.3%	49.8%	100%
実験4	31.8%	96.6%	10.5%	38.1%
実験5	53.1%	96.9%	46.3%	100%

タイプ別結果を見ると，明白型は独立変数３群，公表型は独立変数２群，はっきり型は独立変数１群と３群，話者判断提示型は独立変数３群によって，正答率が上がった。

最後に，実験５による分類の結果と決定木を示す。

表５．実験５の分類

観測値		予測値				
		公表	判断提示	明白	はっきり	正答率
学習	公表	2276	0	76	0	96.7%
	判断提示	0	153	0	141	52.0%
	明白	420	0	553	0	56.8%
	はっきり	0	0	0	1342	100%
評価	公表	2238	0	70	0	96.9%
	判断提示	0	133	0	154	46.3%
	明白	448	0	508	0	53.1%
	はっきり	0	0	0	1332	100%

表５の通り，話者判断提示型と明白型における誤分類の比率が高いことが明らかになった。また誤分類の傾向として，公表型と明白型の間で誤分類の問題が発生し，話者判断提示型とはっきり型で誤分類の問題が発生することはあっても，公表型と話者判断提示型で誤分類は発生しないことが明らかになった。このことは， 述語相当の振る舞いをする公表型と明白型，修飾節を構成する話者判断提示型とはっきり型で意味のグループが存在することを示唆している。

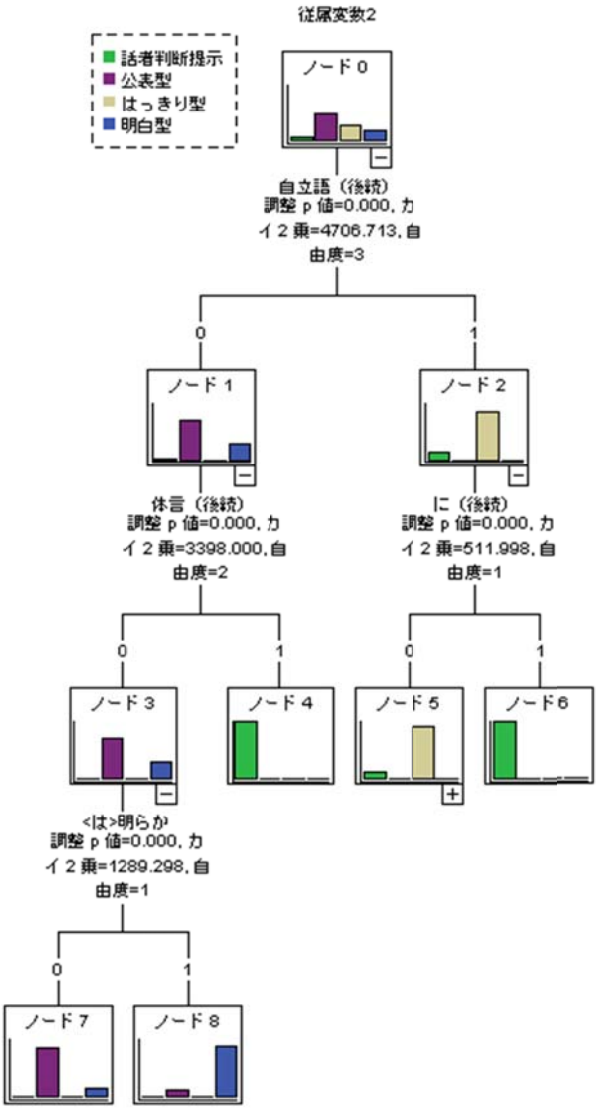

図１．決定木

図１を見ると，全体のデータの分岐で，後続要素が自立語かどうかで公表型・明白型とはっきり型・話者判断提示型で分岐している（$\chi^2(3)=4706.13$, $p<.0001$）。次に後続要素として，体言が来るがどうかで一部の話者判断提示型とそれ以外の要素で分岐している（$\chi^2(2)=3398$, $p<.0001$）。次に，後続要素として，助詞の「に」が来るかどうかで，はっきり型と話者判断提示型が分岐している（$\chi^2(1)=511.998$, $p<.0001$）。そして，公表型と明白型の分岐には，先行要素として「は」が出現しているかどうかで分岐していることが明らかになった（$\chi^2(1)=1289.298$, $p<.0001$）。

5. まとめ

本研究では，「明らか」を例に決定木を用いた多義語分析の方法を紹介した。分析の結果，「明らか」の曖昧性解消においては，後続要素が重要であること，各タイプによって曖昧性解消において貢献する変数が異なることが明らかになった。

* 謝辞

本研究は，科研費基盤研究（C）「感覚語彙の歴史的変化における構文と意味の相互関係：認知類型論的コーパス対照研究（23520477）」による補助を得て行った。

【参考文献】

進藤三佳(2009)「視覚形容詞から強調詞への意味変化：文法化の対照言語学的研究」、山梨正明 他(編)『認知言語学論考 Vol. 8』、157-190、ひつじ書房.

Shindo, M. (2010) “A Cross-linguistic Study of Constructionalization in the Grammaticalization of Adjectives.” Paper Presented at the 6th International Conference on Construction Grammar (ICCG-6). Prague: Charles University (Czech Republic).

進藤三佳・李在鎬・渋谷良方（2011）「感覚形容詞の語用論的意味変化に見る統語構造の影響」『日本語用論学会　第13回大会発表論文集 第6号』, pp. 57-64.

玉岡賀津雄 (2006)「「決定木」分析によるコーパス研究の可能性: 副詞と共起する接続助詞「から」「ので」「のに」の文中・文末表現を例に」『自然言語処理』13(2), 169-179.

李 在鎬(2008)「移動動詞に対する実験的分析」、『言葉と認知のメカニズム』、87-101、ひつじ書房.