コーパスとしての教材：語学教師のための分析とツール

竹井　光子
広島市立大学大学院
情報科学研究科
yamuram@nlp.its.hiroshima-cu.ac.jp

相沢　輝昭
広島市立大学
情報科学部
aizawa@its.hiroshima-cu.ac.jp

藤原　美保
ウィラメット大学
日本語・中国語学科
mfujiwar@willamette.edu

1　はじめに

コーパス分析は，言語研究において様々な役割を果たしている。語学教師とコーパスの関わりにおいては，「学習者コーパス」の分析が言語習得研究や教育に示唆を与える点で重要である一方，「教材コーパス」の分析も効果的な指導を行う上で有効である。

本稿では，「コーパスとしての教材」の分析を難易度や困難点の判定という観点から論じる。特に，判定基準として，一般的な語彙・統語的な要素に加え，日本語の省略現象であるゼロ代名詞（以後，ゼロ）に注目することを提案するとともに，Yamura-Takei (2005) で行ったコーパス分析結果の中から，その重要性を示すデータを提示する。また，分析補助ツールとして，ゼロディテクターも紹介する。

2　読解難易度とゼロ

テキストの読み易さ (readability) については，語彙や構文，文の長さなどを基準とした研究が先行しているが，最近では談話の結束性 (cohesion)，一貫性 (coherence) に注目した研究も進んでいる (McNamara et al., unpublished)。日本語には，ゼロという結束性を形成し一貫性に貢献する談話現象があるが，その不可視性のためか，教師の教育的関心は他の文法項目ほどゼロに向いていないように思われる。

しかしながら，ゼロの種類やふるまいの多様性は想像以上に大きく，それが読解難易度に関係していることは明らかである (Fujiwara and Yamura-Takei, 2003; 2004)。また，あらかじめゼロの分布状況を知ることは学習者が直面しそうな困難点を予測することにつながる。

ゼロに注意を向けることの必要性を実証するため，コーパス分析の結果を，ゼロの (1) タイプ (type)，(2) 推論量 (inference load) の 2 つの観点から考察する。

2.1　タイプ (type)

まず，第 1 のタイプとして，(1) **ゼロ連用項**と(2) **ゼロ連体項**に分ける。(1) は，動詞の統語的必須項の省略であり，(2) は，名詞の意味的不完さ[1] によって想起される要素である。それぞれの例を挙げる。

(1) 昨日は，［　が］カレーを食べた。
(2) ［　の］身長は 50 センチだ。

(1)では，「食べる」という用言（動詞）が要求する要素のうち欠けている主格を**ゼロ連用項**[2]とする。(2)では，「身長」という体言（名詞）が意味的に完結するために必要とする「何（誰）の」という情報を欠いている。この省略されている連体格要素を**ゼロ連体項**[3]と定義する。

第 2 のタイプは，指示対象によって次の 8 つに分類し，その分布状況を調査した。

(i)　**local**: 指示対象が直前の節にある
(ii)　**global**: 指示対象が直前の節より前にある
(iii)　**intra-clausal**: 指示対象が同一節内にある
(iv)　**cataphorical**: 指示対象が後方文脈にある
(v)　**event**: 指示対象が事象や陳述である
(vi)　**situational**: 指示対象が文脈内に存在しないが文脈から推測可能である
(vii) **indeterminate**: 指示対象が特定の事物ではなく一般的な人・物である
(viii)**time/weather**: 時や天候を表す

2.2　推論量 (inference load)

談話における一貫性と推論量および指示表現の選択との関係をモデル化した理論としてセンタリング理論 (Grosz et al., 1995) がある。センタリングは，談話内の発話（節）ごとに更新される「話題の焦点」を CENTER (CB)，その「更新のされ方」を TRANSITION として規定している。TRANSITION には CONTINUE(CON), RETAIN(RET), SHIFT の 3 つがあり，その連続の仕方によって，談話の一貫性の度合，談話の解釈に必要となる推論量が異なるとしている。

Grosz らによると，CON-CON の並びは RET-RET, SHIFT-SHIFT よりも好まれる。また，RETAIN は後続の CENTER の変化を予測させる状態であり，RET-SHIFT は望ましい連続である。さらに，CENTER の移行が完了した状態から新たな CENTER が連続することは自

[1] 例えば，西山(2003) の「非飽和名詞」がこれに含まれる。

[2] ゼロ連用項はさらにゼロ主格（ガ格），ゼロ対格（ヲ格），ゼロ与格（ニ格），その他に分類できる。

[3] ゼロ連体項はゼロ属格（ノ格）のみを含む。

然であり，SHIFT-CON も妥当な流れである。したがって，CON-CON は一貫性のきわめて高い連続，CON-RET-SHIFT-CON は理想的な CENTER 移行の流れであると言える。

そして，このような好ましい流れの中ではゼロが使われるであろうということが仮定できる。この仮定を統計的に実証するため，11 の TRANSITION の連続パターン[4]を想定し，それぞれの連続でゼロが CENTER となっている割合を調査した。

3 コーパス分析

コーパスは，8 つの日本語教科書中の読解教材 87 テキスト（2,007 節）を使用し，人手によりゼロの検出，タイプの分類，TRANSITION の計算を行った。

また，読解難易度に関連して，テキストのジャンルによるゼロの分布の違いにも注目する。コーパスの中から「物語文 (narrative)」と「論説文 (expository)」の 2 つのサブコーパスを選択し，比較に用いた。

分析の目的は，ゼロの分布やふるまいの多様性，およびジャンルやテキストによる傾向や特徴を示し，ゼロ指導への示唆を導き出すことである。

3.1 タイプ分布

まず，コーパス中に現れたゼロ（総数 1,382）の項（格）タイプによる分布を見ると（図 1），ゼロ連用項の中でもゼロ主格（すなわち主語の省略）が圧倒的に多いことがわかる。一方，ゼロ対格やゼロ与格はわずかで，ゼロ連体格がゼロ主格についで高い頻度を示している。

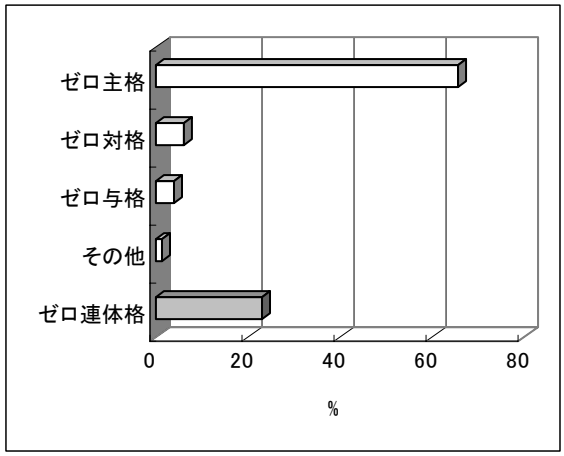

図 1：ゼロ項（格）タイプの分布状況

[4] CON-CON, RET-CON, SHIFT-CON, NULL-CON, CON-RET, RET-RET, SHIFT-RET, NULL-RET, CON-SHIFT, RET-SHIFT, SHIFT-SHIFT の 11 である。NULL-SHIFT は理論上存在しない。

ゼロ連体項は，ゼロ連用項に比べると，省略現象の一つとして扱っている研究は少ない。[5]しかし，言わないことによって結束性を生じさせるという機能はゼロ連用項と同様であり，話題の一貫性を維持する役割としては重要である。

ゼロ連用項とゼロ連体項のどちらが学習者にとって認識しにくいかの問題は，今後の検証研究を待つべきであるが，ゼロ連用のみで CENTER を構成している場合とゼロ連用とゼロ連体が関わり合いながら CENTER を構成している場合とを比べた場合，後者の方が結束性を認識しにくいのではないかと想像できる。統語的理由から生じるゼロ連用項と違い，意味的補足を必要とする名詞にのみ現われるゼロ連体項は，その存在自体を確認しにくいと考えられるためである。このゼロ連用項とゼロ連体項の分布状況を見てみると，コーパス内には表 1 に示すように両者の場合が存在する。

ジャンル	Text	ゼロ連用 CB	ゼロ連体 CB
物語	A	7	0
	B	2	5
論説	C	13	0
	D	7	6

表 1：テキスト別のゼロ連用・連体ゼロの出現傾向

物語 B，論説 D のような項タイプ混合のテキストは，ゼロ連体項が関与している分，ゼロ連用項のみが使われたテキストよりも話題の連続がとらえにくいと考えられるので，それぞれのテキストの特徴をふまえた指導が重要であろう。

次に指示対象のタイプによるゼロの分布状況を表 2 に示す。

タイプ	個数（率）
local	887 (64.17%)
global	146 (10.56%)
intra-clausal	130 (9.41%)
indeterminate	104 (7.53%)
situational	56 (4.05%)
event	21 (1.52%)
cataphorical	20 (1.45%)
time/weather	18 (1.30%)
計	1,382 (100%)

表 2:指示対象のタイプ別ゼロ分布

隣接する文間の結束関係を形成するゼロの主機能を象徴するように，local が過半数を占めている。一方で，文脈からの類推を必要とする situational や一般

[5] 本研究でゼロ連体項と定義しているものの内，「建物」－「（建物の）入り口」などの例は「間接照応」として論じられている場合がある。

的な人を示す indeterminate，前述の文全体や時には談話節全体の内容を指すこともある event などが，低頻度ながらも使われている。

遠距離照応の global は，先行詞が見つかる範囲によってさらに分析できる。図 2 に，コーパスに現れた 146 個の global ゼロの先行詞との距離を示す。距離 1 とは，直前の節のことであるが談話節（ここでは段落）の境界をまたいでいる場合である。距離 2 とは，ゼロと先行詞の間に節が 1 つ存在する場合である。

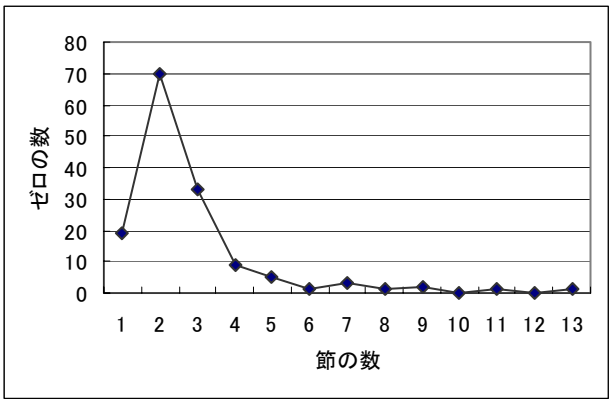

図 2：global ゼロと先行詞との距離

Global の大半が直前の節の 1 つ前に先行詞を見つけることが出来るが，10 節以上前にさかのぼらなければならない例もあることは注目すべき点である。

次に，指示対象別の分布を 2 つのジャンルで比較すると図 3 のようになる。

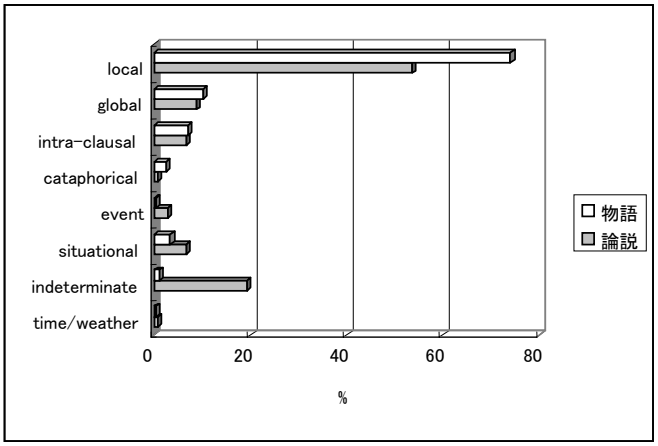

図 3：ジャンル別指示対象タイプの分布

特徴的なのは，物語における local と，論説における indeterminate の高頻度である。これは，ジャンルとしての特徴（主登場人物を中心に出来事を述べる物語と事象や概念などを解説する論説）と直感的に一致する結果である。

しかし，同じ論説文であってもテキストによって各タイプのゼロの分布は異なる。表 3 に 2 つの論説タイプの読解テキストにおけるゼロの分布の相違を示す。

ゼロの種類	テキスト A	テキスト B	コーパス全体
local	66.67	14.29	53.59
global	26.67	7.14	8.86
event	6.67	7.14	2.95
situational	0	7.14	6.75
Intermediate	0	64.29	19.94

表 3：論説タイプ 2 テキストにおけるゼロの分布(%)

この表から，同じ論説文であっても，local, global が大半を占めるテキスト(A)と，indeterminate, situational が多用されているテキスト（B）があることが分かる。また，これらはコーパス全体の分布平均からも逸脱しており，ゼロのタイプ分布の観点からは特徴的なテキストであることがわかる。

論説文には indeterminate，物語文には local などと，テキストの種類によって頻出するゼロの種類の一般的傾向を述べるには，さらに大きなコーパスでの検証を必要とするが，テキストのタイプによる頻出ゼロの傾向が分かれば，教師の教材研究に役立ち，問題点に焦点を絞った指導ができるであろう。

3.2 推論量

センタリング理論によって予測される推論量とゼロの使用との関係を，11 の TRANSITION の連続列において CENTER (CB)がゼロによって実現されている割合で示したのが図 4 である。

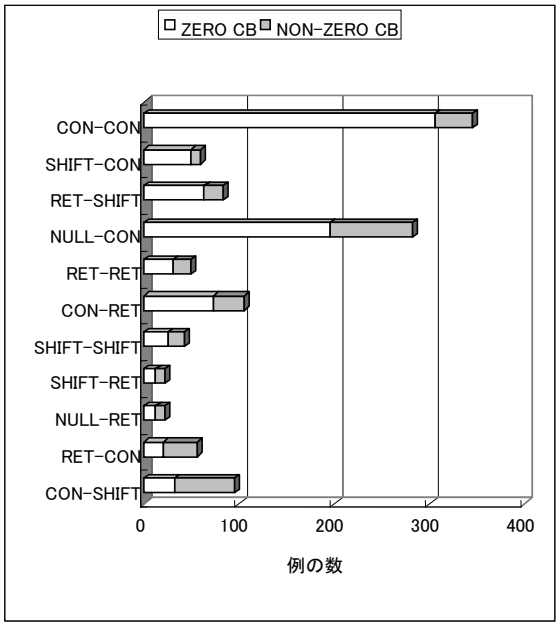

図 4：TRANSITION の連続とゼロの分布

ゼロ CENTER (CB)の割合が高いものから順に並べてある。CON-CON の 88.44%から CON-SHIFT の 33.68%までの広がりがある。

前述の望ましい TRANSITION の連続 (CON-CON, RET-SFHIT, SHIFT-CON) は，すべて上位にある。理想的な CENTER (CB)の遷移環境においては，実際にゼロが使用されやすいということがわかる。

一方で，好まれない（すなわち多くの推論量を要求すると予測される）TRANSITION の連続[6]においてもゼロが CENTER (CB)となっている場合がある。この事実は，学習者が困難と感じる，または誤った解釈をしてしまう可能性を示唆している。これらの環境で曖昧さを感じることなくゼロを正しく解釈するためには，センタリング以外の推論のための情報源[7]が必要であり，教師はそれを把握しておくことが必要であろう。

4 コーパス分析ツール

ゼロは表層上見えない。見えないために，コーパスの統計的な分析を行ったり，分布状況から困難点を判定したりするのが極めて難しい。特に日本語母語話者教師にとって，ゼロは自然な存在であり，普段あまり意識することがない。前節で述べたような視点でゼロを観察・分析するためにはまずゼロの存在を認識することが必要であるが，その補助ツールとしてゼロの自動検出・明示システムであるゼロディテクター(ZD)を開発した(Yamura-Takei, 2005)。

ZD の出力をもとに，ゼロの分布や種別の出現傾向を観察し，その結果を目的に合った教材の選択や困難点の予測に活用することを提案する。

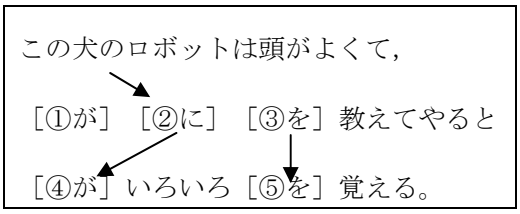

図 5：ZD 出力の利用例

図 5 は ZD 出力の利用例であるが，指示関係を矢印で示すことにより，①③が先行詞のないゼロ (situational)であることがわかる。

[6] 例えば，RET-CON は CENTER の変化を予感させながらそれをくつがえすという点で，CON-SHIFT は CENTER 変化の前兆なしに唐突に移行するという点で好ましい流れではないとされる。

[7] 推論量が高いとされる環境におけるゼロの使用を観察した結果，選択制限，主題化，視点，平行構造，文脈知識，世界知識，文化的背景などが影響していることがわかった。

5 まとめ

実際に教室で使われている読解教材をコーパスとして，ゼロの分布状況を調査した。その結果，ゼロの種類や出現環境の多様性，ジャンル・テキストによる傾向の違いを示すことが出来た。

ZD の活用例として，竹井ら(2004)では，第二言語習得理論に基づき，ゼロの習得促進を目的とする教材作成のための補助ツールとしての役割を論じた。それに加え，本稿では，教材難易度判定のための補助ツールとしての ZD の役割を論じるとともに，ゼロ分析の必要性を実証するデータを示した。

ゼロの分布状況にもとづく難易度の判定基準の妥当性については，実証研究を待たなければならないが，ある程度教師の直感や経験に基づいて行うこともできるであろう。ただし，その土台としてまず教師がゼロの存在を認識し，また意識することが必要である。本稿で示した提案や分析結果が実際の読解指導に活用されることを期待したい。

参考文献

Fuijiwara, Miho and Mitsuko Yamura-Takei. 2003. Accessing the difficulty of Japanese reading materials: A zero pronoun perspective. Presented at ATJ Seminar 2003, March 27, 2003, New York City.

Fujiwara, Miho and Mitsuko Yamura-Takei. 2004. Zero anaphora analyses to access reading difficulty in Japanese text. Presented at the 4th Biennial International Conference on Practical Linguistics of Japanese (ICPLJ2004), April 3-4, San Francisco.

Grosz, Barbara, Aravind Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21/2, 203-225.

McNamara, Danielle, Max. Louwerse, and Art Graesser. Unpublished. Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension. Grant proposal.

竹井光子，磯江健史，相沢輝昭．2003. 第二言語習得におけるインプット強化と自然言語処理技術．言語処理学会第１０回年次大会発表論文集，東京．

西山佑司．2003．日本語名詞句の意味論と語用論－指示的名詞句と非指示的名詞句．ひつじ書房．

Yamura-Takei, Mitsuko. 2005. Theoretical, Technological and Pedagogical Approaches to Zero Arguments in Japanese Discourse: Making the Invisible Visible. Doctoral thesis (in review), Hiroshima City University.