『現代日本語書き言葉均衡コーパス』に対する節の意味分類情報アノテーション ： 基準策定，仕様書作成の必要性について

著者	松本 理美，浅原 正幸，有田 節子
雑誌名	言語資源活用ワークショップ発表論文集
巻	1
ページ	336-346
発行年	2017
URL	http://doi.org/10.15084/00001489

『現代日本語書き言葉均衡コーパス』に対する節の意味分類情報アノテーション —基準策定、仕様書作成の必要性について—

松本 理美（立命館大学大学院言語教育情報研究科）
浅原 正幸（国立国語研究所コーパス開発センター）
有田 節子（立命館大学大学院言語教育情報研究科）

Clause Class Annotations on the 'Balanced Corpus of Contemporary Written Japanese' -- Necessity of Developing Fine-grained Criteria and Specification

Satomi Matsumoto (Graduate school of Language Education and Information Science, Ritsumeikan University)
Masayuki Asahara (National Institute for Japanese Language and Linguistics)
Setsuko Arita (Graduate school of Language Education and Information Science, Ritsumeikan University)

要旨

本発表では、「現代日本語書き言葉均衡コーパス」に対する節の意味分類情報アノテーションについて報告する。多様な形式を持ち、文脈の中でその意味が解釈される日本語文中の従属節の意味分類については、人手による分類が不可欠である。そこで、我々は「鳥バンク」基準互換 (池原 2009) の節の意味分類情報アノテーションを進めている。しかし、現行の作業においては、節の認定、タグ付け箇所、作業者の言語感覚に頼るところが大きい意味分類判断など、作業上の揺れも多く、改善が求められる。作業効率と信頼性の向上に繋がる基準策定と仕様書作成が必要であり、そのためには現行作業での問題点を整理することが必須であると考える。本発表では、人手による節境界アノテーション・節の意味分類タグ付け作業についての基準策定と仕様書作成が今後のコーパス開発に資することを主張し、現在の作業における問題点に焦点を当てた考察を行う。

1. はじめに

本発表では、我々が進めてきた「現代日本語書き言葉均衡コーパス」に対する節の意味分類情報アノテーションについて紹介する。

人手による節境界アノテーション・節の意味分類タグ付け作業（以下、作業とする）における問題点を整理し、作業上の基準策定や作業仕様書の必要性を示すことを本発表の目的とする。アノテーション整備のゴールが“正確な”情報付与でないことは言うまでもない。設計に関わる諸分野の専門家と多種多様な利用者による問題提起や議論、一方通行ではない情報の授受というアノテーションを介在したコミュニケーションが、言語現象に関する新たな発見とコーパス言語学の発展をもたらす可能性は大きいと考える。

そこで、節認定・タグ付け箇所・意味分類のいずれにも判断の揺れが生じることを許容したうえで、人手による作業だからこそ可能となる判断の許容範囲の探求を試みる。作業者間、あるいは解析器と人間の間に生じた齟齬の分析により技術的、理論的な問題を明らかにする。

実際のアノテーションにおける齟齬について紹介すると次のような傾向がある。補足節

における齟齬の発生は他の節と比較して少なく、特定の節に偏って齟齬が見られるということもない。作業者間で多くの齟齬が生じる名詞修飾節は他の節と認定されることは極めて少ないが、名詞修飾節内の下位分類間での齟齬が大量に発生している。同じく作業者間で多くの齟齬が生じる副詞節は、並列節と認定される齟齬が目立つ。同様に並列節は専ら副詞節との間に齟齬を生じる傾向にあり、副詞節 VS 並列節の様相を呈している。本稿では、これらの各論について節認定・タグ付け箇所・意味分類の三つの観点から議論する。

一言で節認定・タグ付け箇所・意味分類の齟齬といっても、従属節の一つ一つの個性が、その齟齬にも反映される結果となっている。本研究は、工学的な要求からコーパスを設計する工学研究者、文法的視点から節認定や意味分類を検討する日本語学研究者、節境界アノテーション・節の意味分類タグ付けを行った作業者による共同発表であるが、それぞれの観点で「どの程度の齟齬を許容すればよいか」、その“落としどころ”をどこに求めるかは、非常に難しい問題である。それでも敢えて、作業者間、作業者内での揺れを想定した作業上の基準策定や仕様書作成の必要性を主張する。それにより作業の効率化と、データの信頼性向上の可能性があると考えるからである。

次節以降、解析器への実装を前提とした作業上の問題点を論じる。

2. 作業における基準と問題点

本作業では、節を「複文を構成するところの、述語を中心とした各まとまり」(益岡・田窪 1992 : 4) と定義する。また、従属節の意味分類については、益岡・田窪 (1992)、益岡 (1997) を参考に、「実際の文型パターンに関する用例分析の結果に基づき」作成された池原 (2009) の分類体系を基準とする。

作業開始時の確認事項は、述語を含む 2 文節[1]以上からなるものを節とし、節末の接続表現の右端にタグをつけるということであった。作業においては、「節間意味分類体系」(池原 2009) のみを手掛かりに作業を行った。

以下作業の概要について示す。

最初に「鳥バンク」のパターンを UniDic 品詞体系に対応させた節自動解析器により、可能な節境界をすべて枚挙する (浅原ほか 2015: https://github.com/masayu-a/clause_pattern)。この可能な候補を見ながら、作業者 2 名 (作業者 A, 作業者 B) がお互いのアノテーションを見ずに 1 次タグ付け作業を行う。その後、作業者 A が機械の出力および 2 名分の 1 次タグ付け作業結果を見ながら 2 次タグ付け作業を行う。対象は「現代日本語書き言葉均衡コーパス」の新聞記事コアデータの一部 (PN) 54 ファイル (優先順位 00001～00054: A 集合) とした。

この作業過程において、作業者間、作業者内での判断の不一致や揺れが多数生じた。以下、特に、節認定、タグ付け箇所、意味分類で生じた作業者間の不一致や揺れについて、実例を挙げ、その問題が発生した原因について考察を行う。

3. 節の意味分類アノテーションにおける問題点と策定するべき基準

以下では、「句」ではなく「節」として認定するか否か (節認定)、「節境界」をどの位置に設定するか (タグ付け箇所)、「節」の意味分類としてどのラベルを付与するか (意味分類) についてそれぞれ示す。

[1] 文節の定義は、小椋他 (2011) の文節認定規定に従う。

3. 1　節認定について

節認定についての問題点と、それを踏まえた基準策定について述べる。

3. 1. 1　節認定の問題点

節認定の問題に関しては、2 つに分けて論じる必要があると考える。1 つは、何を節と認定するかという問題であり、2 つめは、どこからどこまでを節と認定するかという問題である。

まず、何を節と認定するかという問題について述べる。現行の作業では、「述語を含む 2 文節以上を節とする」という基準で作業をしているが、この場合、以下の例文のような節認定に問題が生じる。

なお、以下の例文について、出典を示していないものは、筆者の作例である。

(1) 彼らは一晩中、飲んだり食べたりしていた。
(2) 彼らは一晩中、酒を飲んだり、つまみを食べたりしていた。

(1) について、「彼らは一晩中、食べたりしていた。」を主節とすると、節認定には 2 文節以上を必要とする現行の作業基準では、「飲んだり」は節と認定しないが、(2) の「酒を飲んだり」は節と認定をする。(1) を節認定せず、(2) を節認定するという妥当な根拠はなく、現行の基準には問題があると考える。

また、丸山他 (2016) が指摘する通り、名詞修飾節の認定にもいくつかの問題がある。

(3) 青いビンを見つけた。
(4) ふたが青いビンを見つけた。

上記の例文では、(3) (4) ともに「青い」は「ビン」という名詞を修飾しているが、現行の作業基準では、(3) は「青い」が単独で「ビン」を修飾しているため、名詞修飾節とは認定されない。それに対し (4) の「ふたが青い」は、2 文節であることから節と認定される。このように、文節数のみを節認定の基準とすることには議論の余地があると考える。

(5) 放置された車がある。
(6) ガレージに放置された車がある。

(6) の「ガレージに放置された」を節として認定することには問題なさそうであるが、(5) の「放置された」の節認定は、どのように判断すればよいか。現行の基準で判断するならば、1 文節の「放置された」は節と認定されないことになる。しかし、「放置された」という、過去の出来事を描写している意味的な特徴を無視して、単独の文節であるという形式を根拠に節認定しないことには、問題がある。さらに「放置された」は、形態素にわけると「放置する＋れる＋た」となることからも、同じ単独文節である (3) の「青い」の節認定の問題とは性質を異にすると考えられる。

丸山他 (2016)　においては、寺村 (1981) が「主体が文脈からわかること、その述語にテンスの意識があること、という 2 点を満たす場合にのみ『節』を認める、という立場をとっている」ことに言及し、「対象の属性を規定する名詞修飾表現は (タ形をとっていても) 連

体節とは認められないことになる[2]。しかし、このような意味的な違いを表層の単語列から判定するのは極めて難しい。」(丸山他 2016 : 1116) としている。この点については、同意するところが多く、特に形式からだけでは、「テンスの意識」があるかどうかの判断は不可能である。

本作業の基準から、属性についての名詞修飾節認定を論じると、「曲がった木」の「曲がった」は単独の文節であることから節認定されず、「先が曲がった木」であれば 2 文節であることから節認定されることになる。このように、属性を表す名詞修飾表現であっても 2 文節以上で現れることはあり得ることであり、この基準にも議論の余地があると言える。

また、どこからどこまでを節と認定するかということについても、以下のような問題が指摘できる。

(7) この部屋は隅々まで掃除するのに 3 日かかった。
(8) この部屋は隅々まで掃除するのにどうしてあの部屋はしないの。

(7) では、「この部屋は」を主題とし、「隅々まで掃除する」を形式名詞「の」に係る名詞修飾節としたうえで、「隅々まで掃除するのに」を「かかった」という述語の補足節と認定するのが、厳密には正しい節認定であろう。ただ、本作業では、作業の手がかりである「節間意味分類体系」 (池原 2009) が、「のに」を補足節の「節間キーワード[3]」としているため、「隅々まで掃除するのに」を補足節と認定し、名詞修飾節の認定は行っていない。この点については、仕様書があることで作業上の揺れはなくなると思われるが、どこまでが実装の際に必要な情報であるかということには、検討の余地があると考える。

(8) については、「この部屋は」を対照主題とし、「隅々まで掃除するのに」は逆接の副詞節である。この「のに」は南の 4 分類[4]において、C 類の副詞節の節形式とされているものであり、「この部屋は」のような主題句は C 類の副詞節の構成要素になりうる。

このように、 (7) と (8) は、表層上は同じであっても、全く異なった文法機能を持った節であるといえる。

(9) 公園を散歩していた時、その事件を目撃した。

(9) では、「公園を散歩いていた時」を、時間を表す副詞節 (時間節) と認定することも、「公園を散歩していた」を「時」の名詞修飾節と認定することも可能である。

以上のように、事態性か属性かという述語の性質、意味、機能を表層上区別する手立てはなく、厳密な節認定を行う方向で議論すると、認定は益々複雑となることが予想される。そこで、文法上の議論を踏まえたうえで、実装を前提としたときの許容範囲を有した妥協点を探索する必要があると考える。その意味でも、揺れを前提とした基準と作業仕様があること

[2] 丸山他 (2016) では、寺村 (1981) のように「お茶がほしい人」を「節」として認め、「やせた人」を「節」ではなく「句」とするという立場をとると、節認定しないことになる例として、「飲むヨーグルト」「やせている男」「曲がった道」を挙げている。

[3]池原 (2009) の用語で、「従属節と主節の意味的な関係を決める接続表現部分」を指す。補足節では形式名詞「こと」「の」「ところ」などを節間キーワードとしている。

[4] 南 (1974) は、日本語の従属節について、節の構成要素を述語的部分の要素と述語的部分以外の要素に分けて論じている。そして、それぞれの要素が節内に存在することの可否を根拠に、従属節を A 類、B 類、C 類、D 類に 4 分類している。

で、齟齬が見られた節を要注意節として、実装時に役立てることができると考える。

次に、節認定における齟齬の発生割合が、従属節によって異なることについて述べる。

一方の作業者は節と認定し、他方の作業者は節と認定していない (タグ付けを行っていない) という節認定齟齬が各従属節で生じている。

作業者のいずれか一方だけが節認定をした節の総数は 631 であった。以下の表 1 に従属節の種類別にその頻度を示す。

表 1　作業者の一方だけが節認定をした従属節の頻度　(　)内は割合(%)

補足節	名詞修飾節	副詞節	並列節
102 (16)	275 (44)	207 (33)	47 (7)

表 1 の割合をグラフで表わしたものが、図 1 である。

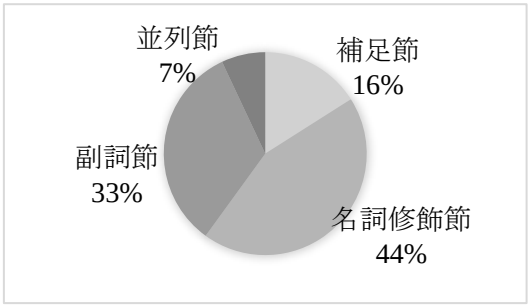

図 1　節認定に齟齬が発生する従属節の割合

作業者の一方のみが節認定した場合を見ると、齟齬の発生割合が従属節の種類によって異なることがうかがえる。作業者の一方が並列節と認定したが、他方はそれを節と認定していないという齟齬の割合が最小で、作業者の一方が名詞修飾節と認定したが、他方が節認定をしなかったという齟齬の割合が最大であった。

これを従属節の種類別にみると、作業者の一方のみが補足節と節認定した齟齬は、その 60%が引用節 : 間接引用において生じている。作業者の一方のみが節認定し、他方は節認定しなかったもので、名詞修飾節の認定における齟齬の 75%が内容節、補足語修飾節で生じ、副詞節では 50%が独立、付帯状況・様態、因果関係の原因において生じている。また、並列節における同様の節認定の齟齬のうち、70%が総記の並列で確認されている。

前述したとおり、述語の性質、意味から節を認定することは極めて難しく、節を文法上厳密に認定しようとすると文法機能上の判断も必要となるため、どの段階までの節情報を付与るすかという設計に関しても、これらの情報は有効であると考える。

3. 1. 2　節認定において策定するべき基準

文節数だけではなく、表層上の形式に着目しつつ、揺れを想定した基準を示すことが現実的であると考える。作業基準の許容範囲の幅をどの程度とするか、またどの深さまで分類する必要があるかは今後の課題である。

基準策定の際、述語になることができる品詞ごとに基準を設けることが考えられるが、本発表は作業上の問題点の整理であるため、例のみの提示とし、具体的な策定基準に関しては

次稿以降に譲る。

例） 動　詞・・・文節数や、活用の有無にかかわらず節認定を行う。

形容詞・・・補語を含まない基本形である場合のみ、節認定をしない。

文法上、節と認めることに議論の余地がある事例に関しても、広めの許容範囲を認め、節情報タグ相当の用法を検討して付与する方向での基準を策定する。

3. 2　タグ付け箇所について

タグ付け箇所についての作業上の問題点と、それを踏まえた基準策定について述べる。

3. 2. 1　タグ付け作業の問題点

タグ付け箇所の齟齬は、作業開始当初、同じ節認定をしているにもかかわらずタグ付け箇所が異なるなど多数生じていたが、作業が進むにつれ、改善、安定がみられ、節の認定上の齟齬がない場合のタグ付け箇所の齟齬は解消されている。

タグ付け箇所の齟齬は File00001 ～ 00014 (5 桁の数字は優先順位を表す https://github.com/masayu-a/BCCWJ-ANNOTATION-ORDER 参照) では 148 件確認されたが、file00015～00054 では 54 件に止まった。作業当初のタグ付け箇所の齟齬のほとんどは、名詞修飾節において発生しており、被修飾名詞の直前につけるか、被修飾名詞につけるかというものであったことを考えると、作業者の認識不足が原因であったと思われる。

3. 2. 2　タグ付け作業において策定するべき基準

開始当初と作業途中の差を考えると、仕様書に明記することにより、節認定が一致した場合は、作業者間のタグ付け箇所の齟齬は解消されると考える。

3. 3　意味分類について

意味分類についての問題点と、それを踏まえた基準策定について述べる。

3. 3. 1　意味分類の問題点

意味分類においては、節機能の齟齬[5]か、意味の齟齬かに分けて考える。

まず、節機能の齟齬について述べる。

作業者双方が節と認定をしていたものについて、作業者 A,B が付与したラベルの Level 1 における節機能の分類の結果を表 2 に示す。

表 2　作業者 A,B が付与したラベルの節機能分類 (Level 1)　数字は頻度

A \ B	補足節	名詞修飾節	副詞節	並列節
補足節	8	4	7	8
名詞修飾節	11	179	4	0
副詞節	12	6	85	125
並列節	1	0	26	2

作業者 A を基準にして考えたとき、file:00015~00054 で、作業者間で節機能分類において齟齬があったのは、補足節で 70%、名詞修飾節で 8%、副詞節で 63%、並列節で 93%であ

[5] 池原 (2009) において、副詞節とその他の 3 つの節では、Level 2 以下の分類基準が異なっており、これら全てを「意味分類」と表現することには、議論の余地があると考える。しかし、本稿では便宜上、文法的に従属節を分類した Level 1 を節機能の分類とし、Level 2 以下を意味分類とする。

った。

つまり、作業者 A と作業者 B で、認定した従属節の種類が異なるという節機能分類における齟齬は名詞修飾節で最も少なく、並列節で最も多かったということである。換言すれば、名詞修飾節は、節機能の齟齬は少ないが、名詞修飾節内での意味分類の齟齬が多く生じており、並列節では、従属節間の齟齬が大半を占めているということになる。

表 2 に示した通り、副詞節－並列節間での齟齬は多く見られるが、丸山他 (2016) が採用している分類体系では、「従属節を大きく 3 つに分けて、並列節を連用節の下位に位置づけて」(丸山他 2016 : 1114) おり、丸山他 (2016) の分類体系に基づくと、機能的なレベルでの作業者間の意味分類の齟齬はほとんど見られないことになる。

次に、作業者間において Level 1 の機能分類は一致していたが、Level 2 以降の意味分類で齟齬が生じた節について述べる。

Level 2 以下の意味分類の齟齬は、並列節、補足節ではほとんど生じておらず、齟齬の頻度が 5 を超える節はなかった。以下に、作業者間の齟齬の頻度が 10 以上であった節について、実例を挙げて詳説する。

まず、作業者双方が Level 1 で名詞修飾節と認定したが、それ以下の分類において作業者間に齟齬が生じたものを表 3 に示す。

表 3　名詞修飾節における作業者間の意味分類齟齬

頻度	作業者 A		作業者 B	
	Level 2	Level 3	Level 2	Level 3
46	補足語修飾節	限定的	補足語修飾節	非限定的
27	補足語修飾節	非限定的	補足語修飾節	限定的
23	内容節		補足語修飾節	限定的
14	縮約形修飾節		内容節	
14	縮約形修飾節		補足語修飾節	限定的
10	補足語修飾節	限定的	内容節	

表 3 に示した順に、名詞修飾節において Level 2 以下の意味分類に齟齬が生じた節から 1 例ずつ挙げる。

作業者が付与したラベルの内容を、作業者 A Level 2 (level 3) – 作業者 B Level 2 (level 3) と示し、例文を挙げて齟齬の解釈を行う。なお、例文の下線部は作業者間に意味分類の齟齬が生じた節である。

補足語修飾節 (限定的) －補足節修飾節 (非限定的)

(10) 雪舟作と伝えられる花鳥図屏風は、10 点余りが知られている。
[PN2b_00002, file:00019]

(10) は、補足語修飾節という判断で両者は一致しているが、「花鳥図屏風」が指す対象が一定しているか否かの判断において、作業者間の認識が一致しなかったため、Level 3 のタグに齟齬が生じたものである。

補足語修飾節 (非限定的) －補足節修飾節 (限定的)

(11) 警察当局が危険人物と認定した九百三十二人に対し、W杯開催の五日前までにパスポートを警察署に預ける命令 が出ているが、 [PN2c_00002, file:00020]

(11) も、(10) と同様に、補足語修飾節という判断で両者は一致しているが、「九百三十二人」が指す対象が一定しているか否かの判断において、作業者間の認識に不一致があったため、Level 3 のタグに齟齬が生じたものである。

内容節－補足語修飾節 (限定的)

(12) 再建計画に数値基準を設けた中間報告の中核的な考えに反映されている。 [PN1g_00002, file:00018]

(12) は、下線部が「中間報告」の内容を表し、修飾節と被修飾名詞が同格にあると考えた作業者 A に対し、作業者 B は、下線部は「中間報告」の指し示す対象を限定していると判断したといえる。

縮約形修飾節－内容節

(13) 診療所存続をめぐる話題が一本の柱だ。[PN1b_00003, file:00033]

(13) では、「診療所存続をめぐる」と「話題」における格関係の有無の判断により齟齬が生じた例である。作業者 A は修飾節と被修飾名詞に格関係はなく、「意味的に間接的な関係にある」 (池原 2009：293) と判断し、作業者 B は格関係を認めたものと考えられる。

しかし、益岡・田窪 (1992) では、池原 (2009) が縮約形修飾節としているものは全て内容節に含めており、格関係の有無に関係なく、「内容節とは、被修飾名詞が指し示す対象の内容を表す」(益岡・田窪 1992：203) と広く定義している。

このように、意味分類はその定義によっても判断が異なることがあり、どの深さまで分類するかは、議論の余地のあるところである。

縮約形修飾節－補足語修飾節 (限定的)

(14) 北朝鮮の核不拡散条約 (NPT) からの脱退宣言が 10 日に 90 日間の「脱退通告期間」切れとなるとされるのを受けた協議だが、（後略） [PN3a_00002, file:00024]

(14) において、作業者 A は修飾節と被修飾名詞は、格関係にも同格関係にもなく、間に「上で行われた」などが省略されていると判断し、作業者 B は修飾節が「協議」の指し示す対象を限定していると判断したものである。

補足語修飾節 (限定的) －内容節

(15) だが上品な画題とは似ても似つかぬ印象は、この屏風が型破りな作品を生涯描き続けた雪舟のまさに真筆だと明かしているようにも思う。[PN2b_00002, file:00019]

(15) で、作業者 A は、修飾節が被修飾名詞「印象」の指し示す対象を限定していると判断し、作業者 B は修飾節が被修飾名詞の内容を表すものであり、両者が同格であると判断している。

以上の齟齬の例をみると、名詞修飾節においては、修飾節と被修飾名詞の関係に判断の齟齬が見られるものと、被修飾名詞の性質についての判断に齟齬が見られるものがあることが明らかになった。

次に、作業者双方が Level 1 で副詞節と認定したが、それ以下の分類において作業者間に齟齬が生じたものを表 4 に示す。

表 4　副詞節における作業者間の意味分類齟齬

頻度	作業者 A		作業者 B	
	Level 2	Level 3	Level 2	Level 3
33	手段		因果関係	原因
12	付帯状況・様態	付帯状況	因果関係	原因

表 4 に示した順に、副詞節において Level 2 以下の意味分類に齟齬が生じた節から 1 例ずつ挙げる。

作業者が付与したラベルの内容を、作業者 A Level 2 (level 3) – 作業者 B Level 2 (level 3) と示し、例文を挙げて齟齬の解釈を行う。なお、例文の下線部は作業者間に意味分類の齟齬が生じた節である。

手段－因果関係 (原因)

(16) 他派閥からも引き抜いて三十人から五十人の新派閥をつくることができるんだ」。
[PN2e_00002, file:00021]

池原 (2009) では[6]、手段を表す副詞節は「主節の内容を行う前提」(池原 2009 : 300) を表す従属節であり、因果関係（原因）は、「従属節の内容が原因となって主節の内容が起こる」という「従属節と主節で表される事態間の因果関係を表す」(池原 2009 : 296) としている。

「主節の内容を行う前提」は当然のことながら、主節で表される事態に因果関係を持つものであることから、この齟齬は節の意味における分類の難しさ、曖昧さから生じるものであると考える。

付帯状況・様態 (付帯状況)－因果関係 (原因)

(17) 小泉内閣は、細川、橋本、小渕の各政権の積み残しを一手に引き受けて、そのすべてを処理するという重荷を背負っている。[PN1b_00004, file:00046]

これは作業者 A が、「引き受けて」を「処理する」に係る節と判断したことにより、付帯状況・様態 (付帯状況) のラベルを付与し、一方作業者 B は「引き受けて」を「背負っている」に係る節と判断したため、因果関係 (原因) のラベルを付与したと推測される。(17) は、作業者間でかかり受けの判断が異なったため、意味分類に齟齬が生じたと考える。

従属節の分類の難しさについては、南 (1974) が次のように述べていることからもうかがえる。

[6] 池原 (2009) は、益岡・田窪 (1992) および益岡 (1997) を参考にしているが、本稿では池原 (2009 : 292-303) の 8 章の付録表「3. 主節従属節間の意味分類体系」から引用している。

> (前略) 問題の従属句がどの類に属するかは、必ずしも、テとかナガラ、バといった接続助詞のいかんによってあらかじめきめられるものではないということである。それは、その句を構成しているすべての要素およびその句の文中での文法的性格による。
>
> (南 1974：130)

副詞節における意味分類は、主節と従属節の関係を文脈から判断することも多く、かかり受けや、節関係の判断に作業者の言語感覚や文法理解のレベルなどによって齟齬が生じることは必須である。

3. 3. 2 意味分類において策定するべき基準

池原 (2009) の意味分類体系においては、Level 2 以下に機能、意味、形式の混在が見られところがある。意味分類においての基準の統一について、その必要の有無も含め、どのように解決していくかという問題がある。

また、例えば、節間キーワードの「ところに」は、補足節の「トコロ型」にも、時を表す副詞節にも同義のものが示されており、作業者の混乱を招くと考えられる。ただ、これは意味分類体系に問題があるわけではなく、複数の意味に解釈される節末接続形式は存在するため、ラベル付与作業における工夫が求められる。

さらに、池原 (2009) の節間キーワードは、「約 1000 件の文型パターンを対象に、従属節と主節とを連結するキーワードとその意味に着目した用例分析を行い、分類を詳細化した」(p.259) ものである。キーワードは、形式に着目したものであるため、助詞が多いものの、助動詞も相当数見られ、動詞も混在しているなど、抽出方法に文法的一貫性がない。一般化、法則化が難しいようであれば、池原 (2009) の分類体系にパターンを追加、修正していくことが求められると考える。

意味分類において、池原 (2009) は、「彼らは会えば必ずけんかする。」の「ば必ず」を、時を表す副詞節の節間キーワードとしている。一方で、法則的条件を表す副詞節の定義として「ある事態が起こると法則的に必ず別のある事態が起こる」ことを表すとしている。ここには矛盾があり、文法的にも「会えば」を時間節とすることには議論の余地がある。他にもいくつか同様のものが見られ、若干の修正が必要ではないかと考える。

4. おわりに

本稿では、節境界アノテーション・節の意味分類タグ付けに関する現行作業の問題点を整理し、より効率的に、信頼性の高い結果を得るための議論を行った。そして、作業において、文法的に厳密で正確な判断を追及することを目的としたものではなく、文法上の議論の余地があるものについても、アノテーションとしていかに記号化していくかという観点で、許容範囲を探り、節認定や意味分類の揺れを整理することを試みた。

副詞節・並列節間の節認定の齟齬は、これらの節の分類の解消か、機能の相違点を明示し厳密に分類するかという議論をもたらす。また、従属節内の意味分類 (level 2 以下) で頻度の高い齟齬を見ると、文脈からの判断が必要なものもあり、名詞修飾節における補足語修飾節の限定か非限定かの分類や、内容節－補足語修飾節－縮約形修飾節間の齟齬を見ると、何を目的にどこまでの分類を行うか、解析器にどこまでの分類を求めるかという議論が必要

であることも示唆される。

節認定にせよ、意味分類にせよ、浅い分類であれば安定しやすいが、深くなるほど分類が細かくなり、語の性質、意味、機能などの解釈による齟齬が生じやすくなる。そこで、実装にはどこまでの情報付与が求められているかということも勘案しながら、作業の基準策定をしていくことが求められるであろう。

一旦は許容範囲を広く認め、取りこぼしなく節情報を付与することで、次の段階の意味処理において、目的に応じた分類を行うなど、利用範囲も広がると考える。

文法上の正確さと、解析器への実装という問題の最適の妥協点を探るために、現行の作業による作業者間の齟齬を中心とした問題点を述べたが、これらを基にした作業基準を策定し、作業仕様書を作成することができれば、作業の効率化も図れ、データの信頼性も向上すると考える。また、一定の基準に基づいた作業における齟齬が、節の意味機能に関する新たな発見とコーパス言語学の更なる発展に繋がることも期待できる。

謝　辞

本研究は JSPS 科研費(課題番号 : 15K12888, 研究代表者 : 浅原正幸)の助成を受けている。

文　献

浅原正幸・小西光・田中弥生・加藤祥 (2015) 「品詞列・係り受け部分木に基づくラベリングツールの設計と実装―節境界ラベリングを例に―」 第 8 回コーパス日本語学ワークショップ pp.83-92.

池原悟 (2009) 『非線形言語モデルによる自然言語処理』 岩波書店

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕 (2011) 「『現代日本語書き言葉均衡コーパス』 形態論情報規定集第 4 版（上)」 特定領域研究「日本語コーパス」平成 22 年度研究成果報告書

寺村秀夫 (1981) 『日本語の文法（下)』日本語教育指導参考書(5) 国立国語研究所

益岡隆志 (1997) 『複文』くろしお出版

益岡隆志・田窪行則 (1992) 『基礎日本語文法』くろしお出版

丸山岳彦・佐藤理史・夏目和子 (2016) 「現代日本語における節の分類体系について」言語処理学会第 22 回年次大会発表論文集 pp.1113-1116

南不二男 (1974) 『現代日本語の構造』大修館書店