MCN コーパスにおける条件節のアノテーションに向けて

宇佐美文梨† 大町麗奈† 松本晴香† 吉久怜子† 田中リベカ† 戸次大介†‡§
†お茶の水女子大学 ‡科学技術振興機構 (CREST) §産業技術総合研究所
{g1320505, g1320507, g1320533, g1320540, tanaka.ribeka, bekki}@is.ocha.ac.jp

1 はじめに

自然言語の意味処理において、自然言語で記述されたテキストが、述語項構造が表す情報を越えた意味を表す場合、機能表現の意味を捉える必要がある。そのためには、機能表現が意味しうる内容とはそもそも何か、各機能表現が持ちうる「異なる意味」にはどのようなものあるか、そしてテキストにおける機能表現の各出現がどの意味を持つのか、といった問題に答えていくことが要求される。

MCN コーパス [2] は日本語の機能表現にアノテーションを付与した言語コーパスであり、機能語のための言語リソースを提供することを目的としている。MCN コーパスのガイドラインでは、様相表現・条件表現・否定表現といった情報の確実性に影響する表現を対象としている。これらの各表現が持ちうる様々な用法・意味をリストアップし、テキストにおける該当表現がどの用法で使用され、どのような意味を表すのかを、「言語学的テスト」を用いることで特定するという方針を採っている。ここでの言語学的テストとは、文に含まれる表現の容認性や適切性を判定するものであり、これによって必ずしも言語学のバックグラウンドのないアノテータにも、「異なる意味」の区別ができることを意図している。しかし一方で、そのような言語学的テストを導入したガイドラインを各機能表現に対してどのくらいのコストで作成できるかという点については課題が残っている。

本研究では、情報の確実性に影響する機能表現のうち条件表現に着目し、「たら」「れば」「なら（ば）」「と」の 4 つの表現についてガイドラインを作成し、アノテーションを実施した。

2 対象とする表現について

本研究では、機能表現として条件節を導入する「たら」「れば」「なら（ば）」「と」を対象とする。これらの表現のもつ用法の多様性については、日本語学の研究において知見が蓄積されている [1, 3]。ただし、本研究でガイドライン作成とアノテーション作業の対象とした事例は、日本語学の条件文研究において対象とされている範囲と必ずしも一致しない。

まず、一般に「と」は条件節を導入する表現とはみなされず、他の 3 表現とは区別されるが、ここでは「と」のもつ仮定的な用法に着目し、条件節を導入する表現として対象表現に含めた。また現代日本語書き言葉均衡コーパス (以下 BCCWJ)[1] において以下の品詞情報をもつ表現全体を対象としており、個々の出現中には、条件文研究の対象とされている用法とは必ずしも対応しないものも含まれる。

タラ：助動詞タのうち、表層形が「たら」の要素
レバ：述語の仮定形に接続し、表層形が「ば」の要素
ナラ：助動詞ダのうち、表層形が「なら」の要素
ト：接続助詞のうち、表層形が「と」の要素

アノテーション対象表現をこのように定めることで、コーパス上の出現について、見た目上互いに区別しづらいものをすべて分析対象に含めている。

3 条件節のアノテーション

本節では、各表現について作成したガイドラインの詳細について述べる。各ガイドラインについて、BCCWJ の書籍ドメインのテキストを用いて、実際にアノテーションを行った。それぞれアノテーション件数は 100 件、アノテータは学部生 2 名で行い、単純一致率 [2] を測定した。

3.1 タラ節

一般に、タラは条件節を導く表現として知られている。以下の例文はどちらも「たら」を含む例であるが、「もし」が挿入できるか否かの違いが生じる。

[1] http://www.ninjal.ac.jp/corpus_center/bccwj
[2] 単純一致率 = カテゴリ一致数 / 全体数

(1) もし 新聞だったら ここにあるよ。

(2) *もし 本を読んでい たら 電話が鳴った。

本研究では有田 (2007)[1] に従い、前件が真とは断定できない場合に「もし」が挿入可、そうでないならば挿入不可であるとしている。

以下の例は「もし」が挿入不可である（つまり、仮定を表さない）ような例である。

(3) 傘をもっていったら いい。

(4) 帰ったら うがいをしないといけません。

(5) 本を読んでい たら 電話が鳴った。

(6) その話を聞いたときの母の顔といったら！

(7) 彼、甘いものとなったら 目がないんです。

(3) と (4) は前件が成立した上での話者の判断や考えを後件で表しているが、(4) は時系列を示唆しているのに対し、(3) はそうではない。(5) と (6) は前件が発話時に対して過去に生じた事象 (すなわち真である命題) を表すが、やはり時系列の示唆の有無が異なる。(7) は話者の判断も時系列の示唆も含まず、前件が真の命題でない。

このようにタラは様々な用法を持ち、それぞれに特徴がある。たとえば (3) は遠回しな提案を、(7) は後件における話題設定を表している。

これらの現象をもとに、本研究では、「たら」の用法分類を作成した。今回作成したガイドラインの一部抜粋を表 1(次頁) に示す。

アノテーションの単純一致率は 0.70 であった。「もし」挿入の容認可否が異なったのは全体の 15%であり、一致しなかったアノテーションのうち半分を占めた。

「もし」挿入テストは、挿入位置によって容認の可否が異なる場合がある。以下の例では「もしお寿司が…」だと容認できないが、「もし気に入ったら…」とすると、容認できる場合がある。

(8) (説明文) お寿司が流れてくるので、気に入ったら 取って食べてください。

「もし気に入ったら…」の場合、容認可否が分かれる。お寿司を食べなくても良い状況なら「もし」挿入が容認できるが、お寿司を食べることが決まっているときは必ず気に入ったものを見つけなくてはいけないので「もし」挿入不可である。

このように、単純一致率では一定の成果を得たものの、テストの設計には依然課題が残っている。

3.2 レバ節

ガイドライン作成に当たって、最初に着目したのは因果関係の有無である。

[レバ 1] この問題を解決すれば、後の仕事が楽になる。

[レバ 2] 雨が降るようであれば、傘を持っていく。

[レバ 1] では、前件によって後件となることの必然性に焦点が当たっているのに対して、[レバ 2] の方では前件によって後件がどう決まるかという対応関係に焦点が当たっている。この二つの用法をアノテーション時に特定するため、レバを「〜ことで」に置換するというテストを採用した。

(9) この問題を解決する ことで、後の仕事が楽になる。

ただし、このテストは後件に動詞が含まれない場合には使用できない。

(10) 一ミリでも手元が狂えば即死だ。

この文は、因果関係があるにもかかわらず「〜ことで」置換をすると文としては不自然となる。このため「〜ことで」置換のほかに、動詞が含まれない文については後件に「その結果は」を挿入するという手法を導入した。

さらに [レバ 2] の特殊な例として、[レバ 3][レバ 4] の分類を設けた。

[レバ 3] 国民は国に税を納めなければならない。

[レバ 4] 刻限までに届ければいい。

[レバ 3] は「〜なければならない」のような形を取り、義務や強い意志を示している。[レバ 4] は「刻限までに届ける」という達成すべき最低限の条件を提示している。両者とも前件に示した条件を無条件に否定または肯定しているという構成であるが、単純な条件という意味合いからは逸脱しているため、[レバ 2] の特殊な例として分類される。この二つは表現が限られているため、ガイドラインでは表層の形式で判別するという基準を提示している。ただし、「〜ればいい」に関しては [レバ 4] のほかに以下の用法も見られる。

[レバ 5] そこのスイッチを押せばいいよ。/押せば？

この場合、前件で何らかの案を上げ、それを相手に提示するという意味合いになる。この二つの用法の判別方法として、とりたて助詞である「さえ」の挿入によるテストを採用した。[レバ 4] の場合では、挿入すると「刻限までに届け さえ すればいい」となり文の意味は変わらないが、[レバ 5] の場合だと「そこのスイッチを押し さえ すればいいよ」となり、「さえ」によって提案の意味が薄れている。

	例文	特徴	もし挿入	テスト
タラ 1	こんな暗いところで本を読んでいたら目が悪くなりますよ。	前件が会話時点より過去もしくは同時刻で、既に起こっている事象である。後件が前件を受けた話者の判断であり、その始点は前件の始点よりも未来である。	不可	前件「このように」挿入可能かつ後件「かもしれない」挿入可能であり、後件「そのあと」もしくは「すぐに」、かつ「と思う」が挿入可能のときこのカテゴリである。
タラ 8	甘いものとなったら目がないんです。	前件が真偽を表すものでなく、時系列がなく、話者の判断でもない。話の導入などに用いられる。	不可	前件「このように」挿入不可または後件「かもしれない」挿入不可であり、後件「そのあと」かつ「すぐに」挿入不可、「と思う」が挿入不可のときこのカテゴリである。
タラ 9	新聞だったらここにあるよ。	前件が話者の仮定を示し、後件が真である。	可	(後件の否定+たら+前件の否定) が偽であり、成り立たない。

表 1: タラのガイドライン

本ガイドラインでの一致率は 0.73 となった。結果が一致しなかったケースでは、以下のような問題が生じた。

まず [レバ 1] の置換によるテストが不十分性が挙げられる。本アノテーションでは「〜ことで」置換によって因果関係の有無を調べようとした。しかしアノテーションの結果、[レバ 1] であるものがテストによって [レバ 2] に分類されるケースが 38 件中 10 件あった。[レバ 2] に対応するものが [レバ 1] に分類されるケースは 26 件中 1 件のみだったため、「〜ことで」置換に関して [レバ 1][レバ 2] で何らかの傾向の違いがあることが示唆されるが、テストとして十分に機能しなかったことが原因と考えられる。同じような問題として、[レバ 4] と [レバ 5] の「さえ」の挿入の可否の不一致が挙げられる。

もう一つの課題として、「どうすれ [[ば]] よいか」等に出現する [レバ] の処理については、今回は疑問詞や終助詞「か」を含む形式のままそのままテストケースに当てはめるという手法を取った。しかし、疑問形をとっていることによってテストが意図した通りに機能しない可能性は十分にある。この対策として、

(11)　この部分をどう解釈すればよいかわからない。

(12)　この部分をこう解釈すればよい。

のように、疑問形を解消した上でのテストの適応を検討中である。

3.3　ナラ節

ナラが前件にとりうる品詞は名詞、形容詞、節と様々である。ここでは、前件が名詞であるときのナラの用法は限られていることに注目して、大きくわけて名詞・名詞以外で分類し、更に各々に見られるナラの用法に細分化していく。まず前件が名詞の場合、主に以下のような二つのケースに分けられる。

[ナラ 1] あいつなら、今元気だよ。

[ナラ 2] 甘いものなら、何でもいい。

[ナラ 1] は、ナラによって「あいつ」が話題として提示・強調されているが、実質的にこれは後件の主語にあたり「あいつは今元気だよ」と言い換えることができる。一方で [ナラ 2] は、名詞とナラの間に「という条件」や「が成り立つ」を挿入しても意味が通るような文である。ところが以下のようなどちらにも分類できる文が存在する。

(13)　5+1 なら、子供が計算できる。

「5+1 は子供が計算できる」と言い換えると [ナラ 1] に分類されると考えられるが、一方、「(与える問題が) 5+1 という条件なら、子供が計算できる。」と言い換えることもでき、[ナラ 2] に分類されるとも考えられる。このような場合、聞き手の解釈が文脈によって異なり、一意に用法を定めるのは困難である。

続いて、前件を文とするナラの用法について見ていく。ナラが導く文には、[ナラ 3] のような慣用表現や、[ナラ 4] のような並列・反復の用法がある。

[ナラ 3] なぜなら、君が悪いから。

[ナラ 4] 花子は何でもできる。勉強ができるなら、スポーツもできる。

これらの用法を除き、前件が文である場合、前件が広義の仮定・前提を表し、後件でその評価・推測をするというものが多い。典型的な仮定の例を以下に示す。

[ナラ 5] 明日忙しいなら、ライブにはいけないだろう。

前件で「明日が忙しい」という状況を前提とし、後件でその結果起こりうる事態を述べている。このようなナラの用法を本ガイドライン中では純粋仮定と呼ぶ。この純粋仮定を典型例ととらえ、それ以外の周辺的な用法については前件と後件の真偽的なつながりや順序関係に基づいて分類した。

[ナラ 6] なるなら、エンジニアがいい。

[ナラ 6] は「どうせ」「どっちみち」が挿入可能であるケースであり、前件に述べられる状況の他に選択の余地がない場合を仮定する用法である。たとえば [ナラ 6] の文では、「何にもならないという選択ができない状況」を仮定して、なりたいものの対象を後件で述べている。他の用法については割愛する。

本ガイドラインに基づくアノテーションの一致率は 0.79 であった。結果の不一致が生じるのは主に以下のような場合であった。

(14)　花子を叱るなら、太郎も叱る。

これは前件と後件が等価であるので、並列・反復の用法だと思われる。しかし、アノテータの回答は「どうせ」の挿入が可能であることに基づき [ナラ 6] と同じ用法という判断であった。これは、話し手が「花子も太郎も平等に叱る」という意味合いで用いたのか、「どうせ花子を叱るなら、ついでに」という意味合いなのか、先述したような理由で複数の解釈が可能なためである。

3.4 ト節

条件を意味するトの例として、まず以下の 4 つの例文を考える。

[ト 1] スイッチを押すと電気がつく。
[ト 2] よく見るとここに傷がある。
[ト 3] 甘いものとなると目がないんです。
[ト 4] 何であろうと大丈夫よ。

これらは時間の経過に着目すると 2 つのグループに分類される。[ト 1] [ト 2] において後件の出来事は前件の後に生じる。しかし [ト 3] [ト 4] でははっきりとした時間の経過を確認できない。

一方、因果関係の観点からも分類が可能である。[ト 1] [ト 2] では、[ト 1] のみ因果関係を確認できる。[ト 1] で電気がついたのはスイッチを押したからである。しかし、[ト 2] で傷があるのはよく見たからではない。よく見なくてももともと傷はあったと考えられるからである。また、[ト 3] [ト 4] では [ト 3] のみ明確な因果関係を確認できる。目がないのは甘いものだからである。

さらに、前件が否定形の場合については別の用法だと考えて独立に分類する。

[ト 5] もう寝ないと怒られる。
[ト 1'] こたつで寝ると怒られる。

[ト 5] は [ト 1'] とよく似ているが、寝なければならないという使命感が加わっていると思われる。ただしこれはトが否定形に接続された場合に生じる新たな役割なのか、条件内の否定形の役割なのかを今後検討する必要がある。

本ガイドラインをもとに行ったアノテーション作業の結果、アノテータ間の一致率は 0.85 となった。しかしアノテーション結果より本ガイドラインは以下の点でまだ不十分であると思われた。

因果関係、時間経過で分類する際、明確な判断が困難なケースがみられた。

(15) クラスメートに、出身地はどこかと聞かれて、「大阪」(大阪の中国語) と答えると、「オーサカ」と非常に納得され、その納得ぶりに納得がいかない私である。

(16) 顔色をうかがっていると彼は私をにらみつけ...

(15) が因果関係について明確な判断が困難である例である。大阪と答えたから納得されたと考えると [ト 1] に分類されるが、大阪以外のどこを答えても適当に納得されただろうと考えると [ト 2] に分類される。(16) は時間経過がはっきりしない例である。顔色をうかがい終わった後ににらみつけられたと考えると [ト 3] に、うかがっているのとにらみつけられたのが同時並行で起こっていると考えると [ト 4] に分類される。これらは、文章から推察する以外ないが、話し手の意図は一意に決まるにもかかわらず、情報の不足から解釈が定まらないものである。そのため、両方の解釈が必要なものとして新たなカテゴリーをつくるべきかは熟考しなければならない。

4 おわりに

本稿では条件節を導入する表現「たら」「れば」「ならば」「と」を対象としたガイドライン作成作業について報告する。またアノテーションの試行結果を示し、多義性の把握・アノテーション時の曖昧性解消に関して、現段階での問題点を考察した。現状では問題点が多く残されているが、今後改良を加えていく予定である。

参考文献

[1] 有田節子. 日本語条件文と時制節性. くろしお出版, 2007.

[2] 川添愛, 齊藤学, 片岡喜代子, 崔栄殊, 戸次大介. 言語情報の確実性に影響する表現およびそのスコープのためのアノテーションガイドライン ver. 2.4. Technical report, Technical Report of Department of Information Science, Ochanomizu University, OCHA-IS 10-4, 2011.

[3] 蓮沼昭子, 有田節子, 前田直子. 日本語文法セルフマスターシリーズ 7 条件表現. くろしお出版, 2001.