# MCNコーパス：

# 言語学的テストに基づくモダリティ・アノテーションの理論と実証

田中リベカ（お茶の水女子大学理学部）
川添愛（国立情報学研究所）
戸次大介（お茶の水女子大学大学院人間文化創成科学研究科／国立情報学研究所）

# MCN Corpus:

# Methodology on the Modality Annotation based on Linguistic Tests

Ribeka Tanaka (Faculty of Science, Ochanomizu University)
Ai Kawazoe (National Institute of Informatics)
Daisuke Bekki (Graduate School of Humanities and Sciences, Ochanomizu University
/ National Institute of Informatics)

## １ はじめに

自然言語処理においては近年、文の表層的な形式から得られる浅い情報にとどまらず、その興味の対象が深い意味に移行してきている。言語の意味に関しては、母語話者間で理解が共有されており、その構造には何らかの法則性があることが示唆される。このため、そのような話者間に共通の意味の認識が反映されているデータの需要が今後ますます高まることが予想される。

しかし、意味を対象としたリソースを作成する方法論は確立していない。意味アノテーションにおいては、言語表現の多義性解消や用法の特定などにおいてアノテータの一貫した判断が求められるため、ガイドラインにおいて判断基準を明確に提示する必要がある。ガイドラインがこのような判断基準や手段の提示を欠いている場合や、提示に失敗している場合、アノテータ間での判断の不一致を引き起こし、本来共有しているはずの意味への認識が正しく反映されない恐れがある。このような判断の不一致を避けるために、しばしばアノテーション作業を複数人でなく一人で行うという方針がとられることがある。しかしこのような方針においては、作成されたリソースが一人のアノテータの主観によるものではないとは言い切れない。またその判断の正当性を第三者が検査するのも困難である場合が多い。

そこで田中ら(2012)[2]では、言語学的テストを用いた意味アノテーションの方法論を提案した。ここで言う「言語学的テスト」とは、理論言語学の理論構築および検証に用いられるテストで、文や文の一部の容認性や適切性を判定するものである。言語学的テストは、容認性や適切性を左右する条件を特定するために、検証したい部分以外の条件をほぼ同じにした二つ以上の文からなる群として提示されることが多い。

本論文では以下、第二節で田中ら(2012)[2]における意味アノテーションの方法論を紹介する。特に、意味アノテーションの作業を一種の「分類タスク」と考えた場合に、適切な言語学的テストの有無が分類を決定する際の判断の一貫性を左右することを実例を交えて論じる。第三節では、アノテーションに有効な言語学的テストの設計についての一般的な

方法論を提案する。第四節でその適用例を見た後、第五節では言語学的テストにおけるアノテータの判断の一致度に関して筆者らが行った調査について述べ、実際にガイドラインを作成する際に注意すべき技術的な問題について考察する。そして最後に、本手法のもつ言語学的な意義について論じる。

なお、本論文において「意味アノテーション」の具体例として用いるのは、「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」（川添ら（2011）[1]）に基づくアノテーション作業である。このガイドラインは、様相表現・条件表現・否定表現など、言語情報の確実性に影響する表現とそのスコープにアノテーションを付与し、機械による確実性判断の基盤となるコーパスを構築するために作成されたものである。

## 2 コーパスアノテーションにおける言語学的テスト

意味アノテーションに限らず、コーパスに対するアノテーション作業の多くは、スキーマの設定者があらかじめ設定したラベルをアノテータがテキストの一部に対して付与していくものである。これはある意味、実際のテキスト中の表現を用意されたカテゴリに分類していく作業であるととらえることもできる。アノテーションガイドラインにはそのような分類を行う際の判断基準についての指示が多かれ少なかれ含まれるわけであるが、その細やかさ・適切さのレベルは様々である。

たとえば、テキストに現れる「(と）いう」という表現のうち、他人による認識あるいは主張を表すものにのみアノテーションをしたいとする。このとき、ガイドラインには以下のようにアノテーション対象のカテゴリの説明のみを提示し、テキストに現れる個々の「(と）いう」の出現がそのカテゴリに属するものかどうかの判断はアノテータに委ねる、という方法が一般的である。

**他人の認識【(と）いう】**
分類：他人の認識（他人の報告する事柄や、命題の真偽に関する他人の判断を表す表現）

しかしこのような方法は、どのような基準で「他人の報告する事柄」と見なせば良いのかということや、そもそもどのような表現を「命題」とみなすかなどについても明確には定められていないという点で、表現を分類する基準や詳しい知識をアノテータが持っていることを前提としていると考えられる。このため、母語話者であっても専門的な知識を有していないアノテータには判定が困難である。

これよりもう少し細やかな指示としては、以下のように例文を示した上で、実際の表現がその例文と同じ用法で使われているかどうかを判断させるような方法がある。

**他人の認識【(と）いう】**
分類：他人の認識（他人の報告する事柄や、命題の真偽に関する他人の判断を表す表現）
例：今冬のインフルエンザの流行は全国的に遅れているという。

上のように例文を提示した場合、ほぼ同じ構造の文に関してはアノテータは専門的な知識を要さずに判定することができる。しかし、表現の形式が変わったり、少しでもニュアンスが異なるような表現に遭遇したりすると途端に判断が困難になる。上の例は「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」にある記述を内容を変えずに再編集したものであるが、筆者らが実際にアノテーション作業を行ったところ、アノテーションの可否が判断出来ない、又はアノテーションの結果が一致しない例に直面した。そのような例の一部を以下に示す。

1. 呼吸困難に陥る可能性があるという。
2. 「彼を絶対に許さない」と言う。
3. 太郎が責任をとるべきという人はどうかしている。
4. 太郎は結婚したという話だ。
5. インスリンというホルモン。
6. 車がガタガタという。
7. お前という人間が信じられなくなった。
8. サプリというサプリは試した。

1.については、アノテーションガイドラインに記載されている例文と非常によく似ていることから、「他人の認識を表す『(と)いう』」としてアノテーションできると想像がつく。他方、2.-8.については決定的な判断基準がない。例文と似ているかどうかをどのような点に着目して判断するのかについて明らかにされていないため、判断基準がアノテータの主観的な理解に委ねられてしまうのである。筆者らの行ったアノテーション作業においては、特に2.3.4.の表現についてアノテータ間でアノテーション可否の判断が大きく分かれた。特に困難であったのは、3.4.の判断である。これらはどちらも名詞句を修飾していて見た目には同じ構造をしており、一見するとガイドラインの例文とは異なる形をとっている。しかしガイドラインの設計者は、3.は他人の認識としてアノテーション可能であり、4.はアノテーション不可能であることを意図している。このように、例文ベースの指示では例文との類似度の判断は困難であり、アノテータ間でも判断が一致しにくい。

そこで筆者らは、田中ら(2012)[2]において言語学的テストを導入し、ガイドラインを以下のように改善した。

**他人の認識【(と）いう】**
分類：他人の認識（他人の報告する事柄や、命題の真偽に関する他人の判断を表す表現）
例：今冬のインフルエンザの流行は全国的に遅れているという。
テスト１：「(と）述べる（述べられる)」に置き換えても意味が変化しない。
テスト２：名詞句を修飾する場合「～との」に置き換え不可

テスト１，２はともに表現の「置き換え」に基づく言語学的テストである。これらのテストを先に挙げた例に適用すると、結果は以下のようになる。

1. （テスト１）呼吸困難に陥る可能性があると述べる（述べられる)。/(テスト２：適用不可)
2. （テスト１）「彼を絶対に許さない」と述べる。/(テスト２：適用不可)
3. （テスト１）太郎が責任をとるべきと述べる人はどうかしている。/(テスト２)*太郎が責任をとるべきとの人はどうかしている。
4. （テスト１)*太郎は結婚したと述べる（述べられる）話だ。/(テスト２）太郎が結婚し

たとの話だ。

5. (テスト１)*インスリンと述べる（述べられる）ホルモン。/(テスト２)*インスリンとのホルモン。
6. (テスト１)*車がガタガタと述べる（述べられる)。/(テスト２：適用不可)
7. (テスト１)*お前と述べる（述べられる）人間が信じられなくなった。/(テスト２)*お前との人間が信じられなくなった。
8. (テスト１)*サプリと述べる（述べられる）サプリは試した。/(テスト２)*サプリとのサプリは試した。

上のように、実際に該当箇所の表現を置き換えた文について考え、別の表現に置き換え可能かどうかを判定させる言語学的テストを導入すると、このテストについて複数のアノテータのYES/NO判定は一致しやすく、結果としてアノテーション可否の判断の一致度も向上した。実際に、テスト１、テスト２のいずれかでNOの判定がでた 4.-8. の表現についてはアノテーション不可能であると判断できるようになった。また、言語学的テスト導入前には困難であった 3. と 4. の区別も、それぞれのテストで異なる結果が出たことで明確になった。

また、筆者らが上記のテストを用いて行った実際のアノテーション作業において、注目すべき現象がみられた。「置き換えが可能」という判断よりも「置き換えが不可能」という判断の方がアノテータ間で一致しやすかったのである。このことから、ガイドラインにおいては、「置き換え可能ならばアノテーション対象表現である」という指示よりも「置き換え不可能ならばアノテーション対象表現ではない」という指示を用いる方が、一貫性のあるアノテーション結果を得るのに有効であるとの結論に至った。

## ３ 言語学的テストの設計方針

前節で述べた作業は筆者らが行った具体的なアノテーション作業であるが、アノテーションガイドラインを設計するに当たっての重要な示唆を含んでいると考えられる。本節ではこれらの考察を元に、理想的なガイドラインを作成するにはどのようなテストを設計すべきかを論じる。

ここで、ある表現Ｅに対して、その用法として$A_1, A_2, \cdots, A_n$のｎ個の分類を考える。ここで、実際のテキストに表現Ｅの出現Ｅ’があったとき、このＥ’がｎ個の分類中どこに属するか（つまりＥ’が表現Ｅのどの用法になっているか）を判定するテストを考える。

表現の出現Ｅ’が特定の分類$A_i$に属するかどうかの判断基準となるテストを、ここでは「個別テスト」と呼ぶ。また分類$A_1, A_2, \cdots, A_n$の個別テストを集めたもの全体を「テストセット」と呼ぶこととし、以下において個別テストとテストセットの２段階における性質を考える。

### ３．１ 個別テストの構築

まず個別テストの内容については、別の表現に実際に置き換えた上で「意味が変わるかどうか」「文自体が意味不明になるかどうか」をYES/NOで判定する形式や、「活用しているか」「コト節を受けているか」など表層上明らかな性質を条件に設定するのが望ましい。前節で見たように、「この表現と同じ用法か」という形式をとった場合、判断は個々人の独自の解釈に左右される。しかし置き換えた上で意味が変わるかどうかの判定は言語直感に訴えるものであり、母語話者を対象とする限りは専門的な知識を要求しないと考えられる。

無論、テストを作成する際にはどのような語に置き換えるよう指示するかについても慎重にならねばならない。たとえば慣用的に使われている表現を用いると、必ずしも本質的ではない点で混乱を生む可能性がある。細心の注意を払ってもなお、設計者が意図しない解釈がされた場合には個別テストを修正する必要があるが、これに関しては後述する。

前節で述べたように、「置き換えが可能」という判断よりも「置き換えが不可能」という判断の方がアノテータ間で一致がしやすいという現象が見られた。この性質をテストに反映するために可能な個別テストのパターンは次の２通りが考えられる。

1)　表現Ｅ’を表現$e_i$に置き換え不可能ならば、表現Ｅ’は分類$A_i$に属さない。
2)　表現Ｅ’を表現$e_i$に置き換え不可能ならば、表現Ｅ’は分類$A_i$に属する。

本手法ではこのうち、1）の形式のテストを採用する。このような形式を満たすテストを作成するためには、それぞれの分類$A_i$について、

3)　表現Ｅ’が分類$A_i$に属するならば、表現Ｅ’を表現$e_i$に置き換え可能である。

といえるような表現$e_i$を見つけ、3）の対偶をとれば良い。これにより、1）の形式のテストが得られる。仮に 2) の形式のテストを採用すると、テストを作成するためには、上と同様の方針に基づいてそれぞれの分類$A_i$について

4)　表現Ｅ’が分類$A_i$に属さないならば、表現Ｅ’を表現$e_i$に置き換え可能である。

といえるような表現$e_i$を探すことになる。しかし、$A_i$以外の分類に属するすべての表現と置き換え可能であるような表現$e_i$を探すことは、対象が広すぎるため非常に困難である。これに対して 3) を満たすような表現$e_i$については、ある特定の分類$A_i$についてその用法がもつ性質に着目して探せば良いため、対象が絞られるだけでなく言語学においてなされた用法の分析の成果を有効に活用することが可能である。以上を踏まえて、本手法では 1) の形式を個別テストの基本的な型とし、これを「ネガティブテスト」と呼ぶ。適切なネガティブテストには、その表現に置き換えられなかったら特定の分類に属さないのだと言い切れるような強い条件を用い、判断に迷うような紛らわしい例を避けることが重要である。

なお、「置き換え不可能である」という判断がアノテータ間で一致することの背景については、次のように考えられる。「置き換え可能」であるという判断は、ニュアンスの変化を許すか、規範的な日本語のみを許すかなど、どの程度なら置き換え可能とみなすかについて個人差がある。また、自分が発話する表現と理解できる表現にも差があり、日常生活ではそれらの程度の異なる言語事象が混在しているため、「置き換え可能か」と問われた場合にはどこを基準にするべきかが明確に定まらない。しかし一方、「置き換え不可能」である表現はコミュニケーションにおいて使用できない（正しく意味をなさない）表現である。このため、本来置き換え不可能な表現を強引に置き換えた文を提示されると、母語話者は言語直感によってその文が正しくないこと、すなわち「置き換え不可能」であることを判定することが可能である。逆にいうと、テストを作成する際には母語話者なら誰もが「置き換え不可能」と判断するであろう表現を意図的に選ぶことが重要である。

### ３．２ テストセットの構成

上のような方針で個別テストを設計することを前提とした上で、テストセットとしてどのような性質を満たすべきなのかを考える。

一つの個別テストにおいては、表現$e_i$に置き換え不可だった場合は分類$A_i$に属さないと言えるが、置き換え不可だと言い切れないと判断された場合はその分類に属すとも属さないとも判定されない。これは、表現$e_i$に置き換え不可能であるということが、分類$A_i$に属さないことの十分条件ではあるが、必要条件にはなっていないことによる。つまりネガティブテストは単独では分類を決定しない。そこで、このような個別のネガティブテストを全ての分類について作成する。個々の個別テストでは「置き換え不可能なので、この分類には属さない」又は「置き換え不可能ではないので、この分類に属さないとは言い切れない」の２パターンの判定がなされうる。「この分類には属さない」の判定が出た場合、表現の出現Ｅ’がその分類に属す可能性はなくなるので分類先の候補から除くことができる。また、「この分類に属さないとは言い切れない」という判定が出た場合には、その時点ではこの分類に属すか否かは判定できないので分類先の候補として残す。ここで、網羅的な分類と適切なネガティブテストから構成された理想的なテストセットを作成すると、全ての分類の個別テストを試した結果、分類先の候補は１つしか残らないはずである。この残った１つの候補は、個別テストにおいて「置き換え不可能ではないので、この分類に属さないとは言い切れない」と判定された唯一の分類であり、すなわち表現の出現Ｅ’の正しい分類先であるといえる。このように網羅的な分類と各々の分類についての適切な個別テストからなるテストセットを構成し、消去法により表現の出現Ｅ’の分類先を特定する。

テストセットを構成する際には、任意の２つの分類$A_i$, $A_j$の個別テストを比較したときに一方が他方を含まないようにする必要がある。仮に$A_i$の個別テストが全て$A_j$の個別テストに含まれていると、「分類$A_j$に属さないとは言い切れない」と判定された場合には必ず分類$A_i$についても同様の判定がなされ、必然的に分類先が１つに特定されない。無論、複数の個別テストが同じ内容であることも同様の理由により避けねばならない。ガイドラインの設計者は、異なる用法として分類を分けるならばその根拠となる性質をテストに反映しなくてはならない。

テストセットにおいて、個々の個別テストは独立に振舞う。決定木のように、あるテストの判定が次のテストを適用する条件になっているわけではなく、どの個別テストから適用するかの順によらず分類先は決定する。そのため、一部徐々に条件を狭めていくような状況を作りたいときには、個別テストで意図的に実現する必要がある。例えば、分類$A_i$に分類される出現は表現$e_a$, $e_b$の両方に置き換え可能なのに対し、分類$A_j$に分類される出現は表現$e_a$には置き換え可能だが表現$e_b$には置き換え不可能であるという状況を考える。このとき、２つの分類の個別テストはそれぞれ次のように作成される。

【分類$A_i$の個別テスト】

_ 表現Ｅ’を表現$e_a$に置き換え不可能ならば、この分類ではない。

_ 表現Ｅ’を表現$e_b$に置き換え不可能ならば、この分類ではない。

【分類$A_j$の個別テスト】

_ 表現Ｅ’を表現$e_a$に置き換え不可能ならば、この分類ではない。

_ 表現Ｅ’を表現$e_b$に置き換え不可能ならば、この分類である(可能性がある)。

ここで、下線部のように「この分類である」という表現を意図的に用いることにより、$e_a$に置き換え可能であるという共通の条件のもと、$e_b$に置き換え可能か否かでどちらの

分類に入るかが決定するような形式を実現することができる。(なお下線部のような表現は、先述したように条件が見つけにくいため個別テストの基本的な型としてはふさわしくないと考えられるが、このように十分対象が狭まったと考えられる状況下では避ける理由はない。)また、「可能性がある」としたのは個別テストの独立性を保つためである。分類$A_j$の個別テストとしては、「表現E'を表現$e_b$に置き換え不可能」という条件のみからは、まだ「この分類である」と決定できないことは明らかである。一方、「$e_a$に置き換え不可能」というテストとは切り離して考えたいので、このような表現を用いるのが適切だと考えられる。「可能性がある」としておくことで、他のケースと同様、消去法の後に分類が決定することになる。

このような形式で個別テストを作成することは一見冗長であり、個別テストを互いに独立なものとしたことの弊害にも見える。しかし以下で述べるようにこの方法論においてはテストの修正が重要であるため、個別テストが互いに独立であることによって一部の個別テストの変更がそれ以外の箇所には影響せず、修正を容易にしているということは大きなメリットである。

### ３．３ テストの修正

本手法では、網羅的な分類と適切な個別テストでテストセットが構成されると、理想的には消去法で分類先が１つ特定される。しかし、ガイドラインの設計者が最初から用法を網羅的に全て把握しているわけではなく、また設計者にとって明らかな個別テストであっても、実際に使用してみるとアノテータにとっては紛らわしいこともある。このため分類先が１つに特定できないということが起きた場合には、アノテータは分類先が特定出来なかったことをガイドライン設計者に報告し、設計者はテストセットの修正を行う必要がある。

**消去法の結果、候補が複数残った場合**

消去法を行った結果、分類先の候補が１つだけ残ることが理想であるが、複数残ってしまう場合がありうる。これは本来その表現が分類されるべき分類の個別テスト以外にも、「置き換え可能である」との判定が出てしまった個別テストが存在したことを意味する。置き換える表現の選択が悪かったことが原因であるため、意図せず「置き換え可能である」の判定がされてしまった分類の個別テストの内容のみを修正する。この際、それ以外の分類に対しては「置き換え不可能である」との判断が正しくなされたのは事実であるので、修正をする必要はない。またスキーマの設計者が意図する本来の分類先についても、「置き換え可能である」の判断がされたことは自然な現象であるので修正をする必要はない。

**消去法の結果、候補が１つも残らなかった場合**

全ての個別テストで、「置き換え不可能である」という判定が出てしまったときには候補が１つも残らないことがある。このとき、２つの原因が考えうる。

１つは、本来の分類先の個別テストが適切でなかったため、ガイドラインの設計者が「置き換え可能である」として設定したにも関わらずアノテータが「置き換え不可能である」と判断した場合である。このような場合には、「置き換え可能である」の判定が出るように本来の分類先の個別テストを修正する。

もう１つは、その表現の出現が本質的にどの分類にも属さない用法であった場合である。この場合は分類が網羅的でなかったことが原因であるため、ガイドライン設計者は新しい分類を増やしその個別テストを作成する。

なお他の可能性として、消去法の結果候補が１つに絞られ分類が決定されたが、それが

ガイドライン設計者の意図とは異なる分類である場合がありうる。このような場合には、アノテータによる問題点の報告がなされないためテストの修正はなされない。このことは、複数のアノテータが同じテキストに対してアノテーションをして結果を比較するような過程をとらない限り、避け得ないことである。しかし、これはガイドライン設計者の意図する分類の個別テスト、およびアノテータによる分類先の個別テストの二箇所において設計の誤りがある場合にのみ起こることである。また、それ以外の場合には上に述べたように、網羅的な分類と適切な個別テストが最初から得られていなくても、ある程度適切なテストセットであるならば実際のアノテーション作業を通して問題箇所がピンポイントに浮き彫りになり、理想的なテストセットへと修正していくことが可能である。

## ４ 適用例

前節の考察に基づき、適切なネガティブセット、および表現の意味の網羅的な分類を含むテストセットの例を表１に示す。ここでは、「（と）いう」をその意味に応じて８つのカテゴリに分け、テキストにおけるこの表現の個々の出現例がどのカテゴリに属すかを判定するためのネガティブテストを８つ用意している。

表１：テストセット構築例

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| iu | 1 | いう | 言う | MODAL | hearsay | 話者にとって真偽が未知、確実性判断なし | 動作主が明示されず、「世間一般」「人々」「専門機関あるいは情報ソース」である。「～によると」と共起することが多い。<br>ソースが「世間一般」の場合は、「いう」を「いわれる」に置き換えてもほとんど意味が変化しない。 | ニュースによると、インフルエンザが流行っているという。<br>世界には自分と同じ顔の人間が７人はいるという。<br><br>東京で一番おいしいという焼肉屋へ行ってみた。 | Negative test5:「いわれる」「いわれている」に置き換え不可、あるいは置き換えて意味が変化する（尊敬の意味になる等）の場合はこのカテゴリではない。 | [命題(S)]と　いう |
| iu | 2 | いう | 言う | MODAL | hearsay | 話者にとって真偽が未知、確実性判断なし | 「述べる（述べられる）」「主張する（主張される）」「報告する（報告される）」と同義。命題をとる。 | 花子は、彼を絶対に許さないという。<br>太郎が責任をとるべきという人は、どうかしている。 | Negative test1：「述べる（述べられる）」に置き換えて意味が変化する、あるいは置き換え不可の場合は、このカテゴリではない。<br><br>Negative test3：過去形「（と）いった」に置き換え不可な場合は、このカテゴリではない。 | [動作主(NP)]が［命題(S)]と　いう。 |
| iu | 3 | いう | | ー | | | 「呼ばれる」「名づけられた」と同義。 | 山田さんという人が来た。<br>血糖値を下げるのは、インスリンというホルモンだ。 | Negative test4:「呼ばれる」に置き換えて意味が変化する、あるいは置き換え不可の場合は、このカテゴリではない。<br><br>Negative test8:「という～」を除けない、あるいは除いて意味が変化する場合はこのカテゴリではない。 | [NP]という |
| iu | 4 | いう | | ー | | | 「音や声を発する」の意味。副詞、あるいは「～と」を伴って擬音語あるいはそれに類する表現をとる。 | あの人はぶつぶつ（と）言っている。<br>車がガタガタ（と）いう。 | Negative test3：過去形「（と）いった」に置き換え不可な場合は、このカテゴリではない。<br><br>Negative test6:副詞、あるいは「～と」を伴って擬音語あるいはそれに類する表現をとっていない場合は、このカテゴリではない。<br><br>補助的テスト：このカテゴリに属するものは、「ぶつぶつ」「ガタガタ」のような擬音語的な表現を取り去ると意味をなさない。（十分条件）<br>*あの人は言っている。（比較：あの人はぶつぶつと言っている。）<br>*車がいう。（比較：車がガタガタいう。） | [擬音語]という |
| iu | 5 | いう | | ー | | | 「話」「噂」「見解」などを修飾する。 | 太郎が結婚したという話だ。<br>時期尚早であるという見解を示した。 | Negative test2：「との」に置き換え不可な場合は、このカテゴリではない。<br><br>Negative test3：過去形「（と）いった」に置き換え不可な場合は、このカテゴリではない。 | 「命題(S)]という話/噂/見解・・ |
| iu | 6 | いう | | ー | | | 名詞句を修飾。補足的な意味を付け加える機能を持つ。 | お前という人間が信じられなくなった。<br>長年住んでいるが、東京という町には親しみがわいてこない。<br>今日という日を忘れないようにしよう。<br>それが男というものだ。 | Negative test8：「という～」を除けない、あるいは除いて意味が変化する場合はこのカテゴリではない。 | [NP]という |
| iu | 7 | いう | | ー | | | 前後に同じ形の名詞をとる。 | サブリというサブリは試した。<br>医者という医者には相談したものの、どうにもならなかった。 | Negative test4:「呼ばれる」に置き換えて意味が変化する、あるいは置き換え不可の場合は、このカテゴリではない。<br><br>Negative test7:前後に同じ形の名詞をとっていない場合は、このカテゴリではない。 | [NP]という |
| iu | 8 | いう | | ー | | | | 太郎の友人であるという人物が現れた。 | | 「命題(S)]という[NP] |

分類１から７に関しては、どの２つの分類のテストを比較しても互いに異なるように設計した。ただ一箇所、分類３と分類６では一方が他方を含むようになってしまっているため、この分類６に別の個別テストを増やすなどして修正することが必要である。

また、分類８の個別テストは空欄になっているが、これは分類８がごく最近見つかった分類であることによる。ネガティブテストを適用した結果、どの分類についても「この分類ではない」という判定が出たため、新しく追加されたものである。今後この欄に新しい

個別テストを追加する予定である。

「(と) いう」に関してこれが完全に網羅的な分類であるかは現時点では不明であるが、この分類のみに関して言えば、明確な方針のもと修正を行った後、ネガティブテストの組み合わせによってどのカテゴリに属するかを判断することが可能である。

## ５ 言語学的テストにおけるアノテータの判断一致度の調査

筆者らは本手法に基づき、ネガティブテストにおけるアノテータの判断の一致度を調査した。調査は、大学1年生～大学院生を含む日本人学生 40 人程度を対象に行った。被験者には用法の分類などの詳細については一切明らかにせず、「(と)いう」が出現する文を提示し、「(と)いう」の箇所を別の表現 e に置き換えると意味が変化するか、という形式の問いに対してYES/NO/わからない、で回答するようにした。

この調査結果については現在分析中である。本手法を用いる際、どういう状況ならアノテータの判断が一致したと言って良いのか、また実際に何人以上の回答が意図に合わなかった場合にテストを修正するのかについて、調査結果を元に仮説を構築する予定である。

調査を行う中で、特に言語学的テスト（ネガティブテスト）の提示方法の改良について、いくつかのノウハウが得られた。

まず、「表現Ｅ’を別の表現 e に置き換えると意味が変化する」と問うた場合には被験者による YES/NO の判断の一致度は悪かったが、実際に表現Ｅ’を表現 e に置き換えた文を提示し、「この文は置き換える前の文と比較して意味が変化する」と問う形式に変更すると一致度が良くなった。このことから、一部の被験者は置き換え可能かどうかを判定する際、誤って都合の良い置き換えをしてしまうことが示唆された。

また、文の外見上自明な特徴について問う設問も交ぜたところ、YES/NO で答えず「わからない」という回答をする被験者が多かった。あまりにも自明な事柄を問うと、設問者の意図について必要以上に意識してしまうためであると考えられる。

提示する文が短いと、置き換え後の意味が変化するかどうかの判断が困難となり、判断が一致しにくかった。一方、前後の一文程度を併せて提示すると、提示しないときと比較してアノテータの判断は収束した。実際のアノテーション作業ではどの程度前後の文を読むかはアノテータによって差が生じることが考えられるため、意味の変化を考える際に文脈をどの程度考慮するかについては個人差が生じる恐れがある。

今後ガイドラインの改良を進めていくにあたっても、これらの要因について考慮していく必要があると考えている。

## ６ 言語学的な意義

本手法は、自然言語処理におけるコーパスの意味アノテーションのための手法であるが、この手法を通して得られるテストセットは言語学的にも意義があるものであると考えられる。

通常、表現の用法を全て列挙することは容易ではない。一般に国語辞典には表現の意味の用法が記載されているが、筆者らの行ったアノテーション作業の結果によれば、実際の言語現象においては辞典に記載されている用法よりも更に細かく意味が区別されていることが明らかである。また、「（と）いう」のような複合語については、「と」と「いう」の意味を単純に合わせたものではない独自の意味をもつにも関わらず、辞書の項目にないことも多い。したがって本手法を通して表現の網羅的で重複のない分類が得られること、複

合語のような表現に対してもその意味が分析されることは大きな長所である。

また、本手法においては第三者が分類の正当性を確認できるという点も重要である。用法の分類や表現を区別する根拠はガイドライン設計者によって考えられたものではあるが、その正当性は常にアノテータによって確認される。アノテータの判断が一致し、ガイドライン設計者が意図した分類に分類されるということを以って、正しい分類であることが裏付けられる。逆にアノテータの判断が一致しなかった場合には、設計者は自分の考えた分類や個別テストを再検討することになり、設計者が誤った分析をしていた場合でも修正が行われる。ガイドライン設計者の主観ではなく、母語話者が共通して持つ言語の意味への客観的な認識が、用法の分類と各個別テストに反映されるのである。

## 7 おわりに

以上、意味アノテーションにおける言語学的テストの利用、およびその際の方法論について論じた。

上で述べた方法論に基づき、「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」に含まれるものを中心に、340 種類の様相表現、13 種類の否定表現、48 種類の条件表現の網羅的な分類とネガティブテストの構築を行い、ガイドラインの改良を行う予定である。


## 謝辞

国立情報学研究所人工頭脳プロジェクト「ロボットは東大に入れるか」NLP コア定例ミーティングにて、メンバーの方々に貴重なコメントをいただいた。


## 文献


[1]川添愛、齊藤学、片岡喜代子、崔栄殊、戸次大介(2011)「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 of Department of Information Science, Ochanomizu University, OCHA-IS 10-4.

[2]田中リベカ、小池恵里子、戸次大介、川添愛(2012)「言語学的テストに基づく意味アノテーションのガイドライン設計―確実性判断に関わる表現を中心に」言語処理学会第 18 回年次大会発表論文集,pp.401-404.