E-016

# 多言語話し言葉のコーパスの節単位認定―
# 分割しにくい単位の分類
# Parsing into Clauses in Multilingual Spoken Corpora:
# Classification of Non-prototypical Cases

エフィーモワ・ゾーヤ†

Zoya Efimova

## 1．導入

現在、二つ以上の言語を対象にする多言語話し言葉コーパスが言語学習と類型論などの研究にはよく使われている。その際、音声記録を有効に使用できるように、表記と様々な研究用情報を音声に付与して音声を文字に書き付けることが必要になる。話し言葉コーパスを構文解説や翻訳処理などに使用するためには、文章を構成する基本的な要素である節単位に区切ることが必要である。話し言葉コーパスの節単位認定の問題についてはこれまでにも研究があるが（参考、高梨等 2003, Kibrik and Podlesskaya 2006)、それらの研究で提案された節単位認定のルールは具体的な一つの言語を対象としており、全般的なものではない。それに対して、多言語コーパスの場合は、全般的な分割のルールを作成しないと、異なる言語のデータを適切に比較することができない。

本稿で取り上げる節単位認定の問題は、多言語の話し言葉コーパスに対するものである。先行研究の結果と「多言語の話し言葉語りのコーパス」(参考、以下（３）)のデータに基づき、どのような言語においても節単位認定しにくい単位を明らかにする。

†ロシア国立人文大学／千葉大学（日本学術振興会特別研究員)。本研究は日本学術振興会（P07304）の支援によって行われた。

## 2. 節単位認定の問題

節[1]とは、一つの述語とそれに付随する要素を含む組み合わせと見なされている。以下の例１は節の分割に合わせて改行したものである。

（例１）　今日は休日です。
　　朝ごはんを食べてから
　　スキーに行きます。

しかし、自発的な談話においては節単位認定がいつも明確であるとは言えない。例1は作例であるが、以下の例２は実際の発話の例である。

（例２）[2]　はい、
　　今日は、えー、休日です。
　　えーと、朝ごはんを食べてから
　　スキーに行くこ＝
　　彼はスキーに行くことにしました。

例２には上述の「節」の定義に合わない単位が認められる。例えば、感動詞のみの「はい」や二つの述語を含む「彼はスキーに行くことにしました」という節がそうである。さらに、自発的な話し言葉では逡巡（例２の３行目のフィラー「えーと」)、言い直し、言い止め（例２の４行目）な

[1] 文章の構造を構成する単位は本稿に「節」と呼ばれるが（参考、高梨等 2003、Chafe 1982)、他研究では基本談話要素‘elementary discourse unit’(Carlson et al 2003, Kibrik and Podlesskaya 2006)などの術語も現れる。

[2] 本稿の例は、特に指定しない場合は「多言語の話し言葉語りのコーパス」からのデータである。全ての例において理解の妨げになる表記の特別な記号は使わないようにした。また、例１、２、１６と１７以外の例には、節の区切りを指示していない。分割のルールは具体的なコーパスの目的によって構成されるべきである。

どの要因により節の範囲が認定される場合が少なくない。本研究の目的は、以上のような節単位認定しにくいケースを分類することである。

## ３．研究対象

研究のデータとして「多言語の話し言葉語りコーパス」を用いる。このコーパスはロシア国立人文大学のプロジェクトで、日本語、ロシア語、英語、フランス語、アラビア語などの話し言葉の語りを含む。しかし、このコーパスの全言語のデータは扱いにくいので、予備研究の対象として日本語とロシア語のデータを選択した。それは、この二つの言語が類型論的に異なり、かつ同族語でないからである。

検討に際しては、４０の日本語（約１時間）と４０のロシア語（約１時間）の話し言葉の語りを選んだ。先行研究（高梨等2003, Kibrik and Podlesskaya 2006）が提案した節単位認定のルールを応用してデータの分割を実施した結果、分割しにくいケースは以下のように分類された。

## ４．分類

節に分割しにくい単位の中には、世界の諸言語において見られる言わば代表的なタイプがあると思われる。その単位は文法構造に関係するものと談話の自発性に関係するものに区別できる。前者は、二つ以上の動詞を含む構文である。後者は、自発的な談話においては発話プランが途中で変更される場合、あるいは話し言葉に特徴的な述語のない単位である。これらのタイプについて以下に順次述べていくことにする。

### 4.1. 文法構造に関係するもの

- 連体節構文（construction with adnominal clause）とは、名詞を修飾する節である。

この構文においては、連体節なのか連体句なのかの区別が問題になることがある。以下の例では、例３がもっとも典型的な連体句、例６がもっとも典型的な連体節であるが、例４と例５は中間的である。

（例３）目の前に良さそうな車がありました。

（例４）帰る時間がやってきました。

（例５）散歩の好きなおじさんがいて、

（例６）でも、さっきまで飲んでいたワインのせいで、・・・

話し言葉のロシア語においても、形容詞は修飾する名詞の直後に置かれた場合、イントネーションによって連体句とも連体節とも解釈できる（Podlesskaya and Kibrik in press）。

連体句と連体節の間の区別はどこにあるか、つまり、連体の表現を独立した単位として認定するかしないかは、節単位認定のルールによって決められるべきである。

- 引用節構文（quotation construction）

引用節構文の分割の問題は少なくとも２つある。一つは引用節とそれを埋め込んでいる節の境界の位置、もう一つは、話し言葉では「言う」などの引用を表す動詞はしばしば省略したり引用内部の中に埋め込んだりすることである。例７ではロシア語の'govorit'という引用の動詞が引用の中に埋め込まれている。一方、日本語の２つの訳には、引用の動詞の省略と境界の位置の問題が見て取れる。

（例７）Ona, - govorit, - uzhe ushla.

彼女 言ってる もう 出かけた

‘彼女もう出かけちゃったって（省略）／彼女もう出掛けちゃった（境界？）って（境界？）言ってる。’

- 連動詞構文（serial verb construction）とは、２つあるいはそれ以上の動詞が、連結するための不変化詞や接語などを伴わずにつなげられている構文（オクスフォード言語学辞典2009）。

日本語では、「～たり～たりする」や「持って来る」のような例をこの構文に含めることが出来る。ロシア語でも同様の動詞の連続が見られる。

（例７） on hodil vybiral

彼 歩いていた 選んでいた

‘彼は歩き回って選んでいた’

• 認識様態性構文（epistemic constructions）とは、事実的必然性、可能性などを示す動詞を含む構文である。認識動詞は節に付き、２つ以上の動詞を含む構文になる。その例として日本語の「〜かもしれません」、「〜でしょう」とロシア語の「kazhetsya」、「mozhet」という動詞の表現が挙げられる。

• 名詞節構文（noun clause constructions）とは、名詞もしくは名詞句と同様の統語的役割を担う節である。（オクスフォード言語学事典2009）。

名詞節を含む構文は多岐にわたるので、コーパスの各言語の名詞節構文を詳細に調べてから節単位認定のルールを構成するべきである。以下の例には日本語の名詞節を含む構文が現れている。

（例８）長距離を歩くことが好きだ。／歩くことが好きだ。／食べ歩きが好きだ。

（例９）長距離を歩くことが出来る。／歩くことが出来る。／食べ歩きが出来る。

（例１０）長距離を歩くことにした。／歩くことにした。／食べ歩きにした。

以上の各例では、最初の文がもっとも名詞節らしく、最後の文がもっとも名詞句らしい。また、例８から例１０を順に見れば、形式的に名詞節を埋め込んでいる主文がまだ主文であるかすでに構文の一部かという疑問が出てくる。

### 4.2. 自発的な話し言葉の特徴に関係するもの

• 「勘体句とは、主に名詞句あるいは間投詞からなっていって、なんらかの種類の感動の意味を持ったものである。」（南1974）

体言、呼びかけ、感動詞、フィラー、表題などの述語のない要素は話し言葉によく現れるものである。

以下の例１１〜例１３は様々な日本語の勘体句の種類を表す。ロシア語のコーパスでも同じような機能を持つ単位がある。

（例１１）ねえねえ、君達、子供にはどんなおみやげがいいかなあ。

（例１２）ほら、これがプレゼントだよ。

（例１３）えーと、彼の名前はセバスチャンと言います。

勘体句は統語的にもイントネーション的にも談話機能の面でも独立した単位なので、それらが節単位で認定される研究が少なくない（高梨等 2003、南 1974）。しかし、フィラーのような単位はよく他の節の中に挿入されるため、フィラーを独立した節として扱っていない研究も存在する（Kibrik and Podesskaya in press）。いずれにしても、具体的な多言語コーパスにおいては全言語に対して勘体句の分割ルールを一致させることが重要である。

• 後置（afterthought）

話し言葉においては語順の逆転がかなり頻繁に見られる。SOV語順の日本語では、後置の場合には文節がその係り先となる述語句の直後に置かれる。

（例１４）その後、んー、紅茶とピザを食べました、朝ごはんに。

SVO語順のロシア語では、後置を表すために語順の逆転に加えてイントネーションの手段もよく使用されている。

(例１５) vypil horosho tak

酒を飲んだ かなりよく

‘彼はかなりよく飲んだ’

後置の文節は、形式的には独立した節ではなく統語的に直前の節に係るにも拘らず、イントネーションとその談話機能により、独立した単位として扱われることが多い。

• 言いさし、言い直し （corrections）

自発的な話し言葉では「挿入や発話の変更により、発話が途中で言いやめられたり、文の構造が発話途中から変ってしまう」（高梨等2003）ことがよくある。その場合、言い直した部分をどう分割するかは、言い直した部分の長さや話者の意向等にかかっている（Kibrik and Podlesskaya in press）。例１６と例１７は両方とも言い直しを表すが、分割の仕方は異なっている。

（例１６） 車を、== （言いさし）

子供が車を欲しいと （言い直し）

言ったので、

（例１７） 酔っぱらってしまっ＝しまいました。

例１６では、話者は言い始めた節を止めて、新しく言い直した節を発話しているので、言いさしと言い直しの間に節の境界を認定した。また、例１７では話者は前に発話した部分を落とさずに言い直しをしたので、言いさしと言い直しの間に節の境界を認定していない。

このように、言い直しにおける節分割のルールは言い直しの種類により作成しなければならない。

## ５．結語

本稿では、ロシア語と日本語の話し言葉のコーパスに基づいて節単位認定しにくい単位を分類した。その単位は文法構造に関係するものと談話の自発性に関係するものに区別できる。前者の範疇には連体節、引用節、連動詞、認識様態性、名詞節という２つ以上の動詞を持つ構文が含まれ、後者の範疇には勘体句、後置句、言い直しという話し言葉の代表的な単位が含まれる。

分類の記述に言語学の普遍的な概念を使用したことにより、ここで提案された分類は全般的なものであると言える。多言語コーパスの節単位認定のルール設定に向けた広範な利用が期待できるだろう。

## 文献

「オクスフォード言語学事典」2009　中島平三・瀬田幸人監訳、朝倉書店

高梨克也・内元清貴・丸山岳彦　2003「『日本語話し言葉コーパス』における節境界認定」(Version1.1)http://www.kokken.go.jp/katsudo/seika/corpus/public/manuals/clause.pdf

南不二男 1974「日本語の構造」大修館書店

**Carlson L., Marcu D., Okurowski M.E. 2003.** Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory // van Kuppevelt J., Smith R. (eds.) Current directions in discourse and dialogue. Kluwer Academic Publishers. pp. 85-112.

**Chafe, Wallace. 1982.** Integration and involvement in speaking, writing, and oral literature. In: D. Tannen(ed.) *Spoken and written language:Exploring orality and literacy*. Norwood: Ablex, 35-54.

**Kibrik A.A., Podlesskaya V.I. 2006.** Problema segmentatsii ustnogo diskursa i kognitivnaya sistema govoryaschego (話し言葉談話の分割の問題と話者の認知仕組み) // Soloviev (ed.) Kognitivnye issledovaniya. No 1. Moscow: Institut psihologii RAN. pp. 138-158

**Kibrik, Andrej A., Vera I. Podlesskaya (eds.) In press.** Night dream stories: A corpus study of spoken Russian discourse.