LN-002

ロシア語タグ付きコーパスの作成
Development of the POS Tagged Corpus of Russian Language

成定久美子†　　佐野　洋‡
NARISADA, Kumiko　　SANO, Hiroshi

1. はじめに

本稿では，ロシア語タグ付きコーパスの作成と，作成したタグ付きコーパスを使った無人称構文の分析について述べる。

述語副詞からなる無人称構文では，述部を構成する述語副詞と動詞不定形の体の間に連関関係が存在している。しかし，一般流通の辞書の述語副詞に関する記述は，そうした関係の記述が不十分である。筆者等は，述語副詞と動詞不定形の体の共起関係を，タグ付きコーパスを用いて調査した。

2. ロシア語タグ付きコーパス

2.1. 入手可能なコーパス

現時点で公開されており，日本でも入手可能なタグ付きコーパスは，The Uppsala Corpus of Modern Russian Texts (Uppsala University が提供)のみである。このコーパスは 600 ファイルのロシア語テキスト(合計 100 万語)から成る。テキストは情報伝達文書(25 分野，1985-1989 年)と文学作品(40 作家，1960-1988 年)から収集されている。

しかし，このコーパスには，単語単位での品詞情報などが付与されていない。したがって，文法情報の組み合わせを利用した検索が行えず，効果的に統語論研究に役立てることができない。

2.2. タグ付きコーパスの作成

筆者等はロシアで公開されているタグ付与ツール Mystem を利用してロシア語タグ付けコーパスを作成した。

2.2.1 Mystem の利用

Mystem はロシア語のタグ付けツールで，Илья Сегалович 氏 (もしくは Виталия Титовый 氏)に連絡後[1]，指定された FTP サイトからダウンロードして利用する。Mystem を使ってテキストを解析すると，品詞のほかに，性，数，格などの形態に基づいた詳細なタグが，語単位で付与される。

このソフトウェアの長所は，文法タグが見える形で付与されること，11 種類の出力オプション(表 1を参照)があり，用途に応じた結果を得ることができること，解析処理速度が速いことなどが挙がる。

表 1. Mystem のオプションの例

コマンド	機能
-n	入力語 1 語に対して 1 行ずつ出力する。
-c	スペース(_)，改行(\n)なども含めて，原文の情報を全て出力する(原文を復元できる)。\は\\になる。
-i	文法情報を含めて印刷する。

タグ付与の手順は次のとおり。

1. 出力時のオプション，入力テキスト名，出力テキスト名を指定する。テキスト名は，ファイル名か絶対パスを入力する。コマンド入力の例を以下に示す。

```
Y:\mystem –nic Y:\belkin.txt Y:\belkin.tag
```

2. 実行後，入力ファイルにタグを付与した結果が指定されたファイル名で出力される。

Mystem はファイル単位でのタグ付けを行う。大規模な文章テキストを扱うことを目的として，ディレクトリ単位で処理を行えるように Mystem を内部で起動するスクリプトファイルを作成し，ディレクトリ単位のタグ付けを行った。

2.2.2 電子テキスト

電子テキストは Библиотека Максима Мошкова[2]からダウンロードしたファイル(総数 1161 ファイル，約 2500 万語，データ容量 172MB)を利用した。テキストは評論(123 ファイル)と文学作品(1038 ファイル，19-20 世紀)，インタビュー記事(2 ファイル)から成る。

これらの電子テキストに付与されていた HTML タグは削除し，文字コードは windows1251 から unicode(UTF-8)に変換した。

2.2.3 タグデータの形式

タグ付け後，プログラムによる用例抽出を容易にする目的で 1 形態素(単語)1 行を出力形式とした(-n オプション)。また，形態素には文法情報を付与し(-i オプション)，スペース(_)，改行(\n)など，原文の情報も全て出力する設定にした(-c オプション)。Mystem によるタグ付けの例を表 3 に示す。

表 2. 入力テキストの例

```
П. Беликина, предлакаумых ныне публике,
```

表 3. Mystem 出力結果

```
П{П??}
._
Белкина{беликин?=S,муж,од=род,ед}
,_
предлакаемых{предлакать=V=непрош,род,мн,прич,несов,страд}
_
ныне{ныне=ADV=}
_
публике{публика=S,жен,неод=дат,ед|публика=S,жен,неод=пр,ед}
,_
```

単語の文法情報は，入力語の後ろの{ }カッコ内にタグ情報列として表示される。曖昧性がある単語には，可能性のある全てのタグ情報列が出力される(表 3の

[1] 利用は無料である。連絡先など詳細は以下を参照のこと。
http://corpora.narod.ru/mystem/mystem.html

[2] http://lib.ru

прудлагаемых 部分を参照)。タグの区切りは|である。未知語については見出し語の後ろに"?"が付与される。未知語は 2 種類の処理が施される。規則に基づいて推測された文法情報が付与される場合(表 3の Белкина 部分を参照)と，全くの未知語としてタグ情報列の代わりに"?"が付与される場合(表 3の П 部分を参照)である。

Mystem で付与されるタグの種類を表 4にまとめる。

表 4. Mystem のタグ一覧

タグの種類	タグの値
見出し語形	—
品詞名	V(動詞), S(名詞), A(形容詞), ADV(副詞), PR(前置詞), INTJ(間投詞), CONJ(接続詞), PART(小詞)
数	ед(単数), мн(複数)
性	муж(男性), жен(女性), сред(中性)
格	им(主格), род(生格), дат(与格), вин(対格), твор(造格), пр(前置格)
人称	1-л(1 人称), 2-л(2 人称), 3-л(3 人称)
活動体	од(活動体), неод(不活動体)
体	несов(不完了体), сов(完了体)
時制	непрош(非過去形), прош(過去形), инф(時制なし／不定形)
変化形	деепр(副動詞), прич(形動詞), изъяв(直説法), пов(命令法)
その他	страд(被動相), кр(短語尾形)

3. 述語副詞からなる無人称構文

3.1. 無人称構文

無人称構文とは主格がない文を指し，述語副詞からなる無人称構文は一般化した評価(判断，態度)を叙述する際に使用される文である。例えば，英語の"It is easy to please John."は，主語に虚辞の it を用いた非人称構文である。これに類する文で，その構文パターンを表 5に示す。

表 5. 構文パターン

<与格名詞> [быть 変化形[3]] <述語副詞> <動詞不定形>

動詞不定形の意味的な主語(Agent)は，主格名詞としてではなく与格名詞として表層に現れる。ロシア語では文脈上明らかな場合，日本語と同じように主語や目的語が省略されるので無人称になる。主語が義務的に省略されることで，評価(判断，態度)を示す主体が現れず，そのため，一般化された評価(判断，態度)の意味を表す。

3.2. 述語副詞

述語副詞は状況叙述の機能を有し，対象状況に対して評価(判断，態度)を表明する。英語の"easy"や"need"のような語である。語彙分類は，文献[3][8]の無人称述語副詞(無人称文の述語として用いられる特殊な副詞)に拠った。

機能的に述語副詞は，文の構造を決定する(動詞の出現形に不定形を指定し，その動詞の意味的な主語を与格名詞で出現させる)だけではなく，動詞不定形の体の決定にも重要な役割を果たす。ただし，無人称構文という文構造としての体の指定能力はなく，述語副詞の意味と体の意味が結合して，一つの意味が決まると考えられる。

文献[7]においても「それぞれの体に一定の叙想的意味が固定される結果，同一の語と結合する不定法の体の使い分けが，その語のもつ複雑な意味を区分する手段として役立つことがある(p.83)」ことが指摘され，述語副詞を含む無人称構文は「同一の語と結合する不定法の体の使い分け」の典型例である。体の選択には述語副詞が強く関係していると考えられる。

述語副詞が完了体か不完了体かのいずれかを選択し(範疇選択)，同時に述語副詞と体の結合により意味が決定する(意味選択)。この範疇選択と意味選択の二種類の共起出現には密接な関係が見られることが予想される。また，体の意味の個別性(表 6参照)や，述語副詞の意味が評価や判断，態度という一般的な用語でしか纏められないことから，述語副詞と動詞の体の選択関係が一般的には規定できず，述語副詞ごとに固有の体の選択基準があることが予想される。

3.3. 動詞の体

ロシア語では体が不可避的に表され，完了体か不完了体のいずれかの形態をとる[4]。この二項対立では，完了体が意味的に有標で不完了体が無標である。完了体は，動作を全一的に把握することにより一定の追加的意味特徴を持ち，それにより不完了体から区別される[5]。それに対し，不完了体は「一定の意味的特徴を欠いており，なんらの制限無しに使用することができる[6]」のである。そのため完了体が伝達できない動作の経過，展開，無限の反復といった意味内容を表現する際には不完了体が用いられる。体のもつ意味を表 6に示す[7]。

表 6. 体の意味

	完了体	不完了体
一般的意味	動作を全一的，一体のものとして，全過程(初め-過程-終了)を圧縮・収斂して把握	その動作そのものが開始されたか否かのみを表す(過程中か終了したかは状況・文脈で判断)
個別的意味	1) 具体的事実。具体時定的動作の一回のみの生起の唯一性 2) 一括化 3) 例示的。一例を通しての一般化	1) 一般的事実。動作の具体的・特定的状況について言及しない。 2) 過程 3) 反復

完了体の付加的意味には，結果の達成，結果の残存，結果の評価，期待された動作，完了した動作の結果の不可能(危惧)，潜在的(不規則的反復の)可能性をふまえた一般化，無意識的動作がある。

一方，不完了体の付加的意味には，動作の名指し，動作事実の有無の確認，経験の有無，回想的過去，結果の消滅，未然の動作，結果達成を希望しない，禁止，不必要，否定的意図・嫌気，着手，様態，努力，所要時間，予定，同時性，規則的，(頻発的)反復の可能性をふまえた一般化がある。

[3] 英語の be 動詞のような動詞で Tense を表す。現在時制では省略される。

[4] 完了体と不完了体の関係は，英語の「単数と複数のような関係」にある(文献[5], p.28)。

[5] 文献[7], p. 31-32

[6] 文献[7], p.32

[7] 文献[7], p.1-27

4. タグ付きコーパスを使った実証分析

取り上げた問題は，述語副詞からなる無人称構文における，動詞不定形の体(完了体か不完了体かの違い)と述語副詞の関係性である。

この関係性については，一般的な説明が文法書や辞書記述になく，また，個々の述語副詞における辞書記述には，記載内容に偏りがみられる。つまり，各述語副詞について，完了体か不完了体かの指定が明示されている項目と，曖昧な表現でしか記述されていない項目がある。

また，関係性について記述がある場合でも，執筆者のロシア語運用の経験則に基づいた説明であることがほとんどのようである。大規模で広範な文章を使った調査による統計的な裏づけが求められる。

以下では，先行研究として文法書や辞書の記述内容を調査した結果を示す。

4.1. 述語副詞と体に関する辞書記述

本節では，述語副詞に続く動詞不定形の体について，ロシア語文法書，日本人学習者用教科書，上級者用露和辞典で調査した結果の一部を示す。

4.1.1 ロシア語文法書

述語副詞 нечего に関する記述[8]

> 「なお，後者の意味，つまり「ーする必要はない」という意味で使われた場合は，不定形は不完了体しか使えません。」

文法書では，述語副詞ごとに説明記述がある。意味的に体の形が固定する場合には，上記のような説明でもよいだろう。しかし，述語副詞の大部分は，どちらの体も後続し得る。そのような述語副詞について，不定形不完了体が，完了体に対してどのくらいの割合で使われるのか，といったことは言及されていない。

4.1.2 日本人学習者用教科書

述語副詞 нельзя に関する記述[9]

> нельзя＋不完了体：禁止
> нельзя＋完了体：不可能

上記に引用したように，説明記述は僅かに 2 行である。学習者は，ある述語副詞が登場するたび毎に，自ら辞書を引き，対応する用例を観察し，その結果を基に述語副詞と体の用法の規則を作っていくことになる。

4.1.3 上級者用露和辞典

述語副詞 нужно に関する記述[10]

> не＋нужно＋不定形の形において
> 「不必要・禁止の意なので，不定形はかならず不完了体」
> 「動作実現の必要性が意識されているので完了体」

上記の引用例で示すように，述語副詞に続く動詞不定形の体についての説明記述は少なく，学習者にとってこの記述説明は不十分だと思われる。また，説明は特定の述語副詞に限られていて，殆どの述語副詞では，後接の動詞不定形が存在する程度のことしか言及されていない。

[8] 文献[5], p.217, l.15
[9] 文献[1], p.57
[10] 文献[3], p.1223

4.2. 調査目的

前節で述べたように，体が決定されなければ動作状況を言語化できないにもかかわらず，述語副詞ごとの体の選択規則が不明である。そこで，まずは意味の確定における述語副詞の体の選択規則を明らかにする。

4.3. 調査対象

次の文型は調査対象から除外される。

- 語順が入れ替わっていたり，述語副詞と動詞不定形の間に別の単語が入っていたりする場合
- -o で終わる述語副詞[11]の変化形＋動詞不定形

4.4. 調査方法

述語副詞 271 語について，述語副詞と動詞不定形からなる構文を調査し，動詞不定形部分の体の違いを用例数で計測した。用例数の比率を計算し，辞書記述がある述語副詞については，説明されている関係性と調査結果を比較し，説明の妥当性を調べた。

抽出の際は，Mystem のタグの種類の制限から，副詞と無人称述語副詞を区別することができないので，文献[8]に基づいて作成した述語副詞辞書との照合を行い，「述語副詞＋動詞不定形」文抽出の精度を高めた。

5. 調査結果

5.1. 調査結果(1)

抽出された用例「ADV＋動詞不定形」の数は 39425 例であった。また，4.2節で定義した述語副詞 228 語のうち，用例が抽出されたのは 125 語であった。

本節では抽出したデータを基に，述語副詞＋動詞不定形に関する辞書記述を検討する。調査には文献[3] を使用した。この辞書で「無人述」の項目に不定形の体が指示されている述語副詞(9 語)に関して，その記述内容を調査結果と比較検討した。その一部を以下に示す。

5.1.1 надо (～しなければならない)

надо＋動詞不定形は，全部で 9108 用例抽出された。9108 用例中，不完了体は 5455 例(約 60%)，完了体は 3653 例(約 40%)であった。若干，不完了体の方が多い。一方，文献[3]に掲載されている例文は全 9 例で，うち 8 例が不完了体で，完了体は 1 例のみであった。辞書記述例として不完了体に偏った例文を提示されると，辞書を参照する者は，完了体が後置する場合も多くあるということは分からない。

> надо の体に関する辞書記述
> 「不必要を意味する не надо のあとでは不完了体不定形を用いる」

не надо＋不定形の形は，9108 例中 602 例あり，そのうち不完了体と判断されたものが 575 例，完了体と判断されたものが 27 例あった。その中からエラー文[12]を取り除いた結果，上記の規則に当てはまらない例，つまり не надо＋不定形(完了体)である例が 8 例見つかった。

5.1.2 поздно(～するには遅い)

поздно＋動詞不定形は，全部で 82 用例抽出された。73 用例中，不完了体は 40 例(約 55%)，完了体は 33 例(約

[11] これらは性質形容詞から作られた副詞で，様態の副詞同様，比較級，最上級をもつ。
[12] 体の判定が間違っている，もしくは体の判定の曖昧性により重複して抽出された文。

45%)であった。若干，不完了体の方が多い。一方，文献[3]に掲載されている例文は全 4 例で，そのうち完了体は 1 例のみであった。

> поздно の体に関する辞書記述
> 「不定形はふつう不完了体を用いるが，否定文において動作実現を可能と見ている場合は完了体」

「ふつう」という語から得る印象からすると，多数の例は不完了体であることが予想できる。不完了体と完了体の割合がそれほど変わらないことから，「ふつう不完了体を用いる」という記述には再考の余地があろう。

また，動詞不定形が完了体だと判定された 33 例のうち，否定文でないものは 21 例あった。その中から поздно が副詞用法で使われている文(慣用句 рано или поздно の構成素となっている文)16 例と，エラー文 1 例を除いた結果，上記の規則に当てはまらない例，つまり肯定文において完了体を使用している例が 4 例見つかった。この 4 例が辞書記述の説明を逸脱する例外なのかどうかは，今後子細に検討したい。

5.2. 調査結果(2)

文献[3] において，不定形の体に関する指示が記されていない述語副詞は，157 語ある。そのうち「述語副詞＋不定形」を含む例文が提示されているものは 71 語あった。

このような述語副詞に関して，抽出結果における体の比率と，辞書の例文における体の比率とを比較し，辞書の例文が実際の用法を反映しているかどうか検討した。抽出結果における体の比率と，辞書の例文における体の比率とを比較した結果の一部を表 7に示す。

表 7. 体の比率の比較結果

	抽出結果		辞書の例文		検証結果	
単語	不	完	不	完	傾向	用法
должно	87%	13%	100%	0%	○	×
нельзя	50%	50%	80%	20%	×	×
охота	75%	25%	0%	100%	×	×
пора	61%	39%	100%	0%	○	×
трудно	28%	72%	40%	60%	○	○
угодно	54%	46%	0%	100%	×	×

表中，「傾向」は，不完了体と完了体の比率の傾向が等しいかどうかを表す。「用法」は，実際の用法で不完了体・完了体ともに結合する(もしくはどちらかとしか結合しない)ことが，例文に反映されているかどうかを表す。

体の比率の傾向が反映されている副詞は 71 語中 48 語(68%)，用法が反映されている副詞は 71 語中 14 語(20%)だった。この結果から，例文中の体の比率は実際の用法をほとんど反映していないことが分かった。理由としては，辞書の紙面に限りがあること，執筆者の主観によって例文が選択されていることが考えられる。

しかし，後接する体が具体的に指示されていない場合，学習者は例文を参照して結合規則を推測する。偏った例文を記述例として提示した場合，誤った規則が学習者の中で化石化する危険性がある。そのような事態を避けるために，英和辞書だけではなく露和辞典においてもコーパスを利用して，例文に実際の用例を反映することが必要あると考える。

6. 実証分析の成果

6.1. 分析効率の向上

タグ付きコーパスを利用することにより，大量のデータを対象として，タグ名や辞書登録形などを指定した言語調査を効率的に行うことができるようになり，分析にかかる労力と時間が大幅に短縮された。

6.2. タグ付きコーパスの有用性

コーパスから抽出した用例は，述語副詞と動詞不定形の体に関する辞書記述の妥当性を判定するのに役立った。また，用例を分析した結果，述語副詞と動詞不定形の体について，幾つかの示唆的な関係性を見いだすこともできた。こうした調査と分析から，タグ付きコーパスの有用性が検証できたと言える。

7. おわりに

この結果をもとに，述語副詞に固有の体の選択規則や，述語副詞の意味分類で一般化できる規則を明らかにしたい。そうして母語話者の無意識の認知に基づく運用規則を明らかにしたい。

こうした取り組みは，言語運用を基本とする言語現象の解釈の枠組みの構築に大いに役立つだろう。また，いままで曖昧なままになっていた文法事項に統計的な裏づけを与えることは，言語学的研究や教育方法への寄与も大きいと考えている。

謝辞

本研究は平成 14-16 年度文部科学省科学研究費(基盤研究(B)(2))「全電子化検定済み教科書データの解析と大規模日本語コーパスの構築」(研究代表者　佐野洋)の助成を受けた。本論文をまとめるにあたり，有用な御指導とコメントを頂きました査読者の方々に感謝いたします。

参考文献

[1] 佐々木照央，「速修ロシア語(増補版)」，研究社，1999.
[2] 城田俊，『現代ロシア語文法　中上級編』，東洋書店，2003.
[3] 東郷正延ほか(編)，「研究社露和辞典」，研究社，1988.
[4] 成定久美子，佐野洋，「対照言語学の視点を考慮した多言語コーパスの作成とその利用(2) ロシア語」，第 2 回情報科学技術フォーラム後援論文集，2003.
[5] 原求作，「ロシア語の体の用法」，水声社，1996.
[6] 原求作，「ロシア語文法の要点」，水声社，1996.
[7] O.P.ラスドーヴァ，磯谷孝(訳)，「ロシア語動詞　体の用法」，吾妻書房，1975.
[8] А.А.Зализняк., “Грамматический словарь русского языка”, Русский язык, 1977г.
[9] Б.П.Кобрицов., “Морфология и синтаксис в проекте 'Русский стандарт' (создание корпуса грамматически размеченных русских текстов)”, 2003г.
http://www.dialog-21.ru/Archive/2003/Kobricov.htm
[10] Г. И. Кустова, В. А. Плунгян., “Краткое описание проекта 'Русский стандарт'”, 2002г.
http://rscorpora.narod.ru/zay.html
[11] И.М.Пулькина, Е.Б.Захава-некрасова., «Учебник русского языка для студентов-иностранцев», Адзума сиобо, 1968г.