

日本学術振興会科学研究費補助金 基盤研究(C)

2017年度－2019年度

(研究課題番号 17K02926)

日本語母語学習者データに基づく
ロシア語学習者コーパス構築の
基盤研究

研究成果最終報告

2020年3月31日

科研費プロジェクト

「日本語母語学習者データに基づくロシア語 学習者コーパス構築の基盤研究」

(2017.04-2020.03)

研究成果最終報告

— 学習者言語分析の可能性と将来的展望 —

林田 理恵

(大阪大学大学院言語文化研究科)

I. 本研究の概要

1990年代以降、英語を中心にコーパス言語学の応用研究領域として、学習者コーパスが相次いで構築され、日本語学習者コーパスについても2009年あたりから整備が進んでいる。ロシア語学習者コーパスに関しては、これまではモスクワの National Research University, Higher School of Economics (以下、HSE) 人文学部言語学コースのスタッフが中心となって2013年に公開した Русский учебный корпус (Russian Learner Corpus = RLC, <http://web-corpora.net/RLC>) が唯一のものであり、日本語を母語とする学習者のデータについてはいまだ整備されていない状況であった。

また、学習者コーパスにおいては、特に誤用情報の付加が重要な役割を果たすが、これまでの研究では誤用情報の分類・構成に関する明確な指針は確立されてこなかった。誤用情報には語彙レベル、形態レベル、構文レベル等、多様なタイプが混在し、これらをどのように分類し、タグとして構成していくかにコーパスの信頼度、有効性の成否が左右される。学習者言語の分析、ひいては教育面に資する有益なデータ取得に役立つコーパス作成のためには、その点の改善が急務とされている。

本研究は、国内外で初となる日本語を母語とする学習者データに基づくロシア語学習者コーパス構築に向け、理論的・技術的な基盤整備を行うことを目的として開始した。作業はロシア・コーパス開発研究チームとの協同研究体制を前提に、以下の3点を最終目標に進められた。

- 1) 信頼性の高いアノテーション済み学習者コーパスを作成するための、付加する情報タグの有効な分類・構成のあり方を検討、新たな付加情報タグ・ガイドラインを提案。
- 2) 収集されている作文試験結果の電子データに基づき、アノテーション、すなわち情報タグの付加を試行、公開されている品詞タグ付けシステムを利用し、日本語母語学習者データに基づくロシア語学習者コーパスのパイロット版を作成。
- 3) パイロット版により試験的にコーパス分析を行い、試作したコーパスによってロシア語学習者の言語使用特徴がどの程度、明示化されるのか、提案した付加情報タグ・ガイドラインの有効性も含めて検証。

コーパスで収録対象としたデータは、研究代表者の所属する機関で実施している TORFL 試験 — CEFR (Common European Framework of Reference for Languages : Learning, teaching, assessment) 基準によるロシア教育科学省主催「外国人のためのロシア語検定試験 (The Test of Russian as a Foreign Language = 以下, TORFL)」 — のうち、1,2年次における A2 (Basic level), B1 (Intermediate level) レベルの作文試験結果、約 1400 人分、約 2800 編の作文である。

15 年間に亘って収集されたこれらの大規模データは、作文のトピックや作成年、習熟度等、テキストや学習者属性情報の内容が明確であるという特徴を持つ。このような特徴によって、客観的・実証的観察、記述、分析を可能とするコーパス構築の実現が可能となった。また、収録された作文試験結果はすべて、教員による添削済みデータであり、学習者コーパスの成否を左右する誤用情報タグの付加についても、その基礎となる情報がすでに準備された状態であった。加えて、添削は TORFL で定められた客観的評価項目に基づいて実施されたものであり、それらは、誤用情報タグの分類・構成を考察する際にも有効な情報として活用できるものであった。

II. 研究活動の流れ

2017 年度～2019 年度、3 年間に亘って調査・研究活動、コーパス構築作業は以下のような流れで実施された。

【2017 年度】

1) 他言語における既存学習者コーパスに関する情報収集と先行研究の整理 — すでに構築されている英語、日本語をはじめとする学習者コーパスについて情報収集を行い、開発・分析に関わる先行研究の知見を整理、これまでの成果と問題点、不十分点等を明らかにした。その上で次の 3 点について研究調査を進めた。

2) HSE コーパス開発部門 (ロシア連邦・モスクワ) における研究調査 — 現存する唯一のロシア語学習者コーパス RLC を構築した HSE コーパス開発部門 (ロシア連邦・モスクワ) に滞在し、研究協力者である部門長をはじめ、開発に関わった研究チームから開発の経緯、具体的な技術的問題、現行の稼働状況等、詳細についてヒアリングを行った。また、日本語母語学習者のデータ利用に基づくロシア語学習者コーパス・パイロット版構築に向けた協力体制、作業分担等の打ち合わせを行い、RLC システム上に下位コーパスとして Japanese RFL Learner Corpus (以下, JRFL Corpus) を開設することになった。

3) データの電子化作業の開始 — 収集されている TORFL 作文試験結果について、データの電子化作業を HSE 側の助力を得て開始した。

4) HSE コーパス開発部門でのアノテーション作業に関する研修に参加 — HSE での研修において、誤用タグ付けの手順についての説明を受け、また実際の作業プロセスに参加、JRFL Corpus アノテーション作業の参考となる知見を得た。

【2018 年度】

1) HSE コーパス開発部門が構築した RLC システム上に下位コーパスとして JRFL Corpus 開設の準備を進め、以下の JRFL Corpus 概要画面を設定した。

JRFLLC

RLC Subcorpus

概要

日本語を母語とするロシア語学習者コーパス — JRFLLC は、言語学・ロシア語教育を専門とする研究者、林田理恵 (大阪大学) と E. ラリーヒナ (National Research University, Higher School of Economics) の共同研究によって構築されたものである。

また、本研究は日本学術振興会科学研究費助成事業による助成金 (学術研究助成基金助成金・基盤研究 C—課題番号 17K02926) を受け実施されているものである。

JRFLLC は日本語を母語とするロシア語学習者のデータに基づく学習者コーパスである。本コーパスにより日本語母語学習者のロシア語使用における多面的な言語特性が明らかになり、その分析に基づいて、日本での現行ロシア語教育における教材や指導案、カリキュラムについて抜本的な見直し、改良を図っていく可能性が広がると期待する。と同時に、RLC のサブコーパスとして、学習者の各母語の違いによって、学習者言語の特性や言語ストラテジー特性にどのような異同点が観察されるか、実証的で正確な記述と分析を可能とし、類型学的研究にも将来的に資するものになると考える。

JRFLLC とは？

JEFLL Corpus は日本の大学で、ロシア語を専攻する大学生の作文データに基づいて作成されている。これらの作文データは、15 年間に亘って収集された CEFR 基準によるロシア教育科学省主催「外国人のためのロシア語検定試験、A2 (Basic level), B1 (Intermediate level) の 2 レベルの作文試験結果データで、現在、約 1400 人分、約 2800 編の作文がデータとして使用可能な状態にある。目下は、その一部に品詞情報、誤用情報タグ等が付加され、すでに公開されているが、作業は続行中で、できるだけ早い段階で全データについて、情報タグが付加された状態で公開する予定である。また、データ数そのものも、今後毎年の試験実施に伴い、増やしていくことが可能である。

メタデータについて

JRFLLC は当初、該当する教育機関の必要性によって構築が始まったが、コーパスが持つ設計基準、豊富なパラメータ上の優位性によって、今後、様々な研究・分析上の課題を解決する可能性をもたらすものと考えられる。コーパスに収録されている各テキストは以下のメタデータが付加されている。：

1. 学生名 (仮名)
2. 性別
3. 学習者の母語及び学習歴を有する言語
4. 学習者の学習言語習得レベル
5. 作文作成時点での学習歴
6. 作文作成に際しての制限時間の有無及び制限時間数
7. 作文ジャンル

8. 作文のスタイル (叙述, 論証等)

これらのカテゴリーは各作文の Header Identification Box (Header ID) に反映されている。このカテゴリーを使って自動的にさらなるサブコーパスを作成することも可能である。

また, JRFLLC 内の全ての語について品詞・文法性・格・アスペクトの情報をもつ形態タグが付与されている。さらに, 作文内の誤りについては, 誤用情報タグが付加されている。これらの情報はコーパスインターフェースにより, 文法的・語彙的カテゴリー別による検索が可能である。

JRFLLC の使用について

JRFLLC は web 検索ツールを介してフリーで公開する資料である。JRFLLC Corpus を用いたあらゆる研究成果物には, 本コーパスを利用した旨, 必ず記載すること。

— . — . — . — . — . — . —

JRFLLC Corpus の構築が進めば, ロシア語初級レベル (A1-A2) における習得困難項目である名辞類格変化や動詞人称変化などは, 本コーパスを使用して瞬時にその誤用の傾向, また学習者がその時点で作り上げている学習者言語体系の実像が明らかになると考えられる。また JRFLLC Corpus 開発プロジェクトでは, 比較コーパスとして, 収録作文の作成者と同世代のロシア国内大学生のデータを扱っており, トピックも類似している CoRST を利用することを検討している。ロシア語母語話者コーパスと比較することで, 日本語母語学習者における語彙の過剰・過小使用など, 母語話者と非母語話者の言語使用特性の差がさらに明らかになると期待している。

今回は収集された T P K И 作文タスクのデータを使ったコーパスの企画であるが, 同じ T P K И の口頭能力試験における発話データについても, 現在, 15 年間分が収集されており, モノログ, 会話, テキスト要約などの音声データが利用可能である。将来的に作文データでの運用が軌道に乗った時点で, JRFLLC Corpus に音声データ及び文字化データを追加していくことが考えられる。作文データと並行して発話データを収録することで, 「書く」活動, 「話す」活動それぞれにおいて学習者の習得プロセスにどのような差が観察されるかなど, 教育面のみならず言語学的にも興味深い実証研究を行うことができるようになるであろう。

JRFLLC RLC Subcorpus

Японский учебный корпус русского как иностранного, или JRFLLC, является результатом коллаборации исследователей в области лингвистики и преподавания – Хаясида Риэ (Осацкий университет) и Рахилиной Екатерины (НИУ ВШЭ).

Данный проект реализован при финансовой поддержке Японского общества содействия науке (JSPS) в форме гранта на научные исследования (№ проекта – 17K02926).

Что такое JRFLLC?

JRFLLC представляет из себя учебный корпус, содержащий образцы речи учащихся, изучающих русский язык и чьим родным языком при этом является японский. Данный корпус позволяет увидеть разнообразные языковые особенности порождения русской речи японскими учащимися. Мы также надеемся, что анализ данных особенностей поможет расширить возможности на пути к пересмотру и совершенствованию содержания

обучающих материалов, методических пособий, а также учебных программ, используемых в реальной преподавательской практике в Японии. Кроме того, поскольку JRFLLC является подкорпусом Русского учебного корпуса (RLC), благодаря чему позволяет проводить точный и практически подтверждаемый анализ того, какие схожие черты и отличия наблюдаются в особенностях порождаемой русской речи и особенностях языковой стратегии учащихся, говорящих на разных языках, мы также смеем полагать, что в будущем корпус окажется полезным в том числе и с точки зрения типологических исследований.

Корпус JEFLL составлен из данных, полученных из сочинений студентов специальности «Русский язык» японских университетов. Эти данные были получены по результатам компонента «Письмо» теста по русскому языку как иностранному (The Test of Russian as a Foreign Language, TORFL) для двух уровней – A2 (Basic level), B1 (Intermediate level), который проводится Министерством образования и науки РФ на протяжении 15 лет и основывается на стандартах CEFR (Common European Framework of Reference for Languages : Learning, teaching, assessment). На сегодняшний день доступно около 2800 сочинений от порядка 1400 человек. Пока метаинформацией (часть речи, неправильное употребление и т.д.) снабжена только часть их тех сочинений, которые представлены в открытом доступе, однако работа над корпусом продолжается, и уже в самом ближайшем будущем мы планируем снабдить метаинформацией абсолютно все имеющиеся данные. Кроме того, объем самих данных впредь планируется увеличивать с каждым новым экзаменом в году.

Метаданные

Несмотря на то, что JRFLLC изначально был разработан с целью удовлетворять локальные (местные) нужды, богатство критериев, согласно которым он был спроектирован, порождает разнообразные возможности для поиска ответов на исследовательские вопросы, а также для анализа. Каждый текст снабжается следующими метаданными:

1. фамилия, имя студента (псевдоним),
2. пол,
3. период изучения языка ранее, опыт владения им,
4. языковой уровень студента,
5. временная отметка (на каком курсе было написано данное сочинение),
6. ограничение по времени, в течение которого должно было быть написано данное сочинение,
7. тип текста,
8. функция текста (повествование, аргументирование и т.д.).

Все эти категории для каждого текста в корпусе отражены в Header Identification Box (Header ID). Возможно автоматическое создание подкорпуса на основе этих данных.

Все слова в JRFLLC снабжены морфологическими тэгами, которые содержат информацию о части речи, поле, падеже, виде. В случае, если слово было написано с ошибкой, тэг будет содержать пометку «нестандартная форма» (т.е. неправильная). Интерфейс корпуса допускает возможность поиска по грамматическим и лексическим категориям.

Использование корпуса

Дальнейшее наполнение корпуса JRFLC позволит искать и мгновенно находить общие тенденции в неправильном и сложном для «неносителей» языка склонении имен (существительных и т.д.) и спряжении глаголов для уровней А1-А2, а также поможет нарисовать реальную языковую картину, которую в тот или иной момент рисует в своей голове студент, изучающий русский язык как иностранный. Кроме того, в рамках проекта по разработке корпуса JRFLC рассматривается также вопрос использования в качестве сравнения корпуса CoRST, где собраны данные сочинений русскоговорящих учащихся-ровесников японских студентов, а темы схожи с теми, что представлены в Японии. Мы надеемся, что сравнение с данными корпуса носителей языка сделает возможным проследить разницу в особенностях использования языка его носителями и «неносителями», в частности чрезмерное (и, наоборот, дефицитное) использование японскоговорящими студентами той или иной русскоязычной лексики.

Корпус изначально планировался как хранилище данных по результатам заданий «Письмо» экзамена ТРКИ, однако по факту на протяжении последних 15 лет собиралась и накапливалась также информация по устным высказываниям студентов в рамках ТРКИ (часть «Говорение»), в результате чего на сегодняшний день доступны аудиозаписи с монологами, диалогами, пересказом прочитанного текста и т.д. В ближайшем будущем, когда использование письменных текстов корпуса встанет на рельсы, мы планируем добавить к корпусу JRFLC также аудиоданные и данные культурного характера. Мы полагаем, что размещение данных устной речи параллельно с данными письменной – поможет проводить более глубокие и подтверждаемые на практике (не только педагогические, но и лингвистические) исследования того, чем отличаются друг от друга процессы изучения студентами русского языка в обеих сферах – письмо и говорение.

Материалы JRFLC распространяются бесплатно через поисковые системы сети интернет. Если вы использовали корпус JRFLC в своей научной работе, вы обязаны указать это в готовом докладе, статье и т.д.

2) 収録予定である研究代表者所属機関が実施している TORFL 試験, 1,2 年次における A2 (Basic level), B1 (Intermediate level) レベルの作文試験結果, 約 1,400 人分, 約 2,800 編の作文のうち, かなりの部分を HSE との協同作業でデータ電子化を行った. また,

3) 収集したデータのアノテーションに向け, データテキスト及び学習者の属性情報タグ, 品詞情報タグ, 誤用情報タグの分類・構成について, 前年度の調査・ヒアリングで得た知見を基に検討し, アノテーション・ガイドライン試案の設計を進めた. すでに RLC 本体データとしてアップロードされているものについては, RLC が採用している綴り字, 語彙, 構造, 形態, 統語, その他という大分類, さらに下位分類からなる誤用タグ (Rakhilina et al. 2016) での検索が可能な状態へとアノテーション作業を進めた. 同時に, 日本人学習者に特有の間違いなどに対応した誤用タグを追加することについて可能性を検討した.

【2019 年度】

1) アノテーション・ガイドライン試案に基づき, すでに電子化が完了しているデータについて, テキスト及び学習者の属性情報タグ, 品詞情報タグ, 誤用情報タグのタグ付け作業を進め, JRFLC Corpus パイロット版を作成, オンライン公開を行った (<http://www.web-corpora.net/RLC/jrflc>). タグ付けについては, ロシアで開発・公開されている品詞タグ付けシステム Mystem (<https://tech.yandex.ru/mystem/>) を利用し, ロシア側研究チームが主体となり RLC での基準により誤用情報

タグの追加付加作業が行われた。

2) 作成された JRFL Corpus パイロット版を使い、A2-B1 レベルでとりわけ習得に負荷がかかるとされるアスペクト、ヴォイス、さらに従属複文の3領域において、学習者言語の傾向、特徴を試験的に分析し、パイロット版におけるタグ分類・構成上の問題点や技術面での問題点を抽出した。

3) 2019年12月にロシア語教育研究集会において研究成果報告会を開催、英語、日本語コーパス研究において豊富な経験と実績を持つ石川慎一郎氏（神戸大学）を招き、「学習者コーパスの歴史：学習者のL2使用を総体としてとらえるために」と題する講演会を実施した。講演では石川氏の経験をふまえた学習者コーパスの設計・構築・公開・分析にかかる諸問題について、詳細かつ具体的な事例とその対処法が展開され、フロアとの活発なやり取りが行われた。

4) 最終研究報告書を上梓した。

学習者コーパスは、非誘導型・大規模データという特性によって、学習者言語における特定言語素材の過剰・過少使用や母語転移、回避方略などの客観的特徴を明らかにし得る有効な手段である。3年間の研究・調査活動、コーパス構築作業により、ロシア語学習における日本語母語学習者の学習者言語の傾向や特性に関して、より客観的・実証的データに基づいた分析に道を開き、教育現場における教材開発や指導案作成等に対し、これまでになく有効かつ信頼性に足る情報の還元が展望できるようになった。

以下、JRFL Corpus パイロット版を使ったアスペクト、ヴォイス、従属複文の3領域における、学習者言語傾向・特徴の試験的分析を概観する。

III. JRFL Corpus 分析の可能性と意義

III-1. アスペクト領域における Japanese RFL 学習者 (A2-B1 レベル) 言語の特徴

ロシア語においては、大多数の動詞が完了体／不完了体というアスペクトの文法的表示を担う2項対立システムに組み込まれている。そういったアスペクト体系を持たない言語を母語とする学習者にとって、この現象の存在は習得上の高い壁になることが一般的には予想される。

今回の JRFL Corpus パイロット版による試験分析によって、未来表現におけるアスペクト領域の誤用、特に完了体人称変化形に対する誤用の割合が収録データの約30%に観察されることが分かった。

現代ロシア語において未来の出来事表現では、不完了体合成未来形は基本的には途中経過と多回の強調表現として機能しており、限定された使用領域しかもたない。筆者の過去の分析データでも、完了体人称変化形と不完了体未来形の出現度数の割合は約8:1となっている。しかしながら授業現場では、形態的に習得が容易であるとされる不完了体合成形が最初に導入され、完了体人称変化形による未来表現は習得に負荷がかかると判断され、後回しにされる。恐らく不完了体合成未来形が学習の初期段階で導入されることで、その後、完了体人称変化形が導入された後も、未来表現において学習者はすでに習得済みの、よりたやすく再生できる不完了体合成未来形をもっぱら選択するという現象が起こっていると考えられるのである。

試験的分析でのデータの計量的分析によって初めて、1) 未来表現のアスペクト領域での誤用出現

に極端な偏りがあるという事実が明らかになり、データが示す誤用に関して、単にこの時制でのアスペクト形態・機能の未習得という点にとどまらず、2) 不完了体合成未来形の過剰使用、さらには3) 学習者の完了体人称変化形に対する回避方略の可能性が見えてきたのである。

III-2. ヴォイス領域における Japanese RFL 学習者 (A2-B1 レベル) 言語の特徴

A2-B1 レベルのロシア語学習において、ヴォイス関係、特に受動表現並びに再帰動詞を用いた中動相表現は、アスペクトと並んで重要な習得項目である。しかしながら今回の試験的分析結果では、設題テキスト内で使用されていた表現をそのまま書き写したものの以外で、学習者自身が自律的に産出した受動文表現はごくわずかしか観察されなかった。

受動文表現が未出現であるという事実について、項目が未習得なのか、それともコーパスサイズやトピックの制約によるものかを特定することは、コーパスからだけではできない。該当の設題トピックに関する発話・作文の際に、そもそも話者のイメージ形成において受動を選択する動機が存在するかどうかの一つの手がかりとなる。その検証のためにロシア国内大学生の作文データを収録した CoRST (Corpus of Russian Student Texts, https://www.hse.ru/org/hse/cfi/corpora/krut_http://web-corpora.net/CoRST/search/index.php?interface_language=ru) で類似トピックデータが抽出されるような検索条件を設定し、ロシア語母語話者作文との比較を試みた。その結果、トピックやテキスト種類、テキスト分量 (CoRST では1編の語数情報等は表示されない) 等で制約条件が異なっており、あくまで目安としての比較であるが、類似トピックにおいて母語話者の分詞形受動文使用頻度は相当高い (1編平均、約5例) と言え、日本語母語学習者データとの差異が際立つ。

また、コーパスでデータが出現しなかった項目については、その理由を特定するための補完的な実験調査が不可欠である。

受動文使用について、学習時間約350時間 (A2 レベルテスト合格者) の学習者を対象に別の機会に文法性判断テスト実施した。そこでの誤用率は60% 近い数値を示し、先の CoRST の数値も踏まえれば、上記 TPKI-1(B1) 作文タスク・データで受動文の自律的使用がほとんど観察されないという事実が新たな様相を帯びる。つまり、問題となる現象は単なるコーパスサイズやトピックの制約によるものではなく、原因として学習者の回避方略の可能性が示唆されるのである。

誘導型の文法性判断テストでは、学習者の半数以上が未解答や誤用、さらには再帰形の過剰一般化といった結果を出す。B1 レベルテスト一学習時間540時間の学習者対象一における非統制の作文タスクでは、ほとんどの学習者があまり自信のない受動文表現を回避し、能動文などで置き換えて表現するという行動をとっていると考えられる。

受動文表現は事実伝達という側面からは、一見、能動文や不定人称文と同内容の意義を持つと考えられ、それらで代替するという回避方略がとりやすい項目とされる。ただ、受動文表現にはテキスト結束性や主題選択にかかわる重要な機能があり、さらには分詞形受動文の場合は、伝達される事実内容そのものも能動文や不定人称文とは異なるケースが多い。その点を考慮すると、受動文表現を射程したアカデミック・ライティングの特別な訓練など、学習者の受動文回避克服のための方策の必要性が浮上してくる。こうした方向性は、文法性判断テスト結果の分析だけでは導き出しにくいものであり、コーパス・データ分析と統合することで見出すことが可能になる。

III-3. 従属複文における Japanese RFL 学習者 (A2-B1 レベル) 言語の特徴

アспект, ヴォイスと並んで B1(Intermediate level)レベルでの重要習得項目である従属複文についても, 試験的分析により学習者の言語特性を分析した. その結果, 次のような特徴が観察された.

- 1) まず, 観察された従属複文の使用総数は 1427 例に対し, 誤用総数は 232 例で, 使用総数の約 16% という以外に低い数値を示した.
- 2) 問題文での従属複文使用の量的・質的差が, 学習者作文における従属複文使用に影響を与えていると思われる傾向が観察された.
- 3) 誤用は言語間で論理の枠組みにズレが観察される場合に特徴的に起こっている.

1) については従属複文で表現される文の論理構成が, 一部の例外を除いて言語によってあまり異同が大きくないことが原因と考えられる. また形態面でも, 接続詞は比較的複雑ではなく, この点でもその他の文法項目に比べて誤用数が少なくなっていると考えられる. 従属複文の意味関係は語彙などとは異なり, 人間に共通のより抽象度の高い論理的意味内容を表し, 社会文化的な違いがあまり反映されず, 言語間での異同が少ないということになる. したがって成人における言語習得の場合, 母語ですでに形成された抽象次元での意味的論理力は, ある程度はそのまま学習言語使用の場合にも援用できているということが示唆されるのである.

従属複文は習得にさほど負荷がかからないという今回の分析結果を踏まえれば, 成人の場合の異言語学習では, より大胆に学習の初期段階から複文項目を使用した教材, タスクを導入できるのではという仮説が成り立つ. そういった方略を採用すれば, 学習者の知的レベル, 興味関心に見合った内容のテキスト等をインプットすることもでき, さらにはそれを踏まえて, 学習者が表現したい内容のアウトプットへとつなげられる. 学習者言語が教えてくれるこうした傾向は, 学習活動にとって軽視できない重要な情報を含んでいる.

2) の問題文で使用されている複文タイプについて, 学習者作文で目立って使用例が増えるという現象は, 統語的プライミング効果 (Syntactic Priming; SP) による影響によって生じていると考えられる. 問題文中の文をそのまま繰り返すのではなく, 問題文で得た情報としての構文タイプを, 異なる語彙・文脈で自らのアウトプットに無意識に利用するというこの傾向は, 学習プロセスにおいて活用できる重要な点として留意する必要があるだろう. すなわち, 学習活動においてコミュニケーション実践の単なる量的な積み重ねだけではなく, 文法や構文等, ターゲットとなる項目を含み, 且つ内容のある事前インプットを意識的に導入し, 直後に学習者による自律的アウトプット活動を促すことで, 学習者の表現域の拡大, より精度の高いアウトプットが期待できる可能性が示唆される.

今回の分析データ上で特徴的な誤用は, 条件節における言語間での論理枠組みの差に起因する誤用, 説明の従属複文における接続詞 *что/как* の意味的誤用であり, いずれも 3) で示したように, 言語間で論理の枠組みにズレが観察される場合に起こっていることがわかる.

IV. 今後の課題と将来的展望

JRFL Corpus 開発プロジェクトの残された課題として, ロシア語母語話者, 日本語母語話者の双方から収集された作文を比較コーパスとして利用するという点がある. 比較コーパスを利用すれば,

学習者の学習言語特性に母語や学習環境など、どのようなファクターが反映しているのか、あるいはしていないのかなどについて、実証データによってそのヒントが与えられる可能性も出てくる。開発プロジェクトでは、日本語を母語とするロシア語学習者に、収録予定の作文データと同じトピックで、かつ同条件で母語による作文作成を依頼し、収集されたデータを比較コーパスとする予定である。また、参照コーパスとして、収録作文の作成者と同世代のロシア国内大学生のデータを扱っており、トピックも類似している上記の CoRST を利用することも可能である。

また、今回のプロジェクトでは収集された ТРКИ 作文タスクのデータを使ったコーパス構築であったが、同じ ТРКИ の口頭能力試験における発話データについても、現在、15年間分が収集されており、モノローグ、会話、テキスト要約などの音声データが利用可能である。将来的に作文データでの運用が軌道に乗った時点で、JRFL Corpus に音声データ及び文字化データを追加していくことが考えられる。作文データと並行して発話データを収録することで、「書く」活動、「話す」活動それぞれにおいて学習者の習得プロセスにどのような差が観察されるかなど、教育面のみならず言語学的にも興味深い実証研究を行うことができるようになるであろう。

さらには習得レベル別の定量比較分析や同一学習者の継時的変化を観察できる縦断的コーパスとしての活用も視野に入れることができる。

JRFL Corpus の開発・構築によって、今回の試験的分析でも明らかなように、ロシア語学習における日本語母語話者特有の傾向を分析することが可能となった。その結果を単なる分析に終わらせず、今後、教材や指導案、カリキュラムについての見直し、改良へといかにつなげていけるか、学習者コーパスの真価がまさにその点において問われることになるであろう。

参考文献

- 林田 理恵 2017. 「ロシア語学習者コーパス構築の可能性と意義」『言語文化研究』43号, 大阪大学言語文化研究科.
- 林田 理恵 2018. 「学習者言語を探る－科研費プロジェクト「日本語母語学習者データに基づくロシア語学習者コーパス構築の基盤研究」2017年度研究成果報告総括－」『日本語母語学習者データに基づくロシア語学習者コーパス構築の基盤研究・2017年度研究成果報告』(http://kyoiku.ru.org/news_all/kaken20190327-13/).
- 林田 理恵 2019. 「ロシア語の学習者言語を探る－A2-B1 レベル学習者の従属複文使用」『ロシア語教育研究』第10号.

参照サイト

- <https://ling.hse.ru/> (最終閲覧日：2020年3月20日)
- <http://web-corpora.net/RLC> (最終閲覧日：2020年3月20日)
- <http://web-corpora.net/RussianLearnerCorpus/search/> (最終閲覧日：2020年3月20日)
- <https://tech.yandex.ru/mystem/> (最終閲覧日：2020年3月20日)
- <http://www.ruscorpora.ru/search-main.html> (最終閲覧日：2020年3月20日)

Japanese Russian as Foreign Language Learner Corpus の外国語教育への活用

佐山 豪太

(上智大学外国語学部)

0. はじめに

本科研¹は、National Research University Higher School of Economics（以下、HSE）が運営する Russian Learner Corpus²（以下、RLC）の検索システムを借用し、日本人ロシア語学習者の作文を収集したコーパス（Japanese Russian as Foreign Language Learner Corpus / 以下、JRFLLC）を完成させることを目的としている。

JRFLLC の設計に関しては、ゼロからコーパスの検索システムやプラットフォームを作成するのではなく、すでに HSE で作成・公開しているものに日本人学習者の作文をアップロードするという形をとる。JRFLLC は、言語学・ロシア語教育を専門とする林田理恵（大阪大学）と E. ラリーヒナ（National Research University, Higher School of Economics）の共同研究によって作成された³。

石川（2008: 202）が述べているように、「学習者コーパスは、外国語教育，中間言語教育，学習者研究のいずれの見地からも大いに活用しうる」が、英語に関してはこれまで日本人学習者コーパスは数多く作成・整備されてきた（cf. 赤野他 2014）。実際，学習者コーパスの研究結果は，辞書の編纂や教科書の作成を含む外国語教育の分野で大規模に利用されている（cf. マケナリー，ハーディー 2014: 124）。

一方で，高度な研究に耐え得るロシア語学習者コーパスは存在していなかった。そのため，JRFLLC の分析を通じて，これまで現場の教員が経験則で感じていた学習者の典型的な間違い，日本人ロシア語学習者に観察される中間言語の有り様を客観的に確認できることが期待される。

1. RLC と JRFLLC の概要

JRFLLC は，大阪大学でロシア語を専攻する大学生の作文データで構成されている。これは約 15 年間にわたって蓄積された CEFR (Common European Framework of Reference for Languages : Learning, teaching, assessment) 基準による「外国人のためのロシア語検定試験」，A2 (Basic level), B1 (Intermediate level) の作文試験結果のデータである。JRFLLC のテキストが保存されている母体，RLC にはロシア語学習者のテキストが検索可能な状態でアップロードされているが，他言語の母語話者によるロシア語作文も含まれている。

¹ 「日本語母語学習者データに基づくロシア語学習者コーパス構築の基盤研究」科研費 基盤(C) (2017, 4-2020, 3) 課題番号 17K02926.

² URL 以下の通りである: http://web-corpora.net/learner_corpus/search/

³ 本研究は日本学術振興会 (JSPS) 科学研究費助成事業による助成金 (学術研究助成基金助成金・基盤研究 C-課題番号 17K02926) を受けている。

表 1. RLC に含まれるテキストの全体 (アクセス日 : 2019 年 6 月 22 日)

	学習者の母語	テキスト数		学習者の母語	テキスト数
1	カザフ語	503	8	フランス語	528
2	スウェーデン語	178	10	オランダ語	18
3	ノルウェー語	28	11	フィンランド語	1,231
4	イタリア語	115	12	中国語	54
5	英語	3,148	13	ドイツ語	284
6	韓国語	203	14	セルビア語	16
7	日本語	1,572	15	不明	310

RLC (JRFLLC) の検索システムは、ロシア語ナショナルコーパスといった通常のコーパスに採用されているそれとほぼ同じである。

図 1. RLC (JRFLLC) の検索画面

ただし、上記の通常の検索に加えて、RLC には誤用タグ (тэги ошибок) の検索が実装されており、これが他のコーパスと一線を画す、学習者コーパスに特徴的な要素である。

図 2. RLC (JRFLLC) の誤用タグ一覧

JRFLLC はこの RLC の検索システムをそのまま借用するが、その理由は次の 2 点に基づく決定である：(1) 当システムはこれまで議論を重ねて改変されてきた経緯をもつ。そのため、現在の検索システムや誤用タグは言語学的な根拠に基づいて設定されている、(2) 他言語の母語話者との誤用の比較が可能となる。

本科研は、HSE 側に日本人ロシア語学習者（約 1,700 名分）の作文のテキストを提供しているが、現在、電子化されて RLC 上にアップロードされているものはその一部に限られる。手書きの作文を人力で電子テキストに変換しているため、残りの全てが RLC にアップロードされるにはまだ時間がかかると思われる。さらに、誤用タグ付けの作業は現在進行中であるが、ネイティヴスピーカーによって一つ一つ誤用とそのタイプを確認していかなければいけないため、かなりの時間を要する。ただ、JRFLLC のテキストに誤用タグが付与された場合、例えば、日本語母語話者の中間言語の分析を客観的に行うことが可能になる。

2. 中間言語分析⁴

中間言語とは、母語から目標言語に移行する過程に出現する言語システムを指す。学習者は L2 の規則を常に修正していくため、中間言語は可変性があり、また、それに基づいて言語を使用するという点で体系性があると言える（小池、河野（編） 2003: 151）。

学習者の中間言語に特徴的なのは、母語話者とは異なる逸脱的な言語特徴が観察されることである。その一部は学習が進んでも修正されずに、L2 の言語システムに残り続けてしまうが、この現象を化石化と言う。中間言語分析では、学習者独自の個別的な言語システムに固定化されている状態（化石化）を研究対象とする。また、中間言語分析が外国語教育へもたらした変化としては、学習者の L2 運用を本格的に研究テーマに設定したことと、中間言語に関して 5 つの心的過程を提唱したことが挙げられる。

(1) 学習者の L2 習得における 5 つの過程 (Selinker 1972)⁵

- a. 過剰一般化
- b. 言語転移
- c. 訓練転移
- d. 学習方略
- e. コミュニケーション方略

a. は、学習者が新規に習った文法・形態ルールを、それが適用される範囲を超えて使用してしまうこと（例：-ed を go に付けて *goed という形態を作り出してしまう）。

b. は、L1 のパターンを L2 に持ち込んで使用する現象を指す（例：開音節である日本語の母語話者が英語を発音する際、子音連続に母音を挿入してしまう場合がある。また、ロシア語も閉音節の構造をもつ言語であるため、日本人学習者の発話においてこの負の転移は観察される）。

⁴ 中間言語の解説は、主に石川 (2017: 59-64) を参照している。

⁵ (1) の解説は、小池、河野（編）(2003: 151-152) と石川 (2017: 61-62) を参照している。

c. は、教師や教材が教育的配慮を優先して誤った、もしくは不自然な L2 言語モデルを提示したり、そうした内容のドリルを実施する、教師があいまいな説明をするなどが原因で、L2 習得に悪影響を及ぼすことを指す（例：ロシア語の会話の授業において問いに答える際に、答えだけでなく問いの部分を繰り返すという練習が観察されるが（“Почему вы изучаете русский язык?”という問いに対して、“Я изучаю русский язык, потому что...”と答えてしまう。実際の発話では потому что 以降から始めるのが一般的であろう）、教育的な配慮から不要な部分も学習者に発話させる。

d. は学習者が自身の L2 学習を成功させるために特定の方略を使用することを指す（例：日本語学習者が「～のほうがいい」をチャンクで覚えて、「*飲んだのほうがいい」という誤用を産出してしまう）。

e. は既存の言語知識が不十分であっても、L2 で学習者がコミュニケーションをとらなければならない場合に、そのギャップを埋め合わせるために使用される手段を指す（言い換え、身振り、回避など / 英語の完了相が苦手な日本人学習者が過去形しか使用しない、など）。

上記で言及した中間言語の現象を分析する際に、客観的なデータを示してくれるのが、学習者コーパスである。

3. JRFLLC を用いた分析

学習者コーパスは、誤用タグを用いて学習者のミス傾向を掴む際に効果的なデータを提示してくれる。ただ、前述の通り、現在、それを可能にするタグ付けは JRFLLC 内で完了していない。そのため、ここでは教員が経験則から感じている学習者の典型的な誤用をいくつか提示し、その客観的な確認方法として JRFLLC のデータが活用できることを参考までにいくつかの例を提示する⁶。

принять の現在変化は第1変化に属すが、その現在語幹である прим- は不定形語幹とは形態が異なる。だが、делать などの最も基本的な変化（現在語幹は дела-）を принять にも過剰に適用して、*приняю という誤った形態を作り出すと考えられる（(1) の過剰一般化に相当）。実際、JRFLLC では以下の用例が確認できる。

(2) * Или ты не знаешь тот, кто приняю участие?

「もしくは、君は参加する人を知らないのですか？」

(JRFLLC より引用：アクセス日 2020 年 3 月 20 日)

なお、現代ロシア語には нять という語根は存在しないが (Исаченко 1960: 142-143), обнять, снять, поднять など数多くの高頻度語内に含まれている。そのため、このタイプの動詞の変化は学習者にとって重要であると言える。

ロシア語では不完了体動詞の未来を表す際に、「быть+不完了体動詞の不定形」という分析的な形を用いる。日本人学習者は、不完了体動詞の未来形を作る際、誤って「быть+完了体動詞の不定形」を用いてしまう場合が観察される。以下に、JRFLLC で確認された例を挙げる。

(3) * Я не буду сказать нечего плохого о Сколково.

⁶ ここで挙げる例は、(2) と(6) 以外は 2017 年度の本科研の報告書でも言及している。これらの例を体系的に、大規模に調査するためにも、誤用タグ付けの作業を完了させることが急務である。

「私はスכולコヴァについて何も悪いことは言わないだろう」

(JRFLLC より引用：アクセス日 2020 年 3 月 20 日)

正しくは完了体ではなく不完了体の動詞でなければならないが (буду говорить), 学習者は быть と完了体動詞の不定形を用いた(3)のような例を産出してしまう場合がある。

子音で終わる大半の男性名詞の複数形は、末尾に-ы を付加することで得られる (例：телефон → телефоны)。だが、子音終わりの男性名詞の中には、一般的・規則的な複数形の作り方が適応されないものも存在する。例えば、друг「友人」は子音終わりの男性名詞であり、上記の作り方 (と正書法の規則) を適応すると*други となるが、正しい複数形の形態は друзья である。ただし、子音終わりの男性名詞の複数形の作り方を друг にも用いた други という形態は、JRFLLC 内において数多く確認される。

(4) * Но по-английскому, они - други.

「でも、英語で彼らは友人である」

(JRFLLC より引用：アクセス日 2018 年 3 月 21 日)

同様に、брат「兄弟」が*братьи となっている例も確認される。一般的・規則的な複数形の作り方が適用されない語は、基本語の中に多く含まれているため (例：брат, учитель「教師」, стул「机」など)、このようなミスが散見されるようであれば、教材や授業でその点に言及する必要があると言えよう。

また、ロシア語ではアクセントのない母音 o と e は、それぞれ a と и に近い音として発音されるが、学習者はこれらの母音を書く際にそれぞれを混同しがちである。

(5) * В первом сценарий главная героиня на корабле сидит за писательным столом <...>

「最初のシナリオでは、船においてメインヒロインが机に向かっている」

(JRFLLC より引用：アクセス日 2018 年 3 月 21 日)

ここでは、下線部の「船」の綴りは корабле とするべきであるが、母音の弱化の影響で*корабле となって現れている。基本語の中にもこのようなミスは散見される。

動詞の体は日本人学習者にとって習得が極めて難しい文法カテゴリーである：надо「～する必要がある」は動詞の不定形と結びついて用いられる。だが、否定の助詞 не を含んだ не надо「～する必要がない」となると、基本的には不完了体動詞と結びつく。

(6) * Мне не надо связаться прямо с владельцем.

「私は所有者と直接連絡を取る必要はありません」

(JRFLLC より引用：アクセス日 2020 年 3 月 21 日)

(6) の例では下線部の動詞が不完了体ではなく完了体である。「連絡を取るの是一次である」という考えが念頭にあるため、完了体の方を選んだのであろう。日本人学習者は、この種類の間違いをおかしやすいと考えられる。

4. 今後の課題

今年が最終年度である本科研は、今後、RLC の検索機能やタグ情報を引き継いだ JRFLLC 独自のページを作成し、日本人ロシア語学習者の中間言語の研究が可能になるように、引き続き改良をしていく予定である（日本人ロシア語学習者を想定した誤用タグの作成など）。そのためにも学習者コーパスを用いた研究で最も重要である誤用タグ付けの作業を完了させる必要がある（当作業は、HSE と協力して進行中である）。

参考文献

- 赤野一郎, 投野由紀夫, 堀正広 2014. 『英語教師のためのコーパス活用ガイド』, 東京: 大修館書店.
- 石川慎一郎 2008. 『英語コーパスと言語教育: データとしてのテキスト』, 東京: 大修館書店.
- 石川慎一郎 2017. 『ベーシック応用言語学』, 東京: ひつじ書房.
- 小池生夫, 河野守夫(編). 2003. 『応用言語学事典』, 東京: 研究社.
- マケナリー, T., ハーディー, A. 2014. 『概説コーパス言語学: 手法・理論・実践』, 東京: ひつじ書房.
- Rakhilina, E., Vyrenkova, A., Mustakimova, E., Ladygina, A. and Smirnov, I. 2016. “Building a learner corpus for Russian”, *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC 2016*.
- Selinker, L. (1972) “Interlanguage. “International Review of Applied Linguistics”, *Language Teaching*, 10(3), pp.209-231.
- Исаченко, А.В. 1960. *Грамматический строй русского языка в сопоставлении со словацким. Морфология., Ч.2*, Братислава: Издательство Словацкой Академии наук.
- Рахилина Е. В. 2016. “О новых инструментах описания русской грамматики: корпус ошибок”, *Русский язык за рубежом*. № 3. С. 20-25.