言語研究に適したロシア語コーパス・頻度辞書
～ 均衡コーパスとモニターコーパスの比較を通して ～

Русские корпуса и частотные словари, подходящие для языковых исследований
– На основе сопоставления сбалансированного и мониторного корпусов –

佐山　豪太
Гота Саяма

Аннотация

Данная работа посвящена анализу того, какой корпус или составленный на его основе словарь подходит для языковых исследований, особенно выбору наиболее часто употребляемых слов. Для этой цели мы сопоставляем 2 корпуса разных направлений: Национальный корпус русского языка (англ. *Russian National Corpus*) является сбалансированным корпусом (англ. *balanced corpus*), состоящим из текстов разного характера в определенной пропорции и предоставляет достаточно достоверную информацию о частоте слов в письменном языке. Данный корпус состоит примерно из 92 миллионов словоупотреблений. Sketch Engine представляет собой мониторный корпус (англ. *monitor corpus*), постоянно пополняющийся текстами. На данный момент объем Sketch Engine составляет 14,5 миллиардов словоупотреблений, намного больше, чем Национальный корпус русского языка.

В Национальном корпусе или в его частотном словаре представлен срез всего потенциально бесконечного множества текстов, функционирующих в современном русском языке. В то время как, цель структуры Sketch Engine – собирать как можно больше текстов, не определяя их пропорции.

Тексты мониторного корпуса носят односторонний характер, так как он

состоит преимущественно из текстов, взятых из интернета. Предыдущая литература упоминает о том, что данный дисбаланс исправляется с пополнением текстов. Для подтверждения этого мы сопоставляем Sketch Engine и Национальный корпус русского языка и исследуем, ближе ли первый корпус (мониторный) ко второму (сбалансированному) по наиболее часто употребляемым словам и их частоте.

1. 本稿の背景

言語研究における分析は、主としてテキストや音声などの言語資料に基づいていると言える。例えば、新聞記事・ネット記事、文学作品、学習者の作文といった書き言葉のテキストや、録音した音声を書き起こした話し言葉のテキストは、言語研究において重要な役割を担っている。

コーパス言語学とは「言語研究を行うための一連の分析手順・手段に焦点を当てた領域である」（McEnery and Hardie 2012: 1）。McEnery et al.（2006: 7）が述べているように、「コーパス言語学は、音声学、統語論、意味論、語用論と同じ意味での独立した言語学の分野というよりも方法論」的な側面を持ち、言語研究のあらゆる調査に用いられ得る。したがって、「コーパス言語学はそれ自身で完結する閉じた学問体系というよりも、様々な言語研究分野とゆるやかに連関する学際的な研究分野」（石川 2012: 3）である。

テキストの集合体であるコーパスは、ある言語現象に対してなんらかの証拠を提示してくれるが（Копотев и Мустайоки 2008: 10）、ロシア語の言語研究にコーパスが用いられるようになってすでに久しい。しかし、コーパスに基づくロシア語の頻度辞書や学習用語彙リストは、英語に比べると遥かに数が少ないと言える。これは、コーパスの普及という点でロシア語は英語に大きく遅れていたことに起因する。実際、ロシア語は電子コーパスの公開が遅く、British National Corpus（以下、BNC）に相当する“National”コーパス、すなわち Russian National Corpus（以下、RNC）が一般的に使用できるようになったのは 2004 年であった[1]。

1　なお、BNC の公開は 1995 年 2 月である（cf. 石川 2008: 22）。

ただし、現在ではこの状況は改善され、オンライン検索インターフェイスを有したロシア語コーパスもいくつか存在する。現段階で高度な研究に耐え得るロシア語コーパス（とそれに基づく頻度辞書）は、均衡コーパスである RNC とモニターコーパスである Sketch Engine の 2 つであろう（詳細は 2. で後述する）。両者はコーパス規模や収集されたテキストの性質の点で異なるが、数多くの研究の情報源として用いられている。本稿ではまず両コーパスの概要に言及する。次に、**応用言語学的な観点からどちらのコーパスが学習用の語彙選定に最適か**を考察する。なお、本稿では専攻過程でロシア語を学ぶ大学生を想定している。

2. ロシア語の巨大コーパス

近年では、コーパスをオンラインのデータベース形式で公開する研究者が増えており、その場合、コーパスと検索機能は表裏一体である（投野 2015: 13）。しかし、ただ検索機能が付いていれば良いというわけではなく、高度な言語研究に耐え得るコーパスはその規模が大きくなくてはならない。例えば、コーパスの頻度データに基づいた語彙リストを作成するには小規模コーパスは適していないと考えられる（cf. Саяма 2017）。

現在、学習を念頭に置いた語彙の選定に使用できる頻度データを提供できるコーパスは、その均衡性と規模を考慮した場合、RNC か Sketch Engine の 2 つしかないと考える。ロシア語の言語研究において、RNC は均衡コーパス、Sketch Engine はモニターコーパスの代表格である。

2.1. Russian National Corpus の概要 [2]

RNC[3] は検索機能を有する大規模コーパスである。RNC の作成は、2001 年に始まったロシア科学アカデミー V.V. ヴィノグラードフ名称ロシア語研究所による巨大プロジェクトに端を発する。National という語を冠したコーパスを作成するにあたって、このプロジェクトは英語の British National Corpus を手本とし、規模が巨大で、かつ様々なジャンルのテキストを含んだコーパスの完成を目指した（cf. Плунгян 2005;

2　RNC の概要は、主にПлунгян（2008）, Ляшевская и Шаров（2009: V-XXII）の記述を参照した。
3　RNC の URL は以下の通りである：http://www.ruscorpora.ru/index.html

Ляшевская и Шаров 2009)。そして、RNC は 2004 年 4 月末に一般公開された。

RNC は高度な言語研究に耐え得る検索機能を備えたコーパスである。まず、タイプ（type）やレマ（lemma）単位[4]での検索や、2 語以上からなる語連続の検索も可能である。その際、検索対象である語 A と語 B が、何語離れた状態で共起しているのかも指定して調査できる。また、正規表現（検索条件を絞り込むための特別な表記コード）を使用することで、より詳細な検索も可能である。例えば、任意の文字列を表す「*」を用いて *чита/čita**[5]（*читать/čitat'*「読む」の現在･過去変化の語幹）とすると、*чита/čita* と「ある文字」を含む語が検索できる（*читал/čital*、*читаю/čitaû* など）。

くわえて、RNC では①文法特性、②意味特性、③テキストの追加パラメーター、④語形成といった項目にチェックを入れることで、検索に細かな制限を設定することができる：①文法特性には、品詞、性、格、法、時制、体などの特性が存在する。例えば、*читать/čitat'*「読む」に対して過去形の制限をかけると、*читать/čitat'* の過去形だけが検索結果として表示される。②意味特性では、語彙素の意味的な特徴を検索条件として追加できる。例えば、レマの検索ワードを **ŭmu/jti* とし、「移動」という特徴を追加すると、*УЙТИ/UJTI*「去る」、*ВЫЙТИ/VYJTI*「出る」の各語形がヒットする。③「テキストの追加パラメーター」では、コンマの前後や文の最初 / 最後など、特定の位置において分析対象の語を検索できる。例えば、*кажется/kažetsâ* は、動詞として「見える」を意味する *казаться/kazat'sâ* の 3 人称単数現在の場合と、「らしい」という挿入語の場合がその語形からは想定されるが、「コンマの前」という条件を追加すると、基本的には挿入語の例が検索可能である。④「語形成」では、[接頭辞] や [接尾辞] といった条件が検索に追加できる。例えば、「文法特性」で動詞の完了体を、「語形成」で [接頭辞]、[*y-/u-*] を検索条件に追加すると、接頭辞 *y-/u-* の付いた完了体動詞の検索結果を出力できる。

上記の他にも、コロケーションの検索・頻度の調査をすることも可能

4　タイプとは、同じ語形を 1 語として数える単位で、レマは各語形を 1 つに集約して数える単位である。詳細は石川 (2008) を参照されたい。

5　スラッシュで区切った左右にキリール文字表記とその翻字（ISO 9 方式）を表記する。

である。なお、RNC の検索結果は excel 形式等で出力が可能ではあるが、全てではなく一部の結果しか保存できないといった制限がある。

一般公開から 14 年が経過した 2018 年 11 月現在、RNC は多種多様なコーパスを提供している。

表 1. RNC における各部門の総語数（アクセス日：2018 年 11 月 28 日）

	コーパスの部門	総語数
1	メインコーパス	283,431,966 語
2	統語コーパス	1,031,675 語
3	新聞コーパス	228,521,421 語
4	英露パラレルコーパス	76,759,952 語
5	教育コーパス	664,751 語
6	方言コーパス	285,281 語
7	詩コーパス	10,967,173 語
8	話し言葉コーパス	12,113,491 語
9	アクセントコーパス	31,733,748 語
10	マルチメディアコーパス	4,751,153 語
11	マルチメディア・英露パラレルコーパス	124,104 語
12	古ロシア語コーパス	504,382 語

RNC を代表するメインコーパスは最も総語数が多く、テキストジャンルの比率を考慮して構築された書き言葉均衡コーパス（現代ロシア語を志向）である。

2.1.1. Ляшевская и Шаров（2009）の頻度辞書

Ляшевская и Шаров（2009）の頻度辞書は RNC のメインコーパス（以下、RNC-M）に基づいて作成された。ただ、この頻度辞書が作成されたのは 2009 年であり、その時点での RNC-M の総語数は約 9,200 万語であった。したがって、Ляшевская и Шаров（2009）の頻度辞書は、総語数 9,200 万語の書き言葉均衡コーパス[6] に基づく。

6　均衡コーパスとは、「母集団となる元データの諸特徴を「均衡」的に取り込むことで、母集団全体を「代表」する標本（sample）となるよう意図された」（石川 2012: 21）コーパスである。

一方で、2009 年当時（RNC-M）と 2018 年現在（現行 RNC-M）で大きく総語数が異なることからわかるように、RNC はモニターコーパスとしての側面も有している。つまり、一定のサンプリング比率を保ちながら、RNC の各コーパスは規模を拡大し続けているのである。現在の RNC-M（以下、現行 RNC-M）のコーパス規模は約 2 億 8,300 万語である。Ляшевская и Шаров（2009）が編纂された際の RNC-M は総語数が 9,200 万語であったため[7]、現行 RNC-M は当時の約 3 倍のコーパス規模を有している。ただ、RNC-M と現行 RNC-M は規模こそ違えど、ほぼ同じサンプリング比率で構築されている。したがって、RNC-M は、英語の Corpus of Contemporary American English と同様に、コーパスを構成するテキストジャンルがある一定の比率を保ちながら、総語数を増やし続けるという均衡コーパスとモニターコーパス[8]の両性質を併せ持っていると言えよう。

その総語数は 9,200 万語ではあるが、Ляшевская и Шаров（2009）の頻度データは巨大コーパスから得られたものである（それまで言語研究に使用できた均衡コーパスは、Лённгрен（1993）の 100 万語が最大規模であった）。そのため、Ляшевская и Шаров（2009）の登場は画期的であった。なお、現行 RNC-M の頻度辞書は発表されておらず、その頻度データを包括的に確認することはできない。したがって、現在でも最大規模の均衡コーパスに基づくЛяшевская и Шаров（2009）の情報は利用価値が依然として高い。

2.1.2. RNC-M のコーパス規模とテキストサンプリング

概要でも言及したが、RNC-M は現代ロシア語の断面を再現し、提示する頻度データに信頼性が伴うように設計された（Ляшевская и Шаров 2009: V）。Ляшевская и Шаров（2009: VI）はコーパス規模とテキストサンプリングに関して次のように述べている：「語の生起頻度に関してより

7　なお、Ляшевская и Шаров（2009）の頻度辞書の総語数 9,200 万語とは、句読点などを抜かして数えたものである。句読点を語として換算した場合の総語数は約 1 億 1,500 万語である。

8　モニターコーパスは絶えず規模を拡張し続け、できるだけ多くのデータを収集することを目的としている（石川 2012: 21; McEnery and Hardie 2012: 6）。「モニターコーパスとは、コーパスの内部構成を厳密化せず、その代わりに圧倒的に膨大な量のデータを集めたものであると言えよう」（マケナリー、ハーディー 2014: 9）。

信頼性の高い情報を提供するには、コーパスは規模が大きく、データの包括性において代表的、つまり、ある一定の比率で様々なジャンルや文体のテキストを含んでいなければならない。この点においてロシア語ナショナルコーパスは、British National Corpus<...> などのナショナルコーパスのよい見本に比肩する」。

高頻度語のデータの信頼性はコーパスの質と量で決まる。まず規模に関して言うと、Ляшевская и Шаров（2009）の頻度辞書が基づく RNC-M は、前述の通り 9,200 万語から成る。RNC-M はロシア語の均衡コーパスとしては最大級の規模を誇る。モニターコーパスであれば RNC-M より総語数の多いコーパスは存在するが、均衡コーパスに議論を限定した場合、RNC-M の次に規模が大きく、新しいテキストで構成される Uppsala Corpus（Лённген 1993）ですら総語数が 100 万語にとどまる。

次に、RNC-M を構成するテキストのサンプリング比率を挙げる。

表 2. RNC-M におけるテキストのサンプリング比率[9]

	テキストの機能領域（ジャンル）	比率	総語数	テキスト数
1	芸術文学	39.04%	35,150,521	2,418
2	社会・政治評論	42.21%	39,739,644	27,390
3	芸術文学以外の文献	16.96%[10]	15,478,151	7,495
	— 教育・学術（教育的・学術的に人気のある論文や書籍、教科書、講義等）	— 11.30%		— 3,994
	— 公的・業務文書（法律、法令、声明等）	— 1.62%		— 1,075
	— 電子媒体でのやりとり（メール等）	— 1.49%		— 133
	— 教会・神学	— 1.44%		— 488
	— 広告	— 0.57%		— 1,232
	— 実生活（手紙、日記等）	— 0.48%		— 439
	— 製造技術（解説書、仕様書等）	— 0.26%		— 134
4	パブリックではない口頭会話	0.88%	758,407	1,005
5	その他	0.90%	827,580	61
合計		100%	91,954,303	38,369

9　表 2 はЛяшевская и Шаров (2009: VI) を著者が日本語に訳し、一部加工を加えたものである。なお、「芸術文学以外の文献」の比率の合計は、おそらく計算ミスにより 16,96% にならない（実際は 17,16%）。結果、合計値は 100.19% となり、この誤差は四捨五入によるものとは考えづらいが、表では原文のままとした。

10「—」の後ろの数字は 16.96% の内数を示す。

なお、現行 RNC-M は絶えずテキストが追加されていくため、テキストの正確なサンプリング比率は挙げられない。だが、前述の通り、当コーパスはおおむねЛяшевская и Шаров（2009）で採用された比率に基づいてその規模を拡張しているようである。

2.1.3. RNC-M の語彙リスト

Ляшевская и Шаров（2009）は高頻度 20,000 語（レマ換算）の頻度リストを、頻度順・アルファベット順に記載している。他にも、品詞毎の頻度リスト、タイプ単位の頻度リストが備わっている。ここでは RNC におけるレマ化の規則について言及する。

まず、Ляшевская и Шаров（2009）における語彙の形態的分類はЗализняк（1977）の記述に基づいており、大半の語は一般的なレマ化の規則によって処理される。ただ、その中には例外的なレマ化の規則がいくつか存在する。以下で、レマ化に関わる特筆すべき事項（①体、②特定の語形、③複数の語から成る単位）について触れるが、その内容は主にЛяшевская и Шаров（2009: XIII-XV）の記述や著者が自ら現行 RNC-M の検索機能を用いて確認した事項に則している。

① Ляшевская и Шаров（2009）における完了体・不完了体

頻度の計算に際して、通常、完了体･不完了体は別々の語として扱われ、コーパス準拠の語彙リストに限らず、学習用の語彙リストにおいてそれぞれが 1 つの語として認識されている。例えば、*делать/delat' – сделать/sdelat'*「する」は別の語として認識されている。

② Ляшевская и Шаров（2009）における特定の語形

形動詞・副動詞は基本的に動詞の語形として処理され、元の動詞に集約して頻度が計算されている。例えば、能動形動詞現在 *читающий/citaûšij*「読んでいる」と不完了体副動詞 *читая/citaâ*「読みながら」は、それぞれに見出し語が与えられているのではなく、レマ化に際して元の動詞である *читать/citat'*「読む」にまとめて扱われる。だが、特定の語形が固定的に用いられ、異なる統語的機能を果たしている場合はこの限りではない（例：*говорящий/govorâšij*「話し手」や *трудящийся/trudâšijsâ*

「勤労者」)。

③ Ляшевская и Шаров (2009) における複数の語から成る単位

Ляшевская и Шаров (2009) において、2 語以上の組み合わせにより使用される語連続の単位は存在しない (cf. Ляшевская и др . 2005)。複数の語から成る単位は各構成要素に分解され、それぞれが別々の見出し語を有している。例えば、*может быть/možet byt'*「もしかして」は、*мочь/moč'* と *быть/byt'* に分解されて頻度が計上される。

2.2. Sketch Engine と ruTenTen11 の概要

Sketch Engine はコーパスの分析に必要な、多種多様な機能を提供してくれる。当初、Sketch Engine は英語コーパスのために企画・作成されたが、現在では様々な言語に対応している：Sketch Engine は、英語の enTenTen11 や日本語の jpTenTen11 に加えて、これまで小規模なコーパスしか存在しなかった言語の巨大モニターコーパスも備えている。スラヴ語は、ロシア語 (ruTenTen11) だけでなくウクライナ語、チェコ語、ポーランド語、ブルガリア語などの TenTen シリーズも作成され、web 上で検索可能な状態で公開されている。どのスラヴ語の TenTen シリーズも総語数は 1 億を超えており、言語研究に応用が可能である。したがって、Sketch Engine の登場が言語研究や教材研究にもたらした寄与は大きい。

2.2.1. Sketch Engine の機能

Sketch Engine は、RNC と類似の検索機能にくわえて発展的なコーパス分析ツールを提供しており、かつ、その分析結果のデータを excel や txt ファイルで保存する機能も搭載している。

まず、Word list という機能では Sketch Engine 上で公開されているコーパスを分析対象とし、タイプ、レマなどの単位での頻度分析が可能であり、かつ、その結果をリストとして出力することができる。以下に、ruTenTen11 のデータを基にした Word list を挙げる。

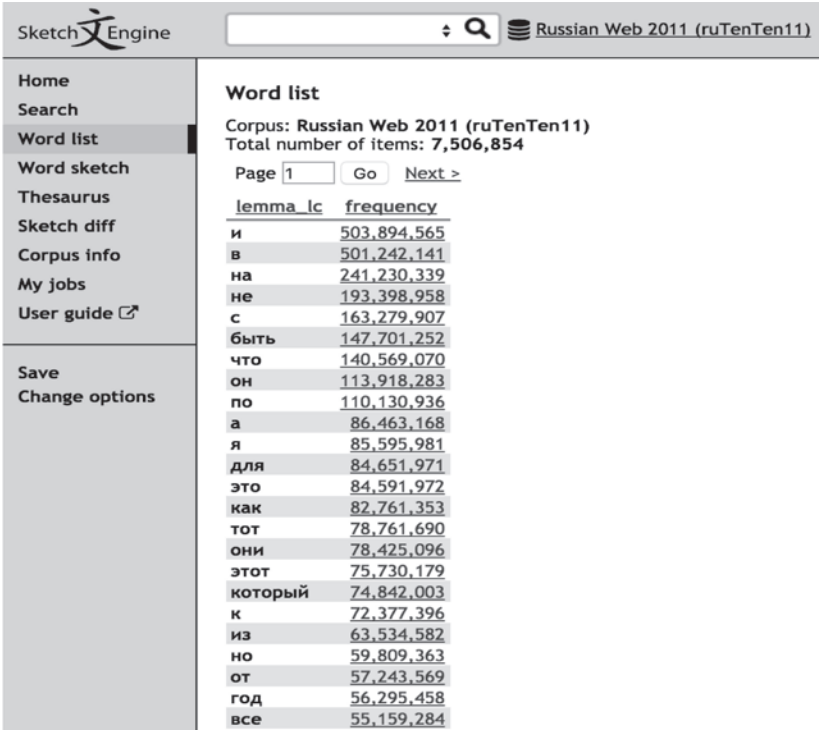

図 1. ruTenTen11 における Word list 機能の出力結果例（語の単位：レマ）

ある語がテキスト内でどのような語と共起しているのかを分析してくれる Word Sketch も、Sketch Engine が提供する有益な機能の 1 つである。例えば *uòmu/idti*「進む、進行する、行われる、降る」という動詞が、ruTenTen11 においてどのような語と共起しているのかを調べると、以下のような結果が得られる。

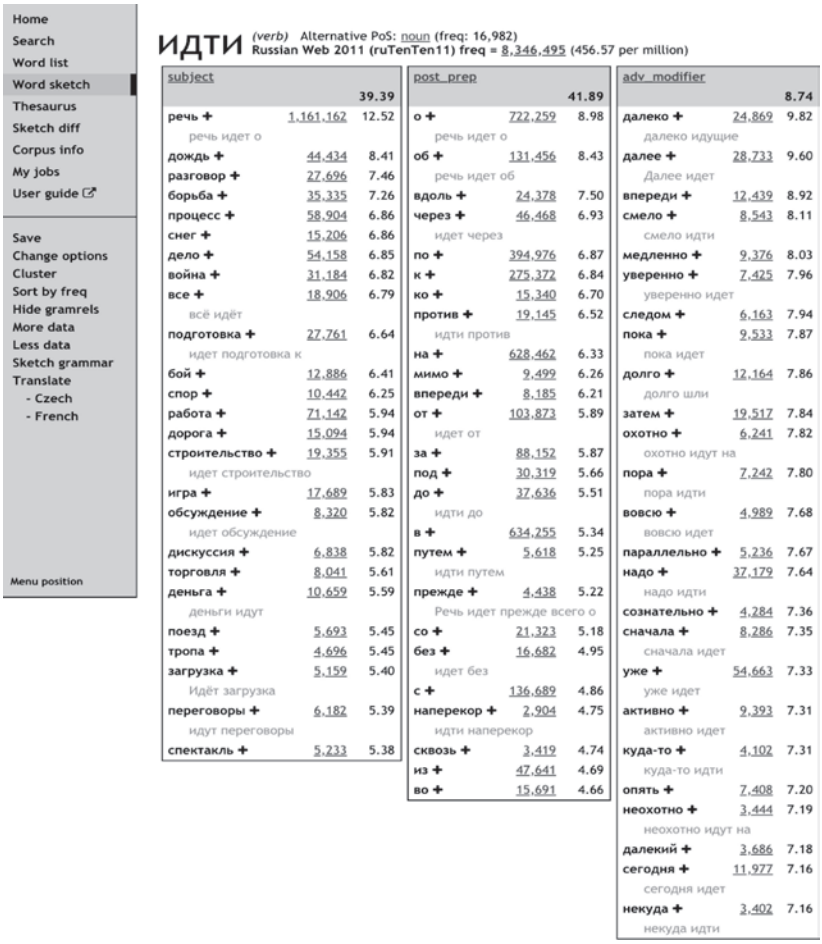

図2. ruTenTen11における Word Sketch 機能の出力結果例(*идти/idti* を例に)

図 2 からは動詞 *идти/idti* がどのような主語と共起しているのかといった情報が得られる。例えば、*идти/idti* は *речь/rec'*,「話」、*дождь/dozhd'*「雨」、*разговор/razgovor*「会話」といった名詞（主語）と共起する頻度が極めて高いことがわかる（この場合、*идти/idti* は「(～について話が) 進む、(雨が)

降る」の意味で生起している)。また、前置詞を分析対象とすると、この動詞は *o/o*「〜について」と頻繁に共起することが確認できる。

Thesaurus と Sketch Difference は、shared triple（共有 3 元）に基づいて統語的に似た振る舞いをする語を抽出する（cf. スルダノヴィチ, 仁科 2008)。例えば、名詞 *учитель/učitel'*「先生」を ruTenTen11 において Thesaurus の分析にかけると、以下のような語群が浮かび上がる。

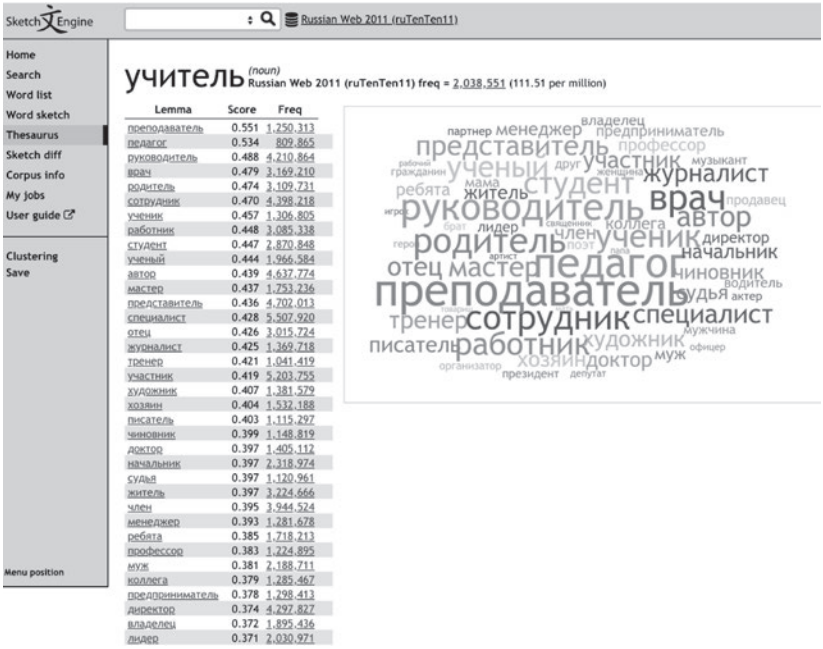

図 3. ruTenTen11 における Thesaurus 機能（*учитель/učitel'* を例に）

Thesaurus の機能は類似度の高さを示すスコアだけでなく、その結果を視覚的に図として提示してくれる。*учитель/uchitel'* と類似度の高い語として *преподаватель/prepodavatel'*「講師」、*педагог/pedagog*「教育者」といった名詞が挙げられる。

このように、Sketch Engine は様々な有益な機能を提供してくれるが、言語研究における Sketch Engine の最大の利点は自作のコーパスを web

上にアップロードし、かつ、そのコーパスに対してここまでに言及した分析機能を活用できる点にある。例えば、自分で作成したコーパスを Sketch Engine にアップロードし、Word list 機能によってその頻度リストを作ることができる。この機能により、より多くの研究者がコーパス言語学的なアプローチを取れるようになった。

2.2.2. ruTenTen11 のコーパス規模とテキストサンプリング

ruTenTen11 は、約 145 億 5,400 万語から成る巨大モニターコーパスである。TenTen シリーズのコーパスは、web をコーパスとして用いる Web as Corpus（cf. Kilgarriff and Grefenstette 2003; McEnery and Hardie 2012）の進化した形である（石川 2012: 18）。

ただし、圧倒的な総語数を誇る ruTenTen11 であるが、コーパスを構成するテキストのサンプリング比率は公表されていないだけでなく、そもそも不透明である。ruTenTen11 のテキストは、SpiderLing というプログラムによって集められたが、これは web 上を自動で巡回して web ページを収集するものである。どのようなジャンル比率でテキストを収集するかを考慮していては、これほど巨大なコーパスは構築できない。テキストのサンプリング比率にこだわらずに、web を介して自動でテキストを収集するからこそ、145 億といった規模のコーパスが構築できたのである。

2.2.3. ruTenTen11 の語彙リスト

現段階では Sketch Engine は、ruTenTen11 に基づく頻度リスト（語彙リスト）を公開していない。前出の Word list 機能を用いれば自ら頻度リストを作成できるが、ruTenTen11 を分析対象とした場合、高頻度 1,000 語までのリストしか入手できないという制限が設定されている（ただし、追加料金を支払えばこの制限の緩和が可能である）。

Sketch Engine のレマ化は、基本的には RNC のそれとほぼ同じ基準を採用している。ただ、別の統語的機能を獲得した特定語形の扱いが他と若干異なる：*может/možet*「かもしれない」、*кажется/kažetsâ*「らしい」といった語形は、RNC-M などでは 1 つの項目として扱われていたが、Sketch Engine のレマ化に際しては元の *мочь/moč'*、*казаться/kazat'sâ* に

まとめられて頻度がカウントされている。

3. 分析

Sketch Engine の ruTenTen11 はコーパス規模の点で他の追随を許さない。ここでは、このコーパスが学習用の高頻度語の選定に使用可能かを考察する。2.2.2. で言及した通り、ruTenTen11 は Web as Corpus を体現したコーパスであり、web 上のサイトからテキストを大量に収集している。そのため、ruTenTen11 は均衡コーパスのように構成要素であるテキストの比率を考慮しておらず、現代ロシア語を代表しているとは言い難い。一方で、モニターコーパスはその規模が大きくなるにつれて、テキストの偏りが是正され、均衡的になるという指摘がある。仮にそうであれば、RNC-M よりも規模が大きい ruTenTen11 を高頻度語の選定に用いる方が良い。

そこで、まず 3.1. にてコーパス規模が高頻度語の選定にどの程度影響を与えるかに言及する。続く 3.2. ではモニターコーパスの ruTenTen11 が、高頻度語の頻度分布の点で均衡コーパスである RNC-M とどの程度類似性を有しているかを調査する。

3.1. コーパス規模と高頻度語の関係

一般的には、規模の大きなコーパスからは安定して信頼性の高い頻度情報が得られるとされる（統計学では一般的に母集団の推定精度は標本サイズの平方根に比例すると言われる（石川 2012: 18））。したがって、コーパスはできるだけ巨大であることが望ましい。

これまでコーパス規模と高頻度語の関係は断片的には言及されてきた：Sinclair（1991: 18）はコーパス規模はできるだけ大きくなくてはならないとし、Kennedy（1998: 68）は 50 ～ 100 万語で高頻度語の調査が可能であると試算している。Biber（1993）の意見では、コーパス規模は調査対象によって確定され、低頻度にしか観察されない言語単位の調査には大規模なコーパスが求められる。Reppen（2010: 55）は、あらゆる分析に適したコーパスサイズは規定できないと述べている（例：言語教育には数百万規模のコーパスより、小規模でも代表性のあるコーパスの方が

適している)。Kilgarriff and Grefenstette (2003: 336) は、1 億語のコーパスでは語彙の大半の生起頻度が 50 を下回るため、これでは統計的に安定した結論を導くことはできないとしている。他にも、語彙の大半は低頻度にしか確認されないため (cf. Zipf 1935)、これらの頻度データの分析にはBNCなどの1億語のコーパスでも不十分であり、数十億規模のコーパスが不可欠であるという指摘もある (Pomikalek et al. 2009: 4-5)。

上述の意見は主に英語を念頭に置いていると考えられる。ロシア語に関して Šteinfeldt (1973: 14) は、40 万語のコーパス規模は高頻度 1,100 語から 1,300 語を抽出するのには十分であると述べている。また、Пиотровский и др. (1972) は数学的な観点から、1,600 ～ 1,700 の高頻度語を選定するのであれば、40 万語のコーパス規模で十分であると試算している。だがЛяшевская и Шаров (2009: VII-VIII) はШтейнфельд (1963)、Засорина (ред.) (1977)、Лённгрен (1993) といったコーパスを意識して、100 万語以下のコーパス規模は頻度の分析には不十分であると述べている。また、Sharoff et al. (2013: 4) も同様に、近年の基準では 100 万語のコーパスは小規模であり、そこから得られる語彙リストの信頼性は低いと述べている。Захаров (2005: 5) とЩипицина (2015: 62) も部分的にコーパス規模に関して言及をしている：前者は Brown Corpus や Uppsala Corpus を例として挙げ、現在ではコーパスの総語数は 100 万語以上でなければならないとし、後者は 1 億語を越えるコーパスは様々な調査に適用可能であると述べている。このように高頻度語の選定に求められるコーパス規模の記述は一致を見ていないが、いずれにしても 100 万語コーパスは現在の基準では小規模であると言える。

なお、Саяма (2017) では、現在では比較的小規模とされる総語数 100 万語のコーパスを分析対象とし、そこから得られる頻度情報を用いてどれだけの高頻度語が安定して選定できるかを調査した。その確認方法として、RNC-M (Ляшевская и Шаров 2009) に含まれるテキストのジャンル比率をそのまま採用し、5 つの 100 万語コーパスを作成した。つまり、5 つの自作コーパスは規模こそ 100 万語と小さいが、内部構造は RNC-M と同じである。次に，RNC-M と自作コーパスの間でどれだけ語彙項目が重複しているかを分析した。9,200 万語の RNC は信頼性が高く、安定した頻度情報を提示してくれる。規模の大きい RNC-M と規模の小さい

100 万語の自作コーパスは内部構造の点では同じであるため、両者を比較した際の語彙重複率の安定性は規模に大きく依存する。結果、5 つすべてのコーパスの分析を通して高頻度 1,500 語までは RNC-M と自作コーパスの語彙重複率は高いことがわかった。すなわち、100 万語コーパスは 1,500 位までの高頻度語を抽出するのに用いることができる。逆に、高頻度 1,500 語以上を選定する際には不十分であると言えよう。高頻度語をどの程度学習者に導入するか次第であるが、上記の結果から、100 万語のコーパスから得られた頻度データは、語彙選定には不十分であると言える。

3.2. モニターコーパスの規模と均衡性の関係

3.1. で述べたように、高頻度語の選定や頻度分析においても求められるコーパス規模も大きいことが望まれる。では､日本人の大学生のように､書き言葉の標準的な現代ロシア語を学ぶ学生を対象として語彙リストを作成する場合、どのコーパスを分析に採用すべきであろうか。この場合、規模の点では ruTenTen11 には劣るが、現代ロシア語のテキストを均衡的に含んだ RNC-M が最適であろう。ruTenTen11 を構成するテキストはインターネットのサイトから抽出したものであるため、そこに含まれる語彙項目や頻度情報は偏っている可能性がある。

ただ、モニターコーパスであっても、その規模が大きくなる過程で次第に偏りが自己解消され、母集団がおのずと均衡的に再現されるという指摘がある（cf. 石川 2012: 40; マケナリー , ハーディー 2014: 9-10)。したがって、巨大モニターコーパスである ruTenTen11 からは、均衡コーパスである RNC-M と類似の頻度情報が得られる可能性がある。もしそうであれば、コーパス規模がより大きく、より多くの用例が確認できる ruTenTen11 を言語研究に用いた方が良いであろう。

それを確認するため、ここでは RNC-M や ruTenTen11 を含む複数のコーパスを分析対象とし、これらの高頻度語の頻度を比較する。分析に用いるコーパスを以下の表に挙げる。

表 3. 主要ロシア語コーパス・頻度辞書の一覧

	コーパス	規模	コーパスの特徴
1	Засорина (ред .) (1977) のコーパス	100 万語	均衡コーパス (現代ロシア語を代表)
2	Лённгрен (1993) の Uppsala Corpus	100 万語	均衡コーパス (現代ロシア語を代表)
3	Ляшевская и Шаров (2009) の RNC-M	9,200 万語	均衡コーパス (現代ロシア語を代表)
4	現行 RNC-M	2 億 8,300 万語	均衡コーパス (現代ロシア語を代表)
5	RNC- 話し言葉コーパス	1,200 万語	話し言葉のテキストによるコーパス
6	Sharoff et al. (2013) の Internet Corpus	1 億 5,000 万語	web ページで構成されたコーパス (語彙学習用に作成 / 学習者が触れるロシア語を代表するように設計)
7	ruTenTen11	145 億 5,400万語	モニターコーパス

上記の通り、Засорина (ред .) (1977)、Лённгрен (1993)、Ляшевская и Шаров (2009) のコーパスは現代ロシア語の書き言葉を代表するように設計された。だが、同じ現代ロシア語を志向していても、採用されたテキストのサンプリング比率はそれぞれ異なるため、結果として得られる頻度情報は異なると推測される。Sharoff et al. (2013) は語彙学習に活用するために作られ、口語に近い個人的なやりとりを多く含んでいるとされる。

頻度の点でこれらのコーパスは互いにどの程度近いのかをクラスター分析 (cluster analysis) で確認する：クラスター分析とは「当初の分類基準が何もないときに、主に量的変数を用いて何らかの対象を幾つかの塊、グループに分類する探索的分析方法である」(小田 2007: 148)。ここでは、7 つのコーパスに共通して高頻度に生起する 50 の内容語の頻度を量的変数とし、分類される対象を各コーパスとした。石川他 (編) (2010: 184) が述べているように、「一般に、基本語頻度はコーパス種別を問わず安定しているとされるが、様々な研究でも示されているように、無作為に選んだ上位 50 語や 100 語であっても、多くの場合、かなり高いテキ

ストの弁別力を持つ」。つまり、これらの語数をもってしてテキスト（コーパス）の分類は可能であると思われる。クラスター分析の結果は以下の通りである。

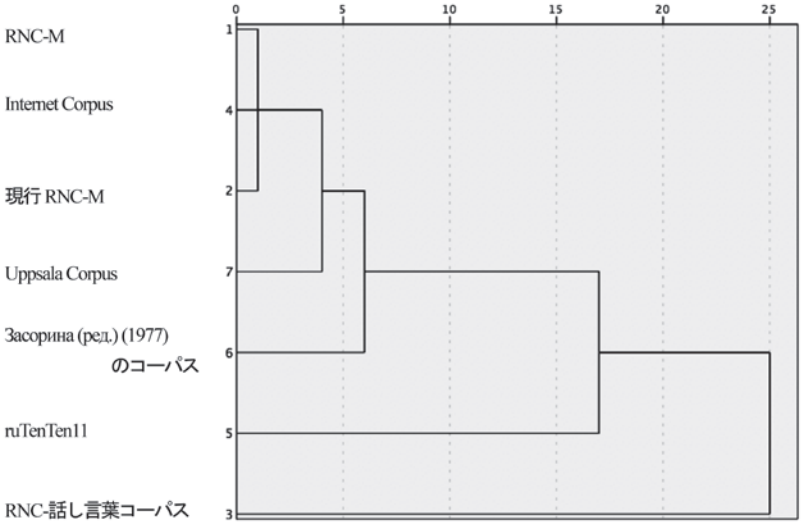

図 4. ロシア語コーパスを対象としたクラスター分析（SPSS: 最遠隣法、相関係数 [11]）

図 4 からは、Ляшевская и Шаров（2009）の RNC-M、Sharoff et al.（2013）の Internet Corpus、そして現行 RNC-M（縦軸の 1.、4.、2.）が早い段階でクラスターを形成していることがわかる（横軸は、分析対象のコーパスがクラスターを結合した際の距離を表す）。また、Лёнгрен（1993）の Uppsala Corpus やЗасорина（ред .）（1977）のコーパスも上記 3 つのコーパスと近い距離にあり、比較的早い段階でクラスターを形成する。RNC-M、現行 RNC-M、Uppsala Corpus、Засорина（ред .）（1977）のコーパスは現代ロシア語を代表するよう設計された書き言葉均衡コーパスである。そのため、これらのコーパス間に高い類似性が確認される

11 高頻度語をケースとした、変数のクラスター化が目的であるため、小田（2007）、水本（2007）、水本 , 野口（2009）、石川他（編）（2010）を参考としてここでは相関係数を使用した。

のは妥当であると言えよう。ただ、Засорина（ред.）（1977）のコーパスが他のコーパスと遠い位置にあるのは、テキストのサンプリング比率に偏りがあったためであると推測される[12]。

なお、Sharoff et al.（2013）は、自身の Internet Corpus は個人的なやりとりを多く含んでおり、RNC-M などの伝統的なコーパスより学習に向いていると述べているが（Sharoff et al. 2013: 5）、クラスター分析からは逆に書き言葉コーパスに近いという結果が得られた。すなわち、Internet Corpus は RNC-M や現行 RNC-M と非常に距離が近く、かつ、RNC- 話し言葉コーパスからはかなり遠くに位置している。

そして、図 4 は巨大モニターコーパス ruTenTen11 が書き言葉均衡コーパスとは高頻度語の頻度の点でかなり異なることを示しており、これらとクラスターを組む段階がかなり遅いことがわかる。したがって、ロシア語の ruTenTen11 と他の書き言葉均衡コーパスを比較した場合、ruTenTen11 において均衡性が実現されているとは言えないと考えられる。

4. 総括

現在、高度な言語研究に使用可能なロシア語のコーパスとして、均衡コーパスでは RNC-M、モニターコーパスでは Sketch Engine を挙げることができる。モニターコーパスは、内部を構成するテキストのジャンル比率を厳密に規定せずに、数多くのテキストを収集することを目的としている。

モニターコーパスはその規模が大きくなるにつれて次第に偏りが自己解消され、母集団がおのずと均衡的に再現されるという意見がある（cf. 石川 2012: 40; マケナリー , ハーディー 2014: 9-10）。したがって、巨大モニターコーパスである ruTenTen11 と均衡コーパスである RNC-M は類似の頻度情報を示す可能性がある。

そこで、本稿は ruTenTen11、RNC-M に加えて、いくつかのロシア語コーパスの高頻度 50 語を分析対象とし、クラスター分析を行なった。そ

12 Засорина（ред.）（1977: 9）では、学術領域のテキストとしてソビエトの著名な研究者の論文（物理学、化学、生物学、鉱物学、歴史学）が採用された。また、社会評論テキストの分野にはレーニンや著名な共産主義者の論文、発表、共産党大会の資料が含まれている。

の結果、ruTenTen11 から得られる頻度情報は、均衡コーパスから得られるそれとはかなり異なることがわかった。したがって、ruTenTen11 に関して言えば、コーパス規模が増えていっても（総語数 145 億語の今の段階では、もしくは本稿の分析からは）、均衡性は実現されてはいないと言える。

参考文献

石川慎一郎 . 2008.『英語コーパスと言語教育 : データとしてのテクスト』, 東京 : 大修館書店 .

石川慎一郎 . 2012.『ベーシックコーパス言語学』, 東京 : ひつじ書房 .

石川慎一郎 , 前田忠彦 , 山崎誠（編）. 2010.『言語研究のための統計入門』, 東京 : くろしお出版 .

小田利勝 . 2007.『ウルトラ・ビギナーのための SPSS による統計解析入門』, 長野 : プレアデス出版 .

スルダノヴィチ , E.I., 仁科喜久子 . 2008.「コーパス検索ツール Sketch Engine の日本語版とその利用方法」『日本語科学』23, 59-80 頁 .

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 .

投野由紀夫 . 2015.「コーパスの英語教育への応用」投野由紀夫（編）『コーパスと英語教育』, 東京 : ひつじ書房 , 1-16 頁 .

水本篤，野口ジュディー . 2009.「多変量解析を用いた PERC コーパスの領域分類」『コーパス言語研究における量的データ処理のための統計手法の概観（統計数理研究所共同研究リポート）』232, 85-106 頁 .

Biber, D. 1993. “Representativeness in corpus design”, *Literary and linguistic computing*, 8(4), pp.243-257.

Kilgarriff, A., Grefenstette, G. 2003, “Introduction to the special issue on web as corpus”, *Computational linguistics*, 29(3), pp.333-347.

Kennedy, G. 1998. *An introduction to corpus linguistics*, London, New York: Longman.

McEnery, T., Hardie, A. 2012. *Corpus linguistics: Method, theory and practice*, Cambridge, Tokyo: Cambridge University Press.

McEnery, T., Xiao, R., Tono, Y. 2006. *Corpus-based language studies: An advanced resource book*, London: Routledge.

Pomikalek, J., Rychly, P., Kilgarrif, A. 2009. "Scaling to billion-plus word corpora", *Advances in computational linguistics. Special issue of research in computing science*, 41, pp.3-14.

Reppen, R. 2010. *Using corpora in the language classroom*, New York: Cambridge University Press.

Sharoff, S., Umanskaya, E., Wilson, J. 2013. *A frequency dictionary of Russian: Core vocabulary for learners (Routledge frequency dictionaries)*, Oxford: Routledge.

Sinclair, J. 1991. *Corpus, concordance, collocation*, Oxford: Oxford University Press.

Šteinfeldt, E. 1973. *Russian word count: 2500 words most commonly used in modern literary Russian: Guide for teachers of Russian*, Moscow: Progress Publishers.

Šteinfeldt, E. 2003. *Russian word count: 2500 words most commonly used in modern literary Russian: Guide for teachers of Russian (reprinted)*, Honolulu, Hawai: University Press of the Pacific.

Zipf, G.K. 1935. *The psycho-biology of language: An introduction to dynamic philology*, Boston: Houghton Mifflin.

Зализняк, А.А. 1977. *Грамматический словарь русского языка: Словоизменение: Около 100,000 слов*, М.: Русский язык.

Засорина, Л.Н. (ред.). 1977. *Частотный словарь русского языка: Около 40000 слов*, М.: Русский язык.

Захаров, В.П. 2005. *Корпусная лингвистика: Учебно-метод. Пособие*, СПб: СПбГУ.

Копотев, М.В., Мустайоки, А. 2008. "Современная корпусная русистика", *Slavica helsingiensia*, 34, С.7-24.

Лёнгрен, Л. 1993. *Частотный словарь современного русского языка. (With a summary in English: A frequency dictionary of modern Russian)*, Uppsala: AUU.

Ляшевская, О.Н., Плунгян, В.А., Сичинава, Д.В. 2005. “О морфологическом стандарте корпуса современного русского языка”, *Научная и техническая информация, сер. 2. Информационные процессы и системы*, 6, С.29.

Ляшевская, О.Н., Шаров, С.А. 2009. *Частотный словарь современного русского языка на материалах Национального корпуса русского языка*, М.: Азбуковник.

Пиотровский, Р.Г., Бектаев, К.Б., Пиотровская, А.А. 1972. *Математическая лингвистика*, М.: Высшая школа.

Плунгян, В.А. 2005. “Зачем нужен Национальный корпус русского языка?”, *Национальный корпус русского языка 2003–2005. Результаты и перспективы*, М.: Индрик, С.6-20.

Савчук, С.О. 2005. “Метатекстовая разметка в Национальном корпусе русского языка: Базовые принципы и основные функции”, *Национальный корпус русского языка: 2003-2005. Результаты и перспективы*, М.: Индрик, С.62-88.

Саяма, Г. 2017. “Влияние объёма корпуса на определение наиболее часто употребляемых слов: Анализ частотных данных из пяти корпусов”, *Русский язык в научном освещении*, 34(1), С.70-91.

Штейнфельд, Э.А. 1963. *Частотный словарь современного русского литературного языка*, Таллин: [s.n.].

Щипицина, Л.Ю. 2015. *Информационные технологии в лингвистике: Учебное пособие. 2-е. изд.*, М.: ФЛИНТА, Наука.