ロシア語の語形変化と語の頻度

Russian Inflection and Word Frequency

上原　順一

Abstract

Russian neuter nouns ending in -мя have a unique declension. Only two of them are introduced in classes of elementary Russian though grammars describe more these words. This may be explained in terms of word frequency. Students of third and fourth grades attending the author's seminar course will learn several kinds of programs or applications and will ensure that elementary knowledge of the language must correlate with word frequency. Similar kinds of problems are concerned with the declension of adjectives. The author hopes that this paper will help senior students of Russian classes to handle data processing and to deepen their knowledge of Russian grammar.

Keywords: Russian, inflection, word, frequency

1. はじめに

筆者は専攻としての「ロシア語」を担当している。1 年生の授業では語形変化を講じることがたびたびある。その一方で、上級生（3、4 年生）の授業「専攻演習」（いわゆるゼミ）などでは、単純なデータ処理を行うことで得られる知識があることを示してきた。本稿では一例として、高頻度の語に注目して、語形変化の現れ方を再確認するプロセスを示したい。これによって、初学者向けには頻出語彙を得られること、その方法論を示すことで上級生を簡単なデータ処理へ誘えることが期待される。読者の環境にも依存するが、基本的には筆者が打ち込んだ命令をそのまま利用できることを想定している。文章は授業メモを整理したものであり、ロシア語の学生向けのロシア語学ないしはロシア語を素材とするデータ処理の教材として利用できるようであれば、教材研究の出発点としたい。文章の区切りには、「学生への課題」を設けた。表示の煩雑さをさけるため、語義の日本訳は可能な限り省略する。

利用したソフトなどは、統計解析で使われるソフト R（R Core Team 2021）、コマンド類は MacOS に標準で用意されているターミナル、その中で動作する grep、sort などである。

2. 名詞

2.1 -мя の名詞

「-мя で終わる中性名詞は独特の変化をする」－これは筆者が 1 年生の授業で繰返し述べてきたことであるし、正しい説明である。おそらく、それまでに学習済みである -я で終わる名詞（この大多数は女性名詞であるが、男性名詞でも女性名詞として変化

する語を含む）との違いを述べる段階である。また、-мя の中性名詞は、-o, -e, -ë なる語尾をもつ中性名詞に対する例外であることも重要である。

しかし、「-мя で終わる中性名詞」という表現は学生に誤解を与えるかもしれない。というのも、この言葉遣いは、-мя の男性名詞や女性名詞がある可能性を排除しないからである。

ロシア語で -мя で終わる語をすべて抽出してその文法的属性を調査するには、ザリズニャクの『文法辞典』（Зализняк А.А. 1980）が便利である。収録語数は約 10 万である。この辞典は、語の変化パターンを記号などで記述する際に、語の並び順がいわゆる逆順になっている。通常の辞書では、語頭の文字、その次の文字などをアルファベット順に並べているが、この文法辞典は語末の文字、その前の文字などを並べる基準にしている。したがって、-мя で終わる語がすべてまとめて記載されることになる。この『文法辞典』によれば、-мя で終わる語は全部で 26 ある。そのうち、名詞が 14 語、他は副詞 12 語の記載がある。なおかつ、その名詞 14 語すべては中性名詞である。このうち、1 語だけは、他とは異なった変化をする語であり、また、標準的な文体には出現しにくい。方言とされる語が 1 語ある。さらに、-мя 名詞に接頭辞がついているだけで、もとの名詞と変化が同一の名詞が 2 語ある。総じて、-мя の名詞は 10 語であると言えよう。そのうち、1 年生が学習する語の数はさらに少ない。『文法辞典』と文法書とは異なり、教科書が扱う語は限られている。これを明確に上級生に述べるには、語の頻度をもとに、よく使う語とそうでない語を明示的に示すのがよいと考えられる。

ロシア語の頻度辞典には、データがインターネットで公開されているものがある。なかでも、リャシェフスカヤ O. N. とシャロフ S.A. が作成したもの（Ляшевская О. Н., Шаров С. А., *Новый частотный словарь русской лексики*）は、ロシア科学アカデミーロシア語研究所によるロシア語コーパスを素材としており、収録されている語の数は 52000 を超える。量的にも、質的にも、優れた資料であろう。以下、『頻度辞典』と記す。

この『頻度辞典』は、タブ区切りのデータであり、見出しは、

Lemma	PoS	Freq（ipm）	R	D	Doc
語	品詞	頻度			

である。Lemma“語”は、厳密には語の基本形、見出し語を意味する。Freq（ipm）“頻度”はデータが仮に 100 万語あった場合の値である。R、D、Doc はここでは触れない。また、入力などを単純にするために Freq（ipm）は単に Freq と書き換えて利用する。参考文献に記した URL から得られるファイルを展開すると、データファイルが見られる。そのファイル名は、“freqrnc2011.csv” である。これに記載されている Lemma“語”はアルファベット順に配列されている。これを Freq“頻度”の大きな語が上に、小さな語が下になるように並べ替え、かつ、行頭の番号を記した形にして、別ファイル “my_list.csv” に保存する手順を示す。

> のあとには実際に打ち込む命令を、# のあとにはそれに対するコメントを記す。記号がないところは得られる結果である。まず、上のファイルを対象にする。

> sort -k 3 -r -n freqrnc2011.csv | nl -n rz > my_list.csv
sort はファイル “freqrnc2011.csv” の行を並べ替える命令である。一般的に、ファイルが存在するフォルダで命令を実行するのが便利である。筆者の経験によれば、これが多くの学生には未知のことがらである。sort は Mac OS には標準で用意されており、ターミナルから実行できる。-k 3 は 3 番目の key つまり Freq を基準にすること、-r は大きい値が上に来る、いわば逆順にする指示である。-n は値を数値として扱うことを示している。
上で得た結果を次に渡すのが | の役割である。
nl は行に番号をつける命令である。番号付けのフォーマットを与えるのが -n であり、wz は右に揃えて各桁には 0 を入れる指示である。
> とファイル名で、ファイルに保存することを意味する。
この方法では、ファイルの見出し Lemma, PoS などがファイル末にきてしまう。これを行頭に移動させる必要がある。また、行番号で示される番号は頻度順になる。これを Rank “ランク” ということにする。

この手順で得た my_list.csv の一部を示す。

（ファイルの先頭）

(Rank	Lemma	PoS	Freq	R	D	Doc)
000001	и	conj	35801.8	100	99	37704
000002	в	pr	31374.2	100	98	37865
000003	не	part	18028.0	100	97	33999
000004	на	pr	15867.3	100	98	36748
000005	я	spro	12684.4	100	95	17116
000006	быть	v	12160.7	100	98	34184
000007	он	spro	11791.1	100	95	28132
000008	с	pr	11311.9	100	99	35700
000009	что	conj	8354.0	100	98	32419

（ファイルの最後）

（Rank	Lemma	PoS	Freq	R	D	Doc）
052130	не-е-ет	part	0.4	27	80	30
052131	ну-у	part	0.4	16	60	29
052132	тс-с	intj	0.4	19	75	27
052133	зы	part	0.4	5	38	14
052134	ет	part	0.4	18	66	19
052135	сю	intj	0.4	15	68	17
052136	ль	conj	0.4	27	80	34
052137	т-клеточный	a	0.4	6	52	19
052138	д-да	part	0.4	20	74	30

準備できたファイルから、-мя で終わる名詞を抽出するには次のようにする。

> grep 'мя\b' my_list.csv | grep 's'
grep は sort と同様に、Mac OS には標準で用意されており、ターミナルから実行できる。1 番目の引数は検索したい語である。\b（語の区切り）などの特殊な表現（いわゆる正規表現）を含む場合は ' で囲む。'мя\b' と書くことで語末が мя である文字列を含む行を抜き出す。
その結果に対して | の右側を実行する。
s はこの『頻度辞典』の記号で名詞を意味する。正規表現でなくてもよい。名詞だけを抽出する工程である。

（Rank	Lemma	PoS	Freq	R	D	Doc）
000052	время	s	2015.7	100	97	24342
000241	имя	s	401.7	100	97	9402
003831	пламя	s	29.5	98	91	1025
004033	знамя	s	27.8	99	93	1077
004142	семя	s	26.9	100	86	881
004668	племя	s	23.1	99	92	838
007537	бремя	s	12.3	98	92	798
017831	имя-отчество	s	3.6	85	92	215
020080	темя	s	3.0	74	89	194
022958	стремя	s	2.4	58	80	102
027929	вымя	s	1.7	57	87	99

имя を含む複合語 имя-отчество「名と父称」を除けば 10 語である。これら -мя の名詞はすべて中性名詞である。время「時」と имя「名」とそれら以外は Rank“ランク”、Freq“頻度”の点で大きく異なる。пламя「炎」ですら 4000 語レベルであり、学生がふだん出会うことの少ない語であると考えられる。作業の過程や結果を実際に見ることで、学生はこれまでに教科書などで習った -мя 名詞を復習するとともに、学習の上で極めて重要なのは、事実上 время と имя のみであることを再確認するだろう。

学生への課題：『頻度辞典』を素材にして、-мя で終わる名詞を抽出し、頻度の大きな語が既習であることを確認してください。

2.2 前置格が -ии で終わる名詞

あとひとつ取り上げたい名詞がある。それは、単数前置格の語尾が -е ではなく、-и になるタイプである。具体的には、-ие, -ия, -ий で終わる、それぞれ中性名詞、女性名詞、男性名詞である。これらを抽出する手順を示す。

（-ие の中性名詞について）
> grep 'ие\b' my_list.csv | grep 's' | head

ファイルから語末が ие である語を検索する。
その結果から s すなわち名詞を選ぶ。文法的には余分なプロセスであるが、他を明示的に排除したい意図がある。
head は結果の冒頭を示す。

（Rank	Lemma	PoS	Freq	R	D	Doc）
000164	отношение	s	557.4	100	93	11695
000210	решение	s	453.4	100	90	11312
000261	развитие	s	372.6	100	89	9208
000267	условие	s	368.1	100	91	10779
000304	действие	s	329.3	100	90	8471
000349	состояние	s	294.4	100	92	8231
000357	внимание	s	286.0	100	97	9306
000370	предприятие	s	281.6	100	90	6408
000392	положение	s	268.2	100	92	7742
000412	управление	s	256.5	100	91	7225

（-ия の女性名詞について）
> grep 'ия\b' my_list.csv | grep 's' | head
手順はすぐ上と同様である。

（Rank	Lemma	PoS	Freq	R	D	Doc）
000096	Россия	s.PROP	952.2	100	92	16797
000216	история	s	443.9	100	94	9150
000249	компания	s	392.7	100	91	8542
000325	организация	s	312.4	100	88	8542
000341	ситуация	s	298.8	100	93	9729
000390	информация	s	269.2	100	90	7314
000407	федерация	s	258.9	90	65	4175
000422	партия	s	250.4	100	92	4489
000611	армия	s	186.4	100	92	3483
000623	территория	s	182.9	100	90	5941

s.PROP は固有名詞を指すと考えられる。

（-ий の男性名詞について）

（Rank	Lemma	PoS	Freq	R	D	Doc）
001349	Юрий	s.PROP	93.0	100	92	3177
001587	Василий	s.PROP	79.6	100	92	1649
001653	Дмитрий	s.PROP	76.1	100	82	2424

001660	русский	s	75.7	100	95	2103
001713	Евгений	s.PROP	72.7	100	92	2393
001733	Анатолий	s.PROP	71.8	98	90	2084
001940	рабочий	s	64.2	99	94	2014
002406	сценарий	s	50.2	99	88	1503
002556	критерий	s	47.1	92	84	1848
002636	Георгий	s.PROP	45.2	99	91	1290

-ие の中性名詞、-ия の女性名詞ともに高頻度の語が並んでいる。これとはやや異なるのが -ий の男性名詞である。形容詞の変化をする名詞として русский「ロシア人」が学習語彙として扱われるのは自然なことである。ただ、普通名詞である сценарий「シナリオ」、критерий「基準」は 1 年生の授業には登場しない。上級生には、むしろ、高頻度語にロシア人の名 Юрий, Василий などが含まれること、これらが -ий の男性名詞の変化表にあまり付記されないことが新鮮に感じられるかもしれない。

学生への課題：-ий の名詞として、ロシア人の名があることを『頻度辞典』をもとに確認し、それらの語形変化を書いてください。

3. 形容詞

形容詞変化には硬変化と軟変化がある。さらに、語幹末子音（いわゆる後舌音：-г, -к, -х とシュー音 :-ж, -ч, -ш, -щ）や、アクセントの位置（語幹、語尾）により、硬変化語尾と軟変化語尾が同一の変化表に登場する、いわゆる混合変化がある。このうち、語幹末子音が -ж, -ч, -ш, -щ の形容詞は、語幹アクセントなら基本的に軟変化、語尾アクセントなら基本的に硬変化と考えてよい。これらの混合変化は 1 年生にとっては相当理解しにくいものになっていると思われる。ややこしさのもとになっている正書法の規則がまだ完全に把握できていない段階にあるからであろう。

これらの変化類型を改めて上級生に説明したのち、実際にはいくつの形容詞がどの変化に該当するか、示すことができる。これにより、変化類型や文法表の重要度が必ずしも等しくはないと、ある程度は感じていたことが、より明確にイメージできるだろう。ここでは、語幹末子音がシュー音になっている形容詞の数を数える。

```
>grep 'жий\b' my_list.csv | grep -c 'a'
# 語末が жий すなわち語幹末子音が ж であり、語幹アクセントである形容詞を抽出する。
# 後半の grep のオプション -c は数を返す。
36
```

他の場合も含めて同様に求めると、次のようになった。

（語幹アクセント）

語幹末	数
жий	36
чий	87
ший	169
щий	364

ただし、これらに該当する медвежий, птичий などは、典型的な軟変化とはやや異なる。

（語尾アクセント）

語幹末	数
жой	1
чой	0
шой	3
щой	0

上でみたように、シュー音で語尾アクセントの形容詞はおどろくほど少ない。どのような語がこれに属するか調べてみる。

```
> grep 'жой\b' my_list.csv | grep  'a'
（Rank    Lemma       PoS   Freq    R    D    Doc）
000730    чужой       a     159.6   100  95   3627

> grep 'шой\b' my_list.csv | grep  'a'
（Rank    Lemma       PoS   Freq    R    D    Doc）
000097    большой     a     944.4   100  97   17589
000577    небольшой   a     196.6   100  96   6721
051563    меньшой     a     0.4     23   76   28
```

この形容詞のうち学習者の目に触れるのは、чужой「他者の」、большой「大きな」の2語のみであると考えてさしつかえないだろう。このためだけに独立した文法表を設ける意義は形容詞変化の体系を概観することであろうが、1年生の段階ではこの2語を不規則変化として特別に扱うのも自然なことである。

学生への課題：形容詞の混合変化がいろいろな教科書によってどのように説明されているか、比べてみましょう。そして、чужой, большой の変化がどのように記述されているか確認してください。

4. カバー率

ここでは、語の頻度をまとめる方法を記載する。具体的には、どのランクまでの語彙がテキスト（文章）の何パーセントをしめるかという指標である。これをカバー率と

名付ける。

はじめに、計算しやすいような架空のコーパスを取り上げて、計算を検算する。それは、357 語からなるテキストであり、最も Freq“頻度”が大きい Lemma“語”が word1 で、その頻度は 100、次が頻度が 80 の word4、最も頻度の小さい語は word6、頻度数は 2 である。語の順序は、アルファベット順ではなく、頻度数の大きなものが上、小さなものが下になっている。

例として、下の “lemma_test.csv” というファイルを用意した。

Lemma	Freq
word1	100
word4	80
word8	70
word10	50
word9	20
word7	15
word3	10
word5	5
word2	5
word6	2
（Freq 合計	357）

カバー率の計算は、word1 のみの場合は、その頻度数 100 / 頻度数の合計（テキストの語数）357 となる。word1 と word4 とのカバー率は、これらの頻度数の合計（100+180）/ 357 である。同じように、word8 までのカバー率は（100+180+70）/ 357 で得ることができる。それぞれの語までのカバー率は、その上までで合計された値にその語の頻度を足した値（累積和）を頻度数の合計で割ることで求められる。これを下にまとめた。

Lemma	Freq	累積和	累積和 / Freq の合計
word1	100	100	0.280
word4	80	180	0.504
word8	70	250	0.700
word10	50	300	0.840
word9	20	320	0.896
word7	15	335	0.938
word3	10	345	0.966
word5	5	350	0.980
word2	5	355	0.994
word6	2	357	1.000
合計	357		

これで手作業での計算方法が確認できた。

語数の大きい『頻度辞典』を扱う場合は、なんらかのソフトが便利である。R を使えば次のようになる。

> lemma_test <- read.csv（"lemma_test.csv", header=TRUE, sep = "\t"）
read.csv 関数を用いて、ファイル read.csv の内容を lemma_test という変数に読み込む。header = TRUE は見出し行があることを示している。sep のところはデータ区切りをタブにする指定である。

> ruiseki <- cumsum（lemma_test$Freq）
cumsum 関数を用いて、上で得た lemma_test の Freq の累積和を求め、これを変数 ruiseki に格納する。

> ruiseki_p <- cumsum（lemma_test$Freq）/sum（lemma_test$Freq）
それぞれの累積和を、sum 関数で得られる頻度数の合計で割って、その値を ruiseki_p という変数に入れる。ruiseki と ruiseki_p の値は、それぞれが番号付きで格納される、いわば配列のようなデータになっている。

> ruiseki
[1] 100 180 250 300 320 335 345 350 355 357
ruiseki（それぞれの累積和）の 1 番目が 100、以降 2 番目が 180 であることなどを示している。

> ruiseki_p
[1] 0.2801120 0.5042017 0.7002801 0.8403361 0.8963585 0.9383754 0.9663866
[8] 0.9803922 0.9943978 1.0000000
ruiseki_p も 1 番目から順に値が並んでいる。[1] の行が 1 番目から 7 番目、[8] の行が 8 番目から 10 番目の値である。

累積和などの計算は上の例と同じように R で行える。が、頻度順（正確には逆順）に並べたファイルはターミナルで作っておいても便利である。

（ターミナルでの作業）
> sort -k 3 -n -r freqrnc2011.csv > sorted_lemma.csv
sort 命令を使う。-k3 は見出しの 3 番目つまり Freq を扱うこと、-r はこれを数値として扱うこと、-r は逆順で並べる指示である。最後の > のあとには、命令が実行された結果を格納する適当なファイル名を書く。なお、この命令では、見出しがファイルの最後尾に来るのでこれを第 1 行目に書き直す。

（R での作業）

```
> Lemma <- read.csv（"sorted_lemma.csv", header=TRUE, sep = "\t"）
> ruiseki <- cumsum（Lemma$Freq）
> ruiseki_p <- cumsum（Lemma$Freq）/sum（Lemma$Freq）
```

たとえば、頻度数が 500 番目に大きな（つまりランクが 500 である）語の頻度は次のように示せる。

```
> ruiseki_p[500]
[1] 0.5556647
```

これを 500 も含めて適当な区切りになる順番で行えば良い。この結果に語も書き込むと次のようになる。

500	0.5556647	опыт
1000	0.639955	сестра
1500	0.6926246	преступление
2000	0.7303106	научиться
2500	0.7588872	тоска
3000	0.7813894	вертолет
3500	0.7998819	вдвоем
4000	0.8157208	оформить
4500	0.8293469	гордый
5000	0.8411811	пропадать

などとなる。カバー率 80%、85%、90% を超えたところにある語を示す。

3504	0.800018	подхватить
5423	0.8500295	диагноз
8969	0.9000027	сжаться

本稿では省略するが、横軸にランク、縦軸に累積和 / 頻度数の合計をプロットしたグラフを教員が示して、たとえば、1000 語レベルを超えると単語学習の効果が薄れた感じがすることなどを学生が再確認することもできる。

学生への課題：ロシア語専攻で設定された到達度目標ないしはロシア語の試験には語彙レベル（ランク）が説明されていますか。もしそうなら、そのランクまでの語彙リストを、『頻度辞典』を用いて作成してみてください。

5. おわりに

学生が学習した語彙のうち、どの語に、どの語類に頻繁にでくわすのか…これを学生自身が一定のデータとコマンドを用いて、これまでには感覚的になんとなく把握し

ていたことを、語の頻度をもとにはっきりと理解してもらえればとの願いから小論を記した。名詞、形容詞のさらなる興味深い現象や他の品詞にも話を広げていくのが展望の可能性である。

筆者自身はいわゆる情報系の専門教育を受けた経験がなく、ソフトやコマンドなどを恣意的に理解している可能性もある。自らの浅学を恥じるばかりである。

参考文献

Зализняк А.А.

1980 Грамматический словарь русского языка: Словоизменение. 2-е изд., стереотип. М.: Русский язык.

Ляшевская О. Н., Шаров С. А.

Новый частотный словарь русской лексики

URL: http://dict.ruslang.ru/freq.php（2022 年 10 月 30 日 最終閲覧）

（同著者の Частотный словарь современного русского языка（на материалах Национального корпуса русского языка）. М.: Азбуковник, 2009. のデジタル版）

R Core Team

2021 *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

URL: https://www.R-project.org/（2022 年 10 月 30 日最終閲覧）

木村 彰一

1974 『ロシア文法の基礎（改訂版）』, 白水社 , 東京（第 34 刷 2005 年を参照した）.