『現代日本語書き言葉均衡コーパス』のロシア語翻訳データの構築とその日露対照研究への活用の可能性

国立国語研究所論集 (NINJAL Research Papers) 19: 167–185 (2020)
ISSN: 2186-134X print/2186-1358 online

『現代日本語書き言葉均衡コーパス』のロシア語翻訳データの構築とその日露対照研究への活用の可能性

宮内拓也[a]　　プロホロワ マリア[b]

[a] 東京大学／国立国語研究所 共同研究員
[b] 東京外国語大学大学院 博士後期課程

要旨

『現代日本語書き言葉均衡コーパス』（の一部のデータ）には，既に英語，イタリア語，インドネシア語，中国語の翻訳データが構築されているが，新たにロシア語の翻訳データを構築した。対象となる起点テキストは『現代日本語書き言葉均衡コーパス』新聞（PN）コアデータ 16 サンプル（総語数は短単位で全 16,657 語）とし，ロシア語目標テキストの総語数は 13,070 語となった。本データの構築にあたっては，日本語からロシア語へ人手による翻訳を行ったが，日本語とロシア語の言語構造の違いや表現の違い等により，翻訳に困難が生じた箇所もあった。本稿では，翻訳データの構築方法，翻訳の際の留意点の詳細を述べる。また，原文の日本語テキストと翻訳先のロシア語テキストは人手で文単位のアライメントを取り，各文には ID を付与した。その作業方法についても記述する。翻訳データの構築，アライメント作業により，起点テキストと目標テキストは簡易的な日露パラレルコーパスとして利用可能となり，日露対照研究や類型論研究に活用できると考えられる。本稿では，このような活用の可能性を示すために，ケーススタディとして日本語の文末表現を取り上げ，ロシア語と対照させて同異を議論する*。

キーワード：『現代日本語書き言葉均衡コーパス』，対訳コーパス，ロシア語，文末表現

1. はじめに

『現代日本語書き言葉均衡コーパス』（Maekawa et al. 2014; 以下, BCCWJ）の一部のデータには，既に英語，イタリア語，インドネシア語，中国語の翻訳データ（BCCWJ-Trans）が構築されている（浅原・森田 2015）。今回，新たに人手による翻訳により，BCCWJ の 16 サンプル分のロシア語の翻訳データを構築した。本稿では，翻訳データの構築方法，翻訳の際の留意点や翻訳データの活用の可能性を報告する。

以下，2 節では翻訳対象としたデータと構築したデータの概要について概観する。3 節では翻訳の方法と翻訳の際の留意点について，4 節では日本語の起点テキストとロシア語の目標テキストのアライメント（対応付け）について，それぞれ述べる。5 節では翻訳データの日露対照研究

* 本稿の一部は国立国語研究所コーパス開発センターの共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」（プロジェクトリーダー：浅原正幸）の研究成果である。加えて，本研究は JSPS 科研費「顕在的な冠詞がない言語における名詞句の統語構造と意味解釈の研究」（課題番号：17J07534，研究代表者：宮内拓也），および「名詞句における統語構造と定性・特定性の意味解釈の相互関係に関する研究」（課題番号：19K23073，研究代表者：宮内拓也）の助成を一部受けている。

また，本稿の内容は「言語資源活用ワークショップ 2018」（2018 年 9 月 4 日，国立国語研究所）における口頭発表「『現代日本語書き言葉均衡コーパス』のロシア語翻訳データの構築」（宮内・プロホロワ 2018）に基づき，その内容を修正，増補したものである。発表の際に参加者の方々から貴重なご意見を頂いた。ここに感謝申し上げる。

への活用の可能性について，ケーススタディとして日本語の文末表現を取り上げ，ロシア語と対照させて論じる。6 節は本稿全体のまとめと今後の展望を示す。

2. 翻訳対象のデータと翻訳データの概要

翻訳の対象は BCCWJ の新聞（PN）コアデータ [1]16 サンプルである。サンプルは PN コアデータの内から BCCWJ-ANNOTATION-ORDER[2] に基づき選択された。構築されたロシア語翻訳データの総語数は 13,070 語であり，文数は 848 文であった。基本的なデータとして，表 1 に対象となる日本語起点テキストの語数（短単位数），文節数，文数，およびロシア語目標テキストの語数，文数をサンプルごとに示す。

表 1　対象となる日本語の起点テキストとロシア語の目標テキストのサイズ

サンプル名	日本語			ロシア語	
	短単位数	文節数	文数	語数	文数
PN1c_00001	784	236	42	568	40
PN1d_00001	783	235	34	555	54
PN1e_00001	763	219	35	653	34
PN1f_00001	797	181	38	568	56
PN2e_00001	750	214	27	639	33
PN3b_00001	975	311	43	787	39
PN3g_00001	2,640	919	142	2,227	167
PN4a_00001	1,244	425	51	1,012	62
PN4b_00001	758	246	26	548	23
PN4c_00001	737	250	31	587	33
PN4f_00001	1,047	297	40	761	38
PN4g_00001	905	296	36	674	36
PN1a_00002	1,797	611	93	1,409	106
PN1b_00002	1,024	277	38	763	57
PN1d_00002	734	206	28	564	29
PN1e_00002	919	272	35	755	41
合計	16,657	5,195	739	13,070	848

既に述べたように，『現代日本語書き言葉均衡コーパス』（の一部分のデータ）には，英語，イタリア語，インドネシア語，中国語の翻訳データがある（浅原・森田 2015）。これらの各言語の翻訳データの総語数 [3]，対象のサンプル数をまとめたものが表 2 [4] である。

[1] コアデータとは，自動解析結果を人手修正したより精度の高いデータサブセットである（山崎 2011）。
[2] BCCWJ-ANNOTATION-ORDER とは BCCWJ コアデータサンプルにおけるアノテーション優先順序である。以下を参照のこと。https://github.com/ masayu-a/BCCWJ-ANNOTATION-ORDER
[3] ただし，中国語は字数をカウントしている。
[4] 宮内・プロホロワ（2018: 5）においては，インドネシア語翻訳データの総語数を「51 語」，サンプル数を「1 サンプル」としていたが，これは誤りである。ここにお詫びし訂正する。

表 2　ロシア語翻訳データと他の言語の翻訳データの比較

言語名	総語数	対象のサンプル数[5]
ロシア語	13,070 語	16 サンプル
イタリア語	6,563 語	16 サンプル
英語	4,840 語	6 サンプル
インドネシア語	297 語	6 サンプル
中国語	7,852 字	6 サンプル

ここからわかるように，他の BCCWJ 外国語翻訳データと比較すると，ロシア語が現状の規模としては最も大きいものとなった。

3. 翻訳方法

翻訳データの構築にあたっては，日本語からロシア語へ人手による翻訳を行った。翻訳者は，（当時）東京外国語大学大学院博士前期課程の，翻訳家を志望するロシア語母語話者の学生（第 2 著者）[6] である。翻訳にあたり，起点言語の日本語と目標言語のロシア語における言語構造の違い等により，翻訳に困難を生じさせるであろう箇所が多くあることが予想されたため，翻訳に際しては一定の方針を設定した。具体的には（1）の方針で翻訳を行った。

（1）　翻訳の方針のまとめ
- a. 単文レベルではロシア語の自然さを追求するが，談話レベルでは追求しない
- b. 企業名は特定の場合を除きラテン文字で表記する
- c. 企業名や地名は特定の場合を除き上位概念を表す語を追加した上で曲用させない

以下，（1a）の方針について 3.1 節で，（1b）について 3.2 節で，（1c）について 3.3 節で，それぞれ詳細を述べる。

3.1 方針（1a）：単文レベルではロシア語の自然さを追求するが，談話レベルでは追求しない

談話レベルではロシア語としての自然さは失われてもよいとしたが，単文レベルでは自然なロ

[5] 具体的に対象となっているサンプルは以下の通りである。
・英語：　OY04_ 00001 / OC01_ 00001 / PM25_ 00001 / PB12_ 00001 / PN1c_ 00001 / OW6X_ 00000
・イタリア語：　OC01_ 00001 / OW6X_ 00000 /OY04_ 00001 / PB12_ 00001 / PM25_ 00001 / PN1c_ 00001 / OC02_ 00001 / OY12 00005 / OC03_ 00001 / OY09_00008 / OC04_ 00001 / OY15_00014 / OC05_ 00001 / OY04_00017 / OC06_ 00001 / OY04_00027
・インドネシア語：　OC01_00001 / OY04_00001 / OC02_00001 / OY12_00005 / OC03_00001 / OY09_00008
・中国語：　OY04_ 00001 / OC01_ 00001 / PM25_ 00001 / PB12_ 00001 / PN1c_ 00001 / OW6X_ 00000
OY はブログ，OC は知恵袋，PM は雑誌，PB は書籍，PN は前述のように新聞，OW は白書の各レジスターをそれぞれ表す。なお，対応する日本語文の数は英語，中国語が 319，イタリア語が 436，インドネシア語が 27 である。

[6] なお，第 2 著者は，本データの翻訳以前にも多数の日露，露日商業翻訳の経験がある。

シア語になるような翻訳を行った。これは，今回の翻訳では日本語とロシア語の対応付けは単文レベルで行うことから，単文としての日本語との対応を優先し，談話としての自然さを犠牲にするという判断を行ったためである[7]。

例えば，日本語では非過去形（ル形）と過去形（タ形）[8]が混ざっている文章が多々ある[9]が，ロシア語だと特定の文脈がない限りどちらかで統一するのが一般的である。よって，日本語の非過去形と過去形の混ざった文章を各文の時制を保ったまま翻訳するとロシア語の談話としては不自然なものとなる場合が多い。日本語からロシア語への文芸翻訳の場合，時制はロシア語の談話として自然な形に直されることが多い。例えば，現代小説の『真鶴』からの引用である（2）を見られたい[10]。

（2） 風はおさまりかけている。ホームにあがると，じきに各駅停車が来た。（川上弘美『真鶴』）

（2）の各文で描写されている出来事は現実的にはほぼ同じ時間で起きているが，それぞれ非過去形と過去形が用いられている。しかし，ロシア語においては，このように動詞の非過去形と過去形を交互に用いることは「時制の不一致（raznoboj vido-vremennyx form glagola[11]）」などと言われ，規範的には誤りとして扱われることもある。例えば，Markova（2013）は，「動詞形式の体（アスペクト）・時制の不一致」に関して，意味的に結束性を持つ部分において動詞は同一の体，時制の形式を用いるべきだと指摘している。さらに，Balašova and Dement'ev（2005）は「時制の一貫性（vremennaja svjaz')」として，特別な意図なく過去時制と現在時制を織り交ぜて用いることを誤りとしている[12]。実際に，『真鶴』のロシア語翻訳では，上記（2）の部分は（3）[13]のように翻訳されている。

---

[7] もちろん，ロシア語の談話として自然なものを優先するべき場合もあるだろう。談話としての自然さを優先するべきか，単文としての日本語との対応を優先するべきかについてはコーパスを利用する目的によって異なると考えられる。なお，具体的なアライメントを取る作業については 4 節で述べる。

[8]「ル」と「タ」の対立は，必ずしも純粋に非過去，過去といった時制の対立を意味するわけではないが，以降，本稿では便宜的に文末が「ル」で終わる形式を非過去形，「タ」で終わる形式を過去形と呼ぶこととする。

[9] 例えば，野田（1992）は，日本語の小説において，過去の事象を表す際，過去形のみを使う文体よりも非過去形と過去形を混ぜて使う文体の方がはるかに多く用いられていることを報告している。

[10]（2）以降の例文において，下線部は注目すべき要素を表す。

[11] 本稿では，キリル文字（ロシア文字）はラテン文字に翻字する。翻字は以下の通りである：A=A，Б=B，B=V，Г=G，Д=D，E=E，Ё=E，Ж=Ž，З=Z，И=I，Й=J，K=K，Л=L，M=M，H=N，O=O，П=P，P=R，C=S，T=T，У=U，Ф=F，X=X，Ц=C，Ч=Č，Ш=Š，Щ=Šč，Ъ=”，Ы=Y，Ь=’，Э=È，Ю=Ju，Я=Ja。ただし，必要に応じてキリル文字による表記を併記している箇所もある。なお，脚注 14 も参照。

[12] これは逆に言えば，特別な意図があれば誤りではないことを意味している。実際に，ロシア語においても，文章におけるあるまとまりの内部で異なる時制が混交された形で用いられることもあり得る。具体的には，ルポルタージュの現在（nastojaščee reportažnoe/nastojaščee dinamičeskoe），歴史的現在（nastojaščee istoričeskoe）（Apresjan 1988 など）が挙げられる。しかし，結論を出すには詳細な検討が必要であるものの，これらは意味，用法の点で日本語文章における非過去形，過去形の混交とは性質を異にすると考えられる。

[13] ロシア語の例文にはグロスを付す。本稿で用いる文法情報の略記は以下の通りである：NOM= 主格，GEN= 属格（生格），DAT= 与格，ACC= 対格，INS= 具格（造格），LOC= 前置格（処格），M= 男性，F= 女性，N= 中性，SG= 単数，PL= 複数，PRS= 現在，PST= 過去，PTCP= 分詞（形動詞），GER= 分詞（副動詞），INF= 不定形，3=3 人称。なお，ロシア語においては性の対立は単数の場合にのみ見られることから，ある語句に性が示されていれば当該の語が単数であることを意味する。複数の場合には性は示さない。

(3) Veter načal ponemnogu stixat'. Ja podnyalas' na
wind-NOM.M start-PST.M little_by_little abate-INF I-NOM.F ascend-PAST.F to
platformu, vskore podošla élektrička.
platform-ACC soon approach-PST.F train-NOM.F

(Mironova, Ljudmila O. (訳) *Manadzuru*)

原文の（2）と異なり，(3）の翻訳においては，下線部の動詞は両方とも過去形となっている。(4)，(5）で示すように『真鶴』と同じ著者の『神様』についても，翻訳者は異なるが同様の現象が見られる。

(4) くまは，魚をわたしの目の前にかざした。魚のひれが陽を受けてきらきら光る。釣りをしている人たちがこちらを指さして何か話している。（川上弘美『神様』）

(5) Medved' podnes rybinu prjamo k moemu licu. Ee plavniki v
bear-NOM.M bring-PST.M fish-ACC straight to my-DAT.N face-DAT.N its-PL fin-PL in
lučax solnca otlivali serebrom. Vse rybolovy na rečke,
ray-LOC sun-GEN shine-PST.PL silverly all-NOM.PL fisherman-NOM.PL at river-LOC
sgrudivšis' v kuču, pokazyvali na nas pal'cami i čto-to
crowd-GER in heap-ACC indicate-PST.PL us-ACC finger-INS.PL and something-ACC
družno obsuždali.
in_concert discuss-PST.PL

(Dutkina, Galina B. (訳) *Medvežij bog*)

文芸翻訳では他にも同じような例が多く見られる。

このように，日本語・ロシア語間の（少なくとも文芸）翻訳においては，日本語原文の時制をロシア語文において維持せず，ロシア語としてより自然な時制に直す傾向が見られる。しかし，今回の翻訳では，単文レベルでアライメントを取ること，および談話の自然さという観点は翻訳者による主観が入りやすいことを考慮して，翻訳元の日本語の文の時制を，ロシア語文でも用いることとした。

例えば，(6）の文を見られたい。翻訳元の日本語文（6a）では，非過去形と過去形が共に用いられている。ロシア語文（6b）ではすべて過去形とする方がロシア語の談話としては自然であるが，今回の翻訳では日本語文（6a）の形式に合わせて翻訳された。

(6) a. [...] 二，三年時に担任だった池田弘子先生（七十五）は違った。「そんな薄いかばんじゃ遊び道具も入らないよ」「体育や部活では，危ないからピアスをはずしたほうがいい」。やんわり語りかける。

b. Xiroko Ikéda (75 let), kotoraja byla ee klassnym rukovoditelem
[Hiroko Ikeda]-NOM.F 75 ages who-NOM.F be-PST.F her [homeroom teacher]-INS

na vtorom i tret'em godu obučenija, byla ne takoj. Ona
on second-LOC and third-LOC year-LOC education-GEN be-PST.F not such she-NOM
razgovarivaet s devočkoj mjagko: «V takuju ploskuju sumku daže
talk-PRS.SG.3 with girl-INS softly in such-ACC thin-ACC bag-ACC even
igry ne vlezut», «Na fizkul'ture i v sekcijax lučše
toys-NOM.PL not go-PRS.PL.3 on physical_education-LOC and in section-LOC better
snimat' serežki, éto opasno».
take_off-INF earrings-ACC it_is dangerous （読売新聞 [BCCWJ: PN1c_00001]）

3.2 方針（1b）：企業名は特定の場合を除きラテン文字で表記する

ロシア語では，ロシア国外の企業名やその商標，国外の新聞の名前などの固有名詞はキリル文字（ロシア文字）[14] で表記される場合もあれば，ラテン文字で表記される場合もある（Ermolovič 2001, Prokopčuk 2017 [15] など）。また，Prokopčuk（2017）によれば，表記上の選択は，文章の作者のセンスや具体的な語がどちらの表記で定着しているかなどによる。例えば，ロシアでなじみのある企業名ほどキリル文字で表記することが多く，また，インフォーマルな文章ほどキリル文字を使う場合が多い。

今回の翻訳では，サンプルの文章が新聞 [16] であり文体的にフォーマルに近いことを考慮し，企業名等は特定の場合を除き，ラテン文字で表記することとした。例外となる特定の場合とは，ロシア語で正式名称のあるものである。例えば，ロシア語で正式名称のある国外の新聞は（7）で示すとおりである。

（7） 例外的にキリル文字表記する国外の新聞名
「人民日報」= Žén'min' žibao (Жэньминь жибао)
「ルモンド」= Mond (Монд)
「タイムズ」= Tajms (Таймс)
「エスタド・デ・サンパウロ」= Éštadau (Эштадау)

なお，このキリル文字とラテン文字の表記については，単に表記の問題というわけではなく，文法面にも影響を与える問題である。一般に，ラテン文字表記となる場合にはその名詞は曲用しないことになるが，キリル文字表記であれば名詞の性・数・格に応じて曲用する（Prokopčuk

[14] ロシア語で用いられるキリル文字はロシア文字と呼ばれ，キリル文字はロシア語では用いられない字母をも含むことになる。しかし，ロシア文字という呼称よりもキリル文字という呼称の方が一般によく知られていると考えられるため，本稿では以降，一律キリル文字と表記する。
[15] Prokopčuk（2017）は，具体的にはペレストロイカ後から，国外の商標についてキリル文字でもローマ字でも表記されるようになったことを指摘している。
[16] Ermolovič（2001）は，ビジネスや新聞実務において，外国企業の名前のラテン文字表記が多いことを指摘している。

2017: 322）[17]。例えば，*Toyota*「トヨタ」であれば，ラテン文字表記の場合は，格により形態が変化することはないが，キリル文字で表記した場合，（8）で示すように曲用する。

（8） a. Tojota
Toyota-NOM
「トヨタが」

b. Tojoty
Toyota-GEN
「トヨタの」

c. Tojote
Toyota-DAT
「トヨタへ」

d. Tojotu
Toyota-ACC
「トヨタを」

e. Tojotoj
Toyota-INS
「トヨタによって」

f. o Tojote
about Toyota-LOC
「トヨタについて」

ロシア語では固有名詞でもキリル文字表記であれば曲用してしまうため，ラテン文字表記にして曲用させない方がもともとの名称を同定しやすいという利点もある。

3.3 方針（1c）：企業名や地名は特定の場合を除き上位概念を表す語を追加した上で曲用させない

固有名詞には，上位概念を示す名詞（例えば，*kompanija*「会社」，*proizvoditel'*「メーカー」など）が同格句として前置される場合もよくある。今回の翻訳では，（9a）のように上位概念を示す名詞が日本語で表示されていても，（10a）のように表示されていなくても，ロシア語では（9b），（10b）で示すように上位概念を示す名詞を加えることとした。

[17] しかし，3.3 で述べるように，上位概念を示す名詞が同格句として固有名詞に前置される場合も多い。この場合には，固有名詞はキリル文字表記であっても曲用させないことが多い（Graudina et al. 1976）。新聞を対象に調査をした Graudina et al.（1976: 174）によれば，以下の（i）は共に可能ではあるものの，（i a）のパターンが 95.79%，（i b）のパターンが 4.12% となることが示されている。

（i） a. v gazete «Pravda»
in newspaper-LOC Pravda
b. v gazete «Pravde»
in newspaper-LOC Pravda-LOC
「プラヴダ紙に」

なお，上位概念を示す名詞が同格句として付加されない場合には，（ii a）のように固有名詞も曲用させる（Graudina et al. 1976）。

（ii） a. fotokonkurs «Pravdy»
photo_contest Pravda-GEN
b. fotokonkurs gazety «Pravda»
photo_contest newspaper-GEN Pravda
「プラヴダ紙の写真コンテスト」

(9) a. 米ガートナー・グループ傘下の調査会社データクエスト
b. kompanija Dataquest, prinadležaščaja amerikanskoj Gartner Group
company-NOM.F Dataquest belong_to-PTCP.NOM.F American-DAT Gartner Group
（産経新聞 [BCCWJ: PN1d_00002]）

(10) a. ファミリーマートは 9 日，[...] 発表した。
b. 9 čisla kompanija FamilyMart soobščila [...].
9 number company-NOM.F FamilyMart announced-PST.F
（産経新聞 [BCCWJ: PN1d_00002]）

（10b）では，上位概念を示す名詞として *kompanija*「会社」が付加されている。もし，これが示されていない場合，*FamilyMart* を知っているロシア語母語話者は一つの店舗としての *FamilyMart* を想定してしまう可能性が高い[18]。この場合は，(11）のように上位概念を示す名詞を付加するならば，*magazin*「店」を想定することになる。店舗を想定してしまう場合は，*magazin* が表示されていなくても，動詞は形式上これと一致し，男性形となる。

(11) 9 čisla (magazin) FamilyMart soobščil [...].
9 number shop-NOM.M FamilyMart announced-PST.M
「ファミリーマート（店舗）は 9 日，[...] 発表した。」

なお，上記の（10）において [⋯] で省略されている部分の内容は（12）に示す通りである。

(12) 店舗内に設置されているマルチメディア端末「Fami ポート」で，シャープの携帯情報端末「ザウルス」のコンテンツを 5 月中旬から販売すると
（産経新聞 [BCCWJ: PN1d_00002]）

これを見ると，その内容からここでの「ファミリーマート」は 1 店舗[19] ではなく，会社全体[20] として解釈するのが妥当であることがわかる。（10b）のように，上位概念を示す名詞として *kompanija*「会社」が付加されていることで，（11）の解釈を排除することができる。

以上に述べたように，上位概念を示す名詞の違いによって，内容的な齟齬をきたすだけではなく，文法面にも影響が出る。適切な上位概念を示す名詞を付加することで曖昧性を排除することができ，文がわかりやすくなるという効果がある。

[18] ロシア語においては，子音で終わる名詞は男性名詞である点からも，動詞は男性形を想定してしまいやすく，店舗としての *FamilyMart* が想起されやすい。
[19] 例えば，「ファミリーマート立川北駅店」などの具体的な店舗を指す。
[20] つまり，「株式会社ファミリーマート（Family Mart Co., Ltd.）」を指す。

以上の方針（1b, c）のもとでは，（8）で示した *Toyota*「トヨタ」は以下（13）のように翻訳，表記されることになる[21]。

（13） a. kompanija Toyota
company-NOM.F Toyota
「トヨタが」

b. kompanii Toyota
company-GEN.F Toyota
「トヨタの」

c. kompanii Toyota
company-DAT.F Toyota
「トヨタへ」

d. kompaniju Toyota
company-ACC.F Toyota
「トヨタを」

e. kompaniej Toyota
company-INS.F Toyota
「トヨタによって」

f. o kompanii Toyota
about company-LOC.F Toyota
「トヨタについて」

この場合，ラテン文字で表記されている *Toyota*「トヨタ」の部分は曲用せず，付加されている上位概念を示す *kompanija*「会社」が曲用する。よって，固有名詞自体の形態が変化することなく，もとの名称がわかりづらくなることもない[22]。さらに，*kompanija*「会社」は曲用するため，この部分で性・数・格という文法的な情報も保持される[23]。

企業名はラテン文字で表記する方針としたため，上記の（9），（10），（13）のように，企業名を示す名詞が曲用することはない。しかし，キリル文字表記される従来のロシア語の語であれば上位概念の言葉があっても固有名詞は曲用することがある。例えば，地名はラテン文字表記することは難しく，（14a）で示すように，上位概念を示す名詞があってもロシアの地名は普通曲用する[24]。

---

[21] ここで，本稿ではキリル文字がラテン文字に翻字されているのでわかりづらいが，*kompanija*「会社」の部分はキリル文字，*Toyota*「トヨタ」の部分はラテン文字で表記されることになる点に注意されたい（つまり，*компания Toyota*, *компании Toyota*, *компанию Toyota* …）。

[22] なお，前述の（10），（11）における「ファミリーマート」の話と並行的であるが，以下（iii）のように，ロシア語の *Tojota* には「トヨタ車」として具体的な車を指す，シネクドキ的な用法も持つ。

（iii） Ja kupil Tojotu.
I-NOM.M buy-PST.M Toyota-ACC
「私はトヨタ（車）を買った。」

kompanija「会社」の付加によって，この解釈ではないことを明白に表すことができる。

[23] ただし，実際には（13）の例の *kompanija*「会社」は単数であれば，属格，与格，前置格が同形となる。

[24] ただし，以下（iv）で示すように，河川の名称などは曲用する場合もしない場合もある。

（iv） a. na reke Volge
at river-LOC.F Volga-LOC.F
「ヴォルガ川で」

b. na reke Enisej
at river-LOC.F Yenisei-NOM.M
「エニセイ川で」

上位概念を示す名詞（ここでは *reka*「川」）と河川の名称を示す名詞の性が一致している場合，曲用し（iv a），一致しない場合は曲用しない（iv b）とされるが，詳細はここでは述べない。

これと同様の方針で翻訳すれば，例えば「静岡県」は（14b）のように曲用させることになる。

（14） a. v gorode Moskve
in city-LOC Moscow-LOC
「モスクワ（という都市）で」
b. v prefecture Sidzuoke
in prefecture-LOC Shizuoka-LOC
「静岡県で」

しかしながら，この場合では，*Sidzuoka*「静岡」は *Sidzuoke* と変化してしまっているため，もともとの名称が分かりにくい[25]。よって，今回の翻訳では，上位概念を示す名詞を付加し，（15）のように地名そのものは曲用させない方針とした[26]。

（15） a. 静岡県出身。
b. Rodom iz prefektury Sidzuoka.
birth-INS from prefecture-GEN Shizuoka-NOM （西日本新聞 [BCCWJ: PN3g_00001]）

ただし，今回の翻訳の対象は PN コアデータであり，出典が新聞からのテキストであることから，地名が多く出現する。地名が出るたび，そのすべてに上位概念を示す名詞（*gorod*「都市」，*prefektura*「県」など）を付加していくと文章がかさばってしまう上に，場合によっては単文レベルでも自然さを欠くことになる[27]。よって，（16）で示す地名についてはロシア語母語話者にとっても十分になじみのある地名と考え，日本語文に示されていない限り，このような語は付加しないこととした。

（16） 例外的に上位概念を示す語を付加しない地名
「東京」= Tokio (Токио)
「大阪」= Osaka (Осака)
「京都」= Kioto (Киото)
「広島」= Xirosima (Хиросима)

[25]（14b）の場合，日本語が一般に開音節言語であることを考慮しても，論理的には元の名称は *Sidzuoka* 以外に *Sidzuoko* や *Sidzuoke* といった可能性が想定され得る。
[26] なお，最近のロシアではロシア語母語話者が聞き慣れないロシア国外の会社名や地名が増えてきており，それらを曲用させてしまうと元の名称がわかりづらくなってしまう。そのため，上位概念を示す語があっても固有名詞を曲用させるという規範はマスメディア等でも緩んできており，あえて曲用させないことが増えているようである。
[27] 1 文の中に複数回同一レベルの地名が現れること（「静岡県と長野県」，「立川市と武蔵野市」，「緑町と高松町」など）も十分に想定でき，その場合は同じ上位概念を表す語を複数回付加するのは単文レベルでも不自然となってしまう。

例えば，（17）で示すように，日本語文に上位概念を表す語が示されていないため，翻訳文の *Tokio*「東京」にも上位概念を表す語は付加されていない。

（17） a. 東京のヨドバシカメラ新宿西口本店
b. v glavnom magazine Yodobashi Camera v Tokio u zapadnogo vyxoda
in main-LOC shop-LOC Yodobashi Camera in Tokyo at western-GEN exit-GEN
so stancii Sindzjuku
from station-GEN Shinjuku （朝日新聞 [BCCWJ: PN4a_00001]）

この場合，*Tokio*「東京」や *Kioto*「京都」など *-o* で終わる地名は不変化名詞（indeclinable noun）となるため，曲用もさせないことになる。しかし，*Osaka*「大阪」や *Xirosima*「広島」のように *-a* で終わる地名は（18）で示すように原則的には曲用させることになる[28]。

（18） a. 広島，大阪各高裁長官を経て
b. zanimal post glavy Vysšego suda Xirosimy, zatem
was_engaged_in post-ACC head-GEN high-GEN court-GEN Hiroshima-GEN then
glavy Vysšego suda Osaki
head-GEN high-GEN court-GEN Osaka-GEN （西日本新聞 [BCCWJ: PN3g_00001]）

ここで，*Osaka*「大阪」と *Xirosima*「広島」はそれぞれ属格形 *Osaki*「大阪の」，*Xirosimy*「広島の」となっている。

実際，現代ロシア語においては，ロシア国外（特にアジア圏）の固有名詞の表記には大きな揺れがあり，例外もでてきてしまうことが一般に多く見られる。しかしながら，今回の翻訳ではなるべく誤解釈の少なくなる，そしてできるだけもとの名称がわかるような方法として，上記（1b, c）の方針を設定した。

4. 日本語文とロシア語文のアライメント

本稿で報告した翻訳については，日本語起点テキストとロシア語目標テキストの間で，文レベルで対応付けが行われており，各文に ID が付与されている。本節ではそのアライメント作業に

[28] 実際には，日本語の *-a* で終わる地名は曲用しない傾向を見せているとの指摘があり，Voroncova（2011: 74）では以下（v a）も（v b）も可能であることが示されている。

（v） a. v Osake
in Osaka-LOC
b. v Osaka
in Osaka
「大阪で」

しかし，本翻訳では，より規範的な原則に沿い，これらは曲用させることとした。

ついて述べる。

4.1 アライメント作業

日本語原文とロシア語翻訳文の間のアライメントを取る作業は MS Excel を用いて，人手で各文を対応させるという方法でなされた。図 1 は MS Excel 上での各文の対応の 1 部を示している。

サンプル名	文番号	識別番号	日本語文	ロシア語文
PN1c_00001	01		ALBUM　私の先生	ALBUM Мой учитель.
PN1c_00001	02		キャスター　蓮舫さん「おしゃべり」才能後押し	Телеведущая Рэнхо. Развитие способностей к «болтовне».
PN1c_00001	03		東京都生まれ。	Родилась в Токио.
PN1c_00001	04		九十五―九十七年、中国・北京大に留学し、帰国後に双子を出産。	С 1995 по 1997 год была на стажировке в Пекинском университете, а после возвращения в Японию родила близнецов.
PN1c_00001	05		子育てのかたわらテレビ、ラジオなどで活躍中。	Совмещает воспитание детей с выступлениями на телевидении, радио и т.д.
PN1c_00001	06		三十三歳。	33 года.
PN1c_00001	07		幼稚園から大学まで通った青山学院では、とにかく活発で、目立つ生徒だったという。	В образовательных учреждениях Аояма Гакуин, которые она посещала с детского сада и вплоть до университета, Рэнхо, по её словам, всегда была активной ученицей, выделяющейся среди других.
PN1c_00001	08		高等部では自由な校風もあって、流行に乗ってかばんを薄くつぶしたり、ピアスをしたり。	В старшей школе, благодаря довольно свободным школьным порядкам, она следовала модным течениям: расплющивала школьную сумку, чтобы сделать её более плоской, прокалывала уши и так далее.
PN1c_00001	09		呼び出して注意する先生もいたが、二、三年時に担任だった池田弘子先生（七十五）は違った。	Некоторые учителя вызывали Рэнхо к себе и делали замечания, но Хироко Икэда (75 лет), которая была её классным руководителем на втором и третьем году обучения, была не такой.
			「そんな薄いかばんじゃ遊び道具もえらないよ」「体育や部	Она разговаривает с девочкой мягко: «В такую плоск

図 1　アライメント

意味的に対応付けされた日本語原文とロシア語翻訳文は，共通の ID を与えることでその対応を明示化した[29]。

アライメントは文単位で取られたが，実際には言語学的に厳密に文を定義するのは容易ではない。本研究では，日本語原文については，『現代日本語書き言葉均衡コーパス』に対する係り受け・並列構造アノテーションである BCCWJ-DepPara（浅原・松本 2018）を利用し，EOS（End of Sentence）が振られている単位を 1 文とした[30]。翻訳対象が新聞記事のデータであることから，見出しやリストといった文書構成要素は単に名詞句の形式となっていることもある。それらは言語学的には文と考えてよいのか検討しなければならないが，上記の指針に沿って文とみなしている。ロシア語翻訳文については，実際に書かれたものを対象にアライメント作業を行っているため，正書法の観点から文の定義をすることが可能となる。つまり，単純に大文字で始まり，ピリオドで終わる単位を文と定義した。

当然，日本語の原文で 1 文であるものがロシア語の翻訳で 2 文となっていたり，またその逆であったりということが起こり得る。これについては，4.2 で対応を述べる。

[29] ID の付与については 4.2 で述べる。
[30] EOS がどのような基準で振られているかについては，浅原・松本（2018）および，その文境界について述べた小西他（2013）を参照されたい。

4.2 ID の付与

対応付けされた日本語原文とロシア語翻訳文は，共通の ID を与えることで日本語，ロシア語の各言語のテキストの対応付けが明示化された。つまり，日本語の起点テキストとロシア語の目標テキストは文レベルで同一の ID によって紐づけされている関係になる。

具体的には，文ごとに文番号を設定し，共通 ID として「BCCWJ のサンプル名 - 文番号」という形式の ID を付与している。例えば「PN1c_00001-05」であれば，サンプル「PN1c_00001」の 5 つ目の文を意味しており，この ID を与えられた日本語の原文とロシア語の翻訳文が意味的に対応することになる。

実際には，日本語の原文で 1 文であるものがロシア語の翻訳で 2 文となっていたり，またその逆に日本語原文で 2 文であるものがロシア語翻訳では 1 文となっていたりすることがあり得る。このような場合には，複数文になる方に対し文番号に加えてその下位の識別番号を設定し，それらをもとに識別可能な形とした。つまり，すでに述べた「サンプル名 - 文番号」により日本語原文とロシア語翻訳文の意味的な対応を保ちつつも，下位の識別番号により日本語原文とロシア語翻訳文がそれぞれ 1 文対 1 文で対応しているか，2 文対 1 文で対応しているかなどの統語的な識別が可能となる。

具体例として（19）を見られたい。

（19） a. [<PN1c_00001-25-1> 池田先生も，蓮舫さんにアドバイスしたことを覚えていた。]
[<PN1c_00001-25-2>「生意気という人もいたけれど，私は，彼女のようにモノをはっきり言えることがこれからは大切だと思っていました」。]

b. [<PN1c_00001-25> Učitel'nica Ikéda tože pomnit, kak davala
teacher-NOM.F Ikeda too remember-PRS.SG.3 how give-PST.F
Renxo sovety: «Nekotorye govorili, čto ona prosto naxalka,
Renho-DAT advice-ACC.PL some-PL say-PST.PL that she-NOM simply saucebox-NOM
no mne kazalos', čto teper' miru ponadobjatsja ljudi,
but I-DAT seem-PST.N that nowadays world-DAT be_needed-PRS.PL.3 people-NOM.PL
kotorye mogut vyskazyvat'sja otkrovenno – takie, kak
who-PL can-PRS.PL.3 express_one's_opinion-INF openly such-NOM.PL as
eta devočka».]
this-NOM.F girl-NOM.F （読売新聞 [BCCWJ: PN1c_00001]）

（19）は，日本語原文で 2 文であるものがロシア語翻訳文では 1 文となっている例である。（19a）の日本語原文には，それぞれ文単位で <PN1c_00001-25-1> という ID と <PN1c_00001-25-2> という ID が与えられ，（19b）のロシア語翻訳文には <PN1c_00001-25> という ID が与えられている。「PN1c_00001-25」が共通であるため，これらの日本語文とロシア語文が意味的に対応していることがわかる。日本語文に与えられている ID の末尾の 1, 2 が上記の識別番号である。これによ

り，ロシア語の「PN1c_00001-25」に対応する日本語文が 2 文あることが分かり，その前後関係についても番号で示されていることになる。

5. 日露対照研究への活用の可能性：日本語の文末表現とロシア語の対応形式

3 節で述べたロシア語翻訳データの構築，さらに，4 節で述べたアライメント作業の実施により，起点テキストと目標テキストは簡易的な日露パラレルコーパスとしての利用が可能となった。このパラレルコーパスは日露対照研究へ活用できると考えられる。以下では，その様な可能性の一例として，日本語の文末表現について，簡単にロシア語と対照させて論じる。

（20a）–（22a）[31] の日本語の各文の文末表現と（20b）–（22b）のロシア語におけるその対応部分（下線部）を見られたい。なお，日本語において名詞で文を終える体言止めを用いている箇所とそれに対応するロシア語の形式は太字で示されている。

（20） a. [...] 非常通報装置が**作動**。[...] 商品のビデオカメラ六十台とノートパソコン四台 [...] が盗まれていた。

b. [...] **srabotala** sistema signalizacii. [...] ukradeno 60
worked-PST.F system-NOM.F signaling-GEN stolen-PTCP.N 60-NOM
videokamer i 4 noutbuka iz čisla tovarov [...]
video_camera-GEN.PL and 4-NOM laptop-GEN.SG from number products-GEN
（中日新聞 [BCCWJ: PN4f_00001]）

（21） a. [...] 異国の食文化をどん欲に**吸収**。[...] パスタもレパートリーに加えた。[...] 心を奪われた。[...] 日本語学校にも通い始めた。

[31]（20b）の例文において，60 に続く *videokamer*「ビデオカメラ」は複数属格となっており，4 に続く *noutbuka*「ノートパソコン」は単数属格となっている。ロシア語において，名詞が数詞に後続する際，名詞の数と格は先行する数詞，および名詞を含む数詞句全体の格によって異なる。すなわち，数詞句全体が主格（とそれと同形の対格）の場合に限って言えば，数詞が 1 の時は名詞が単数主格となり，2 〜 4 の時は単数属格となり，5 以上の時は複数属格となる。ロシア語の数詞（句）は極めて複雑な振る舞いを見せるため，詳細はここでは述べない。また，60 *videokamer* i 4 *noutbuka*「60 台のビデオカメラと 4 台のノートパソコン」は受動分詞 *ukradeno*「盗まれた」に対する主語になっているが，*ukradeno* は単数中性形となっている。（1 を除く）数詞，数量詞を含む句が主語となる場合，過去時制では述語の形態は中性形か複数形があり得る。つまり，以下の（vi a）も（vi b）も共にあり得る文となる。

（vi） a. Prišlo neskol'ko čelovek.
arrive-PST.N some-NOM people-GEN.PL
b. Prišli neskol'ko čelovek.
arrive-PST.PL some-NOM people-GEN.PL
「数人が来た。」

Graudina et al.（1976）では，新聞において（vi a）のパターンが 74.57% を占め，（vi b）のパターンが 25.43% を占めると述べられているが，数詞，数量詞を含む句が主語となる場合の述語形態には様々な要因が関係しており，ここで詳細を述べることはできない。これについての数量的な調査は Graudina et al.（1976）や Akiyama（2014）等を参照のこと。

b. [...] on [...] žadno **vpityval** kulinarnye tradicii čužoj
he-NOM.M greedily absorb-PST.M culinary-ACC traditions-ACC foreign-GEN
strany. On takže dobavil v svoj kulinarnyj
country-GEN he-NOM.M also added-PST.M into self's-ACC culinary-ACC
repertuar pastu [...] on byl neverojatno očarovan [...]
repertory-ACC pasta-ACC he-NOM.M was-PST.M unbelievably fascinated-PTCP.M
on daze načal poseščat' školu japonskogo jazyka.
he-NOM.M even start-PST.M visit-INF school-ACC Japanese-GEN language-GEN
（読売新聞 [BCCWJ: PN4c_00001]）

（22） a. 日本政府は [...] 無形文化遺産保護条約を締結した。[...] 佐藤禎一大使が [...] 締約受諾書を提出した。同条約は [...] 採択された。締結はアルジェリアなどに続いて**三カ国目**。

b. [...] pravitel'stvo Japonii prinjalo Konvenciju ob
government-NOM.N Japan-GEN accepted-PST.N convention-ACC about
oxrane nematerial'nogo kul'turnogo nasledija, [...] Posol [...]
protection-LOC intangible-GEN cultural-GEN heritage-GEN ambassador-NOM.M
Téjiti Sato pred"javil [...] document o soglasii na
[Teiichi Sato]-NOM.M presented-PST.M document-ACC about agreement-LOC on
prinjatie konvencii. Èta konvencija byla
acceptance-LOC convention-GEN this-NOM.F convention-NOM.F was-PST.F
utveržden [...] Vsled za Alžirom i t.d. Japonija **stala tret'im**
approved-PTCP.F following Algeria and so on Japan-NOM.F became-PST.F third-INS
gosudarstvom, prinjavšim konvenciju.
nation-INS taking-PTCP.INS convention-ACC （西日本新聞 [BCCWJ: PN4g_00001]）

（20a）–（22a）の日本語文の文末表現に注目すると，名詞で文を終える体言止めを用いている箇所がある。一方，ロシア語では日本語文で体言止めとなっている箇所の対応部分は（20b）–（22b）で示すように動詞の過去形（*-l/-la*）となっている。

もし，日本語文で体言止めとなっている箇所を完全な文の形式とするのであれば，動名詞（verbal noun）で体言止めにされている（20a），（21a）では「- した」を追加し，（22a）では例えば「- であった」とコピュラを追加することになろう。このようにした場合，例えば（21a）は（23）のようになり，文末では過去形のみが続き「タ」が連続することで文章が単調になってしまう。

（23） [...] 異国の食文化をどん欲に吸収した。[...] パスタもレパートリーに加えた。[...] 心を奪われた。[...] 日本語学校にも通い始めた。

このような「タ」の連続による単調さを避けるために，適宜体言止めが用いられているものと考えられる[32]。ロシア語にも，日本語の動名詞を用いた構文のように体系的に動詞を名詞化させる方法は存在するが，（20b）–（22b）で示すように，そのような形式はここでは用いられていない。（20a）–（22a）の日本語文の文末表現に対応する部分は（20b）–（22b）のロシア語では 1 例を除き全て動詞の過去形（*-l/-la/-lo*）となっている[33]が，ロシア語の基本語順は SVO であり（Isačenko 1966 など），基本的に述語動詞は文末に位置しない[34]ため，動詞の時制により文末の形式が固定されることはない。そのため，文末表現を多様化するという日本語のような理由では，動詞の名詞化の構文は用いられないと考えられる。

以上のように，ロシア語では，文末に述語が位置しないことが多く，動詞の時制が文章内で同一であっても文末の形式が固定されることはない。一方，日本語では述語がほぼ必ず文末に位置するために，動詞の時制が過去で統一されてしまうと，文末の形式が「タ」に固定されてしまう。日本語では過去形の連続を避け，文末表現を多様化するために，体言止めが適宜用いられるといえる。もちろん，これについて確実な結論を出すためにはより広範で詳細な検討が必要となるが，ケーススタディとして翻訳データの活用法を示すという目的は果たせたものと思われる。

6. おわりに：まとめと今後の展望

本稿では，BCCWJ のロシア語翻訳データの構築について述べた。

BCCWJ の新聞（PN）コアデータ 16 サンプルを対象に日本語からロシア語へ人手による翻訳を行った結果，ロシア語翻訳データの総語数は 13,070 語となり，既に構築されていた英語，イタリア語，インドネシア語，中国語の各翻訳データと比べると最大の規模となった。

翻訳の際は，日本語文の時制をロシア語文でも用いること，企業名等はラテン文字で表記すること，固有名詞にはできる限り上位概念を示す名詞を付加すること，「東京・大阪・京都・広島」などのロシア人もよく知る地名に限り上位概念を示す名詞を付加しないこと，等を方針とした。また，原文の日本語と翻訳先のロシア語のアライメントは文単位で取っており，それぞれにサンプル名と文番号を合わせた ID が付与された。

---

[32] 翻訳対象は PN コアデータであり，出典のテキストが新聞であるため，字数の制限等も関係してくる可能性がある。また，これについて結論を出すためにはより詳細な検討が必要となる。
[33]（20b）の *ukradeno*「盗まれた」のみ受動分詞である。
[34] ただし，ロシア語は文の要素の語順は自由であるため，それぞれのニュアンスは異なるものの，以下（vii）で示すように，主語，述語動詞，目的語（対格補語）の 3 つの要素であれば 3!（=6）通りの語順すべてが文法的となる。よって（vii b）や（vii e）のように述語動詞を文末に位置させることも可能ではある。

（vii）a.	Ivan 1	jubit	Mariju.	(SVO)
	Ivan.NOM	loves-PRS.3.SG	Maria-ACC	
b.	Ivan Mariju ljubit.			(SOV)
c.	Ljubit Ivan Mariju.			(VSO)
d.	Ljubit Mariju Ivan.			(VOS)
e.	Mariju Ivan ljubit.			(OSV)
f.	Mariju ljubit Ivan.			(OVS)

「イヴァンはマリアを愛している。」

コーパス研究が盛んである今日でも日本語・ロシア語のパラレルコーパスは希少であり，BCCWJ の 16 サンプルであっても，その基礎となり得るデータを構築したことは日露対照研究，さらには類型論研究に対し一定の意義のある言語資料を提供できたといえる。本稿では，翻訳データの日露対照研究への活用の一例として，日本語の文末形式について，簡単にロシア語と対照させて論じた。ロシア語とは異なり，日本語の新聞には体言止めがしばしば使用されることを指摘し，その要因は文末形式を多様化するためであるとした。

Soejima（2017）は文学作品とその翻訳作品を用いて日露語のパラレルコーパスを構築し，不特定の動作主が関わる意図的な出来事が日本語とロシア語でどのように表現されるかについて検討している。その結果，過程を表す場合は，日本語では受動文がよく用いられる一方でロシア語では不定人称文[35]がよく用いられるとしている。結果を表す場合は，日本語では自動詞文がよく用いられ，ロシア語では受動文や自動詞文など多様な形式が用いられるとしている。Soejima（2017）は文学作品でこの結論を導いているため，新聞という異なるレジスターでも同様の結果が得られるか今後検討する必要がある。

さらに，今回ロシア語に翻訳した日本語のサンプルは情報構造のアノテーションもなされている（Miyauchi et al. 2018，宮内他 2018）。よって，今回構築したロシア語翻訳データにも情報構造のアノテーションを施せば，日本語・ロシア語の情報構造についての対照研究が定量的に行えるようになる[36]。日本語もロシア語も共に顕在的な冠詞のない言語であるため，定性や特定性などの冠詞を持つ言語では冠詞が担う機能をどのように表現するか（またはしないか）[37]を今後詳細に検討したい。

[35] 不定人称文とは，動作の主体に注目せず，動作そのものや・動作の客体に主眼を置いた文である。（viii）で示すように，形式的に主語はなく述語動詞は三人称複数の形態をとる（AN SSSR 1980）。

（viii）a. V Kanade govorjat po-francuzski.
in Canada-LOC speak-PRS.3.PL in_French
「カナダではフランス語が話されている。」
b. Mne zvonili?
me-DAT call-PST.PL
「私に電話がありましたか？」

[36] さらに，Asahara（2017）は，情報構造と読み時間の関係を考察している。ロシア語のデータに対するアノテーションを充実させれば，同様の研究が可能となり，Asahara（2017）で示された成果が日本語特有のものか通言語的なものかを調べることも可能となる。

[37] 一般にロシア語は定性を語順によって表現する傾向があるとされる（Chvany 1973 など）が，あくまで傾向に過ぎない。また，否定属格の現象も定性との関係がよく指摘される（Timberlake 1975 など）が，これも定性以外の要因も影響を与え得るため，はっきりとした規則ではない。

参照文献

Akiyama, Shinichi (2014) Corpus-based analyses of subject and predicate concord in modern Russian. *Rosiago Kenkyuu* 24: 71–89.

AN SSSR (1980) *Russkaja Grammatika*. Moskva: Nauka.

Apresjan, Jurij D. (1988) Glagoly momental'nogo dejstvija i performativy v russkom jazyke. In: Ju N. Karaulov (ed.) *Rusistika segodnja. Jazyk: sistema i ee funkcionirovanie*. 57–78. Moskva: Nauka.

浅原正幸・森田敏生（2015）「コーパスコンコーダンサ『ChaKi.NET』のプロジェクト機能」『第 7 回コーパス日本語学ワークショップ』103–112.

Asahara, Masayuki (2017) Between reading time and information structure. *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation PACLIC 31*.

浅原正幸・松本裕治（2018）「『現代日本語書き言葉均衡コーパス』に対する文節係り受け・並列構造アノテーション」『自然言語処理』25(4): 331–356.

Balašova, Ljubov' V. and Vadim V. Dement'ev (2005) *Kurs russkogo jazyka*. Saratov: Licej.

Chvany, Catherine V. (1973) Notes on 'root' and 'structure-preserving' in Russian. In: Caudi W. Corum, Thomas C. Smith-Stark and Ann Weiser (eds.) *You take the high node and I will take the low node*. 252–290. Chicago, IL: Chicago Linguistic Society.

Dutkina, Galina B. (2003) Medvežij bog. *Ona: Novaja japonskaja proza*. Moskva: Inostranka.

Ermolovič, Dmitrij I. (2001) *Imena sobstvennye na styke jazykov i kul'tur*. Moskva: R.Valent.

Graudina, Ljudmila K., Viktor A. Ickovič and Lija P. Katlinskaja (1976) *Grammatičeskaja pravil'nost' russkoj reči: Opyt častotno-stilističeskogo slovarja variantov*. Moskva: Nauka.

Isačenko, Aleksandr V. (1966) O grammatičeskom porjadke slov. *Voprosy Jazykoznanija* 6: 27–34.

川上弘美（2006）『真鶴』東京：文藝春秋.

川上弘美（1998）『神様』東京：中央公論社.

小西光・小山田由紀・浅原正幸・柏野和佳子・前川喜久雄（2013）「BCCWJ 係り受け関係アノテーション付与のための文境界再認定」『第 4 回コーパス日本語学ワークショップ予稿集』135–142.

Maekawa, Kikuo, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka and Yasuharu Den (2014) Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation* 48: 345–371.

Markova, Viktorija V. (2013) *Stilističeskaja pravka: Učebnoe posobie*. Tjumen': Izdatel'stvo Tjumenskogo gosudarstvennogo universiteta.

Mironova, Ljudmila O. (2012) *Manadzuru*. Moskva: Giperion.

Miyauchi, Takuya, Masayuki Asahara, Natsuko Nakagawa and Sachi Kato (2018) Information-structure annotation of the "Balanced Corpus of Contemporary Written Japanese". *Computational Linguistics: Communications in Computer and Information Science* 781: 155–165. Singapore: Springer.

宮内拓也・浅原正幸・中川奈津子・加藤祥（2018）「『現代日本語書き言葉均衡コーパス』への情報構造アノテーションとその分析」『国立国語研究所論集』16: 19–33.

宮内拓也・プロホロワ マリア（2018）「『現代日本語書き言葉均衡コーパス』のロシア語翻訳データの構築」『言語資源活用ワークショップ 2018 発表論文集』2–11.

野田尚史（1992）「テンスから見た日本語の文体」文化言語学編集委員会（編）『文化言語学：その提言と建設』592–579. 東京：三省堂.

Prokopčuk, Klavdija A. (2017) Inostrannye pragmatonimy v sovremennoj russkoj literature: Latinica ili kirillica? *Studi Slavistici* 14: 309–327.

Soejima, Kensaku (2017) On expressions of agent de-topicalized intentional events: A contrastive study between Japanese and Russian. *Journal of Japanese Linguistics* 30: 107–128.

Timberlake, Alan (1975) Hierarchies in the genitive of negation. *The Slavic and East European Journal* 19: 123–138.

Voroncova, Tat'jana A. (2011) *Kul'tura reči: Učebnoe posobie*. Iževsk: Udmurtskij gosudarstvennyj universitet.

山崎誠（2011）「『現代日本語書き言葉均衡コーパス』の設計」国立国語研究所コーパス開発センター『『現代日本語書き言葉均衡コーパス』マニュアル』2-1–2-8.

Construction of Russian Translation Data for the "Balanced Corpus of Contemporary Written Japanese" and the Possibilities of Using Them in Japanese-Russian Comparative Studies

MIYAUCHI Takuya[a] PROKHOROVA Maria[b]

[a]The University of Tokyo / Project Collaborator, NINJAL
[b]Ph.D. Student, Tokyo University of Foreign Studies

Abstract

A part of the data of the "Balanced Corpus of Contemporary Written Japanese" (BCCWJ) is translated into English, Italian, Chinese, and Indonesian. We added new translation data collected from 16 samples of newspaper (PN) core data to BCCWJ in Russian. The total length of the Japanese source text is 16,657 short unit words, which corresponds to 13,070 words in the Russian target text. The translation was conducted manually by a native Russian speaker. During the translation, various difficulties were encountered due to significant structural and lexical differences between Japanese and Russian. This study introduces the data construction method that we used and some key points that we focused on while translating. We also manually aligned all sentences in the source text with those in the translation and assigned an ID to each sentence; this study provides an explanation regarding this workflow as well. Translation and alignment make the original data and their translation function as a simple Japanese-Russian parallel corpus. This can be useful for Japanese-Russian comparative studies and linguistic typology studies. In this study, we address Japanese sentence endings and compare them with Russian ones as a case study to present the possible ways of using our new translation data.

Key words: "Balanced Corpus of Contemporary Written Japanese," parallel corpus, Russian, expressions at the end of sentences