

名詞句接続の日韓対照研究

— 「NP₁のNP₂」と「NP₁의NP₂」の制約の相違と相対的頻度—
**A Contrastive Study on the Adnominal Constructions with Japanese
 -no and Korean -ui: Their Relative Frequencies and Constraints**

首藤 佐智子[†], 崔 在雄[‡], 原田 康也[†]
 Sachiko Shudo, Jae-Woong Choe, Yasunari Harada

[†]早稲田大学法学学術院, [‡]高麗大学校言語学科
 Faculty of Law, Waseda University, Department of Linguistics, Korea University
 shudo@waseda.jp, jchoe@korea.ac.kr, harada@waseda.jp

Abstract

Similarities between Japanese and Korean are “remarkable” when we consider that the two languages may not be directly related historically the way Spanish and Italian are. One example of such similarities manifests itself in the way two noun(phrases) are combined to form a larger nominal expression, with particles ‘-no’ in Japanese and ‘-ui’ in Korean. However, when we take a careful look into the distribution and relative frequency of the adnominal constructions against noun-noun compounds, we notice an interesting difference between the two languages. Here we report a preliminary result of our survey using a small parallel corpus, corroborated with data from larger monolingual corpora.

Keywords — **adnominal constructions, frequencies, parallel corpus, Japanese, Korean**

1. はじめに^{i ii}

日本語の「NP₁のNP₂」におけるNP₁の指示対象とNP₂の指示対象の意味的な関係に関しては多くの研究がなされている。これを「所有格」ないし「属格」的な関係とみなしたうえで、英語の「s」との比較において「所有」を大きく逸脱するものであることも指摘されているが、この比較は英語の「s」に対応する日本語が「の」であるという一方通行の対応関係にもとづくものであり、「NP₁のNP₂」という名詞句接続の構造あるいは「NP₁の」 という修飾表現の分析には寄与するところが少ない。本研究では、統語法において日本語の「の」と顕著な並行性を有する韓国語の「-ui (의) (以下-ui)」との数量的な比較を行い、その相違を考察する。本研究では、その出現の数量的な相違を示すことを主な目的としたため、データとして、

比較的小規模なタグ付き二言語パラレルコーパスの分析を行った。さらに、それぞれの言語のより大規模なコーパスを用いて、分析の補完とした。

2. 「の」と「-ui」

日本語の「NP₁のNP₂」と韓国語の「NP₁-ui-NP₂」においては、「NP₁の」「NP₁-ui」がNP₂を修飾する。この2つの表現はそのまま双方向に翻訳できることも多いが、NP₁ NP₂ という「の」「-ui」を伴わない接続句の可能性により、一概に双方向に翻訳が可能とは言えない場合が生じる。Chung, et al. (1999) は韓国語から日本語に翻訳する際に韓国語のNP₁NP₂接続句(-uiを伴わない形)を日本語のNP₁NP₂接続句に変換することができず、「の」を挿入する必要があることを指摘する。これは、日本語のNP₁NP₂接続句には韓国語のそれよりも厳しい制約があることを示唆している。本稿では詳述しないが、日本語のNP₁NP₂の制約に関しては、語源種別(例えば「大和言葉」「漢語」「外来語」)にゆるく依拠する制約や、個人による一時的な使用がメディア等を通じて広範囲に認識され、容認度が向上するというような社会言語学的現象も影響する。一方、韓国語において-uiを伴わないNP₁NP₂接続句で表される表現のうちの一部は、NP₁-ui-NP₂でも表すことができるが、-uiを伴う形をとれない表現も多い。

これまでの研究を概観した限りでは、日韓の名詞句接続に関する定量的な対照研究はされていない。本研究では二言語のパラレルコーパスを用いて逆相対的頻度を表し、その制約を考察する上で

の数量的な足がかりを得ることを目的とする。 ⁱⁱⁱ

ファイルのサンプルを示したものである。

3. データ

本研究で用いたデータは Sejong (世宗) 韓日パラレルコーパス (以下「Sejong パラレルコーパス」) と呼ばれるもので、小規模ながら日韓の並行したデータから構成されている。その内訳としては、韓国語で書かれた文献が 21.6%，それを日本語に翻訳したものが 19.2%，日本語で書かれた文献が 27.5%，それを韓国語に翻訳したものが 31.6%となっている (図 1)。

Sejong パラレルコーパスのジャンル構成は、小説が 45.5%，新聞が 27.8%，雑誌が 14.8%，その他の資料が 11.9%となっている (図 2)。

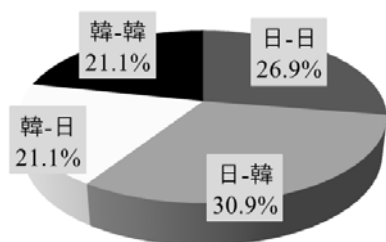


図 1: Sejong 韓日パラレルコーパスの構成

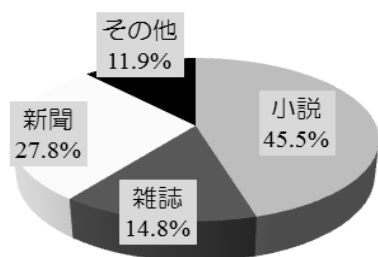


図 2: Sejong 韓日パラレルコーパスのジャンル

言語の対照研究を行う上でしばしば問題となるのは、それぞれの言語におけるデータの相違が分析に影響を与えてしまう可能性である。本研究の場合は、データのジャンルやトピック、内容等が接続名詞句の使用に影響を与える可能性が考えられるが、パラレルコーパスを使うことによって、そのような他の要因を最小限に抑えることが期待できる。

表 1 は Sejong 韓日パラレルコーパスの個別フ

韓国語		日本語	
<p id=1.1.p1>		<p id=1.1.p1>	
<s id=1.1.p1.s1>		<s id=1.1.p1.s1>	
한	한/MM	—	—/NNR
시간의	시간/NGG+	時間	時間/NADP
	의/JKG	の	の/PCS
기적,	기적/NGG+	奇跡	奇跡/NG
	,	,	,/SYP
바다를	바다/NGG+	海	海/NG
	를/JKO	を	を/PJKG
걷는다!!	걸/VV+	歩こ	歩く/VIN
	는다/EF+	う	う/AU
	!/SF+	!	!/SYG
	!/SF	!	!/SYG
</s>		</s>	
</p>		</p>	

表 1: Sejong 韓日パラレルコーパスのサンプル

Sejong 韓日パラレルコーパスのサイズを文単位で表すと、日本語のデータは 4,045，韓国語のデータは 4,038 で、差はほとんどない。一方、形態素を単位にすると、日本語のデータは 89,083，韓国語のデータは 101,548 と大きく異なる。これには、タグ付けによる相違を反映している可能性が高い。表 2 と表 3 は、日韓それぞれのデータの

日本語		
cat	出現数	%
N	29903	33.60%
P	24090	27.00%
S	11539	13.00%
V	11009	12.40%
A	10869	12.20%
C	1402	1.60%
F	170	0.20%
E	100	0.10%
U	1	0.00%
sum	89083	100.00%

表 2

韓国語		
cat	出現数	%
N	29484	29.00%
E	19993	19.70%
J	17109	16.80%
V	13057	12.90%
S	10351	10.20%
X	7147	7.00%
M	4307	4.20%
I	100	0.10%
sum	101548	100.00%

表 3

形態素をカテゴリー別に分類し、頻度の多い順に示したものである。日本語の名詞 (表 2 の N) と

韓国語の体言（表3のN）の出現数は、それぞれ29,903 (33.60%)と29,484 (29.0%)で、大きな差は見られない。一方、日本語の助詞(表のP)が27.0%を占めるのに対し、韓国語でこれに対応するとされる関係言（表のJ）と呼ばれるカテゴリーは16.80%であり、これは明らかにカテゴリーが一对一の対応関係にないことを示している。

4. 「の」と「-ui」の頻度

それでは、まずデータにおける「の」と「-ui」の出現を見てみよう。表4と表5は、日韓それぞれのデータにおける形態素を出現数の高い順に示したものである。日本語のデータでは、形態素の総数89,083に対して、「の」の出現総数は3841回であり、逆相対頻度としては、23.2 : 1となる（表6）。韓国語のデータでは、形態素の総数101,548に対して、「-ui」の出現は1794回で、逆相対頻度は、56.6 : 1となる（表6）。名詞の出現総数に対する「の」と「-ui」の逆相対頻度は、それぞれ7.8:1と16.4:1という数字になっている。「の」と「-ui」をそれぞれ伴う名詞句接続を極めて単純に数量で比較すると、明らかに韓国語の-*ui*を伴わない名詞句接続は日本語のそれよりも頻度が低い。

string	範疇	出現
.	SYP	4636
の	PCS	3841
。	SYF	3364
た	AU	3078
は	PRE	2796
を	PJKG	2678
に	PJKG	2429
だ	AU	2184
て	PJC	2136
する	VIN	2009
が	PJKG	1946
私	NNPG	898
「	SYPO	877

表4: 日本語

string	範疇	出現
.	SF	3757
다	EF	2584
ㄴ	ETM	2174
하	XSV	2016
을	JKO	1931
에	JKB	1866
의	JKG	1794
는	JX	1495
이	VCP	1443
는	ETM	1382
이	JKS	1339
,	SP	1319
은	JX	1181

表5: 韓国語

	日本語	韓国語
形態素総数	89,083	101,548
名詞総数	29,903	29,484
「の」「-ui」の出現総数	3,842	1,795
形態素数/「の」「-ui」	23.2	56.6
名詞数/「の」「-ui」	7.8	16.4

表6: 「の」と「-ui」の出現頻度（逆相対頻度）

Sejong パラレルコーパスでは各言語につき50のファイルが存在する。表7と表8は、日本語と韓国語のファイルから、それぞれ8つのサンプルの状況と、各言語50のファイル全体の平均と標準偏差を示している。「の」と「-ui」の頻度を見ると、日本語のファイルにおける「の」の頻度が比較的安定しているのに比べ、韓国語ではファイルごとに大きな差が見られる。「の」は平均して形態素22.0に1回という逆相対頻度で、その標準偏差は4.9である。これに対して、「-ui」は平均して形態素50.4に1回の逆相対頻度で、標準偏差が31.8という高い数字になっている。ファイルによってこれだけ数値にバラつきがあるという事実は韓国語の「-ui」が文献のジャンルや内容によって影響を受けやすいことを示唆していると言えよう。

file	morph	nn	の	nn/の	mor/の	mor/nn
6JT_12kj	1090	451	74	6.1	14.7	2.4
6JT_13kj	1068	472	64	7.4	16.7	2.3
6JT_14kj	942	400	58	6.9	16.2	2.4
6JT_15kj	661	270	43	6.3	15.4	2.4
6JT_24kj	1008	424	47	9	21.4	2.4
6JT_25kj	1292	651	61	10.7	21.2	2
6JT_26kj	383	176	16	11	23.9	2.2
6JT_27kj	1138	466	50	9.3	22.8	2.4
50 files
総数	89083	29903	3842			
平均	1781.7	598.1	76.8	8.5	22	2.6
標準偏差	4601.7	1226.3	177.4	1.9	4.9	0.4

表7: 日本語データのサンプルと分布

file	morph	nn	의	nn/의	mor/의	mor/nn
6jt_12kk	1072	373	39	9.6	27.5	2.9
6jt_13kk	1062	375	36	10.4	29.5	2.8
6jt_14kk	1089	399	37	10.8	29.4	2.7
6jt_15kk	672	243	25	9.7	26.9	2.8
6jt_24kk	906	274	13	21.1	69.7	3.3
6jt_25kk	1405	500	20	25	70.3	2.8
6jt_26kk	357	144	2	72	178.5	2.5
6jt_27kk	1141	390	7	55.7	163	2.9
50 files
総数	101548	29484	1795			
平均	2031	589.7	35.9	16.4	50.4	3.1
標準偏差	5493.6	1341.2	61.2	11.3	31.8	0.4

表 8: 韓国語データのサンプルと分布

韓国語のファイルではなぜそのように標準偏差が高い数値なのであろうか。表 8 にサンプルとしてあげた 6jt-12kk, 6jt-13kk, 6jt-14kk, 6jt-15kk の 4 つのファイルと, 6jt-26kk, 6jt-27kk の 2 つのファイルをざっと見比べると, 形態素に対する名詞の逆相対頻度はほぼ同等であるにも関わらず, -ui の出現頻度には大きな違いがあることがわかる。前者の形態素に対する逆相対頻度は全て 20 台であるが, 後者のそれは 150 を超えている。一つの説明として, 前者は小説からのデータで, 後者は新聞からのデータであるという可能性がある。対応する日本語のデータ (表 7) を見ると, 日本語でも前者のファイルと後者のファイルにおける「の」の出現頻度には差が見られるが, 韓国語のデータにおけるほどの顕著な差ではない。韓国語における「-ui」の使用に影響を与える要因として, ジャンルあるいは文体という要素を検証することは, 今後の課題である。

5. 補完的データ

パラレルコーパスを使用するメリットに関しては既に述べたが, 我々のデータが小規模であることの問題点は否めない。この欠点を補うべく, パラレルコーパスに加えて, 個々の言語のコーパス

を同様に分析し, 数量の有効性を検証した。日本語は京都大学テキストコーパス 4.0 (以下「京都大学コーパス」) を使用し, 韓国語は意味タグ付き Sejong コーパス (以下「Sejong コーパス」) を使用した。

	日本語	韓国語
a. データ	Kyoto4.0	SJ_semTag*
b. ジャンル	新聞	新聞
	毎日新聞	(主に)朝鮮日報
c. ファイル数	28	255
d. 文単位	38,400	573,447
e. 形態素	972,894	15,425,299
f. e/d	25.3	26.9
g. 標準偏差	14	17.2
h. 「の」「-ui」	53,241	342,650
i. h/d	1.4	0.6
j. 標準偏差	1.3	0.9
k. e/h	18.3	45
l. NNG&NNP	356,629	3,709,032
m. l/h	6.7	10.8

表 9: 大規模データの分布

京都大学コーパスと Sejong コーパスを利用した数量的分析でも, 「の」が現れる頻度は「-ui」よりもはるかに多く, パラレルコーパスの結果を裏付ける結果が得られた (Table 9)。京都大学コーパスの「の」の逆相対頻度は形態素 18.3 に 1 回で, パラレルコーパス (22.0 に 1 回) よりもやや高い。Sejong コーパスの「-ui」の逆相対頻度は, 形態素 45.0 に 1 回で, これも Sejong パラレルコーパス (50.4 に 1 回) よりも高くなっている。京都大学コーパスの「の」の頻度と Sejong コーパスの「-ui」の頻度の差は Sejong パラレルコーパスよりもむしろ大きい。

6. 考察

日本語の「の」が韓国語の「-ui」よりも頻度が高いことを説明する要因は様々である。一つの要因は, 前述したように, 日本語の NP₁NP₂ 接続句には韓国語のそれよりも厳しい制約があることで

ある。すなわち、助詞を伴わない NP₁NP₂ の名詞句連接が各々の言語でどのように制約されているかによって、NP₁ の NP₂ と NP₁-ui-NP₂ の出現に違いが出るという要因である。

Nishiguchi (2009) は日本語の「NP₁ の NP₂」における NP₁ と NP₂ の意味論的關係を 11 のタイプに分類した (表 10)。

Relation	Japanese Possessive
I possession	Naomi-no kaban
II part-whole	Naomi-no kao
III location	Tokyo-no shinseki
IV time	yugata-no koen
	natsu-no kyuka
	7-ji-no nyusu
V accompaniment	kaban-no hito
	boshi-no fujin
VI trade	Kaban-no Kochi
	Ningyo-no Morishige
VII activity	Maaruboro-no kuni
	biiru-no machi
VIII property	chisee-no hito
	osu-no tora
	aoi-me-no ningyo
	tsutsuji-no koen
IX quantity	1-kiro-no pasokon
	3-bon-no pen
X intensional property	nise-no fukahire
	nise-no keisatsukan
XI kind	osu-no tora

表 10: Nishiguchi (2009) から

本稿の筆者たちは、「NP₁ の NP₂」における N1 と N2 の意味論的關係は、Nishiguchi のタイプのようなカテゴリーによって網羅的に示されるようなものであるとは考えないが、Nishiguchi のタイプは、「NP₁ の NP₂」が助詞「の」を伴わない「NP₁NP₂」で表すことができるかを NP₁ と NP₂ の多様な関係に鑑みて検証するためには効果的な

リストである。表 11 は表 10 の表現を「の」を伴わない「NP₁NP₂」で表すことができるかを示したものである。ごく一部の例外を除いて、名詞句連接において「の」を省くことはできないことが明らかである。前述したように、日本語の NP₁NP₂ 連接句は、個人や小集団による一時的な使用がメディア等を通じて広範囲に認識され、容認度が向上するというような社会言語学的現象に大きく影響を受ける。例えば、「干物女」(ひものおんな) という言葉は、漫画で使われ、その漫画がテレビドラマ化されたことから、「恋愛を放棄している、様々な事を面倒くさがり、適当に済ませてしまう女性のこと」(ウィキペディア) を指す一般語として使われるようになった。流行時の 2007 年頃には頻繁に使われ、複合語としてある程度の定着を見せた。しかしながら、このようなケースは稀であり、日本語の NP₁NP₂ に対する制約は韓国語に比べてきびしいことは明らかである。

R	NP ₁ の NP ₂	NP ₁ NP ₂
I	Naomi-no kaban	*Naomi kaban
II	Naomi-no kao	*Naomi kao
III	Tokyo-no shinseki	*Tokyo shinseki
IV	yugata-no koen	*yugata koen
	natsu-no kyuka	*natsu kyuka
	7-ji-no nyusu	*7-ji nyusu
V	kaban-no hito	*kaban hito
	boshi-no fujin	??boshi fujin
VI	Kaban-no Kochi	*Kaban Kochi
	Ningyo-no Morishige	*Ningyo Morishige
VII	Maaruboro-no kuni	*Maaruboro kuni
	biiru-no machi	??biiru machi
VIII	chisee-no hito	*chisee hito
	osu-no tora	osu tora
	aoi-me-no ningyo	*aoi-me ningyo
	tsutsuji-no koen	?tsutsuji koen
IX	1-kiro-no pasokon	?1-kiro pasokon
	3-bon-no pen	*3-bon pen
X	nise-no fukahire	nise fukahire
	nise-no keisatsukan	nise keisatsukan
	osu-no tora	osu tora

表 11: 日本語の NP₁NP₂

それでは、表 10 における「N₁の N₂」を韓国語の「NP₁-ui-NP₂」で表した場合の容認度はどうであろうか。

R	NP ₁ -ui-NP ₂	NP ₁ NP ₂
I	Naomi-ui gabang	Naomi gabang
II	Naomi-ui eolgul	Naomi eolgul
III	Tokyo-ui chincheog	Tokyo chincheog
IV	? jeonyeog-ui gong-won	jeonyeog gong-won
	? yeoleum-ui banghag	Yeoleum banghag
	(?)7si-ui nyuseu	7si nyuseu
V	*gabang-ui salam	?? gabang salam
	?? moja-ui yeoin	? moja yeoin
VI	*gabang-ui kochi	*gabang kochi
	*inhyeong-ui molisige	! Inhyeong molisige
VII	Malbolo-ui nala	Malbolo nala
	maegju-ui dosi	maegju dosi
VIII	*jiseong-ui salam	*jiseong salam (cf. jiseongin)
	? susnom-ui holangi	susnom holangi
	puleunnun-ui inhyeong	puleunnun inhyeong
	? jindallae-ui gong-won	jindallae gong-won
IX	1 kg-ui computer	1 kg computer
	sejalu-ui pen	sejalu pen
X	? gajja-ui sang-eojineuleomi	gajja sang-eojineuleomi
	?? bunjang-ui gyeongchalgwon	! bunjang gyeongchalgwon
XI	? susnom-ui holangi	susnom holangi

表 12: 韓国語の NP₁-ui-NP₂ と NP₁NP₂

表 12 は、左欄が NP₁-ui-NP₂ を、右欄が NP₁NP₂ を示している。興味深いのは、少なくともこのサンプルでは、「-ui」を伴わなければならないもの（右欄の形式だけが容認されていない場合）がなかったという点である。これは日本の「の」を伴う形式と伴わない形式との対立に比べて、大きく異なる。また、「-ui」を伴うことができないもの

（左欄の形式だけが容認されていない場合）が多く存在しているのも日本語の「の」の場合とは決定的な違いである。「-ui」を伴ってもいいが、省くこともできるもの（左右欄の形式が共に容認されている場合）があることは、特筆すべき事実ではないが、「-ui」を伴った形式も伴わない形式も容認されないもの（左右欄共に容認されていない場合）があるということも興味深い。

原田・首藤 (2009) では「NP₁の NP₂」における NP₁ と NP₂ の関係を説明する意味論的な制約は存在せず、その関係は語用論的に制約されていたとした。しかしながら、韓国語の NP₁-ui-NP₂ においてはより厳しい制約が存在している可能性がある。タイプ V の「かばんの人」、タイプ VI の「かばんのコーチ」、タイプ VIII の「知性の人」に対応する韓国語の「-ui」を伴った形は、語用論的に関係が説明されることが期待できるはずだが、容認されていない。11 タイプのうち「-ui」を伴う形が容認されたのは 5 タイプ (I, II, III, VII, IX) の例のみである。他のタイプの例では「NP₁-ui-NP₂」を容認しないという制約が存在するという事実は、本稿のデータで検証した「NP₁の NP₂」と「NP₁-ui-NP₂」の頻度の差をある程度説明することができる。

7. まとめ

本研究では、統語法において日本語の「の」と顕著な並行性を有する韓国語の「-ui」との計数的な比較を行い、その相違を考察した。データとして、第一に日韓の並行したデータから構成されている Sejong 韓日パラレルコーパスを対象として形態素の出現度数の分析を行った。「の」と「-ui」をそれぞれ伴う名詞句接続を極めて単純に数量で比較すると、明らかに韓国語の「-ui」を伴う名詞句接続は日本語の「の」を伴う名詞句接続よりも頻度が低い。また、日本語・韓国語の単一言語コーパスとして京都大学テキストコーパス 4.0 と意味タグ付き Sejong コーパスを使用して補完的な分析を行った。ここでも「の」が現れる頻度は「-ui」よりもはるかに多く、パラレルコーパスの結果を

裏付ける結果が得られた。今後の課題としては、「の」「-ui」を伴う接続句と伴わない接続句に対する制約をジャンルや文体を変数として体系的に比較することで、各々の接続句の分析に貢献することが期待できる。

データ

- [1] Sejong 韓日パラレルコーパス(世宗韓日パラレルコーパス), <http://www.sejong.or.kr/>
- [2] Sejong コーパス(世宗コーパス)(意味タグ付き), <http://www.sejong.or.kr/>
- [3] 京都大学テキストコーパス 4.0
<http://nlp.kuee.kyoto-u.ac.jp/nl-resource/corpus-e.html>

参考文献

- [4] Choe, Jae-Woong, Sachiko Shudo, and Yasunari Harada, (2011) "A Contrastive Study on the Adnominal Constructions in Japanese and Korean: Relative Frequency of '-no' vs. '-ui'" 電子情報通信学会技術報告 (信学技報) IEICE Technical Report TL2011-49, vol. 111, No. 320, pp. 61-66, 社団法人電子情報通信学会
- [5] Chung, Yujin, Namwon Heo, Jong-Hyeok Lee, Tai-Suk Kim, (1999) "Restoration of Japanese particle '-no' in Korean-to-Japanese MT system," Proceedings of the 11th Human & Cognitive Language Technology, pp. 115-120, Cheonbuk Univ, Korea.
- [6] 原田康也・首藤佐智子, (2009) "「の」の意味論と語用論再考: 容認度に反映される文脈への貢献度" 第26回日本認知科学会大会予稿集 pp. 218-219, 日本認知科学会
- [7] Nishiguchi Sumiyo, (2009) Possessive Disambiguation, PhD Dissertation, Osaka University.

i 本研究は科学研究費補助金挑戦的萌芽研究: 課題番号 21652041 『「場の言語学」の構築: 場の意味論と語用論』, ならびに科学研究費補助金基盤研究(C) 23520475 『前提研究の新アプローチ: 前提条件操作の限界事例からの検証』に基づく研究の一部である。またデータとして扱った Sejong 韓日パラレルコーパスを使用するにあたり, Sanghoun Song 氏の協力を得た。

ii 本稿は, Choe, Shudo & Harada (2011)で扱った論点を別の角度から再考したものである。

iii ここでいう逆相対頻度は, 対象となる形態素が一回出現するために平均的にいくつの形態素が必要を示す数字であり, テキストの形態素総数を対象となる形態素の出現度数で割ったものである。テキスト中の出現の割合としての相対頻度の逆数で逆相対頻度が 23.2:1 の場合は相対頻度が 4.3% と考えることができる。