前接名詞に基づく助詞の類似度から見た
日本語母語話者と中国人・韓国人学習者の文体差

鯨井綾希（東北大学）

1. はじめに

文章や文体といった言語学的に大きな単位は、それよりも小さい単位である語や文に比べて、日本語学上の分析方法が確立されていない。日本語教育においてもまた、文よりも大きな文章を視座に据えた書き方の指導については、言語学的な分析結果を反映させた方法が十分には整っていないと言える。したがって、まずは日本語学・日本語教育のどちらにおいても、日本語の文章を対象とした研究方法を様々に模索していく必要がある。加えて日本語教育においては、母語話者の文章と学習者の文章の分析を通して、母語話者の文章のタイプと、学習者の文章のタイプの異同を明らかにしていく必要があると思われる。

以上の問題意識の下、本発表では、母語話者・学習者（特に上位レベルの学習者）に見られる名詞群を用いた助詞の類似度の計量を通して、母語話者と学習者の文章の書き方（文体）の違いがどのような状況で生じているのかを明らかにすることを目的とした考察を行う。

2. 分析資料

本発表では、金澤（編）（2014）の付属 CD-ROM に収録されている『YNU 書き言葉コーパス』を資料として用いた。『YNU 書き言葉コーパス』は、「日本人大学生（30 名）と同大学に所属する留学生（韓国語母語話者 30 名、中国語母語話者 30 名）を対象に、12 の課題による書き言葉の資料、計 1080 編（母語別各グループ 360 編ずつ）を集めたものである」（金澤（編）2014：3）。

また、『YNU 書き言葉コーパス』では学習者の日本語能力を「上位群」「中位群」「下位群」の三つに分けている。本発表では母語話者と対照する学習者の資料として、文法や語法上の誤りが相対的に見られない「上位群」の 20 名を選択した。

『YNU 書き言葉コーパス』で母語話者・学習者に課せられたタスクは全部で 12 種類ある。タスクの一覧を表 1 に、母語話者・学習者（「上位群」）の資料の概略を表 2 に示す。

表 1：『YNU 書き言葉コーパス』で課されるタスク

タスク1	面識のない先生に図書を借りる
タスク2	友人に図書を借りる
タスク3	デジカメの販売台数に関するグラフを説明する
タスク4	学長に奨学金増額の必要性を訴える
タスク5	入院中の後輩に励ましの手紙を書く
タスク6	市民病院の閉鎖について投書する
タスク7	ゼミの先生に観光スポット・名物を紹介する
タスク8	先輩に起こった出来事を友人に教える
タスク9	広報誌で国の料理を紹介する
タスク10	先生に早期英語教育についての意見を述べる
タスク11	友人に早期英語教育についての意見を述べる
タスク12	小学生新聞で七夕の物語を紹介する

（金澤（編）（2014）：53）

表 2：母語話者・学習者の課題ごとの延べ語数と語彙多様度(Guiraud 値)

	母語話者(30名)		学習者(20名)	
	延べ語数	Guiraud値	延べ語数	Guiraud値
タスク1	2788	9.13	2194	10.12
タスク2	1416	8.77	1686	11.01
タスク3	2252	10.41	1646	9.05
タスク4	4225	14.42	3963	16.23
タスク5	8151	17.67	6021	17.54
タスク6	5356	15.62	4058	17.20
タスク7	4269	18.93	3828	17.88
タスク8	2025	8.96	1649	10.86
タスク9	5727	18.10	5405	18.83
タスク10	4042	15.02	3448	14.85
タスク11	3424	15.11	2807	14.42
タスク12	11701	12.63	8006	13.60
合計・平均	55376	13.73	44711	14.30

表 2 の「Guiraud 値」は、語彙の多様性を示す Type-Token Ratio（TTR）をもとに、文章の長さの影響をできるだけ小さくした補正値で、以下の式によって表される。

$$\text{Guiraud 値} = \frac{V}{\sqrt{N}}$$

V は異なり語数、N は延べ語数を表す。Guiraud 値が大きければより多くの語で文章を構成していると言え、小さければより少ない語で文章を構成していると言える。表 2 を見ると、「上位群」の学習者は、母語話者と同等の語彙多様性を示していることが分かる。したがって、分析に際しては、母語話者と学習者が持つそもそもの語彙量の差を考慮しない。

なお、本発表ではコーパスが持つ資料のうち、漢字の誤りや送り仮名を始めとした表記上の問題を適宜修正し、一行一文の形に加工した「補正データ」を利用した。

また、分析資料は、中・長単位解析器 Comainu を用いて長単位と呼ばれる長い単位の語に分割し、形態論情報を付与した上で、その結果を目視で確認して誤解析部分を修正したものを用いた。

3. 分析方法

本発表では、母語話者・学習者（上位群）の用いる助詞とそれに前接する名詞群との共起の傾向を用いて、母語話者・学習者のそれぞれで、それぞれが用いる助詞同士の類似度を計量する。そして、計量された値の違いの観察を通して、母語話者と学習者の文体差がどのような状況で生じるのかを明らかにする。ここで用いる類似度計算は、類似した意味を持つ語同士は類似した言語要素を伴いやすいという分布仮説に基づく（Harris1954）。すなわち、文章中で同じ名詞を前接名詞として多く選んでいる助詞同士ほど、互換性が高い（類似度が高い）と考える。以上を前提として、母語話者と学習者のそれぞれで、助詞に前接する名詞を収集し、ある助詞と別な助詞の前接名詞の類似度を余弦尺度と呼ばれる計量法により算出する。余弦尺度の計算式を以下に示す。

$$\text{余弦尺度} = \frac{\sum v_i w_i}{\sqrt{\sum {v_i}^2}\sqrt{\sum {w_i}^2}}$$

v_i、w_i の i は対象とする二つの助詞 v と w の前接名詞のうち、共通する名詞群の中の i 番目の名詞を指す。v_i、w_i の値は助詞 v、w と要素 i の自己相互情報量（PMI）とした。分子は v_i と w_i の積の総和、分母はそれぞれの二乗和の平方根の積である。助詞 v と助詞 w が共通する名詞を取るほど値が大きくなり類似度が高いと言え、共通する名詞が全くない場合は 0 となり類似してい

ないと言える。上記の計算により前接名詞の共通性から見たときの母語話者と学習者の助詞の用法が定量化され、そこから見た両者の文体差の考察が可能となる。

4. 分析結果

主要な助詞同士の類似度を余弦尺度に基づいて計量した結果を以下の表 3 に示す。注目される部分は太字にしてある。また、母語話者の値から学習者の値を引いた助詞の類似度の差を表 4 に示す。表 4 については、差が0.1以上と大きいものを太字で示し、0.05以上とやや大きいものを斜体で示している。

表 3：母語話者・学習者の使用する助詞同士の類似度

	は･が		は･を		は･に		が･を		が･に		を･に	
	母語話者	学習者	母語話者	学習者	母語話者	学習者	母語話者	学習者	母語話者	学習者	母語話者	学習者
タスク1	0.2173	0.2193	0.0433	0.0774	0.0000	0.0618	0.1400	0.1786	0.0219	0.0427	0.0621	0.0353
タスク2	**0.0000**	**0.1776**	**0.0000**	**0.0247**	**0.0000**	**0.0556**	0.0967	0.1267	0.0031	0.0000	0.0199	0.0163
タスク3	0.0965	0.1842	0.0546	0.0129	0.0912	0.0190	0.0944	0.1188	0.0419	0.0199	0.0424	0.0704
タスク4	0.1086	0.1546	0.0544	0.0275	0.0153	0.0508	0.1412	0.1336	0.0900	0.0744	0.0799	0.0382
タスク5	0.1378	0.1237	0.0957	0.1512	0.1016	0.0833	0.1450	0.1640	0.1077	0.0877	0.1162	0.0823
タスク6	0.1890	0.1444	0.0548	0.0728	0.0758	0.0000	0.1335	0.0890	0.1198	0.0705	0.1102	0.0704
タスク7	0.0446	0.1372	0.1042	0.1110	0.0200	0.1222	0.0784	0.0962	0.0268	0.0631	0.0506	0.0597
タスク8	0.1869	0.1066	**0.0000**	**0.1204**	0.1347	0.0522	**0.2609**	**0.1362**	0.0741	0.0723	0.0177	0.1588
タスク9	0.1454	0.1013	0.1635	0.1618	0.0687	0.0959	0.1616	0.1087	0.0761	0.0311	0.0851	0.0946
タスク10	0.0896	0.0880	0.0630	0.0965	0.0491	0.1068	0.1332	0.1039	0.0945	0.1014	0.0778	0.0630
タスク11	0.1247	0.0874	0.0935	0.1041	0.0566	0.0745	0.1098	0.1072	0.0662	0.0756	0.0566	0.1498
タスク12	0.3050	0.2778	0.1814	0.0808	0.0531	0.1280	0.1926	0.1991	0.0840	0.0977	0.1036	0.0689

表 4：母語話者・学習者の類似度の差（【母語話者の値】−【学習者の値】）

	は･が	は･を	は･に	が･を	が･に	を･に
タスク1	−0.0020	−0.0341	*−0.0618*	−0.0386	−0.0208	0.0268
タスク2	**-0.1776**	−0.0247	*−0.0556*	−0.0300	0.0031	0.0036
タスク3	*−0.0877*	0.0417	*0.0722*	−0.0244	0.0220	−0.0280
タスク4	−0.0460	0.0269	−0.0355	0.0076	0.0156	0.0417
タスク5	0.0141	*−0.0555*	0.0183	−0.0190	0.0200	0.0339
タスク6	0.0446	−0.0180	*0.0758*	0.0445	0.0493	0.0398
タスク7	*−0.0926*	−0.0068	**-0.1022**	−0.0178	−0.0363	−0.0091
タスク8	*0.0803*	**-0.1204**	*0.0825*	**0.1247**	0.0018	**-0.1411**
タスク9	0.0441	0.0017	−0.0272	*0.0529*	0.0450	−0.0095
タスク10	0.0016	−0.0335	*−0.0577*	0.0293	−0.0069	0.0148
タスク11	0.0373	−0.0106	−0.0179	0.0026	−0.0094	*−0.0932*
タスク12	0.0272	**0.1006**	*−0.0749*	−0.0065	−0.0137	0.0347

上掲の二つの表を観察すると、特徴的なタスクが何であるのかが浮かび上がってくる。まず表 3 を見ると、タスク 2「友人に図書を借りる」という状況において、母語話者と学習者の助詞の用法に大きな差が見られる。具体的には、学習者はタスク 2 において、格助詞「が」「を」「に」を取る名詞をある程度「は」によって取り立てている（類似度が 0 ではない）が、母語話者はそれらの格助詞を伴って用いた名詞を「は」によって取り立てていない（類似度は 0）。

また、表 4 を見ると、タスク 8「先輩に起こった出来事を友人に教える」という状況において、母語話者と学習者のそれぞれの助詞の類似度に差が見られやすいことが分かる。具体的には、「は」と「が」「に」および「が」と「を」の類似度が母語話者において高く、「は」と「を」および「を」と「に」の類似度が学習者において高いという傾向が見られる。表 4 と表 3 を合わせて考えると、タスク 8 では、母語話者は「を」に前接する名詞を「は」によって取り立てていないが、「が」

と「を」については学習者の二倍程度、同じ名詞を前接要素として選んでいることが分かる。

以上の差が出た二つのタスクには、次のような特徴がある。まず、資料の概略を示した表 2 を改めて観察すると、各タスクの延べ語数から考えて、タスク 2 は 12 タスク中でひとつひとつの文章が最も短く、タスク 8 もそれに次いで個々の文章が短いことが窺える。すなわち、タスク 2 もタスク 8 も 12 タスク中で文章の長さが短いという点で共通している。このことから、短い文章で用件を述べる場合の文体が、母語話者と学習者で異なっていることを指摘できる。

また、表 1 のタスクの内容について見てみると、タスク 2 とタスク 8 のいずれも、「友人」という親疎関係において「親」であり、かつ上下関係において「対等」であるという、身近な相手に向けて書いた文章である点で共通している。同様の文脈であるタスク 11 においてはそれほどの差が見られないため安易な一般化は避けるべきだが、目上の人物や疎遠な人物に向けて書く「改まった」文章に比べ、普段から親しく接している人物に向けて書く「くだけた」文章の方が、学習者にとって母語話者と同様の文体を作るのが難しいという可能性を指摘できる。

加えて、表 4 において母語話者と学習者の差が大きい、太字および斜体部分を観察すると、「は」と「に」（12 タスク中 8 タスクが太字・斜体）、「は」と「が」（12 タスク中 4 タスクが太字・斜体）の二つで、母語話者と学習者の助詞の前接名詞の類似度に差が見られる傾向にある。これがどのような内容上あるいは情報構造上の特徴に基づいて生じているのかという点について、言語的な具体例に基づいたさらなる分析が必要だが、この点は今後の課題としたい。

5. おわりに

本発表では、日本語母語話者と中国人・韓国人日本語学習者が同一タスクの下で作成した文章のコーパスである『YNU 書き言葉コーパス』を用いて、母語話者と学習者の文章の書き方（文体）の違いが生じる状況を明らかにすることを目的として考察を行った。目的の達成に向けて、本発表では母語話者・学習者が使用する助詞に注目し、助詞の前接名詞の共通性から見た助詞同士の類似度を余弦尺度により計量して、その差異を文体差として分析した。分析結果を以下に示す。

（1）長い文章よりも、短い文章を書く場合において母語話者と学習者に文体差が生じる。

（2）「友人」という親疎関係において「親」であり、かつ上下関係において「対等」であるという、身近な相手に向けて書く場合において母語話者と学習者に文体差が生じる。

さらに、必ずしも文体に関する状況に直結はしないものの、以下の点も指摘できた。

（3）「は」と「に」では 12 タスク中 8 タスクで、「は」と「が」では 12 タスク中 4 タスクで、比較的大きな助詞の類似度の差が母語話者と学習者の間に見られる。

今後は、以上の結果を各タスク内の文章の内容・情報構造に関わる諸々の言語要素にさらに結びつけていき、より包括的な文体差を明らかにしていく必要がある。

【参考文献】

金澤裕之（編）（2014）『日本語教育のためのタスク別書き言葉コーパス』ひつじ書房

Harris,Zellig.（1954）DISTRIBUTIONAL STRUCTURE, *Word*, Vol. 10, pp.146-162

【利用した資料】

『YNU 書き言葉コーパス』（金澤（編）（2014）に付属の CD-ROM 所収）

付記

本発表は平成 27 年度「公益信託田島毓堂語彙研究基金」の助成による研究成果の一部である。