異なる原言語からの翻訳による同義表現の分析
– 韓国語の例 –

白 京姫 † 大竹 清敬 † 山本 和英 †‡
† ATR 音声言語コミュニケーション研究所 ‡ 長岡技術科学大学 電気系
{kyonghee.paik, kiyonori.ohtake}@atr.jp yamamoto@fw.ipsj.or.jp

要旨

同一内容の翻訳で、原言語が異なる 2 つの韓国語旅行会話翻訳コーパスを用いて、原言語の影響が翻訳にどのように表れるのかを調べた。用いられた原言語は日本語と英語である。

1 背景

同一内容を扱っているが原言語が異なる 2 つの韓国語コーパスを利用し、翻訳における原言語の影響を考察する。各コーパスは、 162,308 文から構成される。この 2 つの韓国語コーパスは、日本語と英語の対訳コーパスのそれぞれの言語から翻訳されたものである。したがって、 2 つの韓国語コーパスをつきあわせ、それぞれの文が違う表現である場合、それらは同義表現になる。しかしながら、このようにして得られた韓国語翻訳コーパスは原言語が異なることから、それぞれいくつかの特徴があり、両者は大きく異なる。本稿では、その違いをいくつかの言語現象の観点から分析する。

周知のように英語は比較的固定された語順（SVO）を持ち、主語、目的語などが省略されない。反面、日本語の場合、述部が文末にくるが、それ以外の要素は柔軟な語順を持ち、さらに文脈上明らかな主語、目的語などは省略される。このように構文構造において日本語と英語は大きく異なる言語であると言える。また、語彙論的な観点からも、それぞれの単語によって与えられる意味や、その概念なども相当異なる。これらの点では、韓国語は英語より日本語に非常に近い言語である。

(1) 그가 무례해서 화가 났다.
ku-ka mwuryeyhayse hwa-ka nassta
he-nom rude-because anger-nom rise-past
“His rudeness annoyed/bothered/upset me.
(lit: I am angry because he is rude)”

たとえば、例 (1) に示した韓国語と英語の 2 つの文 [2] は、同じ内容を表しているが、韓国語は複文構造を、英語は単文構造をとっている。これは、英語と日本語の間の翻訳についても言えることであるが、それぞれの言語において自然な表現を相互に翻訳する場合、その構文構造を大きく変更しなければならない場合がある。

したがって、原言語が日本語か英語かによってその翻訳である韓国語文がそれぞれの原言語に大きく影響されると予想する。構文構造を大きく変更して翻訳することは、人間にとっても機械にとっても負担がかかる。以下に示す日本語と英語から韓国語への翻訳は、原言語の違いが翻訳に与える影響をよく示している。

(2) このケーブルカーに乗れば、ホテルに行くことができます。
이 케이블카를 타면 호텔에 갈 수 있습니다.

(3) This cable car will take you to the hotel.
케이블카가 호텔에 데려다 줄 겁니다. [1]

この例から、より自然な文へ翻訳するために大きな構文的変化を要求される場合、そのような構文的変更が行われず、原言語に大きく影響された翻訳が数多く存在していると予想する。

2 原言語が異なるコーパスの比較

本稿では、旅行会話に必要な様々な話題を含んだ ATR 旅行会話基本表現集（BTEC）を用いる。現在、BTEC は日本語、英語、韓国語、中国語などへ多言語化されている。 BTEC は当初、日本語と英語の対訳コーパスの収集から開始されたが、韓国語や、中国語訳を、日本語あるいは、英語を原言語として翻訳することにより拡充してきた。また、本研究で用いる 2 つのコーパスをそれぞれ K_J（日本語から韓国語へ翻訳されたコーパス）と K_E (英語から韓国語へ翻訳されたコーパス) と表記する。 BTEC は全ての言語対において文単位の対応がとれたコーパスである。

本稿では、 K_J と K_E の 2 つのコーパスを比較し、それぞれの性質を詳しく調べる。まずは、 2 つのコーパ

[1] cable car-nom hotel-to take give pred

表 1: 類似度によるコーパスの比較結果

類似度	0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0	Total
のべ	100	1,910	11,006	23,126	33,755	34,888	28,083	17,400	7,693	4,347	162,308
異なり	58	1,243	7,876	19,351	29,053	30,149	24,382	14,946	6,434	3,037	136,529

スの間の類似度の分析から始め、次の節ではいくつかの言語現象についてより詳しく分析する。

2.1 類似度を用いたコーパス比較

まず、2 つのコーパスがどの程度の類似しているのかを表層的な近さを計ることによって求めた。表層的な近さを計るために、編集距離に基づく類似度を用いた。具体的には、2 つのコーパスを utf8 でエンコードし、類似度を求めるためのプログラムを Perl によって作成した。編集距離に基づく類似度を求めるために `String::Similarity` というモジュール [3] を使用した。このモジュールは、Meyers による方法 [4] によって、2 つの文字列間に 0 から 1 までの類似度を与える。全く異なる文字列の場合は類似度 0、同一の文字列には類似度 1 が与えられる。

K_J と K_E のすべての翻訳対毎に類似度を求めた。そして、類似度 0.0 から 1.0 までを 0.1 刻みで 10 のクラスに分類し、各々の類似度に属する文の性質を分析する。使用したコーパスは、それぞれ 162,308 文から構成されるが、それぞれ重複した文を含む。そのため最終的に比較したのは、異なる組みあわせの 136,529 文である。この類似度による分類結果を表 1に示す。

表 1から分かることは、両コーパス全体の中、全く異なる表現と判断された文が 0.1% 以下で、ほぼ同一である文が 3% 以下 (4347/162308) である。使用したコーパスにおいて正書法が一貫してないところがあり、同一と判断されない文が多数存在する。その点を考慮しても、同一と判断できる文は約 8% である。文献 [1] では、BLEU に代表される統計的な翻訳評価指標に関する実験を通して、一つのテキストに対して実に様々な訳が可能であることを議論している。この観点から、我々が用いたコーパスが、原言語が異なるとはいえ約 8% しか同一の文を含んでいないということがそれほど不思議ではない。しかし、彼らが用いたテキストは、聖書と文学作品であり、原言語の影響に関して直接議論しているわけではない。

3 諸言語現象における違い

本稿で扱う言語現象は、敬語表現、訳語選択、ゼロ代名詞である。これらの言語現象が、どの程度コーパスに出現しているかを調べるために表 1に示した類似度分布のうち、0.1 から 1.0 までの範囲からそれぞれ 1% にあたる文 (ペア) を無作為抽出し、それぞれの文がどういった現象を含んでいるかを計数した。集計結果を表 2に示す。

3.1 敬語表現

敬語表現は、表 2から、K_J の方が類似度に関係なく多く使われていることがわかる。韓国語では、日本語と同様に敬語がよく使われる。日本語よりも多段階の敬語レベルがあり、話者、聴者、指示対象との間に社会的地位、年齢、グループ、新密度などを考慮し、使い分ける。詳細は [7] を参照されたい。以下、敬語表現における代表的な K_E と K_J に含まれる例を示す。

(4) 호텔 내에 약국이 **있나요?**: K_E

"Do you have a drugstore in this hotel? "

(5) 이 호텔에 약국은 **있습니까?** : K_J

"このホテルにドラックストアはありますか" (類似度 0.6-0.7 から)

この場合、日本語からの訳がより丁寧である。反面、英語の場合は丁寧さが明らかではなく、どちらにも訳せる。語尾の -요 は 니까 に比べると丁寧度が下がる。しかし、この区別は日本語にはない。実際の会話においては上記の韓国語の訳 (4) は (5) より丁寧さが低いが、日常会話ではよく使われる表現であり、(5) の方は丁寧であるが、かしこまった場面でより相応しいと言える。

3.2 訳語選択

韓国語と日本語は文法の面でも類似しているが、語彙的な面においても非常に近いと言える。ここでは、漢語と外来語に関して分析する。

3.2.1 漢語

韓国語における漢語の七割が日本語にも存在すると言われている [6]。実際に、表 2でも、K_J のほうで漢語がよく使われている。以下に例を 1 つあげる。

(6) **리넨 제품 코너**는 어디에요? [2] : K_E

"Where is the linens section? "

[2]直訳： リネン製品コーナはどこですか。

表 2: K_J と K_E における言語現象とその頻度

類似度	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
無作為抽出数	12	78	192	290	300	243	149	64	30
敬語 (K_J)	1	10	16	82	101	76	32	11	2
敬語 (K_E)	0	8	15	20	32	31	9	2	0
ゼロ代名詞 (K_J)	0	3	8	23	28	22	13	4	1
ゼロ代名詞 (K_E)	0	2	5	7	15	8	9	1	1
漢語 (K_J)	0	7	23	39	54	22	19	4	0
漢語 (K_E)	0	2	10	25	35	19	12	0	0
外来語 (K_J)	0	5	16	14	15	7	6	0	0
外来語 (K_E)	0	0	7	11	9	11	3	0	0

表 3: 同義表現の種別とその頻度

		類似度								
類型	現象	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
文全体	同一	0	0	0	0	0	0	3	0	18
	意訳	6	26	36	13	8	4	4	0	0
	誤訳	3	8	6	7	1	2	2	6	0
訳語選択	名詞	0	30	98	110	98	70	23	15	1
	動詞	0	32	112	193	186	88	37	11	2
	疑問詞	0	2	11	9	10	2	4	1	0
	その他	1	17	68	115	89	40	16	1	1
統語	助数詞	0	2	5	11	6	2	0	0	0
	その他	1	20	71	109	156	96	34	15	5
正書法	表記のゆれ	1	0	1	7	7	0	0	0	0
	数字	0	6	14	18	20	25	10	0	0

(7) **침구 매장**은 어디입니까？ :K_J

“寝具売り場はどこですか。” (類似度 0.3-0.4 から)

3.2.2 外来語

外来語に関しては K_J、 K_E 両コーパスに大きな差はなかった。しかし、 K_E で使われる外来語は “track, main dining room, avenue, check, coupon, dark brown, rent, seat, golf round” といった外来語としては定着度が低い語である。反面、 K_J で使われる外来語は “size, center, tour, ticket, economy car, room music, service, counter, play” のように和製英語及び定着度が高い語が頻繁に表れる。これは翻訳の際に原言語でどのような単語が使われているかによって、翻訳される語彙にも影響を及ぼすと推察できる。 3.2.1の例のような場合でも “リネン製品” より “寝具” の方がより広い意味合いを含んでいる。このような意味のずれが原言語の影響によって明確に表れる。

3.3 統語論的差

表 2から K_J の方がゼロ代名詞を頻繁に使っていることが分かる。 1節で述べたように日本語では、文脈上明白な成分は省略される。次の例は目的語を省略した例である。

(8) ϕ_{subj} 분명히 이 비행기를리컨펌했는데요.[3] : K_E

“I’m sure I reconfirmed this flight. ”

(9) ϕ_{subj} 확실히 ϕ_{obj} 재확인 했습니다. : K_J

“きちんと ϕ_{subj} ϕ_{obj} リコンファームをしました。” (類似度 0.3-0.4 から)

韓国語でも日本語と同様な省略が行われる。上記の例は、原言語が日本語の方がより省略が起こりやすいことを示している。

[3]直訳： ϕ_{subj} 確かに この 飛行機をリコンファームしました。

3.4 同義表現の分類

前節まで敬語表現、ゼロ代名詞、漢語、外来語が両コーパスでどのように使われているのか分析した。しかしながら、2つの韓国語コーパスをつきあわせた結果の約92%の文に不一致表現が含まれる。しかも、それらの文は同義表現であることから、そこにどのような差異が存在するかを調べることによってどのような同義表現を抽出できるかがわかるようになる。そこで、我々が用いたコーパスをつきあわせた結果に対して、そこにどのような違いがあるのかについて調べた。調査の対象とした文の組み合わせは、表2で用いたものと同一である。

表3に調査の結果を示す。結果から、全体的に、名詞または動詞の訳語の違いが非常に多いことがわかる。一方で、誤訳の多さも目立つ。これは、BTECの翻訳が、文脈をほとんど与えられず、一文単位で翻訳されていることに一因がある。さらに、コーパスをまとめる際の誤りによって、誤訳と判断せざるを得なかった文も多く存在する。

また、結果として示していないが、助数詞はK_Jに比べ、K_Eの方でより多くの助数詞が使われていることが分かった。詳しい分析は別稿にゆずる。

類似度がかなり低い(0.0-0.3)文を分析すると、文単位の異表記同義表現が数多く得られると予想する。以下に一例をあげる。

(10) 저는 야행성이에요.[4] : K_E
"I'm a night owl. "

(11) 나는 밤 늦게까지 잠을 안 잡니다. : K_J
"私は夜ふかしです。" (類似度 0.3 から)

このような同義表現の獲得は、換言の研究にも応用できる。単純な換言規則を適用するだけでは、このような同義表現へ相互に換言することは困難である[5]。

3.5 その他

K_JとK_Eは別々に翻訳されたものであり、固有名詞の書き方と数字の書き方等に一貫性がない。たとえば、英語の"hostess"を含む文の翻訳をみると、K_Jでは"호스테스(ホステス)"、K_Eでは"호스티스(ホスティス)"のように表記されている。これは日本語でいう、カタカナの表記の揺れに該当する。また、数字表現では、"6시(時)40분(分)"と"여섯시 사십분"のようにコーパスによって異なる表記を使用している。したがって、このコーパスを用いて、同義表現獲得、あるいは換言規則の抽出などを行う際には、表記の統一を考慮する必要がある。なお、このような異表記はコーパス全体の7%の文に存在する。

[4]直訳: 私は夜行性です。

4 結論

本稿では、同一内容を日本語と英語から翻訳した2種類の韓国語旅行会話コーパスを用いて、原言語が翻訳にどのような影響を及ぼすのかについていくつかの言語現象に着目し、分析した。要約すると、文法及び語彙面において非常に類似している日本語ならびにそれらが相当異なる英語それぞれからの翻訳では、原言語の違いが翻訳に多大な影響を与えている事実を示すことができた。これは人間の翻訳者においても機械翻訳においても同じことだと考えられる。今後はこのような言語差を利用した同義表現の抽出について詳しく検討する予定である。

謝辞

本研究は総務省の研究委託により実施したものである。

参考文献

[1] Christopher Culy and Susanne Z. Riehemann. The limits of N-gram translation evaluation metrics. In *MT Summit IX*, New Orleans, 2003.

[2] Young-Ok Lee. The difference in subject choice between Korean and English. In *English Education in the Era of Information*, 1999. Chungnam National University, Kwangju, Korea.

[3] Marc Lehmann. String::Similarity. Perl Module (`cpan.org`), 2000. (v0.02).

[4] Eugene Meyers. An O(ND) difference algorithm and its variations. *Algorithmica*, 1(2):251–266, 1986.

[5] Kiyonori Ohtake and Kazuhide Yamamoto. Paraphrasing honorifics. In *Workshop Proceedings of Automatic Paraphrasing: Theories and Applications (NLPRS2001 Post-Conference Workshop)*, pages 13–20, Tokyo, 2001.

[6] Kyonghee Paik, Francis Bond, and Satoshi Shirai. Using multiple pivots to align Korean and Japanese lexical resources. In *Proceedings of the Workshop: Language Resources in Asia (NLPRS2001 Post-Conference Workshop)*, pages 63–70, Tokyo, Japan, 2001.

[7] Ho-Min Sohn. *The Korean Language*. Cambridge Language Surveys. Cambridge University Press, 1999.