

言語の推測と類似性：印象が似ている言語・似ていない言語

山口虎太郎

はじめに

言語には互いに似ているとされるものがある。一例として、スペイン語とポルトガル語は、相互理解性があるほど似ている。言語の類似性を見るためには系統を参考にするなどの方法があるが、スペイン語とポルトガル語の系統ははっきりとしない(黒沢, 1998)など常に使える方法というわけではない。

ここで、スペイン語とポルトガル語が近いのであれば、スペイン語の文章を読んでポルトガル語と誤認する人が多いのではないかということが考えられる。このアプローチによれば、アンケートのみで言語同士の印象的な距離を測れる可能性がある。

本研究では、ある未知の言語で書かれた文章に対し何語で書かれていると感じるかを堪えてもらうことで、言語の距離を測ることができるかを検証する。

方法

アンケート概要　日本語を第一言語とする23人の回答者に対し、16の自然言語と3の人工言語について、各言語で書かれた短文を提示した。英語と日本語については解答言語として選ばなかったもののみが提示される。

回答方法　回答者には与えられた文について、どの程度親しみがあるかを1(文意を理解できる)から10(何語か分からない)までで評価してもらい(以下、不案内度)、また何語で書かれているか推測してもらう。

文と言語　全ての自然言語の文は東京外国語大学語学研究所編纂の世界の言語ガイドブック1・2より引用、エスペラントの文はMarina Yaguello(1984, 谷川多佳子・江口修訳, 1990)による。他二つの人工言語は著者自身による。人工言語のうち一つ”言語a”は著者によって作られたものであり文法や単語などの一切を公開していない。

結果1

不案内度が最大となったのは言語aであった。不案内度と解答の種類数の相関係数は0.670、不案内度と最頻の回答の回答数の相関係数は-0.768であった。

結論：読み手が推測する言語には文字が

より強い影響を持ち、系統の影響は限定

的である。

表1 各言語に対する回答者の不案内度の平均値と推測された言語として最も多かったもの、誤答が最頻となったものに色付けした			
言語	文中で使われる文字	平均不案内度	最頻の回答(回答数)
日本語	ひらがな、漢字	1.100	日本語(3)
英語	ラテン文字	1.000	英語(20)
スペイン語	ラテン文字	5.652	スペイン語(12)
イタリア語	ラテン文字	7.348	イタリア語(9)
オランダ語	ラテン文字	7.435	ドイツ語(8)
ラテン語	ラテン文字	7.409	ラテン語(5)
ギリシャ語	ギリシャ文字	6.652	ギリシャ語(13)
スウェーデン語	ラテン文字	7.217	ドイツ語(5)
ポルトガル語	ラテン文字	6.435	ポルトガル語(9)
セルビア語	キリル文字	6.318	ロシア語(15)
タイ語	タイ文字	7.273	タイ語(12)
タガログ語	ラテン文字	8.091	インドネシア語(2)
ヒンディー語	デーヴァナーガリー	6.682	ヒンディー語(14)
ペルシア語	アラビア文字	6.652	アラビア語(18)
マレー語	ラテン文字	7.913	インドネシア語(4)
ハンガリー語	ラテン文字	8.636	チェコ語(2) ポルトガル語(2)
トキポナ	ラテン文字	7.304	トキポナ(4)
エスペラント	ラテン文字	8.043	スペイン語(3)
言語a	ラテン文字	8.696	カザフ語(2) 中国語(2)

表2 ペルシア語の文に対し回答者23人が推測した言語		
言語	回答数	使用文字
アラビア語	18	アラビア
ペルシア語	1	アラビア
タミル語	1	タミル
アフリカーンス語	1	ラテン
ヘブライ語	1	ヘブライ
その他・無回答	1	

表3 キリル文字表記のセルビア語の文に対し回答者23人が推測した言語		
言語	回答数	使用文字
ロシア語	15	キリル
ウクライナ語	4	キリル
ブルガリア語	1	キリル
ベラルーシ語	1	キリル ラテン
ポーランド語	1	ラテン
その他・無回答	1	

表4 ギリシャ語の文に対し回答者23人が推測した言語		
言語	回答数	使用文字
ギリシャ語	13	ギリシャ
ロシア語	6	キリル
ウクライナ語	1	キリル
キプロス語	1	-
その他・無回答	2	

表5 14の自然言語の文に使われた文字と、回答者23人が推測した言語で使われる文字の一致率	
ラテン文字	0.647
ギリシャ文字	0.609
キリル文字	0.913
タイ文字	0.522
デーヴァナーガリー	0.696
アラビア文字	0.826

結果2

ラテン文字で書かれていない自然言語の平均不案内度は6.715で最大不案内度はタイ語の7.273であるのに対し、ラテン文字で書かれた自然言語の平均不案内度は7.348で最大不案内度はタガログ語の8.091であった。

考察

ラテン文字圏の言語においては、不案内度が全般に高かった。これはラテン文字表記される言語で知っているものが多く、文字によって何語であるか想像するのがかえて難しくなったためであると考えられる。

実際の言語と最頻の回答が異なる言語に注目すると、オランダ語とドイツ語は系統的に近いということが知られており(佐藤, 1998)、オランダ語の文をドイツ語だと考えた人が多いという今回の結果はこれに沿っていると言える。またマレー語はインドネシア語の前身にあたり(佐々木, 1998)、これも今回のアンケート結果に合致する。しかしエスペラントや言語aなどの人工言語は回答が多数の言語に分かれてしまい、印象が近いと言える言語は得られなかった。

表2,3の通り、ペルシア語の文にアラビア語と、セルビア語の文にロシア語と回答した人が多いことから、文字が回答に強い影響力を持っていることが推測できる。セルビア語と近いブルガリア語と答えた人は1人にとどまり、他は東スラブ・西スラブの言語である。

これらよりこのアンケートによって言語の距離を測る手法は、日本における言語の認知度と文字に強く影響を受けてしまうと考えられ、表5に示された文に使われた文字と回答した言語で使われる文字の一致率はこの説を支持していると言える。しかし表4に示されたように、これに反してギリシャ語の文にキリル文字圏の言語を回答した人が約30%いることは注目に値する。

おわりに

文の視覚的印象から推測される言語の種類というのは、実際の系統以外に影響を受ける変数が多く、回答結果を言語同士の距離として定量的な取り扱いをするのは難しいと考えられる。

ヨーロッパの言語など範囲を絞って同様のアンケートを取ることで、有用な結果を得られる可能性があり、この点は今後の課題である。

文献

黒沢直俊, ポルトガル語, 東京外国語大学語学研究所編, 世界の言語ガイドブック1, 三省堂, 1998.
佐藤弘幸, オランダ語, 東京外国語大学語学研究所編, 世界の言語ガイドブック1, 三省堂, 1998.
佐々木重次, インドネシア語, 東京外国語大学語学研究所編, 世界の言語ガイドブック2, 三省堂, 1998.