

印欧語話者の英文に内在する言語系統樹

永田 亮† Edward Whittaker††

† 甲南大学知能情報学部 †† Inferret Limited

E-mail: †rnagata@konan-u.ac.jp., ††ed@inferret.co.uk

1. はじめに

本稿では、非母語話者の文章には、母語に対応した言語系統樹が内在することを示す。具体的には、印欧語話者の英文を母語干渉の観点から自動分類し、その結果得られる言語系統樹が印欧語の言語系統樹に極めて類似することを示す。対象とする英文は、ICLE コーパス [5] に収録されている 11 種類の印欧語話者の英文である。更に、本稿では、二つの言語系統樹が類似することに、少なくとも 3 種類の母語干渉が寄与していることを明らかにする。

結論から述べると、二つの言語系統樹は図 1 に示すようになる。上の図は、言語学で知られる印欧語の言語系統樹である（文献 [3] を参考にして 11 言語のみを表記した）。この図より、11 の言語は大きく三つのグループ（イタリック語派、ゲルマン語派、スラブ語派）に分けられることがわかる。下の図は、11 種類の英文から自動生成された言語系統樹である。詳細は 3. で述べるが、両言語系統樹が極めて類似していることは一目瞭然である。なかでも、自動生成された言語系統樹が、11 種類の英文を正しく三つの語派に分類することは特筆すべきである。

従来研究でも、非母語話者の文章の分析は盛んに行われているが、誤り分析と語彙、品詞、文法の使用に関する分析を中心としている（従来研究については、文献 [4] に詳しい）。例えば、Aarts ら [1] は、 χ^2 値を用いて、非母語話者の英文で過度に使用される品詞列の分析を行っている。また、Altenberg ら [2] は、母語話者の英文、フランス語英語、スウェーデン語英語における接続副詞の使用の差異を分析している（本稿では、フランス語英語とは、フランス語母語話者が書いた英語を意味することにする。スウェーデン語英語なども同様である）。しかしながら、我々が知る限り、非母語話者の文章に内在する言語系統樹を明らかにした研究は存在しない。

2. 言語系統樹の生成手法

言語系統樹を生成する手法の基本アイデアは、階層型クラスタリングを利用するというものである。幸い、Kita [7] が、様々な言語を対象としてコーパスから言語系統樹を生成する手法を提案している。Kita の手法では、綴りを通じて様々な言語をモデル化する。具体的には、文字ベースの確率的言語モデルを用いる。言語系統樹の生成は、言語モデルに階層型

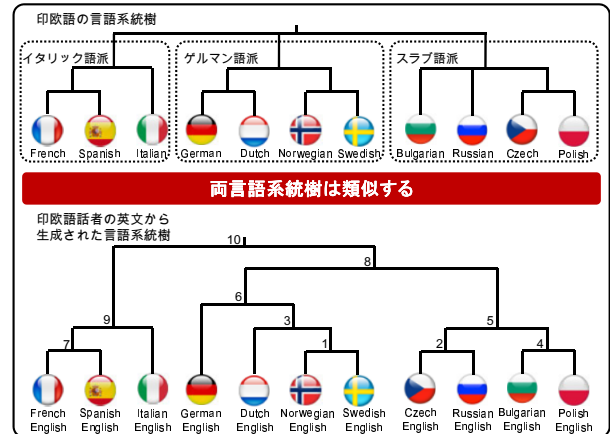


図 1: 上: 印欧語の言語系統樹 (一部). 下: 印欧語話者の英文から自動生成された言語系統樹

クラスタリングを適用することで行う。クラスタリングで使用する距離は、言語モデル間の距離として定義する。Kita の手法には、他の従来手法と異なり、言語学的な事前知識や基礎語彙表を必要としないという利点がある。

本論文の手法は、Kita の手法を基礎とするが、両手法には大きな違いがある。Kita の研究は（また、その他の従来研究も）、様々な言語を対象として言語系統樹を生成することを目的とする。一方、本研究では対象とするのは英語 1 種類のみである。より正確には、様々な非母語話者が書いた英文に内在する言語系統樹を生成することが目的である。この違いのため、従来手法をそのまま本研究に適用することはできない。例えば、本研究で対象とする文章は全て英語の綴り規則に従い書かれているため、綴り規則に基づいた Kita の手法は効果的でないことが予想される。

この問題を解決するため、本研究では、単語ベースの言語モデルを用いる。ただし、単語そのもので言語モデルを構築すると、母語干渉よりも文章の内容から受ける影響が大きくなる恐れがある。そこで、単語と品詞を混ぜた言語モデルを考えることにする。この言語モデルでは、名詞や動詞などの内容語は対応する品詞で置き換える。この処理により、以下に示すように、文章の内容から受ける影響を大幅に減らすことができる。具体的には、言語モデルの構築の前に次の処理を行う。まず、各英文データを文に分割する。次に、品詞解

析を行う。また、全ての単語を小文字に変換する。最後に、各単語を対応する品詞に置き換える。ただし、次の品詞に該当する単語に関しては、単語そのものを品詞として扱う：等位接続詞、決定詞、前置詞、前決定詞、助動詞、所有代名詞、代名詞、疑問副詞。また、固有名詞は、普通名詞として扱う。更に、文頭と文尾には特殊な品詞 B と E があるとする。例えば、この処理により、例文：

The alien wouldn't use my spaceship but the hers.

から、

B the NN would RB VB my NN but the hers . E

が得られる。品詞に置き換えた文は、元の文の内容をほとんど反映していないことがわかる。一方で、*the hers* の部分は、そのまま残されており、依然、母語干渉は反映している。

ここで、手法の定式化を行うため、次の記号を導入する。いま、 i 番目の言語の母語話者が書いた英文データを D_i と表す。ただし、 $0 \leq i \leq 10$ であり、本研究で対象とする 11 種類の印欧語に対応する。また、 D_i に対応する言語モデルを M_i と表す。言語モデル M_i は、Kita の手法にならないトライグラムモデルとする。言語モデル M_i の条件付確率は、上述の前処理を施した言語データ D_i から Kneser-Ney (KN) スムージングを用いて推定する。

このように M_i と D_i を定めると、Kita の手法が自然に適用できる。クラスタリングアルゴリズムは、群平均法を利用した階層併合的クラスタリングを用いる。クラスタリングに必要な距離は、言語モデル間の距離として定義する。言語モデル M_i が言語データ D_i を生成する確率は $\Pr(D_i|M_i)$ で表せる。ただし、トライグラムモデルを考えているので、 D_i 中の各トライグラムに対応する条件付確率を掛け合わせることで近似できる。このとき、言語モデル M_i から M_j への距離を、

$$d(M_i \rightarrow M_j) = \frac{1}{|D_j|} \log \frac{\Pr(D_j|M_j)}{\Pr(D_j|M_i)} \quad (1)$$

で定義する。ただし、 $|D_j|$ は、 D_j 中のトライグラムの数を表す。式 (1) は、元々 Juang ら [6] により提案され、Kita [7] により拡張されたものである。式 (1) の意味するところは、 M_i 、 M_j それぞれが言語データ D_j を生成する確率の比に基づいて M_i から M_j への距離を定義するということである。更に、 $d(M_i \rightarrow M_j)$ と $d(M_j \rightarrow M_i)$ が非対称であることを考慮して、 M_i と M_j の間の距離を両者の平均：

$$d(M_i, M_j) = \frac{d(M_i \rightarrow M_j) + d(M_j \rightarrow M_i)}{2} \quad (2)$$

で定義する。

以上をまとめると、言語系統樹を生成する手順は次のようになる：(i) 各言語データの前処理を行う；(ii) 言語データ

から言語モデルを構築する；(iii) 言語モデル間の距離を計算する；(iv) 計算した距離を用いて、言語データのクラスタリングを行う；(v) 得られた結果を言語系統樹として出力する。

更に、本研究では、もう一つ手法を利用する。この手法では、言語モデルの代わりに、ベクトルにより各国語英語をモデル化する。ベクトルの各要素は、各トライグラムに対応し、値は英文データ中のトライグラムの相対頻度とする。これにより、クラスタリングに用いる距離は、ベクトル間のユークリッド距離として自然に定義できる。言語系統樹の生成手順は、ベクトルに基づいたモデルとユークリッド距離を用いる以外は、言語モデルの場合と同様である。

3. 言語系統樹の生成実験

本実験では、ICLE コーパス [5] を対象英文データとした。同コーパスは、非母語話者（大学生）が書いた英文エッセイ（大部分が argumentative essay）を収録したコーパスである。言語データだけではなく、書き手の母語、家庭での使用言語、収集場所などの補助情報も収録されている。同コーパス内の 11 のサブコーパスが印欧語話者の英文に対応する。したがって、この 11 のサブコーパスを本実験で利用した。

実験条件をそろえるため、各サブコーパスに対して次のような前処理を行った。2 種類以上の言語を母語とする人の英文が存在するため、次の 3 つの条件を満たさない英文をコーパスから削除した：(1) 母語が 1 種類であること、(2) 家庭で使用する言語が 1 種類であること、(3) (1) と (2) の言語がサブコーパスの母語と同じであること。更に、テキスト ID などのタグを除去した。また、記号 ‘と’ は、’ に統一した。前処理後、サブコーパスのサイズは、平均 292 文書、228,193 トークンとなった。

これらのサブコーパスから、言語系統樹を生成した^(注 1)。頻度 5 以下の n グラムは、削除して言語モデルを構築した。

言語モデルに基づく手法で生成された言語系統樹は、既に、図 1 に示した通りである（図 1 の下の系統樹）。各ノードに付された数字は、二つのクラスが併合されたステップを示す。図より、自動生成された言語系統樹は、印欧語の言語系統に極めて類似することがわかる。1. で述べたように、自動生成された言語系統樹は、11 種類の英文を正しく三つの語派に分類する。印欧語の言語系統樹との違いとしては、ゲルマン語派内とスラブ語派内において若干の不一致が見られる。

図 2 に、ベクトルに基づいた手法から生成された言語系統樹を示す。この手法も言語モデルに基づいた手法と同じような振舞を見せていることがわかる。ただし、ポーランド語英語を独立したクラスとして扱っている点に差異が見られる。

(注 1)：言語モデルの構築には Kyoto Language Modeling toolkit: <http://www.phontron.com/kylm/> を使用した。品詞タグ付けには CRFTagger (CRF English POS Tagger, Xuan-Hieu Phan 2006, <http://crftagger.sourceforge.net>) を用いた。

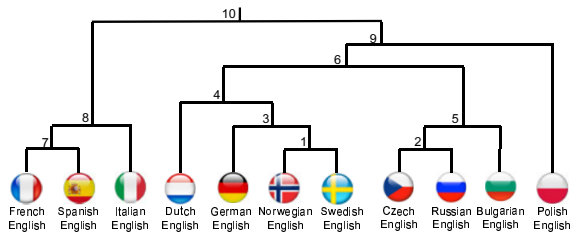


図 2: ベクトルに基づく手法から生成された言語系統樹

4. 考 察

本章では、自動生成された言語系統樹が印欧語の言語系統樹に類似する要因を考察する。効率的に分析を進めるために、言語系統樹の生成に寄与したトライグラムを自動抽出する。基本的な考え方は次のとおりである。この方法では、3種類の英文の比較により抽出を行う。各英文から、あるトライグラムを取り去ると、生成される言語系統樹の形が変わる可能性がある。仮に、言語系統樹に変化がなくとも、各モデル間の距離が逆転する方向に変化するかもしれない。そのようなトライグラムを特徴的であるとして抽出する。いま、抽出方法を定式化するために、トライグラムを t で表す。また、 t の D_i における相対頻度を r_{ti} で表すこととする。このとき、 t を除去したときの、 D_i , D_j , D_k に関する距離の変化は $s = (r_{tk} - r_{ti})^2 - (r_{tj} - r_{ti})^2$ で表される（ただし、ベクトルに基づく手法において）。 $(r_{tk} - r_{ti})^2$ は、 D_i と D_k に関する距離の減少に対応する。同様に、 $(r_{tj} - r_{ti})^2$ は、 D_i と D_j に関する距離の減少に対応する。したがって、 s が大きい順にソートすることで、特徴的なトライグラムのリストを得ることができる。

表 1 に、 D_i , D_j , D_k を、それぞれフランス語英語、スペイン語英語、ロシア語英語とした場合の特徴的なトライグラム上位 12 件を示す。ただし、読みやすさのため、 s と r を、それぞれ 10^6 倍と 10^2 倍している。表から、多くのトライグラムに冠詞が含まれていることがわかる。フランス語英語とスペイン語英語では、冠詞を含む各トライグラムの相対頻度は近い値を示している。一方、ロシア語英語では、相対的に低い。これは、フランス語とスペインには冠詞が存在するが、ロシア語には冠詞が存在しないという事実に対応する。同様な議論が、他のイタリア語英語とスラブ語英語についても成り立つ（例 *the JJ NN* の相対頻度：イタリア語英語 0.82, ポーランド語英語 0.72）。ただし、定冠詞のみ存在するブルガリア語では、定冠詞を含むトライグラムの相対頻度については、その限りではない（例 ブルガリア語英語の *the JJ NN* の相対頻度 0.82, *a JJ NN* の相対頻度 0.60）。

興味深いことに、冠詞を含むトライグラムの相対頻度の相違は、イタリア語英語とゲルマン語英語の間にも存在する

表 1: 特徴的なトライグラム (D_i : フランス語英語; D_j : スペイン語英語; D_k : ロシア語英語)

s	Trigram t	r_{ti}	r_{tj}	r_{tk}
5.14	the NN of	1.01	0.98	0.78
4.38	a JJ NN	0.85	0.77	0.62
2.74	the JJ NN	0.87	0.86	0.71
2.30	NN of the	0.49	0.52	0.33
1.64	...	0.22	0.12	0.05
1.56	NNS . E	0.77	0.70	0.92
1.31	NNS and NNS	0.09	0.13	0.21
1.25	B RB ,	0.25	0.22	0.14
1.22	of the NN	0.42	0.44	0.30
1.17	VBZ TO VB	0.26	0.22	0.14
1.09	B i VBP	0.07	0.05	0.17
1.03	NN of NN	0.74	0.70	0.63

(本研究が対象とするゲルマン語には冠詞が存在する)。例として、表 2 に、フランス語英語 (D_i)、スペイン語英語 (D_j)、スウェーデン語英語 (D_k) とした場合の特徴的なトライグラムを示す。表 1 の場合と同様に、多くのトライグラムに冠詞が含まれており、それらの相対頻度はイタリア語英語より低い。この点をより深く理解するために、各英文における冠詞の分布を図 3 に示す。横軸と縦軸は、それぞれ不定冠詞と定冠詞の相対頻度を表す。ゲルマン語英語は、イタリア語英語に比べて定冠詞の使用が少ない傾向にあることがわかる。これは、イタリア語ではより幅広い定冠詞の用法があることに起因していると推測できる（例 所有代名詞、固有名詞に対する定冠詞の使用）。全体としては、冠詞の分布により、11 種類の英語が三つの語派にほぼ正しく分類されることがわかる。以上の考察により、冠詞を含むトライグラムが言語系統樹の生成に寄与しているといえる。

of を含むトライグラム（例 *NN of NN* や *the NN of*）も特徴的である。フランス語英語とスペイン語英語では、ロシア語英語とスウェーデン語英語に比べて相対頻度が高い。この現象は、次のように説明される。イタリア語では、名詞を連結した名詞句 (*NN NN*) より、前置詞 *of* を利用した名詞句 (*NN of NN*) が好まれる [8]。例えば、英語の *cheese tart* は、イタリア語では *tart of (the) cheese* のように *of* による連結で表現される。一方、スウェーデン語やロシア語では、名詞の連結や類似した構造が許される。その結果、イタリア語英語では相対的に *NN of NN* や *the NN of* などの頻度が高くなる。実際、各英文における *NN of NN* の相対頻度を調べて見たところ、三つの語派で異なった分布域をとることが明らかとなった：イタリア語英語：高域 (0.70~0.74); スラブ語英語：中域 (0.56~0.68); ゲルマン語英語：低域 (0.43~0.57)。以上により、名詞句の構成法も言語系統樹の生成に寄与しているといえる。

表 2: 特徴的なトライグラム (D_i : フランス語英語; D_j : スペイン語英語; D_k : スウェーデン語英語)

s	Trigram t	r_{ti}	r_{tj}	r_{tk}
21.49	the NN of	1.01	0.98	0.54
5.70	NN of NN	0.74	0.70	0.50
3.26	NN of the	0.49	0.52	0.30
3.10	the JJ NN	0.87	0.86	0.70
2.62	. . .	0.22	0.12	0.03
1.53	of the NN	0.42	0.44	0.29
1.50	NN , NN	0.30	0.30	0.18
1.50	B i VBP	0.07	0.05	0.19
0.85	NNS and NNS	0.09	0.13	0.19
0.81	JJ NN of	0.40	0.39	0.31
0.68	. . E	0.13	0.06	0.02
0.63	a JJ NN	0.85	0.77	0.73
0.63	RB . E	0.21	0.16	0.31
0.56	NN , the	0.16	0.16	0.08
0.50	NN of a	0.17	0.09	0.06

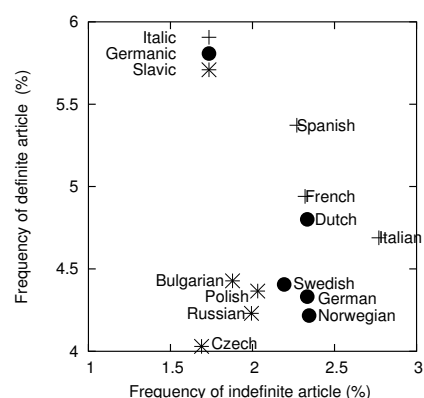


図 3: 非母語話者の英文における冠詞の分布

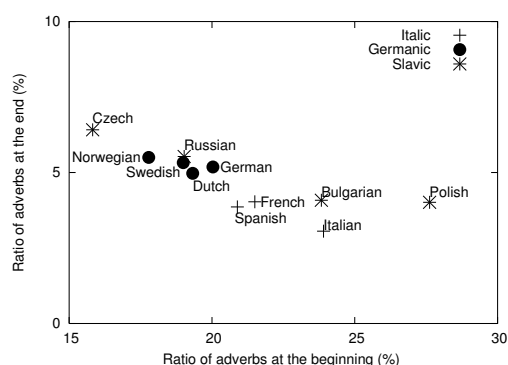


図 4: 非母語話者の英文における副詞の位置の分布

トライグラム B , RB (表 1 で 8 位) や $RB . E$ (表 2 で 13 位) は副詞の使用位置も 11 種類の英文の分類に寄与する

ことを示唆する (このトライグラムは、文頭と文末の副詞に近似的に対応する)。この点を詳細に分析するために、副詞のうち、文頭と文末で使用されている副詞の比率を求めた (図 4)。図より、冠詞の使用や名詞句の構成法ほどではないが、副詞の位置についても、イタリック語英語とゲルマン語英語は一定の傾向を示すことがわかる。一方、スラブ語英語にはばらつきが見られる。しかしながら、他の要因と組み合わせると、イタリック語英語とゲルマン語英語から弁別する手掛りとなる。例えば、冠詞の使用の分布 (図 3) では、ポーランド語英語は、ブルガリア語英語とスウェーデン語英語のほぼ中間に位置し、どちらに近いか曖昧であるが、副詞の位置の分布よりブルガリア語英語に近いと判断できる。

5. おわりに

前節の考察により、自動生成された言語系統樹が印欧語の言語系統樹に類似することは単なる偶然ではないことが明らかになった。少なくとも冠詞の使用、名詞句の構成法、副詞の使用位置について、類似した言語では類似した傾向を示すことが示された。これらの傾向がトライグラムの分布に反映されるため、単語/品詞トライグラムに基づく手法により印欧語の言語系統樹に類似した言語系統樹が生成されたといえる。以上により、印欧語話者の英文には、母語に対応した言語系統樹が内在すると結論付けられる。

謝 辞

次の方々から、本研究に対して有益な助言を頂きました：船越孝太郎氏、早瀬光秋先生、河合敦夫先生、Robert Ladig 氏、Graham Neubig 氏、Vera Sheinman 氏、高村大也先生、David Valmorin 氏。心より御礼申し上げます。本研究の一部は (株) 教育測定研究所の研究助成により実施した。

参考文献

- [1] J. Aarts and S. Granger, Tag sequences in learner corpora: a key to interlanguage grammar and discourse, pp.132-141, Longman, New York, 1998.
- [2] B. Altenberg and M. Tapper, The use of adverbial connectors in advanced Swedish learners' written English, pp.80-93, Longman, New York, 1998.
- [3] D. Crystal, The Cambridge Encyclopedia of Language (2nd ed.), Cambridge University Press, New York, 1997.
- [4] S. Granger, Learner English on Computer, Longman, New York, 1998.
- [5] S. Granger, E. Dagneaux, F. Meunier, and M. Paquot, International Corpus of Learner English v2, Presses universitaires de Louvain, 2009.
- [6] B. Juang and L. Rabiner, "A probabilistic distance measure for hidden Markov models," AT&T Technical Journal, vol.64, no.2, pp.391-408, 1985.
- [7] K. Kita, "Automatic clustering of languages based on probabilistic models," Journal of Quantitative Linguistics, vol.6, no.2, pp.167-171, 1999.
- [8] M. Swan and B. Smith, Learner English (2nd Ed.), Cambridge University Press, 2001.