非母語話者英語における母語の親縁関係の保持

永田　亮†
† 甲南大学知能情報学部
E-mail: †rnagata@ konan-u.ac.jp.

1. はじめに

本稿では，非母語話者の英語における母語の親縁関係の保持について新しい知見を報告する．既に，文献 [5] で，印欧語話者の英語には印欧語の親縁関係が保持されることを示した．具体的には，印欧語話者の英文を単語／品詞列の分布に基づいてクラスタリングすると，印欧語の言語系統樹に類似した系統樹が構築できることを示した．更に，親縁関係の保持にかかわる要因の一部を明らかにした．

一方で，この知見は，親縁関係の保持は普遍的に起こるのかという新たな疑問をもたらす．少なくとも，母語に関する普遍性（親縁関係の保持は印欧語話者の英語以外でも起こるのか？）と言語能力に関する普遍性（親縁関係の保持は英語能力に独立して起こるのか？）について明らかになっていない．前者については，母語干渉が非母語話者の英語に一般的にみられることから，普遍であると予想できる．後者については，英語能力が向上すると母語話者の英語に近づくということを考慮すると，普遍でないと予想できる．以上の議論は，

仮説 I： 親縁関係の保持は母語に依存せず，非母語話者の英語に普遍である

仮説 II： 親縁関係の保持は英語能力に依存する

の二つの仮説にまとめることができる．

しかしながら，両仮説とも次のように反論が可能である．**仮説 I** については，印欧語話者の英語でのみ親縁関係の保持が起こる可能性もある．印欧語は，同系統である英語と言語システムを多く共有するため親縁関係の保持が可能であるのかもしれない．印欧語以外の話者の英語では，親縁関係とは別の特徴が支配的になる可能性は十分にある．また，outer circle(注 1)の英語については，更に議論が難しい．例えば，香港英語は，中国語の影響を受けると予想できる．一方で，expanding circle に位置する中国英語よりも outer circle の他の英語とより関係が深いとも予想できる．したがって，**仮説 I** が成り立つかどうかの確認が必要である．同様に，**仮説 II** についても確認が必要である．仮に，英語能力とは独立した母語干渉が存在すれば，英語能力には依存せず親縁関係の保持が起こる可能性がある．以上を考慮すると，**仮説 I** と **仮説 II** が成り立つかどうかは自明ではない．

そこで，本稿では，この二つの仮説の検証を行う．まず，アジア圏の英語を対象にして言語系統樹の構築実験を行うことで**仮説 I** の検証を行う．もし，アジア圏の英語から，対応する言語系統樹が構築できれば，**仮説 I** を支持する根拠となる．同様に，**仮説 II** の検証のため，英語能力を考慮した言語系統樹の構築実験も行う．更に，「確率的モジュールの存在」という理論を提案し，なぜ**仮説 I** と **仮説 II** が採択／棄却されるのかを説明する．

以下，2. で，仮説を検証するための方法について説明する． 3. で，アジア圏の英語から言語系統樹を構築した実験について述べる． 4. で，英語能力を考慮した言語系統樹の構築実験について述べる． 5. で，実験結果を考察する．

2. アプローチ

言語系統樹を構築する手法は，文献 [5] で提案したものを利用する．この手法では，各国英語を単語／品詞トライグラムベースの確率的言語モデルでモデル化する．この言語モデルに階層型クラスタリングを適用することで言語系統樹を構築する．クラスタリングで必要となる距離は，言語モデル間の距離として定義する．なお，手法の詳細については，文献 [5] を参照されたい．

対象データとして，アジア圏の英語を収録した The International Corpus Network of Asian Learners of English (ICNALE) [3] を使用する．同コーパスは，アジア圏の outer circle と expanding circle の英文を収録している（その他，inner circle の英文も収録）．ICNALE では，各英文に英語能力の情報が付与されている（inner circle は除く）．英語能力は，A_2（最も低い），B1_1，B1_2，B2+（最も高い）の 4 種類に分類されている(注 2)．この情報を**仮説 II** の検証に利用する．表 1 に ICNALE の概要を示す．

3. アジア英語を対象とした系統樹の構築実験

本実験では，ICNALE の全データを使用した．品詞解析には，学習者の英文専用に独自に開発した品詞解析器を使用した．ただし，品詞タグセットは Penn Treebank Tag-set [6]

(注 1)：Kachru [4] は，世界の英語を inner circle，outer circle，expanding circle の 3 種類に分類している．Inner circle とは，英語を母語として使用する区分である．Outer circle は，元植民地などで英語を公用語や国語として使用する区分である（例：香港の英語）．Expanding circle は，それ以外である．

(注 2)：各能力レベルと TOEIC のスコアの対応付けは，http://language.sakura.ne.jp/icnale/に詳しい．

表 1： ICNALE コーパスの概要.

Category	# of essays	# of tokens
Inner circle		
Native	400	88,792
Outer circle		
Hong Kong	200	46,111
Pakistan	400	93,100
Philippines	400	96,586
Singapore	400	96,733
Expanding circle		
China	800	194,613
Indonesia	400	92,316
Japan	800	176,537
Korea	600	130,626
Thailand	800	176,936
Taiwan	400	89,736
Total	5,600	1,282,086

に準拠する．解析済みの ICNALE から言語系統樹を構築した．言語モデルの構築には，Kyoto Language Modeling toolkit(注 3)を利用した．

図 1 に実験結果を示す．各ノードに付された数字は，二つのクラスタが併合されたステップを表す．印欧語に比べるとアジア圏の言語については親縁関係が明らかでない部分も多いが，図 1 の言語系統樹は関連する母語の親縁関係をある程度反映する結果となった．図 1 では，最初に，台湾英語と中国英語が一つのクラスタに併合されている．このことは，台湾英語と中国英語が中国語からの母語干渉を受けるという事実に一致する．言い換えれば，このクラスタは，シナ・チベット語族に対応すると解釈できる．次に併合されたのは日本英語と韓国英語である．両母語は，アルタイ語族に属するとされることが多い(注 4)．日本語英語と韓国英語の次に併合されたのは，タイ英語とインドネシア英語であるが，両地域で使用される言語は，異なる語族に属する；前者はタイ・カダイ語族，後者はオーストロネシア語族である．ただし，タイ諸語は，オーストロネシア語族との関係性が指摘されており，このクラスタも親縁関係を反映している可能性がある．以上の結果は，**仮説 I** を支持するといえる．

興味深いことに，図 1 の言語系統樹は，母語の親縁関係に加えて，三サークル（inner，outer，expanding）の関係性も保持している．このことは，outer circle では，母語の親縁関係ではなく，別の特徴が支配的であることを示唆する；そうでなければ，香港英語がシナ・チベット語族のクラスタ，フィリピン英語がオーストロネシア語族のクラスタに併合されるはずである．したがって，outer circle については，**仮説 I** は支持されない．

（注 3）：http://www.phontron.com/kylm/

（注 4）：両言語がどの語族に属するかは議論の余地が残る．しかしながら，現状では，アルタイ語族とすることが多い [1].

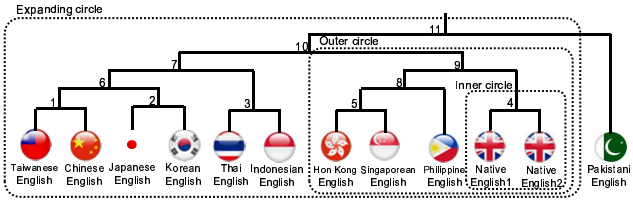

図 1： アジア圏の英語から生成された言語系統樹.

以上を踏まえると，**仮説 I** を部分的に採択し，次のように修正すべきであるといえる：**仮説 I′**：親縁関係の保持は母語に依存せず，expanding circle の英語に 普遍である．

4. 英語能力と親縁関係の保持に関する実験

仮説 II の検証用に，ICNALE から新しい実験データを作成した．図 1 で，最初にペアになっている各国英語において，片方のペアは英語能力が高いエッセイ（B1_2 レベルと B2+レベル）のみを含むように，もう片方は英語能力が低いエッセイ（A_2 レベルと B1_1 レベル）のみを含むように，ICNALE のデータを修正した．例えば，台湾英語と中国英語のペアでは，前者が B1_2 レベルと B2+レベルのエッセイのみを含むように，後者が A_2 レベルと B1_1 レベルのエッセイのみを含むようにした（詳細な組み合わせは，図 2 に示す）．この新しい実験用データから，言語系統樹を構築した（コーパス以外は， 3. の実験と同条件とした）．

この実験の意図は次の通りである．もし仮に，親縁関係の保持が英語能力と独立であれば，新しい実験データから構築される系統樹は図 1 の言語系統樹と類似するはずである．逆に，英語能力に依存するのであれば，異なった形になるはずである．

図 2 に構築結果を示す．図 2 より，新しい実験データから構築された言語系統樹が図 1 の言語系統樹に類似することは明らかである．言い換えれば，図 2 の言語系統樹も母語の親縁関係を保持している．よって，実験結果は，我々の予想に反し，**仮説 II** を棄却する．

5. 考　　察

上述の実験により，単語／品詞列の分布に基づいた手法を用いて，アジア圏の英語からも母語の言語の親縁関係を反映した言語系統樹が構築できることが確認された．この結果および文献 [5] の結果は，いずれも**仮説 I′** を支持する．

ここで，なぜ**仮説 I′** が成り立つかを説明するため，「確率的モジュールの存在」という理論を提案する．この理論では，次のような仮定をおく．まず，人の脳内に確率的な情報を保持するモジュール，すなわち，確率的モジュールが存在する

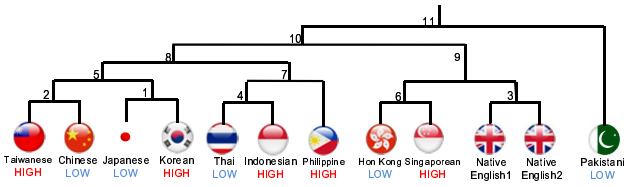

図 2： 英語能力に偏りがある ICNALE から構築された言語系統樹.

と仮定する．この確率的モジュールは，いくつかの確率の組により構成されている．各組は，任意性のある言語項目に対応しており，任意性の度合いは確率値によって表される．任意性のある言語項目とは，複数の候補のうちどれを選んでもよい項目のことである．代表例として，英語における副詞位置を挙げることができる（選択候補として文頭，文中，文尾がある）．任意性のある言語項目における各候補の選択はこの確率値に従う．また，確率的モジュールの確率値は，母語に対応した値に設定されていると仮定する．ただし，親縁関係が深い言語では，確率値は似た値に設定されているとする．その理由は，次の通り説明できる：(1) 祖語において，固有の確率値が発達し，その確率値は子孫となる言語に受け継がれる；(2) 時間と共に，いくつかの確率の組は値が変化する；(3) その確率は，次の子孫となる言語へ受け継がれる．具体例として次のようなケースを考えることができる．印欧祖語の確率的モジュールは，若干の確率値の変化とともに子孫の言語（例えば，ゲルマン祖語やイタリック祖語）に引き継がれたと仮定する．その後，ゲルマン祖語の確率的モジュールは，更なる確率値の変化と共に子孫の言語（例えば，ドイツ語やノルウェー語）に，同様に，イタリック祖語の確率的モジュールは，フランス語やイタリア語に引き継がれることになる．その結果，ドイツ語の確率的モジュールは，フランス語やイタリア語よりも，ノルウェー語の確率的モジュールに確率値が類似することになる．

以上の仮定のもと，非母語話者の英語における母語の親縁関係の保持を次のように説明することができる．非母語話者が，英語における任意性がある言語項目を使用する際，確率的モジュールが使用される．ただし，確率値が母語用に設定された確率的モジュールである．したがって，任意性がある言語項目における候補の選択は，母語での選択傾向を反映する．例えば，文頭の副詞に選好を持つ言語の母語話者は，英語でも文頭の副詞を好む．この選択選好を通じて，確率的モジュールの確率値が単語／品詞列の分布に暗に反映される．例えば，副詞位置であれば，*BOS RB* , や *NN RB* . などの文頭や文尾に対応するトライグラムの分布に反映される．単語／品詞列の分布に基づいたクラスタリングにより，母語の親縁関係を反映する言語系統樹が構築できるのはこのような理由のためである．この議論は，「確率的モジュールの存在」を認めると，どのような母語と対象言語の組み合わせについても成り立つ．必要となる条件は，母語および対象言語に任意性がある言語項目が存在することであり，この条件は全ての言語で成り立つと予想されるからである．

ここで問題となるのは，確率的モジュールの存在根拠をどのように示すかということである．幸い，文献 [5] には，その手掛かりが示されている．同文献では，親縁関係が深い言語では，名詞句の構成法，副詞位置，冠詞の使用に関する分布が類似することを報告している．具体例として，印欧語話者の英語における名詞句の構成法に関する分布を図 3 に示す．図 3 は，印欧語話者の英語から求めたトライグラム「名詞-of-名詞」の相対頻度を示す(注 5)．ここで，英語には名詞句の構成法として，名詞を連結した複合名詞と「名詞-of-名詞」による名詞句の間に任意性があることを指摘しておく（例：*education system* と *system of education*）．図 3 から，「名詞-of-名詞」に選好性があるロマンス諸語（フランス語，イタリア語，スペイン語）の英語では，他の英語に比べ「名詞-of-名詞」の相対頻度が高いことがわかる．逆に，名詞連結を好むゲルマン諸語（オランダ語，スウェーデン語，ドイツ語，ノルウェー語）の英語は，相対頻度が低いこともわかる．更に，「名詞-of-名詞」の相対頻度が，イタリック語派，スラブ語派，ゲルマン語派に対応して，三つのグループに大別されることもわかる．以上のことは，名詞連結と「名詞-of-名詞」の選択に関する確率的な情報が脳内に存在することを示唆する．

同様に，アジア圏の英語においても，「確率的モジュールの存在」に関する根拠を示すことができる．図 4 は，ICNALE における「名詞-of-名詞」の相対頻度を示す．図 4 から，親縁関係が深い英語のペアは，似た相対頻度を示すことがわかる．更に，図 5 は，副詞位置の分布についても同様な傾向にあることを明らかにする；図 5 の横軸と縦軸は，それぞれ，文頭の副詞の割合と文尾の副詞の割合に対応する．この図に

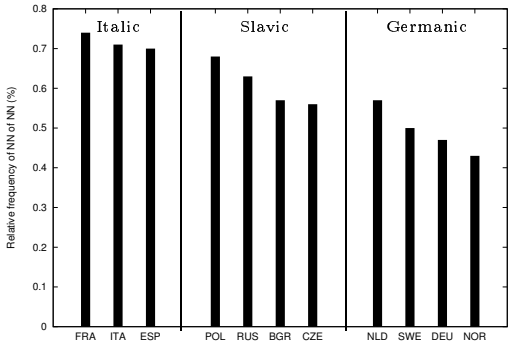

図 3： 印欧語話者の英文における「名詞-of-名詞」の相対頻度.

(注 5)：ICLE コーパス [?] を利用した．また，図 3〜図 6 で使用されているアルファベット 3 文字は国コード（ISO31661 alpha-3 codes）を示す．ただし，NS1（母語話者 1）と NS2（母語話者 2）を除く．

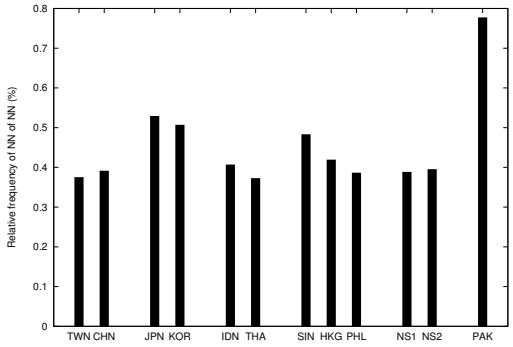

図 4： アジア圏の英語における「名詞-of-名詞」の相対頻度.

おいても，親縁関係が深い言語の話者の英語は，近い位置に分布する傾向がみられる.

図 3〜図 5 は，いずれも親縁関係が深い言語では，分布が類似することを示す．言い換えると，確率的モジュールに相当する情報が脳内に存在し，親縁関係が深い言語では，その値が類似することを示唆する.

「確率的モジュールの存在」は，母親縁関係の保持が英語能力に独立であるという 4. の実験結果も説明する．そのためには，英語能力が向上しても，確率的モジュールの確率値が大きく変化しないことを示せばよい．文法誤りと異なり，任

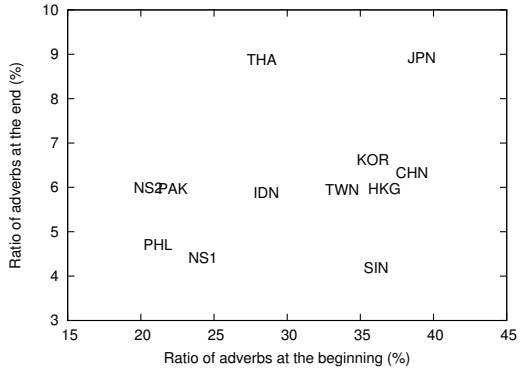

図 5： アジア英語における副詞位置の分布.

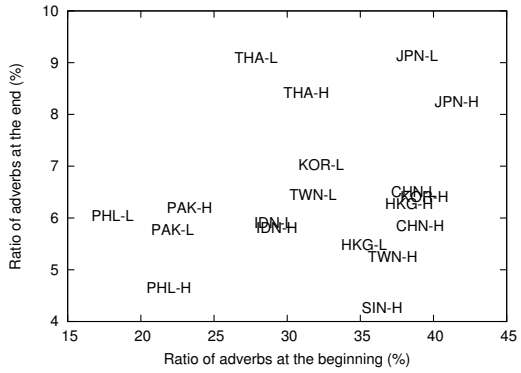

図 6： アジア英語における副詞位置の分布（英語能力別）.

意性がある言語項目の使用に対しては明示的なフィードバック（例えば，教師の指摘）が学習者に与えられることは少ない．なぜなら，任意性がある言語項目では，どの候補を選んだとしても誤りではないからである．例えば，副詞位置の選択で，文頭（例：Already, I have done it.），文中（例：I have already done it.），文尾（例：I have done it already）のいずれも，(若干の意味の違いは伴うものの）基本的には正しい．したがって，学習者は，確率値を変化させる機会がそもそも少ない．仮に，明示的なフィードバックが与えられたとしても，確率値を目的の値に変化させることは難しい．なぜなら，確率モジュール内の確率値を直接観測することができないからである．よって，英語能力が向上したとしても，確率値の変化が少ないことが導かれる．実際に，このことは非母語話者の英文に観測される．図 6 は，ICNALE における副詞位置の分布を英語能力別にプロットしたものである（X-H と X-L は，それぞれ X 英語における能力が高い／低いエッセイ群に対応する）．図 6 より，英語能力の違いにより分布が大きく変化しないことがわかる（同じ国の英語であれば，近い場所にプロットされる）．このことは，英語能力の変化により確率値が大きく変化しないことを示唆する.

6. おわりに

本稿では，非母語話者の英語における母語の親縁関係の保持に関する普遍性について報告した．実験結果は，(I) 親縁関係の保持は母語に依存せず，expanding circle の英語に普遍である，(II) 親縁関係の保持は英語能力に独立に起こる，を支持する結果となった．また，「確率的モジュールの存在」という理論を提案し，(I) と (II) がなぜ成り立つのかを理論的に説明した．更に，「確率的モジュールの存在」を肯定する根拠を示した.

謝　辞

本研究の一部は，私立大学等経常費補助金特別補助「大学連携等による共同研究」により実施した.

参考文献

[1] D. Crystal, The Cambridge Encyclopedia of Language (2nd ed.), Cambridge University Press, 1997.
[2] S. Granger, E. Dagneaux, F. Meunier, and M. Paquot, International Corpus of Learner English v2, Presses universitaires de Louvain, 2009.
[3] S. Ishikawa, A new horizon in learner corpus studies: The aim of the ICNALE project, pp.3–11, University of Strathclyde Publishing, 2011.
[4] B.B. Kachru, The Other Tongue: English Across Cultures, University of Illinois Press, 1992.
[5] R. Nagata and E. Whittaker, “Reconstructing an Indo-European family tree from non-native English texts,” Proc. of 51st ACL, pp.1137–1147, 2013.
[6] B. Santorini, Part-of-speech tagging guidelines for the Penn Treebank Project, University of Pennsylvania, 1990.