言語の表層語形から品詞はどの程度正確に予測できるか？

語彙獲得難度の絶対指標の提唱

黒　田　　航
杏林大学 医学部
kow.k@ks.kyorin-u.ac.jp

1　はじめに

英語の習得で苦労しなかった日本人は滅多にいない．英語の習得は様々な理由から日本人にとって魅力的な目標であるが，多くの人にとって「果たせぬ夢」である．その理由の一つとして，日本で実施されている英語教育の非効率性を指摘することは間違いではない．だが，この指摘が仮に正しいとしても，効率を上げるために何をしなければいけないのか？　母語話者偏重の教育法や会話中心の教育法の効能が大衆メディアを中心に無責任に吹聴されているが，それらに期待されている効果があるかは，大いに怪しい [1)]．

何が欠けているのは，日本人にとって英語の習得が難しい事の認知科学的に妥当な説明である．それがない状態で効率化を目指すのは，「溺れる者は藁をも掴む」の状態で，効果のない教育法に飛びついているだけである危険が高い．これを打開するには，日本人にとって英語の習得が難しい理由を，認知科学的に記述し，それを基にして対策を講じる必要がある．

1.1　言語距離で十分か？

言語間の非類似性を喩えて言語距離 (language distance) と言う．日本語から見れば韓国語は言語距離が近い．これに対し，日本から英語への言語距離は非常に遠い．多くの印欧語は，日本語からの言語距離が遠いが，これは言語の系統が異なるためである．

言語距離は言語の系統で決まるものではない．系統が異なる言語でも，相対的に習得が容易である，あるいは容易だと感じられる言語がある．日本人にとって習得が容易だと感じられる言語は，韓国語，イタリア語，スペイン語，ヘブライ語などがある [2)]．

系統に帰着できない要素を考慮に入れると，言語距離を構成する次元には少なくとも (1) の三つがある:

(1)　a. 語彙的類似性: 語彙素の共有度合いで定義
　　b. 音韻的類似性: 音素集合や音素配列パターンや超分節要素の共有度合いで定義
　　c. 統語的類似性: 語彙要素の配列パターンの共有度合いで定義

言語の系統が強く影響するのは (1a) のみである．

これを基にするなら，日本語から測った英語の（絶望的な）距離は，次の三つの要素に分解される: 言語の系統が違うため，語彙要素が共有されていない．英語には日本語にない音素が多い．英語の基本語順は SVO であり，基本語順が SOV である日本語との類似性は低い (VO か OV かが本質的な違いを生む)．これに，英語自体が混成言語であり (Denning, Kessler, and Leben 2007)，語彙要素の使用に一貫性がないことを加えても良いだろう．

以上の論点は英語の習得し難さを論じる時に良く持ち出されるが，問題がないわけではない．何らかの絶対的基準で英語が他の言語に比べて習得の難しい言語かどうかは，知られていない．

これは，言語距離が二言語間に定義される相対的基準だという点に理由がある．例えば「母語を未獲得な知的エージェントにとって，英語の習得にどれぐらいの負荷がかかるか？」という問題には言語距離を使って答えられない．この種の問題に仮にでも答えを出すには，(2) で定義される量 S を基にして，言語の習得の難易度の順序づけが可能でなければならない．

(2) 母語を未獲得な知的エージェント A にとって，言語 L の習得にどれぐらいの負荷 S がかかるか？

量 S は，言語の習得の難易度の絶対基準となる指標になる．だが，そういう指標 S として何を想定すれば良いだろうか？

1.2　本研究の目的

言語習得の負荷 S が複数の要因 $\Phi = \{\phi_1, \phi_2, \ldots, \phi_n\}$ で複合的に決まるのは明らかである．本研究は (3) を想定し，それに基づいた調査と分析の結果を報告する．

[1)] 効果がないという証拠も上がっていないが，だからと言って効果があるとは言えない．
[2)] この種の主張は逸話が基であり，実験的な証拠が伴っていない．

(3) 言語 L の表層の語形から品詞 (part-of-speech = pos) を予測できる程度 P が，L の習得の負荷 S を決める複数の要因 Φ の一つである．

最終的には英語が認知負荷の観点で習得の難しい言語の一つであることを示すことが目的だが，本研究は予備的なものである．屈折が豊かな言語の例であるチェコ語の語形の品詞予測力を英語のそれと比較し，英語の語形による品詞予測力が低いことを示す．

2 理論と方法

英語が他言語と大きく異なるのは，語彙の混合率の高さと形態論の貧弱さである．多くの言語の習得の試みた筆者の実感として，特に後者に関して英語では語形から品詞を高精度で予測することが困難であると言える [3]．これは学習の際に明らかに障害となる．

直観的には，品詞ごとに典型的な語形 (典型的な接頭辞や接尾辞) があることは，語彙習得を促進する [4]．この直観を次の仮説として定式化する:[5]．

(4) 言語 L の表層語形 f から，f の品詞が高精度で予測できるほど L の語彙獲得は容易である．

(4) の仮説の下では，言語 L の表層語形 f から，f の品詞がどれぐらいの精度で予測できるか？を調査することで語彙習得の難易を近似することが可能になる [6]．より具体的には，次の調査を実施することで，言語 L の習得負荷を近似できる:

(5) L の表層語形 f の部分 n-gram から，f の品詞がどれぐらいの精度で予測できるか？

本研究の提案は，語形 f の語頭か語尾 (=語末) の文字 n-gram で f の品詞をどれぐらい正確に予測できるかを評価 [7] し，それが低いことを Φ の一つと考えようということである．

どの言語でも f の 1-gram で f の品詞を予測できないことは明らかなので，比較的 n の小さい n-gram で品詞を予測できるかを考える．本研究では，語頭や語末の 3-gram と 4-gram で語形の品詞を予測できるかを考える．

文字の共起は音素の共起を反映し，強い制約を受けているので，文字の全組み合わせが実用されることはない．計算量の評価としては，L の表記で実際に使われる語形 $\{f_1, f_2, \ldots, f_N\}$ の語頭か語末の n-gram のすべての異なり数 B を基準にすればよい．経験的評価としては，英語やチェコ語では 2-gram で 100 桁，3-gram で 1000 桁の異なり数が存在する．このため，取りこぼしを覚悟し，使用頻度の少ない語末 n-gram は FCA に与える属性から外している．

2.1 データと分析法

次のような手順でデータを作成し，調査を行った．

(6) a. 基本データの収集: Karel Čapek の R.U.R. の原典 (チェコ語) と英語への翻訳で使われている語形を網羅的に集め，F_{en} (3000 弱個) と F_{cz} (5000 弱個) を得た．これらに次の 2 種類の属性を付与する:

b. 品詞属性の付与: F_x の語形に手作業で Adj, Adv, Conj, Noun, Verb, Prep の品詞を付与した (品詞の排他性は前提としない).

c. 語形属性の付与: 語末 n-gram の全体集合を取得し，一定頻度以上のものを語形の属性として付与した．これは Excel の関数を使って自動付与した．

語形に付与するのは可能な全品詞である．例えば *park* に J = 0, P = 0, D = 0, N = 1, V = 1, Aj = 0, Av = 0 を，*parking* に J = 0, P = 0, D = 0, N = 1, V = 0, Aj = 1, Av = 0 を，*parked* に J = 0, P = 0, D = 0, N = 0, V = 1, Aj = 1, Av = 0 を付与する．

(6b) の作業が完了しなかったので，FCA による分析の対象は，それぞれから無作為抽出した 800 語形を含む 2000 語弱個の語形に限った．

R.U.R. のチェコ語原典中の語形の異なり数は 4,424，語形当りの品詞の平均は 1.04 個であった．R.U.R. の英語訳中の語形の異なり数は 2,645，語形当りの品詞の平均は 1.39 個であった [8]．この指標からも，英語がチェコ語より，語形の品詞予測力が低いことが判る．

2.1.1 Formal Concept Analysis を使った分析

(6) で用意したデータを解析するのに統計的な手法 (例えば機機械学習の成功率) を使うことは可能だが，本研究は形式概念分析 (Formal Concept Analysis: FCA) (Ganter, Stumme, and Wille 2005; Suzuki and Murofushi 2007) を手法に選んだ．本研究は予備的な研究なので，結果の直観的な理解が容易な分析法の

[3] 語形に品詞の手がかり少ない時，文法役割の付与が困難になる．
[4] これは名詞や動詞のような開クラスの語を獲得するために，特に重要な条件となると考えられる．
[5] 品詞は文法機能の良い近似となるとは言え，語形の豊富さは記憶負荷とトレードオフの関係にあると考えられるので，語形が多い事が効果的な語彙習得にとって無条件に好ましいわけではない．本研究では議論の単純化のためにこの点を無視する．
[6] 調査対象は lemma = lexical unit ではなく活用形=実用形である．
[7] 一次近似としては，語頭か語尾を組み合わせて使う必要はない．いずれが有効かは，言語ごとに決まっている．
[8] round が J = 0, P = 1, D = 0, N = 1, V = 1, Aj = 1, Av = 1 で 5 重の属性をもち，曖昧性が最大だった．

方が望ましいと考えたからである．FCA の解析では Concept Explorer (ConExp) の v1.3[9] を使った．

FCA の最大の利点は，データの排他分類を前提とせず，多重分類 (cf. soft clustering) が実現できる点にある．

3 結果と考察

3.1 結果

以下に示す FCA の解析から次の結果を得た:[10]

(7) a. 英語でもチェコ語でも語末 n-gram の品詞予測力は，英語で $n = 4$，チェコ語で $n = 3$ ぐらいで頭打ちになる．

b. 英語の語末 4-gram の品詞予測力は，チェコ語の語末 3-gram の品詞予測力に及ばない．

3.1.1 英語の語末 4-gram と 3-gram の品詞予測力

英語の語末 4-gram と品詞の対応を図 1 から図 3 に，語末 3-gram と品詞分類の対応を図 4 から図 6 に示す．

英語では，語末 4-gram で何とか一部の予測が成立するが，予測精度は低い．紙面の都合で図に示していないが，語末 5-gram でも 4-gram に比べて大きな改善はない．従って，品詞予測力は低い精度のまま，4gram ぐらいで頭打ちになっている．精度の低さは Noun と Verb の重複，Verb と Adj の重複の強さである．語末 3-gram では品詞の予測は成立していない．

3.1.2 チェコ語の語末 3-gram と 2gram の品詞予測力

チェコ語の語末 3-gram と品詞分類の対応を図 7 から図 9 に，語末 2-gram と品詞分類の対応を図 7 から図 9 に示す．

チェコ語では，語末 3-gram で非常に高い精度の予測が成立する．Verb の分離が非常に良い．語末 2-gram でも予測の精度はそれほど低くない．

3.2 考察

3.2.1 *n*-gram が担う情報

字母の情報量は，言語ごとに異なる．英語の表記の字母 (alphabet) を構成する 27 字では 補助記号 diacritic が使われていないが，チェコ語の表記の字母を構成す

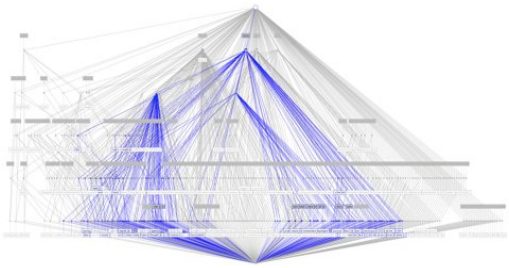

図 1: 英語の語末 4-gram FCA (Verb を選択)

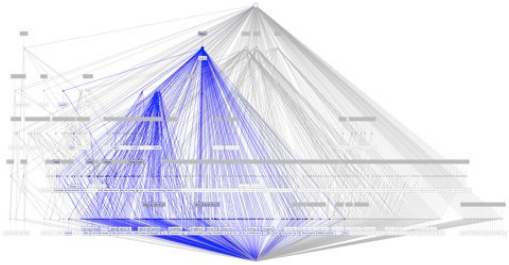

図 2: 英語の語末 4-gram FCA (Noun を選択)

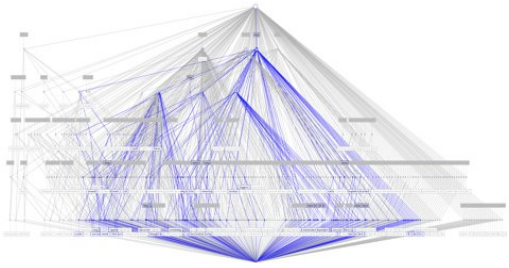

図 3: 英語の語末 4-gram FCA (Adj を選択)

る 40 字には，(8) のように，太字で強調した 14 個の補助記号つき文字が追加されている (大文字と小文字の区別は無視する):

(8) *a, **á**, e, **é**, **ě**, i, **í**, o, **ó**, u, **ú**, **ů**, y, **ý**, b, c, d, **ď**, f, g, h, j, k, l, m, n, **ň**, p, q, r, **ř**, s, **š**, t, **ť**, v, w, x, z, **ž*** czech

チェコ語の n-gram は，英語の n-gram より情報が多い．

本研究の範囲では，チェコ語の 2-gram による予測が英語の 3-gram による予測に，チェコ語の 3-gram による予測が英語の 4-gram による予測に相当すると考えているが，厳密な評価はしていない．

3.3 品詞予測での有標性の役割

品詞の識別は有標性に基づいていると思われる．チェコ語の場合，Noun は無標なクラスであり，動詞や形容詞がそれから分離されるだけでも，語形による品詞の識別は高精度で実現される．

[9] http://conexp.sourceforge.net/ で入手可能．

[10] ConExp 1.3 の Concept Lattice の構築の際に冗長な属性と対象を削除している．

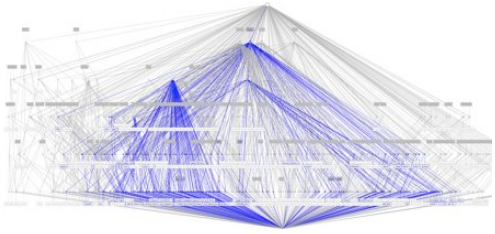

図 4: 英語の語末 3-gram FCA (Verb を選択)

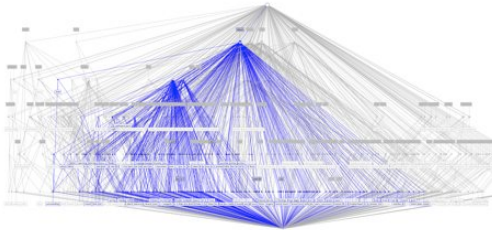

図 5: 英語の語末 3-gram FCA (Noun を選択)

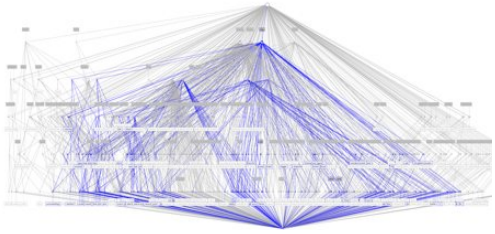

図 6: 英語の語末 3-gram FCA (Adj を選択)

4 まとめと今後の展望

FCA を使った解析の結果から，英語の語形 f の語末 n-gram (n は 4 以下) から，f の品詞が (少なくともチェコ語に比べて) 高精度に予測できないことが確かめられた．これは，表層語形から品詞が高精度で予測できるほど語彙獲得は容易であると想定するならば，英語が語彙獲得の難しい言語である可能性を示唆する．

今後の展望として，ドイツ語とフランス語と韓国語を同じ手法で分析し，結果を比較したい．日本語も同じ手法で分析したいが，表記の問題を解決する必要がある．

参考文献

Denning, K., B. Kessler, and W. R. Leben (2007). *English Vocabulary Elements* (2nd ed.). Oxford University Press.

Ganter, B., G. Stumme, and R. Wille (2005). *Formal Concept Analysis: Foundations and Applications*. Berlin/Heidelberg: Springer.

Suzuki, O. and T. Murofushi (2007, April). Formal Concept Analysis: Introduction, support softwares, and applications. *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics 19*(2), 103–142.

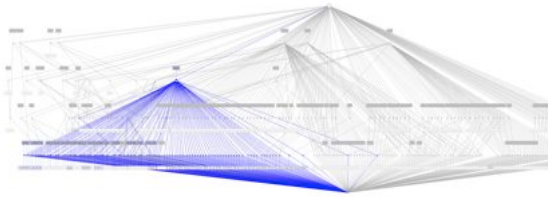

図 7: チェコ語の語末 3-gram FCA (Verb を選択)

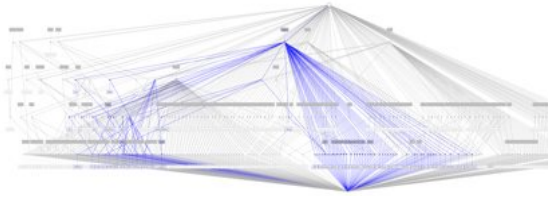

図 8: チェコ語の語末 3-gram FCA (Noun を選択)

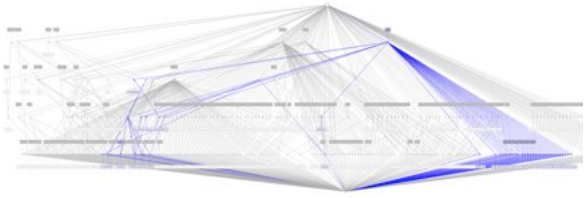

図 9: チェコ語の語末 3-gram FCA (Adj を選択)

図 10: チェコ語の語末 2-gram FCA (Verb を選択)

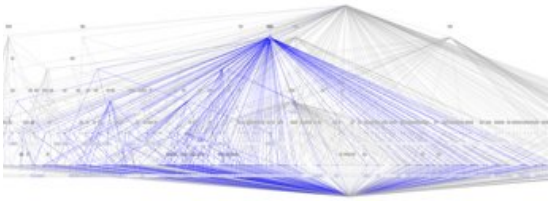

図 11: チェコ語の語末 2-gram FCA (Noun を選択)

図 12: チェコ語の語末 2-gram FCA (Adj を選択)