英語心内辞書データの統計的解析

小林　景　　数理・推論研究系　助教

【研究の経緯】

本研究を始めるきっかけとなったのは，外国語学習を専門とされている折田充教授（熊本大学）から，英語学習者と母語話者とに対して行った実験データを元に，それぞれの心内辞書の相違を統計学的に評価する方法についてご相談を受けたことである．

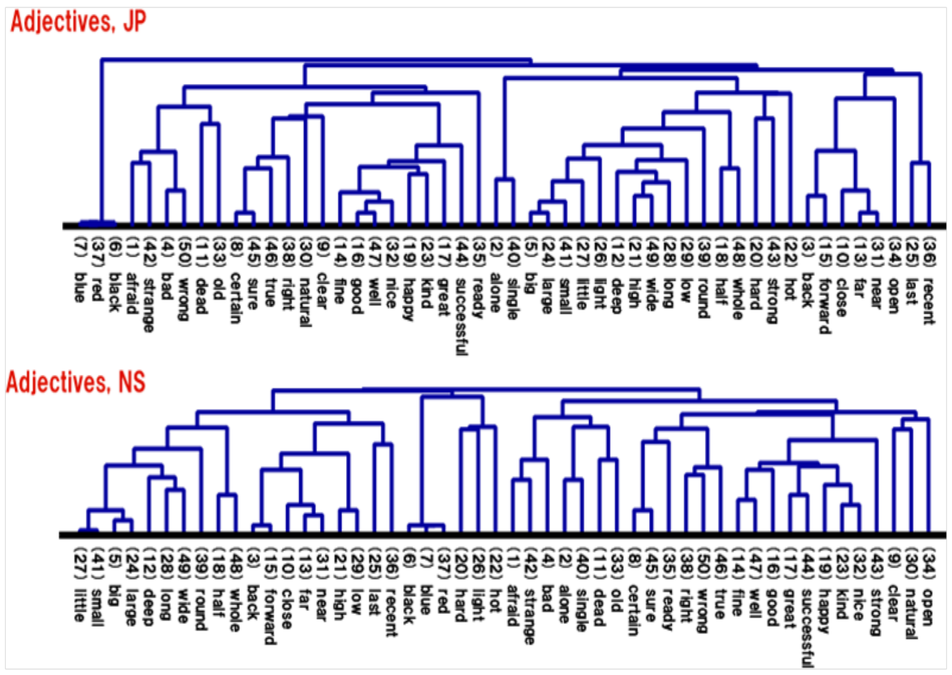

図2：　形容詞の分類に基づいたデンドログラム
（上：日本人，下：英語母語話者）

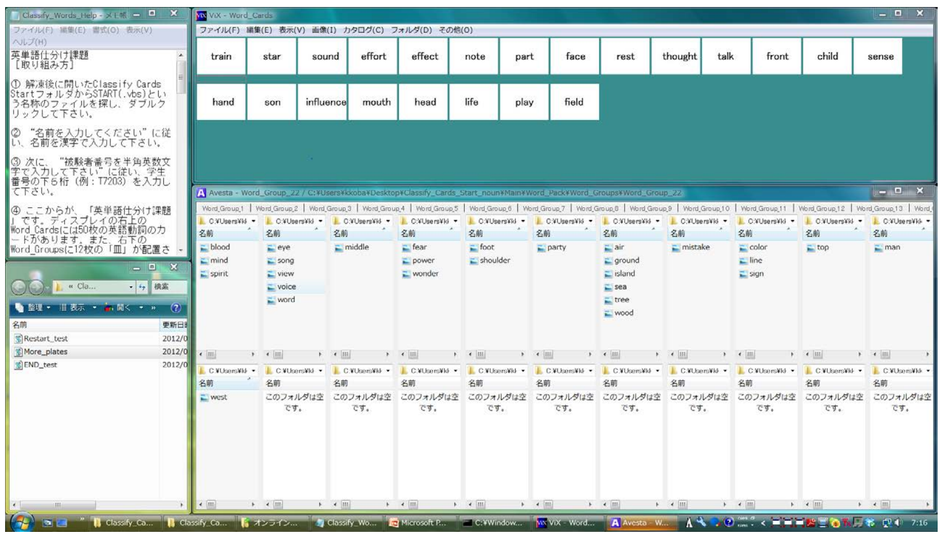

図1：　単語カードの仕分プログラム

【研究対象のデータと解析目的】

具体的な解析データは，日本人英語上級者と英語母語話者の各30名の被験者に，「50個の英単語をあなたが考える意味のまとまりにグループ分けしてください．」という指示を与えて得られたグループ分けの結果である．また英単語の種類を，頻出度の違う単語や，品詞が異なる単語にすることにより，その結果がどのように変化するかということも調べる．

本研究の大きな特徴は，心内辞書の構造としてデンドログラムと呼ばれる木構造（図2）を仮定することである．よって研究の目的は，図2のような2つのデンドログラムの違いを評価することだといえる．

【並べ替え検定を用いた解析手法】

そこで本研究では，並べ替え検定と呼ばれる統計的手法を用い，「2つのグループの心内辞書に差異は無い」という仮説のもとでダミーのデンドログラムを多数生成させ，それらと実際に得られたデンドログラムを比較することにより，定量的に差異を評価する手法を提案した．図3は，ここで用いた並べ替え検定の概略を表したものである．

ここで提案した手法は新しいものなので，その正当性を評価する必要がある．そこで，デンドログラムの構成法として，局所線形性という性質をみたすものを用いれば，提案した並べ替え検定は一致性を持つことを証明した．幸いにも，広く用いられている群平均法などのデンドログラム構成法はこの局所線形性を持ち，我々の手法に適用できることが理論的にも保証された．

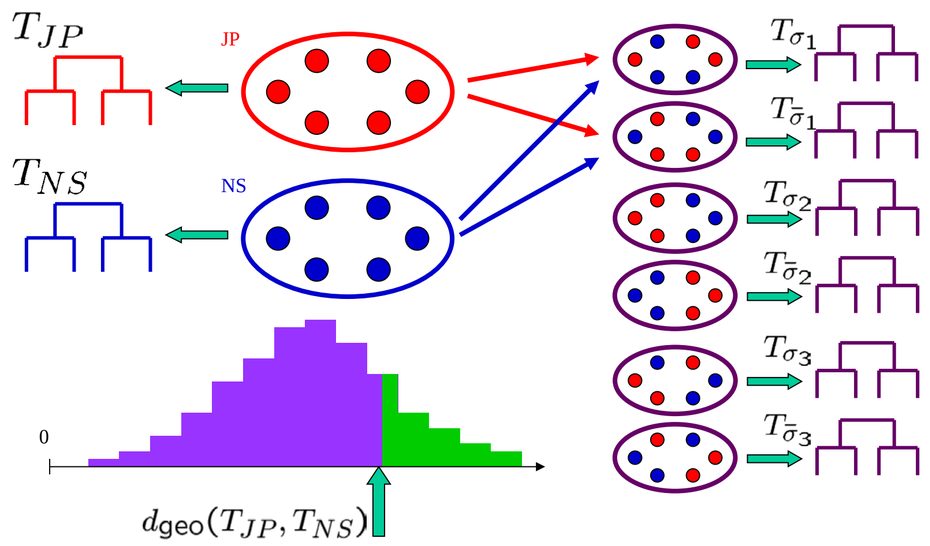

図3：　デンドログラムの並べ替え検定の概要図

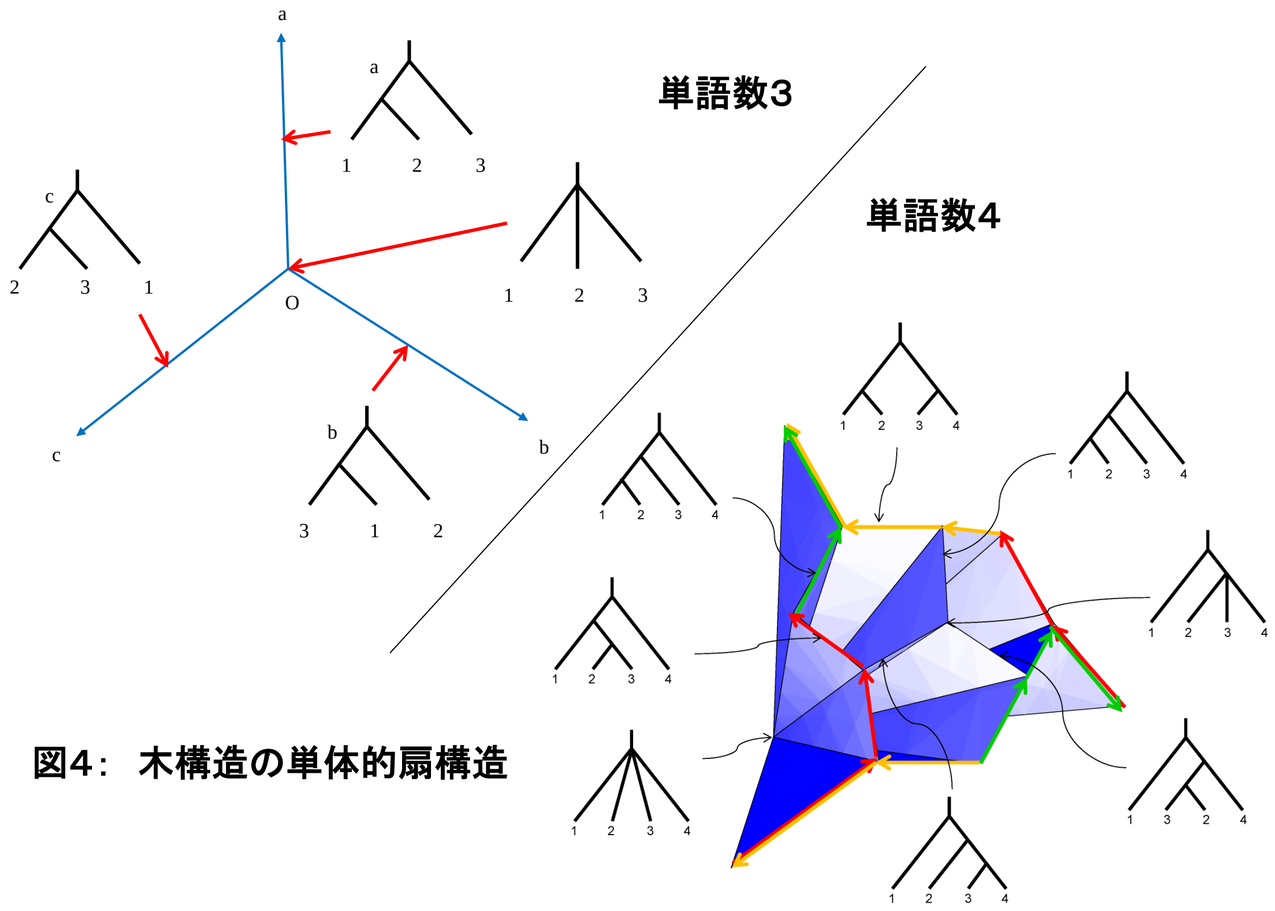

図4：　木構造の単体的扇構造

【単体的扇構造の理論と応用】

並べ替え検定で通常扱われる統計データは，通常はベクトルなどの単純な構造を持っている．ところが，デンドログラムの集合の幾何学は，単体的扇と呼ばれる構造を持っていることが知られている（図4）．一般的な単体的扇は，複雑な構造のため解析が困難だが，デンドログラムの集合は，その中においてCAT(0)とよばれる数学的に良好な性質を持っている．そのおかげで測地線を一意に定義でき，また近年，測地線をO(n^4)で計算する手法が開発された．それを用いて測地距離を用いた新たな並べ替え検定も提案し，通常の場合と比較した（図5）．

【心内辞書データの解析結果】

単語の種類	高頻出	頻出	動詞	名詞	形容詞
	1.76	49.6	1.20	3.88	0.12

- 母語話者と日本人の違いは　形容詞＞動詞＞名詞
- 英語上級日本人は初級日本人より母語話者に近い
- クラスター数は　母語話者＞日本人
- ただし日本人に英単語と日本語訳語を仕分けさせるとクラスター数は　英単語＞訳語

現在，これらの結果も利用した効率的な英語学習法の開発についてのプロジェクトが進行中である．

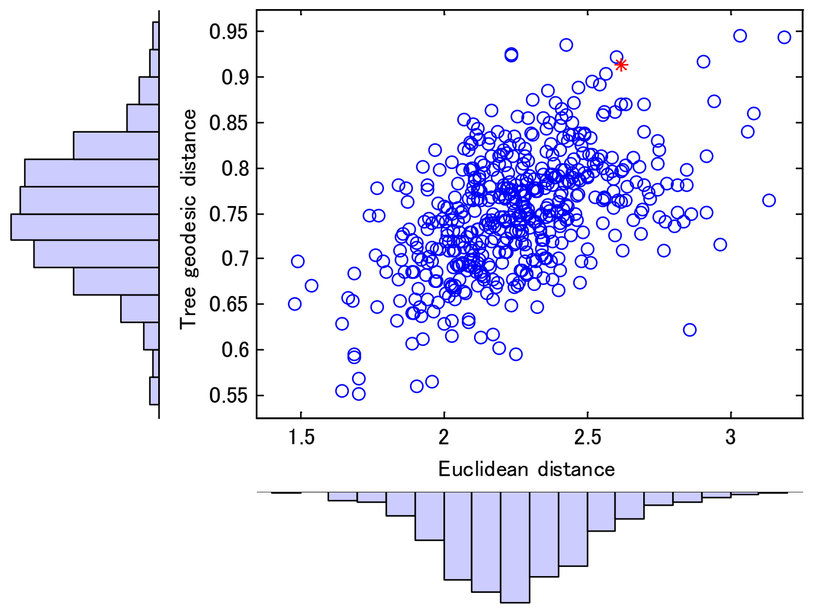

図5：　ユークリッド距離と木測地距離をそれぞれ用いた並べ替え検定のp値比較例