日本語の 2 字熟語ネットワークとその解析

山本健、山崎義弘
早大先進理工、東京都新宿区大久保 3-4-1
yamaken@toki.waseda.jp

Two-kanji compound words in Japanese language as a complex network

Ken Yamamoto and Yoshihiro Yamazaki
Department of Physics, Waseda University, Tokyo, 169-8555, Japan

Abstract

We report some properties of a network of two-kanji compound words. This network has a property of "small-world" and a degree distribution of the network displays power law.We also mention a network formed by common-use kanji. This network is viewed as a subclass of the original network, and also has small-world property. However, a degree distribution of this network has no clear property of "scale-free." We propose a cluster growth model for a selecting process of common-use kanji.

Keywords

network, two-kanji compound word, scale-free, cluster growth

1 2 字熟語ネットワーク

日本語の単語には多くの 2 字熟語がある。2 字熟語を 2 つの漢字を結ぶ枝とみなせばネットワーク構造があらわれる。本研究では 4 冊の国語辞典 "広辞苑（第 4 版）",“岩波国語辞典（第 5 版）", "三省堂国語辞典（第 4 版）",
"光村国語学習辞典（改訂版）" を用い、各々から見出し語として収録されている 2 字熟語のみを取り出してネットワークを構成した（光村は小中学生向けの辞典で、収録されている熟語は全て常用漢字で構成されている)。このネットワークは多重枝および自己ループを含む有向ネットワークであるのだが、本研究ではこれらを無視し、無向ネットワークとして解析する（ネットワークの一部を図 1 に示す)。また、いずれの辞書のネットワークも全ての漢字が 1 つに連結しているわけではなく、いくつかのクラスターに分かれているのだが、その中の最大クラスターを解析の対象とする（約 90% の頂点が最大クラスターに属している)。

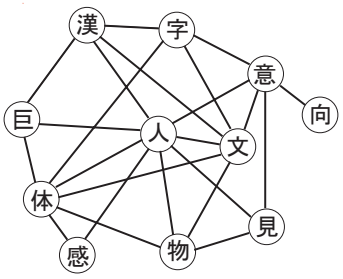

図 1: 2 字熟語ネットワーク

2　結果

4 冊の辞書すべてについて、小さい頂点間距離および大きいクラスター係数をもつというスモールワールドとしての性質がある。また、次数分布は “光村” 以外のネットワークでベキ則を示し、その指数はいずれも -1 に近い値となった（図 2）。

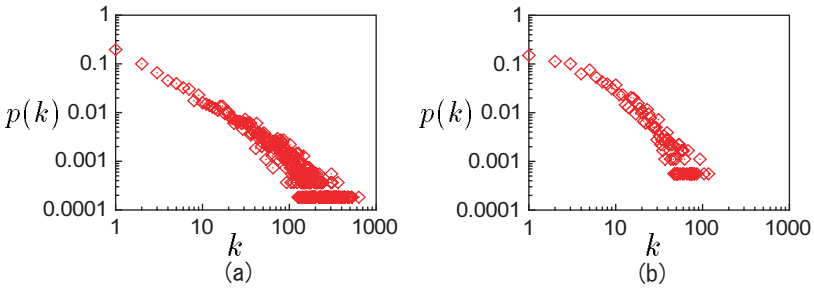

図 2: 次数分布　(a) 広辞苑（ベキ指数 -1.04）、(b) 光村

3　常用漢字のネットワーク

“法令・公用文書・新聞・雑誌・放送等、一般の社会生活で用いる場合の、効率的で共通性の高い漢字を収め、分かりやすく通じやすい文章を書き表すための漢字使用の目安” （昭和 56 年内閣告示第 1 号）を示す目的で 1,945 の常用漢字が制定された。2 字熟語ネットワークを常用漢字に制限したとき、低次数の頂点数は次数によらずほぼ一定となり次数分布はベキ的にならない（図 3）。これは図 2 (b) にもみられる特徴である。

4　常用漢字のモデル

ネットワーク全体から常用漢字を選び出す過程をモデル化する。まず、ネットワークの中からランダムに頂点を選び初期クラスターとする。次に、この頂点に隣接する頂点の中から 1 つの頂点を確率的に選びクラスターに取り込む。クラスターに隣接する次数 k の頂点が次にクラスターに取り込まれる確率は k^{α} に比例すると仮定する（α は実定数）。この操作を繰り返し、成長したクラスターを常用漢字とみなす。定数 α は平均次数が実データと一致する値とし、3 冊の辞書で $\alpha \simeq 1.3$ を得た。このモデルによる結果は実際のデータをよく再現する（図 4）。

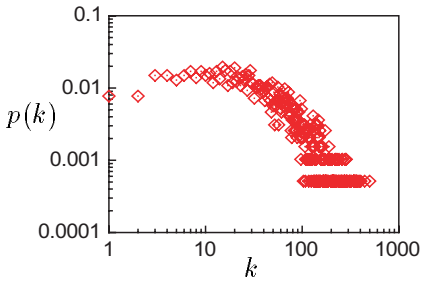

図 3: 常用漢字のネットワークの次数分布（広辞苑）

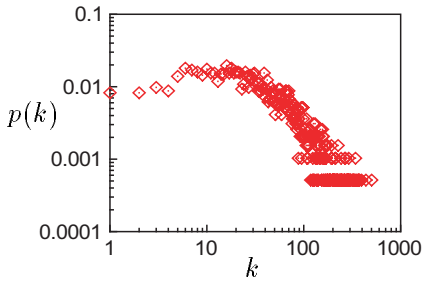

図 4: モデルから得られた次数分布（広辞苑）