# 多要素日本語複合語の構造解析

## Structural Analysis of Multicomponent Japanese Compound Terms

小山　照夫
Teruo KOYAMA

### 何がわかる？

いろいろな分野の論文を探したり、論文の間の関係を発見する手がかりを与えるために、当該分野で用いられる用語の体系化を目指します。
まず、用語として重要な複合語の構造を判定することにより、用語の間の関係を明らかにすることを目的としています。

### どんな研究？

複合語はいくつかの基本語の組み合わせとして用いられる用語ですが、３要素以上の語では、その構造がどのようになっているかが大切です。
用語を構成する要素の統計処理により。３要素の日本語複合語の構造を推定します。

## 研究の目的

**大規模商業施設　は、「大規模商業　施設」？　それとも「大規模　商業施設」？**

３形態素以上から構成される複合語では、語の構成される構造により、複合語の意味が異なってくる。このため、複合語の構造を正しく把握することは、用語の体系化にあたって重要な課題となる。
この研究では、構造の多義性が発生するもっとも簡単な、３つの形態素からなる日本語複合語について、その構造を正しく推定する方法を研究する。

## 研究の方法

テキストコーパスを形態素解析し、あらかじめ定められた、名詞系要素の連接からなる、複合語候補を抽出しておく。

これら候補のうち、３要素のもので、コーパス中に２回以上出現するものについて、どちらの２要素が先に結合していると考えられるかを、各形態素と、２要素連接の出現頻度を手がかりに、統計的手法により推定する。

手順
１．あらかじめ形態素解析をおこなったテキストコーパスから、一定の基準で、用語候補となる名詞関連要素の連接を抽出する。
２．このうち、３要素からなる用語候補で、２回以上出現しているものを選択する。ただし、特異性の高い連接を含むものをのぞく。
３．用語ＡＢＣについて、連接ＡＢおよびＢＣの連接強さを右の基準で推定し、強度の大きいほうが先に結合していると考える。

結合強度

$$S = \frac{F(AB)}{\sqrt{F(A) \times F(B)}}$$

ただし
F(A)：　形態素Aの頻度
F(B)：　形態素Bの頻度
F(AB)：連接ABの頻度

結果
用語候補として抽出されたもの　13,374
特異性の高い候補を除いた結果　10,952
用語としての精度　86%　（ランダムサンプルした300データの評価結果）

結合順序が正しく評価できたもの
順序判定精度　82%　（ランダムサンプルした300データの評価結果
非用語でも結合順序が正しければOKとした）

NII　連絡先：　小山　照夫（Teruo KOYAMA）／ 国立情報学研究所　情報社会相関研究系　教授
TEL：03-4212-2509　FAX：03-3556-1916　Email：t_koyama@nii.ac.jp