部首情報と係り受け単語出現における統計的規則：コーパス分析に基づく客観的指標による検討

Statistical regularities of radicals and word appearances: A corpus based approach

猪原 敬介[1]，上野 泰治[2,3]
Keisuke Inohara, Taiji Ueno

[1]電気通信大学，[2]名古屋大学，[3]日本学術振興会
The University of Electro-Communications, Nagoya University, Japan Society for the Promotion of Science
kei.inohara@gmail.com

Abstract

This study investigated whether radicals, an example of sub-lexical visual information in Chinese/kanji, contribute to computation of character/word meaning. We consulted a noun-noun corpus extracted from Japanese web texts. The analysis showed that nouns including radical friends tended to take more similar nouns than nouns with radical enemies. These findings suggest that characters/words with similar meanings tend to share radicals in kanji, which may explain how children are able to efficiently learn to use the vast number of characters in Chinese/Japanese.

Keywords — semantic radical; predicates; orthography; semantics

1. はじめに

単語の綴りや音からどのようにその意味を計算するかは言語を扱う認知科学において中心的な問題である[1][2]。本研究では，このうち，単語綴りからの意味の計算に焦点を当てた。特に，漢字特有の綴りの一部に関する情報（sub-lexical information）である部首が，意味計算にどのように関わるかを検討した。

部首と漢字の意味の関係について検討したこれまでの研究は，部首の意味と一貫している単語の意味を参加者に直接尋ねる，またはカテゴリ化させるといった主観的な意味類似性評価方法を用いていた[3][4][5]。これに対し，Inohara and Ueno [6] は，言語コーパスに基づく客観的指標により，単語の意味類似性を評価した。具体的には，単語のもつ意味性（semantics）の指標の一つである，係り先（predicates）に注目した。この背後には、似た意味を持つ単語は，似た係り先を持つという考え方がある [7]。結果，「ある名詞が出現し，その係り先としてある動詞が出現するとき，名詞に含まれる漢字の部首と出現する動詞に統計的関連がある」ことを明らかにした。このことは，部首は意味を予測し得ること，部首を考慮した漢字学習が言語処理を促進することを示唆している。

本研究は，「名詞-助詞-動詞」の係り受けコーパスを用いた Inohara and Ueno [6] の知見の一般性を検討するため，新たに「名詞１-助詞-名詞２」の係り受けコーパス [8] を用いて検討を行う。

2. 方法

2.1. 漢字データセット

(a) 近藤・天野 [9] から部首頻度の高い上位5%に入る 13 部首を抽出し, (b) その 13 部首を持つ漢字で，天野・近藤 [10] において漢字頻度が高い上位 25%に入るものを取り出した。結果として，536 漢字が抽出された。それぞれの部首の平均漢字数は 41.2 文字（SD = 21.4, range = 15:82）であった。

2.2. コーパス

日本語ウェブサイト 100 万ページに基づいて作成された名詞 1 - 助詞 - 名詞 2（例　友人 - に - 手紙）の係り受けコーパス [8] を用いた。助詞情報は無視して，名詞１ - 名詞２というペアで分析した。

(a) 出現頻度が 1000 以下のペアを除外，(b) 記号やアルファベットを含む名詞を持つペアは除外，(c) 漢字データセットの 536 漢字を含まないペアは除外，という処理を行い，結果として，102,354

ペア（名詞1-名詞2）を得た。異なり語数は，名詞1が13,783語，名詞2が5,386語であった。

2.3. 手続き

536漢字のうち，19漢字はコーパスに一度も登場しなかったので除外した。また，これらの漢字を含まない名詞ペアを除外した。その結果，5,386語の名詞2の異なり語は2,599語に減少した。

ある漢字が名詞1に含まれたとき，どの名詞2に登場したかを示す漢字（517)×名詞2（2,599)行列を作成した。名詞ごとに出現する頻度が異なるため，行列の要素はその名詞が出現する頻度で除された。すなわち，漢字×名詞2行列の要素は出現確率であった。

3. 結果

すべての漢字ベクトル間の類似度(コサイン)を計算し,部首を共有する漢字との平均類似度,共有しない漢字との平均類似度を計算した。結果，共有する漢字との平均類似度（共有条件：$M = 0.718$, $SD = 0.180$）は，共有しない漢字との平均類似度（非共有条件：$M = 0.714$, $SD = 0.179$)よりも有意に高かった（$t(516) = 2.20$, $p < .05$, $d = 0.10$)。比較のためにInohara and Ueno [6] の結果も表1に併記した。

4. 考察

ある名詞が出現し，その係り先としてある単語が出現するとき,「名詞に含まれる漢字の部首と出現する単語には統計的関連がある」ことが，係り先の単語が動詞の場合 [6] だけでなく，名詞の場合にも一般化された。ただし，表1のように，その効果量はInohara & Ueno [6] と比較して小さかった。この原因として，作成した漢字×係り先行列の性質の違いが考えられる。Inohara and Ueno [6] では,行列の中の0以外の要素の数および合計値が本研究よりも多く，単語間の係り受け関係について，より多くの情報を持っていたと考えられる。今後はより大規模なコーパスを用いた上で行列の情報を圧縮するなど，コーパスやその処理方法についての改善が必要であろう。また、Inohara and Ueno [6] の「名詞 - 助詞 - 動詞」コーパスでは、「手紙 - を - 送る」のように名詞と動詞の係り受け関係が必ず存在したが、本研究の「名詞1 - 助詞 - 名詞2」コーパスでは、「不幸 - の - 手紙」のような係り受け関係があるものと、「友人 - に - 手紙」のような係り受け関係がないものが混在していた。今後は係り受け関係の有無についても分析する必要があるだろう。

謝辞

本研究の表1に掲載されたInohara & Ueno (in press) のデータは、「Proceedings of the 36th Annual Conference of the Cognitive Science

表1 本研究とInohara & Ueno (in press) の結果およびコーパスの比較

	本研究	Inohara & Ueno (in press)
係り受けコーパスの種類	名詞1-助詞-名詞2	名詞-助詞-動詞
共有条件　平均（*SD*）	0.718(0.180)	0.469(0.121)
非共有条件　平均（*SD*）	0.714(0.179)	0.450(0.113)
効果量*d*	0.10	0.47
漢字数	517	528
係り先数	2596	6612
行列の総要素数（漢字数と係り先数の積）	1342132	3491136
行列を0が占める割合	98.9%	97.8%
行列の要素の値の合計	36304	164337

Society」に採択・掲載予定である。

参考文献

[1] Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychological Review*, 111, 662-720.

[2] Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review*, 103, 56-115.

[3] 小河妙子 (2013). 教育漢字を対象とした部品(部首)を共有する漢字群の意味的類似性に関する検討. *東海学院大学紀要*, 6, 217-223.

[4] Shu, H., Chen, X., Anderson, R. C., Wu, N., & Xuan, Y. (2003). Properties of school Chinese: implications for learning to read, *Child Development*, 74, 27-47.

[5] 玉岡賀津雄 (2005). サンズイとイトヘンはどのくらい漢字の意味に影響するか. *広島大学留学生センター紀要*, 15, 11-24.

[6] Inohara, K. & Ueno, T. (accepted). Contribution of sublexical information to word meaning: An objective approach using latent semantic analysis and corpus analysis on predicates. *Proceedings of the 36th Annual Conference of the Cognitive Science Society.*

[7] Jones, G. V. (1985). Deep dyslexia, imageability, and ease of predication. *Brain and Language*, 24, 1-19.

[8] Hayashibe, Y. (2012). Nihongo kakariuke corpus (Japanese modification corpus). nwc2010.verb.th5.bz2 Retrieved from: http://hayashibe.jp/jdc/

[9] 近藤公久 & 天野成昭 (2003). NTT データベースシリーズ日本語の語彙特性 第 2 期: 三省堂, 東京.

[10] 天野成昭 & 近藤公久 (1999). NTT データベースシリーズ日本語の語彙特性 第 1 期: 三省堂, 東京.