1O2-2

word2vecを用いた発達過程における言語獲得の分析

Analysis of Word Acquisition in Development Process with word2vec

永野秀明[*1] 岡田浩之[*1] 山川宏[*2*3] 荒川直哉[*2*3]
Hideaki Nagano Hiroyuki Okada Hiroshi Yamakawa Naoya Arakawa

[*1] 玉川大学
Tamagawa University, Graduate school of Brain Science

[*2] ㈱ドワンゴ 人工知能研究所
Dwango Artificial Intelligence Laboratory

[*3] NPO 法人 全脳アーキテクチャ・イニシアティブ
The Whole Brain Architecture Initiative, a specified non-profit organization

While children acquire words in their development, the relations among words change and come to those of adults. The conceptual space analysis of relations among words for each phase of the development could help understand the word acquisition process. Word2vec, a method for natural language processing, calculates similarity degrees among words in multidimensional vector spaces. In this study, CHILDES, the child speech corpus, was analyzed with word2vec in order to investigate the similarity degree of words relative to the Japanese verb “motsu” ("have" in English). The results indicated difference in the usage of words by children and adults.

1. はじめに

人の発達過程における語の獲得は，その獲得初期においては語と指示対象を結びつけるマッピングを通じて行われる．また，語が指示するカテゴリーの推論も行われる．しかし，語の獲得時には事物全体バイアスや事物カテゴリーバイアスなどのために正しく（熟達した話者＝大人と同じように）語を学習できるとはかぎらない．さらに，語の意味はそれ単独では定まらないため，他の語との境界を調整する必要性も生じる [今井 09] ．これは語の算出の運用パターンなどに現れ，そのパターンは発達に連れて大人へと近づいていく [佐治 11]．著者らは，この変化は発達に応じて形成される概念空間およびその変化を反映していると考え，幼児の発話コーパスの分析を通じてその性質を明らかにすることを試みる．本稿では，近年注目されている word2vec を用い，日本語動詞「持つ」と類似度の高い語を算出することで，幼児と大人における概念空間の違いを分析する．

2. word2vec による幼児コーパスの分析

2.1 分析手法: word2vec

分析には，Mikolov らによる自然言語処理手法である word2vec を用いる [Mikolov 13] ． word2vec では，語を 200 次元（デフォルト値）のベクトルで表現することによって，語と語の距離を計算したり，あるベクトルと別のベクトルの差を演算したりすることが可能である．語と語の距離は，意味的な関連性や類似度を表すとされ，また演算された 2 つのベクトルの差はその 2 つの単語の関係を表すとされる．word2vec には事前にわかち書きされた学習データを読み込ませておき，その結果に基づき上述の処理を施すことが可能となる．本稿では，後述するコーパスを学習データとして用い，その後，日本語動詞「持つ」と類似度の高い語を計算する．

連絡先：永野秀明，玉川大学大学院脳科学研究科，
東京都町田市玉川学園 6-1-1，Tel: 042-739-8326，
E-mail: ngnhi4re@engs.tamagawa.ac.jp

2.2 発話データベース CHILDES

分析対象のコーパスには，幼児の自然発話データである Child Language Data Exchange System (CHILDES) を用いる [MacWhinney 00]．CHILDES には日本語を母語とする幼児のデータもあり，本稿ではそのデータの一部を分析対象とした [Oshima-Takane 98]．また，比較対象には Wikipedia の記事データの一部を用いた．

3. 結果

計算結果を表 1 に示す．幼児における類似度の高い語を左列に，Wikipedia における類似度の高い語を右列に，類似度の高い順に示す．幼児においては，5 番目の「ひろい」を除いて手を用いる動作動詞の類似度が高い結果となっていた．手を用いるということは必然的に身体性の高い語であるといえ，幼児の言語獲得には身体性が重要であることを示唆している．一方，Wikipedia のほうでは，活用が変化した語も類似度の高い語のなかに現れたことや，動作動詞というよりは主語と目的語の関係性や状態動詞が現れたことなどの点が大きく異なっていた．ただ，Wikipedia のデータ自体は自然な発話ではないため，幼児の発話の比較対象である熟達話者のデータと一概に比較できないが，それでもなお同じ動詞に対して幼児とこれほどの違いが現れたことから，word2vec を用いたさらなる分析を行うことで有用な知見が得られると期待できる．

表 1 「持つ」と類似度の高い語

	CHILDES (2-3 years old)	Wikipedia (1/500 of all data)
1	のぼる	持ち／持っ
2	つける	作り出す／作り出し
3	おす	凌ぐ
4	ぬる	施す
5	ひろい	及ぼす
6	はる	もたらす
7	のせる	生み出す

4. おわりに

本稿では，幼児の発話コーパスを対象に word2vec を用いて日本語動詞「持つ」と類似度の高い語を算出した．その結果，幼児においては手を使うなど身体性の高い動作動詞の類似度が高かったのに対し，Wikipedia では状態や関係性を説明する類の動詞の類似度が高かった．今後は，Wikipedia ではなく実際の大人の発話コーパスを対象に分析を行う．

参考文献

[今井 09] 今井むつみ, ことばの意味を類似語の対比とカテゴリーの境界から探る, 2009 年

[Mikolov 13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeff Dean, Distributed representations of words and phrases and their compositionality, 2013 年

[MacWhinney 00] B. MacWhinney,: The CHILDES Project: Tools for analyzing talk. 3rd ed. Vol.2. Tha Database. Mahwah, N.J.:LEA. 2000 年

[Oshima-Takane 98] Oshima-Takane, Y., MacWhinney, B., Sirai, H., Miyata, S., Naka,N. (eds.) CHILDES for Japanese. Second Edition. Tha JCHAT Project Nagoya: Chukyo University. 1998 年

[佐治 11] 佐治伸郎: 母語及び第二言語の習得における語意の再編成過程に関する研究, 慶應義塾大学博士論文, 2011 年.