21

# Finding picture books suitable for a child

## Graph-based similar picture book search with weighted child words

***Abstract*—** Reading picture books to children attracts the attention of educators as a way of aiding children's language and cognitive development. We built a similarity book search system that we can easily find suitable books for our children from the many published picture books. We developed a morphological analyzer that is robust for hiragana and katakana words that are difficult to analyze correctly, and this made it possible to analyze picture books with great precision. Furthermore, taking the words in each stage of an infant's development into consideration, we devised a high-speed similarity search technique by using a graph search algorithm. In future, we expect this approach to be applied to a system for recommending picture books and an educational program to assist children's language development, and a supporting system capable of creating new picture books for each child.

・Reading picture books to children

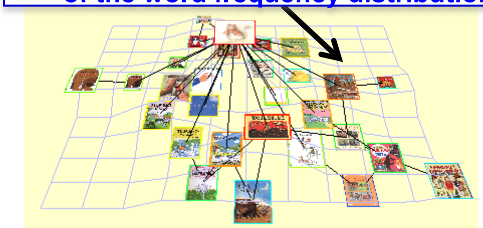
・Good influence for language and cognitive developments

・**Difficult** to select suitable picture books

↓

・A similar picture-book search system for picture-book selection

**A graph is created based on the nearness of the word frequency distribution.**



Users can grasp the relevance of books intuitively.

**Point 1** A morphological analyzer which is robust for *hiragana* and *katakana* words

・Automatic learning of the relations between *hiragana/katakana* and the original *Kanji* characters using large Japanese dictionaries

INPUT（ from picture-book text ）

ぼくらの　もりに　はるが　きたよ　(“Spring has come in our woods.”)

（From R. Kodemari and K. Takasu, “Hajimete-no-mori”、Kinnnohoshi Publishing, 2010）

A conventional morphological analyzer (many errors)

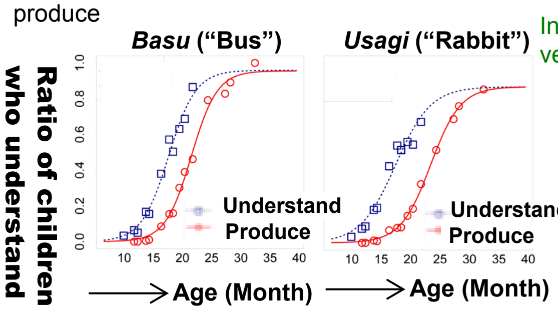
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| ぼく | ら | の | もり | に | はる | が | きた | よ |
| N | SUF | N | V | PRT | V | PRT | N | PRT |
| ぼく | ら | の | もる | に | はる | が | きた | よ |
| | “Ours” | | “piles” | | “sticks” | | “North” | |

Our system (analyzing correctly)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| ぼくら | の | もり | に | はる | が | き | た | よ |
| N | PRT | N | PRT | N | PRT | V | AUXV | PRT |
| 僕ら | の | 森 | に | 春 | が | 来る | た | よ |
| “Our” | | “woods “ | | “spring” | | “come” | PAST | |

**Point 2** Utilizing the result of infant's lexical development research

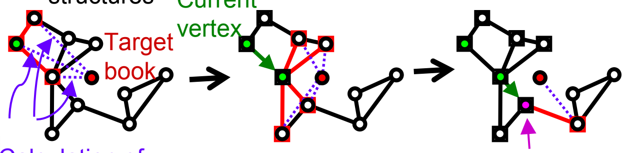
・When and which word infants understand or produce



**Point 3** A similarity search algorithm using graph indexing

・Six or more times faster than a leading conventional method ("hash method")

・Intuitive understanding of search results via graph structures



Calculation of similarities with the target book

A similar book is chosen(➔) and extracted (●)

The most similar vertex is discovered.

### Related works

[1] H. Taira, S. Fujita, T. Kobayashi, “Distribution of the vocabulary which appears in picture-book texts,” in *Proc. JSBS*, p. 92, 2012 (in Japanese).

[2] H. Taira, S. Fujita, T. Kobayashi, “Analysis of the high frequency vocabulary in picture-book texts,” in *Proc. H24 IPSJ Kansai Meeting*, F103, 2012 (in Japanese).

[3] K. Aoyama, K. Saito, H. Sawada, N. Ueda, “Fast approximate similarity search based on degree-reduced neighborhood graphs,” in *Proc. 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD),* pp. 1055-1063, 2011.

### Contact

**Hirotoshi Taira** Linguistic Intelligence Research Group, Innovative Communication Laboratory
E-mail : taira.hirotoshi{at}lab.ntt.co.jp (Please replace {at} with @)