

幼児発話からの生産的統語パターンの獲得

統語発達の計算理論に向けて

吉川 正人 (慶應義塾大学大学院/学振)

補足資料

1 パターンの生産性の計算

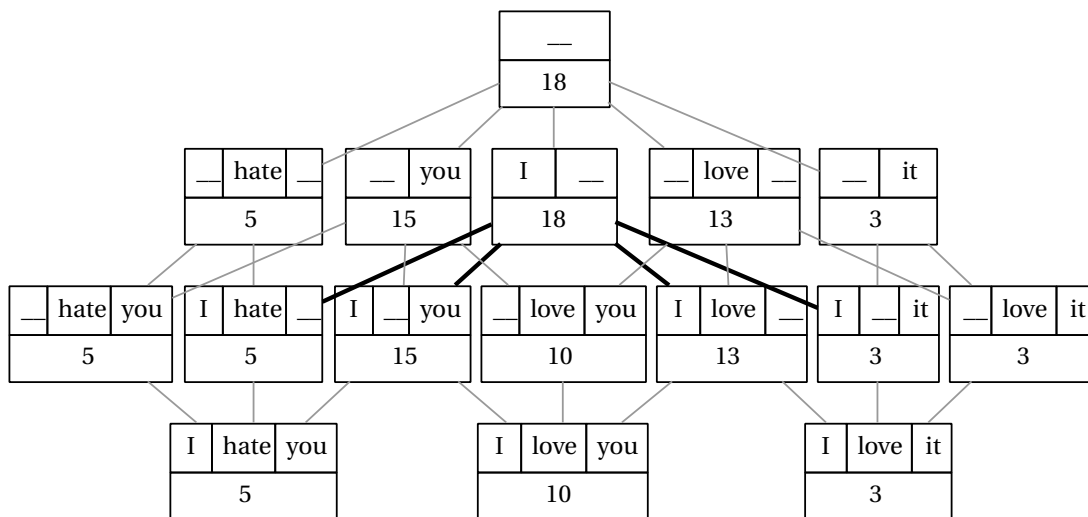
- パターンのバリエーションの豊かさをシャノンエントロピー (H) を応用して算定
- パターン p の生産性 $H(p)$: p の頻度を $f(p)$, p の n 個の直下パターンのうち i 番目のものを q_i として

$$H(p) = - \sum_{i=1}^n \frac{f(q_i)}{f(p)} \log_2 \frac{f(q_i)}{f(p)} \quad (1)$$

⇒ e.g., {*I love you* (10 回), *I hate you* (5 回), *I love it* (3 回)} の場合のパターン [$I _$]:

$$H([I_]) = - \left(\left(\frac{5}{18} \log_2 \frac{5}{18} \right) + \left(\frac{15}{18} \log_2 \frac{15}{18} \right) + \left(\frac{13}{18} \log_2 \frac{13}{18} \right) + \left(\frac{3}{18} \log_2 \frac{3}{18} \right) \right) = -(-1.502427\dots) \approx 1.5 \quad (2)$$

(下図参照)



2 発表資料及びデータのダウンロード

- 以下の URL から本発表のスライド (PDF), 本補足資料及び解析結果データをダウンロードできるようにしておきました:

http://www.yoshikawacademia.com/beyond_sent

- zip ファイル: それぞれの幼児のデータ (ファイル末尾の数値 = ファイルの id)。
⇒ 中のファイル: タブ区切り形式のテキスト; 以下のような構造:

id | ファイルの id | ファイル中の発話の id | 発話の形式 | 発話長 | 発話を構成するパターン数 | 構成パターンのリスト | 生産性の合計

- “_pops.txt”: 分析 [1] で分析対象とした発話形式を含む生起ファイル数の多かった発話の解析結果
⇒ タブ区切り形式; 以下のような構造:

id | ファイル中の発話の id | ファイルの id | 発話の形式 | 発話を構成するパターン数 | 構成パターンのリスト | 生産性の合計