単語に内在する情報量の偏在

田中久美子
東京大学情報理工学系研究科
kumiko@i.u-tokyo.ac.jp

概 要

人間の単語認知について、認知言語学の分野では認知実験を通して二つのことが指摘されている; 1. bathtub 則 : 人は単語の接頭部を接尾部よりも記憶しており、中央部の記憶が最も悪い; 2. 母音より子音を記憶している。本稿では、言語データにおいて「人間は曖昧性の小さい部位を記憶する」との仮説を立て、コーパスにおいて単語の部位別の曖昧性を計測する。そして、この仮説から逆に人間の認知についていえることを探る。

1　はじめに

単語の綴りや音が部分的に誤っていても、人は正しい単語を推測して読み進めることができることが知られている。たとえば、英語「raed tihs wtih quit amzaing esae」は、すべての単語の綴りが間違っているにもかかわらず、英語として推測しながら読むことができる。ここから、単語は文字や音素といった要素の前からの逐次処理により認識されるのではなく、不可分な単体として認識されている可能性があることが指摘されている [12]。

単語の認知について、主に英語の認知実験を通し、単語内の部位によって記憶に偏りがあることが知られており、それは大きく二つに分けることができる。第一は、場所による差異であり、人間は単語の接頭部を最もよく覚えており、つぎに接尾部を覚えており、中央部は最も記憶が曖昧であるという偏りで、Aitchson はこれを bathtub 則と名付けている [1]。たとえば、英字新聞上の綴り誤りのほとんどは、中央部のものであるという。

第二は、子音が母音よりも記憶されているというものである。Nespor らは、母音は韻律と統語情報を担うのに対し、語彙情報は子音が担っていることを指摘している [9]。また、間違った音列を無理矢理単語として聞き取る実験を行うと、子音は保存して母音を変化させることが報告されている [11][4]。子音に重点を置いて単語を識別する傾向はわずか 20 箇月の子供でも見られる [8] のに対し、チンパンジーは子音は発声しないことも知られており [10]、子音は人間の言語活動を特徴付ける要素でもある。

以上は、人間を対象とした認知実験の上で得られている結果であるが、そもそも人間は大量の言語データを浴びて言語を用いるようになる。とすると、上のような記憶にまつわる性質は、そもそも言語データに内在する性質がその原因の一つであってもおかしくはない。むろん、言語データの特徴が認知に何ら影響はない可能性もある。しかし、影響があるなら、なるべく「単語を思い出しやすくする」ところを記憶するのが効率は高いであろう[1]。そこで、本稿では「人間は曖昧性の小さい部位を記憶する」との仮説を立て、単語の異なる部位や子音・母音のもつ情報量を計測し、そこから得られる知見を吟味する。

本稿の報告は、人間の言語活動や言語に内在する性質を知る上での、認知研究と対を成す計量言語学的研究として位置付けることができる。また、単語内の各部位からの単語の予測しやすさを明らかにすることは、工学的にも種々の言語インターフェースや、言語データの圧縮といった分野への知見となりうるであろう。

2　単語推定の複雑さ

上で述べた単語の記憶についての二つの性質は、いずれも単語部分から単語全体を推定することに関する。そこで、本稿では、単語の部位から全体を推定するしやすさを計測する。単語 $w \in W$ をその部分 $c \in C$ から推定する平均的な複雑さは条件付きエントロピーで

[1] 古くより、「言語の効率」については、Martinet といった言語学者が唱えている [7]。

表 1: データ

コーパス・辞書	BNC と CELEX
コーパスサイズ	574MB
単語数	58197
子音数/母音数	31/24
子音の平均割合	63.8%

見積もることができる。

$$H(W|C) = -\sum_{w,c} P(c)P(w|c)logP(w|c) \qquad (1)$$

一般に、$H(W|C)$ が大きければ、部位 C から単語 W を予測する平均的な複雑さは大きい。とすると、前節で述べた「曖昧性が小さいところを記憶する」という仮説は、人間は H の値が小さいところを記憶する、と言い換えることができる。すなわち、この仮説が成り立つなら、過去の認知実験の結果から、H の値は単語の接頭部の方が他の部位よりも小さくなければならないし、母音よりも子音の方が小さくなければならない。

本稿の範囲では、確率の推定については第一次近似として単純な頻度確率を用いる。

3 データ

§1 で報告されている結果は、アルファベットを用いる英語、オランダ語、スペイン語における認知実験に基づくものであり、アルファベットで記述しない言語での実験はあまりみられない。この意味で、人間の単語の記憶についての普遍的性質はまだ議論の余地が大きく残されている。そもそも、英語は、活用が単語の語尾においてのみ起きる strongly suffixing 言語である [5] 特徴を持つため、そのような言語上の特性が記憶に影響を及ぼしている可能性もある。以上さまざまに議論はあるが、認知実験結果が最も多いのは英語であることもあり、本稿でもまずは英語での結果を報告する。英語においても、綴りと音素の二つの実験の可能性があり、認知上の差異に関する報告や [2]、綴りと音素の差異を考慮した認知モデルも提案されている [3]。しかし、本稿では子音・母音に関する実験が眼中にあり、子音・母音は本質的に音素上のものであることや、以降報告する情報量の偏在の観点からは綴りと音素の二つの間に特段の差はみられなかったため、本稿では、音素列における単語の偏りのみ報告する。

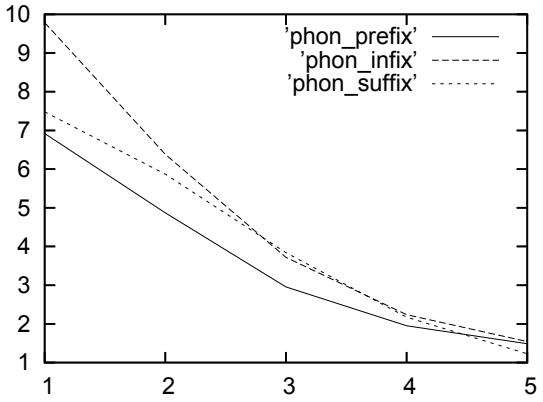

図 1: 接頭部、中央部、接尾部の情報量の差異

用いるデータは、表 1 に記載した。英語は、BNC を用いて単語頻度を計測する。活用の影響も興味の範囲であるので、単語は正規化せずに用いる。これら単語を CELEX データベース[2]を用いて音素列に変換した。たとえば、英単語 clever は-を境界として音素列 k-l-E-v-@-r*に変換される。さらに、子音・母音の影響も見るため、全音素を子音と母音にタグ付けし、単語中のどの音素が子音あるいは母音かがわかるようにした。

表 1 の最後の 2 行には子音・母音の絶対数と、単語内の子音割合が記載されている。子音・母音の数は、二重母音をどのように扱うかなどにも依存して変化するが、英語の場合には、CELEX 上の規定をそのまま用いた。

4 英単語内の情報量の偏在

4.1 部位による偏在

まず図 1 には、単語を接頭部、接尾部、中央部の三クラスに分けた際の計測結果が示されている。部位の長さ n を横軸として、長さ n の接頭部、接尾部、中央部から、単語全体を推測する複雑さが縦軸にプロットされている。たとえば、英語文字の clever の音素列 k-l-E-v-@-r*を推定するのに、n=2 のときには、接頭部 k-l から k-l-E-v-@-r*を推定する複雑さ、接尾部@-r* から k-l-E-v-@-r*を推定する複雑さ、中央部 l-E, E-v, v-@から k-l-E-v-@-r*を推定する複雑さが計測される。これを全単語について考え、式 (1) により部位別の複

[2]http://www.ru.nl/celex/

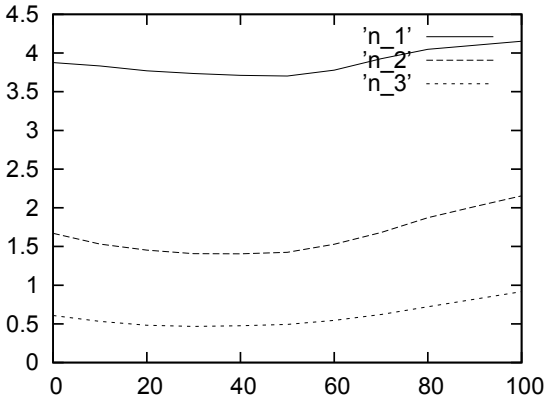

図 2: 単語内の場所における情報量

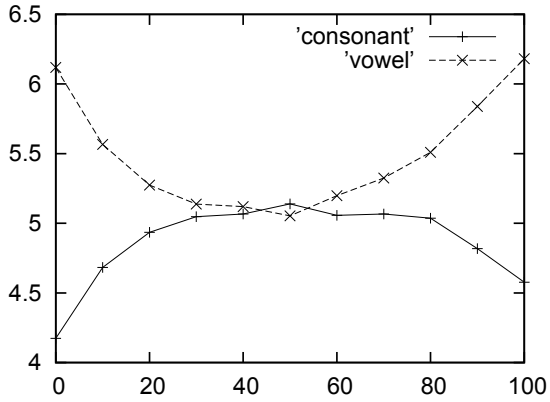

図 3: 単語内の場所における子音・母音の持つ情報量

雑さを算出する[3]。

グラフには 3 本の線があり、それぞれ接頭部、接尾部、中央部に対応する。中央部のプロットが高い位置にあることから、中央部の曖昧性は大きいことがわかる。その原因は、各単語に対して接頭部と接尾部は一つしかないのに対し、中央部は複数存在することにある。中央部の数が多い効果は、特に n が小さいときに顕著であるが、n が大きいときには、複雑さの大きさの順序が入れ替わっており、端であれば曖昧性が小さいとは限らないことを示唆する。

端の曖昧性が必ずしも小さいわけではないことは単語内の位置に対して曖昧性を平均した図 2 に現れている。この図は、横軸は単語内の相対的位置、縦軸に条件付きエントロピーを示している。グラフには線が 3 本あり、上からそれぞれ n=1,2,3 の場合に対応する。たとえば、clever の音素列 k-l-E-v-@-r*を推定する際、n=1 ならば、単語接頭要素 k の位置がグラフの左端に相当し、単語接尾要素 r*の位置がグラフの右端に相当する。中央部の l,E,v,@は、それぞれ横軸の 20,40,60,80%点に位置する。n=2 のときには、k-l が左端、@-r*が右端、l-E, E-v, v-@がそれぞれ 25,50,75%点に位置する。単語長はさまざまに異なるため、まず長さ別に単語の複雑さを計測し、各単語長につき全単語に対する割合の重みでエントロピーを足し込むことで、このグラフを得ている[4]。

プロットは、いずれも中央部下に凸である。これは、部位の正確な位置を考慮して複雑さを計測すると、実は中央部の曖昧性が最も小さいことを示唆し、中央部の複雑さが大きかった図 1 とは、逆の結果となっている。すなわち、英語のクロスワードにおいて、単語中央部分 3 文字がわかっている場合は、単語の端 3 文字がわかっている場合と比べて、データ上は単語を推定しやすいことを意味する。このようになる原因として、単語端は活用したり、定型的な形態素を足したりすることから一定のパターンが現れやすく、単語全体を推定する助けにあまりなっていないことが原因と考えられる。

4.2 子音・母音による偏在

まず、子音のみ、母音のみから単語の推定がどの程度曖昧であるのかを計測した。これは、英単語 clever の音素列 k-l-E-v-@-r*を推定する際、子音 k-l-v から推定するのと、母音 E-@-r*から推定する場合を考えることに相当する。全単語に関して、子音からの推定の複雑さは 0.8881、対する母音は 4.3560 であった。すなわち、英語では、単語内の子音列がわかると単語はかなり絞り込める一方で、母音は単語を特定するのにあまり役には立たない。

図 3 は、n=1 のときの単語内の位置別の子音と母音の複雑さを表している。横軸は前図同様に単語内の相対位置であり、縦軸は条件付きエントロピーである。二本の曲線は、それぞれ子音と母音に対応する。子音の曲線を得るには単語内の全母音をワイルドカードとして、各部位の複雑さを計測しており、母音の曲線を得るには、全子音をワイルドカードとして同様の処理を行った。たとえば、?をワイルドカードとすると、k-l-E-v-@-r* を推定する際、子音 k-l-?-v-?-?から k-l-E-v-@-r*

[3] この実験では、n の計測値を得るのに長さ n+2 以上の単語を用いた。

[4] このプロットを得るのに用いたのは長さ n 以上の単語で、ある長さの総頻度が一定値 (=100) 以上となるものである。このため、n=1 の場合の左端、右端は、前図の接頭部、接尾部の左端とは異なる値を示している。

を推定する複雑さと母音?-?-E-?-@-r*から推定する複雑さを図 2 同様の方法で部位ごとに測定する。

子音と母音のプロットが上下対称になっていることから、子音と母音は互いに補完的な関係にあることがわかる。さらに、子音曲線が母音曲線よりも下部にあり、中央部のわずかな部分を除いて、単語のほぼ全位置において子音の方が母音よりも曖昧性が小さいことがわかる。また、子音のプロットが、接頭と接尾部において小さくなっていることが伺える。

5 議論

「人間は曖昧性が小さい部位を記憶する」との仮説の下で、以上の結果を吟味する。まず、部位別による偏在についてであるが、二つの異なる計測方法—クラス別と位置別—を用いて実験を行い、単語端の曖昧性について反対の結果を得ている。仮説と bathtub 則を整合させるには、二つの計測方法のうちクラス別を採用しなければならない。すなわち、人は単語の中央部の部分列の位置を正確に記憶しているわけではなく、中央部としてひとまとまりとして認知している可能性がある。単語は、音素を前から逐次処理して認知されるわけではないとの説を、本稿の冒頭で紹介したが、本実験は少なくとも中央部についてはひとまとまりの単位として認知している有り様を示唆する。

子音・母音の偏りに関しては、計測結果は、子音が母音に比べて複雑さが小さいことを示しており、仮説の下で認知実験上の結果と整合する。また、図 3 において子音の偏りが単語端においてより低くなっていることから、そもそも bathtub 則の一つの原因が子音にある可能性も示唆される。すなわち、曖昧性の小さい子音を重点的に記憶し、子音の曖昧性が英単語の両端で小さいことを原因として、bathtub 則が成り立っている可能性がある。

6 結語

単語は人間の言語の上で中心的役割を成す記号である。本稿のねらいは、それがどのように認知されるのかの一端を言語データの観点から捉えようとする点にある。単語を思い出す上での効率を鑑みて、「人間は曖昧性の小さい部分を記憶する」との仮説の下、認知実験の報告を手がかりとして、単語内の情報量の偏りを調べた。結果、部位的にも子音の上でも単語の端の情報量が小さくなっており、単語の「輪郭」を重点的に認知していることが示唆された。

記号の輪郭と人間の記憶の関係については、他にもいくつかおもしろい報告がある。Lua は中国人に漢字の部分を見せて漢字を当ててもらう実験をした。漢字の中央部を隠す場合と外側を隠す場合では、中央部を隠す場合の方が正解率が高かったことを報告している [6]。また、芥川は、異なる楽器で同じ音を演奏した録音があるとき、音の鳴り始めと鳴り終わり部分の録音を切り落としてしまうと、何の楽器でその音が演奏されているのか、専門家でも当てることができなくなることを紹介している [13]。すなわち、人は記号を認知する上でその「輪郭」部から大きな情報を得ていることを示唆する。人間において根源的な「記号」がどのように形成され、認知されるのかを明らかにするには、認知実験と計算言語学の両方の分野からさらなる研究が必要である。

参考文献

[1] J Aitchison. *Words in the Mind*. Blackwell, 1994.

[2] A. Buchwald and B. Rapp. Consonants and vowels in orthographic representations. *Cognitive Neuropsychology*, 23(2), 2006.

[3] M. Coltheart, K. Rastle, C. Perry, R. Langdon, and J. Ziegler. DRC : a dual-route cascaded model of visual word recognition and reading aloud. *Psychol. Rev.*, pages 204–256, 2001.

[4] A. Cutler, N. Sebastian-Galles, O. Soler-Vilageliu, and B. van Ooijen. Constraints of vowels and consonants on lexical selection: Cross-linguistic comparisons. *Memory & Cognition*, pages 746–755, 2000.

[5] M. Haspelmath, M. S. Dryer, D. Gil, and B. Comrie, editors. *The World atlas of Language Structures*. Oxford University Press, 2005.

[6] K.T. Lua. Human recognition of chinese characters. *Computer Processing of Chinese and Oriental Languages*, 6(1):75–84, 1992.

[7] A. Martinet. *Éléments de Linguistique Générale*. Colin, 1960.

[8] T. Nazzi and B. New. Beyond stop consonants : Consonantal specificity in early lexical acquisition. *Cognitive Development*, 22(2):271–279, 2007.

[9] M. Nespor, M. Peña, and J. Mehler. On the different roles of vowels and consonants in speech processing and language acquisition. *Lingue e Linguaggio*, pages 221–247, 2003.

[10] S. Savage-Rumbaugh and R. Lewin. *The Ape at the Brink of the Human Mind*. Wiley, 1996.

[11] D.J. Sharp, S.K. Scott, A. Cutler, and R.J.S. Wise. Lexical retrieval constrained by sound structure: The role of the left inferior frontal gyrus. *Brain and Language*, pages 309–319, 2005.

[12] D.D. Wheeler. Processes in word recognition. *Cognitive Psychology*, 1, 1970.

[13] 芥川他寸志. **音楽の基礎**. 岩波新書, 1971. 15 ページ.