人文科学とコンピュータ 56－2
（２００２．１１．１５）

確率的言語モデルに基づくフランス語の使用例の調査
－主語人称代名詞の"on"と"l'on"を例に－

[*]赤間啓之　[*]清水正勝　[**]清水由美子

概要：意味的に等価なフランス語の主語人称代名詞である"on"と"l'on"の使い分けについては、「隣接する単語の音による」という Vaugelas の基準が広く流布したものになっているが、これは絶対的なものではない。本調査では、フランス語の新聞 Le Monde(2000)の電子的コーパスから n-gram を抽出し、人工知能エンジン C5.0 によりその決定プロセスを明らかにすることを試みた。結果、この新聞の書き手は、Vaugelas の示した基準をほぼ忠実に守っていることが明らかとなった。

A research of French usage, based on Probabilistic Language Model
－About the use of impersonal subject pronoun, "on" and "l'on"－

[*]Hiroyuki AKAMA, [*]Masakatsu SHIMIZU, [**]Yumiko SHIMIZU

Abstract: This research aims at examining the uses of "on" and "l'on" which are semantically equivalent as French impersonal subject pronoun. The great grammarian of the 17th century, Claude Favre de Vaugelas, proposed some "euphonic rules" about the discrimination between them. We extracted from *Le Monde*, overall instances of "on" and "l'on" with the preceding and following words, to simulate by the algorithm called C5.0 the hidden rules governing the discrimination process. The result generated by this computation showed the profound influence of Vaugelas upon the writers of this newspaper.

1.はじめに

1-1.研究目的

フランス語の主語人称代名詞"on"と"l'on"は、英語における主語としての one の用法とも重なるが、現代フランス語において、文法上は三人称単数の主語であり、意味的には、ともに非人称の主語として漠然と「人」、「我々」を指す。たとえば、シャンソンの歌詞で有名な、« mais il y a peu de chance qu'*on* détrône le roi des *cons* » （だが、馬鹿者たちの王様から人が王位を剥奪できる可能性はほとんどない）という文は、 « mais il y a peu de chance que *l'on* détrône le roi des cons »とも言いかえても、違いは kon の音の反復が失われるという音調の違いでしかない。"on"と

[*]東京工業大学大学院　社会理工学研究科
[**]武蔵工業大学　環境情報学部
[*] Graduate School of Decision Science and Technology, Tokyo Institute of Technology
[**] Faculty of Environmental and Information Studies, Musashi Institute of Technology

"l'on"は意味の上ではあくまで等価であるが、従来の文法書には、両者を使い分けるためのさまざまな基準が提示されている。しかしそれらは使用場面において決定的に作用するものではない。両者の実際上の選択については、その前後の単語列が含む特徴から、一定の条件付き確率に従って判別可能であると予想される。ただし、それが一般の文法書が記述する基準に合致する保証はない。また Bergen(2002)が、フランス語のリエゾンという発音現象で示したとおり、前後の単語列以外のファクター、とくに話者、書き手の社会言語学的属性が複雑に関与する「変数」としてとらえられる可能性もある。

近年、コンピュータの性能が向上したことにより、大量のテキストデータを処理することが容易になってきた。そこで、文法書のように、一部の作例によって全てを語るというスタンスではなく、実際のコーパスデータから、その使い分けの傾向を明らかにすることが重要だと考える。そして Bergen が上記の発音現象を、条件付き確率の連鎖によるベイジアンネットワークでシミュレーションしたように、さまざまなデータマイニングの手法を認知言語学、統計言語学の領域に導入することが可能である。本論では、新聞コーパス内の全ての"on"と"l'on"を共起語と共に抽出した n-gram データを利用し、両者の使い分けがどのような因果関係によって為されているのかを、人工知能エンジン C5.0 による確率的言語モデルを利用して明らかにするものである。

1-2.先行研究に見る使い分けの基準

"on"と"l'on"という二種類の人称代名詞は、Robert(1985)によれば、「17 世紀の終わり頃までは、定冠詞のついた"l'on"は一般に人間を指示する場合に使用され」ることになっており、また Grevisse(1988)には「古代の書き言葉において、"on"が名詞としての機能を保持している場合には、"on"のより高貴な代用語として"l'on"が使用される」との記述がある。

しかし、現在のフランス語においては、両者は上記のような「意味上」の相違によってではなく、「好音調(euphonie)」という基準によって使い分けられており、この「好音調」に基づく両者の使い分けを提示したのが古典主義の文法家 Vaugelas(1585-1650)であるとされている。

Vaugelas においてキーワードとなるのは、「母音衝突(hiatus)」「不快音調(cacophonie)」「好音調(euphonie)」「言いやすさ・聞きやすさ・書きやすさ・読みやすさ」であり、それらのほとんどは、後の文法書にも踏襲されている。以下、その基準を列挙する。

- et, ou, où, qui, que, quoi, si(aussi), (時に lorsque)の後では l'on が使用される。(母音衝突の回避)（Aristide, 1963・Dupré,1972）
- 「無音の e」以外の母音的発音の直後では、l'on が使用される。(母音衝突の回避)(Vaugelas, 1924)
- 母音衝突が発生している場合であっても、直後の単語が"l"で始まっている場合には、on が使用される。(不快音調の回避)（Vaugelas, 1924・Girodet, 1980）
- qu'on という形は、直後の単語が、"com"あるいは"con"で始まっている場合には使用せず、que l'on を用いる。(不快音調の回避)（Vaugelas,1924・Littré,1885）
- 一文の中で que の連続を避けるために、qu'on を用いる。(不快音調の回避)　また、que が複数出現する場合には、音節数を調整するために qu'on と que l'on を使い分ける。(「言いやすさ・聞きやすさ・書きやすさ・読みやすさ」)(Vaugelas,1924)

- 上記の基準に当てはまらない場合でも、文頭に l'on が使用されることがある。これは「擬古典主義的」あるいは「気取り」のためである。(Dupré, 1972・Girodet, 1980)

しかしこれらの基準は「強制されるものではなく」(Aristide, 1963)、どちらを使用したとしても、文法的に誤っているわけではない。したがって「両者を使い分けるための判断は、まったく使用者の耳に委ねられ」(Littré, 1885)ていることになる。

1-3.先行研究をふまえて

先行研究に見られるような、使い分けの基準は強制的・絶対的なものでないことは上で見たとおりである。つまり実際の使用例を精査すれば、それは確率的な分布を呈するものになることが予想されるのである。今回の我々の調査においては、主に「母音衝突」・「不快音調」・「好音調」に注目し、"on"と"l'on"の前後に出現する語にこれらを変数として割り当て、新聞での使用場面において、それらの変数がどのような因果関係（抑制・強調関係）にあるのかを、調査した。

今回の調査には、Le Monde(2000)の 1 月から 6 月分までを使用した。結果は従来の文法書で提示されている基準を、ほぼ踏襲したような「規範的な」使い分けが為されていることが明らかになった。

2.調査方法

2-1.n-gram の抽出

先行研究の節でも見たように、今回我々が調査しようとしている"on"と"l'on"に関しては、その直前の単語に依存するのみでなく、直後の単語との関係も見なければならない。そのため、n-gram データを抽出するスクリプトを用いて、"on"あるいは"l'on"を中心語とした、前後 2 単語ずつの計 5 単語を出現パターンとしてコーパスデータより収集した。

なお、コーパスデータとして使用したのは、フランスの新聞 Le Monde の 2000 年 1 月から 6 月までの計半年分(総単語数約 1,619 万語)であり、収集のために使用した Perl5.0 のスクリプトにより、"on"あるいは"l'on"を中心語とした全ての出現パターン(延べ 34,607 パターン)が集められた。

2-2.変数の設定

前節で集められた 5gram の単語列に対し、以下の条件に一致するものにそれぞれ数字をあてはめた。我々が当初設定した変数は以下の通りである。

- **変数 1**(句切れをまたぐ cacophonie に関する変数)：**{(1-a)**：直前の単語が、"l"で始まっていれるあるいは、直前の単語が、","・":"・";"であり、その前の単語が"l"で始まっている→**1**、**(1-b)**：それ以外→**0}**

- **変数 2**(直前の単語に関する変数)：**{(2-a)**："・":"・";"である→**0**、**(2-b)**："!"・"?"・"."である→**1**、**(2-c)**："et"・"maintenant"・"si"・"soit"・"sinon"・"ou"・"où"・"qui"・"quoi"である→**2**、**(2-d)**："dont"・"quand"である→**3**、**(2-e)**："～que"である→**4**、**(2-f)**："comme"である→**5**、**(2-g)**："～quel"・"～quelle"・"～quels"・"～quelles"である→**6**、**(2-h)**："tant"・"～ment"である→**7**、**(2-i)**："or"・"car"・"alors"・"d'ailleurs"である→**8**、**(2-j)**："mais"・"plus"である→**9**、**(2-k)**："t-"である→**10**、**(2-l)**：それ以外→**11}**

- **変数 3**(直後の単語に関する変数)：**{(3-a)**："l"で始まっている→**1**、**(3-b)**："con"で始まっている→**2**、**(3-c)**：それ以外→**0}**

- **変数 4**("on"か"l'on"かに関する変数)：**{(4-a)**："l'on"である→**1**、**(4-b)**："on"である→**0}**

2-3.変数の説明

なぜ前節のような条件を変数に挙げたのかを概観しておく。

(1-a)：句切れを越えての cacophonie が"on"か"l'on"（以下、中心語）かの選択に影響があるのかどうかをみるためのものである。

(2-a)(2-b)：句切れや文頭であることと中心語の選択の関係をみるためのものである。

(2-c)：直前の単語が母音で終わっており、後続の単語との間にリエゾン・エリジオン（母音衝突による縮約）が発生しない場合と、中心語の選択の関係をみるためのものである。

(2-d)：これらの単語は、後続する単語が母音で始まっている場合、"t"音でのリエゾンが好まれるとされているものである。リエゾンを好まないのであれば、"l'on"が選択されることになる。

(2-e)：(2-c)とは異なり、この場合にはエリジオンが発生し"qu'on"となることがある。エリジオンを好むか、"que l'on"とし、それを避けるのかをみるためのものである。

(2-f)：発音する際には、アンシェヌマンが起きる場合である。アンシェヌマンを避けるのであれば、"l'on"が選択されることになる。

(2-g)：これらの単語は、発音する際には"l"音で終わる。中心語に"l'on"を選択した場合、「発音上」cacophonie が発生することになる。

(2-h)：これらの単語のあとでは、"on"はリエゾンしないと考えられている。

(2-i)：これらの単語は、"r"音で終わる。書き手が"r"音と"l"音の連続を cacophonie と捉えるか否かをみるためのものである。

(2-j)：(2-h)と同様の理由である。

(2-k)：主語と動詞が倒置されている際の、中心語の選択をみるためのものである。

(3-a)：中心語が"l'on"である場合、cacophonie が発生してしまう。

(3-b)：中心語が"qu'on"である場合、cacophonie が発生してしまう。

これらの変数は、大きく分けて 2 種類に分類される。まず、書き手がリエゾン・エリジオン（母音衝突の際の縮約）・アンシェヌマンといった、母音字と密接な関係にあるフランス語の言語現象を好むか否かというものである。好むのであれば、"on"が率先して使用されるであろうし、好まないのであれば"l'on"が使用されることになるであろう。他方、cacophonie を好むか否かというものである。

そして、母音衝突と cacophonie が同時に発生した場合には、どちらを「より避けたがる」のかを見ることもできると考える。

2-4.決定木を生成するアルゴリズム C5.0

人工知能エンジン C5.0 は、SPSS 社の開発した Clementine 6.0 というソフトウェアの中に収められている分析アルゴリズムであり、結果を決定木の形で表示させることができる。決定木を生成する場合に問題となるのは、どの変数をルートノードとするか、どの変数によって分岐させるかという点である。この点に関し、1986 年 Quinlan によって発表された ID3（Iterative Dichotomiser 3）モデルに用いられている利得基準、C5.0 に用いられている利得比基準について、数学的な概説をしておく。

2-4-1.利得基準

利得基準は、決定木による分岐が生じる場所である親ノードと子ノードとの間で計算される。親ノード内の観測値の集合を O とし、それらの観測値は K 個の水準を持つカテゴリカルな基準変数 CV(criterionvariable の略)によって $\{cv_1,cv_2,\cdots,cv_k,\cdots cv_K\}$ のように分割されているとする。この親ノードから任意の事例を一つ取り出す時に、それが cv_k である確率 $P(O, cv_k)$は、以下のようになる。ただし、集合αに含まれる全ての事例の数を$|\alpha|$、αに含まれる事例βの数を$|\alpha,\beta|$と表記する。

$$P(O,cv_k)=\frac{|O,cv_k|}{|O|}$$

決定木を成長させるためには、単なる確率の高低だけでなくそこに含まれている「情報量」が重要となる。この「情報量」は、底を 2 とする対数で確率を変換し、-1 を掛けた値で定義され、情報量が少ない方がより整理された情報であるとされている。

予測変数（子ノード）を考慮しない場合の、親ノードにおける平均情報量$I(CV)$は、

$$I(CV)=-1\times\sum_{k=1}^{K}P(O,cv_k)\times\log_2\left(P(O,cv_k)\right)$$

と表記することができる。さて、子ノードの候補となる観測値の集合が、L 個の水準を持つカテゴリカルな予測変数 PV(predictor variable の略)によって $\{pv_1,pv_2,\cdots,pv_l,\cdots pv_L\}$ のように分割されているとする。予測変数を考慮した場合の親ノードの平均情報量$I(CV)_{PV}$は、

$$I(CV)_{PV} = -1 \times \sum_{l=1}^{L} P(O, pv_l) \times \left[\sum_{k=1}^{K} P(pv_l, cv_k) \times \log_2 \left(P(pv_l, cv_k) \right) \right]$$

ただし、$P(O, pv_l) = \frac{|pv_l|}{|O|}$、$P(pv_l, cv_k) = \frac{|pv_l, cv_k|}{|pv_l|}$ であるとする。

である。この二式の差が利得基準 $G(CV)_{PV}$ となる。

$$G(CV)_{PV} = I(CV) - I(CV)_{PV}$$

これを候補となる全ての予測変数に関して計算し、値が最大となった予測変数で分岐を行い決定木を成長させてゆく、という方法がとられる。

2-4-2.利得比基準

前節で取り上げた利得基準は、予測変数そのものの平均情報量を考慮しないために、水準の少ない予測変数と、多い予測変数を比較すると、後者の方が親ノード（基準変数）の平均情報量を下げやすいため、結果として後者に有利な判定を下すことなり、単純な決定木を描けなくなってしまうという欠点がある。

そこで予測変数 PV の平均情報量 $I(PV)$ と利得基準 $G(CV)$ との比、すなわち利得比基準 $Gr(CV)_{PV} = \frac{G(CV)}{I(PV)}$ の大きいほうを取るというアルゴリズムが誕生し、1997 年に C4.5 の名で発表され、現在の C5.0 に至っている。

3.結果と考察

3-1.出力結果

C5.0 による出力を、左から右へ親子のノード関係が展開する決定木の形で表すと以下のとおりであった。先に多くの変数を設定したが、実際にルール形成に関与した変数とその組み合わせパターンは、表のように絞られた。しかし、次節で示すように、そこには興味深い交互作用をとらえることができた。

(2-c)：[最頻値：l'on] (該当数：3651)

(3-c)：[最頻値：l'on] (該当数：3377, 0.631) → l'on

(3-a)：[最頻値：on] (該当数：197, 0.975) → **on**

(3-b)：[最頻値：l'on] (該当数：77, 0.688) → l'on

(2-e)：[最頻値：on] (該当数：7094)

(3-c)：[最頻値：on] (該当数：6364, 0.575) → on

(3-a)：[最頻値：on] (該当数：626, 0.995) → on

(3-b)：[最頻値：l'on] (該当数：104, 0.798) → **l'on**

(2-b)：[最頻値：on] (該当数：7638, 0.998) → on

(2-l)：[最頻値： on] (該当数：16224, 0.99) → on

表１：C5.0 による出力結果

3-2.出力結果の分析・考察

我々が設定した変数のうち、変数 1 は決定的な要素ではなく、関与的なのは「直前の単語」と「直後の単語」であることが明らかとなっている。また、直前が母音の場合、母音衝突を避けるために"l'on"が使用される率が高くなる（表中の破線部）のであるが、後続する単語が"l"で始まっている場合（表中の下線部）には、"on"が使用される。つまりこの書き手は母音衝突よりも、cacophonie を「より避けたがる」傾向にあるといえる。一方、同様に母音衝突が発生している場合であっても、"～que"の後（表中の網掛け部）では、"on"を好んで使用しているのであるが、後続する単語が con で始まっている場合（表中の波線部）には、"l'on"が使用していることで明らかなように、ここでも cacophonie を優先して避けようとしていることがわかる。

Vaugelas によれば、「『発音しない e（今回の例では"～que"）』以外の全ての母音字の後には、"l'on"を使用すること、母音衝突と cacophonie が同時に発生した場合には、cacophonie を避ける」ことを奨励しており、今回の結果はほぼそれらを踏襲しているものであるということがいえる。

4.まとめと今後の展望

今回の、n-gram を収集し人工知能エンジン C5.0 により分析するというアプローチによって、調査対象とした Le Monde の書き手は、「Vaugelas が示し、今日まで善くも悪しくも続いてきた」(Dupré, 1972)とされる規範をほぼ忠実に守っているということが明らかとなった。しかし、今回の音韻的要素に基づく調査は、我々が目指す社会言語学的調査の第一段階にすぎない。なぜなら、Web 上のフランス語においては、これらの規範が守られているとは言い難い表記法が使用されているし、Vaugelas によれば、散文や詩などの文学作品においては、「音節数を調節して読みやすくするために、"qu'on"と"que l'on"を適宜使い分けるような使用法」が存在しているとのことである。また、我々の考えた条件のうち、今回調査対象としたコーパスデータでは有効なものとして働かなかったものがいくつかある。これらが無用なものであるのか、調査対象を変えれば有効なものとなるのかは、現段階では不明であり、Le Monde 編集部に使い分けのガイドラインが明文化された形で存在するか問い合わせたり、今後様々なジャンルのコーパスデータを調査したりする必要があると考える。

これからの展望としては、Bergen(2002)がフランス語のリエゾンの成否に関して、性別や出身地・年齢などの社会的要素を変数の中に取り込んで、ベイジアンネットワークを組んだように、この分野においても、そのような社会的な要素を含んだコーパスデータや、文学作品を調査対象として分析することで、調査対象別の、"on"か"l'on"かという選択に至るまでの意志決定のメカニズムの違いを明らかにしていきたいと考えている。そうすることが、我々の目指す確率的言語モデルと、社会言語学あるいは認知言語学の融合のための有効な手段であると確信している。

参考文献

[1] Aristide, *LE FIGARO LITTÉRAIRE 25 mai 1963*, Paris, Figaro, 1963

[2] Benjamin K. Bergen, *Social variability and probabilistic language processing*, International Computer Science Institute Technical Report, (To Appear)

[3] Benjamin K. Bergen, *Of sound, mind, and body: neural explanations for non-categorical phonology*, Ph.D. Dissertation. Department of Linguistics, U.C. Berkeley. (Advisor: George Lakoff), 2001

[4] Benjamin K. Bergen, *Probability in phonological generalizations: Modeling optional French final consonants*. In Alan Yu et al. (eds.), Proceedings of the 26th Annual Meeting of the Berkeley Linguistics Society. Berkeley, Berkeley Linguistics Society, 2000

[5] Jean-Paul Colin, *Nouveau dictionnaire des difficultés du français*, Paris, La Librairie Hachette et de Claude Tchou, 1971

[6] Christopher D. Manning and Hinrich Schütze, *Foundation of Statistical Natural Language Processing*, Cambridge MA and London, MIT Press, 1999

[7] Paul Dupré, *Encyclopédie du bon français dans l'usage contemporain 2*, Paris, Éditions de Trévise, 1972

[8] Jean Girodet , *Dictionnaire du bon français*, Paris, Bordas, 1980

[9] Maurice Grevisse, *Le Bon Usage*, Duclot, 1988

[10] É. Littré, *Dictionnaire de la langue française 2*, Paris, Hachette, 1885

[11] É. Littré, *Dictionnaire de la langue française 3*, Paris, Hachette, 1885

[12] Ph. Martinon, *Comment on parle en français*, Paris, Larousse, 1927

[13] J. Ross Quinlan, *C4.5 : Programs for Machine Learning*, *Morgan Kaufmann*, 1993

[14] P. Richelet, *Dictionnaire françois(1680)*, Tokyo, France Tosho Reprints, 1969

[15] ROBERT, *LE GRAND ROBERT DE LA LANGUE FRANÇAISE Tome VI DEUXIÈME ÉDITION*, Paris, LE ROBERT, 1985

[16] Claude Favre de Vaugelas, *Remarques sur la langue française*, Genève, Droz, 1924

[17] *Dictionnaire de l'Académie française Tome Premier* A-G, Paris, Hachette, 1932

[18] 北研二,『言語と計算 4　確率的言語モデル』, 東京, 東京大学出版会, 1999

[19]豊田秀樹,『金鉱を掘り当てる統計学—データマイニング入門—』, 東京, 講談社, 2001