# コーパスは基礎語彙を確定できるか？
# Est-il possible que le corpus détermine les vocabulaires fondamentaux ?

中尾　浩（愛知大学法学部）

**要旨**

筆者は数年来，フランス語のコーパス分析を通じて基礎語彙や重要語彙と呼ばれるものを確定できるかどうかに取り組んできたが，それは不可能であろうという結論に達した。コーパスが基礎語彙を確定することができないのであれば，どのようにすれば基礎語彙を確定することができるのか，あるいは不可能なのかについて，考察を加えることにする。

**キーワード：**コーパス，基礎語彙，集合知，フランス語

## 基礎語彙とコーパス

フランス語に限らないが，日本語でも英語でも，ネイティブ話者向けであろうと第2言語学習者向けであろうと，何らかの言語を習得することを容易にするために，しばしば「基礎語（彙）」や「重要語（彙）」と呼ばれる一定数の語を集めたリストが利用される。『○○語重要語3,000』のような名称で単語集としてまとめられた形態もあれば，数万語を収めた辞書の中で＊印や色つき文字で強調された数百語から数千語のような形で提示されることもある。印刷物ではなく，単にデジタルデータとしてだけ提供されている場合もある。本論においてはそれらのいずれも「語彙リスト」と称することにする。単語集であろうと辞書の記載であろうと，**何らかの意図の元に選別された一群の語**という共通点のみを指し示すためである。

これらのリストの精度を上げるために，コーパスの利用が大いに期待された。実際，コーパスを利用した語彙リストは既に様々な言語においていくつも作成されて，大いに活用されているばかりか，更なる精度を求めて鋭意研究されている。

コーパスの利用が言語研究に大いに資することは事実であろうと思われる。事実，分野によっては大きな成果を上げている。翻って，基礎語彙や重要語彙と呼ばれる語彙リストの作成についてはどうか。筆者はここ数年，コーパス，もう少し厳密にいうと，生データを形態素解析した結果に基づいたコーパス研究に取り

組んできたが，どのような方法をとっても，またどのような観点から考えてもコーパスが基礎語彙や重要語彙を確定することは不可能である，という結論に達した。これはコーパスの規模を今の10倍や100倍にしてみたところで，あるいはどれほど多種多様なジャンルからコーパスを集めたところで，原理的に不可能である。

## コーパス分析結果の比較

フランス語のコーパス分析結果は探せば意外と何種類もある。残念ながら新しい成果が少なく，過去のものについては母集団が小さなコーパスに基づいているものも少なくない。しかし，それらも含めて，実は大勢には影響がないことも明らかにするつもりである。

入力作業がまだ完全に終わっていないので，今回は私個人が作成したデータ（My Rankと表記），2009年に，Routledge社から出版された*Frequency dictionary of french*（FDFと表記），*Trésor de la Langue Française*を編纂する際に使われた大規模コーパスをEtienne Brunetが分析したもの（TLF-Bと表記）の3つを比較することにした。それぞれのデータについて，もう少し補足をしておく。

## My Rankについて

筆者が個人的に収集した，主としてフランスの新聞（かつてはフランスの新聞データの中にはWeb上で無料で公開されていたものもあった。またCD-ROM等を購入したものもある）を中心としたデータで，語数はおよそ数億語と見積もっている。このデータをTreeTagger[1]という形態素解析ソフトで分析して，TreeTaggerが解析した形態素ごとに集計した。残念ながらいささか信頼度の低いデータであり（たとえばデータ中に重複などが多数ある），優れたアプリケーションではあるが，TreeTaggerの性能の限界により分析精度に限界があり，必ずしも万全の分析結果ではないのだが，実は意外と的確な分析がなされていたことも後ほど紹介する。

## FDFについて

FDFは2,300万語のデータ分析から上位5,000位（5,000語ではない）を収録している。データの詳細を見ると，書き言葉，話し言葉，フランス本国，海外，のみならず，ジャンルに関してもある程度のバランスを考えて母集団を作っているようだ。最近のフランス語コーパス分析においては，それなりに完成度の高いデータ言えよう。

**TLF-Bについて**

TLF（*Trésor de la langue française*）を作成するために構築されたデータに基づいて作成された労作であるが，TLFのデータということからもわかるとおり，主として文学的なデータである。およそ2,500万語のデータを分析し，出現回数が500回以上の6,700語を収録している。FDFとはジャンルが異なるが，母集団の大きさが，おおむね同じである。

以上が今回利用したデータである。出現回数やランクを割り出したデータは他にもいくつかあり，現在入力中ではあるが，おそらく今回の結論を覆すことにはならないと思われる。

**データの検証（1）**

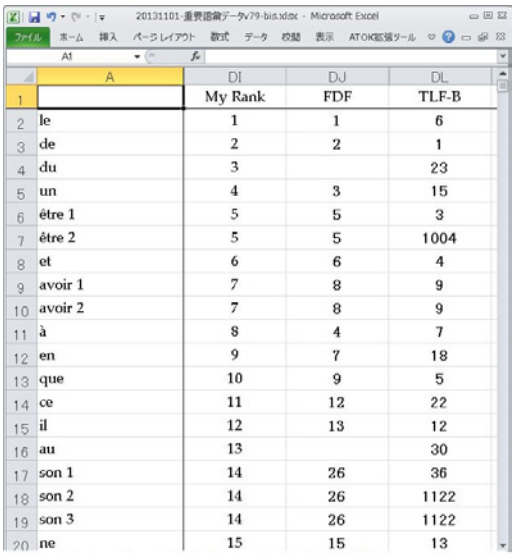
これらのデータを比較するにあたって，まずどの数値で比較するかを決める必要がある。

複数のコーパスの結果を比較する場合，比較検討する数値はいくつか考えられる。まず実際の出現数だが，これはコーパスサイズが異なると比較にならないので不向きである。次に出現率で比較することは可能だろう。コーパスサイズに寄らないので，出現率が10%であればコーパスのサイズにかかわらず10%である。しかし出現率にも難点はある。たとえば，3%という数値がそのコーパス全体においてどのような位置づけなのかがわかりづらく，結果的に異なった結果同士を比べにくくしている。たとえば30%, 20%, 10%, 5%, 3%, 2%, 1%,....というグループと，10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%,....というグループでは同じ3%の出現率でも前者は1位の10分の1にすぎないが，後者では3分の1になってしまうので，同じ3%でも全体の中での意味合いが異なってしまう。

最後に出現順位（ランク）で比較する方法がある。こちらは出現数でも出現率でも，とにかく数値の大きなものから順に並べて順位を決めて，それを比較する方法である。実出現数のランクもあれば出現率のランクもある。FDFは実出現数と分散から割り出したランクを用いている。これなら，先ほどの同じ3%でも，前者のコーパスでは5位，後者のコーパスでは8位となって，コーパス内での相対的な位置を比べやすい利点がある。ただし，ランクにも問題点はある。まず第一にランクの振り方だが，同数のものを一つと数えるか，あくまで上から順に数えるかによってランクの値は大幅に変わってくる。1,2,3,4,5,6,7という振り方をしているものと，1,2,2,4,4,4,7,....や（ExcelのRANK関数がこの振り方をする），1,2,2,3,3,3,4,....といった振り方による違いである。コーパス内に含まれる語彙の使用頻度は一般的にパレートの法則

に従っているので，下位に行けば行くほど同数の要素が大量にある。ランクの振り方次第で1,000番と1,300番といった具合にブレてしまう。

今回はFDFとMy RankとTLF-Bの3つともランク付けの仕方が異なっている。本来，このような異なったランクの振り方を比較することは適切ではないのだが，今回はこのまま放置することにする。というのが，筆者がランクをこのデータに取り込んだ最大の理由は，実はランクそのものに重きを置いていないからである。後ほど明らかになるが，Aという語が100位で，Bという語が200位という数字そのものにさしたる価値がない上に，現実に異なったランクの振り方のデータを比較してみた場合，ランクの振り方に起因すると思われる違いより別の要因の方が大きいことが分かった。したがって，これらのデータはあくまで目安に過ぎず，使用頻度という観点から選択された語彙リストの一種と考えれば，その語彙に一票が投じられているかどうかだけが問題なので，ランクの数値そのものにはあまり大きな意味がない。要するに5,000語や3,000語の語彙リストとだけ見なしておく方が良いと思われる。辞書の＊印なども，第1レベル，第2レベルといった区別がされていても，実際の集計においてはその区別を用いていないという経緯もあるので，今回はさしあたって，大まかな傾向さえ読み取れれば良しとする。



| | A | DI | DJ | DL |
|---|---|---|---|---|
| 1 | | My Rank | FDF | TLF-B |
| 2 | le | 1 | 1 | 6 |
| 3 | de | 2 | 2 | 1 |
| 4 | du | 3 | | 23 |
| 5 | un | 4 | 3 | 15 |
| 6 | être 1 | 5 | 5 | 3 |
| 7 | être 2 | 5 | 5 | 1004 |
| 8 | et | 6 | 6 | 4 |
| 9 | avoir 1 | 7 | 8 | 9 |
| 10 | avoir 2 | 7 | 8 | 9 |
| 11 | à | 8 | 4 | 7 |
| 12 | en | 9 | 7 | 18 |
| 13 | que | 10 | 9 | 5 |
| 14 | ce | 11 | 12 | 22 |
| 15 | il | 12 | 13 | 12 |
| 16 | au | 13 | | 30 |
| 17 | son 1 | 14 | 26 | 36 |
| 18 | son 2 | 14 | 26 | 1122 |
| 19 | son 3 | 14 | 26 | 1122 |
| 20 | ne | 15 | 15 | 13 |

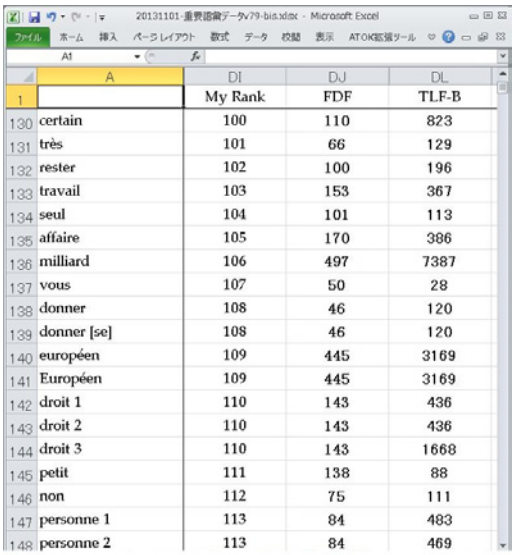
fig-1（My Rankをキーにした場合）

fig-1はMy Rankをキーにして並べ直した結果である。紙面の都合で画面が小さくて見えにくいのが残念だが，My RankとFDFが極めて似通った数値を出していることがわかる。またTLF-Bが若干，他の二つとはズレた値のように見えるからくりは後ほど解説する。

この図を見ると，いろいろなことがわかる。まず第一に，FDFで2か所数値が抜けている。具体的にはduとauである。FDFは恐らくduはde le，auはà leに分解して計算したものと思われる。一口に形態素の数を数えると言っても，duやauをどのように考えるか，あるいはson，sa，sesはすべて個別にカウントするのか，三つをひとまとめにしてしまう

のか，場合によってはsonとsaをひとまとめにして（単数），ses（複数）と区別することも考えられる。分析者の方針によって結果はすべて異なってくる。事実，TLF-Bは動詞については，その過去分詞形が形容詞化しているか否かにかかわりなく，過去分詞でカウントする計測の仕方をしている。その他，たとえばどこまでを派生語と考えるか，多義語をどこまで区別してカウントするかなど，細かい方針の違いの積み重ねが結果的に分析結果の大きな違いをもたらしてしまう。たとえば，TLF-Bだけがêtreを動詞と名詞で分けて計量している。データ中でla 1のように書いてあるものは，語彙リスト入力中に冠詞のlaと代名詞のlaを区別して掲載してある場合などにできるだけ忠実にデータを作るために区別したのだが，当然，見出しとしてはlaだけでその中に冠詞も代名詞も含めてある場合や，それこそleに全てまとめてある場合など，データのジャンルや解析機の精度以上に，分析方針の違いによる影響が無視できないほど大きい。

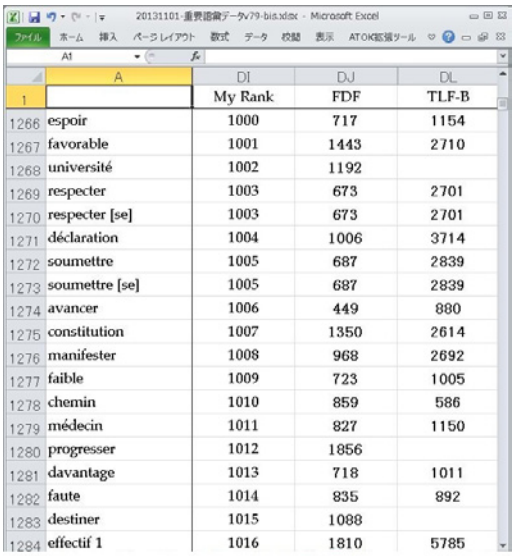
fig-2を見ると，似たような数値を出していたMy RankとFDFだが，100位あたりになると，早くも数値が一致しなくなってくる。



| | A | DI | DJ | DL |
|---|---|---|---|---|
| 1 | | My Rank | FDF | TLF-B |
| 130 | certain | 100 | 110 | 823 |
| 131 | très | 101 | 66 | 129 |
| 132 | rester | 102 | 100 | 196 |
| 133 | travail | 103 | 153 | 367 |
| 134 | seul | 104 | 101 | 113 |
| 135 | affaire | 105 | 170 | 386 |
| 136 | milliard | 106 | 497 | 7387 |
| 137 | vous | 107 | 50 | 28 |
| 138 | donner | 108 | 46 | 120 |
| 139 | donner [se] | 108 | 46 | 120 |
| 140 | européen | 109 | 445 | 3169 |
| 141 | Européen | 109 | 445 | 3169 |
| 142 | droit 1 | 110 | 143 | 436 |
| 143 | droit 2 | 110 | 143 | 436 |
| 144 | droit 3 | 110 | 143 | 1668 |
| 145 | petit | 111 | 138 | 88 |
| 146 | non | 112 | 75 | 111 |
| 147 | personne 1 | 113 | 84 | 483 |
| 148 | personne 2 | 113 | 84 | 469 |

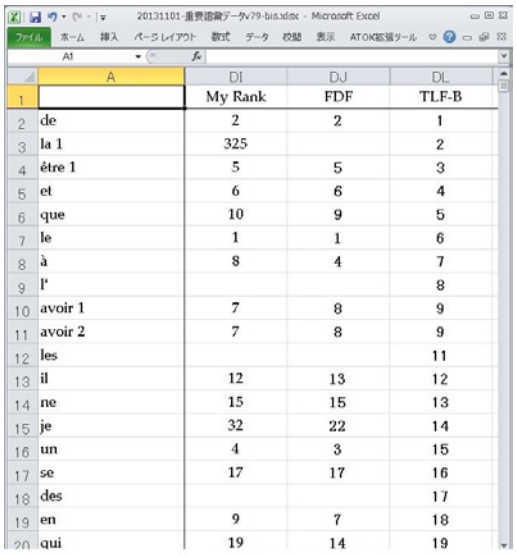
fig-2（100位近辺の図）

先ほどのfig-1の数値は3つのコーパス間で似通った数値だったが，fig-2はかなりばらついた印象を与える。しかし，よく見てみるとMy RankとTLF-B，FDFとTLF-Bはかなりズレているが，My RankとFDFの間にはさほど大きな違いがない。1,000あたりまで検討しても，My RankとFDFは比較的数値がまとまっているのに対して，TLF-Bだけがやや異なった数字を示していることは，1,000位近辺を示したfig-3からもよくわかる。これはTLF-Bの分析基準が個性的な上に，ジャンルがおおむね文学に偏っていることによると思われる。



| | A | DI | DJ | DL |
|---|---|---|---|---|
| 1 | | My Rank | FDF | TLF-B |
| 1266 | espoir | 1000 | 717 | 1154 |
| 1267 | favorable | 1001 | 1443 | 2710 |
| 1268 | université | 1002 | 1192 | |
| 1269 | respecter | 1003 | 673 | 2701 |
| 1270 | respecter [se] | 1003 | 673 | 2701 |
| 1271 | déclaration | 1004 | 1006 | 3714 |
| 1272 | soumettre | 1005 | 687 | 2839 |
| 1273 | soumettre [se] | 1005 | 687 | 2839 |
| 1274 | avancer | 1006 | 449 | 880 |
| 1275 | constitution | 1007 | 1350 | 2614 |
| 1276 | manifester | 1008 | 968 | 2692 |
| 1277 | faible | 1009 | 723 | 1005 |
| 1278 | chemin | 1010 | 859 | 586 |
| 1279 | médecin | 1011 | 827 | 1150 |
| 1280 | progresser | 1012 | 1856 | |
| 1281 | davantage | 1013 | 718 | 1011 |
| 1282 | faute | 1014 | 835 | 892 |
| 1283 | destiner | 1015 | 1088 | |
| 1284 | effectif 1 | 1016 | 1810 | 5785 |

fig-3（1,000位付近の図）



| | A | DI | DJ | DL |
|---|---|---|---|---|
| 1 | | My Rank | FDF | TLF-B |
| 2 | de | 2 | 2 | 1 |
| 3 | la 1 | 325 | | 2 |
| 4 | être 1 | 5 | 5 | 3 |
| 5 | et | 6 | 6 | 4 |
| 6 | que | 10 | 9 | 5 |
| 7 | le | 1 | 1 | 6 |
| 8 | à | 8 | 4 | 7 |
| 9 | l' | | | 8 |
| 10 | avoir 1 | 7 | 8 | 9 |
| 11 | avoir 2 | 7 | 8 | 9 |
| 12 | les | | | 11 |
| 13 | il | 12 | 13 | 12 |
| 14 | ne | 15 | 15 | 13 |
| 15 | je | 32 | 22 | 14 |
| 16 | un | 4 | 3 | 15 |
| 17 | se | 17 | 17 | 16 |
| 18 | des | | | 17 |
| 19 | en | 9 | 7 | 18 |
| 20 | qui | 19 | 14 | 19 |

fig-4（TLF-Bをキーに並べ替え）

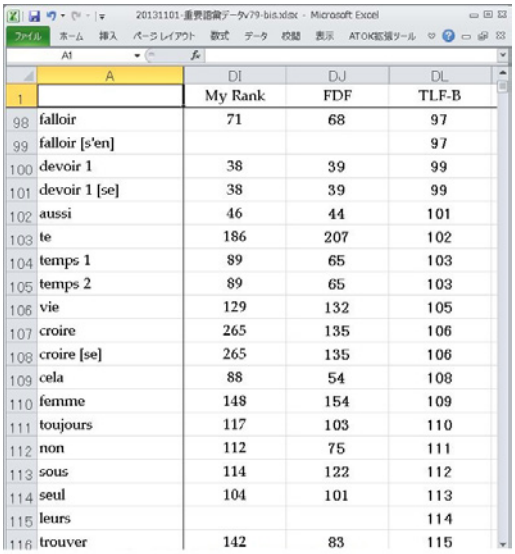
My Rankは必ずしもきれいではない生データを精度に限界のあるツールで分析した割には，ばらつき程度の範囲内に収まっており，結果的にかなり精度が高いと思われるFDFのデータと対応している。もちろん，個別に検討を加えれば問題のある値はいくつもあるのだが，だいたいの傾向はほぼ一致しているということで，実は充分なのである。その理由も後ほど述べる。

FDFで並べ替えてもおおむね同じなので，少しズレのあるTLF-Bをキーにして並べ替えてみる。

fig-4を見ると，TLF-Bをキーにして並べ直した場合，My RankとFDFの数値はばらばらであちこちが空白だらけになっている。しかし語をよく見ると，TLF-Bはl'やles, des, uneなどまで個別にカウントしており，何をどのようにカウントするかによって，結果は大きく異なることの見本である。

さらに，並んでいる語をもう一度観察してみよう。fig-1もfig-4も機能語や文法語と呼ばれる語ばかりが上位に並んでいる。先ほどのfig-1のように主にジャーナリスティックなデータであるMy Rankをキーに並べたときにはは異質のデータのような印象を与えたが，実はかけ離れたデータではないことがわかる。さらに，fig-4を先ほどのfig-2と同じく100位くらいまで下げてみよう。

確かに細かく見ればそれぞれのコーパスごとの傾向のようなものはあるが，あえてそこに拘泥すべきではない。むしろ元のデータがかなり異なっているのに，分析結果はそれほど異なっているわけではないことの方が重要であろう。読者の中にはこれらの違いは大きいのではないかと思われる人もいるかもしれないが，本当に似ても似つかぬ結果とはどのようなものかをこれから御覧に入れたいと思う。

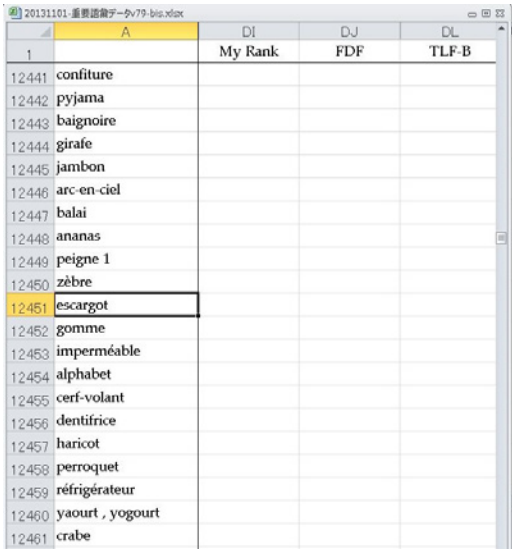


| | A | DI | DJ | DL |
|---|---|---|---|---|
| 1 | | My Rank | FDF | TLF-B |
| 98 | falloir | 71 | 68 | 97 |
| 99 | falloir [s'en] | | | 97 |
| 100 | devoir 1 | 38 | 39 | 99 |
| 101 | devoir 1 [se] | 38 | 39 | 99 |
| 102 | aussi | 46 | 44 | 101 |
| 103 | te | 186 | 207 | 102 |
| 104 | temps 1 | 89 | 65 | 103 |
| 105 | temps 2 | 89 | 65 | 103 |
| 106 | vie | 129 | 132 | 105 |
| 107 | croire | 265 | 135 | 106 |
| 108 | croire [se] | 265 | 135 | 106 |
| 109 | cela | 88 | 54 | 108 |
| 110 | femme | 148 | 154 | 109 |
| 111 | toujours | 117 | 103 | 110 |
| 112 | non | 112 | 75 | 111 |
| 113 | sous | 114 | 122 | 112 |
| 114 | seul | 104 | 101 | 113 |
| 115 | leurs | | | 114 |
| 116 | trouver | 142 | 83 | 115 |

fig-5（TLF-Bの100位近辺）

**語彙リストの分析結果**

このように複数の分析結果を比べてみた結果，たとえば3つのコーパス分析全てに共通するものを第1ランク，2つのコーパスで一致しているものを第2ランク，1つのコーパスでしか出現していないものを第3ランクのようにして，基礎語彙や重要語彙を確定させることができそうだが，その前に，フランス語の初級教科書でもよく見かける，比較的平易と思われる，escargotがどれくらいのランクなのかを確認しておきたいと思う。



| | A | DI | DJ | DL |
|---|---|---|---|---|
| 1 | | My Rank | FDF | TLF-B |
| 12441 | confiture | | | |
| 12442 | pyjama | | | |
| 12443 | baignoire | | | |
| 12444 | girafe | | | |
| 12445 | jambon | | | |
| 12446 | arc-en-ciel | | | |
| 12447 | balai | | | |
| 12448 | ananas | | | |
| 12449 | peigne 1 | | | |
| 12450 | zèbre | | | |
| 12451 | escargot | | | |
| 12452 | gomme | | | |
| 12453 | imperméable | | | |
| 12454 | alphabet | | | |
| 12455 | cerf-volant | | | |
| 12456 | dentifrice | | | |
| 12457 | haricot | | | |
| 12458 | perroquet | | | |
| 12459 | réfrigérateur | | | |
| 12460 | yaourt , yogourt | | | |
| 12461 | crabe | | | |

fig-6（escargotの検索結果）

驚いたことに，escargotはランク外なのである。その他，escargotの前後を見ると，girafe, zèbre, perroquet, crabeといった身近な動物や，confiture, jambon, ananas, haricot, yaourtといった日常生活で毎日のように用いる食べ物名など，いずれもフランス語母語話者であれば誰でも知っていて，第2言語学習者も知っておいて良さそうな語が並んでいるにもかかわらず，これらの語全てが3つのコーパス全てで上位5,000語以上を

とっても含まれていないのである。

先ほども述べたとおり，これら3つのコーパス分析結果の微細な違いを追求することに私があまり意味を見いだせない最大の理由は，このように5,000語程度の高頻度語を選んでもescargotが含まれていない，という事実による。escargotに限らず，先ほど例に挙げた語はもちろん，最重要語ではないだろうが，かといって決して難度の高い語ではない。フランス語が母語の人間で先ほどの語を知らなかったり，難しいと感じる人を見つける方が至難の業であるほど，ポピュラーな語である。ではなぜ3つのデータにescargotがランクインしなかったのか。実はいずれのデータでもたとえばescargotは最低でも一度は使われているはずである。FDFとTLF-Bは追跡のしようがないが，My Rankではescargotは出現回数が1,000回以下で，ランクとしては1万位くらいにある。つまりそれくらい使われない語だということである。確かに新聞でescargotがそれほど出現するとは考えにくい。日本語で考えても，新聞で「カタツムリ」はそんなに頻繁に出現しなくて不思議はないだろう。文学作品であればもう少し出てきそうな気もするが，少なくとも5,700位以下である。

それ以前に，たとえば本論の読者は毎日のように「カタツムリ」という語を使用しているだろうか。最後に「カタツムリ」という語を見聞きしたり自ら口にしたり書いたのはいつだったか思い出して欲しい。それくらい「使われない」のである。それにもかかわらず，我々は「カタツムリ」という語を決して難しい語だとは思わないし，幼稚園以上であれば子供でさえ知っているであろう。

コーパスが基礎語彙確定の決め手にならない理由はここに存在する。それほど使われるわけでもないのに，母語話者であれば誰でも知っているような平易な語が存在する。それは基礎的な語彙ではないのか？

**コーパスは使われた語しか分析できない**

もちろん，たとえば子供用絵本のコーパスや，学校教科書コーパスがあれば，カタツムリはかなり高頻度で出てくるかもしれない。しかし，逆に，カタツムリが高頻度で出てくるコーパスで，法律，経済，国際などといった中学生や高校生以上の学習者にとっては重要と思われる語がどの程度の頻度で出てくるか想像して欲しい。カタツムリとこれらの語が同じ高頻度で出てくるコーパスを想像することは極めて難しい[2]。

コーパスに分析できるのは，与えられたデータの範囲内のみである。コーパスは「存在しないもの」については全く分析できない。さらにescargotのように低頻度で出てきた場合，それが本当にほ

とんど使用されない語なのか，たまたまコーパスの構成上，低頻度で出てきただけなのかを区別することが非常に難しい。

それなら高頻度語だけを扱えばよいという考えも存在するかもしれない。しかし，高頻度語をよく観察してみると，高頻度語の中には3種類あることがわかる。一つは機能語や文法語で，これらはどのようなコーパスにおいても高頻度で出てくる。次は機能語や文法語に準じる語で，いわゆる多義的な語などは高頻度で出現する。たとえばgrand, petit, temps, homme, donner, trouverなどは機能語とは言い難いが，どのようなジャンルであろうと，誰が対象であろうと，頻繁に利用される語である。最後はコーパスに内在する高頻度語で，一般的な新聞データの場合，先ほど見たように，幼稚園児でも知ってそうな動物や身近な食べ物名称がそれほど高頻度では出てこなくて，政治経済関係の語が高頻度で出てくるのは当然であり，さらにそれらの語に固有の言い回し等があれば，ついでにそれらも高頻度で出てくる。しかし，それらは我々が漠然と基礎語彙とか重要語彙ととらえているものに合致するかと言えば，必ずしも合致しないことは容易に想像がつく。高頻度であることが必ずしも我々が漠然と想像している基礎語彙とか重要語彙と一致しているわけではない。

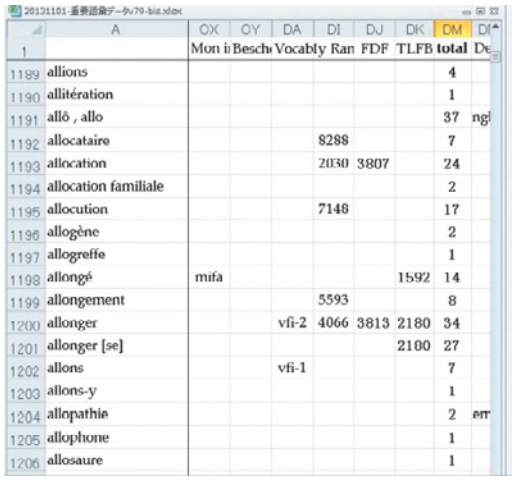
一見したところ，コーパス分析が示す結果は，いかにも重要な語の集まりに思える。事実，重要な語が並んでいる。しかし，基礎語彙や重要語彙を選別する場合，選んだ理由も必要なら，選ばなかった理由も必要なのである。たとえば，homme, femme, chienというたった3語の重要語データがあったとしよう。この場合，たとえばhommeとfemmeは人間なので何らかの類縁性があると自らを納得させることは不可能ではないが，なぜchienが選ばれているのかの理由がはっきりしない。それと同時に，chienを選ぶのならなぜchatも選ばないのかという理由もはっきりしない。基礎語彙や重要語彙と呼ばれるものを確定させる作業はこのような検討の連続である。ところが，コーパスは選ばれたものしか提示しないので，何が選ばれていないのかがわからないし，当然，なぜ選ばれなかったのかを検討する術もないのである。そもそも基礎的とか重要とはいかなることを意味しているのか。

**データの検証（2）**

基礎的とか重要とはいかなることを意味しているかを考察する前に，別の観点からデータを作成して検討してみよう。

筆者はとある必要から，日本で出版されている仏和辞典で基礎語とか重要語と称されている辞書中で＊が付けられてい

る語彙について調べる機会があった。その成果の一部については，日本フランス語フランス文学会で報告し[3]，論文でも発表済み[4]だが，その後，調査範囲を拡大していったところ，きわめて興味深い結果にたどり着いた。

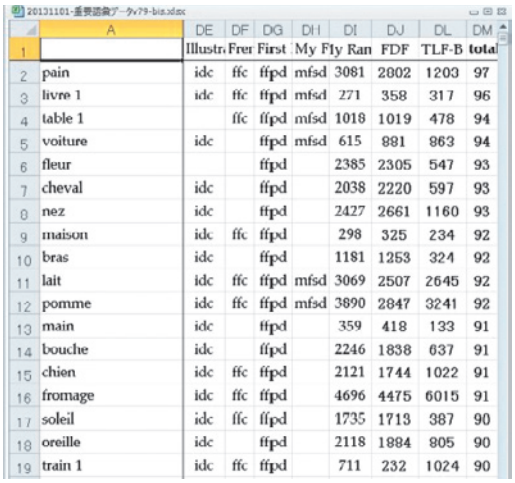


| | A | OX | OY | DA | DI | DJ | DK | DM | DI |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | Mon ii | Besch | Vocably | Ran | FDF | TLFB | total | De |
| 1189 | allions | | | | | | | 4 | |
| 1190 | allitération | | | | | | | 1 | |
| 1191 | allô , allo | | | | | | | 37 | ngl |
| 1192 | allocataire | | | | 8288 | | | 7 | |
| 1193 | allocation | | | | 2030 | 3807 | | 24 | |
| 1194 | allocation familiale | | | | | | | 2 | |
| 1195 | allocution | | | | 7148 | | | 17 | |
| 1196 | allogène | | | | | | | 2 | |
| 1197 | allogreffe | | | | | | | 1 | |
| 1198 | allongé | mifa | | | | | 1592 | 14 | |
| 1199 | allongement | | | | 5593 | | | 8 | |
| 1200 | allonger | | | vfi-2 | 4066 | 3813 | 2180 | 34 | |
| 1201 | allonger [se] | | | | | | 2180 | 27 | |
| 1202 | allons | | | vfi-1 | | | | 7 | |
| 1203 | allons-y | | | | | | | 1 | |
| 1204 | allopathie | | | | | | | 2 | em |
| 1205 | allophone | | | | | | | 1 | |
| 1206 | allosaure | | | | | | | 1 | |

fig-7（データ）

現時点でおよそ100ほどの辞書，単語集などの語彙リストを入力した。そして，その語彙リストにおいて取り上げられていれば1カウントとして集計してあるのがfig-7の右端のtotalの数値である。もちろん，このカウントの中には，先ほどのコーパス分析も語彙リストの一つとして含まれている。

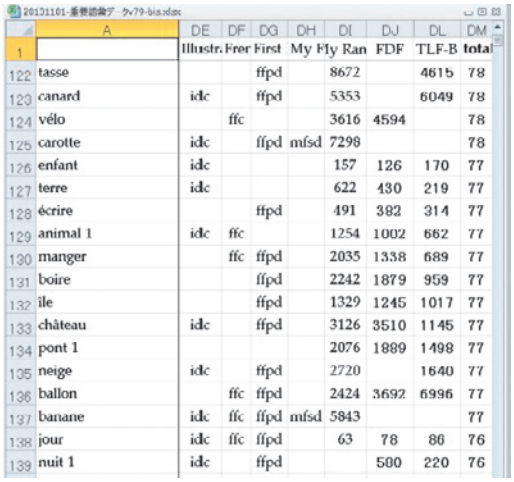
fig-7の場合であれば，たとえばallôは37の辞書等で取り上げられており，allitérationは一つの語彙リストでしか取り上げられていないことになる。言うなれば個々の語彙が得票数を競う形になるわけである。以下のfig-8が得票順に並べた場合の上位20位ほどである。



| | A | DE | DF | DG | DH | DI | DJ | DL | DM |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | Illustr | Frer | First | My | Fly Ran | FDF | TLF-B | total |
| 2 | pain | idc | ffc | ffpd | mfsd | 3081 | 2802 | 1203 | 97 |
| 3 | livre 1 | idc | ffc | ffpd | mfsd | 271 | 358 | 317 | 96 |
| 4 | table 1 | | ffc | ffpd | mfsd | 1018 | 1019 | 478 | 94 |
| 5 | voiture | idc | | ffpd | mfsd | 615 | 881 | 863 | 94 |
| 6 | fleur | | | ffpd | | 2385 | 2305 | 547 | 93 |
| 7 | cheval | idc | | ffpd | | 2038 | 2220 | 597 | 93 |
| 8 | nez | idc | | ffpd | | 2427 | 2661 | 1160 | 93 |
| 9 | maison | idc | ffc | ffpd | | 298 | 325 | 234 | 92 |
| 10 | bras | idc | | ffpd | | 1181 | 1253 | 324 | 92 |
| 11 | lait | idc | ffc | ffpd | mfsd | 3069 | 2507 | 2645 | 92 |
| 12 | pomme | idc | ffc | ffpd | mfsd | 3890 | 2847 | 3241 | 92 |
| 13 | main | idc | | ffpd | | 359 | 418 | 133 | 91 |
| 14 | bouche | idc | | ffpd | | 2246 | 1838 | 637 | 91 |
| 15 | chien | idc | ffc | ffpd | | 2121 | 1744 | 1022 | 91 |
| 16 | fromage | idc | ffc | ffpd | | 4696 | 4475 | 6015 | 91 |
| 17 | soleil | idc | ffc | ffpd | | 1735 | 1713 | 387 | 90 |
| 18 | oreille | idc | | ffpd | | 2118 | 1884 | 805 | 90 |
| 19 | train 1 | idc | ffc | ffpd | | 711 | 232 | 1024 | 90 |

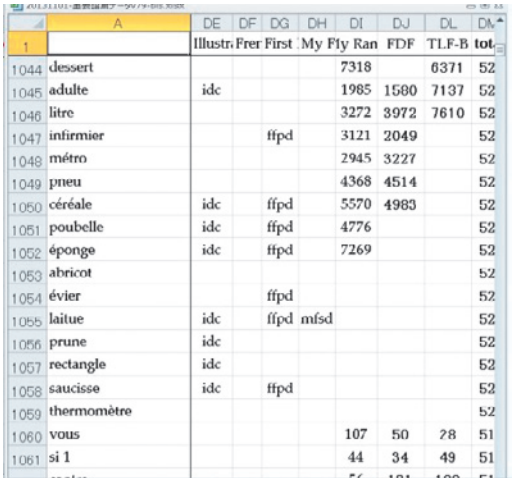
fig-8（得票順）

先ほどのコーパス分析においては，文法語や機能語が上位を占めたが，得票数で集計してみると，平易な名詞が上位を占めている。動詞も最初に現れるのは116番目のacheterであり，次は129番目のmangerである。avoirが現れるのが454番目，êtreは501番目である。平易な語彙，という観点からいえば，acheterやmangerの方が人気があり，文法度が高いavoirやêtreがもう少し後から出てくるのも決して不思議ではない。そして，得票順の100位あたりで早くも先ほどのコーパス分析で出現しなかった語（空白部分）が出始めている。



| | A | DE | DF | DG | DH | DI | DJ | DL | DM |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | Illustr. | Frer | First | My Fly | Ran | FDF | TLF-B | total |
| 122 | tasse | | | ffpd | | 8672 | | 4615 | 78 |
| 123 | canard | idc | | ffpd | | 5353 | | 6049 | 78 |
| 124 | vélo | | ffc | | | 3616 | 4594 | | 78 |
| 125 | carotte | idc | | ffpd | mfsd | 7298 | | | 78 |
| 126 | enfant | idc | | | | 157 | 126 | 170 | 77 |
| 127 | terre | idc | | | | 622 | 430 | 219 | 77 |
| 128 | écrire | | | ffpd | | 491 | 382 | 314 | 77 |
| 129 | animal 1 | idc | ffc | | | 1254 | 1002 | 662 | 77 |
| 130 | manger | | ffc | ffpd | | 2035 | 1338 | 689 | 77 |
| 131 | boire | | | ffpd | | 2242 | 1879 | 959 | 77 |
| 132 | île | | | ffpd | | 1329 | 1245 | 1017 | 77 |
| 133 | château | idc | | ffpd | | 3126 | 3510 | 1145 | 77 |
| 134 | pont 1 | | | | | 2076 | 1889 | 1498 | 77 |
| 135 | neige | idc | | ffpd | | 2720 | | 1640 | 77 |
| 136 | ballon | | ffc | ffpd | | 2424 | 3692 | 6996 | 77 |
| 137 | banane | idc | ffc | ffpd | mfsd | 5843 | | | 77 |
| 138 | jour | idc | ffc | ffpd | | 63 | 78 | 86 | 76 |
| 139 | nuit 1 | idc | | ffpd | | | 580 | 220 | 76 |

fig-9（コーパス欠落語）



| | A | DE | DF | DG | DH | DI | DJ | DL | DM |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | Illustr. | Frer | First | My Fly | Ran | FDF | TLF-B | tot |
| 1044 | dessert | | | | | 7318 | | 6371 | 52 |
| 1045 | adulte | idc | | | | 1985 | 1580 | 7137 | 52 |
| 1046 | litre | | | | | 3272 | 3972 | 7610 | 52 |
| 1047 | infirmier | | | ffpd | | 3121 | 2049 | | 52 |
| 1048 | métro | | | | | 2945 | 3227 | | 52 |
| 1049 | pneu | | | | | 4368 | 4514 | | 52 |
| 1050 | céréale | idc | | ffpd | | 5570 | 4983 | | 52 |
| 1051 | poubelle | idc | | ffpd | | 4776 | | | 52 |
| 1052 | éponge | idc | | ffpd | | 7269 | | | 52 |
| 1053 | abricot | | | | | | | | 52 |
| 1054 | évier | | | ffpd | | | | | 52 |
| 1055 | laitue | idc | | ffpd | mfsd | | | | 52 |
| 1056 | prune | idc | | | | | | | 52 |
| 1057 | rectangle | idc | | | | | | | 52 |
| 1058 | saucisse | idc | | ffpd | | | | | 52 |
| 1059 | thermomètre | | | | | | | | 52 |
| 1060 | vous | | | | | 107 | 50 | 28 | 51 |
| 1061 | si 1 | | | | | 44 | 34 | 49 | 51 |

fig-10（1,000位前後のコーパス欠落語）

tasse, canard, vélo, carotte, neige, banane, nuitなど，確かに新聞などでそれほど出てきそうにはないが，難度は高くないばかりか，フランス語を学習しているのであれば，知っておくべき語ばかりではなかろうか。

得票数上位1,000位あたりになると，コーパス欠落語がさらに増えてくる。

コーパスに欠落していた語（本当に含まれていなかったのか，極度に出現頻度が低いかは問わない）を見ると，abricot, évier, laitue, prune, rectangle, saucisse, thermomètreなど，決して難解な語ではなく，その割に日本のフランス語教科書でも見かけることが少なく[5]，このような語は積極的に教える必要がありそうな語ではなかろうか。fig-3のようにコーパスの順位で並べ直した場合，1,000位でも，数値にばらつきはあるものの，3つのコーパスの中で欠落があるケースはほとんどなかったのと好対照である。

## 集合知としての基礎語彙

ここまで好対照な二つの結果を検討してきた。言語研究においてコーパス利用はまだまだ多くの可能性を秘めており，言語研究全般にコーパスが不適切なので

はない。物事には適不適があり，基礎語彙などを確定する場合には，ここまで検討してきたとおり，コーパスは必ずしも適切な基準になりえないことが明らかであるだけのことである。

では，他方において，投票方式はどうであったか。むしろこの方が基礎的な語を的確に選び出していたように思われる。今まで便宜上，投票方式と呼んできたが，既に読者諸賢にあっては明察のとおり，これは集合知理論に依拠している。まさしく「多くの人が基礎的と考える語彙が基礎語彙である」。

コーパス分析に適不適があるのと同じく，集合知理論にも適不適はある。分析結果に欠落があり，何が欠落しているのかわからないという点でコーパスは明らかに基礎語彙確定に不向きである。逆に集合知理論は基礎語彙確定に向いている。そもそも我々は何が基礎的で，何が重要であるか，それを判定する客観的な尺度を持ち合わせていない。たとえば，「フランス語の母語話者ではない日本人の大学生くらいの学習者にとって基礎的（重要）な語彙」とかなり細かく限定しても，これを判定するための基準はどこにあるだろうか。結局，「大学生ならこれくらいの語は知っていても当然ではなかろうか」とか「大学生くらいの日本人フランス語学習者の多くが知っている」といった目安しか存在しない。

今回分析に用いた語彙リストがそもそも客観的ではないのではないかという考え方もありうるだろう。確かにどのような客観的なデータに基づいて選ばれたのか不明の語を集めた語彙リストが存在する。というより，科学的を謳っている語彙リストであっても，その科学性をすり抜けた低頻度語が多数交じっており，その段階で，客観性を主張できないデータになっていると判断せざるを得ないものもあった。これらの主観性が多かれ少なかれ混じった語彙リストは不適格なのであろうか。

実はそうではない。データとする語彙リストは主観的に作られていても構わない。作成にあたってバイアスがかかっていても構わない。むしろバイアスがかかっている方が好都合である。選択する語数に違いがあっても構わない。その語数が学習者の負担を考えて選択された数であろうと，リスト製作者が思いつく全てであろうと，どちらでも構わない。そのような雑多なリストを集めれば集めるほど，多くの集合的無意識が蓄積されることになる。

実際，データを入力していて分かったことは，多種多様な語彙リストは決してでんでバラバラではない，ということである。もちろん，集計結果にきわめて近い収束率の高い語彙リストもあれば，かなり自由奔放で個性的な語彙リストもあるのは事実である。しかし，だからと言って，完全に集計結果と合致する語彙

リストや，全く見当はずれな語ばかり選んでいるリストは存在しない。程度の差こそあれ，おおむね妥当な選択をおこなっているが，時々，見当はずれなものも交じっている，というのが現実である。その割合に差があるだけだ[6]。互いにその存在を知らないであろうと思われる語彙リストの間にもかなり確固たる類似性が存在している[7]。なぜこのようになるのか。そもそも基礎語彙や重要語彙が「多くの人が基礎的であるとか重要と考える語彙」のことだから，選択者は当然，単に自分自身にとって重要な語彙だけでなく，一人でも多くの読者にとって重要ではなかろうかと思われる語を選ぼうとするであろうし，それが主観的な判断である以上，どうしても他の人は重要とは考えない語が交じることも仕方ない。そして結果的にそうした語彙に票が集まることはなかった。

現在，上限2万語程度のデータまでを入力の対象にしているが，データの総体としてはおおむね3万語くらいの範囲内に収まっている。2万語というのは辞書の対象者から判断すると，フランス語ネイティブ話者の小学生くらいまでで，3万語レベルになると，中高生くらいが対象になる。おそらく3万語レベルの語彙リストを入力していけば，データの総数は4万語くらいになるだろうと予想している。しかし，おおむねそのうちの半数程度は語彙リスト全体で1度しか使われていない語となり，外国人学習者があえて早急に学習する必要度が低い語と考えられる。結果的に2万語前後の語がフランス人ネイティブ話者が義務教育終了時までに習得済みあるいは習得することが望ましい語のコアな集団と予想される。

## 今後の展望

筆者が本当に明らかにしたいのは，2万語前後や3万語程度といったマスの数値ではない。あくまで個々の個別の語彙を確定したいと考えている。基礎語彙と呼ばれる語の集団においては，2万とか5,000といった数値にあまり意味はない。あくまでその中身は具体的にどのようになっているかであって，それが確定しないと基礎語彙の研究が完成したとは言えない。短期的な展望としては，早急にこれらの語の具体的な全貌を明らかにしたいと思っている。

同時に，今回利用したデータはあくまで上限2万語程度までで，これではフランス語ネイティブ話者の小学生レベルである。第2言語学習者として，ある程度その言語を習得したと言えるためには，最低でもネイティブ話者が義務教育終了時までに習得済みあるいは習得が望まれる語まで広げておく必要がある。しかしそのためには3万語レベルのデータを追加していく必要があるのでもう少し時間がかかる。中長期的にはこのレベルまで

は具体的な語を明らかにしたいと考えている。

## 参考文献


・Brunet, Etienne, *Le vocabulaire français de 1789 à nos jours*, I, II, Librairie Slatkine, 1981.

・Deryle Lonsdale and Yvon Le Bras, *A Frequency Dictionary of French*, Routledge, 2009.

・他，語彙リスト多数.


## 注

1　ドイツのミュンヘン大学コンピュータ言語学研究所のHelmut Schmidが開発したTreeTaggerという形態素解析プログラム（解析機と呼ぶこともある）にシュトゥットガルト大学の言語学研究所ロマンス語部門のAchim Steinが作成したパラメータファイルの組み合わせを用いた。フランス語の形態素解析の手法としては最も一般的なものと思われる。フランス語の形態素解析機には有償・無償を含めて他にもいくつかあるが，TreeTaggerが最も妥当な結果を返しているように思われる。

2　バランスド・コーパスや均衡コーパスと呼ばれる，コーパスのブレンディング手法があるが，「カタツムリ」と「経済」が同じくらいの高頻度で出てくれば，それはバランスのとれたコーパスなのだろうか？
別の機会に詳しく論じたいと思うが，基礎語彙や重要語をたとえば1万語や5,000語選び出すことは可能であろう。事実私も取り組んでいるところである。ところがネイティブ話者がそれらを習得する年齢を考慮に入れると，語彙リストは年代によって習得済みの語と習得が望まれる語に分類できることは明白である。年少者になればなるほど，望まれる語が増えて，ある程度の年齢を超えれば，すべてが習得済みになるかもしれない。このように通時的な観点からも考える必要のある語彙組織は単なるバランスの問題ではなく，福岡伸一氏の言うような動的均衡（équilibre dynamique）としてとらえる必要がある。その意味では，たとえ外国語としての学習者であっても，ネイティブ話者の通時的な習得履歴をたどるような学習の仕方は決して無駄ではなかろうと考えている。

3　2008年6月，「学習用仏和辞典をめぐる諸問題－重要語を中心に－」日本フランス語フランス文学会，於青山学院大学

4　「フランス語の基礎語彙画定に関する試論（1）― 量的考察 ―」，『言語と文化』，愛知大学語学教育研究室，第17号，2007年7月。

「学習者用仏和辞典を巡る諸問題」，『言語と文化』，愛知大学語学教育研究室，第25号，2011年7月。

5　日本のフランス語教科書は，まず第一に

文法に中心を置きすぎているし（文法の試験をしたら100点の学生でもフランス語を音読することさえできないことは珍しくない），少なくとも本論との関係からいえば，語彙習得に対する配慮が欠けている。最近でこそ巻末にグロサリーが付くようにはなったが，たとえば本文中で用いられている語にバラエティが少ないとか，その割に，あえて早期に身につける必要のない語彙を用いているなど，どれくらいの語彙が必要なのか，どの語彙から習得を始めるべきかといった配慮が足りないように思われる。

6　もちろん，見当はずれな語をきわめて多く含んだ，学習用単語集としては首をかしげざるを得ないものも多々ある。

7　おそらくウィトゲンシュタインの考える家族的類似性と結びつく概念装置が潜んでいると考えられる。出版された年も土地も違い，対象とする読者も異なり，選んだ語数にもかなり違いがあり，おそらくは互いの存在さえ知らないであろう二つの語彙リストのほとんど大部分が重なり合うという事実は，「選者が重要（基礎的）だと思った語」（家族性）という唯一の共通点しかなく，それによって出来上がった語彙リストが極めて似通っている（類似性）という事実はきわめて興味深い。