論　文

フランス語の基礎語彙確定に関する試論（1）

── 量的考察 ──

中　尾　　浩

要　旨

コーパスの分析から得られる知見には様々なものがある。語彙頻度リストを作成することはコーパス利用の中でも最もありふれたものだが，従来，語彙頻度リストを作成するために必要な語数はあまり明確にされてこなかった。漠然とデータ量は多いほどよいのではないかと思われていたが，本稿ではフランス語の新聞データを元にして，データ量の違いによってどれほど得られる結果にばらつきがあるか（一致率にばらつきがあるか）を明らかにした。今回の調査では1億語レベルまで調査したが，明確に一貫して一致率は上昇したし，語彙頻度リストのばらつきも小さくなっている。語彙頻度リストを作成する場合に最も重要なのは総lemma数であり，母集団からサンプリングした場合には十分なlemma数が得られないで，結果にばらつきが生じる。さらにlemmaは単に出尽くすだけではだめで，その出現が安定するためには，必然的に総token数が多くなければならないと推定できる。

キーワード：コーパス，フランス語，語彙頻度リスト，一致率，データ量

1．はじめに

言語学において従来は不可能だったことで，コーパスを扱うことによって可能になったことは多々ある。コーパスの利用において，何よりも従来の言語学的方法論と異なるのは，

計量的分析が容易になったことである。手作業では一生かかっても不可能だった計算をコンピュータなら瞬時に計算してしまう。

ところが，この計量的分析が簡単なようで難しい。理由はいくつかある。たとえば分析に必要十分な量のデータを集めにくいとか，計量的な分析には述べ語と見出し語を必要とするが，延べ語を見出し語化することでさえも実はそれほど簡単ではない。なお，延べ語はtoken，見出し語はlemmaやtypeと呼ばれることもあり，本稿では以下，特に断りがない限り，延べ語のことをtoken，見出し語のことをlemmaと呼ぶことにする。

言語の計量的分析については，隅から隅まで厳密でなければならないという意見もある。たとえば国立国語研究所の伊藤雅光は伊藤 (2002) の中で，以下のように述べている。

> なお，第Ⅱ部の実践編では，言語処理プログラムを使えば瞬時に終わるような作業を長い時間をかけてコツコツ行うように設定しているが，これは，作業内容を理解するために必要だからである。
>
> 近年，電子化コーパスや電子化テキストが容易に，しかも安価に入手できるようになり，また言語処理プログラムの販売や配付も相次いでいる。そのため，一度も読んだことのないテキストを対象にして，どのような処理をしているかもわからない言語処理プログラムを使って，パソコンに語彙表を作らせるという事例が目につくようになった。しかし，これは果たして，調査や研究といえるのであろうか。人文系の研究者が，読んだこともない文献の「研究」をすることほど，矛盾に満ちたものはない。これは，言語研究の危機である以前に，学問の危機というべきである[1]。

この見解はまったく正しい。しかし，他方においてこのやり方では，データを作成することだけで一生を終わりかねない恐れもある。伊藤氏の警鐘に対しては，筆者としてはせめて，必要なデータは自前で構築して可能な限り目を通し，自分で理解できない処理は加えないという原則を守ることにした。したがって，筆者が管理しているデータは既存の販売されているデータに比べると，はるかに「汚い」データである。しかし，逆に言えば，それこそ我々が普段目にしている言語の姿でもある。また，本稿で使用した言語処理技術も，入力ファイルに対してどのような処理がなされるかがはっきりしているプリミティブなコマンドを組み合わせるのが基本で，その結果出てくる出力ファイルもおそらく誰が見てもわかる処理方法しか用いない，という方針で応えたいと思う。

本稿において以下明らかにしていくが，計量的な研究の中には相当な量の元データが必要と思われる分野がある。クローズドなコーパス，たとえばある作家の作品における分析等であれば，その作家の作品がトータルで30万語であれば，その範囲内で研究するしかない。しかし，オープンなコーパス，たとえば新聞データであるとか，現代小説といった，範囲を定めることの出来ない場合には[2]，10万語，100万語程度では，研究対象によっては正確な結論を出せない。この点については以下の論証で明らかになる。

あるいはそれらの小さな母集団の分析によって導き出した結果に統計的手法を用いて予測しても決して現実を反映しない。それだけ言語は複雑系的な対象である。全体としてはある程度の安定さを持っているが，局所論的には同じ振る舞いを見せないことが多い。統計的手法の重要性は十分に承知しつつ，残念ながら言語のような対象においては，小さな母集団の分析から全体（厳密には言語事象に全体などはありえないが）を推し量ることは，ほとんど不可能であると言わざるをえない[3]。本研究においては，語彙使用の分析に限って，とりわけ量的な問題について扱うことにする。

2．本研究の方針

はじめにでも述べたとおり，言語の計量的研究においては，何はさておいてもtokenとlemmaの両方の情報が必要になってくる。tokenはrawデータを集めさえすれば，tokenの集合になるが，lemmaの方は何らかの処理が必要となる。tokenをlemma化する専用のアプリケーションとしては，lemmatizerと呼ばれるものがあるが，実際にはPOS Tagger等が同時に出力するケースが多い。

フランス語の場合，POS Taggerは決して多くはない。もっとも充実しているのは英語だが，必ずしも多数あるわけではない。フランス語分析用のPOS Taggerとしてはフランスの Synapse Developpement 社の商用ソフトであるCordial Université（現在では Cordial Analyseurと改名），フィンランドのConnexor社のMachinese Phrase TaggerやMachinese Syntax，ドイツのシュトゥットガルト（Stuttgart）大学のコンピュータ言語学研究所で開発されたTreeTaggerに同じくシュトゥットガルト大学のAchim Steinが作成したフランス語辞書を読み込ませて利用するのが主たるところである[4]。そのほか，Brill TaggerにATILFがフランス語の辞書を載せた上でGUI化したツールもあったが，筆者のコンピュータ環境では安定して動作しなかった。どのソフトの出力結果も一長一短で，特に抜きん出ているものがあるわけではないが，辞書のメンテナンスが可能であること，フリーで利用可能なことを考慮した結果，TreeTaggerを採用することにした。

ただし，TreeTaggerを採用するにしてもいくつかの問題点がある。まず辞書の精度だが，lemma化についてはかなりの精度ではあると思える。同綴異義語が多いといったフランス語独特の問題もあり（たとえばest, été, nuit, etc.）おそらく今後どれほど精度が上がっても100%はありえないだろう。これはフランス語に限らないし，TreeTaggerに限った話ではない。さらにPOS解析についても，TreeTaggerのアルゴリズムでは限界があるようだ。こちらも今後どれほど精緻なPOS解析アルゴリズムが見つかっても100%の精度はありえないだろう。そもそも人間が解析しても100%同じPOS解析をするとは限らない以上，コ

ンピュータの解析が100%正確であることなどありえない。

もう一つの問題点はもう少し形而下的な問題で，TreeTaggerは分析させるファイルの中で何らかのシーケンスによってはそこで処理を終了してしまう。ただし，これもTreeTagger固有の問題ではなく，Cordialもよく途中でフリーズした。最も頑健なのはConnexorの製品であった。従って，分析対象となるデータの最後まで分析できている場合もあれば，ファイルの途中で処理が中断されている場合もある。後ほど紹介するTreeTaggerの出力例を見ればお分かりの通り，TreeTaggerで解析させたファイルの容量は，元のtokenにPOSタグとlemmaが付加されるので，おおむね解析元のファイルの３倍となる。従って元のファイルより小さなファイルを出力したり，２倍程度の大きさのファイルで終わっていたら，それはTreeTaggerが途中で解析を停止していると予想できる。そのような場合は解析を中止しているセンテンスを取り除いて再度解析をさせると，たいていは最後まで解析を行う。問題なのは，２倍は超えているファイルである。その大部分は最後まで解析が終わっているのだろうが，本当に最後まで解析しているかどうかは一つずつ検査しなければならない。これはかなり面倒な作業である。上記の事情により，分析の元になるtoken数は結果的にlemma化作業が終わった部分までをtoken数と見なさざるを得ない。データ数が多ければ，この方法でも十分に実用的なので，今回はTreeTaggerが解析を終えたところまでをtoken数と考えることにする。

３．lemmaについて

lemmaを数えると言ってもそれほど簡単ではない。たとえばauは1 lemmaと見なすのか，それともàとleと見なすか（今回は１語とカウントしてある）。mon，ma，mesは合算するか，個別に数えるか（今回は合算してある），人称主語のvousと目的語のvousは同一視するかしないか（今回は同一視してある），固有名詞については，人名や地名はカウントしていないが，たとえばpartie socialeなどは，partieもsocialeも普通名詞（形容詞）だが，partie socialeという連なりになると固有名詞となってしまう。その場合，partie socialeは2語で固有名詞と見なしてカウントしないのか，別々にpartieやsocialとしてカウントするのか（今回はカウントしてある）等，実は必ずしも明確な基準もない。

これらは厳密にはある程度の方針を立てなければならないが，今回はそれらの点についてはあえて目をつむって，基本的にTreeTaggerの出力結果を前提とした。同一の基準で分析したものであれば，結果のばらつき方も同一なので，その範囲内であれば検討に値すると判断した[5]。

4．基礎語彙について

基礎語彙（vocabulaires fondamentaux）とは何かについて，特定の定義はまだない。小池（2003）によると，基本語彙（basic vocabulary）は以下のように説明されている。

> 基本語彙がどういうものであるかという点に関して，決定的な定義はまだないと言ってよい。ただし，基本語彙の特徴として次のものを挙げることができよう。①高頻度（使用頻度が高い），②意味的無標性（意味の上での有標性（markedness）が低い），③高連語可能性（他の語との連結（collection）が比較的自由である），④統語的自由性（文法的な制約が弱く，比較的多くの構文に生起できる），⑤文体的中立性（文字通りの意味が中心で，感情的意味等の余分なニュアンスを持つことが少ない）[6]。

基本語彙の特徴の第一として挙げられているのが高頻度で，使用頻度が高いということは基礎性，基本性の最も重要なものさしと言えよう。いささか古いが田中（1988）では，もっと端的に以下のように説明されている。

> Basic vocabulary《基礎語彙》(言)　一言語の語彙のうち使用頻度の特に高いもの[7]。

基礎語彙の確定において高頻度であることの重要性は伝わるが，おそらく今日ではこれでは通用しないだろう。小池（2003）ほど細かくは分類しないが，もう少しシンプルで妥当と思われる見解が鈴木（2006）で，以下のように説明されている。

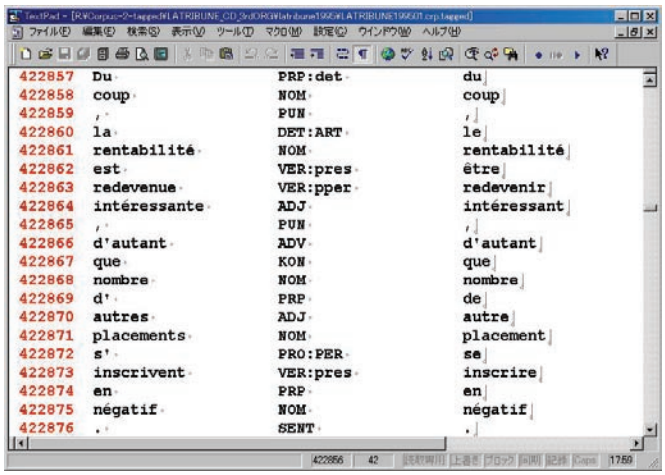
> 「使用度数」(「使用頻度」) の高い語を「高頻度語」というが，それらには，どのような資料においても高頻度で使用される語と，たまたまその資料において高頻度で使用された語の2種類が含まれる。なお，高頻度でかつ「使用範囲」が広い語を「基本語（彙）」とよぶ。上位何語までにするかは目的による。これらは基本的に語彙調査の結果に基づいて定められるものであるが，一方，必ずしも高頻度・広範囲でない語をも含め，その言語を使って生活していく上でどうしても必要となる語群という観点から，演繹的，体系的に定めたのが「基礎語彙」である[8]。

高頻度語はコーパスさえきちんと整備されていれば比較的容易に確定できるし，広範囲高頻度語もコーパスの種類さえ増えれば，各高頻度語から共通部分を抜き出せばよいことになる。基礎語彙確定において難しいのはむしろ「その言語を使って生活していくうえでどうしても必要となる語群という観点から，演繹的，体系的に定め」ることであろう[9]。これらの点については，実際に分析してみてどのような連関があるかをまずは調べるべきで，本稿では主に新聞コーパスというかなり限られた分野ではあるが，成人のかなり多数が日常的に接する可能性が高いコーパスを分析することによって，一分野についてではあるが高頻度語を確定することを目指す。

5．分析の方針

TreeTaggerによる出力結果は以下のようなものである。

図 1　TreeTaggerの出力例



422857	Du	PRP:det	du
422858	coup	NOM	coup
422859	,	PUN	,
422860	la	DET:ART	le
422861	rentabilité	NOM	rentabilité
422862	est	VER:pres	être
422863	redevenue	VER:pper	redevenir
422864	intéressante	ADJ	intéressant
422865	,	PUN	,
422866	d'autant	ADV	d'autant
422867	que	KON	que
422868	nombre	NOM	nombre
422869	d'	PRP	de
422870	autres	ADJ	autre
422871	placements	NOM	placement
422872	s'	PRO:PER	se
422873	inscrivent	VER:pres	inscrire
422874	en	PRP	en
422875	négatif	NOM	négatif
422876	.	SENT	.

このデータを見れば，token数は行数に等しいので，token数を数えるときには行数を数えればよい。もちろん，語数に関係のないponctuation等は取り除いた上で計測している。さらに今回は，固有名詞と思しきものも削除してある。具体的には文中で大文字で始まっている語である。lemmaは第三カラムに出力されているので，ここだけを取り出して，さらに同一行を削除してしまうとlemmaのリストが取り出せる。UNIXではuniqというコマンドがあり，同一行を削除するコマンドだが，削除するだけでなく，同一行がいくつあったかをカウントできるオプションがあるので，sort（並び替え）をした後にuniqをかければlemma数をカウントできるので，再度sortで頻度順に並べ替えておけば，lemmaの頻度データを得ることができる。

基本となるデータはこの二つである。ただし，token数（lemma数）を数える場合に，どのくらいのボリュームを単位とするかが問題となる。そもそも筆者が構築しているデータの単位が均一ではない。一日ごとのデータもあれば，一ヶ月ごとや一年ごとのデータもある。また一ヶ月ごとのデータが一日ごとのデータより小さいことも珍しくない。これはデータの提供方式，収拾方法，処理方法等さまざまな事情による。一ヶ月ごとのデータを分析させたときにTreeTaggerがファイルの先頭近くで処理を終了していたら，一日分のデータより小さい場合もありうる。

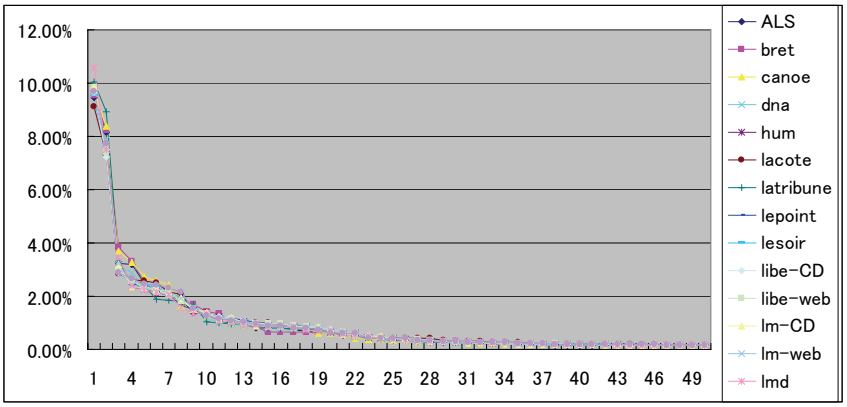
語彙の出現頻度に基づいて基礎語彙を確定したい場合に，元のデータが一日ごとか一ヶ

月ごとか等はまったく問題ではない。同一母集団において，token数がいくつかが問題であって，まとまりが一日ごとであるか一年ごとであるかはまったく関係ない。データ構築上の都合によるものであって，以下のデータ分析において，一日ごとのデータが用いられたり，一ヵ月ごとのデータが用いられたりしているが，token数（lemma数）がいくつかであるかの方に注目していただきたい。

今回の研究において，重要なのは総語数（総token数）とlemma数であるが，もう一つ重要なのは，出現頻度でも出現率でもない。ランクである。基礎語彙は基本的に出現頻度の高いものが採用されるべきで，出現頻度が低いものを基礎語彙に採用するには何らかの理由が必要になる。ところが，出現頻度が低くても基礎語彙に採用すると言う意味は，Aというコーパス（母集団）においては出現率が低いが，Bというコーパスにおいては出現率が高いので，この点を考慮して，Aでは出現率が低くても基礎語彙（重要度高）として採用する，というのが正しい言い方で，単にAというコーパスを見るだけでその中で出現率（≒出現頻度）が低いが重要だ，とは判断できない。今回用いたデータはかなり多くのデータを分析したつもりだが，いずれも新聞等のジャーナリスティックなデータで，Le MondeとHumanitéはもちろん異なったコーパスではあるが，その上位カテゴリにおいては一つの大きなグループを作っているので，その意味では均質なデータ群と言える。もちろん，Le MondeとHumanitéの違いは考えられるが，それほど大きなものではないことはいずれ明らかになる。

出現率を用いなかった理由の一つは，基礎語彙のようなリストを作成する場合，必要なのは出現率ではなくランクであり，ランクさえわかれば，実は出現率はほとんど一対一に対応している。それはジップの法則（Zipf's law）[10]からも明らかであり，ある程度十分に大きなデータなら，どのデータでもほぼ同じジップの曲線を描く。

図2　各コーパスの出現率によるジップ曲線[11]



ランクを用いるもう一つの理由は，語彙頻度リストを作成する場合，ある語の出現率が5%か3%かといった違いはほとんど意味をなさないからである。もちろん，クローズドなコーパス（たとえばある作家の全著作のデータ）においては15%か13%かは意味を持つかもしれないが，語彙頻度リストのような場合は，0.5%であろうと0.3%であろうとランクの100番目くらい，ということさえわかれば十分である。従って，ランクの方も95位か103位かなどということはほとんど意味がない。本稿では1000語をひとまとまりにして一致率をとることにした。そうした上位1000語といったくくりを出現率では行いにくいのも，ランクを用いた理由の一つである。

本稿においては，基本的にランクの一致率を重視した。基礎語彙を確定するに当たって，データ群ごとにどの程度一致しているかは検証しておく必要がある。実際に一致率を検証してみると，確かにジャーナリスティックなデータの場合，とりわけ上位5000語程度までは，1000語ずつ区切って調査してもおおむねデータは一致している。このことが意味することは二つある。ジャーナリスティックなデータを用いる場合，基本的にどのデータを用いようと，結果はだいたい類似するので，どのデータをどれくらい用いるというより，実際にはデータの総量のほうが問題で，それは以下，順次明らかになる。むしろ，総データ量の方が問題なら，Aという新聞もBという新聞も合算の上，語彙頻度を出す方が，下手にブレンディングするより精確なデータが出る可能性がある。第二に語彙の一致率は明らかにデータの分類に役立つ。いかなるデータであってもよく使われる語は実は意外に少ない。だいたい一つの目安が1000語あたりで，この先1000語ずつ増やしていくと，明らかに一致率は下がる。問題はその一致率の下がり方で，急激に下がる場合もあれば，非常に緩やかな下がり方の場合もある。上位1000語程度までは，いわゆる機能語が多数を占めるので，それほど変化がなくて当たり前である。急激に一致率が下がる場合はいわゆる内容語が出てくる段階で，場合によってはまったく異なったジャンルのデータであるような印象を与えられる場合もあり，計量的にそれを裏付けることになる。逆に，新聞は新聞であって，三年前の新聞も昨日の新聞もそれほどたいした違いはないだろう，という印象は，一致率の下がり方が非常に緩やかである，という現象から裏付けられる。実際，今回の分析結果はそうなっている。

一致率を出すに当たっては，先ほど作成したlemmaの頻度リストを用いる。これをheadというコマンドで上位から，1,000，2,000，3,000，. . . etc. と切り出して，切り出したデータ同士を結合させ，再度sortをしてuniqで同一行を計算する。すると，一致している語（行）は２，一致していない語（行）は１がカウントされるので，２の行数を数えれば，どれくらい一致していたかがわかる。

以下のグラフにおいて，各折れ線グラフで，

1）一致率が低いほど，出現頻度にばらつきがある，

2）傾きが大きいほど（＝最大値と最小値の差が大きいほど）一致率が不安定である，

と考えられる。ジップの法則から考えれば，出現頻度の上位語ほど一致するし，下位語ほど一致率がばらけてくることは容易に想像がつく。しかし，データごとに下位に行っても一致率が落ちず，ばらけ（最大値と最小値の差）が少なければ，異なったデータから同じ頻度リストが得られたことになる。つまり，それらの頻度は信用できると言える。

6．分析結果

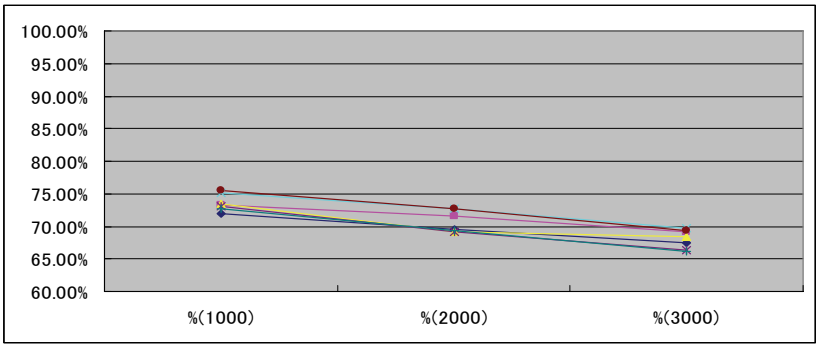
Ⅰ．Le Monde（CD-ROM＋Web）

データとしてはWeb上で販売されているオンライン版とCD-ROMで提供されているものを使用した。オンライン版は1日単位で購入できるので，一日単位で管理しており，CD-ROM版は1ヶ月単位でデータを管理しているので，とりあえずここではオンライン版を使用することにするが，CD-ROM版も後ほど使用する。

上でも述べたとおり，TreeTaggerの出力によってtoken数とlemma数が変更になる可能性があるので，POS解析済みのデータの中から，まず第一にtoken数が比較的近い隣り合った日付ごとのデータと，任意の日付同士のデータの分析を行ってみた。隣り合った日付同士と任意の日付同士でわけたのは，隣り合った日付同士の場合，報道される事件等につながりがあれば，それだけ類似の語彙が用いられる可能性が高くなるかもしれないと考えたからであるが，実際には，ほとんど有意の差は見られなかった。token数はおおむね5万語から8万語前後で，lemma数は5,000語から6,000語あたりである。従って，ランクの比較も3,000語で打ち切りにしたが（6,000語しかないのに5,000語比較したら「ぶれ」が大きくなるのは当然だからである），後ほどどれくらいぶれるかはHumanitéのデータでご覧に入れる。

図3　Le Monde　一日単位（隣り合った日付同士）

	1999/06/10–11	2000/02/05–06	2001/04/21–22	2002/11/26–27	2003/05/17–18	2004/12/03–04	2005/03/01–02
token1	76100	78808	72225	86094	80261	85532	85187
token2	78328	81690	87863	77004	66474	66770	57112
lemma1	6132	6494	6453	6272	6593	6932	6246
lemma2	6644	6640	6987	6226	5821	5769	5191
%(1000)	71.90%	73.30%	73.40%	75.10%	73.10%	75.50%	72.80%
%(2000)	69.45%	71.55%	69.20%	72.80%	69.25%	72.70%	69.40%
%(3000)	67.43%	69.20%	68.33%	69.67%	66.30%	69.30%	66.23%

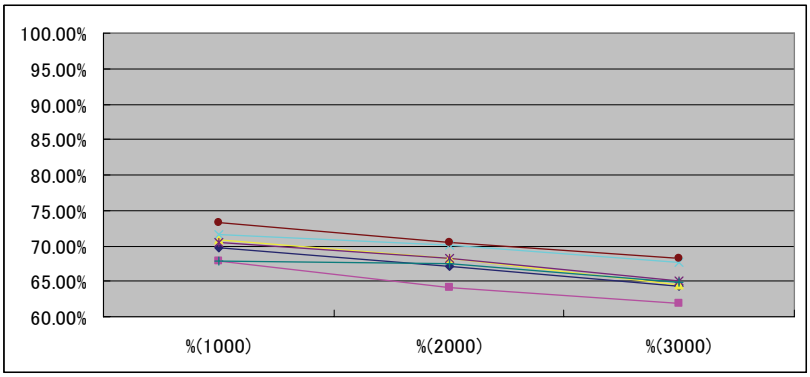


隣り合った日付同士で調べてみると1,000語あたりの一致率は70%を超えているが，3,000語レベルになると70%を維持するものはなくなる。一致率の最大値と最小値の差も5ポイント以上あり，グラフの傾きも大きいことがわかる。

次に任意の日動詞の一日分のデータで調べてみる。

図4　Le Monde　一日単位（任意の日付同士）

	1999/07/03-200511/20	2000/08/16-2004/11/07	2001/10/30-2003/09/24	2002/04/11-2003/04/06	2003/09/13-2001/11/10	2004/02/14-2001/06/03	2005/01/28-2000/05/25
token1	81448	46708	52465	79488	84680	82443	80935
token2	58820	74759	68126	83264	39395	79864	72229
lemma1	6757	5049	4941	6530	6623	6601	6313
lemma2	5402	6225	5915	6287	4367	6509	6072
%(1000)	69.80%	67.90%	70.80%	71.50%	70.40%	73.20%	67.80%
%(2000)	67.05%	64.10%	68.20%	70.10%	68.25%	70.50%	67.50%
%(3000)	64.23%	61.93%	64.47%	67.70%	65.10%	68.23%	64.90%



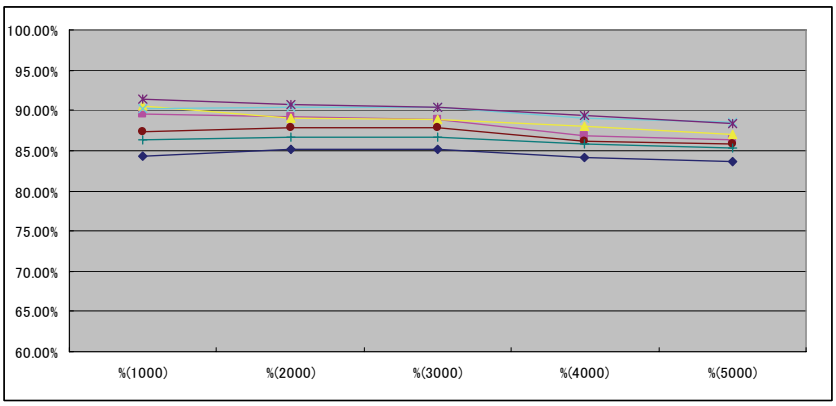
token数が10万以下程度のデータだと，1,000語あたりの頻度ランクの一致率が3,000までででおおむね5ポイント以上落ちていって，3,000語レベルで調べると，一致率は最も高く

て68%，最も低くて61%程度まで落ちてしまう。およそ半数近くが異なるということは，これくらいのサイズのデータから得られたランクは分析の元になったデータ次第で，ばらつきが大きいことになる。

なお，ご覧の通り，隣り合った日付同士と任意の日付同士では特に有意な差は見られない。若干隣り合った語同士のほうが一致率が高いが，10万語前後のデータでは，データ数の違いが過剰に結果に反映してしまうので，その影響を考えるとほとんど差はないと判断すべきであろう。他の新聞も含めてすべてのデータを同じ基準で分析した結果，日付ごとに限らず，月ごとであれ，年ごとであれ，隣り合っているか任意かは頻度ランクの一致率にまったく関係ないので，以下のデータにおいては任意の日月年のデータのみを提示することにする。以下に提示するデータにおいては，データの総量（特にtoken数。それに連動してlemma数）が増えるに従って，どのような違いが現れるかに注目していただきたい。

図5　Le Monde　月単位（任意の月同士）

	1999/10–2005/08	2000/07–2004/11	2001/06–2003/12	2002/03–2003/03	2003/01–2002/05	2004/08–2000/10	2005/04–2000/02
token1	1791003	1868993	1301932	2313076	1891513	1500871	1840958
token2	1226719	1917251	1846479	1902343	2275547	2403184	1473523
lemma1	21816	22486	20052	22647	21747	20581	20928
lemma2	18900	21615	21476	21324	22264	23359	20928
%(1000)	84.30%	89.50%	90.60%	90.20%	91.40%	87.40%	86.30%
%(2000)	85.20%	89.20%	89.05%	90.35%	90.70%	87.80%	86.75%
%(3000)	85.17%	88.83%	88.80%	90.33%	90.43%	87.77%	86.73%
%(4000)	84.08%	86.85%	88.05%	89.00%	89.43%	86.13%	85.88%
%(5000)	83.58%	86.34%	86.98%	88.52%	88.38%	85.74%	85.34%

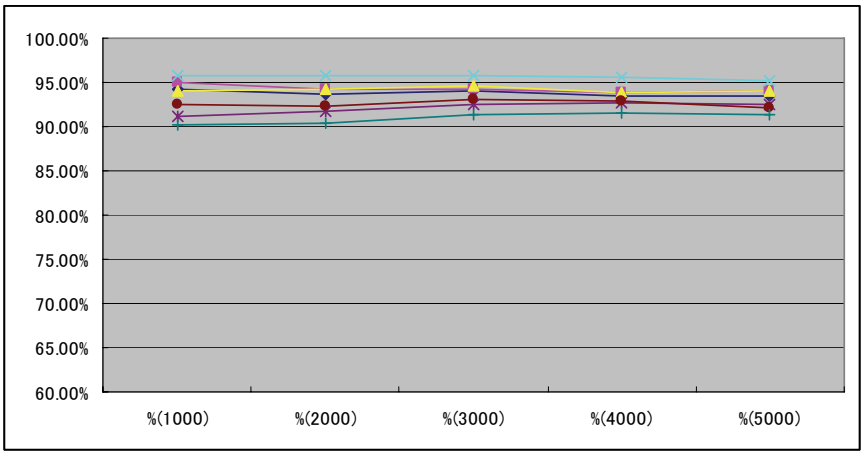


月ごとのデータになると，一日ごとのデータより，25倍を上限として（Le Mondeは毎

週月曜日が休刊のため），目安としておよそ10～20倍程度多い分量になる。token数で言うと，200万語未満くらいだが，一日ごとに比べると，格段に一致率があがっていることがわかる。また，一致率の最大値（おおむね1,000語レベルの値）と最小値（おおむね5,000語レベルの値）の差も３ポイント程度にまで下がっている。つまり分析の元になるデータの量が増えれば，明らかに一致率も上がり，データとしての信頼度も上がっていると考えられる。

図6　Le Monde　年単位（任意の年同士）

	1999–2001	2000–2002	2001–2003	2002–2004	2003–2005	1999–2004	2000–2005
token1	10748472	25084370	17914629	24347171	21351886	10748472	25084370
token2	17914629	24347171	21351886	23079798	20275067	23079798	20275067
lemma1	33250	38199	36041	37461	36193	33250	38199
lemma2	36041	37461	36193	36548	35288	36548	35288
%(1000)	94.20%	95.00%	94.00%	95.70%	91.10%	92.50%	90.20%
%(2000)	93.70%	94.30%	94.20%	95.70%	91.75%	92.30%	90.40%
%(3000)	94.00%	94.30%	94.63%	95.70%	92.43%	93.03%	91.33%
%(4000)	93.45%	93.93%	93.88%	95.58%	92.60%	92.80%	91.60%
%(5000)	93.40%	94.00%	93.96%	95.16%	92.58%	92.04%	91.26%



最後に年ごとのデータの比較である。年ごとになると，一致率がさらに上がっていることがわかる。月ごとだと90%あたりが最大値だったのが，年ごとになると最大で95%程度の一致率になっている。しかも最大値と最小値の差も１ポイント程度で，逆に上昇するケースも見える。年ごとのデータではtoken数は上が2500万語，下が1000万語くらいで，これだけの語数になると，年ごとの違いはほとんどないことがわかる。ただし，これが上

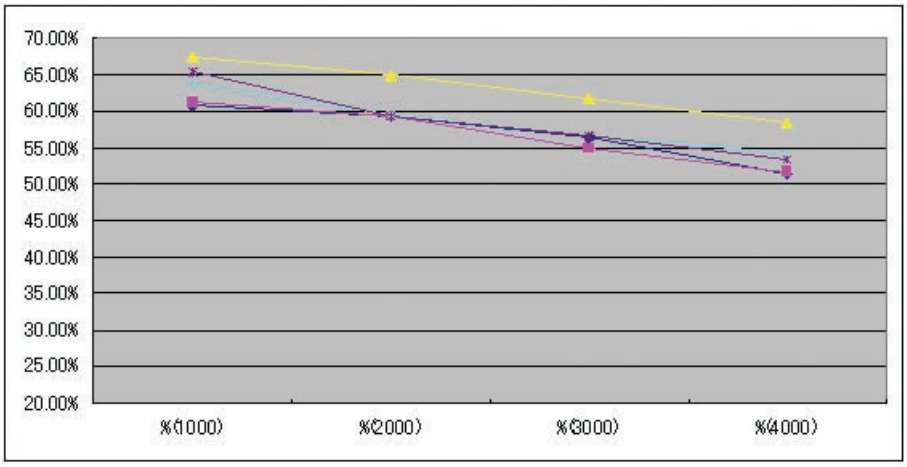
限ではなさそうなことは，後ほど検証する。

II．Humanité（Web）

次に，Humanité紙について見てみよう。Humanitéは1990年以降にWeb上で公開した記事はすべて無料で見ることができる。Web上で記事は1日ごとに提供されているので，データも1日ごとにひとまとめにして管理している。ただし，一日に提供されるデータの分量はさほど多くはなく，token数で3万語前後である。

図7　Humanité　日単位（任意の日同士）

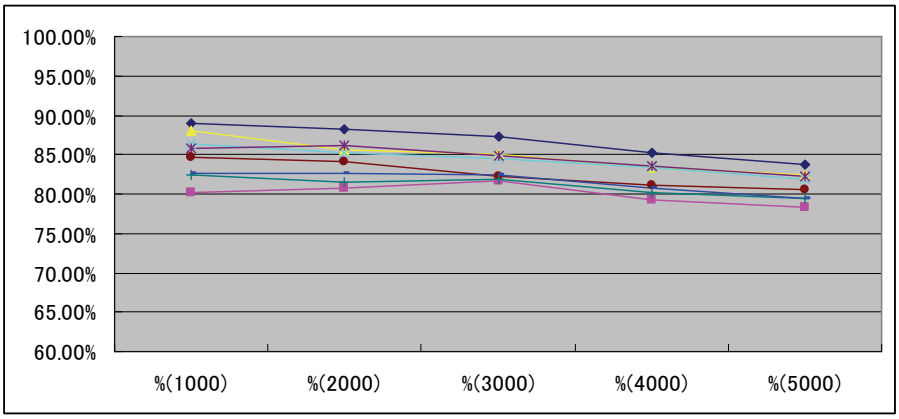
	1990/12/22–2005/01/07	1991/01/07–2004/10/27	1992/09/09–2003/10/18	2002/04/22–1994/02/16	2004/04/05–1992/03/21
token1	53334	35491	37397	47001	29448
token2	32231	28883	69880	35378	39025
lemma1	4287	3768	3960	4671	3738
lemma2	3477	3810	6262	4230	4244
%(1000)	60.70%	61.30%	67.40%	63.80%	65.40%
%(2000)	59.25%	59.40%	64.80%	59.50%	59.40%
%(3000)	56.43%	55.07%	61.60%	56.57%	56.67%
%(4000)	51.43%	51.70%	58.35%	54.20%	53.33%



データ総量が少ないとこのような結果になるといった見本のようなグラフである。一致率は最大でも70%に届かないし，最大値と最小値の差も10ポイントに達するとなると，token数が4～5万語程度のデータでは頻度ランクは出せないと言わざるを得ない。

図8　Humanité　月単位（任意の月同士）

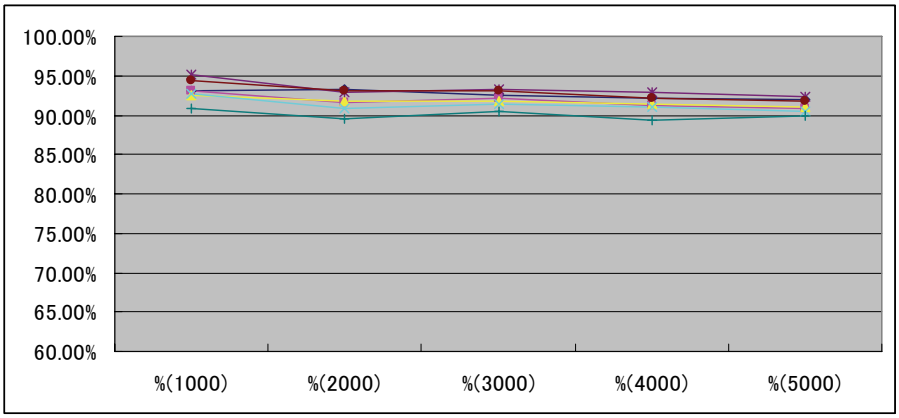
	1998/10–1997/09	1999/08–1996/11	2000/11–1995/08	2001/07–1994/12	2002/09–1993/10	2003/12–1992/07	2004/08–1991/06	2005/04–1990/10
token1	801983	628100	913924	740322	641529	662593	600049	929930
token2	835908	869852	688445	807325	838458	720834	670602	758811
lemma1	15434	17283	17232	16487	14708	14969	14829	15965
lemma2	15409	16020	15785	15796	15751	15305	14326	15130
%(1000)	88.90%	80.20%	88.00%	86.40%	85.80%	84.70%	82.40%	82.60%
%(2000)	88.15%	80.70%	85.70%	85.30%	86.10%	84.10%	81.55%	82.70%
%(3000)	87.30%	81.77%	85.00%	84.47%	84.77%	82.17%	81.83%	82.43%
%(4000)	85.28%	79.33%	83.30%	83.38%	83.60%	81.08%	80.13%	80.83%
%(5000)	83.66%	78.38%	82.42%	81.90%	82.22%	80.54%	79.46%	79.50%



月ごとのデータになると，token数が100万語弱といったところで，このあたりからようやく安定し始める。ただし，この程度の語数でも一致率は90%に達しないし，最大値と最小値の差は５ポイント前後開いている。つまり，この程度の語数でも，どのデータを取るかによって，頻度リストの内容が異なってしまう，ということである。

図9　Humanité　年単位（任意の年同士）

	1990–1992	1990–1993	1990–1994	1990–1996	2000–2002	2000–2004	2000–2005
token1	6366114	6366114	6366114	6366114	11295814	11295814	11295814
token2	10580395	9518041	9452153	10167683	9556092	10660579	8900663
lemma1	27224	27224	27224	27224	31506	31506	31506
lemma2	29875	30104	29671	30508	30004	30894	28627
%(1000)	93.00%	93.00%	92.60%	92.80%	95.20%	94.40%	90.90%
%(2000)	93.25%	91.65%	91.75%	90.85%	92.85%	93.00%	89.50%
%(3000)	92.43%	92.13%	91.80%	91.43%	93.20%	93.03%	90.53%
%(4000)	92.23%	91.20%	91.35%	90.98%	92.90%	92.13%	89.40%
%(5000)	91.96%	90.88%	91.02%	90.48%	92.38%	91.82%	89.86%



年ごとになるとtoken数はおよそ1000万語となり，一致率も最大で95%，最小でも90%弱程度の範囲内に収まり，最大値と最小値の差もおおむね3ポイント程度の範囲内に収まっている。

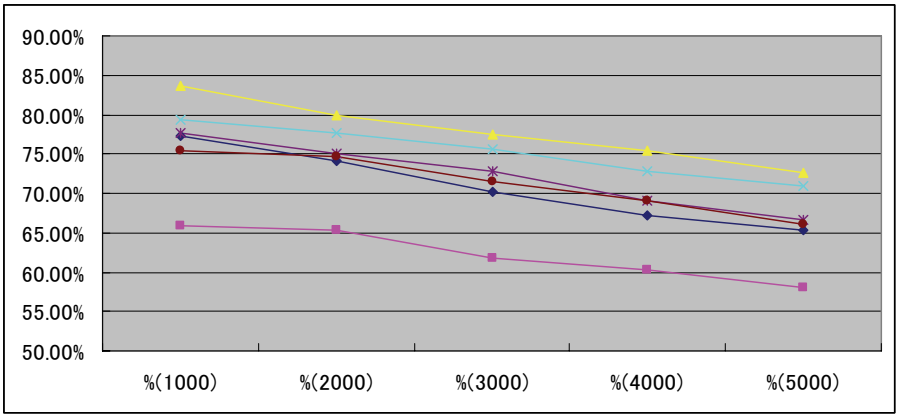
以上，二つの新聞で語彙頻度リスト作成のためにはどの程度の語数が必要か，おおむね判然としていると思われるが，もう一つ例を挙げておこう。Le MondeもHumanitéも中央紙か地方紙かで分類すればパリを発行の中心においている中央紙と考えられるので，地方紙から一つ例を挙げておこう。アルザス地方の有力な新聞であるL'Alsace Le Pays紙（以下，Alsaceと略）を例にして，上記2紙と同じ分析を試みる。

Ⅲ．Alsace（Web）

Alsaceは1996年ごろから2006年まではほとんどすべての記事をWeb上で公開していた。2007年現在では一部の記事は無料で読めるが，大部分は有料となっている。

図10　Alsace　日単位（任意の日同士）

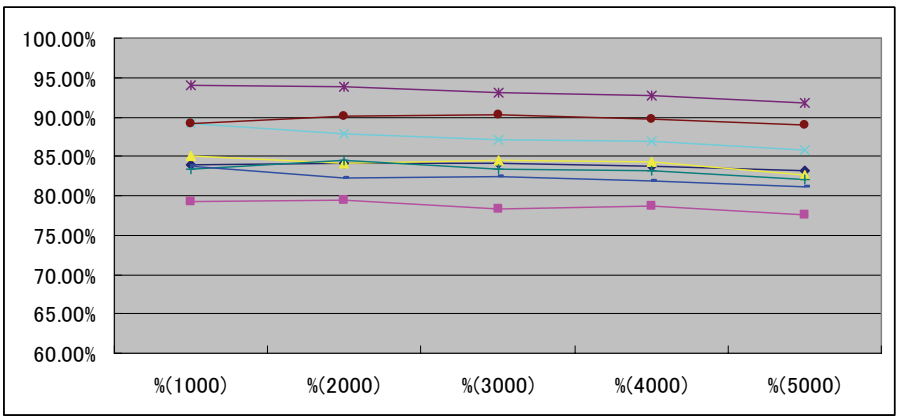
	1996/08/06–1996/09/26	1998/02/07–1998/08/19	2001/06/16–2003/05/15	2000/12/19–2002/01/09	1997/01/30–2005-06-28	1996/09/21–2005/10/28
token1	87212	60741	196661	148291	134728	98966
token2	127885	60638	197978	152406	149568	155879
lemma1	5862	4788	8750	7695	7218	6086
lemma2	6736	5098	8746	7669	7718	7884
%(1000)	77.20%	65.90%	83.70%	79.40%	77.70%	75.50%
%(2000)	74.10%	65.30%	79.90%	77.65%	75.00%	74.65%
%(3000)	70.13%	61.80%	77.57%	75.57%	72.77%	71.53%
%(4000)	67.18%	60.23%	75.50%	72.73%	69.13%	69.03%
%(5000)	65.42%	58.12%	72.56%	70.92%	66.64%	66.06%



このデータも，Humanitéの日ごとのデータと同じく，非常にばらつきが大きい。一つだけ飛びぬけて一致率が低いデータは，明らかに元のデータにおいても，token数が少ないデータで，元データが小さいとどうしてもこのような結果になってしまう。

図11　Alsace　月単位（任意の月同士）

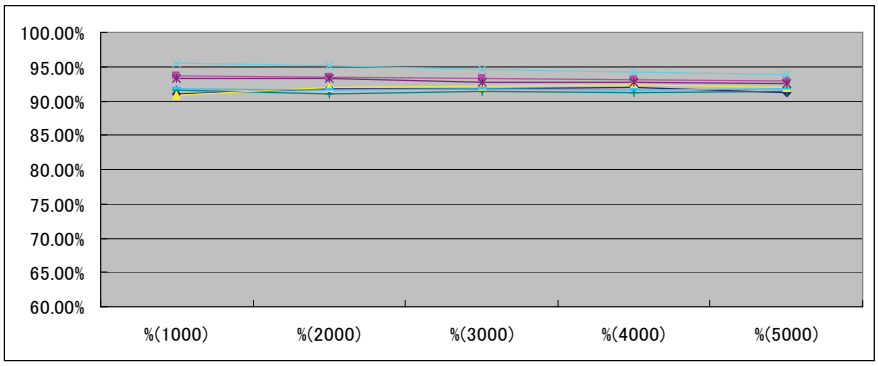
	1996/07–2005/12	1997/08–2004/11	1998/07–2003/12	1999/10–2002/09	2000/12–2001/12	2002/05–2000/07	2003/02–1999/05	2004/06–1998/01
token1	1079800	362764	795830	1267366	4038240	4647252	4835103	4315122
token2	4516636	4581296	4069264	2540111	4405307	3280407	647373	478749
lemma1	15666	10058	14541	16054	22770	21921	23246	22849
lemma2	22895	22977	22337	19926	23067	23392	12554	12001
%(1000)	84.00%	79.30%	85.10%	89.20%	94.10%	89.20%	83.30%	83.70%
%(2000)	84.15%	79.50%	84.20%	87.85%	93.90%	90.15%	84.55%	82.25%
%(3000)	84.20%	78.33%	84.47%	87.07%	93.03%	90.33%	83.30%	82.50%
%(4000)	83.80%	78.68%	84.33%	86.90%	92.78%	89.70%	83.20%	81.85%
%(5000)	83.26%	77.56%	82.68%	85.72%	91.74%	89.00%	82.02%	81.16%



このデータもばらつきが大きいように見えるが，元のデータをよく見ると，データサイズに違いがある場合，すなわち小さなデータが混じると，どうしてもばらつきが大きくなる。たとえば最小値が80%を切ってしまう1997/08–2004/11のデータでは1997年8月分のデータ量が少ない。これはバカンスシーズンであったことに加えて，TreeTaggerでの解析がすべて終わらなかったファイルが多数存在したため，token数が小さくなってしまった例だが，一つでもこうした容量の小さなデータが混じると，即座に一致率に反映してしまう。

図12　Alsace　年単位（任意の年同士）

	1996–2005	1997–2000	1998–2000	1999–2001	2000–2004	2002–1998	2003–1997
token1	15057577	32015700	8482451	10156117	26713374	48322717	37099872
token2	48753069	26713374	26713374	52352789	51359164	8482451	32015700
lemma1	30088	34681	27479	26942	33542	36949	35005
lemma2	36206	33542	33542	37485	36371	27479	34681
%(1000)	91.10%	93.60%	91.60%	90.90%	95.50%	91.70%	93.20%
%(2000)	91.75%	93.40%	91.00%	91.95%	95.10%	91.50%	93.30%
%(3000)	91.77%	93.27%	91.40%	92.07%	94.53%	91.83%	92.80%
%(4000)	92.03%	93.15%	91.30%	92.30%	94.18%	91.63%	92.80%
%(5000)	91.16%	92.92%	91.42%	91.90%	93.80%	91.74%	92.52%



Le MondeやHumanitéと同じ特徴を示している。語彙頻度リスト作成などにあたっては，明らかにデータの総量が品質を左右していると言える。

7．語彙頻度リストと年次による差について

以上のデータから明らかなとおり，どの任意の年度同士を比べても，一致率に影響があ

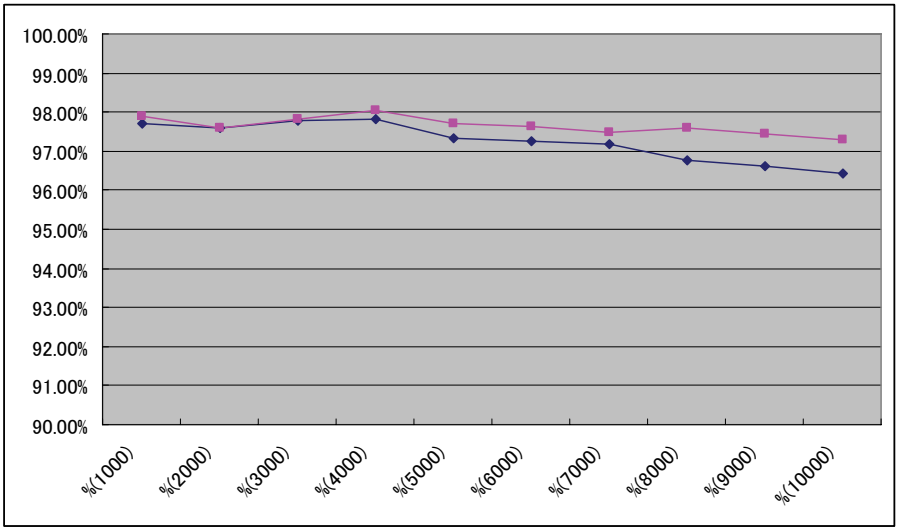
るのはデータの総量だけで，年次による差はまったくといってよいほど現れない。ただし，だからといって年次による差はまったくないとは断言しがたい。

我々の日常の経験知からしても，明らかに語の流行り廃りはある。ただそれらは我々が感知するほど全体としての流通量は多くないこと，従って，もしそれらを検知するような結果を出したい場合には，逆にデータの区分を工夫する必要がある。先にも見たとおり，総量が小さければ，小さな差が顕著に一致率に現れる。今度は逆に一致していない語の方を問題にする必要がある。

ただし，それらのいわゆる流行り廃りは所詮は局所的な現象に過ぎないことも以上の調査から明らかである。たとえば局所的には大統領選挙があったり，戦争があったりして，特定の語彙が集中して用いられることはあっても，それが数千万語レベル以上のデータになると，たちまち単なる局所的な現象となってしまう。

図13　Le MondeとHumanitéの偶数年と奇数年を合算したデータの一致率

	lm-pair-impair	hum-pair-impair
token1	128163238	77198008
token2	130993088	63005479
lemma1	36831	29870
lemma2	35539	27473
%(1000)	97.90%	97.70%
%(2000)	97.60%	97.60%
%(3000)	97.83%	97.77%
%(4000)	98.05%	97.83%
%(5000)	97.72%	97.34%
%(6000)	97.62%	97.27%
%(7000)	97.49%	97.17%
%(8000)	97.61%	96.76%
%(9000)	97.44%	96.62%
%(10000)	97.30%	96.41%



Humanitéでは1万語まで一致率を比較しても1ポイント程度，Le Mondeに至っては0.5ポイントほどしか下がらない。少なくともランク1万位くらいまではデータは98%ほどの割合でほぼ一致しているわけである。

以上の結果から，語彙頻度リストを作成するような場合には，さし当たってはデータの分量は多ければ多いほどよいことになる。ただし，その総量がどれくらいであればよいかは現時点では何ともいえない。数年分まとめたLe Mondeのデータのtoken数は1億語を超えている。Humanitéも7700万語と6300万語である。lemmaの数はおおむね3万語程度でほぼ上限に達している[12]。これで97～8%ほどの一致率が得られることから，これで一致率も上限と見なしてよいのか，さらに語数を増やせば限りなく100%に近づくと考えるべきかは，実際にデータを分析してみないとわからない。

8．語彙頻度リスト作成に必要な語数はいくらか

はじめにでも述べたとおり，コーパスから得られる知見にはさまざまなものがある。調査目的も調査方法も千差万別である。サンプリングで十分な場合もあれば，サンプリングではないほうが品質が高い場合もありうる。

語彙頻度リストを作成するときに，最も大きく影響するのは分析元のデータのlemma数である。先にも見たとおり，lemma数が5,000しかないのに，そこから半分以上の上位3,000語のランクを取ればゆらぎが大きくなるのは当たり前である。データをご覧いただければおわかりのとおり，token数が10倍に増えても，lemma数は10倍も増えない。lemma数

には上限があるからだ。もちろん，データの量が増えればそれだけノイズも増えるので一見したところlemma数は微増しているように見える。しかし，lemma数はおおむね3万語程度で上限に達している。分析元のtoken数を増やせばlemma数がさらに5万語にも7万語にも増えるとは考えにくい。

上記3つのデータを見る限り，どうやらtokenが1,000万語を越えたあたりで（すなわち1年分程度のデータで）lemma数はおおむね上限に達する。しかし，一致率はなお上昇傾向にあるということは，lemmaが出尽くした後で，そのlemmaの出現がこなれてくるには，もう少しtoken数が必要であると考える必要がある。たとえば100万のデータから1万を選んで調査することもサンプリングなら，1兆のデータから1億を取り出して調査することもサンプリングであって，言語の場合，1兆のデータであっても果たしてそこから1億取り出すだけでサンプリングになるのかどうか不明である。もし1兆から1億取り出すのでさえ不足しているかもしれないという疑いがあるのなら，100万から1万取り出すのではあたかも100のデータから1つを調べてサンプリングしたと思っているのと同じようなことになりかねない。現代フランス語や新聞コーパスといっても，その限界がどこにあるかはまだ明らかにされていない。どこまで調べればよいのかといった模索はまだ続ける必要がある。

9．おわりに

以上，語彙頻度リストを作成するためには今回の分析を見る限り，10万語より100万語，100万語より1億語の方が明らかに一致率も高いし，最大値と最小値の差も小さい。すなわち，元データの規模が大きくなればなるほど，そこから得られる語彙頻度リストの質が高くなると考えられる。巨大な語数（token数）が必要なのは，lemmaの出現状況が安定する必要があるからである。

ただし，すでに読者諸賢においてはお気づきの通り，今回はリストの内容について検討を加えていない。今回，リストの内容について検討しなかったのは，いくつか理由がある。その最大の理由は，何がフランス語で何が借用語（外来語）で何が合成名詞なのかの扱いを判定することが難しいことである。cocktail, football, golf, handicap, jazz, match, meeting, stand, stock, stopper, week-end, etc.　これらのうちのどれを外来語とみなし，どれをフランス語（として定着した語）とみなすか。Partie Socialeを1語と考えるか，2語と考えるか。JeanやMarieと同じく固有名詞とみなしてリスト作成上排除して考えるか，分解して普通名詞の集まりとみなすか等，にわかに判断をつけがたかった。今回の結果にはこれらの語はすべて含まれている。次回は語彙頻度リストを作成するに当たって，個々の語の選定にどのような問題があるかについて論じたいと思う。

謝辞

本研究は2006年度愛知大学研究助成　共同研究B（助成番号B-30）の成果の一部である。

参考文献

Stubbs, Michael (2002) *Words and Phrases : Corpus Studies of Lexical Semantics*, Oxford, Blackwell Publishing Ltd.
田中春美（編集主幹）（初版 1988，第5版 1997）『現代言語学事典』東京，成美堂
「日本語の計量研究法」『日本語学』（2001）東京，明治書院.
伊藤雅光（2002）『計量言語学入門』東京，白水社.
小池生夫（編集主幹）（2003）『応用言語学事典』東京，研究社.
鈴木良次（編集委員長）（2006）『言語科学の百科事典』東京，丸善株式会社
マイケル・スタッブズ（2006）『コーパス語彙意味論　語から句へ』東京，研究社.

注

1 伊藤（2005），まえがき── V。
2 もちろん，新聞データであっても，「A新聞の2006年度の朝刊の東海版」くらいまで絞ればクローズドになるかもしれないが，本稿で扱っているような語彙頻度リスト作成などに当たってはこの程度のデータ量では足りない。
3 そもそもサンプリング調査は母集団の範囲がわかっている場合に用いるのであって，母集団の規模さえわからないのにサンプリングなど不可能である。ただし，サンプリングに必要な最低数は予測可能ではないかというのが本稿の論証で明らかになる。
4 今回はAchim Steinが作成したパラメータファイルを用いたが，出力結果を分析してみると，パラメータファイルを作成するのに十分な量のトレーニングファイルではなかったのではないかと思われる部分が多々ある。TreeTaggerはパラメータファイルを作成するためのプログラムも用意されているので，もう少し大きなトレーニングファイルを読み込ませてパラメータファイルを作り直す予定である。
5 あらかじめわかっているlemma化に関する問題点，すなわち，いくつかの誤lemma化とunknownに分類されてしまった語については手動で微調整を行った上で，分析に用いた。
6 小池（2003），p. 657。
7 田中（1988），p. 60
8 鈴木（2006），p. 316。
9 特定のコーパスからしか得られない高頻度語グループが想定できる。たとえば幼児向け書物から作成された高頻度語がたとえば大人向けの文学データや新聞データとどの程度共通部分があるかは疑わ

しい。しかし，幼児向け書物で使用された語彙であれば，たとえ他のコーパス群において使用頻度は低くても，すでに獲得された周知の語彙であると想定できる。その観点においては，もちろん即断はできないが，かなり基礎性は高そうである。いずれ幼児期獲得語彙の分析には着手する予定である。

10 ジップの法則と呼ばれる法則を確立したのはアメリカの言語学者の George Kingsley Zipf（ジョージ・キングスレー・ジフ）で，その名をとってZipf's lawと呼ばれる。なお，ジップの法則は最近では「ロングテール（long tail)」と呼ばれ，Webビジネスのモデルを説明するのにもよく使われる。Amazon.comのように，従来「死筋」と言われていた，めったに売れない本でもそれが数万冊，数十万冊と集まれば，たとえ年に1冊しか売れなくても十分ビジネスとして成り立つというわけだ。これはロングテールのテール部分に注目したビジネスモデルだが，実はコーパス分析においても，テールの部分にはテールなりの意味があり，いずれ稿を改めて論じたいと思う。

11 グラフ描画のために上位50語までしか用いていない。グラフをご覧になればお分かりの通り，上位50語を過ぎれば，出現率は0.1%以下となり，以下，数万語まで限りなく0%に近づくだけである。また，どれがどのデータの線なのかわからないと思うが，要するにどれがどれだかわからないほど，分析に用いる元のデータに関係なく，ジップ曲線はほとんど同じということである。ちなみに，それぞれのデータにおいて分析に用いたtoken数は以下のとおりである。

ALS (L'Alsace Le Pays) 192,146,869 mots
bret (Le Télégramme de Brest) 187,567,238 mots
canoe (Canadian Online Explorer) 1,914,158 mots
dna (Les Dernières Nouvelles d'Alsace) 88,254,359 mots
hum (Humanité) 124,313,168 mots
lacote (La Côte) (Suisse) 1,749,979 mots
latribune (La Tribune) 59,785,163 mots
lepoint (Le Point) 17,639,038 mots
lesoir (Le Soir) (Belgique) 127,264,193 mots
libe-CD (Libération CD-ROM) 64,360,484 mots
libe-web (Libération Web) 10,564,031 mots
lm-CD (Le Monde CD-ROM) 106,961,292 mots
lm-web (Le Monde Web) 124,846,007 mots
lmd (Le Monde Diplomatique) 10,953,025 mots
yonne (l'Yonne Républicaine) 8,964,778 mots

（図版のサイズの都合上，凡例部分で隠れているが，lmdの下にyonneがある）

12 Le MondeとHumanitéの奇数年分と偶数年分の比較のlemma数は平均値である。