『コーパスに基づく言語学教育研究報告』No.3（2009）

N-gram の手法による フランス語の基本的「定型表現」の抽出

－Le Monde と Corpatext を資料として*－

藤村　逸子
（名古屋大学　国際開発研究科）

要　旨

本稿は，フランス語の基本的「定型表現」のリストを日本人学習者のために N-gram の手法によって抽出することを目的としたパイロットスタディである。まず，方法論上の議論として，新聞テクスト（Le Monde 誌 3 年分）と 2700 種の文学・学術テクスト（Corpatext）からなる総語数 1 億語のコーパスの構成を説明し，コーパスに形態素解析を施したうえで N-gram をとる方策に関して詳述する。N-gram の手法で抽出できるのは，「定型表現」のうち「使用頻度の高い表現」であって，「意味的融合のある表現」とは限らない。次に，抽出した表現の観察を行う。両テクストに超高頻度で現れるのは，フランス語のあらゆるテクストに出現する「機能的定型表現」と認められる。一般的な「定型表現」は Le Monde よりも Corpatext から得られやすい。前者は一つの特殊なテクストであるのに対し，後者は種々のテクストの混合だからである。N-gram の手法は，基本的「定型表現」のリスト作成に有効だが，いかなるリストを目的とするかに応じた綿密な計画が必要とされる。

1. はじめに

自然な文を作るためには，単語と文法の知識だけでは十分ではなく，連語の知識が重要であるということが，大規模なコーパスを使った言語研究，言語習得研究の発展とともに，注目されている。言語学では，連語を対象とする phraséologie という分野に関心が寄せられ，語彙論，形態論，統語論，文論，談話やテクスト分析などの従来の枠組みを横断するものとして，研究されている。Stubbs は，次のように言う。A central finding of corpus studies is that the units of language use are longer phraseologycal units. (Stubbs, 2007 :144)。連語の研究は，英語を対象としたものに特にその蓄積があるが，フランス語に関しても，2005 年秋に，ベルギーの Louvain カトリック大学で開かれた*"Phraséologie 2005"*という名の会議が成功

* 本研究の一部は，文部省科学研究費・基盤研究(C)（課題番号：20520379）の助成によって行われた。

をおさめ，*Cahiers de l'Institut de linguistique de Louvain* vol. 31, no.2-4（2005）が，*La phraséologie dans tous ses états. Actes du colloque "Phraséologie 2005"*と題して，300 ページを超えるページ数を会議の論文集に割いている[1]。また, 2006 年には, *Langue Française*, no.150 (2006) が，*Collocations, corpus, dictionnaires* という特集において，この問題を扱っている。

言語習得研究においても，この分野はますます重要視されている。言語の使用者は，一つ一つの「単語」を文法規則に従って組み合わせて，オリジナルな発話を自由に構成しているのではなく，単語の組み合わせには習慣的に定まった傾向があり，そのような大きなユニットの組み合わせおよび，ユニットの微調整によって，発話を構成するという観点が主流になっている。「単語」を数多く知り, 文法テストで満点をとる上級の外国語学習者が，その能力だけで，母語話者と同じような言語使用ができるわけではないことは我々が日常的に気づくことである。Alain REY は，*Dictionnaire des expressions et locutions*（以下では *DEL* と略記）の冒頭で次のように言っている。“Parmi les éléments de la langue qu’il faut acquérir pour s’exprimer figurent non seulement les mots, mais aussi des groupes de mots plus ou moins imprévisibles, dans leur forme parfois, et toujours dans leur valeur ”(Rey & Chantreau, 2007 : p.v) .

本稿は，このような背景の中で，フランス語で頻繁に用いられる基本的な連語，すなわち，定型的表現（以下，「定型表現」）のリストを日本人学習者のために作成することを目的として，理論的および方法論的な検討を行う。それに際しては，具体的に，コーパスを用いてリストを作るという作業を行う。すなわち，約 1 億語からなる大規模コーパス（3 年分の Le Monde（合計 6486 万語）と Corpatext（文学・学術テクストなどからなる無料のコーパス, 3670 万語)）を用い, 以下で説明する N-gram の手法を使って，コーパスから「定型表現」を抽出する。この作業を行う中で生じる種々の問題を考察し，どのようにすればよいリストを得ることができるかを検討する。

本稿において N-gram とは, コーパスから採取した n 個の語の連続のことを言う[2]。また，N-gram を採取すること，およびその方法を N-gram の手法という。たとえば，Le Monde 誌（1999 年）1 年分をコーパスとして，4-gram（4 個の語の連続）を抽出すると，上位 10 位は図 1 のとおりになる。数字は出現頻度である。各単語は，形態素解析を経て，辞書形に変換されている[3]。

[1] これはフランス語で発表された論文のみを収録しており, 他に, 英語で発表された論文のための論文集がある。*Phraseology: an interdisciplinary perspective*. S. Granger & F. Meunier (ed). John Benjamins, 2008.
[2] 文字を単位として N-gram の手法を使うこともできる。
[3] 実際には，この他に，<数詞＋million de euro>などの数字を含む表現が上位にいくつか含まれているが，ここでは省いている。

4935	il ne y avoir
4300	ce ne être pas
3291	de plus en plus
3290	il se agir de
2931	le président de le
2911	il ne être pas
2682	ne y avoir pas
2673	il y avoir un
2600	le ministre de le
2546	ne être pas le

図 1：Le Monde（1999）の最高頻度の 4-gram

N-gram は大規模コーパスから，N 個の語の連続を抽出するという単純な手法によって得ることができる。N-gram の抽出のために，特別のプログラムは必要ではないし，統計学的知識も専門的な情報工学の知識も必要ではない[4]。この方法によって，フランス語で頻繁に用いられる語の連続のリストが目の前に現れるのを見ることは刺激に満ちた経験である。一見して，N-gram の手法は言語学，言語習得研究のさまざまな分野に応用が可能であると思わせる[5]。しかし，このようにして抽出された，生の N-gram のリストは，あまりに多様なものを含んでいる。本稿の目的である「定型表現」に戻るなら，N-gram のリストがそのまま，フランス語の基本的な「定型表現」のリストに一致するわけでないことは，上の例からも明らかである。「定型表現」を抽出するためには，分類や選別を行う必要がある。

特に，次の二つの点を考察しなければならない。第一点目は，特定のコーパスから抽出された高頻度の N-gram の中から，言語学的に有意の「定型表現」を適切に取り出すためにはどのようにすればよいかという問題である。特定のコーパスから得られた高頻度の N-gram のリストには，直感的に「定型表現」ではないと思われるものが多量に含まれている。可能な限り恣意的でない方法によって，適切な「定型表現」のみを残し，不要な連続を排除せねばならない（図 2）。この作業を行うためには，「定型表現」の定義を考える必要がある。この問題はまた，存在するデータから不要なものを排除する仕方にかかわるので，方法論的，あるいはテクニカルな議論とも考えられる。

[4] N-gram 抽出のプログラムも作成され発表されているようであるが，筆者は未見である。
[5] 日本語に関しては，近藤泰弘，近藤みゆき（2004）において紹介されているように，N-gram の手法を活用した業績が上げられている。

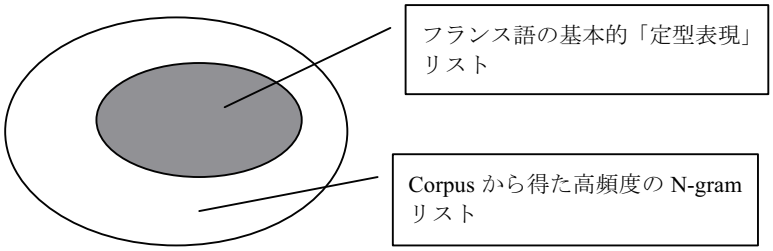

図 2：あるコーパスの N-gram リストとフランス語の基本的「定型表現」リストの関係

第二点目は，特定のコーパスから N-gram の手法によって抽出された高頻度の N-gram のリストが，フランス語の基本的な「定型表現」を網羅しているとは限らないという議論に関するものである。ある特定のコーパスにおける高頻度の N-gram が，フランス語全体の基本的な「定型表現」を含んでいる保証はない。また，単純に高頻度であるという理由によって，「定型表現」を認定してよいかという点も考察せねばならない。すなわち，本稿の方法によって抽出される「定型表現」の限界を認識することが必要である。この問題は，本稿のデータの中には存在しない，欠けたデータに関する議論なので，理論的に，あるいは，他に存在するリストに基づいて考察することになる（図 3）。

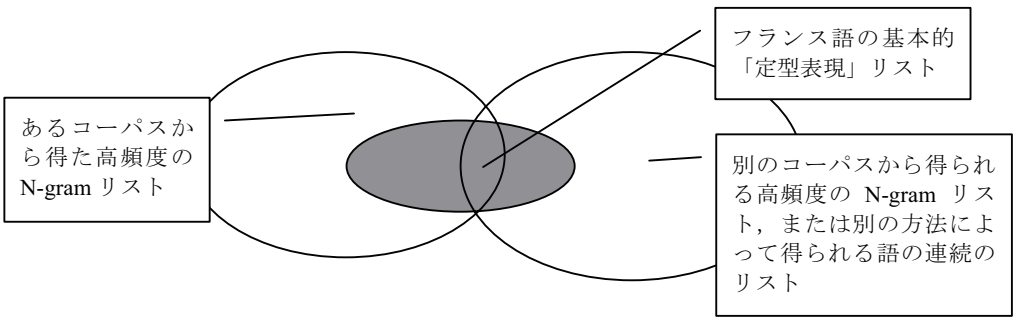

図 3：複数の高頻度の N-gram リストとフランス語の基本的「定型表現」のリストの関係

以上の二つの点の検討にあたって本質的な問題は，基本的な「定型表現」とは何かということである。もしも，「定型表現」のリストがすでに存在するのであれば，研究は方法論的なものになろう。リストを作成するための効率のよい方法を検討すればよい。しかし，そのようなリストは存在しないので，この研究は試行錯誤的なものになる。「巨大コーパス」と N-gram の手法を利用し，これまでに筆者が行った研究，および，種々の文献的知識に基づいて，適切と認められる基本的「定型表現」のリストの作成を試み，また，そのリストの限界について認識することが本論文の課題である。

2. 「定型表現」と N-gram

2.1. 「定型表現」

「定型表現」とは，複数の語の組み合わせがひとつの言語学的なまとまりを構成している場合と定義できる。「単語」はフランス語の場合，書記法上のスペースによって定義されているが，言語学的な単位は「単語」に一致するとは限らない。単位としての「形態素」は「単語」の内部に認定されたが，逆に，複数の「単語」が集まって構成される単位が「定型表現」である[6]。どのような言語にも「定型表現」は存在し，話し言葉においても，書き言葉においても重要な役割を果たしている（Sinclair 1991）。フランス語にも，たとえば，<jeune fille> や <petite annonce> のような「複合語（mot composé）」，<en face de> や <tout de suite> のような「(前置詞，副詞）句 (syntagme (prépositif, adverbial))」，<avoir besoin de> や <aller bien> のような「熟語」，<il n'en reste pas moins que + 節>などの「構文（construction）」，<Chacun son goût> のような「慣用句（expression）」，<La nuit porte conseil> や <prendre la mouche> のような「イディオム（idiomatique）」，<passer un examen>, < entraîner des conséquences > などの「コロケーション（collocation）」などといわれるさまざまなタイプの「定型表現」が存在するが，それぞれのタイプに，明確な定義が与えられているわけではなく，均一なものとして認知されているとは考えられない。言語学的単位としての「定型表現」の認定を助ける，母語話者の直感以外の方法に関しても，研究は端緒についたばかりである[7]。

フランス語の辞書には，母語話者用，学習者用を問わず，「定型表現」に関する記述がある。その意味において，「定型表現」のリストはすでに存在している。どのような辞書にも，単語ごとの見出し語の下位には，「定型表現」が挙げられている。また，たとえば *Le CD-ROM du Petit Robert* には，単語ごとの見出し語の他に，膨大な数の locution（＝「定型表現」）ごとの見出しが設定されている。もちろん，*DEL* のような，「定型表現」専門の大きな辞書も存在する。

しかし，これらの辞書の「定型表現」のリストや記述は，日本人学習者が習得するためや，実際に使用するために参照するにあたって，十分に役立つとは言えない。その理由は，辞書の設計などのせいではなく，「定型表現」の習得は，第二言語学習者にとって，単語や文法の習得よりもさらに困難であるという問題に由来すると思われる[8]。母語話者むけの単語辞典に比べて，母語話者むけの「定型表現」辞典が，学習者の役に立たない度合いには著しいものがある。筆者のようにフランス語暦が 30 年になろうとする者にとっても，*DEL* に収録されている「定型表現」の大半は，記憶の片隅にもないものである。Gougenheim（1963）の *Dictionnaire fondamental de la langue française*（以下では *DFLF* と略記）に収録されている，基礎的であるはずの「定型表現」の中にも，筆者がまったく知らないものは少なくはない（例えば，il m'a payé en monnaie de singe）。

---

[6] フランス語における，言語学的な単位の認定に関する議論は Léon, J. (2001)を参照のこと。

[7] 例えば，Habert & Jacqemin(1993), Degand & Bestgen(2003), Fujimura & Nakao (2005), François & Manguin (2006) が方法論を論じている。

[8] 学習者にとって，文法的性の習得が困難であることもこれに関連していると思われる。<la table>, <le sable>は，「定型表現（連語）」の一種であろう。

すなわち，学習者のためには，様々なタイプの「定型表現」を大量に含んだ辞書ではなく，学習者の目的に応じて作られた「定型表現」リストが必要である。そのリストは，第一に，項目の数が限られていること，第二に，重要度に関する情報が付与されていること，第三に，コンテクストやスタイルに関する情報を付与されていることが必要である。重要度は，後で述べるように高頻度に等しいと考えられる。たとえば，<plus ou moins>は高頻度の「定型表現」なので，最初に習得すべきリストに収録すべきであるが，<tôt ou tard>の頻度ははるかに低いので同じリストには入らない。個別単語は，基本単語などの選別がなされていて，それとわかるように辞書に記載されていることも多いが，「定型表現」はそうはなっていない。単語と同じように，情報を付加することが必要である。コンテクストや文体に関する情報も必要である。話し言葉の定型表現とメディアの定型表現は同じではないはずである。日本語で，例えば「蓼食う虫も好き好き」という表現を使うことのできる場面や文脈は限られている。また，文体と文脈の点で適切であったとしても，コンテクストを構成する他の表現が，当該の定型表現の難易度にふさわしい完成度に達していなければ，その使用は滑稽である。

日本では，日本語母語話者向けに，佐藤（1986），石野（1998）などの学習のための「定型表現」のリストが出版されている。それぞれ，1000 ほどの「定型表現」が収録されていて，学習辞書の「定型表現」の項目の使いにくさを補っていると考えられる。しかし，これらのリストには，上の条件のうち，重要度の情報とコンテクストや文体に関する情報は付されてはいない。リストに挙がっている「定型表現」の選択は，各種文献に基礎を置きつつも，最終的には著者が判定していると考えられるので，実際の使用と対照させてみるのは興味深い試みである。本稿では，N-gram の手法によって「定型表現」リストを作成することによって，これらの問題の解決を図り，日本語を母語とする学習者の便宜に供するリストを作成するための方法を考察する。

2.2. N-gram

N-gram モデルは，Claude Elwood Shannon が確立した情報学の基礎理論であり，言語情報を言語単位の連続の確率として分析するモデルである。本稿では N-gram は，単純に n 個の単語の連続のことをいう。コーパスから N-gram を抽出して，それぞれの頻度を計算すると，観察者の恣意性を遮断して，網羅的なリストを頻度付きで作成できるというメリットがある。長尾，森（1993）では，N-gram の手法が日本語を対象に文字を単位として用いられ，文字の連続としての「単語」を統計的に認定する試みがなされている。日本語では「単語」がスペースによって区切られていないため，「単語」の認定に N-gram の手法が用いられているが，同じ作業を，フランス語の場合に，スペースで区切られた単語を単位として行うと，単語の連続としての「定型表現」を認定できると予測できる。

一つのテクストから，N-gram を取り出す原理は単純である。コーパスが小さければ，エディタと表計算ソフトを使って簡単に作成することができる。例えば Le Petit Prince（総語数 17780 語）の 3-gram の作成のために Excel を利用すると図 4 のようになる。まず，テクストを一語ごとに改行して，語が垂直に並んだ状態にし，Excel のワークシートの左端に

ペーストする（図 4 の 1 列目）。次に，同じものをもう一つ作り，最初の語を削除して，ワークシートの 2 列目にペーストする。さらに，同じものをもう一つ作り，最初の語を再び削除して，ワークシートの 3 列目にペーストする。図 4 を行ごとにみると 3-gram のリストができている。ソートをかけ, 同一の語の連続からなる 3-gram の頻度を計算すると, Le Petit Prince の 3-gram のリストを頻度データつきで得ることができる。図 5 がそれであり，頻度順に結果が提示されている。 最も高頻度の 3-gram は<le petit prince>であり，頻度は 149 回である。

J'	ai	ainsi
ai	ainsi	vécu
ainsi	vécu	seul
vécu	seul	,
seul	,	sans
,	sans	personne
sans	personne	avec
personne	avec	qui
avec	qui	parler
qui	parler	véritablement
parler	véritablement	,
véritablement	,	jusqu’à

図 4：*Le Petit Prince* の 3-gram

149 le petit prince
60 , dit le
49 petit prince .
43 petit prince ,
40 dit le petit
38 . c' est
27 n' est pas
27 , c' est
20 . le petit
19 . J' ai

図 5：*Le Petit Prince* の最高頻度の 3-gram

このテクストに <le petit prince> が多数出現することは予想できるが，その次に出現が多いのが　<, dit le>　であることは, N-gram をとってはじめてわかることである。しかし，

これらの高頻度の 3-gram の中で，言語学的に意味があると思われる連続は，27 回の出現頻度の<n'est pas>のみである。他は，主人公の名前が含まれているか，句読点を含んでいて，言語的な特徴とは言えない。

N-gram の手法を，個別のテクストの特徴ではなく，言語学的に意味のある「定型表現」を抽出する目的で使おうとするなら，以下の点に注意を払わなければならない。

1. コーパスの規模
   それぞれのテクストの特殊性を排除するためには，コーパスは巨大であるほどよい。n 値が大きくなるにつれて，コーパスの特殊性は強調されて現れると言われているので，長い定型表現を得ようとするなら，数千万語以上からなる巨大なコーパスが必要である。Excel で N-gram の頻度を計算するのはこの意味で現実的ではない。
2. コーパスの構成
   コーパスを構成するテクストの種類によって，得られる N-gram は異なるので，コーパスを構成するテクストの特徴は知っておく必要がある。言語学的に意味のある「定型表現」を認定するという目的のためには，異なるタイプのテクストから抽出された N-gram の複数のリストを相互に比較するとよいと考えられるが，現実に使える大規模コーパスには偏りがある。「定型表現」は話し言葉にも多数出現すると考えられるが，数千万語からなる，大規模な話し言葉のコーパスを期待するのは現状ではむずかしい。
3. コーパス内での反復の排除
   テクストの形式的な特徴が N-gram の頻度に影響を与えていないかという問題には，注意を払う必要がある。文学作品では詩のリフレインなどがそれに当たるし，新聞などのメディアのテクストでは，記事の重複や，見出しと本文などの繰り返しがある。また，コーパスの精度も確認せねばならない。雑に構築されたコーパスでは，同じテクストが誤って何度も収録されている可能性がある。インターネット上には，莫大な量のテクストが存在し，大規模という条件を満たしていて N-gram にとっては都合がよいが，インターネットのテクストには繰り返しが多量に含まれているので慎重に取り扱わねばならない[9]。
4. lemma 化の必要性と問題点
   フランス語の「定型表現」は辞書の見出し語の形で表示されるのが普通である。それは動詞の活用などによって「定型表現」であるかそうでないかが変わるわけではないという考えに基づく。それに倣うなら，N-gram 抽出のためのコーパスは，生のデータではなく，形態素解析を施し，lemma 化して，辞書の見出し形に戻したデータを用いる必要がある。しかし，プログラムを用いて機械的に行う lemma 化に信頼をおくことはできない。重大な問題を回避するためには，使用するプログラムの特徴をよく把握しておかねばならない。また，lemma 化しない形の方が「定型表現」

[9] Google は，ウェブページから得た巨大な N-gram を公開している。
http://googlejapan.blogspot.com/2007/11/n-gram.html

として正しい場合も存在する。例えば，<il y a＋時間表現 >の場合はそうである。

5. 大文字と小文字の問題

N-gram の頻度を出す際に，文字列の大文字と小文字の区別を行わないオプションを選んで加算することは一般には可能である。上の例では<le petit prince>と<Le Petit Prince>は区別されていない。しかし，フランス語のアクセント記号付きの文字の場合，大文字と小文字を同一のものとして計算するのは簡単でない。大文字と小文字を同じに扱いたいのであれば，あらかじめ，大文字を小文字に変換した上で，N-gram をとることも必要である。ただし，大文字と小文字は，人名（Pierre）と普通名詞（pierre）などのように，別に計算するほうがよい場合も多い。

2.3. N-gram から「定型表現」を得ることはできるか？

N-gram は，語の単純な連続であり，そこに付された頻度は単純頻度である。このようなデータから，言語学的に意味のある「定型表現」が果たして本当に得られるのかという問題は検討せねばならない。

単純に高頻度であることは，「定型表現」の特徴とみなされる二つの性質のうちの一つに合致するものである。すなわち，単純頻度に基づく本稿の研究は，「定型表現」のうち，「使用頻度の高い表現」を明らかにすると考えられる。一方，「定型表現」の別の重要な特徴は，意味的融合（figement sémantique）である。N-gram の手法では意味的融合を含む「定型表現」のみを抽出することはできない。意味的融合とは，複数の語が，それぞれの語に切り離すことのできない一つの意味的な統一を構成する状態を言う。たとえば，pied noir（植民地からフランスへ戻った人）や La moutarde lui monte au nez（怒りがこみ上げる）などがそれに当たる（Gross, 1996）。単純に高頻度であることは，意味的融合を引き起こすとは限らないし，意味的融合のある連語が高頻度で出現するとは限らないからである。

筆者は Fujimura & Nakao（2005）において，フランス語の＜形容詞＋名詞＞の連続に前置される不定冠詞複数形の des と de の交替に関して，＜形容詞＋名詞＞が複合語か否かということと，この交替の間の関係を考察した（たとえば，des petites annonces, de grandes maisons）。複合語か否かということは，形容詞が意味的に依存的か独立的かという問題であると考えた。大規模な新聞・雑誌コーパスを用いた研究の結果明らかになったのは，形容詞と名詞の間に意味的融合があり，形容詞の独立性が失われているほど，des が使用される傾向が強く，意味的融合がなく，形容詞の独立性が保たれているほど，de が使用される傾向があるということであった。その際に，「複合語度」を計る指標として用いたのは，MI スコア，T スコアなどであったが，不定冠詞 de/des の選択と相関関係があるのは，MI スコアであって T スコアではなかった。

この研究の過程で確認できたのは，繰り返し言われているように（齊藤, 中村, 赤野（1998），Habert & Jacqemin（1993）など），MI スコアは単純頻度とは無関係に，語の間の特別な結びつきを計るのに適する指標であって，意味的融合の指標でもあるのに対して，T スコアは語の連続の生起頻度と深い関係があり，意味的融合とは関係がないということであった。表 1 と表 2 は，Fujimura & Nakao（2005）で扱ったデータのうちで，＜形容詞＋名詞＞の

MI スコアと T スコアの値が上位 10 位以内のものである[10]。タイトル行の「共起頻度」とは，＜形容詞＋名詞＞の単純生起頻度であり，「名詞の頻度」と「形容詞の頻度」は，それぞれの名詞と形容詞の単純生起頻度である。表 2 を見ると，＜形容詞＋名詞＞の T スコアの順位は「共起頻度」の順位とほとんど違わないことがわかる。T スコアが一位の grandes entreprises は，単純頻度も一位である。一方，表 3 を見ると，MI スコアの順位は単純共起頻度の順位とは関係がないことがわかる。MI スコアが上位なのは，「名詞の頻度」に占める「共起頻度」の割合が高いものである。たとえば，名詞複数形の riens はコーパスに 316 回しか出現しないが，そのうちの 167 回は petits と共起しているために，petits riens の MI スコアは高い。2 語間に特殊な結びつきが生じ，意味的融合が生まれ，形容詞の独立性が失われると，不定冠詞複数は des が選ばれやすくなる。MI スコアの上位に並ぶ＜形容詞＋名詞＞は，意味が不透明な特殊な連続であるものが多い。一方，T スコアの上位に並ぶのは，grandes entreprises, grands groupes, nouvelles technologies, grandes villes のように意味的に透明な連続のものが多い。

表 1：T スコア上位 10 位

順位	形容詞＋名詞	T スコア	共起頻度	名　詞	名詞の頻度	形容詞	形容詞の頻度
1	grandes entreprises	75.1	5800	entreprises	129576	grand(e)s	168313
2	grands groupes	69.3	4858	groupes	46061	grand(e)s	168313
3	nouvelles technologies	67.7	4595	technologies	16485	nouve(aux/lles)	145379
4	grandes villes	65.1	4279	villes	28858	grand(e)s	168313
5	grandes lignes	57.3	3308	lignes	19727	grand(e)s	168313
6	grandes surfaces	52.9	2803	surfaces	5342	grand(e)s	168313
7	grandes banques	51.6	2735	banques	54638	grand(e)s	168313
8	grandes écoles	49.5	2471	écoles	16798	grand(e)s	168313
9	petites entreprises	49.1	2472	entreprises	129576	petit(e)s	66361
10	bons résultats	48.6	2380	résultats	58514	Bon(ne)s	36749

[10] コーパスは，Le Monde など 5 種類の新聞・雑誌からなり，総語数は，2 億 7 千万語である。データは，＜à や dans などの 14 種類の前置詞＋de または des＋形容詞（ancien, beau, bon, grand, nouveau, petit）＋複数形名詞＞という形で生起した＜形容詞＋名詞＞のみを対象にしている。表の数値は，このようにして選択した＜形容詞＋名詞＞の MI スコア，T スコアを，同じコーパスを使って手作業で計算した結果である。単純生起頻度が少ない場合に，MI スコアは極端に高くなるなどの問題は調整済みである。詳しくは，Fujimura & Nakao (2005)を参照のこと。

表 2：MI スコア上位 10 位

順位	形容詞＋名詞	MI スコア	共起頻度	名　詞	名詞の頻度	形容詞	形容詞の頻度
1	bonnes volontés	11.2	379	volontés	1194	Bon(ne)s	36749
2	petits riens	11.1	167	riens	316	petit(e)s	66361
3	petits boulots	11.0	172	boulots	355	petit(e)s	66361
4	nouveaux venus	10.7	350	venus	407	nouve(aux/lles)	145379
5	belles empoignades	10.6	40	empoignades	315	be(aux/lles)	21643
6	petits déjeuners	10.6	290	déjeuners	768	petit(e)s	66361
7	nouveaux arrivants	10.4	548	arrivants	759	nouve(aux/lles)	145379
8	petites touches	10.3	409	touches	1339	petit(e)s	66361
9	petits pois	10.3	214	pois	702	petit(e)s	66361
10	petites annonces	9.9	869	annonces	3665	petit(e)s	66361

石川（2006）は，英語を対象に，粗頻度（単純頻度）・T スコア，対数尤度比，共起頻度比，MI スコアを検証して，「5 つの指標は，およそ，＜粗頻度・T スコア・対数尤度比＞と，＜共起頻度比・MI スコア＞という 2 つのグループにまとめられることが明らかになった。」としている。また，「前者は頻度にウェイトを置いた指標で，生起頻度の高い一般的なコロケーションの検出に強く，後者は，頻度情報にあまり依存しない指標であり，生起頻度は少ないものの，際立った特徴性・共起傾向性を示すコロケーションの検出に強い。」として，単純頻度からは T スコアと似た，一般的な高頻度の「定型表現」が得られるという筆者の主張と同じことを述べている。

以上から言えるのは，N-gram によって得られる単純頻度に基づくデータからは，T スコアから得られるものと類似して，「使用頻度の高い表現」を抽出することはできるが，複合名詞や慣用句のような意味的融合を伴う表現を抽出する目的には合致しないということである[11]。言語現象において，出現頻度は重要な役割を果たしていると筆者は考えている。「使用頻度の高い表現」には，質的な意味があり，言語現象の一つのまとまりを成す。しかし，それは，MI スコアの高い表現と同じような意味の融合を生むとは限らない。

3. N-gram の抽出

以下では，実際に N-gram を抽出する。

[11] Degand & Bestgen（2003）は，慣用的表現を抽出するための 3 つの指標を挙げている。そのうちの一つの，「neighbours の少なさ」というのも，MI スコアと類似した考え方に基づいている。

3.1. コーパス

主たるコーパスとして，1999 年の Le Monde 誌一年分を用いたが，テクストによる結果の違いを検討するために，2000 年と 1996 年の Le Monde，および，異なったテクストジャンルのものとして，Corpatext を使用した。これらのそれぞれを以下では，サブコーパスと呼ぶ。

Le Monde は，ELRA（European Language Resources Association）が作成，販売しているものを用いた[12]。このバージョンの Le Monde は，学術的用途のために整形され，テクストの構造を示すタグが施されたものである。たとえば，図 6 は 2000 年 1 月 1 日の記事の冒頭である。本研究では，可能な限り無用の重複を避けることが目的として，このうち，#TEX というタグによってマークされた記事本文部分のみを抽出し，コーパスとして用いた。

@ ARTICLE
#ACC(1)=682668
#CAT(1)=SUPPLEMENT,MANCHETTE
#DAT(1)=20000101
#DOC(1)=DRX
#ETA(1)=BASE
#FAB(1)=ADI12/0101TI
#NTE(1)=(Doc : Avec 13 dessins de Plantu résumant l'année mois par mois)
#NUM(1)=20000101-1
#PUM(1)=QUO
#REF(1)=2-S01-01
#SEC(1)=SPA
#SOT(1)=« Le Monde » présente un bilan de l'année illustré par Plantu. L'intervention de l'OTAN au Kosovo a dominé l'actualité internationale. En France, sur fond de cohabitation, de nombreux groupes comme Elf ou Paribas ont été rachetés
#SYE(1)=1999,BILAN
#SYF(1)=1999,BILAN
#TAI(1)=66
#TIC(1)=SUPPLÉMENT DE 24 PAGES IMPRIME TETE BECHE AVEC LE QUOTIDIEN DU MEME JOUR TITRE "BONJOUR 2000, ANNEE SYMBOLE"
#TIJ(1)=Adieu 1999,année extrême
#TEX= INAUGURÉE dans une certaine gaieté avec le lancement de l'euro, l'année 1999 se termine dans une méditation songeuse, à Seattle, sur les avantages et les inconvénients de la mondialisation. D'une monnaie encore largement virtuelle, l'euro, au bon vieux roquefort, l'année a donc balancé entre européisme convaincu et

図 6：Le Monde（ELRA)のタグと構成例

[12] http://catalog.elra.info/product_info.php?products_id=438&language=fr

Corpatext は，Paris5 大学の Lexique というサイトに置かれた大規模コーパス（Corpatext1.02）であり，無料で入手可能である[13]。フォーマットは単純なテクストファイルであって，自由に加工できる。内容は，16 世紀から 20 世紀までの文学作品と学術的文献であり，約 2700 種類のテクスト，3670 万語からなる。元のデータは，WordTheque[14]におかれた多様なフリーのテクストである。書誌情報はこのサイトで入手することが可能である。Lexique のサイトの説明によると，テクストの数は 2667 であり，"$$$$$"という印によって，テクストの区切りが示されているはずである。しかし，数えたところ，“$$$$$”の数は 2709 個ある。そしてそのあとに，テクストのタイトルなどが表示してあるが統一はとれていない。テクストの数が合わないし，整形がきちんとなされていないなど，問題もあるのではあるが，凡その辻褄は合っているし，重複が含まれている様子もないので，本研究のサブコーパスとして適切であると判断して利用することにした。以下に挙げるのは，Corpatext に採取されているきわめて多様なテクストのうちの数例である。

- *Essai*, Montaigne, 1592
- *Médecin malgré lui* , Molière, 1666
- *Déclaration Universelle des droits de l’homme*, Nations Unies, 1948
- *Chantefables et chantefleurs*, Desnos, Robert, 1944
- « Le moi dans le bouddhisme japonais au moyen-age : En ce concentrant sur Dôgen », *Philosophy East and West* , vol 41, No3. Kimura, Kiyotaka 1991

表 3 は，サブコーパスの規模と特徴を示したものである。すなわち，全てのサブコーパスの述べ語数に加えて，新聞データと文学・学術データの違いを把握するために，Le Monde1999 と Corpatext の異なり語数，TTR（Type-Token Ratio），Guiraud Index[15]を挙げてある。コーパスの総語数は，約 1 億語である。異なり語数は，品詞情報付きの Lemma（= 辞書の見出し語）ごとに計算してある。すなわち，たとえば，動詞の doit は devoir の他の活用形とともに一つの type にまとめられて計算されているが，名詞の devoir は別の type として計算されている。新聞データと文学・学術データともに，異なり語数は 20 万語を超えている。TTR と Guiraud Index は，どちらも語の使用の多様性に関する指標であるが，新聞データと文学・学術データとの間で数値に違いはなく，この点に関して，この二つのサブコーパスの傾向は似通っている。

[13] http://www.lexique.org/public/corpatext.php
[14] http://www.logoslibrary.eu/pls/wordtc/new_wordtheque.main?lang=fr&source=search
[15] 藤村（2008）を参照のこと。Guiraud はフランスの言語学者の Pierre Guiraud である。

表 3：コーパス

サブコーパス （Le Mondeは本文のみ）	語数 （=述べ語数（token））	異なり語数 （type）	TTR	Guiraud Index
Le Monde1999	22,585,884	214,787	0.00951	45.19
Le Monde2000	22,331,031			
Le Monde1996	19,942,447			
Le Monde 計	64,859,362			
Corpatext	36,694,907	284,694	0.00776	47.00
総計	101,554,269			

3.2. コーパスの整形

コーパスの整形としては，1）大文字の小文字化，2）品詞・形態素解析の 2 つの作業を行った。

3.2.1. 小文字化

コーパスは，Perl スクリプトによって，大文字を小文字に整形して用いた。文頭から始まる語の連続（たとえば，On ne peut pas）と文中にあらわれる連続（on ne peut pas）の頻度を合算することが目的であったが，アクセント付きでない文字の場合には，別の方法によって問題が簡単に解決することに後で気づいた。アクセント付き文字の場合には，色々と厄介であるので，この問題の解決は今後の課題とする。

3.2.2. 品詞・形態素解析

すでに述べたように，フランス語の場合，N-gram の抽出の前には，コーパスに形態素解析を施して，全ての語を辞書の見出し語の形に変えなければならない。そうしなければ，N-gram の頻度は, 種々の語形変化のそれぞれに拡散してしまう。たとえば, 5-gram の<ce qui ne empêcher pas>はコーパス全体のなかで 266 回出現し，同じく 5-gram の<ne pouvoir se empêcher de>は 710 回出現する。前者は Le Monde に多く，後者は Corpatext に多いという特徴を見ることができる。もしも品詞・形態素解析をせずに，生のままのデータを検索したとすれば，前者は，<ce qui n’empêche pas>, <ce qui n’empêchait pas>, <ce qui n’empêchera pas>などに分かれ，後者は，pouvoir の様々な活用形に分散するので，上のような特徴を見つけることはできなくなる。

コーパスの品詞・形態解析は，TreeTagger を使って行った[16]。TreeTagger のフランス語版は，筆者の所属する名古屋大学国際開発研究科国際コミュニケーション専攻のコーパスサーバ[17]にインストールされていて，登録ユーザはこのツールを使うことができる。

図 7 は，TreeTagger の出力例である。

[16] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[17] http://dicom2.gsid.nagoya-u.ac.jp

```
[fujimura@dicom2]$ tree-tagger-french petit_prince.txt
        reading parameters ...
        tagging ...
Je      PRO:PER je
demande VER:pres        demander
pardon  NOM     pardon
aux     PRP:det au
enfants NOM     enfant
```

図 7：TreeTagger の出力例

TreeTagger は無料のプログラムであり，個人のコンピュータにインストールすれば，自分のコンピュータ上でも使える。このような形態素解析プログラムを利用することによって，動詞の活用形や名詞や形容詞の語尾変化によって細分されることなく N-gram の頻度が計算できることは大きなメリットである。しかし，同時にプログラムの問題点も認識しておく必要がある。

1. 機能語の分析には注意が必要
   最頻出の語である機能語の分析には，解析ミスが多いし，設計上の問題もあるので注意が必要である。たとえば，que は，代名詞と接続詞の間でミスが頻繁である。性・数の分析には設計上の問題がある。代名詞の性は男女が別々に分析され，il と elle, le と la は別の lemma とされる。複数の ils は il, elles は elle に分析される。目的語複数の les は le|la という，別の lemma に分析され，一貫性がない。また，冠詞には性・数の区別はなされず，le，la，les，l'は le と解析される。des はすべて du と同じに扱われ，前置詞とされる。「否定の冠詞」の de も前置詞と分析される。N-gram をとるという目的から考えるならば，同じ形態の語が，あやまって複数の lemma に分析されることは問題となる。
2. 動詞と名詞間で解析ミス多数
   フランス語には，語尾が-e, -es で終わる名詞が多いと同時に，第 1 群規則動詞の語尾も-e,-es で終わるため，動詞と名詞間で分析のミスはよく起こる（danse, change など多数）。en fait の fait は, faire の活用形と分析されることがある。président が présider と分析されることもある。
3. 動詞の複合形は，二つの動詞と分析される
   たとえば, <avoir faire le objet de un >は，全体で 593 回出現する 6-gram であり, <faire le objet de un>は，2577 回出現する 5-gram である。<faire l'objet de >は，「subir 受ける」という意味の高頻度の表現であるが，単純形と複合形を区別する意味はなく，この「定型表現」の本当の価値を計るためには，融合させることが必要であろう。
   - Installé en Angleterre, il fait l'objet d'une enquête des services de police britanniques.

- Sur cet ensemble, 29 téléfilms, unitaires ou issus de collections, ont fait l'objet d'une sortie cinéma.

4. au, du などの縮約形

TreeTagger は，au や du を，<à le>や<de le>とは分析しない。一方，<à la>, <de la>は<à le>, <de le>と lemma 化され，aux は au に，des は du に lemma 化される。したがって，たとえば，<au sein de>，<au milieu de>，<dans le cadre de> の頻度を厳密に比較したければ[18], それぞれ, <au sein du>と<au sein de>, <au milieu du>と<au milieu de>, <dans le cadre du>と<dans le cadre de>を加算したものを比較せねばならない。Le Monde 3 年分においては，表 4 のように，それぞれ，合計，12420 回，3385 回，7532 回であり，Corpatext では，1059 回，8810 回，370 回である。

表 4：au sein de/du, au milieu de/du, dans le cadre de/du

	Le Monde	Corpatext
Au sein de/du	12420	1059
au milieu de/du	3385	8810
dans le cadre de/du	7532	370

Le Monde では <au sein de>と<dans le cadre de>，Corpatext では<au milieu de>が高頻度で使用されている。Le Monde においては，<au sein de >のあとには，l'Union Européenne，le groupe， le conseil など，特定の組織を示す名詞が続く。<dans le cadre de>は，<dans le cadre de l'enquête sur>が目立つ。<au milieu de> は<au milieu des années + 数詞>が多い。厳密にいうとこの通りであるが，大まかな傾向のみを得たいのであれば，du には目をつぶり，de で終わるもののみを比較しても大勢に影響はないだろう。いずれにせよ，TreeTagger からの出力結果をどのように使うか決定せねばならない。フランス語をよく知らないまま使うと，落とし穴にはまる危険性も高い。<aux yeux> は<au oeil>として出力されるなど他にも注意すべき点は多々存在する。

3.3. N-gram

TreeTagger による品詞解析のあと，そのデータをもとに N-gram を抽出する。コーパス全体を対象に，頻度が最も高い 4-gram を抽出した結果は，図 8 のとおりである。

24044	PRO_il ADV_ne PRO_y VER_avoir	(il ne y avoir)
19910	PRO_ce ADV_ne VER_être ADV_pas	(ce ne être pas)
12778	PRO_il ADV_ne VER_être ADV_pas	(il ne être pas)
11378	PRO_il PRO_se VER_agir PRP_de	(il se agir de)
11316	PRO_il PRO_y VER_avoir DET_un	(il y avoir un)

図 8：コーパス全体から得られた高頻度の 4-gram（上位 5 種）

[18] Petit Robert には，これらが同義語であるかのように書いてある。

計算は 4-gram の頻度順一覧を得るための次のプログラムをもとに, 必要な修正を施して行った[19]。

```
#! /bin/sh

#copy left 2004-12-12 sugiura@nagoya-u.jp

perl -pe 's/¥s+/¥n/g' $1 > ngramtmp1.txt
tail +2 ngramtmp1.txt > ngramtmp2.txt
tail +3 ngramtmp1.txt > ngramtmp3.txt
tail +4 ngramtmp1.txt > ngramtmp4.txt
paste ngramtmp1.txt ngramtmp2.txt ngramtmp3.txt ngramtmp4.txt |
sort | uniq -c |
sort -nr
rm ngramtmp*.txt
```

図 9：4-gram をとるためのプログラム

具体的な手順は以下のとおりである。

1. Le Monde1999 をもとに，2-gram から 10-gram のリストを作成。
2. コーパスの規模および，データベースソフトとして使った Excel 2003 が扱えるデータの量を考慮して，出現頻度の下限は，図 10 のとおりとし，これより低頻度のものは削除。

```
2gram:  100
3gram:  100
4gram:   60
5gram:   50
6gram:   50
7gram:   30
8gram:   20
9gram:   20
10gram:  20
```

図 10：各 N-gram の頻度の下限

[19] このプログラムも名古屋大学国際開発研究科国際コミュニケーション専攻のコーパスサーバにインストールされている。

3. N-gram の文字列中に，カンマ，ピリオド，クオテーションマークなどの記号が含まれているものは削除。
4. Excel を使ってデータベースを作成。
   データの総数は 50243 であり，50243 種類の N-gram が抽出されている。
   1. それぞれの N-gram を各行に割り当てる。50243 行のデータベースになる。
   2. それぞれの N-gram に種々の情報を付与する。
      (ア) Le Monde の 1996 における頻度データを付ける。
      (イ) Le Monde 2000 と 1996，および Corpatext における頻度データを追加。
      (ウ) その他，分析のために必要な種々の情報を付与（たとえば，動詞の複合形を含むか否か，数詞を含むか否かなど）。
5. 集計
   集計は，目的に応じて行う。様々な角度からの検討ができる。ただし，50000 件のデータの全てを有効に利用することは容易ではなく，第一に考えたいのは，データを絞るということである。
   たとえば，４種類のサブコーパスの全てにおいて，上に挙げた回数以上の出現があり，しかも，数詞と，動詞の複合形を含まないものだけを抽出すれば，データの件数は図 11 のとおりになる。8-gram，9-gram，10-gram は消える[20]。

2-gram:	8762
3-gram:	4562
4-gram:	1535
5-gram:	297
6-gram:	21
7-gram:	9
総計:	15186

図 11：全てのサブコーパスにおいて図 10 の下限頻度を
設定した場合のデータの数

4. ケーススタディ

以上の手順によって得られた N-gram のデータベースをもとに，いくつかの結果をケーススタディとして以下に紹介する。

[20] Le Monde1999 で最高頻度の 10-gram は次のものである。このような固有名詞的な表現は，Corpatext には多くは現れないため，データから排除される。
l'Organisation pour la sécurité et la coopération en Europe 欧州安全保障協力機構，105
Ministre de l'aménagement du territoire et de l'environnement 国土整備・環境大臣，104

4.1. 7-gram

表 5 には，サブコーパスのどれかに 100 回以上出現し，どのコーパスにも 20 回以上出現する 7-gram を全て挙げた。この条件に合うものは 4 個しか存在しない。

全てのサブコーパスに 100 回以上出現したのは，<que il ne y avoir pas de>のみであり，群を抜いた高頻度である。

- Le patronat doit reconnaître qu'il n'y a pas de réussite économique durable sans réussite sociale.

Le Monde のサブコーパスのどれかに 100 回以上出現し，Corpatext に 20 回以上出現するのは，<ne avoir rien à voir avec le>と<il ne se agir pas de un>である。この二つはメディアによく見られる表現であるとともに，日常的にもよく用いられる。

- la chanson populaire n'a rien à voir avec le chant classique.
- Il ne s'agit pas d'un festival en plus.

表 5：7-gram（サブコーパスのどれかに 100 回以上出現するもの）

番号	7-gram	例	合計	LM	Corpatext
1	que il ne y avoir pas de	qu'il n'y a pas de	1032	838	194
2	ne avoir rien à voir avec le	n'a rien à voir avec le	455	428	27
3	il ne se agir pas de un	il ne s'agit pas d'un	434	409	25
4	ce que il y avoir de plus	ce qu'il y a de plus	463	106	357

逆に Corpatext に 100 回以上出現し，Le Monde のサブコーパスのすべてに 20 回以上出現しているのは，<ce que il y avoir de plus>のみである。この表現は，＜tout ce qu'il y a de plus + 形容詞＞として，辞書などに掲載されている表現の一部であるが，tout を伴わない場合もある。意味は，extrêmement である。

- Cela leur serait peut-être tout ce qu'il y a de plus désagréable.
- Aimer, c'est ce qu'il y a de plus beau.

7-gram のような長い「定型表現」の場合，タイプの違いをこえて，どのようなテクストにも恒常的に用いられるものは数が限られる。<que il ne y avoir pas de>は，構成する全ての語が機能語の「定型表現」であり，あらゆるフランス語のテクストで使われる「機能的定型表現」と考えられよう。

4.2. 6-gram

6-gram は数が増える。全てのサブコーパスにおいて，出現頻度が 50 回以上の 6-gram を表 6 には掲載した。全部で 21 あるうちで，<il y a>を含む表現が 8 個，<il s'agit de>を含むものが 4 個である。<il y a> は言うまでもないが，<il s'agit de>も，高頻度の連語である。

表 6：6-gram（全てのサブコーパスに 50 回以上出現するもの）

番号	6-gram	例	計	LM	Corpatext
1	il ne y avoir pas de	ii n'y a pas de	5697	4276	1421
2	que il ne y avoir pas	qu'il n'y a pas	1771	1324	447
3	ce que il y avoir de	ce qu'il y a de	1332	293	1039
4	il ne se agir pas de	il ne s'agit pas de	1319	1132	187
5	il ne y avoir plus de	il n'y a plus de	1316	823	493
6	que il se agir de un	qu'il s'agit d'un	870	791	79
7	de le autre côté de le	de l'autre côté de la	794	510	284
8	il ne y avoir que un	il n'y a qu'un	791	353	438
9	il ne être pas question de	il n'est pas question de	622	559	63
10	il ne y avoir pas un	il n'y a pas un	589	280	309
11	ne avoir pas le droit de	n'a pas le droit de	567	386	181
12	ne avoir pas le intention de	n'a pas l'intention de	512	459	53
13	se il ne y avoir pas	s'il n'y a pas	445	315	130
14	ne être pas tout à fait	n'est pas tout à fait	427	287	140
15	il ne se agir que de	il ne s'agit que de	399	249	150
16	mais il ne y avoir pas	mais il n'y a pas de	342	235	107
17	il ne être pas possible de	il n'est pas possible de	336	237	99
18	jusque à le fin de le	jusqu'à la fin de la	329	276	53
19	se il se agir de un	s'il s'agit d'un	296	234	62
20	que il ne avoir pas le	qu'il n'a pas le	285	203	82
21	Du point de vue de le	du point de vue de la	281	227	54

どのサブコーパスにも 100 回以上出現するのは，（1）<il ne y avoir pas de>，（2）<que il ne y avoir pas>，（4）<il ne se agir pas de>，（5）<il ne y avoir plus de>，（7）<de le autre côté de le>，（8）<il ne y avoir que un>，（11）<ne avoir pas le droit de>の 7 個である（数字は表の番号に対応）。Le Monde に多いが，Corpatext に少ないのは，差が大きい順に，（6）<qu'il s'agit d'un>，（9） <il ne être pas question de>，（12）<ne avoir pas le intention de>，（4）<il ne se agir pas de>，（18）<jusqu'à la fin de > である。反対に Corpatext に多く，Le Monde に少ないのは，（3）<ce que il y avoir de> である。これは，7-gram で見た< (tout) ce que il y avoir de plus>の一部でもあるが，extrêmement の意味を持つ場合もあるし，そうでない場合もある。

- Ce qu'il y a de sûr, c'est qu'elle doit avoir pleuré. (Corpatext)
- Je ne vois pas bien ce qu'il y a de constitutionnel dans tout cela. (Le Monde)

4.3. 2-gram から 5-gram まで（上位 10 位）

n 値が減ると，データの数はますます増える。表 7 には，Le Monde と Corpatext に分けて，2-gram から 5-gram までの表現を，頻度順に上から 10 位まで掲げた。結果を見ると，

Corpatext で上位のものは全て，フランス語の基礎的な「機能的定型表現」といえるものである。テクストジャンルに固有のものではない。Le Monde では，上位に，「機能的定型表現」以外に，メディアに特徴的な，<président de le>,<du droit de le homme>, <le fin de le année> などの表現が含まれている。その傾向は n 値が大きくなるほどはっきりしている。

表 7：2-gram から 5-gram（上位 10 種類）

	Le Monde	頻度	Corpatext	頻度
2-gram				
1	de le (=de la/l') [21]	1021539	de le (=de la/l')	290407
2	à le (=à la/l')	410277	dans le (=dans le/la/les/l')	148210
3	dans le (=dans le/la/les)	327646	à le (=à la/l')	135034
4	de un (=d'un/une)	304627	de un (=d'un/une)	111674
5	sur le (sur le/la/les/l')	230882	ce être (=ce/c' êtreの各活用)	109952
6	et le (et le/la/les/l')	207661	et le (=et le/la/les/l')	100253
7	que le (que le/la/les/l')	205244	que le (=que le/la/les/l')	95852
8	par le (=par le/la/les/l')	202659	de son (=de son/sa/ses)	82870
9	pour le (=pour le/la/les/l')	155349	sur le (=sur le/la/les/l')	79188
10	ce être (=ce/c' êtreの各活用)	154569	il ne (il/ils ne)	69459
3-gram				
1	ne être pas	72413	ne être pas	31832
2	ne avoir pas	52525	il y avoir	25851
3	il y avoir	44869	ne avoir pas	21821
4	et de le	33039	ce être le	16601
5	ce être le	32358	ce être un	15749
6	ce être un	21441	que il ne	13757
7	président de le	20751	ce ne être	12947
8	le un du	20223	il ne y	12314
9	ce ne être	17156	ce que il	11758
10	que il ne	16167	ne y avoir	11542
4-gram				
1	il ne y avoir	13543	il ne y avoir	10501
2	ce ne être pas	12149	ce ne être pas	7761
3	il se agir de	9751	il ne être pas	4343
4	de plus en plus	9517	je ne avoir pas	4041

[21] カッコの中に書き入れたのは実現形である。2-gram に対してのみ書き入れたが，以下同様である。

5	le président de le	8829	il y avoir un	3900
6	il ne être pas	8435	il ne avoir pas	3371
7	ne y avoir pas	7471	ne y avoir pas	3369
8	il y avoir un	7416	il y avoir du	3113
9	ne être pas le	7117	ne être ce pas	2815
10	le ministre de le	6911	que il y avoir	2604
5-gram				
1	il ne y avoir pas	7162	il ne y avoir pas	3110
2	ne y avoir pas de	4339	que il ne y avoir	1736
3	du droit de le homme	4171	il ne y avoir que	1672
4	le président de le république	3383	ce que il y avoir	1593
5	il se agir de un	3297	ne y avoir pas de	1456
6	ce ne être pas le	2967	il ne y avoir plus	1234
7	à le fin de le	2727	ce ne être pas le	1199
8	le fin de le année	2658	que il y avoir de	1158
9	dans le cadre de le	2498	ce ne être pas un	916
10	faire le objet de un	2423	que il ne être pas	857

4.4. locutions prépositives

最後に，「定型表現」のリストの作成の例として，高頻度の locution prépositive のリストの抽出を試みる。locution prépositive とは，例えば<en face de>のような複合前置詞のことであり，「前置詞＋（別の語）＋名詞＋（別の語）＋前置詞」という形式を持つものである。手順は以下のとおりである。

1. Excel に保存されたデータベースを使い，TreeTagger の出力データをオートフィルタ機能によって検索し，PRP*NOM*PRP で始まる N-gram を抽出する。<PRP*NOM*PRP>は Excel の検索式であり，単語に付与された品詞タグが，PRP（= préposition）で始まり，任意の文字列のあと，NOM（= nom）を含み，さらに任意の文字列の後に再び PRP（= prépositon）が現れる文字列のことである。例えば，この検索式によって，PRP_au NOM_milieu PRP_de がヒットする。
2. １の出力に対して，再検索をかける。1 の条件を満たすものの中で，Le Monde において頻度の高い N-gram（ここでは 3 年間合わせて 1500 回以上）を抽出する。
3. 同じく，1 の条件を満たすものの中で，Corpatext において頻度の高い N-gram（ここでは 500 回以上）を抽出する。
4. <dans le maison de>などが抽出されてしまうので，不要例を直感に基づいて排除する。
5. 結果として，表 8 に挙げた 58 個の N-gram が採取される。

6. 表 8 では，Le Monde と Corpatext の使用の差を提示するために，コーパスのサイズの違いを調整したうえでの比を掲げてある。「LM/Corpatext（調整ずみ）」の列がそれに当たる。

表 8：locutions prépositives

番号	n 値	N-gram（代表例に変換済み）	総計	LM 合計	Corpatext	LM/Corpatext（調整ずみ）[22]	閾値を超えて生起するテクスト
1	3	au cours de	8667	8015	652	6.95	両方
2	3	au sein de	8390	7719	671	6.51	両方
3	3	en matière de	7702	7215	487	8.38	ＬＭ
4	3	au milieu de	6860	1483	5377	0.16	Corpatext
5	3	au lieu de	6721	3262	3459	0.53	両方
6	3	par rapport à	6441	5980	461	7.34	ＬＭ
7	3	en cas de	5818	5260	558	5.33	両方
8	4	dans le cadre de	5724	5463	261	11.84	ＬＭ
9	3	au bout de	5325	2506	2819	0.50	両方
10	3	en raison de	5213	4790	423	6.41	ＬＭ
11	4	à la fin de	4899	4184	715	3.31	両方
12	3	en faveur de	4690	4072	618	3.73	両方
13	3	au nom de	4400	3273	1127	1.64	両方
14	3	à cause de	4150	1833	2317	0.45	両方
15	4	à la suite de	4093	3584	509	3.98	両方
16	3	ｅｎ　fin de	3657	3450	207	9.48	ＬＭ
17	3	au fond de	3639	748	2891	0.15	Corpatext
18	3	au début de	3500	3222	278	6.56	ＬＭ
19	3	à propos de	3393	2490	903	1.56	両方
20	4	à la tête de	3373	2886	487	3.35	ＬＭ
21	3	en dépit de	3161	2785	376	4.19	ＬＭ
22	3	au terme de	3137	2972	165	10.20	ＬＭ
23	3	au mois de	3124	2757	367	4.25	ＬＭ
24	4	à l'égard de	2947	2381	566	2.38	両方
25	3	au coeur de	2922	2586	336	4.35	ＬＭ
26	3	au moment de	2915	1997	918	1.23	両方

[22] 数字は，コーパスの規模に合わせて調整済みである。LM/Corpatext が 3 というのは，Le Monde では，Corpatext の 3 倍の頻度で使用されるということである。例えば，<en raison de>と <à cause de >の Le Monde と Corpatext での使用頻度を比べると，<en raison de>は Le Monde では Corpatext の 6 倍使用されるが，<à cause de >は半分以下しか使用されない。すなわち，この二つの「定型表現」には，テクストジャンルによる頻度差に，12 倍の開きがある。

27	4	à l'occasion de	2872	2682	190	7.99	LM
28	3	à côté de	2795	1152	1643	0.40	Corpatext
29	3	en vue de	2652	2088	564	2.09	両方
30	4	à l'âge de	2597	2302	295	4.41	LM
31	3	du côté de	2581	1418	1163	0.69	Corpatext
32	3	au bord de	2466	1221	1245	0.55	Corpatext
33	4	de la part de	2417	1738	679	1.45	両方
34	3	au centre de	2275	2001	274	4.13	LM
35	3	au profit de	2227	1997	230	4.91	LM
36	3	au service de	2173	1821	352	2.93	LM
37	4	à l'origine de	2151	2051	100	11.60	LM
38	3	en mesure de	2145	1942	203	5.41	LM
39	3	au lendemain de	2084	1984	100	11.22	LM
40	3	au pied de	2077	697	1380	0.29	Corpatext
41	3	en dehors de	2059	1342	717	1.06	Corpatext
42	3	aux yeux de	1936	1059	877	0.68	Corpatext
43	3	à coup de	1890	1071	819	0.74	Corpatext
44	3	sous forme de	1831	1612	219	4.16	LM
45	3	au long de	1814	1688	126	7.58	LM
46	3	en état de	1804	887	917	0.55	Corpatext
47	4	à la recherche de	1774	1501	273	3.11	LM
48	3	en face de	1772	557	1215	0.25	Corpatext
49	4	dans le domaine de	1752	1561	191	4.62	LM
50	3	en présence de	1660	1003	657	0.86	Corpatext
51	3	à force de	1614	664	950	0.40	Corpatext
52	3	en vertu de	1565	708	857	0.47	Corpatext
53	4	sous le nom de	1494	924	570	0.91	Corpatext
54	4	sur le point de	1337	828	509	0.92	Corpatext
55	3	au moyen de	1299	512	787	0.36	Corpatext
56	4	d'un coup de	1284	527	757	0.39	Corpatext
57	3	au sujet de	1227	697	530	0.74	Corpatext
58	3	de mainière à	1138	441	697	0.36	Corpatext

表 8 の N-gram には次の特徴がある。

1. 3-gram と 4-gram のみであり，5-gram は含まれていない。4-gram が 14 個，3-gram が 44 個である。一般に高頻度の N-gram は 4-gram までの範囲に収まり，それを超えると急に少なくなる。
2. 表 8 の表現はすべて，どのテクストにも高頻度の「定型表現」であるが，テクストジャンルの違いによって，頻度に偏りがあるものとそうでないものとがある。

3. Le Monde において多用されるのは，順に，（1）<au cours de>，（2）<au sein de>,（3）<en matière de>,（6）<par rapport à>,（8）<dans le cadre de>である（上位 5 位）。Le Monde のサブコーパスでは，どれもこの 5 つが同じ順序で上位に並ぶ。Corpatext では，（4）<au milieu de>,（5）<au lieu de>,（17）<au fond de>,（9）<au bout de>,（14）<à cause de>が上位にあり，Le Monde とはまったく別の 5 つである。
4. Le Monde において 1500 回，Corpatext では 500 回以上現れ，両者ともに頻度が高いのは，（1）<au cours de>,（2）<au sein de>,（5）<au lieu de>,（7）<en cas de>,（9）<au bout de>,（11）<à la fin de>,（12）<en faveur de>,（13）<au nom de>,（14）<à cause de>,（15）<à la suite de>,（19）<à propos de>,（24）<à l'égard de>,（26）<au moment de>,（29）<en vue de>,（33）<de la part de>の 15 個であり，基本的と認定しうる。
日本で出版されている『フランス基本熟語集』（以下『熟語集』と略記）は，このうち，（2）<au sein de>,（12）<en faveur de>,（13）<au nom de>,（24）<à l'égard de>を採取していない。
5. Le Monde に多く，Corpatext には少なく，その間の差が大きい（約 10 倍の差がある）のは，頻度差の順に，（8）<dans le cadre de>,（37）<à l'origine de>,（39）<au lendemain de>,（22）<au terme de>,（16）<en fin de>である。<en fin de>を除いて，『熟語集』には採取されていない。
    - Mais elles travailleront dans le cadre de notre culture.
    - Elle est à l'origine de 63 morts et 22 avortements.
    - Or le marché immobilier parisien ne semble pas capable d'absorber une telle offre au lendemain de la quinzaine Olympique.
    - Au terme de sa carrière militaire, il y a six ans, il fut acheminé à Helsinki par Subexpo où il servit comme musée-restaurant.
    - Ce devrait être fini en fin de matinée.
6. 逆に，Corpatext に多く，Le Monde は少なく，その間の差が大きい（3 倍から 7 倍）のは，頻度差の順に（17）<au fond de>,（4）<au milieu de>,（48）<en face de>,（40）<au pied de>, (58)<de manière à >である。『熟語集』には，<au fond de>, <au milieu de> , <en face de>が採取されている。
    - Gardez son image au fond de votre coeur.
    - Comme elle était tourmentée d'une idée particulière, au milieu de la conversation la plus générale, elle ne restait jamais parfaitement calme.
    - Dans la rue, en face de vos fenêtres, dans l'embrasure de cette porte: un homme enveloppé dans un manteau
    - Il descendit au jardin, et commença à creuser la terre au pied de l'églantier,...
    - j'ai disposé les choses de manière à tout prévoir.
7. フランス語の古典的な基本語辞典である *DFLF* に熟語として採取されず，例にも挙げられていないのは，以下の 16 個である。
    （3）<en matière de>,（8）<dans le cadre de>,（16）<en fin de>,（21）<en dépit de>,

（22）<au terme de>,（23）<au mois de>,（30）<à l'âge de>,（35）<au profit de>,（36）<au service de>,（37）<à l'origine de>,（39）<au lendemain de>（42）<aux yeux de>,（44）<sous forme de>,（49）<dans le domaine de>,（53）<sous le nom de>,（54）<sur le point de>。

- il était aux yeux de ma famille, qui le citait toujours en exemple, le type de l'homme d'élite, prenant la vie de la façon la plus noble et la plus délicate.

Corpatext で頻度が高い表現は，*DFLF* にほとんどが採取されている。30 位までのうちで，採取されていないのは，<aux yeux de>と<sous le nom de>の 2 個のみである。Le Monde での頻度の高さと *DFLF* への採取とは無関係である。30 位までのうちで 10 が採取されていない。採取されていないものの中でもっとも頻度が高いのは，Le Monde において 3 位の<en matière de>である。意味は「〜に関して」である。

- Nos opinions publiques demandent des progrès en matière de normes sociales et de respect de l'environnement.

5. 結論

本稿では，N-gram の手法によってフランス語の基本的な「定型表現」のリストを作ることを目的とし，そのための予備的研究を行い，以下のことを明らかにした。

1. N-gram の手法によって，目的を果たすことは可能と考えられるが，限界もある。N-gram によって得られるのは，「使用頻度の高い表現」であって，「意味的融合のある表現」ではない。「意味的融合のある表現」は，意味論の研究対象として，知的に興味深いが，言語使用の面では，そのうちの頻度の低いものは習得しても実用的な利益は少ない。N-gram に基づく「定型表現」は，実用的価値が高いと考えられる。
2. Corpatext と Le Monde の高頻度の N-gram のリストには，思った以上に，大きな差があることが明らかになった。どちらのコーパスにも多数含まれるものは，「機能的定型表現」であり，内容的な単位を結ぶ連結部の働きをすると想定される。Le Monde と Corpatext のリストの表現を観察すると，Corpatext では「機能的定型表現」が上位に多数出現し，テクストの特殊性は排除されていると考えられる。この二つのコーパスに関する限り，基本的な「定型表現」の採取のためには Corpatext が適している。その理由は，Corpatext は多様なテクストによって構成されているが，Le Monde は特殊な単一のテクストで構成されているからであると思われる。
3. Le Monde からはメディアで多用される表現を得ることができるが，専門的な熟語の採取を目的としないのであれば，今回行ったように，違ったテクストジャンルのコーパスを利用して，フィルターをかけ，特殊用語は排除するようにする必要がある。Le Monde のみを使った場合には，固有名詞のような経済・政治用語が数多く採取される。Le Président de la République は，基礎的な「定型表現」と言ってよいだろうが，l'Organisation pour la sécurité et la coopération en Europe はそうではない。後者を排除

し，前者を残すためには，他のジャンルのテクストにどの程度現れるかを調査せねばならない。

4. 品詞解析と lemma 化は，フランス語の場合，避けられない過程であるが，lemma 化された「定型表現」を，生の形の「定型表現」と比較する作業も同時に行うべきである。同じ lemma であっても実現形は，どれか特定のものに偏っている可能性が高い。また形態素解析プログラムの限界を知り，最適な利用法を検討すべきだろう。
5. N-gram の最大のメリットは方法が単純だということである。特別のプログラムを使わず，言語研究者が自分の手で N-gram のリストを作ってそれを観察すれば，さまざまな新しい発見が可能であると考えられる。例えば，<il être 形容詞 que>を含む高頻度の N-gram にはどのようなものがあるかとか，定冠詞・不定冠詞を従える「表現」の違いなどを，研究者の創意によって調査することが可能である。

本稿の N-gram データベースは 50000 例以上からなる。4 つのサブコーパスの全てにおいて 100 回以上出現する例は 15000 例存在する。本稿においては明らかにできることの一端を示した。このデータベースをさらに様々な角度から観察することによって，他にも様々な有意の結果を得ることができるだろうと考えている。

参考文献

藤村逸子(2008)「フランスの特徴的なコーパス研究―語彙研究と政治ディスコース研究―」,『英語コーパス研究』15: 45-56.

石川慎一郎（2006）「言語コーパスからのコロケーション検出の手法－基礎的統計値について－」『統計数理研究所共同研究レポート』 190: 1-14.

石野好一（1998）『パターンで覚えるフランス基本熟語』, 白水社.

近藤泰弘, 近藤みゆき（2004）「N-gram の手法による言語テキストの分析方法」,『漢字文献情報処理研究』 No5: 50-55.

長尾真，森信介（1993）「大規模日本語テキストの n グラム統計の作り方と語句の自動抽出」,『自然言語処理』96-1: 1-8.

齊藤俊雄，中村純作，赤野一郎（1998）『英語コーパス言語学―基礎と実践―』, 研究社出版.

佐藤房吉（1986）『フランス基本熟語集』,白水社. ［本文中では『熟語集』と略記］

Degand, Liesbeth & Yves Bestgen (2003) "Towards Automatic Retrieval of Idioms in French Newspaper Corpora", *Literary and Linguistic Computing*, 18(3):249-259.

François, Jacques & Jean-Luc Manguin (2006) "*Dispute théologique*, *discussion oiseuse* et *conversation téléphonique* : les collocations adjectivo-nominales au coeur du débat". *Langue Française* 150 : 50-65.

Fujimura, Itsuko et al. (2004) "*De* et *des* devant les noms précédés d'épithète en français : problème de *petit*", *Le Poids des mots* vol.1, Presses Universitaires de Louvain: 456-48.

Fujimura, Itsuko & Hiroshi Nakao (2005) "Le choix de l'article indéfini *de* et *des* dans les noms composés en français", *Cahiers de l'Institut de Linguistique de Louvain*, 31.2-4: 27-44.

Fujimura, Itsuko et al. (2007) "Opposition entre *de* vs *des* devant les noms précédés d'épithète en français: portée du "poids"", *Texte et Corpus 2003 : Acte des JLC3*: 131-141.

Gougenheim, G.eorges (1963) *Dictionnaire fondamental de la langue française*, Modern Asia Editions. ［本文中では *DFLF* と略記］

Granger, Sylviane & Fanny Meunier (ed) (2008) *Phraseology: an interdisciplinary perspective*. John Benjamins.

Gross, Gaston (1996) *Les expressions figées en français: noms composés et autres locutions*, Ophrys.

Habert, Benoît & Christian Jacquemin (1993) "Noms composés, termes, dénominations complexes; problématiques linguistiques et traitements automatiques", *Tal* 34-2: 5-41.

Léon, Jacqueline (2001) “Conceptions du ‘mot’ et débuts de la traduction automatique.” *Histoire Épistémologie Langage* 23-1; 81-105.

Rey, Alain & Sophie Chantreau (2007) *Dictionnaire des expressions et locutions*, Le Robert. ［本文中では *DEL* と略記］

Sinclair, John (1991) *Corpus, Concordance, Collocation*. Oxford University Press.

Stubbs, Michael (2007) “On texts, corpora and models of language”, in *Text, Discourse and Corpora, Theory and Analysis*, Continuum: 127-61.