# フランス語品詞タグ付きコーパス作成における若干の問題点について

中尾　浩

キーワード

フランス語，コーパス，レキシコン，形態的解析，構文解析

**要約**

コンピュータの出現によって，言語研究は今や大きく変化しようとしている。今後，大規模コーパスの利用に基づかない言語研究はありえない状況にさえなりつつある。英語や日本語ではすでに大規模コーパスが多数作成され，研究に利用されているのに対して，フランス語やドイツ語ではデータの段階ですでに遅れを取っている。本論ではまず最初にコーパスに関して概観し，次いで筆者が試みているフランス語コーパスおよびレキシコン作成過程における若干の問題点について述べた後で，フランス語の計量分析結果について報告する。

## 1. はじめに

現代言語学におけるコンピュータ利用は分野によってはコンピュータサイエンスの最先端と歩を同じくし，遅れている分野ではインフラ整備すらままならない，という状況である。フランス語学においてはどうかというと，残念ながらフランス本国はもとより日本でももちろん，デー

タの整備さえ整っていない[1]。むしろスイスのジュネーブ大学[2]やアメリカのシカゴ大学[3]やバージニア大学[4]などが積極的にデータや成果を公開しているといったありさまである。

ますます精緻を極める言語研究において，コンピュータなしで研究を続けられる日は，研究対象にもよるが，早晩終わりを告げることは間違いない。そのときに，どのように基盤を構築していたかが大きな問題点として立ちはだかる。単にフランス語が入力されたデータがありさえすればよいだろうか。分析するためのツールやノウハウはそろっているか。その差が厳しく問われる日は遠からずやってくる。

ここでは筆者の試みを紹介しながら，フランス語研究におけるコンピュータ言語学について若干の考察をおこなう。

## 2. コーパスとは何か

コーパス（Corpus）に関する一般的な定義は以下の通りである。

「コーパスもしくは電子コーパスとは，言語研究に使用されることを想定して，実際に書かれたり話されたりした言語をコンピュータ上で利用可能にしたテキストの集合体のことである」[5]。

もちろん，コーパスとは何もコンピュータ上で利用可能なテキストの集合体とばかりは限らない。コンピュータの出現以前からコーパスは存在したが，今日ではコーパスといえば一般に電子コーパスのことを指す。

（電子）コーパスというと，単にフランス語だけや英語だけが入力されたものを想像しがちだが，それはコーパスの一部に過ぎない。そうした生のデータは Raw Corpus と呼ばれる。以下が，Brown Corpus による Raw Corpus の例である。

```
A01 0010    The Fulton County Grand Jury said Friday an investigation
A01 0020 of Atlanta's recent primary election produced "no evidence" that
A01 0030 any irregularities took place.    The jury further said in term-end
A01 0040 presentments that the City Executive Committee,  which had over-all
A01 0050 charge of the election,  "deserves the praise and thanks of the
A01 0060 City of Atlanta" for the manner in which the election was conducted.
```

ちなみに，こうしたRaw Corpusは通常，処理しやすいように適当に整形される。以下はLOB Corpusの例である。

```
A01    1 ** [001 TEXT A01**]  A01    2 *<<*'*7STOP ELECTING LIFE PEERS**'*>>
A01    3 *<<*4By TREVOR WILLIAMS*>>
A01    4     |^A *0MOVE to stop ¥0Mr.  Gaitskell from nominating any more Labour
A01    5 life Peers is to be made at a meeting of Labour  {0M P} s tomorrow.
A01    6     |^¥0Mr.  Michael Foot has put down a resolution on the subject and
A01    7 he is to be backed by ¥0Mr.  Will Griffiths,   {0M P}  for Manchester
```

しかし，コーパスの構築を少しでも試みたことのある人なら，このようなRaw Corpusだけでは不充分であることにすぐに気づくだろう。

筆者が最初にコーパス作りに着手したのはもう10年以上前になる。そのとき，ソシュールの言語事実（fait linguistique）に関する論文を準備していた筆者は，ソシュールの原資料をワープロ（残念ながら当時はまだパソコンを所有していなかった）に打ちこんで，ある程度データがたまった段階で検索をかけたところ，たちどころに困難にぶつかった。ソシュールが言語事実について論じている場合，必ずしもfait linguistiqueと表現しているとは限らないので（単にun faitやles faitsとしか記述されていない場合も多い），必然的に，「fait」で検索をかけざるをえない。すると，il fait（faireという動詞の現在形）や，il a fait（同じく過去分詞）を大量に出力して，そこから名詞のfaitだけをまた選びなおさなければならなかった。

つまり，フランス語のような同形異義語の多い言語では，単にフラン

ス語が入力されているだけだと，単なる検索ですら，余計なものを大量に拾って，目的の用例に到着することさえ一筋縄では行かないのだ。

あるいは，半過去の用法を研究している人がいるとしよう。筆者が調べたところでは，8000 以上もあるフランス語の動詞の半過去だけをどうやって探し出そうか。半過去は一つの動詞につき，人称ごとに 5 つの形態の差があるが，4 万回も検索を繰り返さなければならないのだろうか。

確かに，Raw Corpus の検索だけで十分な場合もある。しかし，Raw Corpus だけでは形態素検索しかできないので，研究の範囲が狭まってしまう。この問題を解決するには二つの方法がある。

1：データにタグをつける。
2：適当な電子辞書を参照させながら検索する。

ここでは 2 については論じないので，先に済ませてしまうと，辞書参照型検索の場合，原則としてデータを選ばないという利点はある。場合によっては特定の形式でなければならないこともあるが（たとえばコンコーダンスを生成するために，ページ番号が特定の形式で含まれていなければならない，など），一般に Raw Corpus であればよい。ただし，辞書とアプリケーションを作成するのに膨大な労力を必要とし，しかも辞書に記述されていない文字列に対してはエラーを返すしかないので，柔軟性にも欠けるという欠点もある。ここではこれ以上は論じないが，だからといって辞書参照型検索が不要なわけでも，タグ方式より劣っているわけでもない。

## 3. タグ付きコーパス（Tagged Corpus）

説明が前後してしまったが，以上のような状況を考えると，データにタグを付ける方式は辞書参照型に比べると，比較的少ない労力で大きな成果が期待できる。事実，タグ付きコーパスは英語や日本語では同一のコーパスに対して研究目的に応じて何種類も作成されている。タグ（tag（英），étiquette（仏））とはもともと（荷）札のことを指すが，簡単に言ってしまえば，データに対して付加情報を与えることである。たとえば，先ほど例に出した，fait についても，N を名詞，V を動詞と定義して，fait_N，fait_V といったきわめて原始的なタグをつけるだけで，名詞の fait と動詞の fait を区別することができる。

具体的なタグ付きコーパス例を紹介しよう。LOB Corpus のサンプルである。

```
A01   2 ^ *'_*' stop_VB electing_VBG life_NN peers_NNS **'_**' . _.
A01   3 ^ by_IN Trevor_NP Williams_NP . _.
A01   4     ^ a_AT move_NN to_TO stop_VB ¥0Mr_NPT Gaitskell_NP from_IN
A01   4 nominating_VBG any_DTI more_AP labour_NN
A01   5 life_NN peers_NNS is_BEZ to_TO be_BE made_VBN at_IN a_AT meeting_NN
A01   5 of_IN labour_NN ¥0MPs_NPTS tomorrow_NR . _.
A01   6     ^ ¥0Mr_NPT Michael_NP Foot_NP has_HVZ put_VBN down_RP a_AT
A01   6 resolution_NN on_IN the_ATI subject_NN and_CC
A01   7 he_PP3A is_BEZ to_TO be_BE backed_VBN by_IN ¥0Mr_NPT Will_NP
A01   7 Griffiths_NP , _,  ¥0MP_NPT for_IN Manchester_NP A01   8 Exchange_NP . _.
```

ここで紹介したタグ付きコーパスは品詞タグだが，その他に構文タグや意味タグなど，研究対象に応じてさまざまなものが存在するし，タグ付けの方式にもさまざまなものが工夫されている。それは個別の処理に対して最適な形にしてあるだけで，基本的な考え方は全て同じである。

こうしたデータなら，日常使いなれているアプリケーションでも簡単に過去形だけの検索などもできる。上の LOB Corpus なら，_NP で固有名詞が，_VBN で動詞の過去分詞が検索できる。簡単な例だが，百聞は一見にしかずで，Raw Corpus に対して Tagged Corpus の方が付加価値が格段に高いことはお分かりいただけよう。

このように生のデータに付加価値をつけたのがタグ付きコーパスだが，それを構築することは容易なことではない。英語などではすでに何種類もタグ付きコーパスが作成されているにもかかわらず，タガー（Tagger）と呼ばれるタグ付与ソフトが今でも開発されていることからもわかるとおり，精度の高いタグを付与するためには，まだ研究の余地がある。ましてや，フランス語やドイツ語では，スタンダードとなるタガーが存在しないので，一日も早く開発する必要がある。英語用に開発されたタガーを応用する方法もあるが[6]，タグ付けは個別言語の難しさをそのまま反映してしまうので，必ずしも精度が高いとは言えない。一般にタグ付けは以下の順序でおこなわれる。

1：タグセットの決定
2：形態的解析によるタグ付け
3：構文解析によるタグ付け

一六九

データに対して品詞タグを振るとは言っても，あらかじめそれを定義しておかないと，泥縄式に途中でタグが混乱してくる。その場合，フランス語だと「動詞」というタグだけを振ってもほとんど使い道がないので，「直説法現在」や「接続法過去」といったタグも付けなければならないのだが，そうなると，品詞タグと構文タグの境界線がいささか曖昧になってしまうが，それは仕方ない。あまりに何もかも一つのデータに

盛り込みすぎるのは使い勝手が悪くなるので，必要なタグは振りながら，適度にタグセットも分ける，と考えるのが妥当だろう[7]。もちろん，特殊な研究分野になれば，それに応じて特殊なタグセットを必要とすることは言うまでもない。

タグ付けの実際については章を改めて論じることにする。

## 4. 形態的解析の問題点

タグセットが決定すればコーパスに対してタグを付与する下準備が整ったことになる。ただし，実際にはタグを与えやすいようにコーパスを整形したり，タグを与えるためのプログラムを書かなければならないが，今回はその部分については割愛する。むしろ，なぜ形態的解析によるタグ付けをおこなってから構文解析によるタグ付与をおこなわなければならないかについて論じてみたい。

理論的に考えると，実際には構文解析によるタギングだけでコーパスに対するタグ付けはほぼ完了する[8]。しかし，これでは構文規則が無意味に複雑長大になってしまうばかりか，人間の言語認知過程からもかけ離れた方法になってしまう。もちろん，何も人間と同じ認知過程をコンピュータにシミュレートさせなければならない理由はないのだが，現実問題として，il fait の fait が動詞，un fait が名詞と決定できるのは，fait に先行する il や un が「代名詞主語」や「冠詞」であることが分かっているからであって，結果的に形態的にのみ決定できるものをもアルゴリズムの中に取り込まなければならなくなる。donner という動詞の条件法現在形の三人称複数 donneraient は形態的に動詞でしかありえない。多くの単語はこのようにして品詞が決まる。それなら，形態的にのみ品詞が決定できるものにはあらかじめ前処理として単純な検索・置換によってタグを与えてしまってもかまわないわけだし，むしろ，同形異義

語に絞って品詞情報を与える方がプログラムが短くなる上に，構文解析のための手がかりが与えられているので，精度が上がると期待できる。

ところがここに重大な問題がある。

あらかじめ形態的解析により品詞情報を与えるのはよいが，どんな言語にも必ず同形異義語がある。しかも，フランス語の場合，その同形異義語はかなり多いと予想される。動詞の時制に限っても，一部の不規則動詞を除いて，ほとんど全ての動詞は，たとえば -er 動詞なら直説法現在と接続法現在の一人称単数と三人称単数は同形といった具合に，必ずバッティングが生じている。半過去にいたっては全ての動詞に関して必ず一人称単数と二人称単数が同形である。ところが話はそこで終わらない。副詞の plus と動詞の plaire の直説法単純過去 je plus，tu plus も同形である。動詞の devoir の現在分詞と前置詞の devant も同形であるばかりか，devant は名詞として使われることもある。このように，単に動詞の変化形同士のみならず，他の品詞にまで同形異義語が存在するとなると，いったい，何と何が同形異義語で，構文的解析によってでなければ品詞タグを付与することができないのかをあらかじめリストアップしておかないと，形態的解析さえ不可能である。副詞の plus と plaire の活用形である plus を比べれば，前者の方が出現頻度ははるかに高いことは経験的に分かるが，だからと言って，いつどこで動詞の plus が出現するか分からない以上，形態的タグ付与のエントリーからははずしておく必要がある。そうでないと，出現数が少ないということは，珍しい用例と考えられるので，せっかくの用例を逃してしまうことになる。



では，具体的に，何と何が形態的バッティングを起こしているのか。寡聞にしてフランス語に関してそのようなリストが存在することは聞いたことがない。従って，フランス語に自動的に品詞タグを付与するためには，まず，ここから出発せざるをえない。ここで，コーパスを作成するための道具の一つとしてのレキシコンの問題が現れる。

## 5. レキシコンとは何か

コーパスと同じく，レキシコンも何も電子レキシコンであるとは限らない。語彙集ならはるか昔から作成されてきた。しかし，近年，自然言語処理の分野で，機械可読辞書としてのレキシコンという名称はほぼ定着してきた感がある。

ここまで見てきたとおり，コーパスに品詞タグを与える場合，形態的解析から入るほうが経済効率がよい。いきなり構文的解析だと無駄が多い。しかし，形態的解析から入ろうと，構文的解析からいきなり始めようと，そもそも品詞情報を記述したリストがなければ話にならない。そのリストがレキシコンである。

レキシコンにもさまざまな種類のものがある。品詞情報のためのものもあれば，意味情報のためのものもある。こうしたレキシコンは英語や日本語ではかなり整備されてきて，たとえば情報処理振興事業協会が作成した，IPAL 辞書[9] などは情報処理関係でよく利用されている。利用目的として「IPAL（アイパル）は，機械翻訳をはじめとする，日本語ワードプロセッサ，日本語インタフェース，音声入出力などの計算機システム以外に，日本語教育，日本語研究などにも利用いただいております。」[10] とあることからもわかるように，レキシコンの応用範囲はきわめて広い。ここで目的としているようなフランス語の品詞タグ付けのためにもどうしても最初にレキシコンの作成から入らざるをえない。

コーパスを作成する前にレキシコンを整備しなければならないことは分かったが，だからといってレキシコンの方もそう簡単に作成できるものではない。現時点でフランス語のレキシコンとしては，筆者に確認できた限りでは ABU（Association de Bibliophiles Universels）が作成したものをインターネット上で入手可能である[11]。以下のような形式で作成

されている。

```
abbevillien   abbevillien   Adj:Mas+SG
abbevillien   abbevillien   Nom:Mas+SG
abbevillois   abbevillois   Adj:Mas+InvPL
abcédant      abcédant      Adj:Mas+SG
abcédant      abcder        Ver:PPre
abcède        abcéder       Ver:IPre+SG+P1:IPre+SG+P3:SPre+SG+P1
```

左端が見出し語，中央がレンマ，右端が品詞情報で，それぞれのフィールドはタブで区切られている。ただし，このレキシコンは問題がないわけではない。多少，正体不明の語が混じっている上に，たとえば最後のabcéde がそうだが，品詞情報をまとめてしまっているのは整合性がとれていないと言わざるをえない。紙幅の都合で実は少し後半をカットしてあるのだが，左から直説法現在一人称，三人称，接続法現在一人称と続いていると見当がつくので，接続法現在三人称と命令法二人称の二つが実はこの後に続いていると読者にも想像がつくだろう。事実そのとおりである。データ容量を少しでも圧縮するためには仕方のない措置だったとは思うが，これをフランス語品詞タグ付与に使うためにはさらに整形を施さなければならない。このような場合は，多少冗長になっても１見出し＋１レンマ＋１品詞情報に分割しておくべきだと考える。

フランス語レキシコンについては，筆者が独自に作成途中なので，完成したら報告したいと思う。



## 6. フランス語の同形異義語

作成途中のレキシコンを使って，問題のフランス語の同形異義語について最後に報告しておきたい。筆者に確認し得た限りでは，以下の動詞

は活用をさせると，そのどこかで必ず形態上のバッティングを引き起こす。ただし，そのバッティングの中には他の品詞とのケースは含まれていない。

accroire, accroître / aller, ailler / ardre, arder / braire, brayer / claqueter, claquetter / comparer, comparoir / croire, croître / décroire, décroître / dépeigner, dépeindre / embatre, embattre / faillir, failler / faillir, falloir / fondre, fonder / levreter, levretter / lire, liser / mégir, mégisser / moudre, mouler / ouvrir, ouvrer / patir, patisser / peigner, peindre / plaire, pleuvoir / poigner, poindre / rassir, rasseoir / rayer, raire / recouvrir, recouvrer / remoudre, remouler / rentraire, rentrayer / repeigner, repeindre / saillir, s ailler / être, suivre / surfer, surfaire /(se) tapir, tapisser / taveler, taveller / tripolir, tripolisser / vivre, voir / bruir, bruire, bruisser

あまりなじみのない単語も多いが，すでに述べたとおり，あまり見かけない単語であれば，貴重な出現を見過ごすことのないように，品詞タグ付与システムを慎重に設計しなければならない。この中で最も形態上のバッティングが少ないと思われる単純未来についてのみ調べたところ，以下の通り，同形異義語が含まれていた。

claquettera_Ver+IndF+P3+SG
claquetterai_Ver+IndF+P1+SG
claquetteras_Ver+IndF+P2+SG
（claqueter と claquetter）

faudra_Ver+IndF+P3+SG
（falloir と faillir（古い活用形で））

levrettera_Ver+IndF+P3+SG
levretterai_Ver+IndF+P1+SG
levretteras_Ver+IndF+P2+SG

(levreter と levretter)

surfera_Ver+IndF+P3+SG
surferai_Ver+IndF+P1+SG
surferas_Ver+IndF+P2+SG
surferez_Ver+IndF+P2+PL
surferons_Ver+IndF+P1+PL
surferont_Ver+IndF+P3+PL
(surfaire と surfer)

tavellera_Ver+IndF+P3+SG
tavellerai_Ver+IndF+P1+SG
tavelleras_Ver+IndF+P2+SG
(taveler と taveller)

以上の結果において，claqueter と claquetter，levreter と levretter，taveler と taveller はいわゆる異綴り（spelling variant）である。特に 19 世紀あたりまでの語彙にはこのような異綴りが多い。現代フランス語でもよく見かける代表的なものは clef と clé である。こうした異綴りのうち，どちらが現代にまで使われているのか，あるいはどちらも使われるのか，などの調査も含めて，他の活用形や他の品詞と同形になっているものがあるのかないのかなど，さらに広範囲に調査をする必要がある。

## 7. まとめ



コーパスとレキシコン。具体的な言語事実としてのデータの集合体がなければならないのと同時に，そのデータを分析するためのもっとも根幹となる辞書データも同時に必要である。厳密な話をすれば，実はレキシコンはコーパスから自動生成して，次にそのレキシコンを当てはめてコーパスを解析することによってレキシコンの信頼性を高めていくとい

う相互作用が必要なのである。ただし，一度に大規模なコーパスから巨大なレキシコンを作成することは難しいので，少しずつ精度を上げていくというステップバイ方式を取るのが妥当な方法である。

言語の自動処理の研究は英語や日本語ではすでに多くの成果をあげている。それはそれだけの基盤整備ができているからであって，たとえばコーパスに対して品詞情報を与えるだけでもどれだけの下準備が必要かがお分かりいただけたと思う。次回はさらにコーパスとレキシコンを整備して，その成果を報告するつもりである。

## 注

1) たとえば，Benoit Habert et al. (1997). pp17-18 で紹介されているタグ付きコーパスは，英米系で Brown，LOB，Susanne，London-Lund，Lancaster/IBM Treebank，Helsinki，Archer，BNC，Penn Treebank の 9 つに対して，フランスは Menelas，Mitterrand1，Enfants の 3 つだけで，いずれも非公開となっている。そのほかいくつかの研究機関でデータが作成されているようだが，残念ながら非公開のところが多い。それでなくてもデータ蓄積が少ない上に，非公開が多いのでは，事実上，存在しないのと同じである。
2) http://un2sg4.unige.ch/athena/html/francaut.html
3) http://humanities.uchicago.edu/ARTFL/ARTFL.html
4) http://etext.lib.virginia.edu/french.html
5) 齋藤俊雄他編著『英語コーパス言語学：基礎と実践』研究社，1998 年，p. 17.
6) Eric Brill が開発した，通称，Brill Tagger を応用する場合が多い。INaLF も Brill Tagger を利用している。
7) 従って，同じデータに対して何種類ものタグ付きデータが存在することになる。
8) ただし，現実には形態的解析によっても構文解析によっても品詞を決定できない場合がある。たとえばフランス語だと，Je vis avec Marie. という文で，vis が vivre の直説法現在なのか，voir の単純過去なのかは，文脈を見ないと決定できない。従って，厳密に言えば形態的解析，構文解析に続いて，文脈解析が必要になってくるのだが，ここではとりあえず

論じないでおく。というのが，現実のコーパスでそのような文脈依存でなければ品詞が決定できないような文がどの程度の割合で出現するのか（つまり，そうした例文は単に言語学者による捏造に過ぎない場合もある）を調べた上でなければ，労多くして益少ない研究になりかねない。

9) http://www.ipa.go.jp/STC/NIHONGO/IPAL/ipal.html

10) ibid.

11) http://cedric.cnam.fr/ABU/DICO/

## 参考文献


Karin Aijmer & Bengt Altenberg (Ed.), *English Corpus Linguistics*, Longman, London and New York, 1991.

Jenny Thomas & Mick Short (Ed.), *Using Corpora for Language Research*, Longman, London and New York, 1996.

Anne Wichemann et al. (Ed.), *Teaching And Language Corpora*, Longman, London and New York, 1997.

Benoit Habert, Adeline Nazarenko, Andre Salem, *Les linguistiques de corpus*, Armand Colin, Paris, 1998.

Graeme Kennedy, *An Introduction to Corpus Linguistics*, Longman, London and New York, 1998.

John Lawler & Helen Aristar Dry (Ed.), *Using Computers In Linguistics: A Practical Guide*, Routledge, London and New York, 1998.

岡田毅，『実践「コンピュータ英語学」：テキストデータベースの構築と分析』，鶴見書店，1995 年。

大津由紀雄他，『岩波講座　言語の科学 3　単語と辞書』，岩波書店，1997 年。

大津由紀雄他，『岩波講座　言語の科学 9　言語情報処理』，岩波書店，1998 年。

齋藤俊雄他，『英語コーパス言語学：基礎と実践』，研究社，1998。

上田博人，『パソコンによる外国語研究（Ⅰ）数値データの処理』，くろしお出版，1998 年。

上田博人，『パソコンによる外国語研究（Ⅱ）文字データの処理』，くろしお出版，1998 年。

溝口理一郎他，『大規模知識ベースに関する調査研究――オントロジー工学に関する調査研究――』報告書，財団法人　日本情報処理開発協会，1998 年。