

情報ファイル

オンラインデータベース— Frantext, Wordbanks*Online*, TLFi —

藤村逸子・内田充美

1.はじめに

2002年夏、オンラインで言語資料を利用できる Collins Wordbanks*Online* のフランス語版サービスが始まった。Collins Wordbanks*Online* はこれまで Cobuild*Direct* の名で英語のオンラインデータベース提供を行ってきたが、フランス語、スペイン語などへとサービスを拡大するとともに改称したものである。2001年春には、フランスの国家プロジェクトとして長い歴史のある Frantext が、1830年以降のテキストのほとんどに品詞情報を付けたバージョン(Frantext version catégorisée) を公開した。その結果、Frantext はフランス語研究のためのデータベースとして格段に利用価値が高まった。さらに2002年春には、Trésor de la langue française がインターネット上で無料で公開されるようになり(TLFi), この辞書を経由して Frantext の用例や Académie Française の辞書が検索できるようになった。

本稿は、フランス語を研究する上で、最近急速に利用しやすくなったこれらのデータベースについての情報を提供することを目的とする。実用的な情報に加えて、実際にこれらを利用したケーススタディを行い、データ利用の意義について検討する。

2. Frantext と Wordbanks*Online* の利用

まず、Frantext と Wordbanks*Online* 利用についての情報を表1に示す。

表1： Frantext と Wordbanks*Online* 利用情報

名称	Frantext	Wordbanks <i>Online</i>
情報ウェブページ	http://atilf.inalfr.fr/frantext.htm	http://titania.cobuild.collins.co.uk/wbinfo.asp
規模(約) ¹⁾	21400万語(品詞情報なし) 12700万語(品詞情報つき) 1950年以降は、3600万語(品詞情報なし), 3100万語(品詞情報つき)	7800万語
資料の内容	小説、学術書、戯曲など。 1500年代からの資料が、時代ごとに均等に存在する ²⁾ 。	新聞・雑誌(1990–2001年) ³⁾ と書籍45点 ⁴⁾
加入主体	大学など研究機関 ⁵⁾	研究目的であれば個人単位でも可
費用	1年間: 304.90ユーロ (2000仏フラン)	3カ月間(試用版): 50英£, 6カ月間: 300英£, 1年間: 500英£
登録方式	1契約で50個のIPアドレスを登録。 固定アドレスの振られた大学などのコンピュータからのみ利用可能。	1契約でアカウントが10個発行される。 契約者と同じ団体に属する人に割り振ることができる ⁶⁾ 。
接続方式	ウェブブラウザ	telnet接続 ⁷⁾ またはウェブブラウザ(専用プログラムのインストールが必要)

検索方法	Frantext独自の検索プログラム ⁸⁾ を用いる。検索式の書き方は独自のものだが一般的な正規表現 ⁹⁾ と共通する点が多い ¹⁰⁾ 。リスト作成という方法をとれば、語形を正規表現で指定することも可能。絞り込みはできない。	WordbanksOnline独自の検索式。アクセサン付きの文字の入力方法とともに簡単な説明が画面上に表示される。結果に対し、正規表現による絞り込み(grep)が可能。
活用	一括検索可	一括検索可
品詞情報	あり	なし
文脈情報	400字(キーワード含)	512字まで(キーワード含)
キーワード	マークつきで出力	マークなし。別途テキスト加工が必要。
検索結果の保存	ウェブブラウザからダウンロード	サーバ上に保存し、ftp ¹¹⁾ で取り込むことが可能。結果を指定のメールアドレス宛に送付するサービスもある。

いずれのサービスも、規模・内容ともに十分なレベルのものであると言えよう。言語研究の資料として見た場合の長短はおおむね表2のようにまとめられる。

表2：FrantextとWordbanksOnlineの特長と改善を期待したい点

	Frantext	WordbanksOnline
特長	<ol style="list-style-type: none"> 1. フランス 唯一の大規模データベースである 2. 著者名・書名などが明確 3. 品詞情報がある 4. 著作権のからんだ現代の資料を利用できるのはこのみ 	<ol style="list-style-type: none"> 1. 大規模である 2. ジャンル・資料の出所が明確 3. telnet 接続なので、環境を選ばない 4. 検索だけでなく、統計的分析の計算結果表示、コロケーション表示などができる 5. 研究機関単位でなくても契約可能 6. アカウントを振り分けられるので、実質的な費用負担は比較的少ない
改善を期待したい点	<ol style="list-style-type: none"> 1. 詳細な出典情報へのアクセスが面倒 2. 品詞情報の精度 3. サーバ上に何も保存できないので、接続のたびに全ての指定を新たに行う必要がある 	<ol style="list-style-type: none"> 1. 品詞情報がない 2. ジャンルが新聞雑誌に偏っている 3. 動詞活用情報の不備があるなど、ソフト面での問題¹²⁾

3. ケーススタディ: <de / des + 複数> の検索とその結果

ここではケーススタディとして、「複数形容詞+複数名詞」に先立つ不定冠詞についてデータベースを利用した調査の結果を報告する。「複数形容詞+複数名詞」に先立つ不定冠詞については、「desは複数形容詞の前ではdeとなるのが原則」(朝倉 2002: 66)とされてきた。しかし、実際にはdesが用いられる場合も少なくないことは周知の通りである。

初級の文法で取り上げられることも多い割には、実態があまり明確に記述されていないこの現象を、表3のデータを使って調査した。

表3：ケーススタディに用いた資料一覧

ジャンル	Frantext		WordbanksOnline	
	roman	traité ou essai	journal	revue
語数(約)	1600万語	1200万語	1450万語	510万語
内容	小説 (1950年以降)	専門的すぎない 学術書とエッセー (1950年以降)	<i>Le Monde</i> (1997年と2001年) <i>Libération</i> (1992-1993年)	<i>Actuel</i> (音楽雑誌) (1990-1992年) <i>Marie Claire</i> (女性誌) (1990-1993年)

2つのデータベース、4ジャンル合わせて約4760万語から、[avoir, être, prendreの各活用形または前置詞à, avec, par, dans, pour, sur] + [deまたはdes] + [beau, bon, grand, petit, nouveau, vieux, joli, jeune, mauvaisの男女の複数形] + [複数名詞¹³⁾] という条件に合致する用例を検索し、手作業で余分なものを取り除いた¹⁴⁾。得られた結果について、ジャンルの違い、形容詞の種類、リエゾンの有無の3つの側面から報告する。

3.1. ジャンル

まず、ジャンルの違いに着目して、de と des の分布を表4と図1に示す。

表4：ジャンル別 de と des の生起状況

ジャンル	Frantext		WordbanksOnline	
	roman	traité ou essai	journal	revue
用例数 de / des (100万語あたり)	785 / 228 (49.1) / (14.3)	540 / 39 (45.0) / (3.3)	737 / 155 (50.8) / (10.7)	342 / 128 (67.1) / (25.1)
des の割合	22.5%	6.7%	17.4%	27.2%

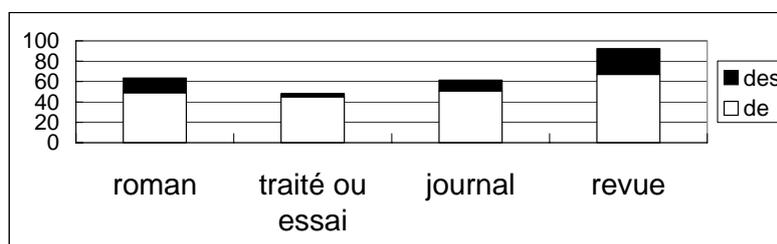


図1：ジャンル別100万語あたりの生起頻度

図1はdeとdesの100万語あたりの生起頻度をジャンル別に示したものである。学術書とエッセー(6.7%)、新聞(17.4%)に比較して小説(22.5%)と雑誌(27.2%)のジャンルでdesが多用されていることがわかる。このことは、比較的インフォーマルな状況でdes

が用いられる傾向があるという従来の記述をうらづけていると言えるだろう。

3.2. 形容詞の種類

次に、形容詞による分布を用例の実数で示す。

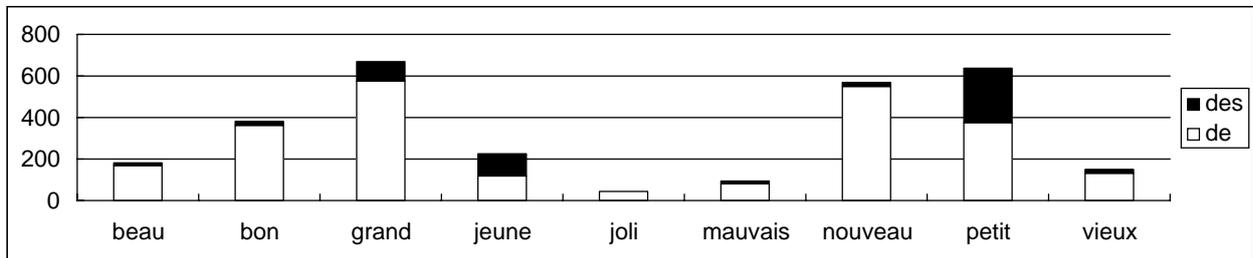


図 2：形容詞の種類と de / des

まず、名詞の前に来る形容詞として *grand*, *petit*, *nouveau* の多いことが確認できる。次に、*des* が用いられる比率に注目すると、*jeune* の 224 例中 105 (46.9%), *petit* の 639 例中 266 (41.6%) が高い割合を示していることが目を引く。一方 *des* の割合が低いのは *joli* の 2.3% (43 のうち 1), *nouveau* の 3.2% (569 のうち 18), *bon* の 4.5% (380 のうち 17) などである。5 件以上の用例が見つかった「形容詞+名詞」の組み合わせ 70 種類のうち、*des* としか共起しない例は、*des jeunes filles* (28 件) のみである一方、*de* としか共起しないものは *de bonnes conditions* (69), *de bonnes raisons* (39), *de grands yeux* (27), *de bonnes chances* (22), *de beaux yeux* (21), *de nouvelles bases* (21) など多数存在した。これらの事実については、形容詞と名詞でひとつの複合語をなす場合に *des* が用いられることがあるという従来の記述と照らし合わせてさらに観察・考察する必要がある。

3.3. リエゾンの有無

今回の調査結果では、従来の記述では触れられていない点として、形容詞と名詞の間のリエゾンの有無との関連を示唆する傾向が見られた。用例の分類にあたっては、名詞の先頭が母音か子音かによって、*h* や *y* で始まる語については、有音か無音かによって判別を行った。分類結果を図 3 に示す。

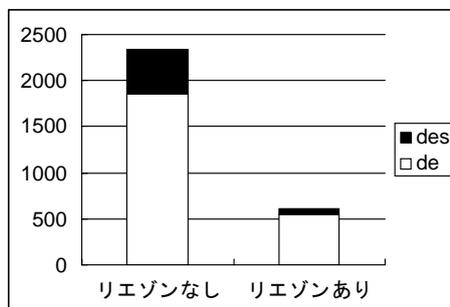


図 3：リエゾンの有無と de / des

形容詞と名詞の間にリエゾンがない用例 2339 のうち *des* をとっているものが 20.8% にあたる 486 例あるのに対し、リエゾンがあるものについては全 615 例のうち *des* を伴うも

のが 64 例と、10.4% にとどまっている。

「複数形容詞+複数名詞」に先立つ不定冠詞 *de* と *des* の交替については、伝統的にはジャンルによる傾向の相違、形容詞と名詞の繋がりや強さなど意味的要因 (Grevisse 1980, § 658–9) による説明などが試みられてきているが、ここでのリエゾンの有無による影響は、従来着目されていなかった音声面の要因の関与の可能性を窺わせる¹⁵⁾。

以上、オンラインデータベース *Frantext* と *WordbanksOnline* を利用して行った「複数形容詞+複数名詞」に先立つ不定冠詞についての実態調査の結果を、3つの側面から報告した。収集・分析の対象は「3つの動詞または6種の前置詞の後に出現した、9種類の形容詞+s, x, z で終わる複数名詞」という、限定された幅の用例ではあったが、*de* と *des* の使用実態についての一定の傾向を捉えることができた。

4. TLFi の特徴, TLFi の記載

TLFi (Trésor de la Langue Française informatisé : <http://zeus.inalf.fr/tlf.htm>) は、16巻、23000ページからなる辞書として1971年から1994年にかけて刊行された *Trésor de la Langue Française* (以下 TLF) のオンライン版である。TLFi は TLF の単なる電子化ではなく、むしろ TLF には、二つの別のバージョンが存在すると考えるべきである。TLFi は、更新され続けること、また、*Frantext* や *Académie Française* の辞書 (第8版と第9版 (A から M まで)) などへのリンクが充実していて、辞書のレベルを超えた用例検索が可能であるという特徴を持つ¹⁶⁾。*Frantext* そのものは、加入者だけが利用できるデータベースだが、TLFi を経由することにより加入者でなくても、*Frantext* のデータのうちで著作権の問題のないテキストを参照することができる。語彙項目の検索しかできないために構文の研究などには不向きという制約はあるものの、変化にとんだ実例を400字の文脈付きで手に入れることができるのは、言語研究者にとって大変魅力的である¹⁷⁾。

TLFi は複数の *de* と *des* に関しては、複合語に関する記載の他、次のように言っている。

Dans la lang. parlée fam. ou chez les écrivains qui l'imitent, *des* est souvent employé pour *de*, parfois avec la nuance de *certain*s. *Pourquoi des jeunes compositeurs se détournent de Beethoven* (BLOCH, *Dest. du S.*, 1931, p. 16). En partic. dans une énumération:

- 4. Il eut **des** danseuses qui le tutoyèrent, **des** bourgeoises qui lui dirent, mon prince; **des** grandes dames qui voulaient pécher sans déchoir. PÉLADAN, *Le Vice suprême*, 1884, p. 40.

レジスターの問題に加えて、「あるいくつか (いく人か) の」という意味や、文脈の関与を示唆している点が注目される。

5. まとめ

Frantext と *WordbanksOnline* は以上のように、フランス語研究においては有効活用すべき貴重な情報源である。リファレンスが明確である点、有料であるためサポートや管理が期待できる点において、その他の手作りのコーパスにはない利点がある。ジャンルの偏りという問題は、*Frantext* と *WordbankOnline* を組み合わせることによってカバーが可能である。

ただし、話し言葉がフランス語のデータベースに相変わらず欠けている点は残念である。TLFi に関しては、無料であり、インターネットに接続できる環境さえあれば、いつでもどこでも使えるのだから、これを利用しない手はないであろう。

(藤村: 名古屋大学、内田: 大阪女子大学)

【注】

- 1) Frantext, WordbanksOnlineとも、資料は更新され続けている。情報は2002年12月段階のものである。
- 2) <http://atilf.inalf.fr/artis/nvlbiblio.htm> に全ての書誌情報がある。
- 3) *Actuel* (1990–1992), *Marie Claire* (1990–1993), *Le Monde* (1997, 1998.4–2001.3), *Libération* (1992–1993)。
- 4) 資料の内容を記述した情報はftp://titania.cobuild.collins.co.uk/pub/WB_French/からダウンロードできる。
- 5) Frantextは年ごとの1年単位の契約を求めてくるが、国公立大学では会計年度単位の支払いしか可能でないため、会計上の困難がある。代理店を介入させることによりこの問題は解決できる。
- 6) 10個のうち1個を「教室利用アカウント」として申請することも可能。「教室利用アカウント」では、同時に複数の利用者がそのアカウントを用いてサービスを利用することができる(ただし、結果をサーバに保存してファイル転送することはできない)。
- 7) ここでは、自分のコンピュータからtelnetソフトウェアを使って遠くにあるWordbanksOnlineのサーバを操作することを指す。
- 8) 加入者でなくても<http://atilf.inalf.fr/agregation.htm>で一部を試してみることができる。(品詞情報はない。)
- 9) 文字情報をコンピュータを使って処理するための一般的な約束。たとえば [a-z] は任意の半角小文字のアルファベットを指す。
- 10) 検索式の例: (maison|palais)d'un(blanc|immacule|sale)bleu(d'azur|profond))
- 11) ここではftpソフトウェアを使ってサーバから自分のコンピュータにファイルを転送することを指す。
- 12) 2002年12月上旬の段階で、動詞faireの活用情報が機能していない、不規則動詞の活用形に実際に存在しない形式が登録されているなどの問題点が見られる。
- 13) WordbanksOnlineでは、品詞指定ができないので、実際には、s, x, zのいずれかで終わる単語という指定で検索し、不必要な用例は手作業で除去した。Frantextでは品詞情報を利用し、&e(g=S) (Sはsubstantif) と指定して名詞のみを検索した。品詞情報の利用に関しては、誤りが必ず含まれるという点に当然ながら注意すべきである。名詞と形容詞の区別は特に問題があり、手作業による修正が欠かせない。たとえば、ce sont des jeunes barbares qui piétinent les plates-bandes! (*Les boulevards de ceinture*, 1972)で、jeunes やbarbaresが名詞なのか形容詞なのかといった点に関して、品詞情報は全く信頼できない。
- 14) 形容詞と名詞がハイフンで繋がれたもの(cf. Grevisse 2001: § 569) については、今回の集計から除外している。不要な例の排除においてもっとも問題だったのは、être + de (前置詞) + 複数形容詞 + 複数名詞の選別であった。該当例は次の2例のみ: Tant que la particule est de petites dimensions par rapport à la longueur d' onde... (*Astrophysique*, 1963), ...quand les pièces sont de petites dimensions et nécessitent une épaisseur d' argenture constante et connue... (*Le verre*, 1966)。
- 15) リエゾンがあると複数性はリエゾンによって音的に示されるが、そうでない場合にはdesという明示的な複数冠詞が必要なのだという説明、あるいはリエゾンが起きると「形容詞+名詞」の音が単複で異なるため複合語としての扱いが難しくなり、その結果、一般の規則どおりdeが選択されるという説明などが可能かもしれない。
- 16) この他、検索の方法や、表示の方法にも工夫が凝らされている。たとえば、syntagmeに色をつけて表示することを選択すると、熟語だけに色が付き、他の部分との見分けが一目で付きやすい。大きな辞書の場合、このような工夫は現実的問題として重要である。
- 17) テキストごとに開くウィンドウから例文をコピー&ペーストして、自分のコンピュータに保存することが可能である。

【参考文献】

- GREVISSE, M. (1980): *Le bon usage*. 11^e éd., Duculot.
— (2001): *Le bon usage*. 13^e éd., Duculot.
朝倉季雄 (2002): 『新フランス文法事典』 白水社。