

# Lexical richness of French and Quebec journalistic texts

Natalia Dankova

Université du Québec en Outaouais

natalia.dankova@uqo.ca

## Abstract

This paper presents some results of a quantitative study that focuses on the variety and word frequency in texts from a comparative perspective. The study aims to analyze and compare French and Quebec journalistic texts on political and cultural topics written in French and recently published in major newspapers such as *Le Monde*, *le Figaro*, *Le Devoir*, etc.

The statistical analysis concerns the number of different words in the text, the number of different adjectives, the number of different verbs (and also passive structures, participles and gerunds which contribute to syntactic and stylistic sophistication), and the number of hapaxes.

French texts from France exhibit greater lexical richness and sophistication: they contain more adjectives, a greater variety of adjectives, as well as more participles and gerunds compared to French texts from Quebec.

The originality of the study lies in the fact that it analyzes variation in French using a lexicometric approach.

**Keywords:** lexical statistics, comparative analysis, journalistic texts, variation in French

## 1 Introduction

This quantitative ongoing study focuses on the variety and frequency of words in texts from a comparative perspective. This exploratory study aims to test the hypothesis regarding differences in the lexicon of verbs and adjectives between the two varieties of French.

The study analyzes and compares French and Quebec journalistic texts, written in French and recently published in major newspapers such as *Le Monde*, *le Figaro*, *France-Soir* (France), and *Le Devoir* and *La Presse* (Quebec).

Quebec French is often perceived as being less rich than standard French (i.e., French from

France). The feeling of linguistic insecurity remains strong among Quebec French speakers. Boudreau (2023) mentions that linguistic insecurity, common in the French-speaking world, stems from the way in which the French language has developed, from the idea of a single standard and a unitary, uniform vision of French

Written Quebec French, especially in texts intended for publication, adheres to a model known as international French, which shows minimal signs of regional variation. These differences are mainly lexical; syntax and grammar conform to the standard shared by all French-speaking countries. Consequently, one would reasonably expect few notable differences between French and Quebec texts.

However, our previous studies on variation in French (Dankova, 2002, 2009, and 2017) highlighted the choices of verbs that differ in oral narratives in French from France and in French from Quebec. These differences concern movement verbs and psychological verbs. Thus, we distinguished the following types: verbs of perception such as *regarder* ‘to look’, *voir* ‘to see’, *entendre* ‘to hear’, and verbs of cognition that describe mental representation and activity such as *comprendre* ‘to understand’, *penser* ‘to think’, *décider* ‘to decide’, and *se rendre compte* ‘to realize’. Verbs of visual or auditory perception are more frequent in narratives in French from France compared to those in Quebec French. This provides another avenue for investigating variations in journalistic texts.

Another study that we conducted on the use of adjectives in two types of narratives shows that French informants use a significantly greater variety and quantity of adjectives compared to Quebec speakers (Dankova, 2024). The linguistic means used are more diversified in the French corpus: we noted a greater number of different adjectives, a greater diversity of verbal forms and temporal adverbs.

In the present study, we aim to verify if the observed results hold true in journalistic texts.

## 2 Data and method

There are many lexical richness measures (Lebart and Salem, 1994, Brunet, 2003). The most popular one is the type/token ratio:  $V/N$  where  $N$  = number of tokens and  $V$  = number of types (Muller, 1979). This is the formula we use in our study. Roeland van Hout and Vermeer (2007) note that “lexical richness measures aim to reveal the richness of lexis used in a text, with the intention also of discovering the richness of the lexicon which produced the text. But all kinds of measures turn out to have reliability and/or validity problems” (Roeland van Hout and Vermeer, 2007: 136).

Although in linguistics there is a distinction between *lexicon* and *vocabulary* (de Saussure, (1995/1916), the term *lexical richness* is traditionally used in lexical statistics to measure the vocabulary rather than the lexicon of a language (see also Thoiron and Arnaud, 2008).

At the exploratory stage of the study, the corpus consists of 90 newspaper articles covering current political and cultural topics (45 articles written in each variety of French). This number will be increased in the next phase.

The corpus in French from Quebec (FQ) contains 35975 words and that in French from France (FF) 36999 words; the average length of the texts is respectively 799 words (FQ) and 822 words (FF).

Although it was not possible to select a large number of articles on a single subject, we opted for articles with common themes (post-Covid-19, major political and social events).

The statistical analysis concerns the number of different words in the text, the number of adjectives, the number of different verbs, and the number of hapaxes.

There is no consensus regarding the delimitation of the category of adjectives, the criteria for adjectivity or the classification of adjectives. The literature devoted to adjectives is abundant and reveals several paradoxes (Pottier, 1985; Goes, 2015; Van de Velde, 2009; Baylon and Mignot, 1995; Noailly, 1999; Riegel, 1985; Picabia, 1978; Wilmet, 1997).

When talking about adjectives, we mainly consider their functions rather than their forms, because only some of the adjectives constitute a class (or part of speech) called adjectives. For the

purposes of this study, we therefore take an inclusive stance. The main function of adjectives is to qualify nouns. There are qualifying adjectives (*une robe noire* ‘a black dress’), relational adjectives which can be replaced by a complement of the noun (*une décision gouvernementale* ‘governmental decision’ vs *une décision du gouvernement* (literally ‘a decision of the government’) ‘a government decision’), and referential or third type adjectives (*une carte bleue* (literally ‘a blue card’) ‘a credit card’).

We consider as adjectives nouns that are used as adjectives, for example, in *situation d’urgence* (literally ‘situation of emergency’), *d’urgence* ‘emergency’ plays an adjective role. On the other hand, adjectives that form part of compound nouns which designate, for example, institutions, political parties, or newspapers are excluded from the calculations: *la Cour Suprême* ‘the Supreme Court’, *la République Française* ‘the French Republic’, *le parti vert* ‘the green party’, *Le Monde Diplomatique* (literally ‘Diplomatic World’, name of a newspaper). Possessive adjectives which are considered determiners in the same way as articles are not taken into account either: *leur décision* ‘their decision’.

When analyzing verbs, we examine the number of different verbs and the most frequently used verbs, passive structures, participles, and gerunds which contribute to syntactic and stylistic sophistication; and types of verbs: psychological verbs which describe perception (*to see, to look*) and cognition (*to think, to realize*), and movement verbs (*to go, to come*).

A long indexing work preceded the analyses. We manually annotated adjectives and verbal forms because of the specificity of morphology. In French, verbal morphology is characterized by a large number of temporal forms, simple and compound, marked by endings and including alternations in verb roots. Some homonym forms can be verbs or nouns, adverbs, or adjectives: *pouvoir* ‘to be able / can’ vs *pouvoir* ‘a power’; *note* ‘(I, he, she) ‘note’ vs *note* ‘a note/a mark’, etc.

Adjectives agree in both number (singular or plural) and gender (masculine or feminine) with the noun or pronoun they modify. A limited number of adjectives have several masculine forms for historical reasons: *vieux (tableau)* ‘old (painting) vs *vieil (homme)* ‘old (man)’.

### 3 Discussion and results

#### 3.1 Adjectives

Adjectives are not always as essential as verbs or nouns. Qualifying adjectives appear when the speaker needs to express an inherent quality of the referent (Baylon and Mignot, 1995) and selects one of the dimensions which constitute the essential characteristics (van de Velde, 2009). Goes (2015) emphasizes that the argumentative context also requires the use of adjectives. Chengyu Fang and Cao (2009: 138) state on empirical basis “that adjective density is significantly correlated to degrees of formality of different text categories”.

The use of adjectives varies in the two corpora: in total 2765 adjectives in French from France and 1860 in French from Quebec, i.e. respectively 61.44 and 41.33 adjectives per text. Adjectives ratio is 0.0747 in FF and 0.0517 in FQ (Table 1).

	FF	FQ
Tokens (N)	36999	35975
Ratio (V/N)	0.0747	0.0517

Table 1: Ratio of adjectives

There is an example of the use of adjectives in an article published in France<sup>1</sup>:

- (1) *Fourbes et lâches* qu'ils sont, *tous*, ces «contrôleurs de la bien-pensance», se régalez à appliquer des «circulaires», jusqu'à l'idiotie *manifeste*, qui n'est pas sans rappeler le zèle *sadique* du pervers *narcissique* persécutant sa victime.

*Deceitful* and *coward* as they are, these “controllers of right-thinking” delight in applying “circulars”, to the point of *manifest* idiocy, which is reminiscent of the *sadistic* zeal of the *narcissistic* pervert persecuting their victim.

Among the most frequent adjectives common to both corpora are *social* ‘social’, *public* ‘public’, *d’urgence* ‘emergency’, *bon* ‘good’, *nouveau* ‘new’. The frequency of the adjective *d’urgence* ‘emergency’ is linked to the theme of Covid-19 in newspapers.

<sup>1</sup> Azalbert, Xavier (2023). Dégénérescence programmée. *France-Soir*, 2023, March, 27.

The corpus in French from France contains 1.55 times more different adjectives than the corpus in French from Quebec.

Our results confirm those of a previous study on the use of adjectives in two types of oral narratives which also shows that French informants use a greater number and variety of adjectives compared to Quebec speakers (Dankova, 2024).

#### 3.2 Verbs

According to the study known as Français Fondamental (‘Fundamental French’), the verbs *être* ‘to be’, *avoir* ‘to have’, followed by *faire* ‘to do/to make’, *dire* ‘to say’, *pouvoir* ‘to be able/can’, *aller* ‘to go’, *voir* ‘to see’, *savoir* ‘to know’, and *vouloir* ‘to want’ are among the most frequent words in literary French (Niklas-Salminen, 2015: 39).

Regarding the number of verbs and the number of different verbs, we observe no differences between the two corpora. The same observation applies to passive structures.

Among the most frequently used verbs in the French from France corpus, there are *être* ‘to be’, *avoir* ‘to have’, *faire* ‘to do/to make’, *pouvoir* ‘to be able/can’, *mettre* ‘to put’, *exprimer* ‘to express’, *tenir* ‘to hold’, *dire* ‘to say’, *devoir* ‘to have to/must’, *rendre* ‘to realize/to give back’, *considérer* ‘to consider’. The translation of the verbs into English does not reflect all the meanings present in French.



Figure 1: Ten most frequent verbs (FF) (designed with *Voyant Tools*)

In FQ corpus, the most frequently used verbs are *être* ‘to be’, *avoir* ‘to have’, *faire* ‘to do/to make’, *devoir* ‘to have to/must’, *dire* ‘to say’, *pouvoir* ‘to be able/can’, *mettre* ‘to put’; *voir* ‘to

see', *prendre* 'to take', *penser* 'to think', *opposer* 'to oppose'.



Figure 2: Ten most frequent verbs (FQ) (designed with *Voyant Tools*)

Verbs of perception (such as *to see*, *to look*, *to hear*) and cognition (*to think*, *to realize*, *to understand*) are 2.5 times more frequent in the French from France corpus (360 occurrences) than in the Quebec French corpus (143 occurrences).

While this does not strictly concern the lexicon, it is noteworthy to mention the use of participles and gerunds. Participles that are not part of compound tenses and do not function as adjectives, along with gerunds, predominantly appear in written texts, enhancing stylistic complexity, as illustrated in the following example<sup>2</sup>:

- (2) Certains parents dépassent les limites, *en exposant* leurs enfants à des applications de lecture.

'Some parents exceed boundaries *by exposing* their children to reading apps'.

The corpus in French from France contains 1.71 times more participles and gerunds compared to that in French from Quebec (Table 2):

	FF	FQ
Tokens (N)	36999	35975
Participles/gerunds (occurrences)	615	360
Number of occurrences per text (average)	13.67	8

Table 2: Participles and gerunds

Moreover, our previous studies on oral narratives in French from France and in French from Quebec highlighted differences in verbal morphology and the types of verbs used: verbs of perception and cognition are more frequent in French from France (Dankova, 2017 and 2009), and the verbal morphology is considerably richer compared to the Quebec corpus. (Dankova, 2002).

#### 4 Conclusion

This ongoing research enabled the identification of avenues for analyzing journalistic texts in two varieties of French. While the genre imposes a certain standard and it is the same language, our results indicate that French texts from France exhibit greater lexical richness and sophistication. They contain more adjectives, a wider variety of adjectives, as well as more participles and gerunds compared to French texts from Quebec.

Increasing the size of the corpora will enable us to validate the results obtained at the exploratory stage. Further analyses will be conducted in the future.

#### References

Christian Baylon and Xavier Mignot. 1995. *Sémantique du langage*. Nathan, Paris.

Annette Boudreau. 2023. *Insécurité linguistique dans la francophonie*. Les Presses de l'Université d'Ottawa, Ottawa.

Étienne Brunet. 2003. Peut-on mesurer la distance entre deux textes? *Corpus*, 2. <https://doi.org/10.4000/corpus.30>

Alex Chengyu Fang and Jing Cao. 2009. Adjective Density as a Text Formality Characteristic for Automatic Text Classification: A Study Based on the British National Corpus, *23rd Pacific Asia Conference on Language, Information and Computation*, pages 130-139 <https://aclanthology.org/Y09-1015.pdf>

Natalia Dankova. 2002. Conceptualisation et expression du temps en français québécois et en français de France. *Langues et linguistique*, 28: 47-70.

Natalia Dankova. 2009. Variétés de français. Conceptualisation et sélection de l'information à transmettre, in F. Martineau, R. Mougeon, T. Nadasdi, M. Tremblay, editors, *Le français d'ici. Études linguistiques et sociolinguistiques sur la variation du*

<sup>2</sup> *France-Soir* (2022). Les assistants vocaux nuisent au développement social et émotionnel des enfants. *France-Soir*, 2022, October 18.

*français au Québec et en Ontario*, GREF, Toronto, pages 15-35.

Natalia Dankova. 2017. Storytelling in French from France and French from Quebec. *Corela*, 15(2). <https://journals.openedition.org/corela/5051>

Natalia Dankova. 2024. Variation en français. Un cas des adjectifs. *Moderne Sprachen*, 66: 79-95.

Jan Goes. 2015. Types d'adjectifs et fonctions adjectivales: quelques réflexions. *Studii de lingvistică* 5: 293-322.

Lucien Lebart and André Salem (editors). 1994. *Statistique textuelle*. Dunot, Paris.

Charles Muller. 1977. *Principes et méthodes de statistique lexicale*. Hachette/Champion, Paris.

Aïno Niklas-Salminen. 2015. *La lexicologie*. Arman Colin, Paris.

Michèle Noailly. 1999. *L'adjectif en français*. Ophrys, Gap/Paris.

Lélia Picabia. 1978. *Les constructions adjectivales en français, systématique transformationnelle*. Librairie Droz, Genève/Paris.

Bernard Pottier. 1985. De l'adjectif. *Travaux de linguistique et de littérature*, 23/1: 301-305.

Martin Riegel. 1985. *L'adjectif attribut*. PUF, Paris.

Ferdinand de Saussure. 1995/1916. *Cours de linguistique générale*. Payot, Paris.

Philippe Thoiron and Pierre Arnaud. 2008. Quelques aspects de la perception de la richesse lexicale, *Cycnos*, 8.

<https://epi-revel.univ-cotedazur.fr/cycnos/392.pdf>

Danièle van de Velde. 2009. *Comment, manières d'être et manières de faire* ». *Travaux de linguistique*, 58/1: 39-61.

Roeland van Hout and Anne Vermeer. 2007. Comparing measures of lexical richness. In: H. Daller, J. Milton & J. Treffers-Daller (eds.), *Modelling and assessing vocabulary knowledge*, pages 93-115. CUP, Cambridge.

Marc Wilmet. 1997. *Grammaire critique du français*. Hachette-Duculot, Paris/Louvain-la-Neuve.

*VoyantTools*

<https://voyant-tools.org/>