Looi, J., Riget, P. N., Boulton, A. & Hassan, R. (in press). From theory to data: Testing introspective claims on synonymous French adjectives “prochain” and “suivant” using corpus-based methods. *International Journal of Corpus Linguistics*.

From theory to data: Testing introspective claims on synonymous French adjectives “prochain” and “suivant” using corpus-based methods

This paper presents a corpus-based study that evaluates variables identified introspectively by Berthonneau (2002) in relation to the alternation between two synonymous French *prochain* (“next” or “upcoming”) and *suivant* (“following” or “next”). Employing a multifactorial and behavioural profiles approach, we explore the asymmetry in their temporal and spatial applications. Our findings highlight distinct uses with temporal nouns, whereas uses with material nouns overlap. Binomial logistic regression and multiple correspondence analysis reveal a general divide between the deictic use of *prochain* and the anaphoric use of *suivant*, corroborating Berthonneau’s descriptions. Additionally, event predictability, which Berthonneau addressed only fleetingly, significantly influences form choice, with *prochain* often associated with general, eventual contexts and *suivant* with specific, predictable events. This study contributes to the understanding of how corpus-based methods can refine existing linguistic hypotheses by illuminating the intersections and divergences between empirical and introspective research.

Keywords: French linguistics, adjectives of the third type, introspection, multiple correspondence analysis, binomial logistic regression

1. Introduction

The debate between corpus linguistics and introspective methods has swung back and forth throughout linguistic history. Chomsky (1979: 57) famously criticised the 1950s approach of linguists as mere 'butterfly collecting', gathering isolated data points without broader analysis. Modern corpus linguistics has significantly evolved from these early criticisms, yet the divide remains, with each approach maintaining its proponents and critics. Efforts to reconcile the two have met with limited success, witness the Bootcamp debate in 2010 in this journal (Worlock Pope, 2010).

The debate persists over the nature of evidence. For instance, Teubert (2010: 356) criticises the reliance on introspective data, arguing that "Corpus linguistics is not about modelling an ever-elusive language system, whether situated in society… or in the mind, as mentalist and cognitive linguists claim." This debate extends across various linguistic disciplines, including generative and cognitive linguistics, which discuss the reliability (e.g. Gibbs, 2006) and evidential status (e.g. Schindler, Drożdżowicz & Brøcker, 2020) of intuitive data.

In French linguistics, the blend of introspection and corpus data is typical; for example in research on the adjectival category 'adjectives of the third type'. The term was first used in the 2002 special issue of *Langue Française* (Schnedecker, 2002), where qualitative analysis of observed data, often without statistical methods, was combined with introspective examples to address data gaps and illustrate impossible/unacceptable adjective usages. This method continues to be used in recent studies (Goes, 2021, 2022; Yamamoto, 2020).

This article re-examines the variables identified in Berthonneau (2002), one of the contributions to Schnedecker (2002), in the context of synonym alternation between *prochain* ("next" or "upcoming") and *suivant* ("following" or "next"). This article investigates the possibility of using solely corpus data and statistical methods to replicate Berthonneau's study and demonstrate how corpus-based methods can be effectively applied in studies of adjectives of the third type and French linguistics. This study thus extends beyond French, comparing introspection and corpus-based methodologies.

1.1 Adjectives of the third type

Feuillet (1991:47) identified a group of adjectives, distinct from the traditional qualitative and relational adjectives, and termed them 'situational' as they denote predominantly place, time,

and existence of the modified entity. Examples (1) to (3) illustrate respectively the three types of situational adjectives *proches, imminent,* and *réelles*.

(1) des zones proches de l'explosion[1]
("zones near the explosion")

(2) un orage imminent
("an imminent storm")

(3) L'IBCT souffre de limitations réelles.
("IBCT suffers from real limitations.")

These adjectives are often considered "other" in traditional grammar (e.g. Noailly, 1999) or omitted in linguistic analysis (e.g. Forsgren, 1978) due to their non-prototypical nature and unclassifiability under the two traditional categories: qualitative and relational adjectives. Qualitative adjectives describe a characteristic of the noun, such as form, dimension, colour, and property (Riegel et al., 2004: 355-356), as in (4) with *fade* and *ennuyeux* describing the properties of the era. Relational adjectives, derived from nouns, express a relationship with the entities those nouns denote (Riegel et al. 2004: 357). For instance, *napoléonniens* in (5) relate to the noun *Napoléon*, the former French emperor.

(4) Cet âge est nécessairement fade et ennuyeux.
("This era is necessarily bland and boring.")

(5) le héros des légendes napoléoniennes
("the hero of Napoleonic legends")

The classifying framework of these two traditional categories appears to be inadequate. 'Situational' adjectives are currently more commonly referred to as 'adjectives du troisième type' or 'adjectives of the third type', a distinct category now formally considered in French

[1] All examples cited were extracted from the 10-million-word reference corpus Corpus d'Étude pour le Français Contemporain (CEFC), unless otherwise specified.

grammar (cf. Riegel et al. 2009). However, the definition of this category remains vague and relies on elimination for the identification of adjectives of the third type:

> "Adjectives of the third type", a convenient label grouping a set of adjectives like *futur* in *notre futur gendre* or *simple* in *simple formalité*, **which are neither relational nor qualitative**, even though they often constitute specific uses of the latter (*une pure calomnie* vs *un coeur pur*). Exclusively attributive and non-gradable, they are generally placed before the noun they modify and do not specify its semantics but rather modulate, each in its own way, the relationship of the nominal group they belong to with its referential counterpart. (Riegel et al. 2009: 599; translation and emphasis added)

Research in relatively recent years into adjectives of the third type focuses on identifying their idiosyncrasies (e.g. Goes, 2021, 2022), commonalities (e.g. Schnedecker, 2002) and differences within the category (e.g. Berthonneau, 2002). One frequently cited example of adjectives of the third type is *prochain* due to its polysemy and complex placement (cf. Benzitoun et al., 2010).

Comparing the referential functions of *prochain* to those of *suivant,* also an adjective of the third type, Berthonneau (2002) identified the asymmetry between the temporal and spatial uses of *prochain* and also the concurrence between *prochain* and *suivant* when modifying material nouns. With material nouns such as *station* in (6) to (8), regardless of its position, *prochain* can be used interchangeably with *suivant* because *la station prochaine* and *la prochaine station* refer to the same entity as *la station suivante*. However, such parallelism is not observed in the temporal uses of *prochain*: *la semaine prochaine* in (9) refers to the week that follows the moment of utterance; *la prochaine semaine de stage* in (10) to the following week during an internship; and lastly, *la semaine suivante* in (11) to the week following Christmas. Evidently, the two adjectives *prochain* and *suivant* produce distinct temporal references and the positional difference of *prochain* also bring about semantic changes.

(6) Je descends à la station prochaine. [2]
("I get off at the upcoming station.")

[2]Examples (6) – (11) were borrowed from and produced by Berthonneau (2002) via introspection.

(7) Je descends à la prochaine station.
("I get off at the next station.")

(8) Je descends à la station suivante.
("I get off at the following station.")

(9) Je viendrai la semaine prochaine.
("I will come next week.")

(10) la semaine prochaine de stage
("the following week of internship")

(11) Paul viendra à Noël et Jules la semaine suivante.
("Paul will come at Christmas and Jules the following week.")

To account for the usages of *prochain* and the choice between *prochain* and *suivant*, Berthonneau (2002) proposes temporal entities lacking salience in communicative situations, temporal nouns' sequentiality, repetitiveness and individuality, temporal entities' predictability, and the closedness of topological inclusion of temporal/material entities as the variables influencing how *prochain* is used and when *suivant* is presented as an alternative to *prochain*.

1.2 Research objectives and questions

The present article revisits key variables identified by Berthonneau (2002) through introspection, employing inferential statistical methods and corpus-based analysis to compare introspective arguments with empirical and corpus-based findings. The study is guided by three research questions:

(i) Are the variables proposed by Berthonneau (2002) statistically associated with the choice between *prochain* and *suivant*?
(ii) How do these variables interact to influence the choice of *prochain* and *suivant*?
(iii) To what extent do the arguments made through introspection align with the findings from statistical analysis and corpus evidence?

The research investigates the alternation between the adjectives *prochain* and *suivant*, highlighting a broader interest in linguistic alternation within contemporary corpus linguistics (e.g. Xu, Li & Szmrecsanyi, 2024; Liu & Dou, 2023). This investigation also fills a gap in empirical research on French adjectives, enriching understanding through advanced statistical methods and corpus analysis. This approach aligns with Gries and Deshors' (2014) endorsement of multifactorial statistical analyses of corpus data to dissect the nuances of French adjective usage and inform broader linguistic theories about linguistic choices.

Following this introduction, the next section summarises Berthonneau's (2002) study on the key variables affecting the choice between *prochain* and *suivant*. It outlines the methodology, leading into sections that discuss results from Bayesian binomial logistic regression and multiple correspondence analysis (MCA). The article concludes with a comparison of introspective and empirical findings.

2. Starting point: Berthonneau (2002)

This section summarises Berthonneau's (2002) study, which underpins this article. It also incorporates examples from the Corpus d'Étude pour le Français Contemporain (CEFC) (see Section 3) to preliminarily assess her introspective descriptions against natural language usage and to help operationalise her proposed variables.

Berthonneau (2002: 105-106) defines *prochain* as deictic, oriented around the moment of utterance (T0) and indicating the future, while *suivant* is anaphoric, tethered to an objective time reference in event chronology, applicable to both past and future. For example, *l'an prochain* in (12) refers to the year after the writing year, dependent on T0, while *l'année suivante* in (13) specifically refers to 1999, following 1998 mentioned explicitly.

(12) La croissance française devrait monter à 3% l'an prochain, au-dessus de la moyenne de la zone euro (2.9%).
("French growth should rise to 3% next year, above the euro zone average (2.9%).")

(13) En 1998,… Il finit 3^{e} du Tour d'Espagne. L'année suivante, il est 5^{e}…
("In 1998,… He finished 3rd in the Tour of Spain. The following year, he was 5th…")

The deictic nature of *prochain* is contingent on three conditions (Berthonneau, 2002: 107-108). It must be postpositional as anteposition changes its meaning from *next* to *upcoming*. The nouns it modifies should suggest a sequential structure, thereby excluding unique entities like *2001* or *le 19ème siècle* and also days or parts of a day. It requires a definite article, except with days of the week where an indefinite article imposes the meaning *upcoming*.

For example, *prochain* in *les prochaines années* (15) loses its deixis when used in anteposition, referring vaguely to years to come, unlike (14). Its non-deictic use is evident when modifying terms like *mort* (16), where it conveys an imminent event rather than a recurring one, underlining the principle of sequentiality. The presence of a definite article is crucial for its deictic function; without it, as in *un jour prochain* (17), the reference becomes non-specific.

(14) Mon dernier va entrer en sixième l'année prochaine.
("My youngest is going into sixth grade next year.")

(15) Nous avons un projet ambitieux pour les prochaines années.
("We have an ambitious project for the coming years.")

(16) Elle prévoit dès lors les malheurs qui la menacent, annonce sa mort prochaine.
("She therefore foresees the misfortunes that threaten her and announces her imminent death.")

(17) M. de Candolle prévoit même qu'un jour prochain la profession de savant et celle de professeur, aujourd'hui encore si intimement unies, se dissocieront définitivement.
("Mr. de Candolle even predicts that one day soon the profession of scholar and that of professor, still so closely united today, will dissociate definitively.")

2.1 Conditions on temporal nouns

For *prochain* to function deictically, it must not only satisfy specific conditions but also modify temporal nouns that inherently suggests a repetitive, sequential structure (Berthonneau, 2002: 110-112). This presupposes a preceding event of the same type, necessitating the topological inclusion of these nouns (e.g. Monday–Sunday), a predetermined order (e.g. Monday preceding Tuesday), and a defined structure (e.g. a week comprising seven days). Example (18) demonstrates *mardi* fitting these criteria. In contrast, nouns like *moment* or *instant*, lacking a

fixed duration and structure, prevent *prochain* from maintaining its deictic meaning, shifting it to imply imminence especially when used with an indefinite article. In the 10-million-word CEFC, *moment* and *instant* are never paired with *prochain*, somewhat corroborating Berthonneau's introspection. Additionally, *moment* is never paired with *suivant*, and *instant* appears with *suivant* only three times, as seen in (19). However, their absence in the corpus does not rule out their possibility.

(18) Leurs petits déjeuners hebdomadaires reprendront mardi prochain.
("Their weekly breakfasts will resume next Tuesday.")

(19) Les positions et les vitesses des trois corps à un instant donné suffisent pour déterminer leurs positions et leurs vitesses à l'instant suivant...
("The positions and velocities of the three bodies at a given instant are sufficient to determine their positions and velocities at the following instant…")

Berthonneau (2002: 119-120) also notes that *prochain* cannot modify non-individualised temporal nouns like *jour* or parts of a day (e.g. *midi, après-midi*) or an hour (e.g. *minute, seconde*). Similarly, *page* differs from *chapitre* in that it lacks individualised status as it is defined by the numbering system rather than T0. This distinction necessitates the use of the anaphoric *suivant* for *page* in (20), while *chapitre* appropriately takes the deictic *prochain* in (21).

(20) Vous trouverez plus d'information à la page suivante.
("You will find more information on the following page.")

(21) Dans le prochain chapitre, nous allons examiner les différentes stratégies d'allocation des coûts de congestion.
("In the next chapter, we will examine the different congestion cost allocation strategies.")

Furthermore, *prochain* only modifies singular nouns to retain its deictic function (Berthonneau, 2002: 115-116). When applied to plural nouns, as in (15), *prochain* is anteposed and loses its deictic property because plural nouns do not cumulate successive entities nor are they

individually identified in relation to T0. In (22), for instance, the committee's review does not suggest a continual daily action but an unspecified future event.

(22) La commission des Affaires étrangères examinera dans les prochains jours le traité Franco-Russe sur l'adoption.
("The Foreign Affairs Committee will examine the Franco-Russian adoption treaty in the coming days.")

2.2 Spatial uses of *prochain* and *suivant*

When used with nouns denoting spatial or material entities considered as places, the usage of *suivant* and *prochain* often overlap, though they may not mean exactly the same when modifying spatial nouns (Berthonneau, 2002: 104), as demonstrated in (6) to (8). Nonetheless, finding exact corresponding examples in a corpus is challenging, if not impossible.

In the CEFC, *prochain* appears just 9 times with spatial nouns out of 1309 instances. A notable pair from the CEFC illustrates the adjective choice with the noun *station*, where in (23), *suivant* is anaphoric, referencing the previous station *Mention-Garavan*, while in (24) *prochain* is deictic, indicating the next station along a waterway from the speaker's current position.

(23) Vous auriez pris le train devant vos amis à Menton-Garavan, mais vous en seriez descendu à la station suivante qui est celle de Menton...
("You would have taken the train in front of your friends at Menton-Garavan, but you would have gotten off at the following station which is Mention…")

(24) On entendit au loin, très loin, tout au fond du val, le son rauque de la trompe... pour prévenir l'éclusier de la station prochaine.
("We heard far away, very far away, at the very bottom of the valley, the hoarse sound of the horn… to warn the lock keeper of the next station.")

2.3 Salience of the referents in communicative situations

Berthonneau (2002: 121-134) attributes the overlapping usage of *prochain* and *suivant* in modifying spatial nouns, as opposed to their distinct applications with temporal nouns, to the lack of salience with temporal nouns. Material entities like train stations are inherently salient,

i.e. immediately accessible in a communicative situation, while temporal entities like days or months lack immediate salience, existing as intangible coordinates identifiable only by names (e.g. *lundi, semaine*) or and needing overt identification.

Salience is a key concept in Berthonneau (2002). She defines a 'salient entity' as a referent that acts as a time marker for anaphoric adjectives when the referent has not been explicitly mentioned before (Berthonneau, 2002: 123). The overlap between prochain and suivant therefore only occurs in spatial contexts where material entities are directly identifiable during communication. Conversely, in temporal contexts, *suivant* requires a specified time reference (Ts) as in (13), distinct from T0, while *prochain* typically marks the first entity following T0. This clarifies why temporal nouns are usually modified by the deictic *prochain* in postposition. However, when a temporal entity is explicitly marked by an external factor, like *des incendies* in (25), *prochain* is used in anteposition and loses its deictic nature. Furthermore, when *prochain* is non-deictic and anteposed, it can be modified by an intensifier like *tout*, shown in (26).

(25) La prochaine saison des incendies, entre novembre et janvier, pourrait également battre les tristes records de 2001.
("The next fire season, between November and January, could also break the sad records of 2001. ")

(26) La barre symbolique de 10% semble devoir être irrémédiablement franchie dans les tout prochains mois.
("The symbolic bar of 10% seems likely to be irremediably crossed in the very following few months.")

Differences in the use and placement of *prochain* between material and temporal nouns stem from whether the nouns denote sequential entities within a closed or open group (Berthonneau, 2002: 118-119). *Prochain* is anteposed to denote a specific entity within a closed group, as demonstrated in (27) where there is a limited number of "Lord of the Rings" episodes, and (28) where a specific weekend during Easter in Brittany is referenced. In contrast, *prochain* is postposed in open group like in (29) where a non-specific weekend among potentially countless weekends is referred to. Noun modification, thus, forms a closed group of entities and also provides an explicit time reference.

(27) Rien ne pourra plus arrêter Frodo ; dans le prochain épisode, Frodo continue à semer la panique.
("Nothing can stop Frodo; in the next episode, Frodo continues to cause panic.")

(28) Il arrive, c'est sûr, en brillante forme pour le Challenge national qui se déroule ce prochain week-end de Pâques en Bretagne.
("He is arriving, for sure, in brilliant form for the National Challenge which is taking place this upcoming Easter weekend in Brittany.")

(29) Le week-end prochain, le tir campagne à Torpes reprendra ses droits.
("Next weekend, field shooting in Torpes will resume its rights.")

Summarising the current section, the dissociation between *prochain* and *suivant* aligns with the broader differentiation between deictic and anaphoric adjectives. *Suivant* is less likely to overlap with *prochain* in temporal contexts due to the lack of salience. However, they can compete in scenarios involving temporal nouns if the temporal position is calculable (as required by *prochain*) and another temporal entity serves as an anaphor (as required by *suivant*) (Berthonneau, 2002: 122).

Berthonneau's key concepts were operationalised into thirteen categorical variables, with production mode added as another variable, detailed in Table 1. The discussion and examples from the CEFC illustrate how these variables were annotated. Given that many of these variables are binary, there may be concerns about their sufficiency to fully capture the subtleties of introspective results; however, they are consistent with Berthonneau's conceptual framework. It is also important to note that the variable meaning is binary, consisting of two categories: *next* and *about to come* for simpler annotation. *Next* refers to an entity *x* that may or may not exist at T0 but is similar to a pre-existing entity *x*; *about to come* refers to the first occurrence of an entity *y* after T0 (Berthonneau, 2002: 116).

Table 1: Variable operationalised based on Berthonneau (2002) with the first level shown in the list as the reference level in the logistic regression model

Variable	**Levels**
form	Dependent variable
	Binary: *prochain*, *suivant*
mode	Binary: spoken, written
meaning	Binary: next, about to come

number	Binary: singular, plural
temporality	Binary: non-temporal, temporal
aspect	4 levels: present, future, conditional, past
determiner	4 levels: absent, definite, indefinite, possessive
sequentiality	Binary: non-sequential, sequential
repetitiveness	Binary: non-repetitive, repetitive
individuality	Binary: non-individualised, individualised
time reference	Binary: T0 (moment of utterance), Ts (time specified)
closedness	Binary: open, closed
salience	Binary: non-salient, salient
predictability	3 levels: confirmed, possible, probable

3. Methodology: corpus and statistical methods

Having discussed Berthonneau's (2002) variables in the previous section, this section describes the corpus and statistical methods used to examine the association between these variables and the choice between *prochain* and *suivant*. The study utilises the CEFC, a balanced 10-million-word reference corpus of both spoken and written French. The corpus is composed of twenty different sub-corpora, each stemming from distinct research projects, yet conceptualised in a broadly comparable way. The 4-million-word spoken component, drawn from fourteen sources, includes recent recordings from over 2,500 adult French speakers across France, Switzerland, and Belgium. It captures a variety of speech contexts, ranging from informal conversations to interactions with public services, as well as public speeches, academic discussions, and corporate meetings. The 6-million-word written component comprises material from six sources, including classic French literature, regional and national newspapers, scientific publications, and more informal forms of writing, such as SMS, tweets, and blogs. For further details on the corpus, please consult the presentation page of the corpus[3].

Data were extracted from the CEFC using TXM, with the keywords *prochain* and *suivant* with their inflected forms (*prochain.e.s.* and *suivant.e.s*) to capture gender and number agreements. The final dataset included 1,309 instances of *prochain* and 1,018 of *suivant*, with their occurrences in the written component nearly equal (117 and 118 per million words, respectively). In contrast, in the spoken component, *prochain* (109 pmw) is nearly as frequent as in the written, while *suivant* is significantly less frequent (25 pmw). A 10% representative sample of these occurrences was randomly selected aligned with the diamesic proportions

[3] For more information about the CEFC, see https://repository.ortolang.fr/api/content/cefc-orfeo/11/documentation/site-orfeo/index.html

using Excel's RAND() function for manual annotation based on the criteria in Table 1. Details of the dataset analysed are summarised in Table 2.

Table 2: Summary of the dataset analysed

Form	Production mode	Original counts	Annotated counts (10%)
Prochain	Written	917	92
	Spoken	392	39
Suivant	Written	927	93
	Spoken	91	9
TOTAL		**2327**	**233**

3.1 Bayesian binomial logistic regression

This section details the analysis of the fourteen variables identified in Section 2, utilising the behavioural profiles approach by Gries and Divjak (2009). The aim is to explore how these variables influence the choice between *prochain* and *suivant*. To address RQ1 regarding the statistical association of these variables, the annotated data underwent a regression analysis.

Given that our dependent variable – the choice between the two adjectives – is binary, binomial logistic regression was selected as the most suitable model. This choice not only fits the binary nature of the data but also facilitates a multifactorial analysis, supporting Gries and Deshor's (2014) advocacy for in-depth linguistic analyses. We opted to use Bayesian statistics, rather than the more conventional frequentist methods, even though it might seem excessive for this study. This decision aligns with our broader goal of pushing methodological boundaries in linguistics and exploring alternative approaches and frameworks. Bayesian inference combines observed data with a fitted model to generate a posterior distribution, effectively incorporating data uncertainty and facilitating simulations of future outcomes (Gelman, Hill & Vehtari, 2020: 113). While frequentist methods focus on the probability of observing the data given the null hypothesis, the Bayesian approach estimates the probability of a hypothesis being true given the data; in other words, Bayesian analysis assesses the likelihood of a hypothesis based on the data, whereas frequentist methods calculate the probability of obtaining a dataset as extreme as the one observed, based on *p*-values (cf. Fornacon-Wood, et al., 2021). This methodology also responds the increasing call to go beyond the conventional *p*-value interpretation (e.g. Wasserstein et al., 2019) by shifting the focus from black-and-white judgments of 'statistically significant' results to a more nuanced assessment of effect sizes, conditional on statistical difference from zero (Gelman et al., 2020: 57-59).

The regression model was developed following a forward selection procedure recommended by Gelman et al. (2020). Starting with a null model that only included the response variable, variables were added incrementally until no further improvement was noted. Leave-one-out (LOO) cross-validation was used at each step to mitigate overfitting. The final model included ten variables and demonstrated significant improvement in fit and explanatory power, as evidenced by the expected log predictive density scores improving from -154.79 to -72.39 and the LOO R^2 values from 0.06 to 0.69.

The regression results, summarised in Table 3, include the variables' median posterior estimates and 95% credible intervals. Our regression assumptions were tested using the R package DHARMa, presenting diagnostics tailored for binary data. Figure 1 displays a boxplot highlighting any deviations from uniformity. Within-group deviation from uniformity test and Levene test for homoscedasticity were performed, both resulted in non-significant outcomes. Additional tests conducted with DHARMa, including Kolmogorov-Smirnov test for normality (p=0.94), dispersion test (p=0.33) and outlier test (p=1), also indicated non-significant findings. Following Gabry et al. (2019), Posterior predictive checks (PPCs) involved comparing observed data (y) with simulated data (yrep) from the model's posterior distribution. Figure 2 shows this comparison, highlighting congruence between the simulated and observed data, indicating a well-calibrated model.

Table 3: Regression results output for the final model predicting the choice between prochain and suivant

Variable	**Estimate (posterior median)**	**MAD_SD**	**95% credible intervals**	**Decision**
(Intercept)	-3.00	1.33	(-5.837, -0.443)	Reject Null
mode-written	-1.96	0.97	(-3.872, -0.042)	Reject Null
meaning-about to come	3.25	0.71	(1.889, 4.779)	Reject Null
number-plural	2.96	0.89	(1.390, 4.827)	Reject Null
temporality-temporal	-4.96	1.31	(-7.700, - 2.523)	Reject Null
aspect-future	-2.10	0.67	(-3.447, -0.862)	Reject Null
aspect-conditional	-3.56	1.18	(-6.060, -1.411)	Reject Null
aspect-past	-2.60	1.04	(-4.717, -0.742)	Reject Null
sequentiality-sequential	2.67	0.99	(0.772, 4.713)	Reject Null
individuality-individualised	1.70	0.88	(0.036, 3.488)	Reject Null
time reference-Ts	7.20	1.20	(4.995, 9.761)	Reject Null
salience-salient	2.36	0.80	(0.908, 3.983)	Reject Null
predictability-possible	-3.06	1.23	(-5.652, -0.900)	Reject Null
predictability-probable	-3.03	0.89	(-5.062, -1.368)	Reject Null

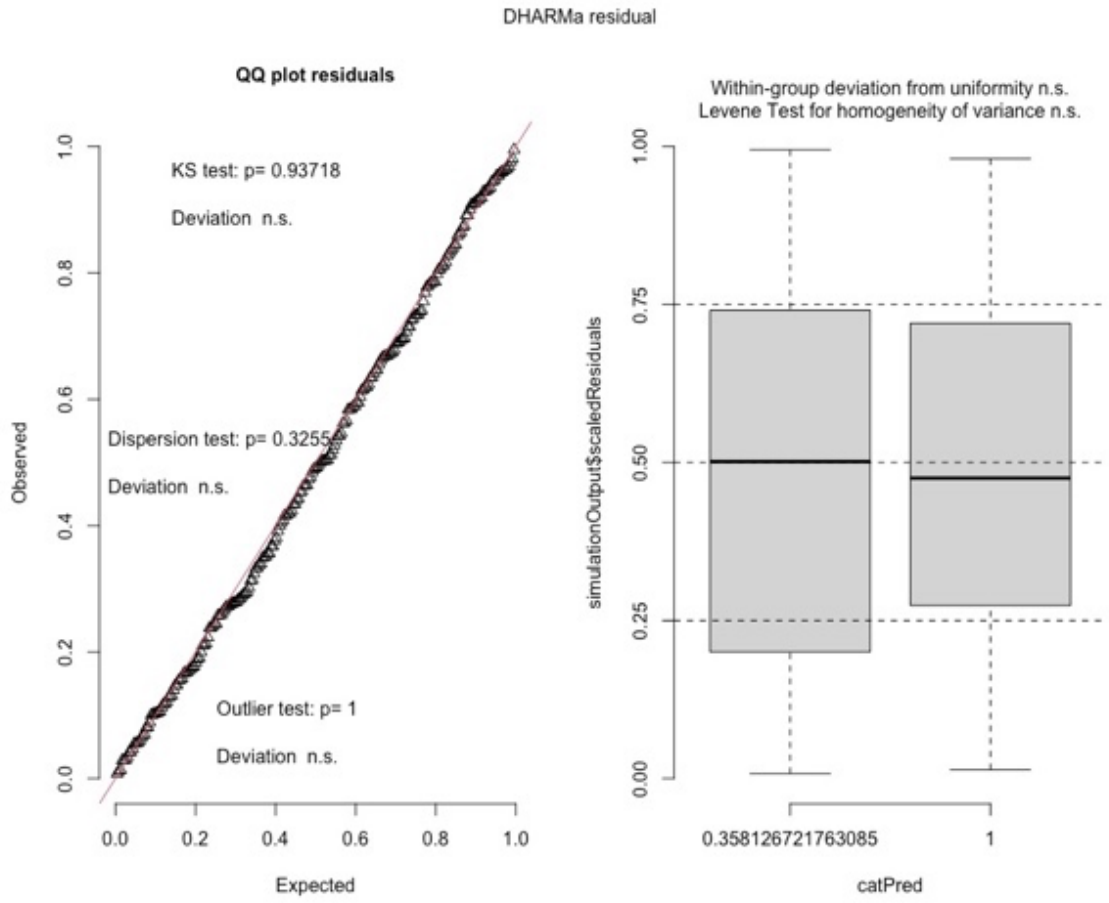

Figure 1: DHARMa residual testing output

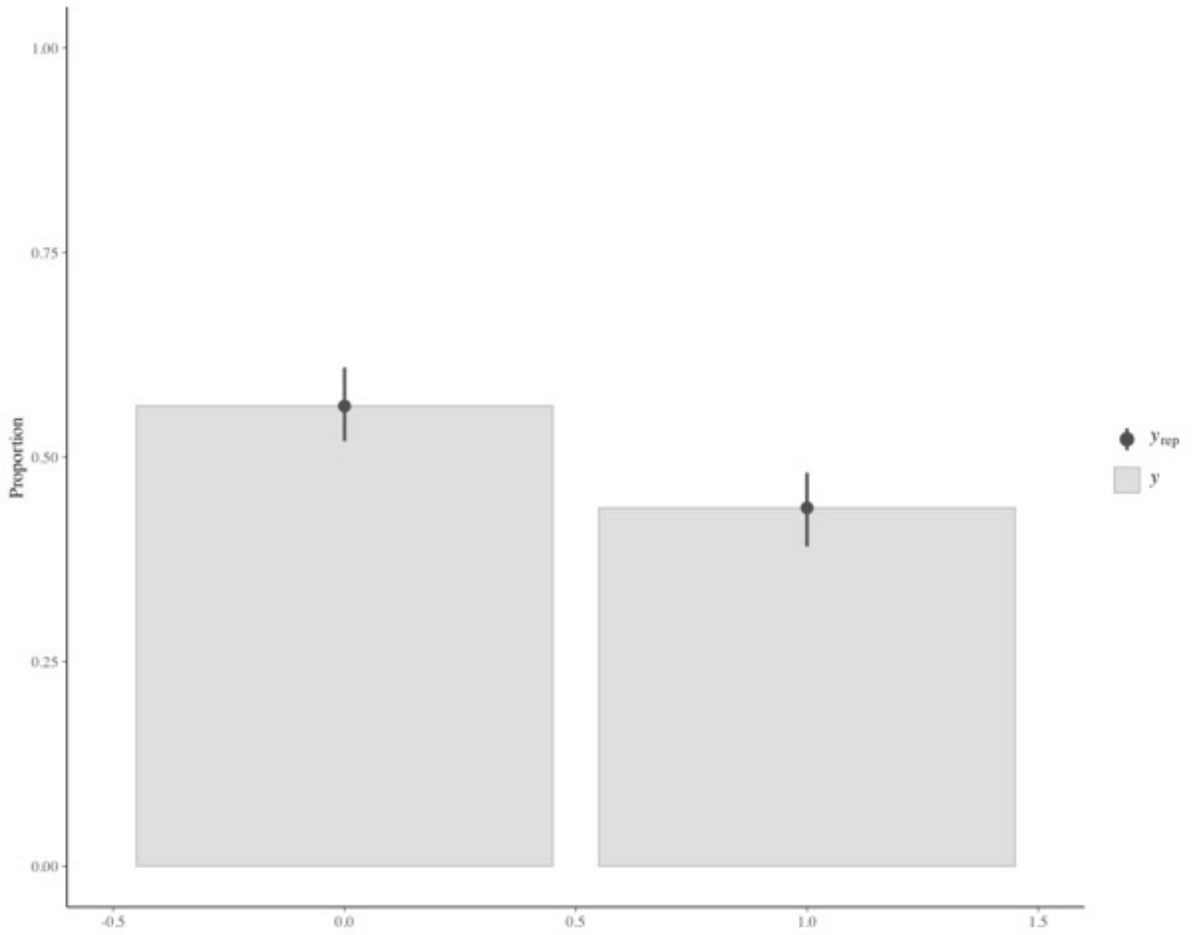

Figure 2: Posterior predictive check (PPC) for the final model as a whole

3.2 Multiple Correspondence Analysis for results summarisation

To address RQ2 regarding how the fourteen variables relate to the choice between *prochain* and *suivant*, Multiple Correspondence Analysis (MCA) was utilised. MCA, an effective tool for examining the relationships between categorical variables (Levshina, 2015: 375-376), summarised and visualised the interactions among variables that showed statistical associations with the form choice in the regression analysis (see Figure 3). By analysing variables retained in the regression model, MCA provided deeper insights into the complex, multifactorial patterns within the data, enhancing the interpretability of the regression outcomes.

MCA typically reveals a low proportion of explained variance due to the inflated total variance (cf. Greenacre, 2017: 145). The eigenvalues show that the first two dimensions account for 39.5% of total variance (see Appendix A), suggesting the need to consider additional dimensions. As a solution, 'adjusted MCA' that offers a more realistic estimate of explained variance (cf. Greenacre, 2007: 149) was applied , which indicated the first two dimensions explain 83.4% of variance (see Appendix A). Following Levshina's (2015: 382-

383) recommendations, a correlation test on the plotting coordinates from both methods confirmed a perfect correlation, justifying the focus on the first two dimensions. Details on each variable's contribution to the dimensions and their directionality are provided in Appendix B.

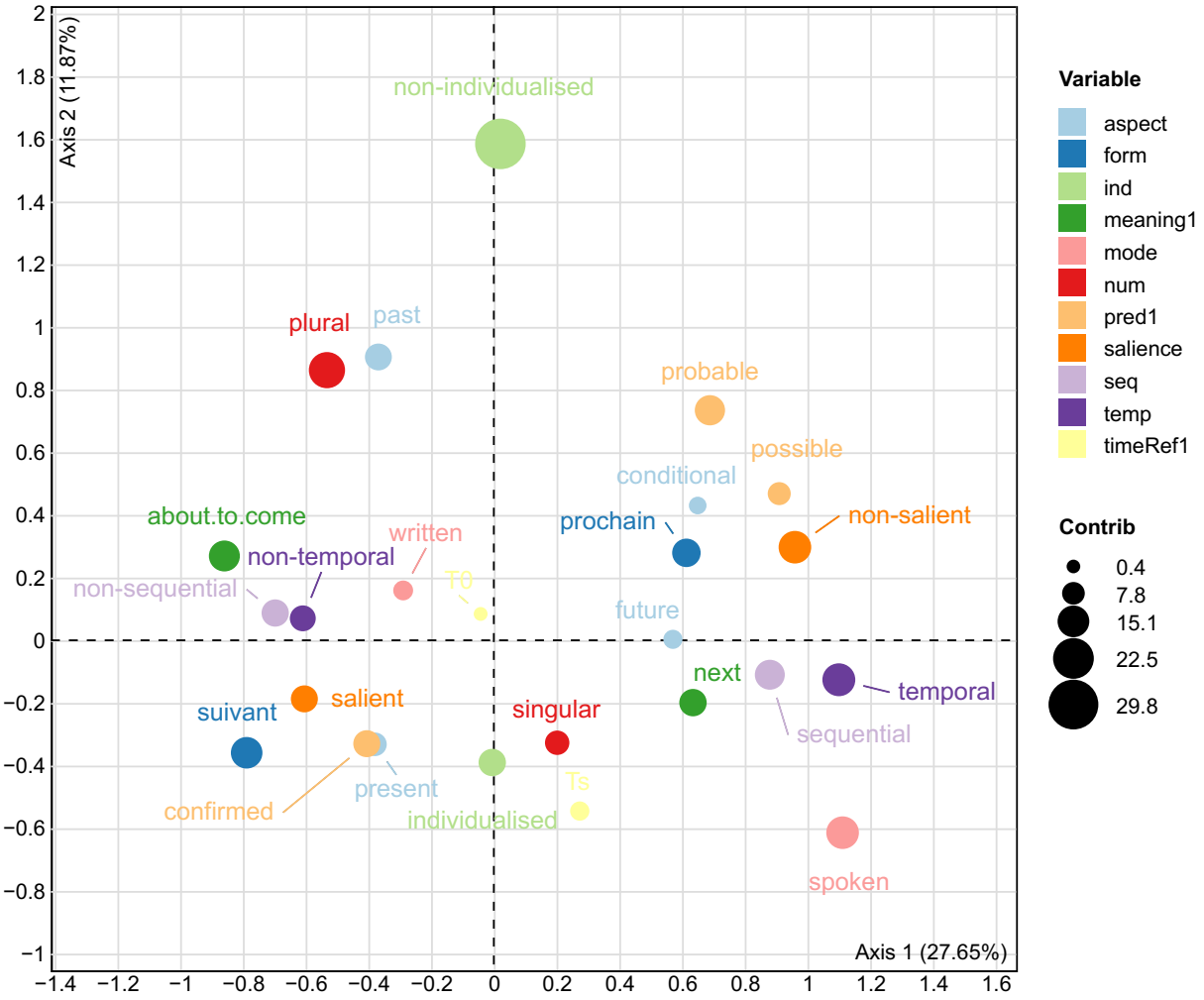

Figure 3: Multiple Correspondence Analysis biplot showing credibly non-zero variables from the regression model

4. Results and discussion: logistic regression and multiple correspondence analysis

In Section 3, a regression model (Table 3) examined the relationship between fourteen variables and the choice between *prochain* and *suivant*, addressing RQ1. The analysis indicated that three variables – determiner type, closedness of topological inclusion, and repetitiveness of the modified entity – showed no statistical association with the choice and were excluded during cross-validation. A review of their distribution highlighted divergences from Berthonneau's (2002) descriptions.

4.1 First divergences from Berthonneau (2002): Variables lacking statistical associations

For *prochain* and *suivant*, the use of the definite article dominates, with 104 occurrences for *prochain* and 97 for *suivant*, representing 79% and 95% of total occurrences respectively. This supports Berthonneau's (2002: 107-108) description that the definite article is essential for *prochain*'s deictic use. Instances without an article, where *prochain* modifies a day of the week (30) or a title (31) and *suivant* modifies a post-nominal modifier (32), accounting for 16 occurrences (12%) and 5 occurrences (5%) respectively.

(30) dès mardi prochain
("from next Tuesday")

(31) Prochaine randonnée: dimanche 12 mai est proposée une randonnée-détente de jour.
("Next hike: Sunday May 12 a relaxing-day hike is offered.")

(32) la densité de probabilité suivante
("the following probability density")

Unlike *suivant*, which only appears with a definite article, *prochain* co-occurs with indefinite articles article (6 occ. or 5%) or a possessive adjective (5 occ. or 4%), as exemplified in (17) and (33) respectively. The use of an indefinite article with *prochain* supports Berthonneau's findings on its impact on deixis; however, the infrequency of this usage and the limited variability in determiner types prevent it from reaching statistical significance.

(33) L'Orchestre de chambre de Toulouse lui a confié la direction de son prochain concert.
("The Toulouse Chamber Orchestra has entrusted him with the management of its next concert.")

The second excluded variable, the closedness of topological structure, shows a preference for *prochain* with open structures (65 occ. or 81%) compared to *suivant* (15 occ. or 19%), but less so with closed structures where *suivant* appears more frequently (87 occ. or 57%). The distribution of *prochain* is even between open and closed structures (65 vs 66 occ.), suggesting the variable's importance in determining *prochain*'s usage rather than the form choice.

Lastly, the repetitiveness of the modified entity lacks statistical association with the choice between *prochain* and *suivant*, despite initial trends showing a preference for *prochain* when entities are repetitive (111 occ. or 76%) like in (34) and for *suivant* in non-repetitive contexts (66 occ. or 77%) like in (35). However, repetitiveness does not significantly influence the form choice, suggesting that other variables might better explain the alternation.

(34) et euh m tu m'avais fait une petite blague noire en me disant que le prochain endroit où tu tu irais ce serait euh au cimetière.
("and uh hm you told me a little dark joke by telling me that the next place you would go would be uh to the cemetery.")

(35) Les quatre problèmes suivants avaient été conçus de manière à être trop difficiles pour les enfants de cet âge.
("The four following puzzles were designed to be too difficult for children of this age")

4.2 Logistic regression results

After excluding three variables, the logistic regression identified ten key factors influencing the choice between *prochain* and *suivant*. The intercept (-3.00) shown in Table 3 represents the log odds of choosing *suivant* when all predictors are at their reference levels (Levshina, 2015: 259). Converting this log odds to simple odds (0.05) suggests that *suivant* is chosen once for every 20 times *prochain* is selected. Using Equation (1), the probability of selecting *suivant* is calculated at 4.74%, indicating a strong preference for *prochain* with all variables held constant at their reference levels.

$$P = \frac{Odds}{1+Odds} \quad (1)$$

4.2.1 *Categories associated with* suivant

Among the ten factors, time reference shows the strongest association with the choice between *prochain* and *suivant*. Specifically, *prochain* is favoured when referring to T0 (124 occ. or 61%), while *suivant* is preferred when referring to Ts (24 occ. or 77%). Holding all other variables constant, specifying Ts increases the odds of choosing *suivant* over *prochain* by a factor or

1,339. This significant change is calculated by exponentiating the median estimate of 7.20 (SD=1.20). To determine the probability of choosing *suivant*, we calculate the log odds, combining the intercept with the coefficient of time reference multiplied by 1 while that of all other variables is multiplied by 0, based on Equation (2), and use the logistic function (Equation 3) to convert the odds to probability. This results in a high likelihood of selecting *suivant* (98.5%).

$$g(x) = b0 + b1x1 + b2x2 + \ldots \quad (2)$$

$$P = \frac{e^{g(x)}}{1+e^{g(x)}} \quad (3)$$

Next, meaning is inevitably important in the alternation between *prochain* and *suivant* (see Section 2.3 for the meanings operationalised). *Prochain* is commonly associated with *next* (101 occ. or 77%) whereas *suivant* is associated with *about to come* (69 occ. or 68%). The regression model indicates that using the meaning *about to come* (Mdn.=3.25, SD=0.71) increases the probability of choosing *suivant* from 4.74% to 56.21%, holding all other predictor variables constant. However, this probability's proximity to 50% suggests that the choice is not solely dependent on their intended meanings.

Grammatical number is also associated with the form selection, where *prochain* is typically used with singular nouns (109 occ. or 64%) and *suivant* is favoured for plural referents (42 occ. or 66%). With all other variables held constant, the model estimates a 49% probability of choosing *suivant* for plural nouns (Mdn.=2.96, SD=0.89), a probability close to the 50% threshold hinting at a more complex decision-making process.

The discussion hereafter focuses on variables that are positively associated with *suivant* but do not counter the general preference for *prochain* due to their relatively smaller estimates. Noun sequentiality (Mdn.=2.67, SD=0.99) shows a positive correlation with *suivant* but does not outweigh the intercept's negative estimate (Mdn.=-3.00, SD=1.33), resulting in a 41.82% probability of choosing *suivant* with sequential nouns. In practice, *prochain* is chosen more frequently (75 occ. or 73%) for sequential nouns while *suivant* (74 occ. or 57%) is preferred by non-sequential nouns.

Salience (Mdn.=2.36, SD=0.80), while a key concept in Berthonneau (2002), exhibits a modest association with *suivant*, with *suivant* preferred (94 occ. or 66%) over *prochain* (49 occ. or 34%) when modifying a salient referent (i.e. a referent with an implicit and clearly

understood antecedent). However, when the referent lacks salience, *prochain* dominates (82 occ. or 91%). With other variables held constant, the probability of choosing *suivant* to modify a salient referent increases to 34.52% from the baseline of 4.74%, indicating that salience alone does not drive the choice.

Finally, noun individuality (Mdn.=1.70, SD=0.88) suggests *suivant* is more likely selected with individualised nouns, albeit its weak association. Holding other variables constant, the probability of choosing *suivant* stands at only 21.42% in such contexts. Data shows no strong trend for individualised nouns with which *prochain* co-occurs in 98 occurrences (52%) and *suivant* in 89 (48%), whereas non-individualised nouns clearly favour *prochain* (33 occ. or 72%) over *suivant* (13 occ. or 28%).

4.2.2 *Categories associated with* prochain

The variable with the strongest association with *prochain* is noun temporality. It shows a strong negative association with *suivant* (Mdn.=-4.96, SD=1.31), indicating that in contexts involving a temporal referent, the probability of opting for *suivant* is reduced to a mere 0.03% with other variables held constant. Empirical data supports this, with *prochain* being the dominant choice for temporal nouns (69 occ. or 83%). In contrast, for non-temporal nouns, *suivant* enjoys a slight preference (88 occ. or 59%).

Next, temporal aspect correlates with the alternation between *prochain* and *suivant*. Temporal aspect operationalised into *present*, *past*, *future*, and *conditional*, *prochain* is more commonly selected in *future* (57 occ. or 80%) and *conditional* (17 occ. or 85%) contexts while *suivant* is preferred in *present* scenarios (70 occ. or 63%). The distribution is nearly even in *past* contexts with *prochain* used in 16 occurrences (52%) and *suivant* in 15 (48%). Regression analysis shows that relative to the reference level *present*, the probabilities of choosing *suivant* over *prochain* in *conditional* (Mdn.=-3.56, SD=1.18), *past* (Mdn.=-2.60, SD=1.04), and *future* (Mdn.=-2.10, SD=0.67) contexts are notably low, at 0.1%, 0.3%, and 0.6% respectively.

The level of event predictability also shows a correlation with the form choice. Within the regression model, predictability is operationalised into three categories: *confirmed, probable*, and *possible*. As predictability decreases, *prochain* is increasingly chosen, evidenced by the estimates of the categories *possible* (Mdn.=-3.06, SD=1.23) and *probable* (Mdn.=-3.03, SD=0.89). The probability of choosing *suivant* in these less predictable contexts is identically low at 2.4%, seconded by the observed data where *prochain* is the preferred choice in 29 occurrences (94%) of *possible* and in 43 occurrences of *probable* scenarios.

Finally, the association between diamesic variation and the form selection is marginal. The regression model indicates a slight preference for *prochain* in written contexts (Mdn.=-1.96, SD=0.97), reducing the probability of choosing *suivant* in written French to just 0.7%, with all other variables held constant. Empirical observation shows a balanced preference for both adjectives in written French where *prochain* occurs in 92 instances (50%) and *suivant* in 93 (50%), but a clear preference for *prochain* (39 occ. or 81%) in spoken French.

In the current section, the discussion underscores that the choice between *prochain* and *suivant* is determined by a range of factors, suggesting the need to consider multiple categories concurrently. For example, the analysis of (36) must incorporate variables such as written mode, plural noun, sequential noun, salient referent, and future aspect. Using the R function posterior_epred() from the rstanarm package, the model predicts *suivant* being chosen at the probability of 68.54%. Since this exceeds the 50% threshold, the model accurately predicts the choice of *suivant*.

(36) Nous mettrons en évidence les différents outils utilisés pour décrire et analyser ces dimensions dans les chapitres suivants.
("We will highlight the different tools used to describe and analyse these dimensions in the following chapters.")

The multifactorial examination of the choice between *prochain* and *suivant* underscores the utility of inferential methods. Attempting to answer RQ1, the regression analysis reveals that not all introspectively identified variables are statistically associated with the choice, specifically determiner type, closedness, and repetitiveness. More importantly, the analysis highlights the probabilistic nature of the choice, contrasting with the binary distinctions often implied in Berthonneau (2002).

4.3 Multiple correspondence analysis results

While using the regression model to calculate probabilities for specific instances like (36) provides valuable insights, it may not fully capture the complex relationships among variables influencing the choice between *prochain* and *suivant*, addressed by RQ2. This section explores insights from an MCA analysis of how these variables are collectively associated with the alternation.

Figure 3 visualises the relationships between the predictors retained in the regression model and the choice between *prochain* and *suivant*. Additionally, two plots integrating all 233 data points were generated to examine the prototypicality of each form. While Figure 4 shows considerable overlap in the usage of *prochain* and *suivant*, Figure 5 introduces centroids that reveal a clear prototypical divergence between the two forms. This divergence sets the stage for interpretating Figure 3, identifying specific variables that drive the differences between *prochain* and *suivant*.

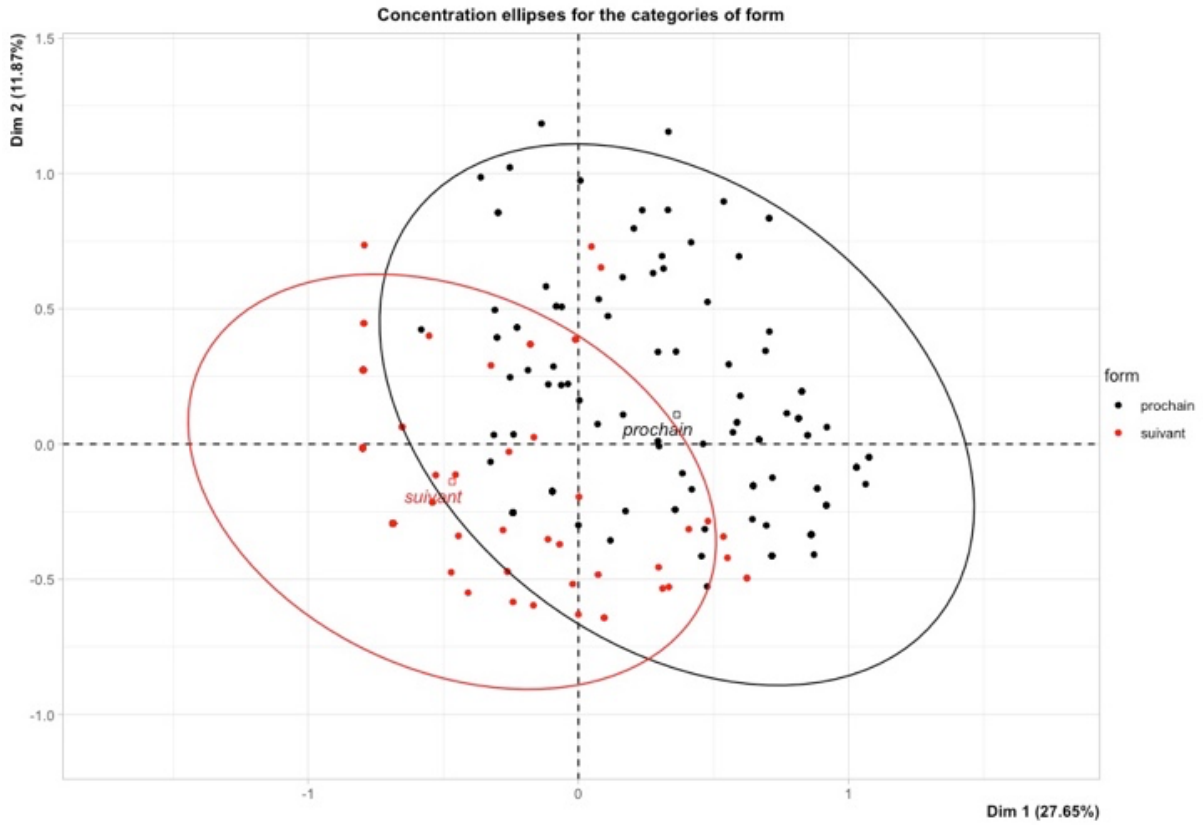

Figure 4: Confidence ellipses arounds the categories of form without centroids

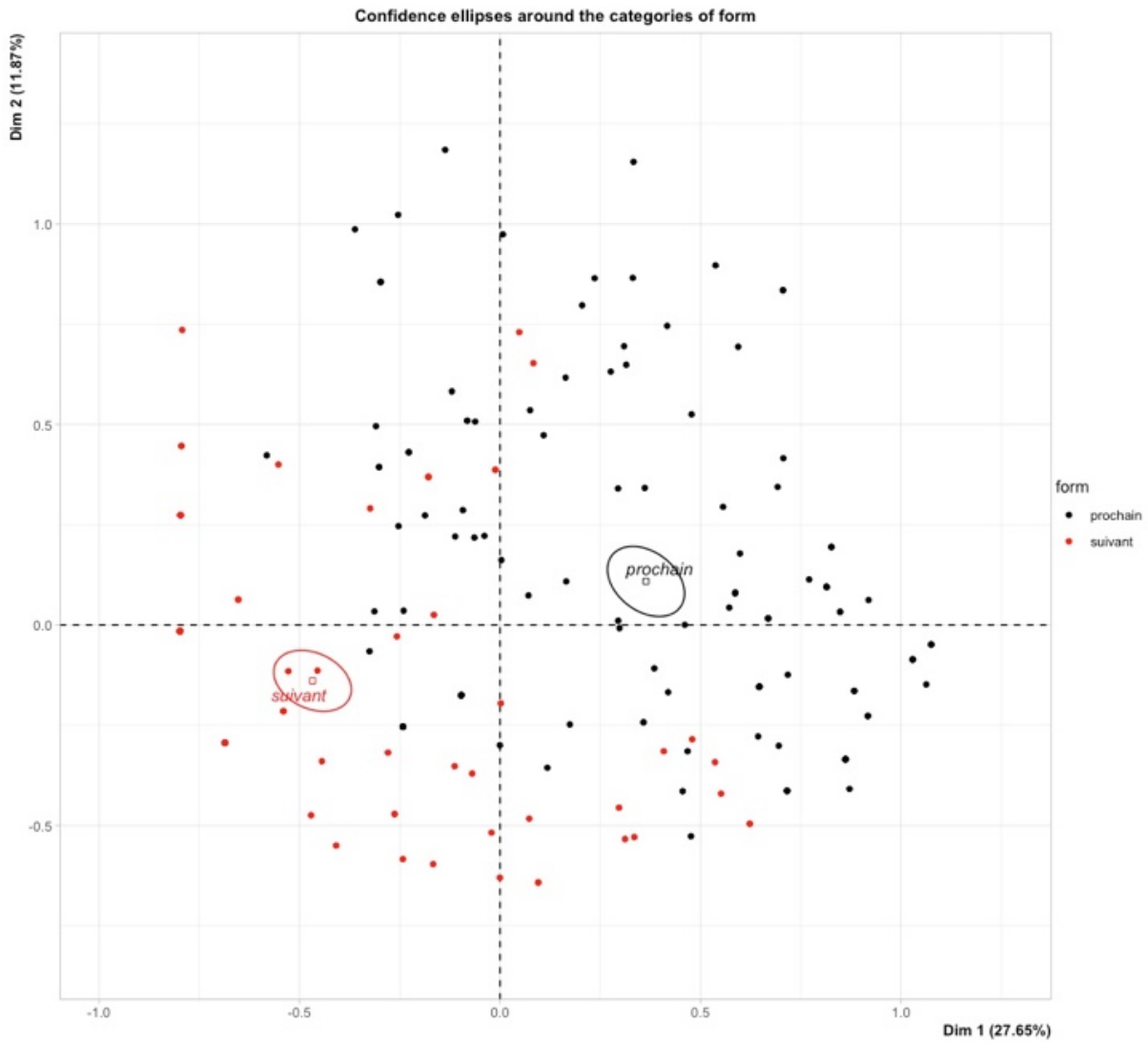

Figure 5: Confidence ellipses arounds the categories of form with centroids added

In Section 3.2, the two dimensions together accounted for 83.4% of the total variance, with the first dimension capturing 79.2% and the second 4.2%. The disparity suggests that interpretation should primarily focus on the horizontal axis, which emphasises the contrast between the deictic *prochain* and anaphoric *suivant*. The second dimension, meanwhile, provides insights into the predictability of an event's occurrence.

4.3.1 *First dimension: distinction between deictic* prochain *and anaphoric* suivant

In Figure 3, the first (horizontal) axis closely aligns *prochain* with the temporal categories of *conditional* and *future*, reflecting its deictic usage as typically pointing towards *future* or post-

T0 references, corroborating Berthonneau (2002: 105-106). In contrast, *present* is plotted near *suivant*, indicating its frequent anaphoric uses in present contexts. However, the distant placement of *past* and *future* categories from *suivant* suggests a weaker association with these times, diverging from Berthonneau's description where *suivant* can refer to both (Section 2).

The MCA biplot also reveals a relationship between event predictability and temporal aspects. Lower predictability categories like *probable* and *possible* are near the *conditional* aspect, also closely associated with *prochain*, aligning with the conditional tense's role in signaling expected but uncertain events, as in (37). Conversely, the *future* tense, suggestive of higher predictability, is positioned further from these categories, indicating a stronger link with more predictable events, as in (38). The highest predictability level, *confirmed*, is plotted alongside *present*, near *suivant*, reflecting scenarios often described with certainty or as factual, like in (39), and indicating confirmed upcoming sections in texts as in (40).

(37) Real Sociedad… dont il pourrait devenir le technicien la saison prochaine.
("Real Sociedad… of which he could become the technician next season.")

(38) Christian Califano rejoindra la saison prochaine ses compatriotes… aux Saracens, club anglais basé à Watford.
("Christian Califano will join his compatriots next season… at Saracens, an English club based in Watford.")

(39) Chaque jour, ces populations rencontrent d'abord une phase de latence… jusqu'au jour suivant, où les cellules recommencent à croître dans le milieu frais.
("Each day, these populations first encounter a lag phase… until the following day, when the cells begin growing again in the fresh medium.")

(40) Dans ce cadre, nous proposons, dans les paragraphes suivants, plusieurs interprétations pouvant expliquer les résultats obtenus.
("In this context, we propose, in the following paragraphs, several interpretations that can explain the results obtained.")

The choice between *prochain* and *suivant* is a phenomenon multifactorially determined by event predictability, temporal aspects, and the natures of the modified noun. Figure 3 shows *prochain* surrounded by the categories *temporal, sequential,* and *non-salient*, indicating its

common usage with time-related, sequentially structured, but non-salient nouns. Examples include *semaine, mois, année,* and *lundi*. Conversely, *suivant* is often associated with nouns that are *non-temporal, non-sequential*, but *salient*, such as *sujet, formule, activités,* or *étudiants*, as reflected by these categories' positioning in the biplot.

Noun temporality and sequentiality, plotted on separate quadrants from *prochain* and *suivant*, show that *temporal* and *sequential* nouns typically align with *prochain* on the first dimension but with *suivant* on the second. *Non-temporal* and *non-sequential* nouns generally pair with *suivant* on the first dimension and *prochain* on the second. Given that the first dimension captures significantly more variance, *temporal* and *sequential* nouns are predominantly linked with *prochain*, while *non-temporal* and *non-sequential* nouns tend to be associated with *suivant*.

The meaning *next* and *about to come* also display distinct preferences: *next* is frequently linked with *prochain* along the first axis but with *suivant* on the second, and vice versa for *about to come*. These meanings are situated near categories representing noun natures, indicating *next* commonly occurs with *temporal* and *sequential* nouns, while *about to come* aligns with *non-temporal* and *non-sequential* nouns.

The current analysis bringing to light these associations with the respective choice of *prochain* and *suivant* corroborates the introspective descriptions of Berthonneau (2002) from a general perspective. On the first, or horizontal, dimension emerges a clear divide between deictic *prochain* and anaphoric *suivant*, manifested through differences in temporal aspect, event predictability, noun temporality, sequentiality, salience, and adjective meaning. This divide is exemplified by the prototypical examples (41) and (42). In (41), the deictic *prochain* meaning *next* is oriented towards the future, accompanied by the temporal, sequential but non-salient noun *semaine* and the use of conditional tense indicating a lower predictability of the speaker arriving on time. In contrast, the anaphoric *suivant* in (42) signifies *about to come* pivoting around the present, as evidenced by the use of simple present and its confirmed occurrence, and co-occurs with the non-temporal, non-sequential but salient noun *raison*.

(41) Je m'excuse pour le retard, la semaine prochaine je serais là plus tôt.
("I apologise for the delay, next week I will be there earlier.")

(42) Eh bien c'est pour la raison suivante euh une élection aura lieu et nous savons ce que nous ne voulons pas comme homme politique.

("Well it's for the following season uh an election will take place and we know what we don't want in a politician.")

However, MCA manifests more nuanced interplay among these variables and between the choice of *prochain* and *suivant*. While this hardly amounts to criticism on Berthonneau's work, which is highly complex and detailed, the lacking quantification in introspective work inevitably treats each variable on largely equal footing, which is clearly not the case as shown by MCA. For instance, grammatical number is described as directly associated with the deixis of *prochain*, the deictic usage of which is confined to singular noun (Section 2.1). However, MCA reveals that the category *singular*, plotted close to the origin, lacks distinguishability on the deictic-anaphoric dimension, in comparison with categories like *temporal* and *non-salient*. Furthermore, variables said to be correlate with the choice can lack statistical association, such as *repetitiveness.*

4.3.2 *Second dimension: distinction between occurrence predictability and eventuality*

The first dimension in Figure 3 emphasises the deictic-anaphoric distinction, while the second dimension seems to focus on event occurrence, with the category *non-individualised* being the most distinctive at the top of the vertical axis. However, individuality categories plotted on the y-axis do not aid in differentiating between deictic *prochain* and anaphoric *suivant*, contrary to Berthonneau's claim (2002: 119-120) that *prochain* requires temporal nouns to have an individualised status for its deictic function.

Figure 3 also challenges Berthonneau's description regarding *T0* and *Ts*. Both categories being positioned near the origin and excluded from the first dimension (see Appendix B) suggests their minimal role in distinguishing between *prochain* and *suivant*, challenging the notion of a simple binary distinction based on time reference, as suggested in Berthonneau's descriptions.

The highest contributors on the second dimension – *non-individualised*, *plural*, and *probable* – all tend to associate with *prochain*. For instance, in (43), the plural noun *mois* implies a non-specific future event as it does not specify an exact month for an expected rise in unemployment rates. This non-individualised and vague context aligns with the probability of the event, indicating an eventual rather than a specific occurrence.

(43) Le chômage, en hausse depuis près de 14 mois, devrait donc augmenter dans les prochains mois.
("Unemployment, which has been increasing for almost 14 months, is therefore expected to increase in the coming months.")

Below the horizontal axis, categories such as *individualised*, *singular*, *present*, and *confirmed* are associated with *suivant*. For instance, in (44), the singular noun *étape* signifies a specific phase in an experiment, denoting individualised status and definite occurrence, as emphasised by the present tense. This represents the prototypical use of *suivant* in contexts where events are specific and certain, contrasting with the more general and eventual contexts associated with *prochain*.

(44) Après élimination du butanol, l'étape suivante consiste à placer la strip en contact étroit avec la surface du gel de concentration.
("After removing the butanol, the following step is to place the strip in close contact with the surface of the concentration gel.")

In this dimension focused on event occurrence, *time reference* and *production mode* show limited associations with the choice between *prochain* and *suivant*. Positioned differently than the adjectives, categories like *T0, Ts, written* and *spoken* suggest preferences in form selection. With the first dimension capturing 79.2% of the variance, spoken language, distinctly on the right, aligns more with *prochain*. Written language, near the origin on the left, shows less distinction but a slight preference for *suivant*.

Time reference, especially *T0*, is the nearest to the origin, indicating a minimal role in differentiating the two forms. *Ts*, while closer to categories denoting certainty, has a weak link with *suivant*, as shown in Appendix B. These findings suggest that while time reference and production mode are associated with the alternation, their association is less pronounced compared to other variables in the analysis.

5. Conclusion

This article commenced with a brief discussion on the evidential status of observed and introspective data within linguistics as a whole, and in the French context in particular. The study itself tested Berthonneau's (2002) introspective account of the choice between *prochain*

and *suivant*, here using corpus-based methods and empirical data. We operationalised fourteen variables identified by Berthonneau and employed Bayesian binomial logistic regression and Multiple Correspondence Analysis (MCA) on data extracted from the CEFC corpus to quantify their associations with the choice between *prochain* and *suivant* and to shed light on the interplay among these variables.

Three research questions were formulated to steer our analysis. RQ1 examines whether Berthonneau's (2002) introspectively identified variables are statistically associated with the choice between *prochain* and *suivant*. Our findings indicate that most of these variables do indeed have a statistical association with the choice, with notable exceptions being the type of co-occurring determiner, the closedness of the modified referent, and the entity's repetitiveness. Additionally, our analysis suggests some discrepancy from Berthonneau's interpretations, particularly in the varying levels of importance given to each variable associated with the form selection. In particular, while Berthonneau considers that salience (or the lack of it) is "key" to explaining the differences (2002: 122), our regression analysis revealed salience as one of the variables exhibiting the weakest associations with the choice of form. Furthermore, MCA also showed that grammatical number is not directly associated with the choice and appears to lack distinguishability from other variables.

RQ2 focused on the specific relationships among the fourteen variables and the usage of *prochain* and *suivant*. Our MCA analysis identified two primary dimensions of usage: the deictic-anaphoric distinction as the main dimension and the expression of event occurrence certainty as the secondary dimension. The first dimension aligns with Berthonneau's (2002: 122) conclusion that there exists "a necessary referential disjunction between *prochain* and *suivant*, and more generally between 'deictic' adjectives and 'anaphoric' adjectives". Our findings refine this account by further showing that variables such as temporal aspect, event predictability, the meanings of the adjectives, and characteristics of the modified noun like temporality, sequentiality, and salience, are crucial in differentiating the deictic use of *prochain* from the anaphoric use of *suivant*. Additionally, the second dimension of our MCA analysis underscores the importance of referent individuality, time reference, grammatical number, and event predictability levels in interpreting how *prochain* and *suivant* convey the certainty of an event's occurrence, which Berthonneau addressed only fleetingly.

Finally, RQ3 examined to what extent the arguments made through introspection align with the findings from statistical and corpus analyses. Our study reveals a substantial agreement between Berthonneau's (2002) introspective insights and our inferential methods and corpus data. Yet at the same time, introspection falls short in quantifying the variable

relationships and prioritising their relative significance, and it might not fully capture the linguistic creativity of users. As Fillmore (1992: 43) explains, corpus data oblige researchers to total accountability. Aarts (1991: 46) also points out that "only linguists who use corpus data themselves will know that a corpus always yields a greater variety of constructions than one can either find in the literature or think up oneself." While Berthonneau (2002: 104) emphasises the division between temporal and material nouns which are "susceptible to be considered as places", our corpus data suggest a rare use of *prochain* with location; such instances do not emerge as statistically collocates of *prochain* (see Appendix C & D). In contrast, *suivant* is predominantly used with abstract nouns like *page, manière,* and *paragraphe* ("page", "manner", and "paragraph") (Appendix E). The use of event-related nouns with *prochain*, such as *élections*, *réunion*, and *mariage* ("elections", "meeting", and "marriage "), prevalent in our data, were overlooked in her analysis but considered in the current study (Appendix C & D).

Taking the discussion further, the classification of 'non-prototypical' adjectives into the third type is not without debate. The ambiguity in its definition and incomplete classification have led Goes (2021) to criticise the tripartite division of adjectives – qualitative, relation, and of the third type – as fragmentary. He introduces the unitary hypothesis, which posits that French adjectives form a single category whose usage varies depending on the accompanying noun. This hypothesis highlights the importance of analysing French adjectives at the phrase level, considering their interaction with nouns.

Our study leans toward this unitary view, given the observed dependencies between the choice of *prochain* and *suivant* and the nature of the nouns in this study. This hypothesis echoes Sinclair's lexical grammar (2000) and co-selection (1998), suggesting that the meanings of adjectives emerge from their syntagmatic relationships with nouns. Some scholars argue that individual words carry only meaning potentials, and that their actual meanings only emerge when they are used and combined in clauses and texts (Hanks, 2013: 65). Moreover, there are parallels with English and other languages which may provide valuable insights for studies on French adjectives (e.g. Scontras, 2023; Wulff, 2003). While this line of inquiry is promising, it merits a detailed exploration in a separate paper.

Our goal in re-visiting Berthonneau's work is not to condemn it out of hand. Rather, her analyses have catalysed this study, which demonstrates how corpus data and statistical methods can interact with linguistic hypotheses, even those formulated through pure introspection. As McEnery and Hardie (2012: 158) state, testing non-corpus-informed theory against corpus data is "a critical test of that theory, not an uncritical reconfirmation of it". Furthermore, we recognise that corpus data, like introspective data, do not encapsulate the entirety of language,

but rather represent a mere sample of it. The presence or absence of occurrences may be influenced by sampling choices or mere chance. We make no claims of discovering the absolute truth of language. Instead, we propose that a closer approximation of this truth can emerge from the interplay of various methodologies, including both corpus-based and introspective approaches. This paper contributes to such endeavours, which are particularly relevant in the French linguistic community. Despite a growing shift towards corpus-based methods over the past decade, introspective techniques remain common and are often the primary approach, with corpus data frequently supplying illustrative examples. Methodologies used in this study can apply broadly to other linguistic features and languages, emphasising the importance of variable quantification and testing introspective claims against corpus data.

Notes

The authors used generative AI ChatGPT version 4.0 during the preparation of this work to check for language errors and stylistic unnaturalness, and to streamline the text. After utilising the tool, the authors thoroughly reviewed and edited the content as necessary and assumed full responsibility for the content.

Bibliographical references:

Aarts, J. (1991). Intuition-based and observation-based grammars. In K. Aijmer and B. Altenberg (Eds.), *English Corpus Linguistics*, (pp. 44-63). Longman.

Benzitoun, C., Bresson, S., Budzinski, L., Debaisieux, J.M. & Holzheimer, K. (2010). Quand un corpus rencontre un adjectif du troisième type. Étude distributionnelle de *prochain*. *Corpus*, 9, 245-264. https://doi.org/10.4000/corpus.1927

Benzitoun C., Debaisieux, J.M. & Deulofeu, J. (2016). Le projet ORFÉO : un corpus d'étude pour le français contemporain, *Actes du colloque Corpus de Français Parlés et Français Parlés des Corpus (Corpus 15)*, 1-18.

Berthonneau, A.-M. (2002). *Prochain/dernier* et compagnie. Les adjectifs « déictiques » à l'épreuve de l'espace ou comment circuler dans le temps, l'espace, le texte. *Langue Française*, 136, 104-125.

Chomsky, N. (1979). *Language and Responsibility: Based on Conversations with Mitsou Ronat.* Pantheon.

Fornacon-Wood, I., Mistry, H., Johnson-Hart, C., Faivre-Finn, C., O'Connor, J.P.B., Price, G.J. (2021). Understanding the differences between Bayesian and frequentist statistics. *International Journal of Radiation Oncology, Biology, Physics*, 112(5), 1076-1082.

Feuillet, J. (1991). Adjectifs et adverbes, essai de classification. In C. Guimier & P. Larcher (Eds.), *Travaux Linguistique du Cerlico 3 : Les États de L'adverbe,* (pp. 35-58). Presses Universitaires de Rennes.

Fillmore, C.J. (1992). "Corpus linguistics" vs. "computer-aided armchair linguistics". In J. Svartvik (Ed.), *Directions in Corpus Linguistics: Proceedings from a 1991 Nobel Symposium on Corpus Linguistics*, (pp. 35-66), Mouton de Gruyter.

Forsgren, M. (1978). *La place de l'adjectif épithète en français contemporain : Étude quantitative et sémantique* [The Placement of Attributive Adjective in Contemporary French: A Quantitative and Semantic Study]. Acta Universitatis Upsaliensis.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M. & Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society*, 182(2), 389-402.

Gelman, A., Hill, J. & Vehtari, A. (2020). *Regression and Other Stories*. Cambridge University Press.

Gibbs, R. W. (2006). Introspection and cognitive linguistics. Should we trust our own intuitions? *Annual Review of Cognitive Linguistics*, 4, 135-151.

Goes, J. (2021). L'adjectif: une partie du discours éminemment syncatégorématique. *Kalbotyra*, 74, 72-87.

Goes, J. (2022). Grand et petit, de simples antonymes ? *Éla. Études de Linguistique Appliquée*, 206, 139-153.

Greenacre, M. (2017). *Correspondence Analysis in Practice* (3rd ed.). Chapman and Hall.

Gries, S.T. & Deshors, S.C. (2014). Using regressions to explore deviations between corpus data and a standard/target: two suggestions. *Corpora*, 9(1), 109-136.

Gries, S.T. & Divjak, D. (2009). Behavioral profiles: a corpus-based approach to cognitive semantic analysis. In V. Evans & S. Pourcel (Eds.), *New Directions in Cognitive Linguistics*, (pp. 57-75). Johns Benjamins.

Hanks, P. (2013). *Lexical analysis: norms and exploitations*. The MIT Press.

Heiden, S. (2010). The TXM platform: Building open-source textual analysis software compatible with the TEI encoding scheme. In R. Otoguro., K. Ishikawa, H. Umemoto, K. Yoshimoto & Y. Harada (Eds.), *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, (pp. 389-398). Institute of Digital Enhancement of Cognitive Processing, Waseda University.

Levshina, N. (2015). *How to Do Linguistics with R: Data Exploration and Statistical Analysis*. John Benjamins Publishing Company.

Liu, M. & Dou. J. (2024). Metaphorical polysemy of the Chinese color term hēi 黑 "black": A corpus-based cognitive semantic analysis with behavioral profiles. *International Journal of Corpus Linguistics*, 29(1), 1-33.

McEnery, T. & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.

Noailly, M. (1999). *L'adjectif en français* [Adjective in French]. Ophrys.

Riegel, M., Pellat, J.-C., Rioul, R. (2004). *Grammaire Méthodique du Français* [Methodical Grammar of French] (3rd ed.). PUF.

Riegel, M., Pellat, J-C., Rioul, R. (2009). *Grammaire Méthodique du Français* [Methodical Grammar of French] (4th ed.). PUF.

Schindler, S., Drożdżowicz, A. & Brøcker, K. (2020). *Linguistics Intuitions: Evidence and Method*. Oxford University Press.

Schnedecker, C. (Ed.). (2002). L'adjectifs sans qualité(s). *Langue Française*, 136.

Schnedecker, C. (2002). Présentation : les adjectifs « inclassables », des adjectifs du troisième type ? *Langue française*, 136, 3-19.

Scontras, G. (2023). Adjective ordering across languages. *Annual Review of Linguistics*, 9, 357-376.

Sinclair, J. M. H. (1998). The lexical item. In E. Weigand (Ed.), *Contrastive lexical semantics*, (pp. 1-24). John Benjamins Publishing Company.

Sinclair, J. M. H. (2000). Lexical grammar. *Naujoji Metodologija*, 24, 191-204.

Teubert, W. (2010). Our brave new world? *International Journal of Corpus Linguistics*, 15(3), 354.358.

Wasserstein, R. L., Schirm, A. L. & Lazar, N. A. (2019). Moving to a world beyond "p < 0.05". *The American Statistician*, 73(sup1), 1-19.

Worlock Pope, C. (Ed.). (2010). The bootcamp discourse and beyond. *International Journal of Corpus Linguistics,* 15(3).

Wulff, S. (2003). A multifactorial corpus analysis of adjective order in English. *International Journal of Corpus Linguistics*, 8(2), 245-282.

Xu, M., Li, F. & Szmrecsanyi, B. (2024). Modeling the locative alternation in Mandarin Chinese: A corpus-based study. *International Journal of Corpus Linguistics*.

Yamamoto, D. (2020). L'adjectivité des épithètes antéposées sale et foutu. *Travaux de Lingustique*, 80, 49-61.

R packages:

Barnier, J. (2023). explor: Interactive interfaces for results exploration. R package version 0.3.10. URL https://CRAN.R-project.org/package=explor

Dogucu, M., Johnson, A. & Ott, M. (2021). bayesrules: Datasets and supplemental functions from Bayes Rules! Book. R package 0.0.2.9000. URL https://github.com/bayes-rules/bayesrules

Gabry, J. & Mahr, T. (2022). bayesplot: Plotting for Bayesian models. R package version 1.10.0. URL https://mc-stan.org/bayesplot/

Goodrich, B., Gabry, J., Ali, I. & Brilleman, S. (2022). rstanarm: Bayesian applied regression modelling via Stan. R package version 2.21.3. URL https://mc-stan.org/rstanarm

Hartig, F. (2022). DHARMa: Residual diagnostics for hierarchical (multi-level/mixed) regression models. R package version 0.4.6. URL https://CRAN.R-project.org/package=DHARMa

Le, S., Josse, J. & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*, 25(1), 1-18.

Lüdecke, D., Ben-Shachar, M., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R Package for Assessment, Comparison and Testing of Statistical Models. *Journal of Open Source Software*, 6(60), 3139. https://doi:10.21105/joss.03139

Nenadic, O. & Greenacre, M. (2007). Correspondence Analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software*, 20(3), 1-13.

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P., Paananen, T. & Gelman, A. (2022). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 2.5.1. URL https://mc-stan.org/loo

Wickham, H. (2011). The Split-Apply-Combine strategy for data analysis. *Journal of Statistical Software*, 40(1), 1-29. URL https://www.jstatsoft.org/v40/i01/.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York.

Appendix A. Variance explained in MCA and adjusted MCA

MCA				Adjusted MCA			
Dimension	Eigenvalue	% of variance	Cum. % of variance	Dimension	Eigenvalue	% of variance	Cum. % of variance
dim 1	0.35	27.7	27.7	dim 1	0.082	79.2	79.2
dim 2	0.15	11.9	39.5	dim 2	0.004	4.2	83.4
dim 3	0.13	10.2	49.7	dim 3	0.002	1.7	85.1
dim 4	0.10	8.0	57.7	dim 4	0.0001	0.1	85.2
dim 5	0.10	7.8	65.5	dim 5	0.00009	0.1	85.3
dim 6	0.09	6.8	72.3				
dim 7	0.09	6.7	79.0				
dim 8	0.06	4.9	83.9				
dim 9	0.05	4.0	87.9				
dim 10	0.05	3.6	91.5				
dim 11	0.04	3.2	94.7				
dim 12	0.03	2.4	97.1				
dim 13	0.02	1.8	9.0				
dim 14	0.01	1.0	100				

Note: The cumulative percentages in adjusted MCA do not add up to 100% (see Greenacre, 2017: 249).

Appendix B. Directionality of associations among categories in MCA

First Dimension (x-axis)			**Second Dimension (y-axis)**		
Category	Estimate	*p*-value	Category	Estimate	*p*-value
Temporal	0.51	< 0.01	**Non-individualised**	0.38	< 0.01
Sequential	0.47	< 0.01	**Plural**	0.23	< 0.01
Non-salient	0.46	< 0.01	**Probable**	0.17	< 0.01
Next	0.44	< 0.01	**Past**	0.25	< 0.01
Prochain	0.42	< 0.01	**Prochain**	0.12	< 0.01
Spoken	0.42	< 0.01	**Written**	0.15	< 0.01
Future	0.27	< 0.01	**Non-salient**	0.09	< 0.01
Possible	0.30	< 0.01	**About to come**	0.09	< 0.01
Probable	0.17	< 0.01	**T0**	0.12	< 0.01
Singular	0.22	< 0.01	**Possible**	0.07	< 0.01
Conditional	0.32	< 0.01	**Conditional**	0.07	0.04
Past	-0.29	0.02	**Ts**	-0.12	< 0.01
Plural	-0.22	< 0.01	**Next**	-0.09	< 0.01
Present	-0.30	< 0.01	**Salient**	-0.09	< 0.01
Confirmed	-0.48	< 0.01	**Spoken**	-0.15	< 0.01
Written	-0.42	< 0.01	**Suivant**	-0.12	< 0.01
Suivant	-0.42	< 0.01	**Present**	-0.23	< 0.01
About to come	-0.44	< 0.01	**Confirmed**	-0.24	< 0.01
Salient	-0.46	< 0.01	**Singular**	-0.23	< 0.01
Non-sequential	-0.47	< 0.01	**Individualised**	-0.38	< 0.01
Non-temporal	-0.51	< 0.01			

Appendix C: List of nouns statistically significant ($p < 0.001$) and occurring at least three times to the right of *prochain*

RightN	Trans.	X^2 values
LT	LT	15925.35
fois	time	8226.79
élections	elections	6524.66
échéances	due dates	5088.8
réunion	meeting	5069.47
mariage	marriage	4476.47
années	years	4179.28
séance	session	3741.72
Conseil Européen	European Council	3311.96
quinquennat	five-year mandate	2540.11
étape	step	2415.9
législatives	legislatives	1744.01
mariages	marriages	1709.86
rendez-vous	appointment	1400.64
édition	edition	1239.4
décennie	decade	1172.06
ramassage	collection	1172.06
semaines	weeks	1119.22
collecte	collection	1039.18
Coupe du monde	World Cup	1039.18
mois	month(s)	1019.49
sommet	summit	1006.5
jours	days	874.5
vacances	vacations	763.23
saison	season	692.89
championnats	championship	690.14
congrès	congress	516.73
foire	fair	475.3
week-end	weekend	399.2
permanence	office hours	394.11
assemblée	assembly	326.15
recherches	research	325.33
élection	election	236.9
manifestation	demonstration	223.49
conférence	conference	220.2
gouvernement	government	149.74
conseil	council	121.58
match	match	107.16
voyage	voyage	103.11
départ	departure	93.34
président	president	86.02
objectif	objective	69.87
occasion	occasion	65.63
invité	invitee	61.89
passage	passage	55.58
sortie	exit	53.41
cours	class	14.91

Appendix D: List of nouns statistically significant *($p < 0.001$)* and occurring at least three times to the left of *prochain*

LeftN	Trans.	X^2 values
semaine	week	163133.87
année	year	37275.26
an	year	22830.34
saison	season	16607.29
automne	autumn	2555.87
week-end	weekend	2210.38
mois	month(s)	922.84
printemps	spring	861.3
rentrée	school reopening	494.08
arrivée	arrival	306.72
jeudi	Thursday	238.7
vendredi	Friday	208.94
lundi	Monday	133.43
ouverture	opening	119.56
mardi	Tuesday	90.47
mercredi	Wednesday	89.08
été	summer	82.98
jour	day	40.31
dimanche	Sunday	39.64
mort	death	38.01
nuit	night	36.38

Appendix E: List of nouns statistically significant ($p < 0.001$) and occurring at least three times to the left of *suivant*

LeftN	Trans.	X^2 values
page	page	16668.93
manière	manner	13766.25
équation	equation	11499.86
paragraphe	paragraph	9192.76
chapitres	chapters	5850.78
année	year	5341.93
forme	form	4637.92
relation	relation	4577.77
chapitre	chapter	4287.78
paragraphes	paragraph	4178.32
notations	notations	3521.65
formule	formula	2352.65
étape	step	2220.89
exemples	examples	1835.11
faits	facts	1792
schéma	scheme	1409.2
semaine	week	1409
ordre du jour	agenda	1398.36
hypothèses	hypotheses	1303.08
relations	relations	1293.5
façon	way	1200.44
essai	trial	1161.99
étages	stages	1089.66
phrase	sentence	1049.65
section	section	1012.36
affirmation	assertion	939.46
critères	criteria	666.18
hypothèse	hypothesis	657.6
question	question	649.93
essais	trials	649.01
proposition	proposal	637.17
colonne	column	626.87
résultat	results	591.93
échantillon	sample	589.65
équations	equations	561.7
protocole	protocol	550.38
propositions	proposals	547.03
jours	days	522.09
secteurs	sectors	510.64
thèmes	themes	510.64
critère	criterion	509.32
étapes	steps	497.7
cible	target	432.18
caractéristiques	characteristics	428.87
points	points	418.92
séquence	sequence	415.61
adresse	address	403.98
domaines	domains	347.81
thème	theme	328.53
questions	questions	326.49
signaux	signals	312.94
éléments	elements	309.34
explication	explanation	299.95
valeurs	values	277.41
conclusion	conclusion	273.75
programme	program	256.47
modifications	modifications	252.35
variables	variables	244.06
stratégie	strategy	242.22
paramètres	parameters	239.22
lien	link	238.33
phrases	sentences	233.44
années	years	226.19
pages	pages	223.65
dimanche	Sunday	221.36
nuit	night	204.56
conditions	conditions	204.1
mois	month(s)	181.87
résultats	results	167.18
lettre	letter	149.37
siècle	century	140.4
structure	structure	132
tableau	table	127.31
partie	part	112.44
études	studies	108.51
phase	phase	103.21
journée	day	98.18
taux	rate(s)	97.84
mots	words	74.44
expression	expression	70.37
instant	instant	68.42
position	position	54.64
jour	day	54.29
article	article	41.91
exemple	example	25.09
fois	time	13.6
point	point	11.92