Refining quantitative methods for Diachronic Construction Morphology: a study of three compound families in French

Jan Radimský
University of South Bohemia in České Budějovice
radimsky@ff.jcu.cz

Florence Villoing
Paris Nanterre University & MoDyCo, UMR 7023
villoing@parisnanterre.fr

1. Introduction

French nominal compounds consisting of two nouns (henceforth, "N+N compounds") (see (1)) have been extensively investigated from a synchronic point of view, but much less attention has been paid to the diachronic perspective.

(1) a. BOUCHERIE-CHARCUTERIE (lit. butcher shop.delicatessen) 'butcher shop'
b. ENFANT-ROI (lit. child.king) 'child as a king'
c. CONTRÔLEUR-QUALITÉ (lit. controller.quality) 'quality controller'

They seem to represent a relatively recent innovation in Romance languages. According to Rainer (2021), the pattern does not display any continuity from Latin compounding and probably stems from a variety of different syntactic constructions whose number is extremely limited in French, at least until the end of the 19th century.

Recently, a large-scale study was conducted on the Attributive-Appositive Noun + Noun compounds (henceforth, "N+N ATAP compounds", see § 2. for the classification) of Italian (Radimský 2024 ; Radimský & Micheli 2025). It investigated the origin and diachronic development of these compounds, with a particular focus on a detailed analysis of the history of three semi-schematic families: N-*modello*, N-*base*, and N-*limite* (see (2) for some examples). These three families have been chosen because of the crucial role they have played in shaping the broader N+N ATAP pattern in Italian.

(2) a. *podere modello* (lit. farm.model) 'model farm', *lettore modello* (lit. reader.model) 'model reader'
b. *prezzo base* (lit. price.base) 'base price', *campo base* (lit. camp.base) 'base camp'
c. *caso limite* (lit. case.borderline) 'bordeline case', *età limite* (lit. age.borderline) 'threshold age'

Since two of the three compound families under investigation – namely N-*limite* and N-*modèle* – also developed in French, the present study replicates Radimský's (2024) research on French data, pursuing several objectives. The principal aim is to further refine a methodology designed to trace the diachronic development and constructionalization of semi-schematic compound families by exploiting large-scale data from Google N-grams. In addition to the N-*limite* and N-*modèle* families, which have evolved in parallel in both Italian and French, we include a third French family, N-*surprise*, which exhibits a comparable overall type frequency but is unattested

in Italian. In a broader perspective, this research contributes to the question of how the Attributive-Appositive (ATAP) N+N compound pattern emerged and evolved in French. We proceed from the hypothesis that it remains an open question whether the ATAP pattern as such constitutes a productive higher-order construction in present-day French, or whether its apparent rise in type frequency is primarily driven by a limited number of lower-level, semi-schematic constructions (Bauer 2017:74; Rainer 2016:2714).

The analysis will be conducted within the theoretical frameworks of *Construction Morphology* (cf. § 2). It is based on an original methodology drawing on very large corpus data (§ 3.) and will be both quantitative (§ 4) and qualitative in nature (§ 5).[1]

2. Noun+Noun compounds in French

The identification of sequences of two nouns as belonging to a morphological composition rather than a syntactic sequence has been the subject of much debate (see, for French, Corbin 1997:84; Fradin 2009: 192-206; Villoing 2012: 35-36; Van Goethem & Amiot 2019: § 3; Amiot 2020: 1920-1921, among others). The difficulty of defining this type of compounds in French partly explains why few studies have been devoted to it. However, adopting the theoretical perspective of *Construction Morphology* (Booij 2010, 2016) renders obsolete the problem of modular boundaries in grammar between lexicalized constructions of syntactic or morphological origin. Thus, we will accept as a binomal compound noun any form-meaning pair with a conventionalized denotating function, notably because these pairs are constructions in the sense of *Construction Grammar*.

Nominal compounds of two nouns in French share some key properties:

(i) it is a productive pattern that forms complex nominal units ;
(ii) it involves two bare common nouns without determiner ;
(iii) there is an implicit relationship between the two nouns, without a preposition ;
(iv) compounding is endocentric, and the compound is left-headed.

Despite these general properties, N+N compounds do not all fall under the same semantic type. As for the typology of compounds, we followed the proposal by Scalise and Bisetto (2009), adapted to Romance languages by Radimský (2015). In this typology, compounds of two nouns appear in three subtypes : Coordinative compounds / Attributive-Appositive compounds / Subordinative compounds. Our research will focus on ATAP compounds.

[1] This article builds on the study by Radimský & Villoing (2025) written in French, which is based on the same dataset. While Radimský & Villoing (2025) focus in detail on the qualitative analysis of individual compound families – an aspect of particular interest to a Francophone readership and only briefly outlined here in Section 5 – the present study offers a more in-depth discussion of the methodological issues related to the quantitative description of these families, presented in Sections 3 and 4. The two studies are thus complementary.

Table 1. Main subtypes of NN compounds in French

Coordinative compounds	**Attributive-Appositive compounds (ATAP)**		**Subordinative compounds (SUB)**	
(COORD)	Attributive	Appositive	Verbal-nexus	Ground
BOUCHERIE-CHARCUTERIE ‘butcher shop .delicatessen’ MONTRE-BRACELET ‘watch.bracelet’	ARTISTE-AMATEUR ‘artist.amateur’ TRAIN-MINIATURE ‘train.miniature’	ENFANT-ROI ‘child.king’ TOMATE-CERISE ‘tomato.cherry’	CONTRÔLEUR-QUALITÉ ‘controller.quality’ PRÉVISION-MÉTÉO ‘forecast.weather’	APPAREIL-PHOTO ‘camera.photo’ ZONE-EURO ‘zone.euro’

These N+N ATAP compounds have two interpretations:

(i) The **Attributive** interpretation which is characterized by the fact that the second noun behaves like a modifier of the first, expressing a property or a quality of the head: ARTISTE AMATEUR is interpretated as an *artist* which is an *amateur*.
(ii) The **Appositive** interpretation which is close to the attributive, except that the attributive relationship is metaphoric: ENFANT-ROI denotes a child (ENFANT) which is like a king (ROI).

The interpretation of these compounds is triggered by the modifier (i.e. the rightmost element, the second noun). They tend to form strong modifier-based families, which is why selected modifiers with highest type frequencies have sometimes also been analysed as nouns with adjectival value (Noailly 1990).

3. Data

3.1 Underlying corpus

Empirically, this study relies on large-scale diachronic data extracted from the Google Books corpus, specifically in the form of raw frequency lists published by Google as the third version of the French and Italian Google N-grams datasets.[2] The total size of the underlying Google Books corpus for the period 1800–2019 amounts to 318,631,007,871 tokens for French and 120,410,089,963 tokens for Italian. However, corpus size varies considerably over time. For French, this diachronic variation is visualized in Figure 1. The smallest corpus size is recorded for the year 1944, with 296,475,658 tokens, while the largest sample corresponds to the year 2016, with 4,594,932,651 tokens.

[2] https://storage.googleapis.com/books/ngrams/books/datasetsv3.html

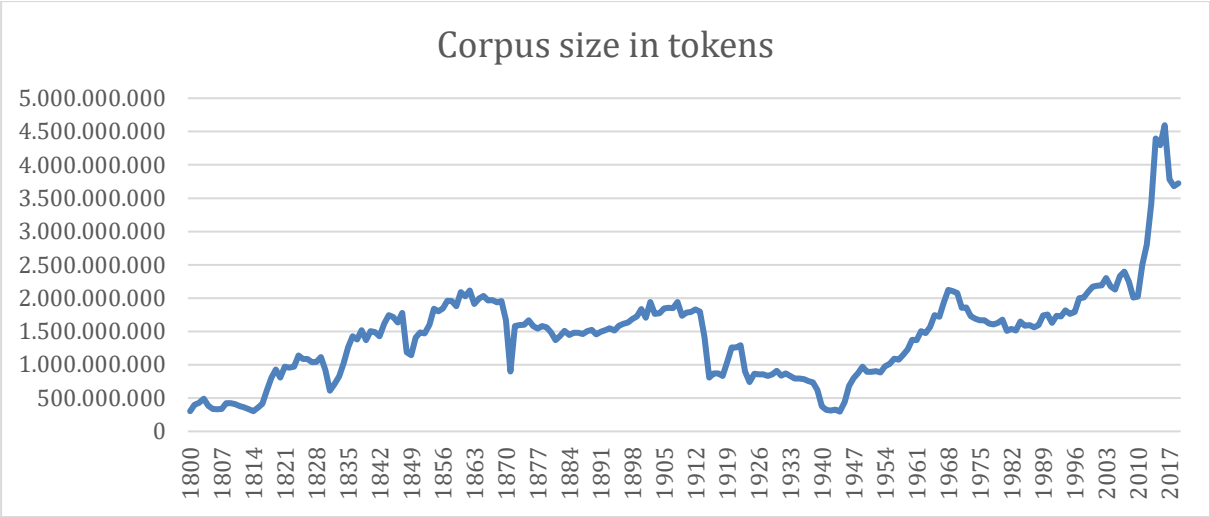

Figure 1. Size of the underlying Google Books corpus for the French Google N-grams frequency lists

These figures demonstrate that we are dealing with data whose magnitude and diachronic granularity are unparalleled in comparison with those typically employed in diachronic linguistic studies to date. A recurring challenge in previous research lies in the limited size of available corpora – often only a few million tokens per diachronic period – and the small number of time slices analyzed (frequently around five). As Hartmann (2018) has shown, this often necessitates the use of complex extrapolation techniques. As noted by Barðdal et al. (forthcoming), "[t]ypically, historical corpora offer a much greater abundance of texts for more recent periods, while the earlier periods are less documented, thus often yielding data scarcity. For instance, [...] in Frantext, the subcorpus for the 13th century is 74 times smaller than the one for the 20th century."

In the dataset used here, diachronic granularity corresponds to the year of publication as recorded in Google Books, which yields 219 discrete time slices (i.e., individual years) for the period 1800–2019.[3] In terms of corpus size per slice, the lowest-resourced year – 1944 – still comprises 296,475,658 tokens, a figure approximately three times larger than that of a so-called "third-generation" synchronic corpus such as the British National Corpus, which has served as a standard benchmark since the 1990s. In this light, while the overall size of the underlying corpus may seem unnecessarily excessive – indeed, it "represents many times the linguistic input that a single person would receive in their lifetime," potentially diluting stylistically marked phenomena (Stefanowitsch 2020: 37) – the sizes of individual subcorpora, especially for the less represented periods, are sufficiently robust to obviate the need for extrapolation.[4]

Nevertheless, the corpus remains quantitatively imbalanced across time slices: the most extensive subcorpus (2016) is 15.5 times larger than the smallest one (1944). It is therefore essential to normalize type frequency values relative to corpus size, as will be discussed below.

This imbalance is not limited to volume but also extends to corpus composition, particularly in terms of genre representation – another longstanding challenge in diachronic research (Barðdal et al. forthcoming). In principle, this issue is intractable for historical corpora, given

[3] Naturally, the assigned year corresponds to the publication date of the document rather than to the date of the text's composition, which means that the re-edition of older sources may affect the accuracy of dating to some extent.

[4] Stefanowitsch (2020: 38) estimates that a corpus generally needs to contain between one million and five hundred million tokens in order to allow for the study of any type of linguistic phenomenon.

that "the texts transmitted to the present represent a random subsample of the whole population, due to largely extralinguistic accidents. Thus, historical corpora can never even remotely capture the full variety of language" (Claridge 2009: 247). Genre balancing is problematic even in synchronic corpora. For instance, the latest conception of the Czech National Corpus (CNC) – one of the largest of its kind – does not define representativeness in terms of proportional genre distribution based on actual usage, but rather as the inclusion of maximal genre diversity, with relative weightings assigned arbitrarily (Cvrček, Čermáková & Křen 2016).

In this regard, the Google Books corpus offers a potentially more representative dataset than traditional historical corpora, due to its greater diversity of text types and genres. Despite its name, it is not restricted to books or literary texts; it includes various types of printed materials archived in libraries, including periodicals, and encompasses not only fiction but also administrative and technical language. This is of particular importance for the present study: Italian data show that N+N compounds began to emerge in the mid-19th century, primarily in the bureaucratic and economic language of the newly unified Kingdom of Italy. This stylistic anchoring explains their absence before the 1950s in diachronic corpora such as CODIT (Micheli 2022), which consist exclusively of literary texts (Radimský 2023).

In conclusion, the uneven distribution of genres across time periods does not necessarily constitute a methodological flaw. Over time, certain genres may naturally emerge, disappear, or change in frequency within the textual population. Thus, an unequal genre representation in large-scale datasets may not indicate sampling bias but may, in fact, reflect structural changes in the textual ecosystem over time.

3.2 Data extraction

The data used for the extraction of N+N compounds were drawn from the bigram and trigram datasets of the third version of the Google N-grams corpus, pre-processed in accordance with the procedure detailed in Radimský (2022). This step was necessary due to the frequent coexistence of two competing orthographic variants of such compounds: the hyphenated form and the spaced form (e.g., *mot-clé* vs. *mot clé*). In the raw data, the hyphenated variant appears as a trigram, while the spaced variant is listed as a bigram, given that the hyphen is treated as a standalone unigram. During pre-processing, these orthographic variants were consolidated into a unified bigram file to allow for consistent treatment. No lemmatization was applied, however, so as to enable future analyses to distinguish between different inflected forms. In technical terms, then, *mot-clé* and *mot clé* (= the singular form in two orthographic variants) on the one hand, and *mots-clés* and *mots clés* (= the plural form in two orthographic variants) on the other, are treated as two distinct compounds in the present analysis.

The pre-processed bigram file comprised a total of 29,697,536 candidate forms, from which actual N+N compounds had to be identified. This identification followed a multi-stage process. First, an automatic filtering was applied to retain only those bigrams whose constituents – word1 and word2 – could theoretically correspond to French nouns. This filtering excluded, in particular, cases where the second constituent (word2) was in fact a primarily adjectival form with only marginal nominal usage. To this end, the *Lexique 3.83* dictionary[5] and part-of-speech frequency information extracted from the *Araneum Francogallicum Minus* corpus[6] were employed to compile a list of nominal forms in French (a total of 48,231 forms), which served as the basis for automatic filtering of both constituents in the pre-processed bigrams. The filtering criteria for word2 were more stringent than for word1: adjectives with only marginal

[5] Available at http://www.lexique.org/ (cf. New et al. 2004).
[6] See http://ucts.uniba.sk/aranea_about/index.html

nominal use (e.g., *cardiaque*), lexicalized adjectival nouns (e.g., *civil* in *un civil américain*, or *exclu* in *les exclus de la société*), as well as demonyms and toponyms (e.g., *espagnol*) were manually removed. This process reduced the candidate list to 3,324,646 bigram types (non-lemmatized).

Subsequently, the identification of genuine N+N compounds – representing only a small fraction of these candidates – was carried out manually. For the present analysis, we selected three N2-based families for detailed study: N- *limite*, N- *modèle*, and N- *surprise*, within which we identified 143, 144, and 145 compound types, respectively. To ensure maximal accuracy in data extraction, each compound type was manually verified by a native speaker against a sample of contextualized occurrences (tokens) in Google Books. Compounds that exhibited a high rate of false positives at the token level were excluded from the final dataset.

4. Diachronic profile of the families

4.1 Methodological preliminaries

The traditional methodological framework for analysing morphological productivity relies on Baayen's (2009) measures of Realized Productivity, Expanding Productivity, and Potential Productivity – two of which depend on hapax legomena. It is important to recall that these metrics were originally developed to infer productivity patterns even in relatively small or strictly synchronic corpora. As Berg (2020: 1118) observes, hapax-based measures offer only indirect approximations of productivity, and their interpretive value diminishes considerably when applied to very large corpora. For these reasons, our quantitative analyses will prioritize direct approaches wherever feasible. However, as Hüning (2019: 485) emphasizes, genuinely diachronic methods for tracing productivity change are still in their infancy and face substantial methodological challenges.

We assume that the diachronic development of constructions can be analysed directly by observing two complementary dimensions: (i) the gradual emergence of new types, observable through their first attestations (a direct approach advocated by Berg 2020), and (ii) changes in type frequency across successive time periods. Although these approaches constitute promising alternatives to hapax-based methods, their practical implementation is far from straightforward, particularly when working with large-scale datasets such as Google N-grams.

Given the nature and limitations of the available data, we therefore propose a set of quantitative procedures focusing on: (i) estimating the period during which the vast majority of types become attested for the first time (the "sample saturation period", opposed to the "onset period"), (ii) examining diachronic changes in realized productivity, including the detection of potential breakpoints, and (iii) identifying salient leader words – i.e., particularly frequent items in specific periods that may have influenced the evolution of the respective constructional patterns. The former two methodological components will be introduced and applied in the following subsections, whereas the identification of leader words will be explored qualitatively in Section 5.

4.2 First attestations: Identifying the onset, expansion and saturation periods of the sample

In theory, it should be possible to directly retrieve the year of first attestation for each compound in our sample from the Google N-grams data. In practice, however, this is not feasible. Despite careful manual selection procedures, each compound still contains a proportion of false positives, and due to OCR-related errors, the likelihood of such false positives increases in earlier periods. Consequently, the directly observed year of first attestation cannot be considered fully reliable. Nevertheless, the diachronic evolution of absolute type frequency

allows us to approximate the initial emergence of each family and to identify key moments in their constructionalization. A suitable procedure is proposed below.

Let us first examine the raw type frequency for the three families – i.e., the absolute number of N+N compounds belonging to each family across time, as visualized in Figure 2.

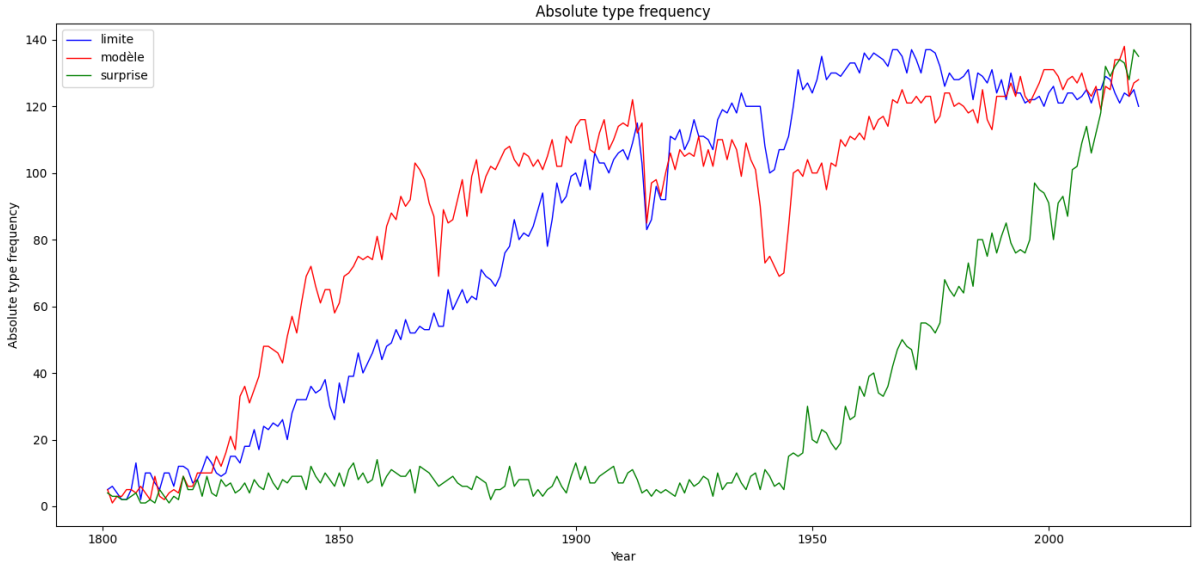

Figure 2. Absolute type frequency of the sample corresponding to the N-*limite*, N-*modèle*, and N-*surprise* families

The specific shape of these curves is not of primary relevance, since the underlying corpus size varies considerably across years (see Figure 1). Moreover, with Google N-grams, it is not possible to control for corpus size per time slice – as recommended by Berg (2020) – because the full underlying text data are not publicly accessible. The presence of false positives also implies that the real number of types is slightly lower than indicated in Figure 2. Nonetheless, the data visualized in Figure 2 offers a reasonably reliable indication of the period in which new compounds within each family first emerged. This emergence period – corresponding to the constructionalization of the respective families – can be divided into three distinct diachronic stages:(i) the *onset period*, (ii) the *expansion period*, and (iii) the *saturation period*.

Let us begin by defining and determining the sample saturation period. To put it simply, sample saturation is reached when nearly all compounds of a given family have already been attested for the first time, so that no new items (types) appear after this date. Given the methodological limitations outlined above, and as a precaution, we define and operationalize the beginning of the saturation period as the time span in which at least 95% of the NNs belonging to each family have appeared in the data in a more systematic manner – that is, over at least three consecutive years. If we take the first year of such a triennium as the milestone, saturation is reached in 1957 for N-*limite*, and in 2014 for both N-*modèle* and N-*surprise*. In other words, after these time points, the respective families ceased to produce new compounds.

The onset period can be viewed as the exact opposite of the saturation period, i.e. as the time span in which the attested compounds are so few that they are not likely to be formed according to the respective family schema which, at this time, does not exist yet. To operationalize the end of the onset period, two factors are to be considered: (i) the minimum number of types that can make up a higher-level construction is to be determined rather in absolute than in relative

figures – the lowest possible figure being 2 or 3 compounds[7]; and (ii) in older time periods the probability of false positives increases, so by precaution the required minimum type frequency should be higher. Let us therefore define the end of the onset period as the first year of the triennium in which at least 10 different compounds are attested in three consecutive years. According to this criterion, the onset period ends in 1816 for N-*limite*, in 1824 for N-*modèle*, and only in 1945 for N-*surprise*.

As a result, the expansion period covers the time span between the end of the onset period and the beginning of the saturation period, corresponding to the time in which the newly established construction (family) produces new compounds. The three periods for the respective families are summarized in Table 1.

Table 2. Onset, expansion and saturation periods for the N- *limite*, N- *modèle*, and N-*surprise* families

	Onset	Expansion	Saturation
N-limite	< 1816	1816-1957	> 1957
N-modèle	< 1824	1824-2014	> 2014
N-surprise	< 1945	1945-2014	> 2014

The longest expansion period is attested for the N-*modèle* family (1824–2014). Conversely, the expansion period of N-*limite* (1816–1957) and that of N-*surprise* (1945–2014) hardly overlap, which suggests that these constructions developed diachronically in two different historical periods.

4.3 Realized productivity

The data presented in Figure 2 reflect the number of N+N compounds for each of the studied families, which broadly corresponds to the notion of *rentabilité* 'profitability' as defined by Corbin (1987: 177), or to Baayen's (2009) concept of *realized productivity*. However, before interpreting the trajectories of these values, it is necessary to normalize them relative to the size of the underlying corpus for each year.

A straightforward normalization approach would involve dividing the absolute type frequency (V) by the number of tokens in the corresponding subcorpus (N). However, as illustrated by Baayen's (2009: 903) *Vocabulary Growth Curve*, the number of new types (V) does not increase proportionally with corpus size (N); rather, the rate of type emergence decreases as corpus size expands. This poses a particular challenge for large-scale corpora such as the one used in the present study. While the largest diachronic subcorpus (for the year 2016) in the French Google Books dataset is 15.5 times larger than the smallest one (1944), the actual difference in the number of new types is expected to be considerably smaller. Consequently, it is more appropriate to normalize type frequency relative to the number of unique word forms within each subcorpus.

For this purpose, we used the number of alphabetic unigram types (i.e., unigrams consisting exclusively of alphabetic characters) per diachronic slice as an estimate of the number of unique word forms. The range between the minimum (925,378 unique forms in 1814) and the maximum (2,849,533 in 1970) corresponds to a factor of approximately three (more precisely, 3.08). A diachronic overview of the number of unique word forms is presented in Figure 3.

OBJ

[7] See Jackendoff & Audring (2020: 224-225).

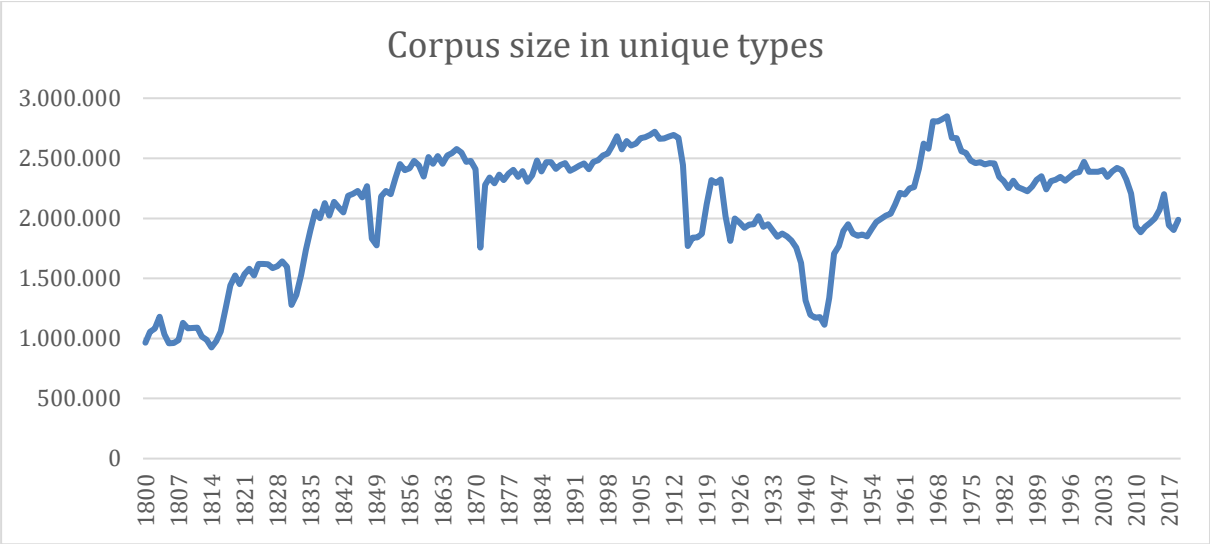

Figure 3. Number of unique word forms in the alphabetic unigrams of the French Google N-grams dataset

It is worth noting that the overall shape of the curve in Figure 3 closely mirrors that of Figure 1, though with considerably less pronounced fluctuations. A particularly interesting divergence appears in the most recent time periods, from 2010 onward: while the token-based curve (Figure 1) shows a steep upward trend reflecting a dramatic growth in corpus size, the curve for unique word forms (Figure 3) remains within the bounds of earlier trends. This suggests that the corpus may have reached a threshold at which the addition of large volumes of texts from similar registers and domains no longer significantly increases lexical diversity.

The normalized type frequencies, calculated relative to the number of unique word forms, are shown in Figure 4.

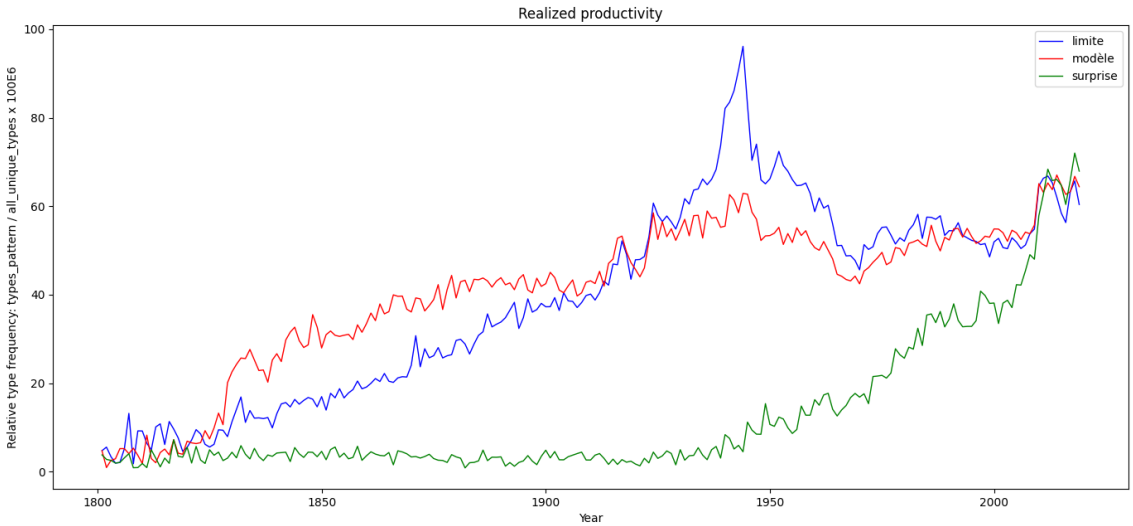

Figure 4. Relative type frequency (realized productivity), normalized by the number of unique word forms

According to Figure 4, the N-*limite* and N-*modèle* families exhibit a continuous increase in type frequency from the early 19th century until 1944, at which point a discontinuity occurs, followed by a decline and then a phase of stabilization or mild increase. After normalization, the apparent dips observed in the absolute values (Figure 2) during the 1920s and 1940s no longer appear, indicating that both families maintained steady growth in realized productivity until the mid-20th century. By contrast, the N-*surprise* family shows a similar trajectory in both

the absolute (Figure 2) and normalized (Figure 4) data, indicating a sustained influx of new forms and a continued growth in productivity since the 1940s.

In the next section, we will show how the shape of these curves can be divided into distinct phases, and how these phases relate to the onset, expansion, and saturation periods identified above.

4.4 Phases in the realized productivity curve

The diachronic trends in constructional patterns, as visualized through realized productivity in Section 4.3, can be further analyzed to identify potential change points marking shifts in their evolution. An initial attempt to apply the variability-based neighbour clustering method proposed by Gries and Hilpert (2008) did not yield satisfactory or intuitively interpretable results for the trajectories observed in our dataset. We therefore adopted an alternative method, tested in previous studies (Radimský 2024, 2025; Radimský & Micheli 2025), to analyze the N-*limite* and N-*modèle* families, which display a pronounced degree of variability.

To visualize diachronic trends via regression lines, we employed the Theil-Sen estimator (Sen 1968), supplemented by the Mann-Kendall trend test (Kendall 1975) to assess the statistical significance of the identified trends (Python implementation based on Hussain & Mahmud 2019). These non-parametric, rank-based methods are suitable for detecting various types of dependencies (not limited to linear relationships), do not require normally distributed residuals, and are robust to outliers. They are therefore well-suited to trend analysis in lexical usage based on large-scale diachronic corpora (see Kovář & Herman 2013). Key change points can be estimated – within a reasonable margin of approximation – through visual inspection of the realized productivity curve. Their validity is then confirmed (or rejected) if the Mann-Kendall test detects a statistically significant change in trend between the segments preceding and following the proposed point.

The realized productivity curve for the N-*limite* family can thus be divided into four phases, delineated by three change points, as shown in Figure 5.

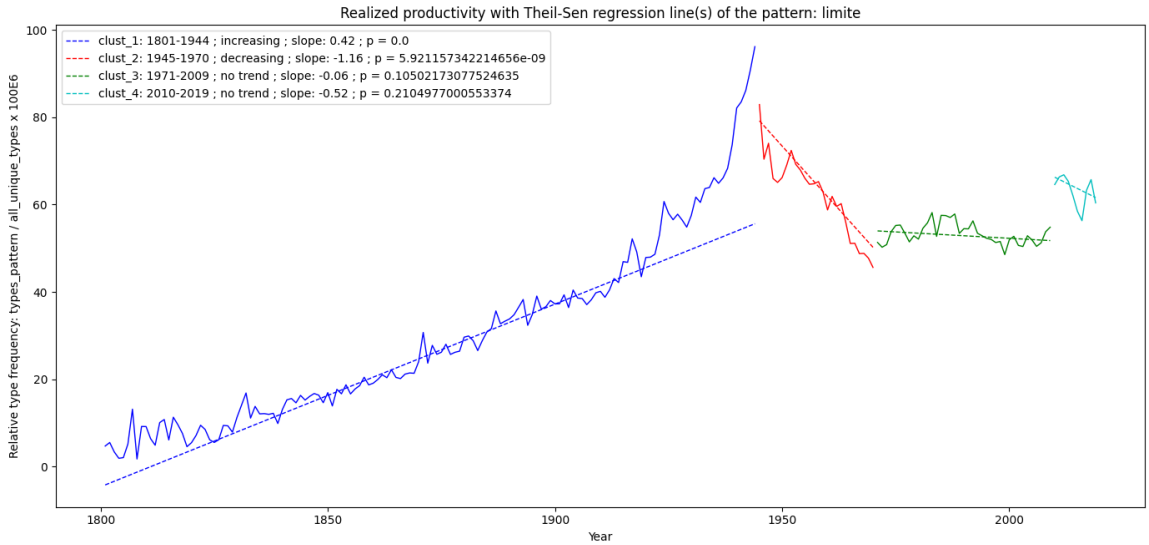

Figure 5. Relative type frequency (realized productivity) of the N-*limite* family, with diachronic segmentation

The first change point, in 1944, corresponds to the curve's absolute maximum, while the second, in 1970, marks its lowest value since the mid-20th century. In other words, after an extended initial period of growth (1800-1944), the curve undergoes a sharp trend reversal. The

second phase, which begins only 13 years before the start of the saturation period identified in the previous section (1957), is characterized by a declining trajectory.

From 1971 onward, two interpretive possibilities emerge. The first is to treat the entire 1971-2019 period as a single phase; in this case, the Mann–Kendall test reveals a mildly increasing trend, although the relatively high *p*-value (p = 0.048) suggests that this trend is only marginally significant. The second possibility, illustrated in Figure 5, isolates a third phase between 1971 and 2009, representing nearly four decades of stability, during which the slope of the regression line is close to zero. The final portion of the curve – post-2009 – is difficult to interpret. This segment is relatively short and marked by considerable variability, such that small adjustments to the boundaries of the cluster yield different trend outcomes in the Mann–Kendall test. A longer diachronic perspective will therefore be required before this final phase can be meaningfully analysed.

The diachronic segmentation visualized in Figure 5 aligns quite closely with the key phases identified in Section 4.2. Although no significant trend change appears at the boundary between the onset and expansion phases (which is not to be expected), the trend reversal in 1944 lies close to the end of the expansion phase (1957), suggesting that saturation effectively takes place during the decade 1945-1955. After this point, no significant increase in type frequency is detected.

The realized productivity curve for the N-*modèle* family can be divided into five phases, delineated by four change points, as shown in Figure 6.

Figure 6. Relative type frequency (realized productivity) of the N-*modèle* family, with diachronic segmentation

Similarly to the N-*limite* family, the N-*modèle* family displays a mild increase during the onset period (cluster 1, from 1800 to 1824) and a decreasing tendency between 1945 and 1970 (cluster 4). However, the expansion period of N-*modèle* is longer and more articulated than in the case of N-*limite*. While the latter exhibits a continuous increase throughout its entire expansion period until the 1940s–1950s, the development of the former is structured into a sharp increase (cluster 2), a slower increase (cluster 3), a decline (cluster 4), and a renewed increase (cluster 5). This confirms the observation made in Section 4.2 that the expansion period of the N-*modèle* family extends into the 2010s.

5. Qualitative analysis: example of the emergence of a productive family, the N-MODÈLE family

The quantitative analyses presented above can be complemented by a detailed qualitative investigation into the history of each compound family. In this case, our methodology involves both an analysis of available dictionaries and an examination of so-called *leader words* (see Burdy 2019), that is, compounds with the highest relative token frequency in different diachronic stages, which most likely played a crucial role in the emergence and development of the respective families. As an illustration, we will briefly discuss the leader words of the N-*modèle* family as identified in the Google Books corpus (for a thorough analysis of the three families under investigation, see Radimský & Villoing 2025).

The first leader word before the 1930s and up until the 2000s was *ferme-modèle* (lit. farm.model) 'model farm', conveying the meaning of the family "organization to imitate" (as in Italian). This compound emerged at the time of the development of modern agriculture in France and Europe (1930-45), then in the colonies (Africa, Asia), and the Near and Middle East. It is found in literary works and in history, geography, and economics.

The second leader word in this family is *école-modèle* (lit. school.model) 'model school'. It emerged at the time of the rise of public and secular education, to designate schools training teachers; then the term spread to describe other types of schools (agriculture, commerce, architecture, military) during the years 1930-45-60, and in other countries (China, the USSR, Israel; African and Asian countries). It remains in literature.

The third leader word is *lecteur-modèle* (lit. reader.model) 'model reader', which emerged in the 1980s. It belongs to the entire semantic subfamily of compounds meaning 'perfect in their kind' that have been present since before the 1930s, with *époux-modèle* (lit. husband.model) 'ideal husband', *femme-modèle* (lit. woman.model) 'model woman', *fille-modèle* (lit. girl.model) 'model girl' and others. But it is the compound *lecteur-modèle* that marks the turning point for the entire semantic subfamily, rising to the status of leader word from 2000. It immediately leads to about ten compounds of the same type. It was completely absent from the top twenty leading words during the four previous periods, but it emerged from the publication of Umberto Eco's novel in which he proposes the notion that becomes extremely popular in the intellectual world. The notion is taken up in works of literary, linguistic, cinematographic, psychoanalytic, and sociological studies. It therefore seems that this compound has caused a real revival of the family of N-*modèle* compounds, in particular those whose N1 refers to a person.

6. Conclusion and future work

This study has explored the diachronic evolution of three semi-schematic compound families in French (N-*modèle*, N-*limite*, and N-*surprise*) using large-scale data extracted from Google N-grams. It complements previous research on Italian (Radimský 2024), where parallel families (N-*modello*, N-*limite*, and N-*base*) developed during the same historical period (1800-2000).

From a coarse-grained perspective, the diachronic trajectories of these three French and Italian families closely mirror each other, both in terms of timeline and domain-specific diffusion. Upon closer inspection, however, important divergences arise, confirming Franz Rainer's intuition that each family has probably followed its own specific historical trajectory. First, not all three families are parallel to the same extent: while Italian exhibits a robust development of the N-*base* family, its French equivalent failed to achieve comparable productivity; conversely, the N-*surprise* family, absent in Italian, demonstrates a sustained

trajectory in French, with increasing productivity well into the contemporary period. Second, the data reveal important differences in diachronic trajectories within each language. The expansion of the French families N-*limite* and N-*surprise* was continuous, but took place in two distinct, non-overlapping periods. By contrast, the expansion of the N-*modèle* family extended over a longer time span and unfolded in two phases, separated by a period of decline between the mid-1940s and the mid-1970s. Google N-gram data also make it possible to identify and qualitatively analyse the different *leader words* that may have played an important role either in the initial formation of the respective families or in their subsequent development. This was only briefly illustrated here by a few examples from the N-*modèle* family.

Beyond its empirical contributions, this study also proposes several methodological innovations in the quantitative analysis of diachronic constructional change. It introduces a set of procedures to identify onset, expansion, and saturation periods for compound families, based on first attestations and type-frequency trajectories derived from Google N-gram data. Furthermore, it normalizes realized productivity using the number of unique word forms – rather than token counts – thereby addressing the limitations of applying traditional frequency metrics to large, genre-heterogeneous corpora such as Google N-grams. The use of Theil-Sen regression and the Mann-Kendall trend test provides a robust statistical framework for detecting change points in constructional development, complementing and refining existing models of productivity based on hapax legomena or vocabulary growth curves.

Future research should extend this approach to a broader set of N+N ATAP families in French and in other Romance languages, as such a systematic cross-linguistic comparison will be essential to fully reconstruct the mechanisms of interlingual influence. Another important issue left for future research, touched upon in Radimský & Micheli (2025), concerns a better understanding of how the expansion of individual families contributes to the emergence of more abstract higher-order constructions, such as the construction corresponding to N+N ATAP compounds.

Acknowledgements

This research was supported by the research project No. 25-15350S, entitled *How do word-formation patterns emerge? An empirical diachronic analysis of Italian N+N compounds*, funded by the Grant Agency of the Czech Republic (GAČR) and conducted at the Faculty of Arts, University of South Bohemia in České Budějovice.

References

Amiot. Dany. 2020. Procédés morphologiques de création lexicale. *Grande Grammaire Historique du Français (GGHF)*. 1894-1927.

Baayen, R. Harald. 2009. Corpus linguistics in morphology: Morphological productivity. In Anke Lüdeling & Merja Kytö (eds.), *Handbooks of Linguistics and Communication Science*, 899–919. Berlin, New York: Mouton de Gruyter. https://doi.org/10.1515/9783110213881.2.899.

Barðdal, Jóhanna, Renata Enghels, Quentin Feltgen, Sven Van Hulle & Peter Lauwers. Forthcoming. Productivity in Diachrony. In *Wiley Blackwell Companion to Diachronic and Historical Linguistics*. Hoboken: Wiley-Blackwell.

Bauer, Laurie. 2017. *Compounds and Compounding*. Cambridge University Press.

Berg, Kristian. 2020. Changes in the productivity of word-formation patterns: Some methodological remarks. *Linguistics* 58(4). 1117–1150. https://doi.org/10.1515/ling-2020-0148.

Booij, Geert. 2010. *Construction Morphology*. Oxford: Oxford University Press.

Booij, Geert. 2016. Construction Morphology. In Andrew Hippisley & Gregory Stump (Eds.), *The Cambridge Handbook of Morphology*, 424–448. Cambridge: Cambridge University Press.

https://doi.org/10.1017/9781139814720.016.
Burdy, Philipp. 2019. On the importance of leader words in word formation: The popular transmission of the Latin abstract-forming suffix *-io* in French. *Word Structure*. 12(1). 42–59. https://doi.org/10.3366/word.2019.0138.
Claridge, Claudia. 2009. Historical corpora. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: an international handbook* (Handbücher Zur Sprach- und Kommunikationswissenschaft = Handbooks of Linguistics and Communication Science Bd. 29.2), 242–259. Berlin, New York: Walter de Gruyter.
Corbin, Danielle. 1987. *Morphologie dérivationnelle et structuration du lexique* (Linguistische Arbeiten 193). Tübingen [Lille]: M. Niemeyer [diff. Presses universitaires de Lille].
Corbin, Danielle. 1997. Locutions, composés, unités polylexématiques: lexicalisation et mode de construction. In Martins-Balter (éd.), *La locution entre langue et usage*, 53-101. Lyon: ENS Éditions. https://doi.org/10.4000/books.enseditions.18713.
Cvrček, Václav, Anna Čermáková & Michal Křen. 2016. Nová koncepce synchronních korpusů psané češtiny. *Slovo a slovesnost* 77(2). 83–101.
Fradin. Bernard. 2009. Compounding in French. In Rochelle Lieber & Pavol Štekauer (eds.), *The Oxford Handbook of Compounding*, 417–435. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199695720.013.0022.
Gries, Stefan & Martin Hilpert. 2008. The identification of stages in diachronic data: variability-based neighbour clustering. *Corpora* 3(1). 59–81. https://doi.org/10.3366/E1749503208000075.
Hartmann, Stefan. 2018. Derivational morphology in flux: a case study of word-formation change in German. *Cognitive Linguistics* 29(1). 77–119. https://doi.org/10.1515/cog-2016-0146.
Hüning, Matthias. 2019. Morphological Theory and Diachronic Change. In Jenny Audring & Francesca Masini (eds.), *The Oxford Handbook of Morphological Theory*, 475–492. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199668984.013.28.
Hussain, Md. & Ishtiak Mahmud. 2019. pyMannKendall: a python package for non parametric Mann Kendall family of trend tests. *Journal of Open Source Software* 4(39). 1556. https://doi.org/10.21105/joss.01556.
Jackendoff, Ray & Jenny Audring. 2020. *The texture of the lexicon: Relational Morphology and the Parallel Architecture*. Oxford, New York: Oxford University Press.
Kendall, M. G. 1975. *Rank correlation methods* (4th ed.). London: Charles Griffin.
Kovář, Vojtěch & Ondřej Herman. 2013. Methods for Detection of Word Usage over Time. In Aleš Horák & Pavel Rychlý (eds.), *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2013*, 79–85. Tribun EU. https://www.muni.cz/en/research/publications/1131909. (4 March, 2022).
Micheli, M. Silvia. 2022. CODIT. A new resource for the study of Italian from a diachronic perspective: Design and applications in the morphological field. *Corpus* (23). https://doi.org/10.4000/corpus.7306.
New, Boris, Christophe Pallier, Marc Brysbaert & Ludovic Ferrand. 2004. Lexique 2 : A new French lexical database. *Behavior Research Methods, Instruments, & Computers* 36(3). 516–524. https://doi.org/10.3758/BF03195598.
Noailly, Michèle. 1990. *Le substantif épithète*. Paris: PUF.
Radimský, Jan. 2015. *Noun+Noun compounds in Italian: a corpus-based study* (Episteme. Theoria). České Budějovice: Jihočeská univerzita v Českých Budějovicích.
Radimský, Jan. 2022. Towards a diachronic analysis of Romance morphology through Google n-grams. *Corpus* (23). https://doi.org/10.4000/corpus.7205.
Radimský, Jan. 2023. Where did the Italian Verbal-Nexus N+N compounds come from? *Proceedings of the Mediterranean Morphology Meetings* 13. 71–82. https://doi.org/10.26220/mmm.4407.
Radimský, Jan. 2024. I composti N+N attributivi in diacronia: le famiglie N-modello, N-base e N-limite. *Études romanes de Brno* (3). 10–34. https://doi.org/10.5817/ERB2024-3-2.
Radimský, Jan. 2025. I composti NN romanzi: una storia di famiglie. Il caso di N-chiave e N-fantasma. In Sara Matrisciano, Johannes Schnitzer, Elisabeth Peters & Franz Rainer (eds.), *Patterns, variants and change: through the prism of morphology: studies in honor of Franz Rainer* (Travaux de Linguistique Romane Morphologie, Syntaxe, Grammaticographie 5), 351–370. Strasbourg: ELiPHi - Éditions de Linguistique et de Philologie.
Radimský, Jan. & M. Sylvia Michelli. 2025. Expanding Horizons: A comprehensive insight into the evolution of Italian Attributive-Appositive Noun+Noun Compounds using Google Books Data. *Lexique*, *Numéro spécial*. https://doi.org/10.54563/lexique.1948
Radimský, Jan. & Florence Villoing. 2025. Le rôle de trois familles semi-spécifiées dans le développement du modèle de composition N-N de type Attributif. *Langue française,* 228. 9-21.
Rainer, Franz. 2016. Italian. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen & Franz Rainer (eds.), *An International Handbook of the Languages of Europe*, Volume 4 Word-Formation, 2712–2731. Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110379082-017.
Rainer, Franz. 2021. Compounding: From Latin to Romance. In Aronoff (Ed.), *Oxford Research Encyclopedia of Linguistics*, 1–31. Oxford: Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.691

Scalise, Sergio & Antonietta Bisetto. 2009. The Classification of Compounds. In Rochelle Lieber, & Pavol Štekauer (eds.), *The Oxford Handbook of Compounding,* 34–53. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199695720.013.0003

Sen, P. K. 1968. Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*, *63*(324), 1379–1389.

Stefanowitsch, Anatol. 2020. *Corpus linguistics: A guide to the methodology*. Zenodo. https://doi.org/10.5281/ZENODO.3735822.

Trésor de la langue Française Informatisé. http://www.atilf.fr/tlfi

Van Goethem, Kristel & Dany Amiot. 2019. Compounds and multi-word expressions in French. In Barbara Schlücker (ed.), *Complex lexical units*, 127–152. Berlin: De Gruyter. https://doi.org/10.1515/9783110632446-005.

Villoing, Florence. 2012. French compounds. *Probus*, *24*(1). 29-60.