Received: May 2024 / Accepted: September 2024
Published online: April 2025

Expanding Horizons: A comprehensive insight into the evolution of Italian Attributive-Appositive Noun + Noun Compounds using Google Books Data

Élargir les horizons : une analyse approfondie de l'évolution des composés N + N attributifs-appositifs en Italien à partir des données de Google Books

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

M. Silvia Micheli
University of Milan
maria.micheli@unimi.it

Abstract

This study builds upon previous research by Radimský (2023) to offer a comprehensive analysis of Italian NN Attributive-Appositive (ATAP) compounds from 1850s to these days. By using an expanded and more accurately annotated dataset, the study unveils the nuanced dynamics shaping the pattern's productivity from a diachronic point of view. Through the Theil-Sen regression and the Mann-Kendall test, the study identifies four distinct clusters delineating the exponential growth of ATAP compounds over time. However, a non-uniform distribution of this growth emerges among lower-level constructions, with the expansion of existing N2-based families driving the pattern's dissemination, especially since 1950s. Remarkably, the study also highlights the subsequent emergence and dissemination of N1-based families that eventually surpass N2-based families starting from the 90s. This complex interplay between N1- and N2-based families indicates the need to analyze both types of families to better understand what mechanisms are at play and what types of generalizations in addition to the formal identity of constituents operate in the formation of new compounds. From a methodological perspective, this paper introduces a novel empirical approach, termed *Relative Family Type Frequency* (RFTF), for evaluating the extent to which

higher-level constructions are covered by corresponding lower-level constructions in compounds.

Keywords: compounding, Italian, diachrony, Diachronic Construction grammar, Relational Morphology

Résumé

Cette étude s'appuie sur les recherches précédentes de Radimský (2023) pour offrir une analyse plus approfondie des composés NN du type attributif-appositif (ATAP) en italien de 1850 jusqu'à nos jours. En utilisant un ensemble de données élargi et plus précisément annoté, l'étude dévoile les dynamiques subtiles qui façonnent la productivité du patron NN ATAP d'un point de vue diachronique. Grâce à la régression de Theil-Sen et au test de Mann-Kendall, l'étude identifie d'abord quatre tranches temporelles distinctes qui caractérisent la croissance exponentielle de la productivité des composés ATAP au fil du temps. Cependant, une distribution non uniforme de cette croissance émerge parmi les constructions de niveau inférieur : l'expansion des familles existantes basées sur le composant N2 lexicalement spécifié apparaît comme moteur principal de la diffusion du patron ATAP, notamment à partir des années 1950. Par la suite, l'étude met également en évidence l'émergence et la diffusion des familles basées sur le composant N1 lexicalement spécifié, dont le nombre finit par surpasser celui des familles basées sur le N2 spécifié à partir des années 90. Cette interaction complexe entre les familles basées sur N1 et N2 lexicalement spécifiés indique la nécessité d'analyser les deux types de familles pour mieux comprendre les mécanismes en jeu et les types de généralisations, en plus de l'identité formelle des constituants, qui opèrent dans la formation de nouveaux composés. D'un point de vue méthodologique, cet article introduit une nouvelle mesure empirique, appelée « *Relative Family Type Frequency* » (RFTF), visant à évaluer dans quelle mesure les constructions de haut niveau sont couvertes par les constructions de bas niveau correspondantes dans les composés.

Keywords: composé, Italien, diachronie, Grammaire de construction diachronique, Morphologie relationnelle

1. Introduction

Italian Noun + Noun compounds (henceforth, NNs) have been a subject of extensive study due to their increasing productivity and diverse patterns within the Italian language (see Radimský, 2015 for a synchronic overview). One specific type, the Attributive-Appositive NNs (ATAP; e.g., *parola chiave* 'keyword' or *luogo simbolo* 'symbolic place'), has garnered significant attention due to their fast dissemination, especially in the journalistic variety of the language (see, e.g., Grandi, 2009; Grandi, Nissim & Tamburini, 2011; Radimský, 2016). However, while the synchronic account of these compounds is particularly rich, there has been a notable gap in understanding their diachronic development. Indeed, this topic seems to be particularly intriguing due to at least two reasons. Firstly, while according to Micheli (2020, p. 120) the pattern has reached a significant productivity and dissemination only since the 21st century, the first results of the diachronic survey by Radimský (2023) suggest that this pattern emerged at least 50 years earlier. Therefore, it seems worthwhile to investigate more thoroughly the period during which the first families (i.e., sets of compound words that share a constituent, such as *romanzo fiume*, *discorso fiume*, *riunione fiume*, etc.) were created and began to spread. Moreover, it remains to be clarified to what extent the success of this pattern can be attributed to individual, particularly type-rich families of compounds, or whether the productivity of the entire pattern can be considered homogeneous.

The present paper aims at expanding on Radimský (2023) by providing a comprehensive and more fine-grained analysis of ATAP NN compounds. We will examine thoroughly the diachronic profile of ATAP NNs based on a large sample of almost 3,000 manually filtered compounds (types) and their diachronic frequency profiles drawn from the Google n-gram data. With reference to the theoretical frameworks of Construction Morphology (Booij, 2010, 2016), Relational Morphology (Jackendoff & Audring, 2020) and Diachronic Construction Grammar (Hilpert & Gries, 2009; Traugott & Trousdale, 2013; Goldberg, 2019; Hilpert, 2021, among others), we will analyze the progressive coinage of constructions at different levels of abstraction (i.e. semi-schematic and schematic) and the relationship between them.

Indeed, it is not very often that a new compounding pattern appears and develops in a modern language, in a diachronic period that is quite richly documented by written sources. Therefore, the analysis of this process will not only make it possible to show the specific situation of Italian NN compounds, but also to discuss general theoretical questions concerning the emergence of compounding patterns within the selected framework, such as "coverage" (Goldberg, 2019) or "structural intersection" (Jackendoff & Audring, 2020), as

well as methodological tools designed for analysis of diachronic corpus data. These tools will include methods that cluster the development of patterns into distinct diachronic stages, such as Variability-based Neighbor Clustering (Hilpert & Gries, 2009), and those that assess the coverage of higher-level constructions by corresponding lower-level constructions, referred to in this paper as Relative Family Type Frequency (RFTF) of nuclear and well-established families.

The paper is structured as follows: Section 2 summarizes existing knowledge about ATAP NN compounds. Section 3 briefly introduces the theoretical models here adopted, while Section 4 describes the data-gathering process as well as the methodology used to analyze the pattern's productivity. In Section 5 we will present the results of our analysis: after providing an overview of the realized productivity of the entire pattern (Section 5.1), we focus on the productivity of individual lower-level schemas (Section 5.2), and then delve into the mechanisms at play in the creation of new families in the most recent period (Section 5.3). Some concluding remarks are provided in Section 6.

2. ATAP compounds: what we know so far

ATAP NN compounds represent one of the three compound categories identified by Scalise and Bisetto (2009) along with subordinative and coordinative compounds. They feature an attributive relationship between the head and its modifier, wherein the latter expresses a "property" or "quality" of the head. According to the classification proposed by Radimský (2015, pp. 158-159), the modifier may have either a metaphoric (in appositive NNs, e.g., *parola*$_{N}$ *chiave*$_{N}$ 'keyword' – the word is 'key', important) or a literal (in attributive NNs; e.g., *luogo*$_{N}$ *simbolo*$_{N}$ 'symbolic place' – the place is a symbol) interpretation. When the N2 is a concrete noun used as a metaphoric modifier, the ATAP compound may be transformed at least indirectly, using an operator such as It. *come* 'like' (e.g., *pesce*$_{N}$ *palla*$_{N}$ 'globefish' → this fish is LIKE a globe). In both cases, the interpretation of ATAP compounds is generally triggered by the modifier (i.e. the rightmost element) and they tend to form strong modifier-based families, which is why selected modifiers with highest type frequencies have sometimes also been analyzed as "noun-clad adjectives" (Grandi, Nissim & Tamburini, 2011). However, as we shall see in more detail in our analysis, the formal identity of the second constituent is not the only factor driving the creation of new ATAP compounds, but the leftmost constituent also plays an important role.

Turning to the history of these compounds, Rainer (2021) pointed out that NN compounds do not display any continuity from Latin compounding and rather stems from a variety of heterogeneous syntactic constructions whose number seems extremely limited in

Italian, at least until the end of the 19th Century. An extensive diachronic analysis of Italian compounds based on the CODIT corpus by Micheli (2020, pp. 91-93) identified only a handful of ATAP NNs in Old and Middle Italian. More specifically, she found 3 ATAP NNs referring to fish species and an artifact in Old Italian texts (i.e., *pescespada* 'swordfish' - fish + sword; *pesceporco* 'grey triggerfish' - fish + pig; *arcamensa* 'large cupboard' - ark + table), and 13 ATAP referring to fish species, plants, vegetables and some concrete substances NNs in Middle Italian texts. However, as already mentioned, the preliminary study carried out by Radimský (2023) revealed that this pattern was already productive at least 50 years before the end of the 19th century. This suggests that further research is needed to better understand the emergence and development of this pattern.

3. Theoretical framework

Construction Morphology and Relational Morphology are usage-based models, which entails that schemas available in the constructicon[1] capture generalizations over a critical mass of already attested words. In a diachronic perspective thus, constructionalization must be based on previous individual innovation (in the sense of Traugott & Trousdale, 2013). Once a critical mass of individual lexical innovations is in place, constructionalization – within the Relational Morphology framework (Jackendoff & Audring, 2020) – consists of two steps. First, relational links between the existing words must be built through the process of structural intersection, and then it is necessary to determine whether these new relational schemas are productive. In the case of compounding, we assume that the structural intersection yields primarily semi-schematic constructions, in which chunks of forms (either the leftmost or the rightmost constituent) are shared. Such a view is consistent with the assumption of Bauer (2017, p. 74) that "it is not the N + N pattern of compounding which is productive, but patterns with individual lexemes within that", as well as with the observation of Rainer (2016, p. 2714) that within Italian N + N compounds, "neologisms tend to follow analogues or series of analogues with the same first or second constituent".

While the creation of individual semi-schematic constructions, or "families", seems relatively straightforward, the formation of higher-level constructions is much more intriguing. According to Goldberg's (2019, pp. 51-73) concept of "coverage", higher-level constructions should correspond to areas where encountered examples cluster. In the case of English noun-participle compounds, Hilpert (2015) suggests that these clusters correspond to

[1] According to constructional theories, the term *constructicon* refers to the network made up of the constructions of a language (Diessel, 2023).

N2-based families, such as N-*based* or N-*carved*. However, we will demonstrate that, contrary to expectations, Italian NN ATAP compounds form families triggered not only by lexically specified N2s but also by lexically specified N1s. Moreover, the progressive emergence of these two types of families follows a diachronic path that is far from straightforward. Therefore, coverage should be assessed for both N1 and N2-based families. This indicates that the process of constructionalization, which leads to the progressive establishment of the higher-level construction of NN ATAP compounds, is more complex than anticipated.

4. Data and methodology

Building upon the foundational work detailed in Radimský (2023), this study extends the diachronic analysis of Italian ATAP NN compounds using an enlarged and updated dataset extracted from the Google books corpus of approximately 120 billion tokens, which is publicly available in the form of raw frequency lists known as the 3rd version of Italian Google n-grams[2]. The original dataset, as described in Radimský (2023), encompassed 1,924 ATAP NNs (non-lemmatized types) derived from 47 modifiers (N2s) and 1,148 head nouns (N1s). The current study expands on these results by incorporating a substantially larger sample size, in particular with regard to the number of modifiers, and introducing a more accurate manual data cleaning procedure performed by a native speaker. The new dataset includes 2,997 ATAP NNs (non-lemmatized types), marking a 1.6-fold increase from the preliminary sample. Notably, the number of N2s has surged by 4.3 times to 204 modifiers, while the N1s have increased to 1,499, a 1.3-fold rise. These increments make it possible to capture a more diverse and comprehensive picture of the modifier and head noun landscape in Italian NN ATAP constructions[3].

Methodologically, the data extraction process mirrored that of the previous study (see Radimský, 2023, for more details). In sum, the initial list of compound candidates was compiled from raw frequency lists of bigrams and trigrams and automatically pre-filtered to isolate potential NN formations. Then, to address the high rate of false positives prevalent in the initial dataset, a stringent manual verification was employed. First, relevant modifiers (N2s) and head nouns (N1s) were identified based on data from previous research and dictionaries and then, all candidates within these modifier-based and head-based families present in the data were checked manually by a native speaker (the co-author M. Silvia

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

[3] The complete dataset of NN compounds referenced in this research will be available in open access at: https://osf.io/46qcd/

Micheli). The entire procedure was recursively repeated several times, as the manual identification of new compounds within N2-based families revealed additional relevant N1-based families, and vice versa.

It must be emphasized that all compounds (i.e. types) were meticulously checked by a human native speaker within real contexts (a sample of tokens) in Google Books. Typically, a preliminary review of a sample comprising 10-20 contextualized occurrences (tokens) is sufficient to ascertain whether a candidate represents an authentic ATAP NN compound or a false positive, the latter often resulting from various factors such as OCR-based errors and issues of morphological or syntactic ambiguity. This rigorous verification process facilitated the elimination of numerous false positives[4]. However, frequency counts for the types classified as true positives will inevitably retain some noise, an artifact of restricted access to the original Google Books data and its vast size. Despite these limitations, it can be asserted that the procedure employed yields the cleanest possible data currently obtainable from Google Books in its present form.

For each identified compound (=type), our dataset provides dated token frequencies from 1800 to the present, with annual precision. Stored within a PostgreSQL database, this dataset can be readily converted into insights regarding the diachronic type and token frequencies of various higher-order constructions, such as semi-schematic constructions (N1- or N2-based families), the fully schematic ATAP construction, or others.

While a range of sophisticated psycholinguistic and corpus-based methods have been developed to determine the existence and the productivity of morphological schemas in the synchronic study of present-day languages, there are no well-established direct methods to ascertain “changes” in productivity, which precisely represent the diachronic aspect of the phenomenon (Hüning, 2019, p. 485). Our data would make it possible to determine the first documented occurrences of compounds, a direct measure of productivity recommended by Berg (2020). However, the risk of encountering false positives might be slightly higher in older diachronic layers, due to lower quality of printed documents resulting in OCR errors, which might introduce a significant bias. We opted therefore for measuring rather the realized productivity, one of the classical synchronic productivity measures introduced by Baayen (2009), which can be calculated with annual precision in our data, allowing for detailed analysis of temporal trends in their change. This is usually achived using Kendall’s Tau, a rank-correlation coefficient advocated by Hilpert and Gries (2009) as suitable for assessing

[4] The rate of false positives in the sample, as identified manually by the native speaker, is at least 10% (300 types).

frequency changes in diachronic corpora. In this research we will employ the Mann-Kendall test, a non-parametric test based on Kendall's Tau recommended for trend analysis across various domains, as a tool suitable to test any form of dependence (not only linear) that does not assume a normal distribution of errors and is not sensible to outliers. This, along with the Theil-Sen estimator for regression analysis, is suggested by Herman and Kovář (2013) as particularly suited for identifying trends in word usage within large diachronic corpora[5].

A known challenge common to diachronic research on productivity is the limited size of corpora (usually units of millions of words per one diachronic period) and the minimal number of diachronic periods analysed (usually less than 10), which makes it necessary to resort to complex extrapolation methods, as demonstrated, e.g., by Hartmann (2018). However, the granular year-by-year data provided by Google N-grams, with substantial sub-corpus sizes ranging from 123 million words per year in 1944 to 1.8 billion words per year in 2013 (which represent, respectively, the smallest and the biggest annual sample from the time span between 1850 and 2019), eliminate the need for extrapolation, and still makes it possible to observe relative frequencies recalculated to 100 million words, which approximately corresponds to the standard size of a third-generation corpus, such as the British National Corpus. The difference between the size of the smallest sample from 1944 and the largest sample from 2013 may appear substantial (approximately a 14-fold disparity). However, given the immense scale of the data, this difference has a relatively minor impact on type frequency values, primarily due to a phenomenon widely recognized as Heaps' law. An analysis of unique word forms revealed that their number differs by only a factor of two between the smallest and largest samples (i.e. 778,163 unique forms in 1944 versus 1,497,732 in 2013). Moreover, the fact that realized productivity is restricted to "past achievement", usually quoted as a drawback of this tool which makes it an indirect measure of productivity when applied to synchronic corpora, does not represent an issue when diachronic data are available. The change of realized productivity, observed in diachronic corpora through the lens of the Theil-Sen estimator and the Mann-Kendall test, may be considered as a valid direct productivity measure in large diachronic corpora.

[5] In practice the Python implementation by Hussain and Mahmud (2019) has been used.

5. Results

5.1 Realized productivity of the ATAP pattern and identification of breakpoints

Let us first make a global overview of the whole sample of 2,997 ATAP NNs that corresponds to the schematic construction (1).

(1) ATAP NN construction
$[N_iN_j]N_k \leftrightarrow [N_i\text{-head È (COME) UN(A) } N_j\text{-non-head}]_k$
$[N_iN_j]N_k \leftrightarrow [N_i\text{-head IS (LIKE) A } N_j\text{-non-head}]_k$

Figure 1 gives a diachronic overview of the realized productivity of the ATAP pattern from the 1850s to 2019. In conformity with Baayen (2009) it has been calculated as the relative type frequency ($f_r = V/N$, where V is the number of types, and N is the corpus size in the respective year), and subsequently multiplied by the constant 10^8 so that it intuitively approaches the order of magnitude of the original type frequency data.

In examining realized productivity, we posed the question of whether its diachronic curve undergoes any significant breaks that would segment its development into distinct stages. To identify these breakpoints, Hilpert and Gries (2009) proposed a method named Variability-based Neighbor Clustering (VNC), derived from studies on language acquisition. This algorithm progressively clusters adjacent periods with similar relative frequencies, ultimately producing several relevant clusters. However, Hilpert and Gries (2009) worked with input data comprising a limited number of diachronic slices (approximately 10). When applied to our data, which is an order of magnitude more fine-grained (over 150 temporal periods), VNC did not yield intuitively meaningful results[6]. Therefore, a different approach was adopted to determine relevant diachronic clusters. First, we visually identified potential breakpoints from the graph. While this step is impractical for the smaller number of diachronic slices used by Hilpert and Gries, it is easily feasible with a fine-grained diachronic dataset. We then fitted the individual periods using the Theil-Sen regression line and calculated the trend for each period using the Mann-Kendall test. The number and boundaries of clusters were subsequently adjusted to ensure that the regression line accurately represented the data points, and that there was a clear difference in the slope of the regression line or even a change in trend (increasing / no trend / decreasing) between the periods. In this way, the realized

[6] Even personal communication with Martin Hilpert did not show the way how to modify the VNC algorithm to make it operational.

productivity curve of the ATAP pattern ranging from the 1850s to 2019 in Figure 1 was divided into 4 distinct clusters.

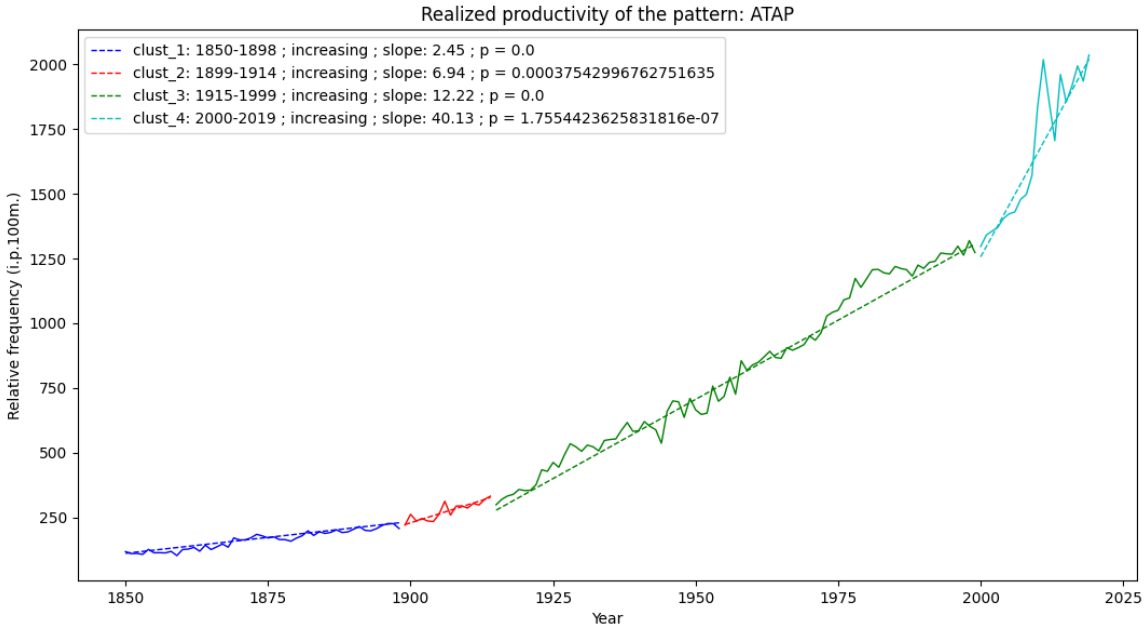


Figure 1. Realized productivity of ATAP NNs in diachrony
(V/N × 10^8, where V is the number of types, and N is the corpus size in the respective year)

Figure 1 illustrates that the realized productivity of the entire ATAP NN pattern demonstrates a growth that closely resembles an exponential increase over the entire period under investigation. Notably, the slope of the regression line for each cluster is approximately twice that of the preceding period. As indicated in the legend, the realized productivity curve shows a significant increasing trend within each diachronic cluster, with the p-value calculated using the Python implementation of the Mann-Kendall test by Hussain and Mahmud (2019).

5.2 Analysis of Coverage: Relative Family Type Frequency

Nevertheless, this overall depiction introduced in Section 5.1 does not fully encapsulate the dynamism of the pattern. Specifically, it fails to reveal the contributions of various lower-level families to this increase; that is, whether the overall rise in type frequency results from the expansion of many diverse lower-level constructions, or from the high type frequency of just a few. In Construction Grammar terms, these two scenarios would signify different extents of "coverage" (Goldberg, 2019) of the ATAP pattern by lower-level constructions. If the growing realized productivity of ATAP compounds were driven primarily by a few semi-schematic constructions (families), such as N-*chiave* (N-'key') or N-*modello* (N-'model'), then the ATAP construction would exhibit uneven coverage by lower-level constructions. In this scenario, the increased productivity of these limited sub-patterns would not enhance the mental

representation of the ATAP pattern as a whole. In other words, new compounds of the N-*chiave* family (2), such as *mercato chiave* (‘key market’), would strengthen the representation of the N-*chiave* family alone, and not contribute to the formation of ATAP compounds from other families, such as *doccia lampo* (lit. ‘shower + flesh’, ‘quick shower’) or *prezzo civetta* (lit. ‘price + owl’, an attractively low price used as a lure for customers, ‘loss leader price’).

(2) ATAP NN semi-schematic construction N-*chiave*
$[N_i$ ***chiave***$]_{Nk} \leftrightarrow [N_{i\text{-head}}$ È chiave/importante$]_k$
$[N_i$ ***chiave***$]_{Nk} \leftrightarrow [N_{i\text{-head}}$ IS key/important$]_k$

Provided that ATAP compounds are known to cluster around modifier-based families – where the modifiers on the N2 position “trigger” the attributive interpretation of the compounds – and that some authors even assume that nominal modifiers in ATAP compounds constitute a relatively limited set of nouns (Baroni, Guevara, & Pirrelli, 2009), we will consider that the primary relevant lower-level constructions within the ATAP pattern are the N2-based families, such as (2). However, in Section 5.3 we will subsequently consider even the possibility that head-based (i.e., N1-based) semi-schematic constructions, such as (3), may also represent relevant lower-level ATAP constructions. Although this may seem a non-canonical approach, the theoretical rationale behind this assumption is drawn from the concept of Structural Intersection within the Relational Morphology framework (Jackendoff & Audring, 2020, pp. 223-225), which will be briefly discussed below in Section 5.3.

(3) ATAP NN semi-schematic construction *pesce*-N
$[$***pesce*** $N_j]_{Nk} \leftrightarrow [$pesce che È COME UN(A) $N_{j\text{-non-head}}]_k$
$[$***pesce*** $N_j]_{Nk} \leftrightarrow [$fish that IS LIKE A $N_{j\text{-non-head}}]_k$

Now, the key question is how to efficiently assess the coverage of the ATAP NN construction by the different semi-schematic constructions, and the coverage changes over time. Hilpert (2015) in his study on English noun-participle compounds, such as *hand-carved* and *computer-based*, achieved this by analysing the token frequencies of the 30 most frequent participle types compared to the token frequency of the whole pattern. Considering that in relative terms “the ‘upper crust’ of 30 highly frequent participle types has been consistently representing a share of about 40% of all usage events” (Hilpert, 2015, p. 130), he concluded that low-frequency types do not account for a progressively larger ratio of tokens, which would entail that the coverage by lower-level constructions does not increase. However, the method that Hilpert applied to a well-established compounding pattern would be difficult to utilize for an emerging compounding pattern, such as Italian ATAP NNs. Additionally, this is an indirect analysis that determines the number of lower-level constructions through their

share of token frequency. Therefore, we will attempt to propose a direct method of analysis that takes into account the changing number of families over time.

Based on the Baayen's concept of realized productivity, which reflects the relative type frequency of the whole pattern and its change over time, we will use an analogous concept called *Relative Family Type Frequency* (RFTF). This new concept mirrors the realized productivity of individual families. In short, RFTF answers the question: how many different N2-based (or N1-based) families are present in the data in a given time period.

One additional problem needs to be addressed: how many different NN types sharing the same N1 or N2 component represent a family that allows speakers to make generalizations? Jackendoff and Audring (2020, p. 222) consider that two types are the strict minimum, but more would be preferable. In this regard, we will distinguish between two types of families. We have set the strict minimum number of members for a family at three, following the principle "tres faciunt collegium". This means that a semi-schematic construction, such as (4), is licensed by at least three different compounds (types), such as (4a-c). Let us call this a *nuclear family*. Conversely, a *well-established family* will be at least twice as large, having at least six different members (types). The number of such families in a sample from a given time period will be referred to as RFTF-3 and RFTF-6, representing the relative family type frequency of nuclear families and the relative family type frequency of well-established families, respectively.

(4) ATAP NN semi-schematic construction N-*fiume*
$[N_i$ ***fiume***$]_{Nk} \leftrightarrow [N_{i\text{-head}}$ È fiume/estremamente lungo$]_k$
$[N_i$ ***fiume***$]_{Nk} \leftrightarrow [N_{i\text{-head}}$ IS EXTREMELY LONG$]_k$

a. *romanzo fiume* 'extremely long novel'
b. *seduta fiume* 'extremely long session'
c. *discorso fiume* 'extremely long speech'
d. *riunioni fiume* 'extremely long meeting'
e. *racconto fiume* 'extremely long story'
f. *lettera fiume* 'extremely long letter'

Let us now examine in detail how the value of RFTF (Relative Family Type Frequency) is calculated. In absolute terms, the family type frequency of, let's say, 3-member N2-based families (FTF-3 of N2s) corresponds to the number of N2 forms whose type frequency (number of identified compounds in which they appear) is greater than or equal to 3 within the specified time span. This result must then be normalized to account for the unequal size of the underlying corpus by dividing it by the number of unique word forms in the given

diachronic sample and multiplying it by a constant (here 10^6) to ensure that the scale aligns with the baseline type frequency values. The resulting formula runs as follows:

$$RFTF\text{-}3\ (N2) = \left(\frac{FTF\text{-}3\ (N2)}{V}\right) \times\ 10^6$$

Where:

- **RFTF-3 (N2)** is the relative family type frequency of at least 3-member N2-based families
- **FTF-3 (N2)** is the number of N2 forms with type frequency ≥ 3
- **V** is the number of unique word forms (in the given diachronic sample)
- **10^6** is a constant, the value of which is determined to be proportional to V (i.e. roughly the mean of the values of V in all diachronic samples).

For the calculation of RFTF, it appears more appropriate, particularly with datasets of this magnitude, to normalize FTF with respect to the number of unique word forms in the diachronic sample (V) rather than the corpus size of the diachronic sample (N), as is the case in the standardized calculation of realized productivity. This is because FTF values are generally very low and more sensitive to the size of the lexicon than to the size of the corpus, meaning normalization by N would significantly distort the results[7].

Figure 2 provides the relative family type frequency of modifier-based nuclear families (i.e. RFTF-3 of N2-based families), showing how the number of 3-member N2-based families changed over time in our sample.

[7] This raises the question of whether, in corpora of such size, the value of realized productivity should also be normalized in the same manner.

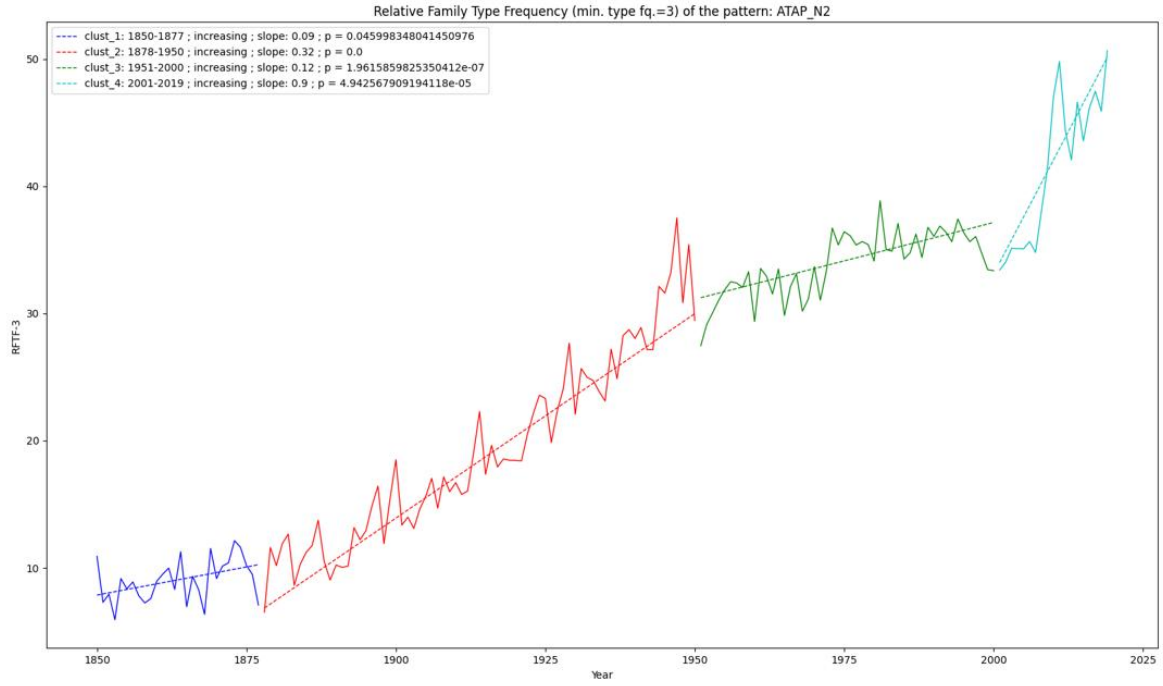


Figure 2. Relative family type frequency of nuclear N2-based ATAP families (RFTF-3) in diachrony

Notice that the graph in Figure 2 has a different shape compared to the graph in Figure 1, particularly for the period covering the second half of the 20th century (1951-2000). During this period, the rapidly increasing type frequency of the ATAP compounds (Figure 1) contrasts with the very slowly increasing RFTF-3 of the N2-based families (Figure 2)[8].

This discrepancy indicates that the increasing number of ATAP compounds is almost exclusively sustained by already existing N2-based families. In other words, the ATAP pattern (1) does not seem to develop during this period; rather, it is only some N2-based families that exhibit growth. This suggests that while there is a proliferation of compounds, they are predominantly variations within established families, rather than the emergence of new (N2-based) families. Consequently, the diversification of ATAP compounds appears limited to certain families that continue to expand, rather than a broadening of the pattern itself through the creation of new N2-based families.

However, this is still not the whole story that can be inferred from the given data, as we will attempt to demonstrate in the next section.

5.3 The interplay between N2-based and N1-based families

The conclusions from Section 5.2 make it possible to ask a follow-up question concerning the inner development of ATAP NNs during the second half of the 20th century: do the established

[8] Notice that in the smaller sample examined by Radimský (2023), the slope of the regression line was even lower, and no trend was observed within this period.

N2-based families develop accidentally, or do the newly formed compounds assemble into some new patterns, perhaps forming new N1-based families? And if so, how important is this phenomenon?

As outlined above, this research question emerges from the concept of Structural Intersection formulated within the Relational Morphology framework in the context of constructionalization in language acquisition (Jackendoff & Audring, 2020, pp. 223-225). In short, the idea behind Structural Intersection is that the relational links between the existing words that make up constructions must be established based on a shared segment of form between these words. In the case of derivation, this is obviously an affix, but in the case of compounds, the shared segment must necessarily be either the leftmost or rightmost word. Therefore, the compounds, such as (5a-f), pertaining to the well-established N2-based families N-*fantasma* (5a-b), N-*satellite* (5c-5d), and N-*simbolo* (5e-f), allow for a new kind of generalization. The N1 *città* ('town'), representing the shared part of (5b), (5d), and (5f), along with the part of speech of the components (N + N) and the attributive interpretation of the relationship between them, licenses speakers to establish the generalized N1-based construction (6), which may, as the case may be, become productive at some point[9].

(5) a. vascello fantasma (lit. 'vessel + ghost', 'ghost ship')
b. **città** fantasma (lit. 'city + ghost', 'ghost town')
c. paese satellite (lit. 'country + satellite', 'satellite country')
d. **città** satellite (lit. 'town + satellite', 'satellite town')
e. luogo simbolo (lit. 'place + symbol', 'symbolic place')
f. **città** simbolo (lit. 'town + symbol', 'symbolic town')

(6) ATAP NN semi-schematic construction *città*-N
[***città*** $N_j]_{Nk}$ ↔ [città che È COME UN(A) $N_{j\text{-non-head}}]_k$
[***città*** $N_j]_{Nk}$ ↔ [town that IS LIKE A $N_{j\text{-non-head}}]_k$

To assess this hypothesis, let us examine the presence of well-established N1-based families within the sample and their diachronic development, compared to the development of well-established N2-based families, on Figure 3.

[9] In an analysis of synchronic data on Italian NN compounds, Radimský (2020) noticed that such family-size effect was prominent with both a specified N1 and N2 for different types of compounds.

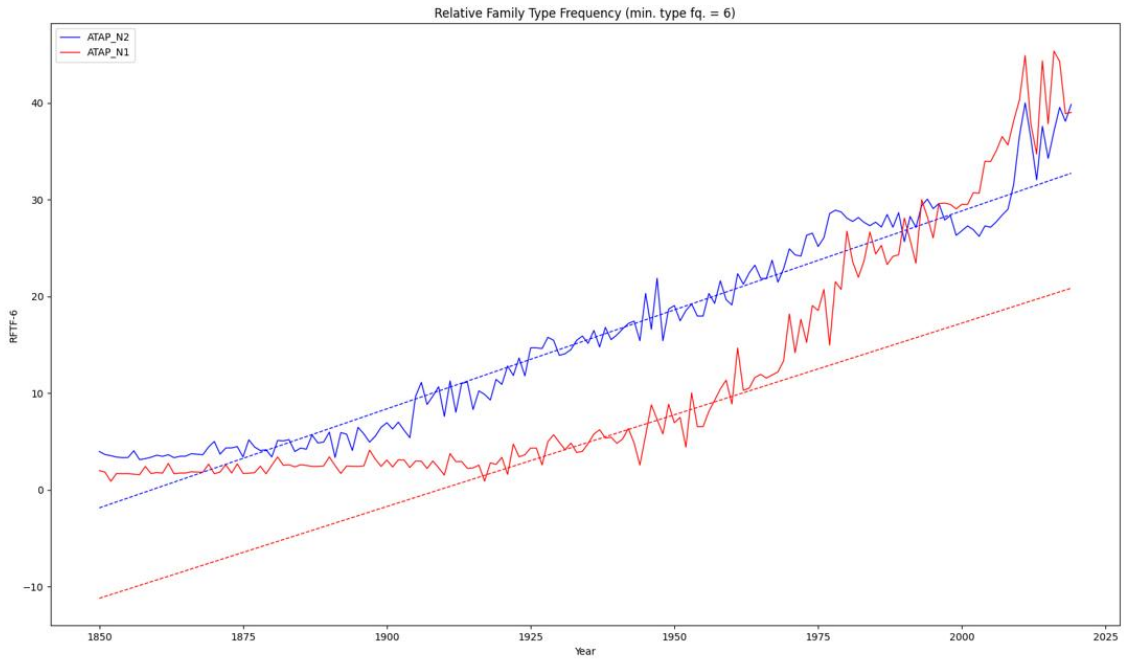


Figure 3. Relative type frequency of well-established N1-based and N2-based ATAP families (RFTF-6) in diachrony

As showed on Figure 3, data provide a very strong support for our hypothesis. While well-established N1-based families had no significant presence in data before 1950s, their number kept constantly growing since then to the point where they even outnumbered N2-based families in 1990s. That is, well-established N1-based families begun to emerge later than their N2-based counterparts, but followed a steeper slope then, which renders the two regression lines virtually parallel for the whole period under investigation.

To complete the picture, let us also observe the number of nuclear N1-based families and their development over time. Figure 4 provides the graph of RFTF-3 of N1-based families that may be directly compared with RFTF-3 of N2-based families, already analyzed on Figure 2. The curve in Figure 4 has been divided into two clusters with the breakpoint situated at the beginning of 1920s.

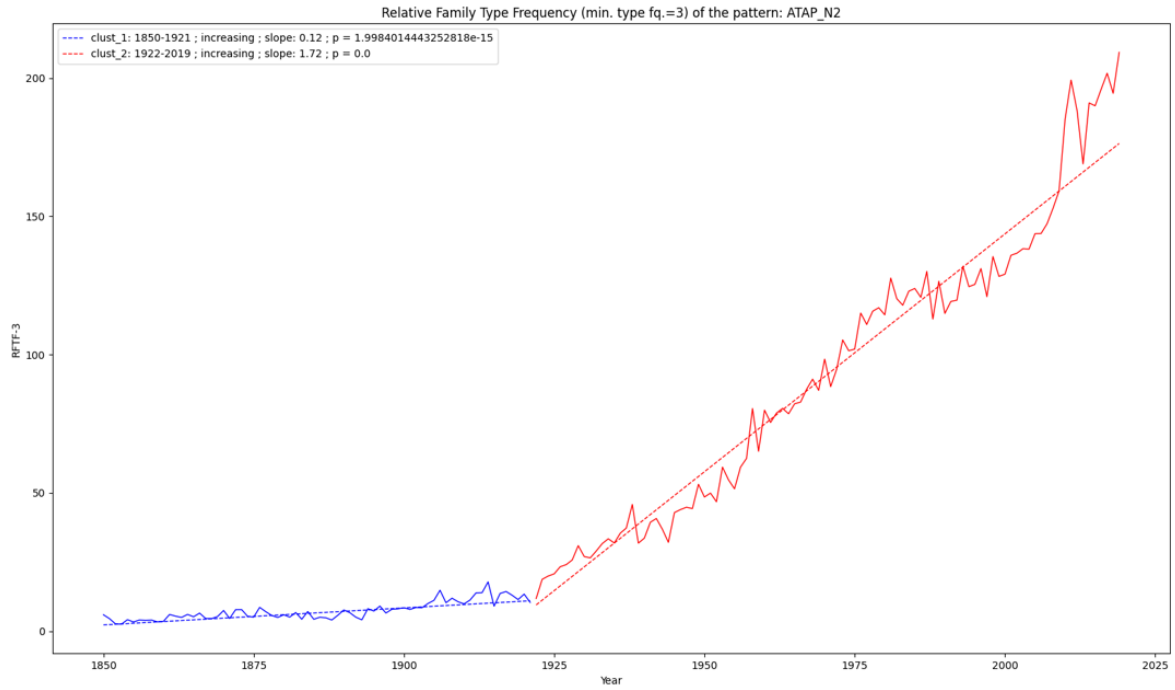


Figure 4. Relative type frequency of nuclear N1-based ATAP families (RFTF-3) in diachrony

Indeed, the RFTF-3 of N1-based ATAP families begun to significantly increase in the beginning of 1920s, that is roughly 40 years later than the RFTF-3 of N2-based ATAP families (see Figure 2), but it keeps increasing since then. Notably, it keeps increasing sharply throughout the whole second half of the 20th century, when the RFTF-3 of N2-based families is experiencing only very slow growth.

This observation leads to important conclusions. First, it means that the established N2-based families do not develop accidentally between 1950-2000. Even though only very few new N2-based families appear in this period, the N1s already present in an existing N2 family tend to expand across other N2 families within the ATAP pattern, giving birth to the progressive coinage of new N1-based families. Indeed, a qualitative look in the data reveals that, besides the old and well-established family *pesce*-N, lying behind the names of more than 50 fishes in Italian[10] and in other European languages, other important N1-based families with more than 10 NN compounds emerge, such as those illustrated in Table 1.

[10] Such as *pesce spada* ('swordfish'), *pesce cane* ('dogfish', 'shark'), *pesce ago* ('pipefish'), *pesce porco* ('grey triggerfish', lit. 'fish.pig'), *pesce sega* ('sawfish'), etc.

N1-based families	Type frequency	Examples
città-N ('city'-N)	43	*città fantasma* ('city-ghost') *città dormitorio* ('city-dormitory')
uomo-N ('man'-N)	33	*uomo bersaglio* ('man-target') *uomo fantoccio* ('man-puppet')
film-N	33	*film evento* ('film-event') *film manifesto* ('film-manifesto')
donna-N ('woman'-N)	27	*donna prodigio* ('woman-prodigy') *donna leader* ('woman-leader')
legge-N ('law'-N)	18	*legge limite* ('law-limit') *legge bavaglio* ('law-gag')
paese-N ('country'-N)	17	*paese modello* ('country-model') *paese vassallo* ('country-servant')
personaggio-N ('character'-N)	14	*personaggio chiave* ('character-key') *personaggio ombra* ('character-shadow')

Table 1. N1-based families with more than 10 NN compounds (types)

Second, we may hypothesize that if some new N1-based families become productive at some point, yielding new instances of ATAP NNs, these do not have to fit only the already established N2-based families. In other words, the new N1-based families, such as (6), may not develop only within the already established N2-based families, but they may cross these borders and give birth to new N2-modifiers and, potentially, to new N2-based families in the future. This is what we actually observe in the final cluster of the graph in Figure 2: a renewed vitality of N2-based families which may be in part due to the productivity of new N1-based families, established within the period 1921-1999. A more detailed examination of this hypothesis would require a fine-grained qualitative analysis of individual families.

Finally, future research should attempt to shed light on the question of whether there might be other types of generalizations that give birth to new lower-level constructions of ATAP-NN compounds than those based on the formal identity of the leftmost or rightmost component. These might include for instance sense-based generalizations based of the word on the leftmost position, such as nouns referring to persons or groups of persons (*uomo* 'man', *donna* 'woman', *personaggio* 'character', *società* 'company', *gruppo* 'group'), nouns referring to places (*città* 'town', *paese* 'country', *regione* 'region', *quartiere* 'quarter'), etc. At the current state of knowledge, it is not clear what kind of constructions might emerge and, especially, how these different constructions interact with each other in diachrony.

6. Conclusions

This study extended the investigation carried out by Radimský (2023) with a larger sample of ATAP compounds and a more rigorous process of verification of data. It revealed that the overall productivity of ATAP compounds shows a near-exponential growth, segmented into four distinct clusters based on significant breakpoints identified through Theil-Sen regression and the Mann-Kendall test. However, this global increase is not uniformly distributed among lower-level constructions. Indeed, the analysis of the Relative Family Type Frequency (RFTF) showed that the dissemination of ATAP compounds in the latter half of the 20th century was predominantly sustained by the expansion of existing N2-based families rather than the emergence of new ones. This indicates that pattern growth is due to the creation of new types within established families rather than the emergence of new families.

Furthermore, the analysis highlighted that the emergence and development of N1-based families (e.g., *città*-N 'city'-N, *donna*-N 'woman'-N, etc.) play a crucial role in the internal evolution of the ATAP pattern. Our data shows that while N1-based families were virtually absent before the 1950s, they experienced significant growth thereafter, eventually surpassing N2-based families in number by the 1990s. This suggests that the establishment and expansion of N1-based families contributed to the overall vitality and diversification of the ATAP pattern. Our findings also hint at the potential for new N1-based families to cross the boundaries of established N2-based families, giving rise to new modifiers and, potentially, new N2-based families. This interaction between N1- and N2-based families points to a more complex process of linguistic innovation and generalizations than would emerge from observing – as done so far – only or mainly N2-based families. This suggests the need to apply measures such as coverage to both types of families.

Future research should delve deeper into the qualitative features of individual families and explore other potential types of generalizations that can lead to the creation of new lower-level schemas, such as sense-based generalizations. Such investigations would enhance our understanding of the mechanisms driving the evolution and productivity of compound constructions.

Acknowledgements

This paper represents a pilot study of 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).

References

Baayen, R. H. (2009). Corpus linguistics in morphology: Morphological productivity. In A. Lüdeling & M. Kytö (Eds.), *Handbooks of Linguistics and Communication Science* (pp. 899-919). Mouton de Gruyter. https://doi.org/10.1515/9783110213881.2.899

Baroni, M., Guevara, E., & Pirrelli, V. (2009). Sulla tipologia dei composti N + N in italiano: principi categoriali ed evidenza distribuzionale a confronto. In G. Ferrari, R. Benatti & M. Mosca (Eds.), *Linguistica e modelli tecnologici di ricerca: atti del XL Congresso internazionale di studi della Società di linguistica italiana (SLI): Vercelli, 21-23 settembre 2006* (pp. 73-96). Bulzoni Editore.

Bauer, L. (2017). *Compounds and compounding*. Cambridge University Press.

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

Booij, G. (2010). *Construction morphology*. Oxford University Press.

Booij, G. (2016). Construction Morphology. In A. Hippisley & G. Stump (Eds.), *The Cambridge Handbook of Morphology* (pp. 424-448). Cambridge University Press. https://doi.org/10.1017/9781139814720.016

Diessel, H. (2023). *The Constructicon. Taxonomies and Networks*. Cambridge University Press.

Goldberg, A. E. (2019). *Explain me this: creativity, competition, and the partial productivity of constructions*. Princeton University Press.

Grandi, N. (2009). When Morphology 'Feeds' Syntax: Remarks on Noun > Adjective Conversion in Italian Appositive Compounds. In F. Montermini, G. Boyé & J. Tseng (Eds.), *Selected Proceedings of the 6th Décembrettes* (pp. 112-124). Cascadilla Proceedings Project.

Grandi, N., Nissim, M., & Tamburini, F. (2011). Noun-clad adjectives. On the adjectival status of non-head constituents of Italian attributive compounds. *Lingue e Linguaggio*, *10*(1). https://doi.org/10.1418/34543

Hartmann, S. (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

Herman, O., & Kovář, V. (2013). Methods for Detection of Word Usage over Time. In A. Horák & P. Rychlý (Eds.), *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2013* (pp. 79-85). Tribun EU.

Hilpert, M. (2015). From *hand-carved* to *computer-based*: Noun-participle compounding and the upward strengthening hypothesis. *Cognitive Linguistics*, *26*(1). https://doi.org/10.1515/cog-2014-0001

Hilpert, M. (2021). *Ten lectures on diachronic construction grammar*. Brill.

Hilpert, M., & Gries, S. (2009). Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, *24*(4). https://doi.org/10.1093/llc/fqn012

Hussain, Md., & Mahmud, I. (2019). pyMannKendall: a python package for non-parametric Mann Kendall family of trend tests. *Journal of Open Source Software*, *39*(4). https://doi.org/10.21105/joss.01556

Hüning, M. (2019). Morphological Theory and Diachronic Change. In J. Audring & F. Masini (Eds.), *The Oxford Handbook of Morphological Theory* (pp. 475-492). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199668984.013.28

Jackendoff, R., & Audring, J. (2020). *The texture of the lexicon: relational morphology and the parallel architecture*. Oxford University Press.

Micheli, M. S. (2020). *Composizione italiana in diacronia: Le parole composte dell'italiano nel quadro della Morfologia delle Costruzioni*. De Gruyter. https://doi.org/10.1515/9783110652161

Radimský, J. (2015). *Noun+Noun compounds in Italian: a corpus-based study*. Jihočeská univerzita v Českých Budějovicích.

Radimský, J. (2016). I composti N-N attributivi nel corpus ItWac. In A. Elia, C. Iacobini & M. Voghera (Eds.), *Livelli di analisi e fenomeni di interfaccia: Atti del XLVII Congresso internazionale di studi della Società di linguistica italiana* (pp. 189-204). Bulzoni editore.

Radimský, J. (2023). Tracing back the history of Italian Attributive-Appositive Noun+ Noun compounds. *Linguistica Pragensia*, *33*(2). http://dx.doi.org/10.14712/18059635.2023.2.3

Rainer, F. (2016). Italian. In P.O. Müller, I. Ohnheiser, S. Olsen & F. Rainer (Eds.), *Word Formation. An International Handbook of the Languages of Europe* (pp. 2712-2731). De Gruyter Mouton. https://doi.org/10.1515/9783110379082-017

Rainer, F. (2021). Compounding: From Latin to Romance. *Oxford Research Encyclopedia of Linguistics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.691

Scalise, S., & Bisetto, A. (2009). The Classification of Compounds. In R. Lieber & P. Štekauer (Eds.), *The Oxford Handbook of Compounding* (pp. 34-53). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199695720.013.0003

Traugott, E. C., & Trousdale, G. (2013). *Constructionalization and constructional changes*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199679898.001.0001