BILINGUAL PUBLISHING GROUP

Forum for Linguistic Studies

https://journals.bilpubgroup.com/index.php/fls

ARTICLE

Fixedness as a Criterion for the Idiomaticity of Nominal Compounds

Xiaoqin Hu

Beijing Language and Culture University, 15, Xue Yuan Road, Hai Dian District, Beijing 100083, China

ABSTRACT

Many studies have argued that idiomaticity judgment should be based on multiple aspects, including both semantic and formal fixedness, also known as inflexibility. However, the results delimiting idiomaticity are far from scarce. Although the validity of the notion of fixedness has been challenged, the fixedness of fixed expressions and collocations are still considered of the same nature, differing only in degree. The difference in fixedness between fixed expressions and collocations is still ambiguous, creating several barriers to further study of idiomaticity. This article aims to define the idiomaticity of nominal compounds by reviewing the fixedness of nominal compounds and that of collocations. The research was conducted on a finite lexical list based on French lexical resources. The parameters for fixedness were defined and annotated on the basis of the corpus provided in Sketch Engine, and the data were processed with the support of statistical models. The results reveal the difference in fixedness between nominal compounds and collocations, challenging the argument that collocations are constructions that stand on the continuum of free to fixed phrases. This study proposes a new perspective on fixedness and provides unique insight into the idiomaticity of nominal compounds.

Keywords: Idiomaticity; Linear Regression; Distribution of Fixedness Degree; Semantics of Nominal Compounds; Paradigmatical Inheritance

*CORRESPONDING AUTHOR:

Xiaoqin Hu, Beijing Language and Culture University, 15, Xue Yuan Road, Hai Dian District, Beijing 100083, China; Email: franhuxqin@hotmail.fr

ARTICLE INFO

Received: 2 August 2024 | Revised: 20 August 2024 | Accepted: 21 August 2024 | Published Online: 11 November 2024
DOI: https://doi.org/10.30564/fls.v6i5.6993

CITATION

Hu, X., 2024. Fixedness as a Criterion for the Idiomaticity of Nominal Compounds. Forum for Linguistic Studies. 6(5): 364–382. DOI: https://doi.org/10.30564/fls.v6i5.6993

1. Introduction

Nominal compounds, or compound nouns, belong to a specific category of compounds and are defined as sequences combined with multiple lexemes, displaying lexical, syntactic, semantic, pragmatic, and/or statistical idiomaticity[1]. Nominal compounds have long been considered fixed expressions that can be discriminated in terms of the degree of fixedness, in both semantic and formal fixedness[2–6]. However, the semantics of compounds are not always opaque, and they share many properties (co-occurrence, semantic transparency/opaqueness, paradigmatical restrictions) with other complex expressions, especially collocations. Statistical criteria essentially define collocations: they are words that tend to appear together more often than would be expected by chance[7–11]. A prototypical collocation is a phrase constructed A+B or B+A, where lexical Item A is the base of collocation selected by the speaker, and B is the collocate of A selected according to the combinatorial properties of A[11, 12]. The boundaries between compounds, collocations, and other complex expressions are often fuzzy; they are seen as somewhat overlapping on a continuum of idiomaticity, and collocations are placed on a free–fixed phrase continuum[13–21].

However, some studies have questioned the validity of using the notion of fixedness to judge idiomaticity[22–25]. These studies were based on observations of certain syntactic behaviors in specific compound nouns, which cannot adequately reflect the overall degree of fixedness of a compound noun or account for the distribution of fixedness across compound nouns and collocations. The degree of fixedness should be measured by considering multiple parameters in terms of both semantics and forms. Nevertheless, no comprehensive quantitative study on fixedness has reviewed its validity in the judgment of idiomaticity. It remains unclear how fixedness is distributed across compound nouns and collocations and whether the degree of fixedness can distinctly differentiate the two (and if so, what the threshold might be—at what point does an expression cease to be a compound noun and become a collocation?). Additionally, the fixedness of fixed expressions and that of collocations has often been considered to be of the same nature, differing only in degree[2, 18, 26–33]. No study has explored the differences in these fixedness parameters—such as compositionality, substitutability, and morphosyntactic transformability—between nominal compounds and collocations. It remains uncertain whether the difference in fixedness between the two is merely one of degree or if it is a fundamental difference in nature. As a result, it is unclear whether idiomaticity can be defined as a scalar value on a continuum.

The ambiguity surrounding the difference in fixedness between nominal compounds and collocations creates several barriers to further study idiomaticity. This study aims to redefine the idiomaticity of nominal compounds by reviewing the validity of the degree of fixedness in distinguishing compound nouns from collocations and then revealing the difference in fixedness between nominal compounds and collocations on the basis of a quantitative analysis. This will provide accurate data and theoretical insights for future research on idiomaticity. This research looks to the corpus for elaborating the annotated data and to the statistical models for computing. The study specifically focuses on adjective + noun (Adj + N) and noun + adjective (N + Adj) nominal compounds and collocations, which are often confused in idiomaticity judgments. The article begins with a literature review of the idiomaticity criterion. The experimental results are subsequently described and analyzed after the research methods used are addressed. Finally, the idiomaticity of nominal compounds is defined by investigating the fixedness of nominal compounds and collocations in semantic and syntactic terms.

2. Literature Review

Early studies inspired by generative linguists equated idiomaticity with noncompositionality[34–36], which means that the meanings of the parts do not add up to the meaning of the whole[37]. Noncompositionality, sometimes referred to as decomposability, analyzability, or transparency, is a gradient[38]. Libben[39] introduced a morpheme-based model for processing the morphology of compounds and considered that transparent compounds could be paraphrased with the pattern "compound (N1N2) being N2," e.g., "a blueberry is a berry." He proposed four degrees of transparency: transparent–transparent (TT), where both constituents are transparently related to the compound meaning (e.g., "blueberry"); transparent–opaque (TO), where only the first constituent is transparent, whereas the second is opaque (e.g., "shoehorn"); opaque–transparent (OT), where the first con-

stituent is opaque, and the second is transparent, such as strawberry. Finally, both constituents are opaque, yielding opaque–opaque (OO) combinations such as bighorn ('a species of sheep'). Nevertheless, Sandra [40] argued that compositionality should be distinguished from transparency. Transparency refers to the relationship between compound and constituent meanings, whereas compositionality refers to the possibility of predicting the overall meaning from the constituent meanings. There may be a semantic relationship between a transparent compound and its constituents, but the overall meaning of the compound is often greater than that of its constituents [41]. Reddy et al. [42] considered both transparency and compositionality as questions of meaning predictability, and the semantic relationship between compound constituents has a neglectable function in whole-meaning prediction. However, noncompositionality cannot serve as a criterion of idiomaticity since both compositionality and noncompositionality occur in idioms and collocations [26], as illustrated in opaque collocations such as *peur bleue* (fear + blue, 'great fear'), *colère froide* (angry + cool, 'suppressed anger'), and *grand café* (big + coffee, 'coffee of high quality'), where the interpretation of the collocate is neither predictable nor transparent, and the collocations are noncompositional [43]. Nunberg et al. [38] argued that noncompositionality fails to recognize several important dimensions of idiomaticity, including figuration and conventionality, unpredictability of meanings on the basis of knowledge of the independent conventions of their constituent parts.

Substitutability, sometimes called collocability, was considered a criterion for idiomaticity in several early studies. Substitutability investigates whether parts of a phrase can be substituted for others without losing idiomatic interpretation and is connected to the question of productivity since substitution leads to the production of new combinations. However, (non)productivity occurs both in compounds and collocations. For example, in the collocation *ferme espérance* (firm + hope), the adjective *ferme* ('firm') cannot be substituted by its synonym fort ('strong'): **forte espérance* ('strong hope') [44], and in the compound *emballage maigre* (packaging + thin, 'skinny packaging'), the adjective maigre cannot be substituted by its synonym *mince* ('thin'). Vašků [33] argued that (non)productivity should be considered a feature of idiomaticity since "speakers must store and retrieve a formation based upon unproductive processes as one unit without segmenting it into its components," whereas (non)productivity should not be a criterion of idiomaticity in English because it does not represent an idiosyncratic phenomenon for two main reasons. First, an overwhelming majority of unproductively formed English words are of origin Latin or Greek loanwords and words formed within English by analogy with these Latin and Greek words, and they exist within a system with certain rules that are not used (normally) to form new lexemes; second, the production of these words is approached in different ways from that of word combinations. It is widely acknowledged that neither noncompositionality nor nonsubstitutability alone can be equated with idiomaticity.

The view that idiomaticity is a complex concept associated with various forms of formal transformability, also called flexibility, has been supported by many researchers [2, 28, 45–49]. Gross [28] introduced the notion of fixedness and argued that compounds are fixed to varying degrees. He attempted to subject compounds to syntactic transformations, such as deletion or replacement of word sequences, lexical and syntactic substitutions with or without semantic shift, and the addition of information for measuring their degree of fixedness. It has been argued that semantic noncompositionality, paradigmatical constraints (substitutability), and morphosyntactic constraints (transformability) are three main features that together characterize fixedness [3, 4, 28]. Anscombre [50] grouped the criteria of fixedness into three categories inspired by Gross's work: referential fixedness, which is the absence of reference to specific entities and, hence, the difficulty of modifying the determinants in a fixed phrase; transformational fixedness, which is the impossibility of passivation, of the pronominal recovery of the constitutive elements, and of modifying the order of the constituents; and finally, semantic fixedness, particularly concerning the synonymous paradigm and the noncompositionality of meaning. From a usage-based approach, Wulff [18] claimed that idiomaticity judgment relies on multiple features, including compositionality, formal fixedness, and flexibility, such as syntactical flexibility, which is tested if (or to what extent) specific syntactic transformations are possible, lexico-syntactic flexibility (primarily defined as lexical insertion), and morphological flexibility regarding morphological variations (e.g., tense, aspect, and negation). Idiomaticity is considered a scalar, complex concept that captures the id-

iosyncrasies of all multiword units, which can be classified on a collocation–idiom continuum [5, 13–17, 19–21, 26].

However, Sliwa [23] declared that analyzing components from the point of view of their fixedness fails to distinguish compound nouns from other expressions. Considering only the syntactical side risks of confusing a collocation with an endocentric compound noun, which has a head transmitting its semantic and syntactic properties to the compound, e.g., *terminal pétrolier* (terminal + petrolic, 'oil terminal'), because the compound nouns are often subject to erasure transformation [51], especially in the case of terms (endocentric compound nouns), which are characterized by the successive addition of modifiers that specify the referenced concept. Examples are the complex term *détecteur d'horizon terrestre* (detector + horizon + terrestrial, 'horizon detector') and the reduced term *détecteur terrestre* (detector + terrestrial, 'horizon detector'). She argued that a compound noun is a designated entity with a preestablished reference, which does not distribute its meanings among its components [38]. Only reference distinguishes compound nouns from other expressions. Chakiri [24] questioned the fixedness in terms of semantics. He concluded that fixed expressions consisted of translating an object or an event by a sequence of lexical units, which assumed a link between a signifier and a signified, and this link was established by denomination. Goes [25] believed that collocations should be considered devoid of fixedness, as most impossibility of transformation in collocations is due to the nature of adjectives themselves, called pseudo-fixedness. By examining the syntactic behaviors of the adjectives in the syntagms N + Adj, he argued that nominal compounds were different from collocations by their denomination and the adjectives in collocations should retain their function of modifying. However, these studies that question fixedness have typically been based on observations of certain factors in compound nouns or collocations, which cannot reflect the overall degree of fixedness of one compound noun nor do they account for the distribution of fixedness degree across compound nouns and collocations.

To summarize, fixedness is an important notion for considering idiomaticity comprehensively, but two issues remain unresolved in the research on fixedness. First, the degree of fixedness should be measured via multiple parameters in terms of both semantics and forms. However, current research [22–25] that questions fixedness has typically been based on observations of some syntactic transformations in certain compound nouns, which cannot provide an accurate and comprehensive assessment of the degree of fixedness of different expressions. Second, there has been no discussion on whether the mechanisms of transformability and substitutability in compound nouns and collocations are different. If fixedness is not only a question of degree, can idiomaticity be defined as a scalar value on a continuum?

In the following sections, a detailed quantitative analysis of the degree of fixedness of compound nouns and collocations is provided and the fixedness-related linguistic behaviors (compositionality, substitutability, and morphosyntactic transformability) exhibited by both are compared using a linear regression model, which provides empirical evidence for revealing the differences in substitutability and transformability between compound nouns and collocations.

3. Methodology

3.1. Resources

The study data were established based on the French lexical resources DELA and LAF. DELA includes 102,073 simple entries (102,073 different lemmas) and 83,604 compound entries (83,604 different lemmas). DELA is described by morphology and inflection. Each entry is registered with its canonical form (lemma) associated with grammatical and semantic codes. The grammatical code indicates the grammatical category of the entry and the grammar, which allows for the generation of different inflected forms from the canonic form. The semantic code indicates the semantic category of the entry, such as HumColl (collective nouns of people), Conc (concrete noun), Abst (abstract) and z1 (general language).[1] In DELA, compound words are distinguished from simple words based on a formal definition: a simple word is a sequence built on the alphabet, and a compound word is a sequence of simple words. The compounds without a blank space or punctuation mark separating the constituents were recorded as simple entries; for example, *deltaplane* is registered as a simple entry, while *delta-plane* is taken as a compound entry. Therefore, many neoclassical compounds,

[1]The taxonomy of the semantic codes used in DELA is available at https://unitexgramlab.org/releases/3.1/man/Unitex-GramLab-3.1-usermanual-en.pdf

such as *homicide* ('homicide') and *cyclomoteur* ('moped'), as well as pseudo-compounds, such as *logiciel* ('software') and *électroménager* ('family appliances') in French, were registered as simple entries in DELA. The compound words were distinguished from free phrases based on the principle that a sequence of simple words is compound if at least one of its syntactic, distributional, or semantic properties cannot be deduced from the properties of its constituents. The compounds were rather distinguished based on the judgment of French native speakers according to the proposed principle. There was no systematic annotation or comprehensive measurement of the degree of fixedness for distinguishing compounds.

LAF is a general public version of the Explanatory and Combinatorial Dictionary of Mel'čuk et al.[52], providing a fine lexicographical description of lexical relationships, including semantic derivations (e.g., synonymy, antonymy, and nominalization) and collocations[12, 53, 54]. All types of collocates, such as adjectival collocates, verbal collocates, and nominal collocates, were collected. LAF includes a total of 20,000 semantic derivations and collocations of French. The collocations in LAF are collected based on the definition of collocations given by Hausmann[11, 55] and Mel'čuk[52]; in other words, collocations are linguistic expressions in which the base keeps its primary meaning while the other element, called collocate, which can be transparent or not, is in an irregular and/or constrained manner. By examining six main collocation dictionaries in English and French based on the size of the nomenclature and the proposed syntactic and semantic treatments, Tutin[56] argued that the collocation precision and coverage of LAF are higher than other collocation resources or language dictionaries in French, the access to collocations is systematic, and the linguistic information associated with each entry is fine and comprehensive.

3.2. Data Collection

The study presented in this article focuses on N + Adj and Adj + N nominal compounds as they are most easily confused with collocations. Both morphological compounds, formed by combining two lexemes of a language without syntactic phenomenon[57–61], and syntactical compounds, lexicalized from syntactical constructions, are considered as nominal compounds in our study. Three types of morphological compounds are distinguished: a) neoclassical compounds, that are formed along two bases of Greek or Latin origin and are not syntactically autonomous in French, connected by a linking element like *i*, *o*, and *a*[62, 63], such as [*micr**o***][*céphale*] (small + head, 'microcephalic') and [*még**a***][*lithe*] (large + stone, 'megalith'); b) native compounds, composed of two lexemes of current French lexicon without linking element, such as *wagon-fumeur* (car–smoker, 'smoking car') and *homme-grenouille* (human–frog, 'experienced diver'); c) and concealed compounds[64] or also called pseudo-compounds[65] formed of at least one truncated French lexeme such as [*afro*]-[*brésilien*] (African + Brazilian, 'Afro-Brazilian') and [*ludi*][*ciel*] (game + software, 'game software')[66]. The neoclassical compounds and pseudo-compounds of N + Adj/Adj + N are excluded from the data for the following reason: words of Latin or Greek origin, loanwords exist within a system with certain rules which are not used (normally) anymore to form new lexemes, and their production is approached in a different manner from that of word combinations[33].

DELA includes 29,199 N + Adj compounds (29,199 different lemmas) and 2,121 Adj + N compounds (2,121 different lemmas). The entries are registered in alphabetic order. For N + Adj compounds, the paradigmatically related compounds with the same head noun are listed adjacent to one another, for example, *crème allégée*, *crème bronzante*, *crème brûlée*, *crème catalane*, and *crème dermique*, while for Adj + N compounds, the compounds with the same collocate are listed adjacent to one another, such as *arrière ban*, *arrière bief*, *arrière boutique*, *arrière bras*, *arrière cabinet*, and *arrière cavité*. In total, 6,464 different head nouns were found. The nouns assigned different semantic codes are distributed throughout the list of N + Adj/Adj + N entries. Thus, 500 N + Adj/Adj + N compounds were extracted using a computational program that read the entries and rewrote 500 entries randomly in the data list. These 500 entries cover compounds of different head nouns and semantic codes. Compositional, partially noncompositional, and noncompositional compounds were included. Subsequently, the head nouns of the extracted compounds were taken as the bases for extracting 500 Adj + N/N + Adj collocations randomly from LAF. However, it was observed that several units were collected in both DELA and LAF due to disagreements among linguists regarding the classification of nominal compounds and collocations. Therefore, the phrases that were registered in both

DELA and LAF were removed to obtain two lists of phrases: a list of nominal compounds that are rarely considered collocations and a list of collocations that are rarely considered nominal compounds. Then, after removing the overlapping part, the collected data were examined to see if they can be clearly distinguished based on their degree of fixedness; if not, how they distribute according to their fixedness values. Finally, 47 overlapping phrases were removed from the data list, and then 490 nominal compounds and 462 collocations were obtained as study data.

3.3. Corpus

The annotation and analysis were executed based on the French Web Corpus provided in Sketch Engine,[2] a platform for exploring corpora in more than 90 languages. Sketch Engine allows to generate concordances (i.e., to display the search word or phrase in context), a frequency list of all the words in a corpus and an n-grams list based on co-occurrence calculation, and to extract the keywords of a corpus. The annotation was conducted based on its function of concordance to verify in context whether a transformed phrase still maintains the idiomatic interpretation. The French Web Corpus is a set of corpora comprising texts collected from the Internet and consists of 20.9 billion words. The French Web Corpus contains many varieties of the French language—European, Canadian, and African French—and the texts were collected between 2019 and 2020 to enter the corpus of Sketch Engine.

3.4. Annotation of Data

Four parameters of fixedness were defined and annotated: compositionality, substitutability, morphological restrictions, and syntactical transformability. The annotation was conducted manually by three linguists who are native speakers. The result was validated only if two-thirds, or all, of the linguists concurred with the annotation of each phrase. Compositionality annotation, substitutability annotation, and syntactical transformability annotation were conducted based on the proposed principles (as presented in the following), while morphological restriction annotation was performed using the morphological and inflectional information provided by DELA. The pseudo-fixedness, resulting from the morphosyntactic properties of adjectives [25, 38], was taken into account in annotation. As the majority (88.87%) of adjectives in nominal compounds and collocations present the pseudo-fixedness in at least one of the defined syntactical transformations, the score of 0.5 was adopted for differentiating the pseudo-fixedness from the true fixedness, assigned the score of 0, and the nonfixedness, assigned the score of 1, during data annotation.

3.4.1. Compositionality Measurement

In this work, compositionality is considered to predict the whole meaning from the constituent meanings. The measurement of compositionality is based on the morpheme-based model of Libben [39] but also considers the role of semantic relations in meaning prediction. The constituent meanings and their sum meanings are compared with the whole meaning of each construction. It is assumed that the opacity or transparency of the head noun of an N + Adj/Adj + N nominal compound has a more significant impact on the meaning prediction of the whole. Five degrees of compositionality are identified:

- CAT0: if neither of the constituent meanings is transparently related to the meaning of NA, NA is considered to be noncompositional, and its degree of compositionality is noted as "0," for example, *tapis franc* (carpet + free) refers to the den where bandits gather;
- CAT1: if the meaning of the head noun N is not transparently related to the whole meaning, or, in other words, the meaning of NA cannot be paraphrased by "NA is N," but the adjective is entailed in the whole meaning, the degree of compositionality is noted as "1", for example, *perle rare* (pearl + rare, "rare-gem") refers to somebody or something rare and valuable due to their or its quality;
- CAT2: if only the meaning of the head noun N is transparently related to the meaning of NA, the degree of compositionality is noted as "2," for example, *humour noir* (humor + black, 'black humor');
- CAT3: if both constituent meanings of NA are transparently related to its whole meaning but the semantic relation between N and A is not predictable, or, in other words, the construction meaning cannot be paraphrased by "NA is N which is A," although both constituent meanings are entailed in the whole meaning, its degree of compositionality is set to "3," for

[2] https://www.sketchengine.eu

example, *abonnés absents* (subscriber + absent, 'absent subscriber') refers to absence all the time when required;

- CAT4: if both constituent meanings of the construction NA are transparently related to its whole meaning and the semantic relation between N and A is predictable, or, in other words, the whole meaning can be paraphrased by "NA is the N which is A," this construction is considered compositional, and its degree of compositionality is evaluated at "4" such as *roquette antichar* (rocket + antitank, 'antitank rocket') and *onde monochromatique* (wave + monochromatic, 'monochromatic wave').

The scores for measuring compositionality range from 0 to 4. A more compositional compound yields a higher score. The same compositionality measurement was applied to Adj + N phrases.

3.4.2. Substitutability

Substitutability is about knowing whether parts of nominal compounds and collocations can be replaced by synonyms or antonyms without losing their idiomatic interpretation. For each nominal compound or collocation NiAj/AjNi in our study data,

- all other compounds or collocations with the same head noun Ni were first obtained from DELA or LAF;
- then, each obtained compound or collocation NiAk was compared with NiAj to determine whether Ak is a synonym or an antonym of Aj; if so, and NiAk keeps the same (if synonym) or reverse (if antonym) meaning as NiAj, it is considered that Aj in NiAj can be replaced by synonyms or antonyms without losing the idiomatic interpretation;
- the meaning of NiAk was judged by annotators based on the French Web Corpus; for certain phrases NiAk whose meanings vary with contexts, only if it can have the same or reverse meaning with NiAj, it is considered that NiAk keeps the same (if synonym) or reverse (if antonym) meaning with NiAj after substitution; for example, *nominale* in the nominal compound *liste nominale* (list + of name, 'name list') is the synonym of nominative in the nominal compound *liste nominative* ('name list');
- the total number of possible substitutions was counted as a substitutability value.

3.4.3. Morphological Restrictions

It is believed that a free phrase has a variation in word morphology between singular and plural in French, while a fixed phrase may receive morphological restrictions, such as *abonnés absents* (subscriber + absent, 'absent subscriber'), *eaux usées* (water + worn out, 'sewage'), and *blouses blanches* (coat + white, 'medical staff'), which must be plural, while some phrases have only singular forms, such as *devoir conjugal* (duty + conjugal, 'conjugal duty') and *instant présent* (instant + present, 'present instant'). It is emphasized that these morphological restrictions affect the whole construction. The head nouns *abonné* ('subscriber'), *eau* ('water'), and *blouse* ('coat') have singular forms and the head nouns *devoir* ('duty') as well as *instant* ('instant') have plural forms. Nevertheless, some simple nouns have only plural forms in French, such as *funérailles* ('funeral') and *obsèques* ('obsequies'), while some only have singular forms such as *beauté*. The morphological restrictions caused by the morphological properties of the head nouns were not considered as true morphological restrictions. For each N + Adj/Adj + N phrase,

- if N has both singular and plural forms but N + Adj/Adj + N phrase only have singular or plural form, the phrase is considered to receive true morphological restrictions and was assigned the score "0";
- if N has only singular or plural form, leading to the pseudo-fixedness in morphology, the phrase was assigned the score "0.5";
- otherwise, the phrase was assigned "1" as the value of its morphological parameter.

3.4.4. Syntactical Transformability

The syntactical transformability annotation was conducted based on the work of Gross [2]. Among the parameters defined by Gross [2], predicativity and ellipsis of an adjective correlate with the question of (non)compositionality, and the fixedness of the first term correlates with the question of substitutability. Therefore, only four transformations were taken into account: a) nominalization of an adjective; b) replacement of an adjective by "de + N" as equivalence; c) insertion of a degree adverb; d) and juxtaposition with an adjective.

Nominalization of an adjective refers to the transformation of the nominalization of the adjective without losing

the original meaning of the phrase. Example (1) shows that the nominalization of the adjective *noir* ('black') is accepted for the noun phrase *robe noire,* whereas it is refused for the nominal compound *abeille noire*.

(1) a. *robe* *noire*
dress black
'black dress'
noirceur *de* *la* *robe*
blackness of the dress
'blackness of the dress'

b. *abeille* *noire*
bee black
'Carniolan honeybee'
**noirceur* *de* *l'* *abeille*
blackness of the bee
'blackness of the bee'

(2) *festival* *musical*
festival musical
festival *de* *musique*
festival *of* *music*
'music festival'

The replacement of an adjective by "de + N" as equivalence aims to replace the adjective with a complementary noun introduced by the preposition *de*, as shown in example (2). Insertion of a degree adverb signifies the possibility of modifying the adjective only if it is gradable with the sequences; cf. Example (3), and juxtaposition with another adjective is effectuated with the support of conjunctions such as *et* ("and") and *mais* ("but"); cf. Example (4).

(3) *un* *livre* *difficile*
a book difficult
'a difficult book'
un *livre* *très* *difficile*
a book very difficult
'a very difficult book'

(4) *une* *fille* *intelligente*
an girl intelligent
'an intelligent girl'
une *fille* *intelligente* *mais* *timide*
a girl intelligent but shy
'a girl intelligent but shy'

The transformability annotation was based on the French Web Corpus in Sketch Engine. For each N + Adj/Adj + N phrase,

- if the transformation corresponding to idiomatic usage of the phrase can be found in the corpus, a score of "1" is assigned; the transformation related to nonidiomatic usage is not considered, although it can be found in the corpus; for example, the nominal compound *disparition totale* (disappearance + total) refers to , while this phrase could be a free combination in certain contexts, for example, *disparition totale du bouton de fièvre*, which means "the fever blisters disappear completely";
- if the transformation is restricted due to the syntactic–semantic property of the adjective that constitutes the construction, it is considered to be pseudo-fixedness and the score "0.5" is assigned;
- otherwise, the score is "0" only if the transformational restriction is not due to the pseudo-fixedness;
- finally, the sum value of the four parameters was added in equal parts as the transformability value of each phrase.

3.5. Data Processing

The values of each parameter were nominalized by the normalization technique Min–Max Scaling which aims to rescale data, ensuring it falls within range 0–1, by subtracting the minimum and dividing by (max–min), as shown in the following formula:

$$Z_i = \frac{x_i - min\,(x)}{max\,(x) - min\,(x)} \quad (1)$$

where Z_i is the *i*th normalized value in the dataset, and x_i is the *i*th value in the dataset. Then, the value distribution of each parameter was first demonstrated by histograms and scatter plots, including the distribution of compositionality and morphological restriction values, the distribution of pseudo-fixedness, as well as the distribution of substitutability and transformability values. The value distribution of nominal compounds was compared with that of collocations attempting to outline the boundaries between nominal compounds and collocations.

Subsequently, the annotated data were represented by a unison plot using the software R. The unison graph in R is focused on the question of classification. The graph plots screwed curves if they are similar, or unscrewed curves if they are different types. The unison graph is a graphical representation of the method proposed by Andrews[67]. The objective is to project a point in a multidimensional space onto the curve of a two-dimensional plane. For a variable of

p-dimensions, X_r is the *r*th observation.

$$X_r^T = (x_{r1}, x_{r2}, ..., x_{rp}) \quad (2)$$

The unison graph was calculated using the following formula:

$$f_r(t) = \frac{x_{r1}}{\sqrt{2}} + x_{r2}\sin t + x_{r3}\cos t + x_{r4}\sin 2t + x_{r5}\cos 2t + ..., -\pi \leq t \leq \pi \quad (3)$$

where *n* observations correspond to *n* curves. The application of the unison graph aims to show whether consideration of all the defined parameters allows discrimination of nominal compounds from collocations.

Finally, the annotated data were used to calculate the linearity between the parameters of nominal compounds and that between the parameters of collocations, which were then compared for exploring whether there exists some difference in fixedness between nominal compounds and collocations that could lead to the difference in linearity. The calculation of linearity aims to determine the correlation between these parameters. A linear regression model was applied and, using the ordinary least squares (OLS) method, calculated as follows:

$$Y_i = {}_0 + {}_1X_i + {}_i, i = 1, 2..., n, E({}_i) = 0, var({}_i) = 2, (i, 1, 2...n) \quad (4)$$

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i \quad (5)$$

The OLS method minimizes the sum of the square differences between the observed and predicted values regarding $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \widehat{\beta} - \widehat{\beta}_1 X_i\right)^2 \quad (6)$$

The vector β of the coefficients can be estimated by the following formula:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}$$
$$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X} \quad (7)$$

The linearity calculation was done with the support of R software. The format of the annotated data aligns with the data format used in R, where each row represents an entry and each column corresponds to a parameter.

4. Results

4.1. Distribution of Data

The distribution of compositionality values of nominal compounds and collocations is shown in **Figure 1**. It is evident that the compositionality values of collocations are mainly distributed in CAT4 and only two collocations were found to belong to CAT3 and CAT2. By contrast, the compositionality values of nominal compounds are distributed in all five categories (CAT0–4), while most of the nominal compounds in the elaborated data are compositional and belong to CAT4. The compounds that belong to CAT0 and CAT1 present the absence of head nouns are exocentric nominal compounds. **Figure 2** shows the value distribution of the morphological restrictions of nominal compounds and collocations. Only 15 nominal compounds (14 endocentric and 1 exocentric) from 490 (3.06%) and 2 collocations from 462 (0.43%) were subject to morphological restrictions. No morphological pseudo-fixedness was found in the study data. The proportion of phrases that receive morphological restrictions in exocentric nominal compounds is higher than in endocentric nominal compounds, and that proportion in nominal compounds is higher than in collocations.

Figures 3 and **4** show the distribution of pseudo-fixedness in nominal compounds and in collocations, respectively, and the distribution of pseudo-fixedness in endocentric nominal compounds and that in exocentric nominal compounds are compared in **Figures 5** and **6**. It is observed that pseudo-fixedness presents a large percentage in nominal compounds and collocations. The adjectives that are nongradable, nondenominal, or impossible to nominalize lead to the pseudo-fixedness in the syntactical transformation: insertion of a degree adverb (INSERT), nominalization (NM), or replacement of the adjective (REPL) respectively. The proportion of true fixedness in collocations is far smaller than in nominal compounds and the nonfixed phrases in collocations present a larger proportion than in nominal compounds. Nevertheless, true fixedness in exocentric nominal compounds presents a larger proportion than in endocentric nominal compounds, and the nonfixed phrases in exocentric nominal compounds present a far smaller proportion than in endocentric nominal compounds.

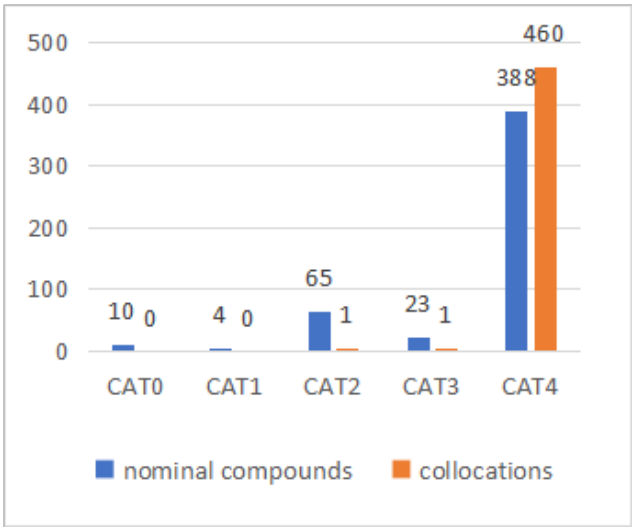

Figure 1. Compositionality values of nominal compounds and collocations.

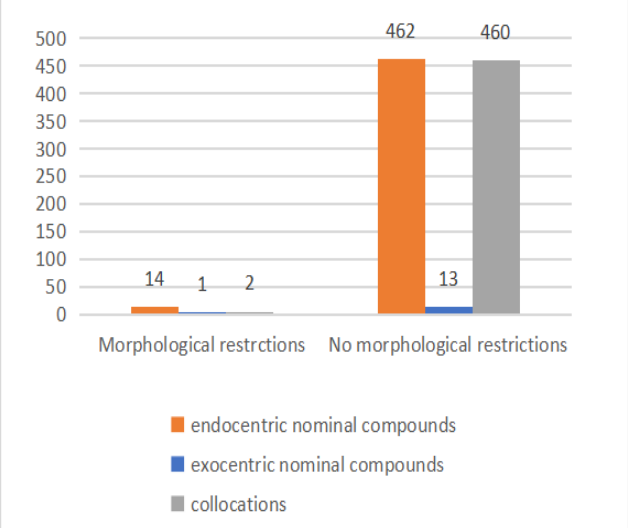

Figure 2. Morphological restrictions of nominal compounds and collocations.

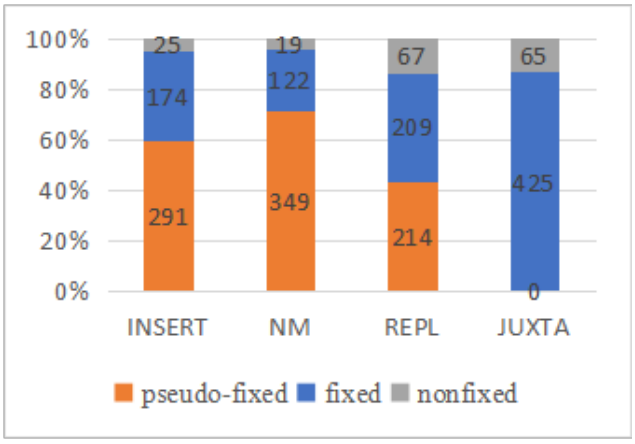

Figure 3. Proportion of pseudo-fixedness in nominal compounds.

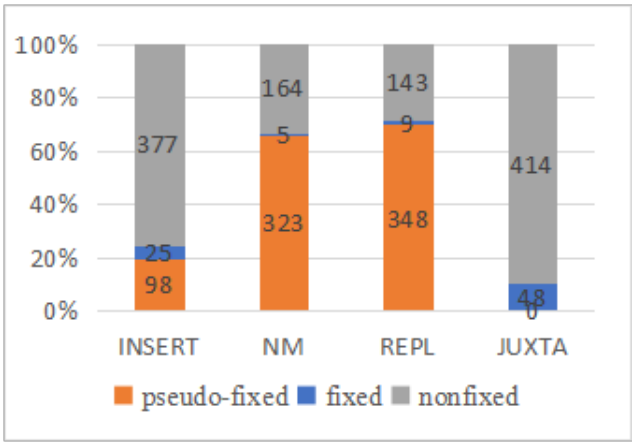

Figure 4. Proportion of pseudo-fixedness in collocations.

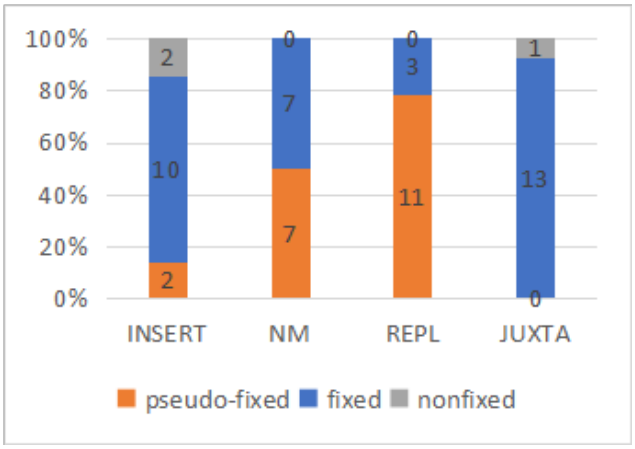

Figure 5. Proportion of pseudo-fixedness in exocentric nominal compounds.

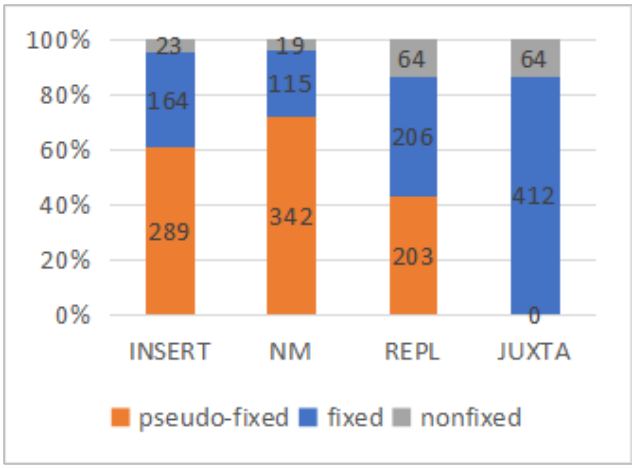

Figure 6. Proportion of pseudo-fixedness in endocentric nominal compounds.

The distribution of substitutability value and the distribution of transformability value of nominal compounds and collocations are represented by scatter plots. In **Figures 7–10**,

the x axis represents the serial number of each nominal compound or collocation, and the y axis represents the sum value of synonym and antonym substitutability (noted as SUMs) or the sum value of transformability parameters (noted as SUMt) of each phrase. It is observed that the substitutability values of all exocentric nominal compounds distribute below 0.1, while those of several endocentric nominal compounds are between 0.1 and 0.4. The substitutability values of nominal compounds are mainly distributed between 0 and 0.1, while the substitutability values of collocations above 0.1 present a far higher percentage. Nevertheless, there also exist a large number of collocations whose substitutability values are between 0 and 0.1 (**Figures 7** and **8**), or in other words, the substitutability values of nominal compounds and collocations mainly overlap between 0 and 0.1. The transformability values of nominal compounds and those of collocations mainly overlap between 0.4 and 0.6. The transformability values of exocentric nominal compounds distribute below 0.2, while those of nominal compounds are distributed across nearly the whole range 0–1. The transformability values of collocations are mainly distributed between 0.6 and 1, several are distributed below 0.6 (**Figures 9** and **10**). Finally, the values of all parameters, including compositionality, substitutability, morphological restrictions and syntactical transformability, were added up as fixedness degree values. A higher value yields a less fixedness degree of the phrase. The distribution of the values of the fixedness degree of nominal compounds and collocations is shown in **Figures 11** and **12**. It is shown that the values of the fixedness degree of collocations are mainly distributed between 0.6 and 1, while the values of the fixedness degree of most nominal compounds are distributed below 0.6. The fixedness degree values of exocentric nominal compounds range between 0 and 0.43, while those of endocentric nominal compounds distribute within a larger range between 0 and 0.87. Although the fixedness degree of most of the nominal compounds is higher than that of collocations, there is still overlapping that cannot distinguish nominal compounds from collocations.

In summary, assessing any separate defined parameter is never enough to distinguish nominal compounds from collocations. Compositionality and substitutability exist in both collocations and nominal compounds. Both nominal compounds and collocations have been subjected to substitutability restrictions and most of the nominal compounds or collocations do not receive morphological restrictions (**Figures 1** and **2**). The pseudo-fixedness presents similar proportions in nominal compounds and collocations, while the true fixedness presents a far larger proportion in nominal compounds than in collocations. Nevertheless, there still exists a quite percentage of phrases that accept at least one of the transformations in nominal compounds and collocations (**Figures 3–6**). Finally, the substitutability values (SUMs), the transformability values (SUMt), as well as the fixedness degree values of nominal compounds and collocations overlap (**Figures 7–10**). It is difficult to set a distinct boundary between nominal compounds and collocations (c.f. **Figures 11** and **12**).

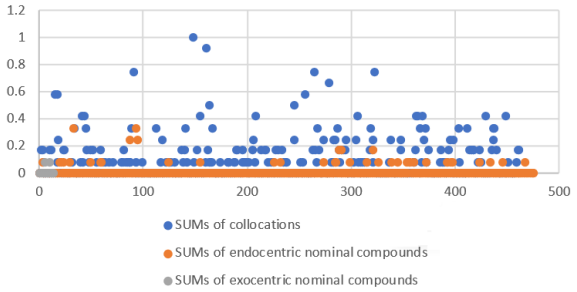

Figure 7. Distribution of substitutability values of nominal compounds and collocations.

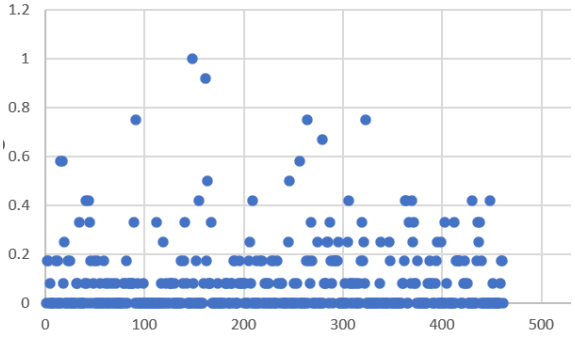

Figure 8. Distribution of substitutability values of collocations.

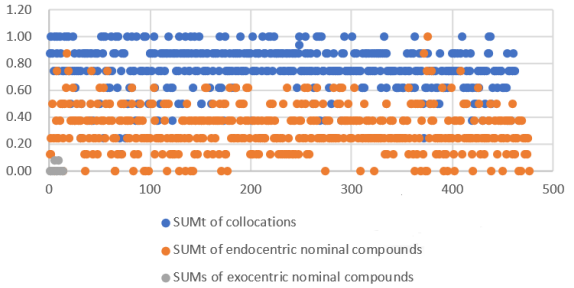

Figure 9. Distribution of transformability values of nominal compounds and collocations.

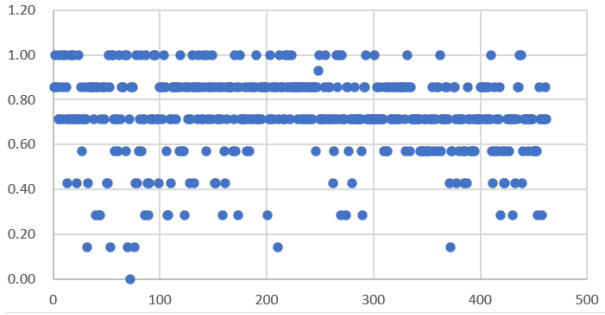

Figure 10. Distribution of transformability values of collocations.

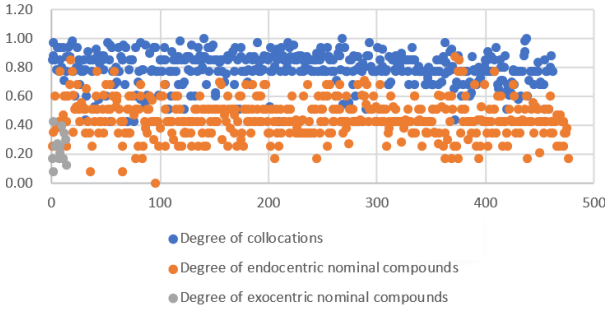

Figure 11. Fixedness degree of nominal compounds and collocations.

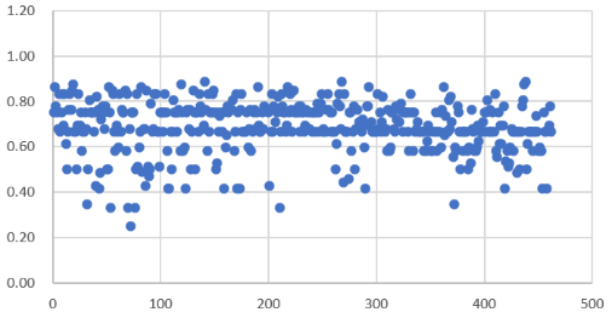

Figure 12. Fixedness degree of collocations.

4.2. Unison Graph

The unison graph focuses on the question of classification. In a unison graph, similar curves are screwed together, and different curves are twisted into different bundles. The computation of unison only considers the parameters of compositionality, substitutability, morphological restrictions and syntactical transformability. The sum value, SUMt, was removed from the data. The values of the parameters were taken as features, and the curves represent the classification based on these features. As shown in **Figure 13**, the curves of exocentric nominal compounds, endocentric nominal compounds and collocations are represented in different colors. To make the graph clearer, we separated the curves of endocentric nominal compounds and collocations in **Figure 14**.

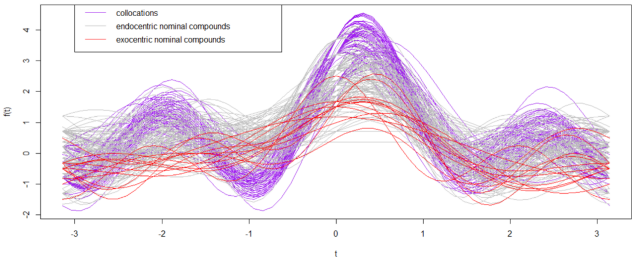

Figure 13. Unison graph of nominal compounds and collocations.

It is observed that the curves of exocentric nominal compounds twist more differently compared with those of endocentric nominal compounds and collocations. Part of the curves of endocentric nominal compounds twist very differently from those of collocations, while there are also some curves of endocentric nominal compounds that overlap those of collocations. In addition, in the curves of nominal compounds (endocentric and exocentric) and collocations, there are curves that are not screwed together. It is still difficult to distinguish nominal compounds from collocations, even considering several parameters in classification.

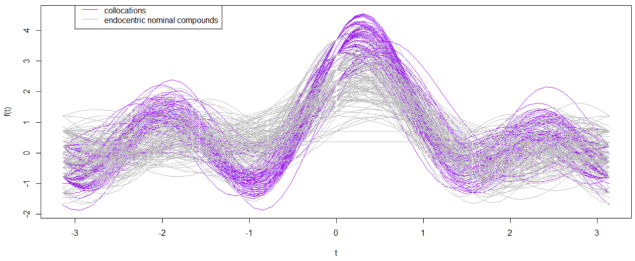

Figure 14. Unison graph of endocentric nominal compounds and collocations.

4.3. Linearity

Linearity between the parameters of the nominal compounds and collocations was calculated and compared to explore the difference in fixedness between these two types of phrases. The difference in linearity between the parameters of nominal compounds and collocations may allow us to reveal their different natures of fixedness in terms of compositionality, substitutability and transformability. As shown in **Tables 1** and **2**, each column shows the linear regression coefficients: estimate ($\widehat{\beta}_0$ and $\widehat{\beta}_1$), standard error, *t* value, R^2, and *p* value (statistically significant). Each row lists a pair of variables—the dependent variable (Y) and the independent variable (X): Y–X, for example, SUMt–C, SUMt is the

dependent variable (Y) that varies with C (the independent variable (X)). ***, **, and * indicate that the linearity is significant at the statistical levels of 0.001, 0.01, and 0.05. • indicates less significant, and no sign means not significant.

Table 1. Linearity between C, SUMs and SUMt of nominal compounds.

	Estimate	Std. Error	t-Value	R^2	Pr(>|t|)
(Intercept)	0.21464	0.03113	6.896	0.01999	1.67e-11 ***
SUMt-C	0.10639	0.03372	3.155		90.0017 **
(Intercept)	0.014273	0.006566	2.174	0.001424	0.0302 *
SUMs-C	−0.005934	0.007114	−0.834		0.4046
(Intercept)	0.308280	0.007866	39.189	0.001424	<2e-16 ***
SUMt-SUMs	0.180575	0.216467	0.834		0.405

Table 2. Linearity between SUMt and SUMs of collocations.

SUMt	Estimate	Std. Error	t-Value	R^2	Pr(>|t|)
(Intercept)	0.4837	0.2528	1.931	0.002633	0.0563 •
SUMt-C	0.2790	0.2532	1.102		0.2710
(Intercept)	−0.08576	0.21975	−0.390	0.001199	0.697
SUMs-C	0.16353	0.22007	0.743		0.458
(Intercept)	0.762834	0.008600	88.703	5.641e-05	<2e-16 ***
SUMt-SUMs	−0.008646	0.053676	−0.161		0.872

First, there is a significant linearity at 0.001 between the dependent variable SUMt and independent variable C of the nominal compounds, while the linearity between the SUMt and C of collocations is less significant. In other words, the transformability of a nominal compound is, to a great extent, linked to its compositionality, but the transformability of a collocation is not. This implies that the transformability of nominal compounds may operate through a different mechanism compared to that of collocations, which makes the transformability of nominal compounds sensitive to compositionality, whereas that of collocations is not. Second, for nominal compounds and collocations, there exists no significant linearity between the dependent variable SUMs and the independent variable C. However, for the regression calculation between SUMt and SUMs, both nominal compounds and collocations represent a low degree of fit between the regression model and actual data according to the R^2 value, and the performance of the regression model between the SUMt and the SUMs of collocations is even worse (**Tables 1** and **2**). This means that for only a small number of nominal compounds, their transformability is linked to substitutability, but for collocations, the number of those whose transformability is linked to substitutability presents a far smaller number, in other words, the transformability of collocations is barely linked to substitutability. This difference implies that the substitutability of nominal compounds may also be different from that of collocations. These differences in fixedness between nominal compounds and collocations are discussed in detail in the following section.

5. Discussion

The results indicate that both compound nouns and collocations exhibit substitutability and transformability, and neither individual nor combined parameters of fixedness can completely differentiate them. However, the differences in the linearity between the parameters of compound nouns and those of collocations somewhat reflect the underlying differences in substitutability and transformability. In this section, the compositionality of nominal compounds on the basis of the notion of reference is investigated[23, 38]. Then, the differences in substitutability and transformability between compound nouns and collocations on the basis of their semantic properties are revealed.

5.1. Compositionality and Semantics of Nominal Compounds

The measurement of compositionality was based on the morpheme model with the help of the paraphrase "NA is N which is A", taking the view that the meanings of compositional nouns are predictable from their constituent meanings[40]. However, even for compositional nominal compounds, their meaning cannot be reduced to "the N which is A." For example, the phrase *un joueur qui est professionnel* ('a player who is professional') does not hold the same meaning as the nominal compound *joueur professionnel* ('professional player'). The latter refers to a profession, implying that being professional is an inherent characteristic of the player, forming a new concept—a subcategory of <profession>—while *un joueur qui est professionnel* merely describes a characteristic of a type of player without forming a new category. In contrast, the meaning of collocation *joueur combatif* can be equated with *un joueur qui est combatif*, both of which specify the characteristics of players. For the compound *jupe entravée*, it refers to a certain style of skirt, whereas *les jupes qui sont entravées* refers only to skirts with narrow and straight properties. If *jupe entravée* (skirt + hobble; 'hobble skirt') is paraphrased as "all N that are A are NA," it is observed that *jupe entravée* refers to a reduced concept compared to the reference of *les jupes qui sont entravées* (the + skirts + which + are + hobbled, 'the skirts which are hobbled'), as not all skirts that are hobbled are considered hobble skirts—a straight skirt that becomes hobbled due to an ill-fitting size remains a straight skirt, but not a hobble skirt.

Indeed, the reference of nominal compounds differs from that of collocations. The reference assigned to the whole unit of a nominal compound is conventional and well-defined. However, for a collocation, although the reference of the collocate can be conventional as *bleue* in *peur bleue*, the entire combination is not assigned a specific reference. Instead, the meaning of the collocation is the sum meaning of the base *peur*, referring to fear, and the collocate *bleue*, adopting its conventional reference "great" or "very". We access the meaning of a collocation by combining the individual references of its components, whereas the reference of a compound functions as a whole that we access through the entire complex form without investigating each individual reference of the component. In a strict sense, the meaning of nominal compounds is not predictable from constituent meanings, since the reference to the nominal compound is conventional, a reduced concept or at least a well-defined concept, and can never be predicted from constituent meanings, which are the summed meaning of a free phrase.

5.2. Substitutability of Nominal Compounds

As shown in **Figure 2**, the substitutability values of compounds and collocations overlap significantly and do not allow us to distinguish compounds from collocations (**Figures 7** and **8**). However, there is a difference in linearity between the transformability and the substitutability of nominal compounds and collocations (c.f. **Tables 1** and **2**). This implies a difference in substitutability between nominal compounds and collocations. For example, *spinal* ('spinal') is a synonym of *rachidien* ('spinal'), and both the nominal compounds *nerf rachidien* (nerve + spinal, 'spinal nerve') and *nerf spinal* in French refer to the same object. There are also pairs of compounds whose adjectives have no synonymous relationship but have the same reference, such as *abeille tueuse* (bee + murderous, 'killer bee') and *abeille africanisée* (bee + Africanized, 'killer bee'). Some pairs of compounds whose adjectives are antonyms do not refer to opposite meanings, such as *eau forte* (water + strong, 'nitric') and *eau douce* (water soft, 'softened water'), in which *douce* ('soft') is the antonym of *forte* ('strong'), but *eau forte* does not mean unsoftened water, contrary to the meaning of *eau douce*—"softened water". Nevertheless, the nominal compound *mouche sèche* has the opposite meaning of the compound *mouche noyée*, and *sèche* ('dry') is the antonym of *noyée* ('drowned'). Unlike collocations, which are subject to selectional constraints on their collocates, nominal compounds are combinations associated with a particular reference, and the formal combination and the associated reference to the whole combination are both conventional.

If N + Adj/Adj + N nominal compounds are considered lexicalized from syntactic constructions N + Adj/Adj + N, the process of lexicalization is, in fact an assignment of the reference to a particular phrase N + Adj, and both that combination and the reference are conventionally selected by the speech community. Once a specific reference is assigned, the construction is lexicalized, and the component of this construction can no longer be substituted as in free phrases. The substitutability of nominal compounds is in fact,

a relation inherited from the paradigmatical relationship that existed between the two phrases before their lexicalization. Furthermore, the paradigmatical relationship between nominal compounds can also be obtained via deconstruction. We speak of deconstruction when the reference of a lexical unit is decomposed and assigned to its components. By deconstruction, the reference of the original compound is no longer considered as a whole but is split and assigned to its components, which enables substitution or transformation. If the recombined reference constitutes a new reference that functions as a whole to be assigned to the substituted form, a new compound is produced. The paradigmatical relationship between these two compounds was adopted via deconstruction. For example, for the two nominal compounds *mouche noyée* and *mouche sèche*, when we selected the form *mouche noyée* for "fishing lure underwater" and *mouche sèche* for "fishing lure out of water", deconstruction occurred because the form *mouche* was preserved along with part of the original reference "fishing lure". Then, only the adjective was substituted by an antonym indicating the difference in reference: "out of water" opposite to "underwater". The reference has been decomposed into two parts and assigned to the components when the adjective was substituted.

In contrast to collocations that received selection restrictions on collocates, nominal compounds could not be selected according to combinatorial properties but were combinations assigned with a given reference. Once a conventional reference is assigned to a particular form or a combination, it makes the form difficult to transform. Therefore, the substitutability of nominal compounds is only the paradigmatical relationship that existed between them before they were lexicalized or adopted via deconstruction. As shown in **Figure 15**, the paradigmatical relationship between nominal compound A and nominal compound B is the paradigmatical relationship that existed between them before they were lexicalized. **Figure 16** shows the paradigmatical relationship between nominal compounds obtained via deconstruction.

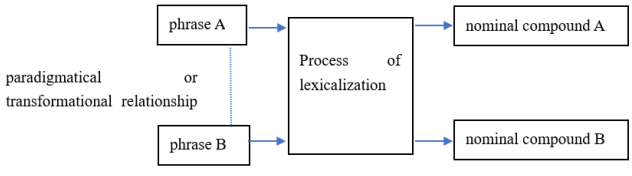

Figure 15. Inheritance of paradigmatical or transformational relationship.

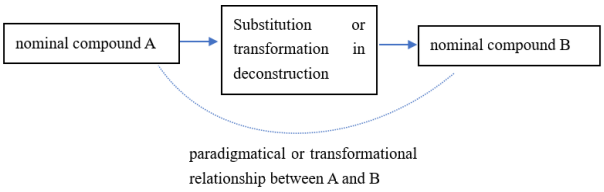

Figure 16. Paradigmatical or transformational relationship adopted via deconstruction.

5.3. Transformability of Nominal Compounds

The results in **Figures 9** and **10** reveal that some of the transformability values of nominal compounds and collocations overlap, and the transformability value cannot distinguish nominal compounds from collocations. Moreover, the transformability of nominal compounds also differs from that of collocations on the basis of the difference in linearity between their transformability and compositionality. For example, the transformation REPL (replacement by de + N as equivalence) was observed in the nominal compounds *livre scolaire* (book + scholastic, 'school book') and *livre de scolarité* (book + of + schooling, 'school book'). As argued in Section 5.2, once a conventional reference is assigned to a form, it is difficult to transform the form; thus, it is reasonable to consider that the compound *livre de scolarité* was not formed from the compound *livre scolaire* via syntactical transformation, but rather both of them were lexicalized from syntactical constructions, and conventionally, they have the same reference. For the compounds *beurre salé* (butter + salted, 'salted butter') and *beurre très salé* (butter + very + salted, 'butter with more salt'), the insertion of adverb *très* is observed. The compound *beurre salé* refers to the butter with salt added during the manufacturing process, and the compound *beurre très salé* refers to the butter with a greater quantity of salt added during manufacturing. This suggests that the intensifier "very" was applied to the reference of the component *salé* in the compound *beurre salé,* which implies the process of deconstruction, i.e., the reference of the nominal compound *beurre salé* has been split into parts and assigned to its components for accepting the insertion of the intensity.

In brief, transformability cannot distinguish nominal compounds, but the transformability of nominal compounds differs from that of collocations. In the case of a nominal compound, it is a form assigned a given reference, which makes the form difficult to transform. The transformability

of nominal compounds is the transformational relationship that exists between two phrases before their lexicalization or is produced via the process of deconstruction (**Figures 15** and **16**). It is supposed that the phrase was transformed when it was free, and it is lexicalized along with its transformed construction. However, collocation is a combination that only receives selection constraints on collocates, without a specific reference assigned to the whole form. Therefore, whether a collocation N + Adj/Adj + N is compositional or not, it accepts the syntactical transformations only if the semantic-syntactical properties of the adjective in it allow these transformations.

5.4. Status of Collocations

Collocations have long been considered a type of construction positioned on the transition between free and fixed phrases and should pass first by collocations before arriving at fixedness[5, 13–17, 19, 25, 26, 28]. However, the degree of fixedness of collocations is not always situated between free and fixed phrases. As shown in **Figures 11–14**, it is difficult to find a distinct boundary between the fixedness values of nominal compounds and collocations. The fixedness values of some collocations are the same or even higher than those of some nominal compounds. The fixedness values of many collocations and nominal compounds overlap. Moreover, as discussed in Sections 5.2 and 5.3, collocations receive restrictions on the choice of the collocate, whereas compounds are conventional combinations associated with specific references without the possibility of choice in combination. The substitutability and transformability of collocations and compounds are different in nature. Furthermore, there is no clear evidence that nominal compounds passed first through the collocation stage before arriving at fixedness. In contrast, it is possible that compounds can be lexicalized from both collocations and free phrases. Nominal compounds are products of lexicalization with fixed forms and fixed references, whereas collocations are not. Collocations are partially fixed; however, they are only restricted combinations, without a conventional reference to the entire form, although the components of collocations can have conventional references. The whole meaning of a collocation is still accessed by investigating the individual meanings of its component. There is no reason to consider that the lexicalization process proceeds in a particular order: first, the fixedness of the form and then the fixedness of the reference. Instead, as argued in Section 5.2, lexicalization is indeed an assignment of reference. Once a conventional reference is assigned to a form, free phrase or collocation, this form is lexicalized and fixed. Therefore, it is reasonable to consider that the form is apparently fixed along with the reference during lexicalization.

In summary, the meanings of nominal compounds can never be predicted from their constituent meanings because the latter is the summed meaning of the free phrase, which is different from the meaning of the compound in terms of reference. The reference to nominal compounds involves a well-defined concept that we access without investigating the meanings of components, whereas that of collocations and noun phrases does not. Moreover, the substitutability and transformability of nominal compounds differ from those of collocations. They are considered paradigmatical or transformational relationships that existed between them before their lexicalization or were produced in deconstruction. It is argued that lexicalization is an assignment of reference, and deconstruction is a split and a reassignment of the reference. Therefore, placing collocations on the transition between the free and fixed phrases is irrelevant. The idiomaticity of nominal compounds and that of collocations are different in all aspects: semantic (compositionality and reference), lexical (substitutability), and syntactical (transformability). Idiomaticity should not be considered a scalar value adopted from multiple factors as argued in previous studies[2, 18, 19, 50, 68], but instead as a discontinuous value that differs from one type of construction to another on various factors. From the semantic to the formal level, nominal compounds thus reveal differences from collocations.

6. Conclusions

This study aimed to define the idiomaticity of a specific category of fixed expressions—nominal compounds. The traditional assumptions of semantic fixedness (compositionality and reference) and formal fixedness (substitutability and transformability) in relation to idiomaticity were questioned. The results demonstrate that neither a single nor multiple parameters allow nominal compounds to be distinguished from collocations. The difference in linearity between parameters of fixedness of nominal compounds and collocations reveals their difference in substitutability and transformability. It is

argued that nominal compounds are combinations assigned a given reference, which is a reduced or well-defined concept that cannot be predicted from constituent meanings. The substitutability and transformability of nominal compounds are paradigmatical or transformational relationships that exist before lexicalization or are produced via deconstruction with the split and reassignment of reference.

This research provides a thorough quantitative analysis of the degree of fixedness of nominal compounds and collocations, which allows a precise examination of the validity of the notion of fixedness in judging idiomaticity and a statistical investigation of the linear relationships between the fixedness parameters of nominal compounds and collocations. The difference revealed by the latter provides experimental evidence for a deeper exploration of the difference in transformability and substitutability between nominal compounds and collocations. The qualitative analysis, based on the notion of reference, clarifies the difference in idiomaticity between nominal compounds and collocations and shows the limits of the concept of the continuum. The idiomaticity of nominal compounds differs from that of collocations in semantics and form but not only in terms of fixedness. The compositionality, substitutability, and transformability of nominal compounds differ fundamentally from those of collocations. Hence, it is argued that idiomaticity is not a scalar value but rather a discontinuous value that differs from one type of expression to another on various factors, both in terms of semantic and formal fixedness. These findings offer a new perspective on the notion of fixedness and provides new insight for future study on idiomaticity judgment.

Funding

This research project is supported by Science Foundation of Beijing Language and Culture University (supported by "the Fundamental Research Funds for the Central Universities") (Approval number: 17YBB14).

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request. The datasets generated and analyzed during the current study are not publicly available due to confidentiality imposed by data providers, but can be accessed by contacting the corresponding author. All relevant data and materials are available to qualified researchers.

Acknowledgment

We would like to thank all the reviewers for their valuable feedback and constructive comments, which greatly improved this manuscript. Our gratitude also extends to all project participants whose contributions were essential to the completion of this study. Additionally, we appreciate the support provided by Beijing Language and Culture University, which facilitated the research and provided the necessary administrative support.

Conflict of Interest

We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with this work.

References

[1] Baldwin, T., Kim, S.N., 2010. Multiword expressions. In: Indurkhya, N., Damerau, F.J. (Eds.) Handbook of Natural Language Processing, 2nd ed. CRC Press: Boca Raton, USA. pp. 267–292.

[2] Gross, G., 1988. Degré de figement des noms composés. In: Les expressions figées. Armand Colin. pp. 57–72.

[3] Lamiroy, B., Klein, J.R., 2010. Les expressions verbales figées de la francophonie. Ophrys, Paris.

[4] Pecman, M., 2004. L'enjeu de la classification en phraséologie. Actes du congrès EUROPHRAS. Schneider Hohengehren Verlag: Baltmannsweiler. pp. 127–146.

[5] Pamies, A., 2019. Phrasal verbs, idiomaticity and the fixedness continuum. In: Szerszunowicz, J., Awier, M. (eds.) Reproducible Multiword Expressions from a Theoretical and Empirical Perspective Special Issue of Intercontinental Dialogue on Phraseology. University: Bialystok. pp. 55–74.

[6] Pamies, A., 2022. The relationship between construc-

tions and units in Spanish phraseology. Islas. 64(202), 132–152.
[7] Palmer, H.E., 1933. Second interim report on English collocations. Tokyo: Kaitakusha.
[8] Firth, J.R., 1957. Papers in Linguistics 1934-1951. Oxford University Press: Oxford.
[9] Sinclair, J., 1991. Corpus, Concordance, Collocation. Oxford University Press: Oxford.
[10] Williams, G., 1999. Les réseaux collocationnels dans la construction et l'exploitation d'un corpus dans le cadre d'une communauté de discours scientifique [Ph.D. Thesis]. Université de Nantes.
[11] Hausmann, F.J., 1979. Un dictionnaire des collocations est-il possible? Travaux de Linguistique et de Litterature 17/1. 187–195.
[12] Mel'čuk, I., 1995. Introduction à la lexicologie explicative et combinatoire. Louvain, Duculot.
[13] Cowie, A.P., Mackin, R., Isabel R.M., 1983. Oxford Dictionary of Current Idiomatic English II (retitled Oxford Dictionary of English Idioms 1993). Oxford University Press: Oxford, UK.
[14] Alexander, R.J., 1984. Fixed expressions in English: reference books and the teacher. English Language Teaching Journal. 38(2), 127–132.
[15] Carter, R., 1987. Vocabulary: Applied Linguistics Perspectives. Routledge: London, UK.
[16] Nattinger, J.R., De Carrico. J.S., 1992. Lexical Phrases and Language Teaching. Oxford University Press: Oxford, UK.
[17] Goldberg, A., 1995. Constructions: A Construction Grammar approach to argument structure. University of Chicago Press: Chicago.
[18] Wulff, S., 2008. Rethinking Idiomaticity: A Usage-based Approach. Continuum Press: London, UK.
[19] Wulff, S., 2012. Idiomaticity. In: Robinson, P. (Ed.) The Routledge Encyclopedia of Second Language Acquisition. Routledge: London, UK. pp. 291–293.
[20] Fellbaum, C., 2019. How flexible are idioms? A corpus-based study. Linguistics. 57(4), 735–767.
[21] Giacomini, L., 2022. The contextual behaviour of specialised collocations: typology and lexicographic treatment. Yearbook of Phraseology. 13(1), 55–80. DOI: https://doi.org/10.1515/phras-2022-0005.
[22] Abeillé, A., 1995. The flexibility of French idioms: A representation with lexicalized tree adjoining grammar. In Everaert, M., Van der Linden, E.-J., Schenk, A., et al. (Eds.) Idioms: Structural and Psychological Perspectives. Hillsdale, Erlbaum. pp. 15–42.
[23] Sliwa, D., 2006. Unités polylexicales dans le cadre du syntagme nominal. Akadémiai Kiadó, Budapest. Verbum Analecta Neolatina. VIII(2), 407–415.
[24] Chakiri, M., 2010. La locution nominale: Entre opacité et transparence. Études et Documents Berbères. 1(29–30), 97–108.
[25] Goes, J.A., 2018. Collocations et dénominations: Quelques réflexions sur le figement adjectival. In Philologia mediana n°10.
[26] Fernando, C., 1996. Idioms and Idiomaticity. Oxford University Press: Oxford, UK.
[27] Sonomura, M.O., 1996. Idiomaticity in the Basic Writing of American English. Formulas and Idioms in the Writing of Multilingual and Creole-Speaking Community College Students in Hawaii. Peter Lang: New York, NY, USA.
[28] Gross, G., 1996. Les expressions figées en français (noms composés et autres locutions). Ophrys: Paris.
[29] Čermáková, F., 2007. Combination, Collocation and Multi-Word Units: Approaches to lexical combinatorics. In Proceedings of EURALEX 2000. pp.489–495.
[30] Hamma, B., 2010. La non-compositionnalité est-elle un indice du caractère locutionnel? Etudes de quelques locutions en sous. Les Cahiers de l'ED 139, Syntaxe et Sémantique grammaticale Descriptives. Publication de Paris Ouest Nanterre-la Défense. pp. 55–86.
[31] Trabelsi, H., 2015. L'idiomaticité fait-elle la langue de spécialité? AL - MUTARĞIM, n° 31.
[32] Ten Hacjen, P., 2017. The semantics of compounding. In Arndt-Lappe S., Bonami O., Ana R. Luís (eds.) Morphology. Cambridge University Press: Cambridge, UK. Volume 27, pp. 721–725.
[33] Vašků, K., 2019. Lexical idioms: What is regular and what anomalous in word-formation. Linguistica pragensia. Karolinum Press: Prague.
[34] Chomsky, N., 1980. Rules and Representations. Columbia University Press: New York, NY, USA.
[35] Fraser, B., 1970. Idioms Within a Transformation Grammar. Foundations in Language. 6, 22–42.
[36] Van der Linden, E.-J., 1992. Incremental processing and the hierarchical lexicon. Computational Linguistics. 18, 219–38.
[37] Polguère, A., 2015. Non-compositionnalité : ce sont toujours les locutions faibles qui trinquent. Verbum. XXXVII(2), 257–280. Presses Universitaires de Nancy.
[38] Nunberg, G., Sag, I.A., Wasow, T., 1994. Idioms. Language. 70 (3), 491–538.
[39] Libben, G., 1998. Semantic transparency in the processing of compounds: Consequences for representation, processing, and impairment. Brain and Language. 61(1), 30–44.
[40] Sandra, D., 1990. On the representation and processing of compound words: Automatic access to constituent morphemes does not occur. The Quarterly Journal of Experimental Psychology Section A. 42(3), 529–567.
[41] Zwitserlood, P., 1994. The role of semantic transparency in the processing and representation of Dutch compounds. Language and Cognitive Processes. 9(3), 341–368.
[42] Reddy, S., McCarthy, D., Suresh, M., 2011. An empirical study on compositionality in compound nouns.

In Proceedings of the 5th International Conference on Natural Language Processing. pp. 210–218. Available from: http://sivareddy.in/papers/files/ijcnlcompositionality_data.tgz.

[43] Tutin, A., Grossmann, F., 2002. Collocations régulières et irrégulières: Esquisse de typologie du phénomène collocatif. Revue française de linguistique appliquée. 7 (1), 7–25.

[44] Hausmann, F.J., 1997. Tout est idiomatique dans les langues. In: Martins-Baltar, M. (Ed.) La locution entre langue et usages. ENS Editions, Fontenay Saint-Cloud. pp. 277–290.

[45] Burger, H., 1998. Phraseologie: Eine Einführung am Beispiel des Deutschen. Erich Schmidt: Berlin, Germany.

[46] Cowie, A.P., (Ed.). 1998. Phraseology: Theory, Analysis, and Applications. Clarendon Press: Oxford, UK.

[47] Howarth, P., 1998. Phraseology and Second Language Proficiency. Applied Linguistics. 19(1), 24–44.

[48] Wulff, 2009. Converging evidence from corpus and experimental data to capture idiomaticity. Corpus Linguistics and Linguistic Theory. 5(1), 131–159.

[49] Sailer, M., 2013. Idiom and phraseology. Oxford Bibliographies Online Datasets.

[50] Anscombre, J.-C., 2003. Les proverbes sont-ils des expressions figées ? Cahiers de Lexicologie, 82, 159–173.

[51] Jacques, M.-P., 2003. Approche en discours de la réduction des termes complexes dans les textes spécialisés [Ph.D. Thesis]. Université de Toulouse II.

[52] Mel'čuk, I., Arbatchewsky-Jumarie, N., Elnitsky, L., et al., 1984. Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques I. Les Presses de l'Université de Montréal: Montréal.

[53] Polguère, A., 2000. Une base de données lexicales du français et ses applications possibles en didactique. Revue de linguistique et de didactique des langues (Lidil). 21, 75–97.

[54] Mel'čuk, I., Polguère, A., 2007. Lexique actif du français. De Boeck Supérieur.

[55] Hausmann, F.J., 1989. Le dictionnaire de collocations. In: Hausmann, F.J., Reichmann, O., Wiegand, H.E., et al., (Eds.) Wörterbücher: ein internationals Handbuch zur Lexicographie. Dictionnaires. De Gruyter: Berlin., Germany. pp. 1010–1019.

[56] Tutin, A., 2010. Le traitement des collocations dans les dictionnaires monolingues de collocations du français et de l'anglais. Congrès Mondial de Linguistique Française. EDP Sciences: Paris, France. pp. 1075–1090.

[57] Matthews, P.H., 1991. Morphology: An Introduction to the Theory of Word-Structure, 2nd ed. Cambridge University Press: London, UK.

[58] Corbin, D., 1992. Hypothèses sur les frontières de la composition nominale. Cahiers de grammaire. 17, 25–55.

[59] Booij, G., 2002. The Morphology of Dutch. Oxford University Press: Oxford, UK.

[60] Haspelmath, M., 2002. Understanding Morphology. Arnold: London, UK.

[61] Fradin, 2003. Nouvelles approches en morphologie. Presses universitaires de France: Paris, France.

[62] Villoing, F., 2009. Les mots composés VN. In Bernard. Fradin, Françoise Kerleroux & Marc Plénat (eds.) Aperçus de morphologie du français. Presses Universitaires de Vincennes: Saint-Denis. pp. 75–197.

[63] Villoing, F., 2012. French compounds. Probus. Int. J. Latin Romance Linguist. 24(1), 29–60.

[64] Fradin, B., 2000. Combining forms, blends and related phenomena. In: Dolsescha, U., Thrornton, A.M. (Eds.) Extragrammatical and Marginal Morphology. Münich, Lincom Europa. pp. 11–59.

[65] Anderson, S.R., 1992. A-morphous morphology. Cambridge University Press: Cambridge, UK.

[66] Dal & Amiot 2008.

[67] Andrews, D.F., 1972. Plots of high dimensional data. Biometrics. 28, 125–136.

[68] Jacquemin, C., 1991. Transformations des noms composés. Thèse en informatique fondamentale, Université de Paris 7.