sciendo
DOI 10.2478/jazcas-2025-0010

IDIOMS IN DISGUISE: HOW GRAMMATICAL PROFILING REVEALS PHRASEOLOGICAL PATTERNS

ANNA VYSLOUŽILOVÁ[1] – DOMINIKA KOVÁŘÍKOVÁ[2]
[1]Institute of Linguistics, Faculty of Arts, Charles University, Prague, Czech Republic
(ORCID: 0009-0001-0670-7993)
[2]Institute of Linguistics, Faculty of Arts, Charles University, Prague, Czech Republic
(ORCID: 0000-0002-4419-6901)

VYSLOUŽILOVÁ, Anna – KOVÁŘÍKOVÁ, Dominika: Idioms in Disguise: How Grammatical Profiling Reveals Phraseological Patterns. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 113 – 122.

Abstract: This article examines how morphological anomalies – specifically, the unusually high frequencies of certain singular noun forms – can reveal idiomatic usage in Czech. Using data from the GramatiKat tool, 1,102 noun lemmas were analyzed, of which 28% participated in idiomatic expressions. The study identifies clear distributional patterns across grammatical cases, with idioms most frequent in the accusative, genitive, locative, and instrumental singular. Monocollocational idioms are distinguished, as they are associated with specific structural patterns. The results show that idiomatic expressions can influence morphological distributions and leave measurable traces in corpus data. The approach is further applicable to other parts of speech, such as verbs and adjectives, suggesting a broader role for grammatical profiling in the identification of idiomatic and phraseological patterns.

Keywords: grammatical profiling, morphological anomalies, idiomatic expressions

1 INTRODUCTION

In morphologically rich languages like Czech, lexemes rarely exhibit uniform frequency distributions across their paradigm forms (Janda and Tyers 2021). Instead, certain grammatical forms often occur with markedly higher frequency than others, creating distinctive grammatical profiles. This paper investigates whether such morphological anomalies – particularly nouns with unusually high frequencies in specific case forms – can serve as reliable indicators of idiomatic expressions in Czech.

Research into the frequency distribution of grammatical forms has a solid tradition in Czech linguistics (Jelínek, Bečka and Těšitelová 1961; Bartoň et al. 2009; Cvrček et al. 2010). These studies have established that grammatical anomalies often correlate with specific lexical combinations found in idiomatic expressions. As Kodýtek (in Cvrček et al. 2010) observes, morphological distributions are influenced by semantic factors: nouns denoting animate entities typically show higher

frequencies in nominative forms, while inanimate nouns often display increased occurrences in genitive and accusative cases.

The analysis builds upon previous research by extending Kováříková's (in press) study of morphological anomalies in the dative singular and incorporating Vysloužilová's (Dittrichová 2024) findings on the relationship between morphological anomalies and multi-word units.

By analyzing 1,102 anomalous noun lemmas from the corpus tool GramatiKat (Kováříková and Kovářík 2021), this study addresses three questions: (1) Can paradigmatic imbalance indicate idiomatic expressions? (2) Which idiom types most frequently underlie such anomalies? and (3) How does the relationship between morphological anomalies and idiomaticity vary across grammatical cases? The findings reveal that over a quarter of lemmas displaying distributional outliers participate in idioms, with proportions exceeding 80% in certain cases, suggesting that morphological anomalies can serve as pathways for identifying phraseological patterns.

2 THEORETICAL FRAMEWORK

This study positions itself at the intersection of corpus linguistics, Construction Grammar, and phraseology. From a corpus linguistics perspective, our approach follows Sinclair's (1991) emphasis on examining actual usage patterns rather than linguistic intuitions, while employing frequency-based criteria for identifying phraseological units as outlined by Gries (2008). By using distributional outliers as the entry point for analysis, we employ a corpus-driven rather than corpus-based approach (Tognini-Bonelli 2001), allowing patterns to emerge from frequency data rather than testing predefined hypotheses.

Within Construction Grammar (Goldberg 1995; Croft 2001), linguistic patterns are understood as form-meaning pairings with varying degrees of fixedness and conventionality. When certain grammatical forms appear with unusual frequency in particular contexts, this often signals their entrenchment in linguistic usage.

The phraseological dimension builds on Čermák's (2007) conception of idioms as involving both formal and semantic anomaly. We developed a modified classification system to accommodate the specific patterns revealed in our corpus analysis. This approach investigates whether paradigmatic imbalance serves as an effective entry point for idiom identification across different grammatical cases.

3 DATA SOURCE AND SAMPLE SELECTION

This study utilizes the corpus tool GramatiKat (Kováříková and Kovářík 2021), which analyzes data from the SYN2015 corpus to provide detailed information on the distribution of word forms across word classes and specific lemmas. It compares

these distributions to class-wide patterns through interactive tables that help identify lemmas with anomalous behavior.

The tool distinguishes between two anomaly types: **upper outliers** (lemmas with unusually high frequency of specific forms) and **lower outliers** (forms with very low or zero frequency). Upper outliers are defined as lemmas whose frequency in a given form exceeds 1.5 times the interquartile range above the 75th percentile; lower outliers typically lack any corpus attestation (Kováříková 2021).

This study focuses exclusively on upper outliers – noun lemmas with disproportionately high frequency in particular singular forms. Using GramatiKat's "Anomalous lemmas" function, we selected the "Noun" category and examined each singular case separately. To ensure comparability and manage sample size, we selected the top 20% of anomalous lemmas for each case based on frequency deviation scores. Lemmas with tied values at the cutoff point were also included.

The selection used GramatiKat version 1. For each case form, we exported, sorted, and thresholded the lemmas. The final sample also contained mistagged or duplicate lemmas, which were kept for transparency but excluded from idiom analysis. In total, 1,102 lemmas were analyzed out of 5,120 anomalous entries (see Tab. 1).

Case	**Total**	**Sample**
Nominative sg.	769	162
Genitive sg.	361	75
Dative sg.	1,169	252
Accusative sg.	321	66
Vocative sg.	839	201
Locative sg.	809	169
Instrumental sg.	852	177
Sum	**5,120**	**1,102**

Tab. 1. Number of anomalous lemmas per case and sample size (top 20%)

4 METHODOLOGY[1]

The analysis focused on identifying and classifying idiomatic constructions associated with the anomalous noun lemmas. All lemmas included in the study were drawn from the *GramatiKat* tool as described above. For each lemma, the specific anomalous form – typically a case form with unusually high frequency – served as the starting point for corpus exploration.

[1] A more detailed description of the methodology can be found in Vysloužilová's thesis (Dittrichová 2024).

The corpus analysis was conducted using the KonText application, drawing on two corpora, SYN2015 and SYNv11 (synchronic corpora of written Czech). Each lemma was searched in the form in which it exhibited the anomaly, using a corresponding CQL query adapted to the grammatical case.

Idiom identification employed two complementary approaches:

1. **Automatic annotation** using the FRANTA annotation tool, which tags multi-word units based on a predefined list of approximately 40,000 phraseological units, mostly from Čermák's *Slovník české frazeologie a idiomatiky* (Čermák 2009; Čermák and Hronek 2009a–c). We queried both the SYN2015 subcorpus within SYNv11 and the full SYNv11 corpus.
2. **Collocational analysis** using KonText's Collocations function with the following parameters: collocation window span of –3 to +3, minimum collocate frequency of 3, and sorting by the logDice association measure.

For idiom classification, we developed a custom typology with nine categories that combined structural and functional criteria, as the traditional tripartite typology of verbal, non-verbal, and propositional idioms (Čermák and Hronek 2009c) was found to be too coarse, while Čermák's detailed typology based on structural components (Čermák 2007) proved too fine-grained for the purposes of this study.

Six primary categories:

- Grammatical idioms (such as multi-word prepositions, e.g. *z hlediska* 'from the perspective of')
- Monocollocational idioms (containing a component with extremely limited collocability, e.g. *být k mání* 'to be available')
- Binomials (characterized by repetition of two formaly similar components, e.g. *alfa a omega* 'the alpha and omega')
- Similes (e.g. *žít si jako v bavlnce* 'to live in cotton wool')
- Contact idioms (e.g. *pozdrav pánbůh* 'God bless you')
- Foreign-language units (e.g. *alma mater*).

Three broader types for remaining idioms:

- Nominal idioms (e.g. *od malička* 'since childhood')
- Verbal idioms (e.g. *hodit zpátečku* 'to shift into reverse')
- Propositional idioms (e.g. *andělíčku, můj strážníčku, opatruj mi mou dušičku* 'little angel, my guardian, protect my little soul').

5 IDIOMATICITY ACROSS GRAMMATICAL CASES

Of the 1,102 anomalous noun lemmas analyzed, 28% (306 lemmas) participated in one or more idiomatic expressions. The distribution of idiomaticity varied

markedly across grammatical cases, revealing significant asymmetries in how cases participate in phraseological patterns (Tab. 2).

Case	Number of lemmas	Idiomatic lemmas	Percentage
Nominative sg.	162	15	9%
Genitive sg.	75	39	52%
Dative sg.	252	47	19%
Accusative sg.	66	58	88%
Vocative sg.	201	10	5%
Locative sg.	169	71	42%
Instrumental sg.	177	66	37%
Total	**1,102**	**306**	**28%**

Tab. 2. Proportion of idiom-participating lemmas by case

The accusative singular exhibited the strongest correlation with idiomatic usage (88% of analyzed lemmas). These idioms frequently involved the preposition *na* and included numerous monocollocational idioms (expressions in which one component appears almost exclusively in that specific phrase) such as *dávat si bacha* (‘to watch out’) and *brát v potaz* (‘to take into account’). Verbal idioms in this case often featured substantivized adjectives, as in *být na pováženou* (‘to be questionable’) or *dát někomu čas na rozmyšlenou* (‘to give someone time to think it over’). Many of these expressions combined with the verb *dát/dávat* (‘to give’), including *dát někomu na srozuměnou* (‘to make something clear to someone’) or *dát někomu něco na požádání* (‘to provide something upon request’). Several idioms also referenced cultural or temporal contexts, such as *na Zelený čtvrtek* (‘on Green Thursday’) or *na doživotí* (‘for life’).

The genitive singular showed idiomaticity in 52% of cases and was associated with binomials more than other cases: *ani vidu, ani slechu* (‘not a trace’) and *bez ladu a skladu* (‘without order or structure’). The genitive also appeared in numerous prepositional idioms with *bez, do*, and *od*, as in *bez prodlení* (‘without delay’), *dostat se do ráže* (‘to get fired up’) or *od malička* (‘since early childhood’).

The locative singular displayed idiomatic usage in 42% of analyzed lemmas and was particularly rich in grammatical idioms, especially multi-word prepositional constructions with *v* or *na*: *v rámci* (‘within the framework of’) and *na základě* (‘on the basis of’). The locative also featured monocollocational idioms like *být ve střehu* (‘to be on alert’) and *v mžiku* (‘in an instant’).

In contrast, nominative (9%) and vocative (5%) forms rarely participated in idioms. When they did, they typically appeared in contact idioms (ty vole – ‘dude’), exclamatory formulas (*pane bože* – ‘oh my God’), or foreign expressions (*alma mater, Ave Maria*).

Idiom type	Nom. sg.	Gen. sg.	Dat. sg.	Acc. sg.	Voc. sg.	Loc. sg.	Instr. sg.	Total
Grammatical idioms	0	1	0	0	0	11	3	**15**
Monocollocational idioms	1	7	3	13	0	10	8	**42**
Binomials	1	3	0	1	0	1	0	**6**
Similes	3	0	0	0	0	1	2	**6**
Contact idioms	3	0	0	4	8	0	2	**17**
Foreign-language units	3	2	0	0	0	0	0	**5**
Nominal idioms	2	17	4	13	0	19	19	**74**
Verbal idioms	2	9	40	27	1	28	31	**138**
Propositional idioms	0	0	0	0	1	1	1	**3**

Tab. 3. Idiom types by case (number of lemmas)

The identified idioms were classified into nine types based on structural and functional properties (see section 4). Verbal idioms emerged as the most prevalent, accounting for 138 lemmas and showing particular concentration in the dative, instrumental, locative, and accusative cases. Though less common, nominal idioms (74 lemmas) clustered notably in the locative and instrumental cases, with significant presence in the genitive and accusative as well. Monocollocational idioms, comprising 42 lemmas, revealed a widespread distribution pattern across multiple cases, particularly favouring the accusative and locative (more about this type in section 6). The analysis uncovered clear case preferences among certain idiom types – grammatical idioms appeared almost exclusively in the locative case, while contact idioms gravitated toward vocative. The remaining categories – binomials, similes, and foreign-language units – appeared infrequently in the corpus, with just 5–6 lemmas each distributed sparsely across different cases. This uneven distribution pattern confirms that idiom types do not spread randomly across grammatical cases but rather reflect underlying structural constraints and functional contexts of language use.

6 MONOCOLLOCATIONAL IDIOMS: STRUCTURE AND DISTRIBUTION

Among the nine idiom types identified, monocollocational idioms represent a particularly distinctive category characterized by containing components that rarely appear outside the specific idiomatic construction, creating strong lexical restrictions that contribute to morphological anomalies. They often contain the verb *být* (‘to be’) or a light verb (e.g. *dát*, ‘to give’, *mít* ‘to have’) combined with a fixed noun phrase, often introduced by a preposition.

The 42 monocollocational idioms identified in our sample exhibited clear distributional patterns across grammatical cases, with the accusative (13 lemmas),

locative (10), instrumental (8), and genitive (7) showing the highest frequencies. This distribution indicates that certain cases offer especially favourable conditions for these fixed expressions, while others – notably the vocative, with no occurrences – do not support this idiom type.

6.1 Case-based distribution of monocollocational idioms

The accusative singular is especially productive for monocollocational idioms, typically following a verb + *na* + noun pattern. These often incorporate substantivized adjectives such as *srozuměnou* or *rozmyšlenou*:

- *dát na srozuměnou* (‘to make clear’)
- *dát na rozmyšlenou* (‘to give time to think’)
- *vystavovat něco na odiv* (‘to flaunt something’)
- *brát v potaz* (‘to take into account’).

The locative singular is also common, typically with *v*:

- *být ve střehu* (‘to be on alert’)
- *zmizet v propadlišti dějin* (‘to disappear into the abyss of history’)
- *v mžiku* (‘to be in an instant’)
- *v hloubi duše* (‘deep down’).

Instrumental singular idioms occur more often without prepositions:

- *zářit novotou* (‘to shine with novelty’)
- *končit fiaskem* (‘to end in a fiasco’)
- *nehnout ani brvou* (‘not even blink’)
- *mít něco za lubem* (‘to have something up one’s sleeve’).

Genitive singular idioms often involve *do*:

- *vyšumět do ztracena* (‘to fade away into nothing’)
- *nemít potuchy* (‘to have no idea’)
- *do třetice všeho dobrého* (‘third time’s the charm’)
- *dostat něco do vínku* (‘to be endowed with something at birth’).

6.2 Productive constructions beyond morphological anomaly

The monocollocational idioms identified in our study revealed several productive patterns, with *být/nebýt k* + noun in the dative singular standing out as particularly notable. Monocollocational expressions such as *být k mání* (‘to be available’), *být k nesnesení* (‘to be unbearable’), *být k popukání* (‘to be hilarious’), and *být k snědku* (‘ready to be eaten’) exemplify this pattern. While these constructions were initially identified as part of our search for morphological anomalies, their recurring formal structure suggested a more systematic phenomenon deserving deeper investigation.

Further analysis, as documented in Kováříková (in print), showed that the *být/nebýt k* + dative construction is far more productive than initially expected. This pattern encompasses dozens of items, many of which do not display the stark

morphological anomalies that first drew our attention or do not qualify as strictly monocollocational. The identified construction *být/nebýt k* + noun is not merely a random collection of idioms but rather a partially schematic template that combines fixed grammatical elements (the verb *být* and the preposition *k*) with a variable nominal component.

This expanded perspective also revealed the existence of additional constructional types with similar syntactic foundations but different preposition-case combinations. For example, constructions with the accusative case and preposition *na* typically express states approaching a limit or breakdown: *být na spadnutí* ('to be about to collapse'), *být na vyhození* ('to be fit for disposal'), *být na vymření* ('to be on the verge of extinction'). In parallel, *být v* + locative constructions like *být v pokušení* ('to be tempted'), *být v ohrožení* ('to be in danger'), or *být v napětí* ('to be tense') typically denote internal or situational states that involve an element of danger, pressure, or tension.

These patterns demonstrate that the *být* + preposition + noun in a certain case frame represents a broader system of idiomatic expressions in Czech, within which the *k* + dative variant stands out for its productivity and formal coherence. This finding illustrates how initial observations about monocollocational idioms can lead to the discovery of more extensive constructional patterns that blur the boundary between grammar and lexicon.

7 CONCLUSION

This study has shown that grammatical anomalies – defined as unusually high frequencies of particular singular noun forms – can serve as useful indicators of idiomatic expressions. In many cases, what first appears to be a morphological irregularity turns out to reflect the influence of fixed multi-word combinations. When a noun occurs disproportionately in one case form, it is often because it regularly appears in a specific idiomatic construction. This tendency is especially clear in the accusative (88%), genitive (52%), locative (42%), and instrumental (37%) singular, whereas the nominative (9%) and vocative (5%) show minimal idiomatic usage. Of the 1,102 anomalous lemmas analyzed, 28% participated in idioms. Verbal idioms were the most frequent (138 lemmas), followed by nominal idioms (74) and monocollocational idioms (42). These findings demonstrate the potential of corpus-based methods to uncover idiomatic patterns that may remain unnoticed in dictionary-based or introspective approaches.

The findings suggest that paradigmatic imbalance can reflect syntagmatic regularity. Idioms and other multi-word constructions appear to influence the frequency of specific forms within a paradigm, shaping usage patterns in observable ways. This distributional signature is measurable through corpus analysis, suggesting that idiomaticity functions not just semantically but also as a morphological phenomenon with quantifiable effects.

While this study focused on nouns, the profiling method used in GramatiKat may also be useful for exploring distributional patterns in other word classes. Preliminary observations point to promising directions: in verbs, certain lexical items appear disproportionately in feminine or masculine forms. For instance, *háčkovat* (‘to crochet’), *zachichotat se* (‘to giggle’), or *proplakat* (‘to cry through’) are more frequent in feminine past tense forms, while *narukovat* (‘to enlist’), *vloupat se* (‘to break in’), or *habilitovat se* (‘to obtain habilitation’) occur more often in masculine animate. In adjectives, anomalies often arise from multi-word terms, where the gender of the adjective is determined by the head noun of complex noun phrases – for example, *akciová společnost* (‘joint-stock company’), *ministerská vyhláška* (‘ministerial decree’), or *vysoká škola* (‘university’). These regularities may be phraseological rather than idiomatic, but they still show how lexical, syntactic, and discursive conventions shape morphological distributions.

By combining computational anomaly detection with careful qualitative analysis, this research contributes to data-driven approaches to phraseology and grammatical profiling. The methodology presented here demonstrates how corpus evidence can complement traditional idiom identification methods, potentially uncovering patterns that might otherwise remain undetected using conventional approaches.

ACKNOWLEDGEMENTS

This work was supported by the European Regional Development Fund project “Beyond Security: Role of Conflict in Resilience-Building” (reg. No. CZ.02.01.01/00/22_008/0004595).

References

Bartoň, T., Cvrček, V., Čermák, F., Jelínek, T., and Petkevič, V. (2009). Statistiky češtiny. Nakladatelství Lidové noviny.

Croft, W. (2001). Radical Construction Grammar: Syntactic Theory in Typological Perspective. Oxford University Press.

Cvrček, V. et al. (2010). Mluvnice současné češtiny. 1, Jak se píše a jak se mluví. Karolinum.

Čermák, F. (2007). Frazeologie a idiomatika: česká a obecná = Czech and General Phraseology. Karolinum.

Čermák, F. (2009). Slovník české frazeologie a idiomatiky. 4, Výrazy větné. Leda.

Čermák, F., and Hronek, J. et al. (2009a). Slovník české frazeologie a idiomatiky. 2, Výrazy neslovesné. Leda.

Čermák, F., and Hronek, J. et al. (2009b). Slovník české frazeologie a idiomatiky. 3, Výrazy slovesné. Leda.

Čermák, F., and Hronek, J. et al. (2009c). Slovník české frazeologie a idiomatiky. 1, Přirovnání. Leda.

Dittrichová, A. (2024). Vyhledávání frazémů na základě anomálie v distribuci tvarů. Diploma thesis, supervisor Kováříková, D. Ústav českého jazyka a teorie komunikace, FF UK, Praha. Accessible at: http://hdl.handle.net/20.500.11956/188429.

Goldberg, A. E. (1995). Constructions: A Construction Grammar Approach to Argument Structure. Chicago: University of Chicago Press.

Gries, S. T. (2008). Phraseology and linguistics theory: A brief survey. In: S. Granger – F. Meunier (eds.): Phraseology: An interdisciplinary perspective, pp. 3–25. John Benjamins.

Janda, L., and Tyers, M. (2021). Less is more: why all paradigms are defective, and why that is a good thing. Corpus Linguistics and Linguistic Theory, 17(1), pp. 109–141.

Jelínek, J., Bečka, J. V., and Těšitelová, M. (1961). Frekvence slov, slovních druhů a tvarů v českém jazyce. Státní pedagogické nakladatelství.

Kováříková, D. (in press). Morfologické anomálie jako klíč k idiomatickým konstrukcím. Studie z korpusové lingvistiky, svazek 30. Nakladatelství Lidové noviny.

Kováříková, D. (2021). Sharing data through specialized corpus-based tools: the case of GramatiKat. Jazykovedný časopis, 72(2), pp. 531–544.

Kováříková, D., and Kovářík, O. (2021). GramatiKat. Nástroj pro výzkum gramatických kategorií a gramatických profilů. FF UK. Accessible at: http://www.korpus.cz/gramatikat.

Křen, M. et al. (2015). SYN2015: reprezentativní korpus psané češtiny. FF UK. Accessible at: https://www.korpus.cz.

Křen, M. et al. (2022). Korpus SYN, verze 11 ze 14/12/2022. FF UK. Accessible at: https://www.korpus.cz.

Machálek, T. (2014). KonText – rozhraní pro vyhledávání v korpusech. FF UK. Accessible at: http://kontext.korpus.cz/.

Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford University Press.

Tognini-Bonelli, E. (2001). Corpus Linguistics at Work. John Benjamins.