Peer-reviewed article

From Norm to Usage: Computational Insights into the Genitive–Accusative Alternation in Ukrainian with Reference to Russian

Inna V. Stupak & Harald R. Baayen

University of Tübingen

inna.stupak@uni-tuebingen.de; harald.baayen@uni-tuebingen.de

Abstract

While sharing a common Slavic origin, Ukrainian and Russian exhibit both grammatical similarities and significant differences, particularly in noun morphology and vocabulary. This study investigates the representation of Ukrainian noun morphology, focusing on grammatical case and number, within distributional semantic vector spaces (fastText embeddings). We pay special attention to the complex relationship between the Accusative (Acc) and Genitive (Gen) cases, known for frequent alternation and syncretism in Ukrainian. Using t-SNE for visualization and DBSCAN for clustering on a dataset derived from the GRAC corpus, we analyze the structure of the vector space for Ukrainian nouns, comparing findings to previous work on Russian and conducting a direct comparison using translational equivalents. Our results show that while grammatical number provides a relatively strong organizing principle for Ukrainian nouns, case representation is less distinct, with substantial Acc/Gen overlap observed across multiple semantic classes, reflecting the pervasiveness of linguistic alternation rules (e.g., partitive, animacy-based, negation). A focused DBSCAN analysis of the ACC/GEN overlap zone revealed mostly diffuse similarity ("noise") but also distinct dense clusters

Slovo 66:1 (2026), 238–259. DOI https://doi.org/10.33063/slovo.v66i1.1234
CC BY 4.0

comprising primarily AccSg, GenSg, and GenPl forms, notably excluding AccPl forms. Furthermore, a comparative t-SNE visualization showed that Russian translational counterparts of the Ukrainian nouns from these dense clusters exhibit a similarly high degree of Acc/Gen mixing. The findings highlight both potentially language-specific organizational principles (number dominance) and shared distributional tendencies between Ukrainian and Russian in representing Acc/Gen functional overlap, suggesting common underlying semantic factors are captured despite finer-grained grammatical differences.[1]

1. Introduction

Within the diverse landscape of Slavic languages, the East Slavic branch, comprising Russian, Ukrainian, and Belarusian, holds a unique position. While their shared ancestry is acknowledged, scholarly perspectives on their precise historical paths diverge. It is generally accepted that these three languages form a subgroup, though practical communication reveals nuances: mutual intelligibility is often reported as higher between Ukrainian and Belarusian than between Ukrainian and Russian. This observation hints at a complex relationship between Ukrainian and Russian, characterized by both grammatical similarities (Shevelov 1993; Timberlake, 1993) and a notable divergence in terms of their vocabulary (Press and Pugh, 2015a). While their grammatical structures share a common foundation rooted in their Slavic heritage, their vocabularies have followed notably distinct trajectories.[2]

This lexical divergence is particularly salient in quantitative comparisons. Contemporary Ukrainian and Russian share only about 62 % of their vocabulary, of which merely 18% consists of lexemes that are formally and semantically identical; an additional 44% comprises items that are recognizably related but not identical (Tyščenko, 2000). This stands in sharp contrast to Ukrainian's relationship with Belarusian (approximately 82% overlap) and Polish (around 70%). These figures underscore that Ukrainian and Russian, despite close genealogical ties, have followed distinct evolutionary trajectories, particularly in their lexical and morphological systems. While Ukrainian and Russian share a core grammatical framework of gender, number, and case, their contemporary realizations

[1] Acknowledgements. This project has received funding through the MSCA4Ukraine project awarded to the first author and supported by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the MSCA4Ukraine Consortium as a whole nor any individual member institutions of the MSCA4Ukraine Consortium can be held responsible for them.
[2] The present analyses rely on standard descriptive grammars of both languages (Shevelov 1993; Timberlake, 1993; Bezpojasko, 1991).

diverge in subtle ways. For instance, Both languages historically treated the masculine as the unmarked member of gender opposition (Corbett, 1991). However, contemporary Ukrainian has developed a highly productive derivational system for feminine professional nouns, such as *inspektorka* ‘female inspector’ or *filolohynja* ‘female philologist’, a system formally codified in the 2019 Ukrainian orthography. Russian also possesses a range of derivational suffixes for feminine professional nouns. However, within Standard Russian (i.e. the codified literary norm), their productivity appears to be relatively more constrained and lexically uneven (Nesset et al., 2022). In many cases, professional nouns lack widely accepted feminine counterparts, or such forms display noticeable stylistic or register-based marking (Kirey-Sitnikova, 2021).

As shown in Table 1, Ukrainian exhibits a high degree of diversification of morphological forms (i.e., inflectional realizations), especially in the singular declension classes. The structural organization of the Ukrainian nominal vector space might be rooted in the inflectional landscape summarized in Table 1. This table serves as a functional map of distributional ’noise’ and ’signal’ for the computational model.

Cases	Singular				Plural			
	1	2	3	4	1	2	3	4
NOM	-a/-ja	∅, -o /-e / -ja	∅, -y	-ja/-a	-y/-i/-ji	-y/-i/ji, -a/-ja	-i	-a
GEN	-y/-i, -ji	-a/-ja, -u/ -ju	-i/-y	-jaty/ aty, -i/-ja	∅, -ej, -iv	-iv/-jiv, ∅, ej	-ej, -iv	-
DAT	-i/-ji	-u/-ju, -ovi/ -evi/-jevi	= GEN	-jati/ -ati, -i	-am/ -jam	-am/-jam	-am/ -jam	-am
ACC	-u/-ju	=NOM or GEN	=NOM	=NOM	=NOM or GEN	=NOM or GEN	=NOM or GEN	=NOM or GEN
INS	-oju/-eju /-jeju	-om/-em / -jem, -jam	-ju	-jam/ -am, -em	-amy/ -jamy, -my	-amy/ -jamy, -yma	-amy/ -jamy	-amy/- jamy,
LOC	= DAT	-i/-ji, -u/-ju, -ovi/-evi/-jevi,	-i	-jati/ -ati, -i	-ah/-jah	-ah/-jah	-ah/ -jah	-ah
VOC	-o/-e /, je, ju	-e, -u/ -ju, -ja	-e, -y/ ∅	= NOM	= NOM	= NOM	=NOM	=NOM

Table 1: Paradigmatic syncretism and inflectional diversification in Ukrainian nouns. The table highlights the formal overlaps (syncretism) and variations that govern the distributional behavior of word embeddings. High diversification in the singular vs. unification in the plural explains the spatial salience of Grammatical Number, while recurrent syncretism in the ACC row predicts the spatial overlap between cases in t-SNE visualizations

Differences in the nominal systems in Ukrainian and Russian extend to the status of the Vocative case. In the standard grammatical tradition, modern Russian is described as having a six-case system, with the historical vocative becoming obsolete by the late medieval period and being functionally replaced by the nominative (Vinogradov 1952; Švedova et al. 1980; Testelec 2001).

While contemporary Russian allows vocative-like address forms such as pap! or Saš!, these so-called "new vocatives" are generally analyzed as phonologically reduced or truncated forms restricted to a small and lexically limited set of kinship terms and personal names. They lack a productive morphological paradigm, do not participate in regular inflectional oppositions, and are not associated with case assignment or syntactic valency (Vinogradov 1952; Arkadiev 2008; Padučeva 1985, 2004).

This situation contrasts with contemporary Ukrainian, where since 1990s the Vocative is a morphologically distinct, productive, and paradigmatically integrated category. Although the Vocative is morphologically codified in Ukrainian, corpus-based evidence suggests it accounts for only approximately 40% of all address tokens, with the Nominative frequently substituted in contemporary practice (Shvedova and Lukashevskyi 2025). For a distributional model like *fastText*, this implies that the Ukrainian Nominative occupies a broader and more diverse set of syntactic contexts than its Russian counterpart.

The category of number further illustrates the importance of a paradigmatic, rather than purely formal, approach. Both languages today operate with a binary singular–plural distinction, but historically inherited a dual number, remnants of which persist in so-called paucal constructions with the numerals 'two', 'three', and 'four'. Both Ukrainian and Russian exhibit complexity in paucal constructions, but differ in their morphological preferences. In Ukrainian, nouns following numerals 2–4 typically occur in the Nominative plural (*dvi knyhy* NOM.PL. 'two books'), with Genitive singular forms attested in specific lexical, semantic, and constructional contexts (Stupak and Baayen, n.d.). The Nominative plural functions as the unmarked and most frequent option (Horpynyč 2004). However, the form (*dvi knyhy* NOM.PL. 'two books') often coincides morphologically with the Genitive singular of the noun, as in *knyhy* (GEN.SING) (see Table 1).

In the Russian grammatical tradition, by contrast, numerals 2–4 are traditionally analyzed as selecting a noun in the GEN.SING (Breu 2020; Franks and House 1982; Isačenko 1960; Švedova et al. 1980; Titov 2016; Corbett 1983; Plungjan 2011). However, Genitive singular forms can be syncretic with nominative plural forms (e.g., *dve knigi* 'two books' or *dva pjatna* 'two spots'). In other words, the syncretic form *knigi* is described as a Genitive singular in Russian grammars, while *knyhy* is described as a nominative plural in Ukrainian grammars.

This is not merely a label change; it represents a fundamental difference in how the language "tags" counting contexts. If Russian processes "plural" contexts (2–4 objects) as "Genitive Singular," the distributional model will logically pull plural vectors toward the singular-case cluster, reinforcing Case as the dominant signal. In Ukrainian, the alignment of these constructions with the plural paradigm may shift their vectors into either plural and singular clusters or a single unified plural cluster.

Taken together, the structural differences in gender, case and number raise broader questions about the interaction between shared grammatical heritage and language-specific morphological organization. These differences make Slavic languages an ideal testing ground for investigating how morphological structure is reflected in language use.

Recent developments in computational linguistics make it possible to approach such questions from a new perspective. Distributional semantic models represent words as vectors in a high-dimensional space derived from corpus data, allowing researchers to examine how grammatical categories are reflected in patterns of usage. Within this framework, morphological distinctions can be studied not only as formal categories, but also as distributional structures, observable through clustering and distances in vector space.

The present study brings together these two perspectives by examining Russian and Ukrainian nominal forms in a shared distributional space. Focusing on non-syncretic forms, which provide an unambiguous mapping between form and grammatical meaning, the analysis investigates whether case and number form coherent structures in vector space and how these structures differ across languages. By combining traditional morphological analysis with computational methods, the study aims to show how grammatical organization is reflected in distributional patterns.

2. Data Collection and Methodology

The methodological goal of this study is not to describe the Ukrainian or Russian nominal system as such, but to test how grammatical number and case distinctions are reflected in distributional semantic representations of noun word forms.

The data for this study were collected and processed in several stages. First, an initial dataset of Ukrainian noun word forms was extracted from the General Regionally Annotated Corpus of Ukrainian (GRAC, version 10) (Shvedova et al. 2017). The selection was based on GRAC's semantic tagging system (Starko 2020, 2021) (for a detailed description of the extraction procedure see (Stupak and Baayen n.d.)).

The entire dataset was morphologically annotated using the *Morphological Analyzer and Generator for Russian and Ukrainian Languages* (pymorphy2) (Korobov 2015). This

analyzer assigns all grammatically possible analyses to each word form independently of context. Due to extensive syncretism in Ukrainian noun inflection, many forms correspond to multiple case–number combinations.

All potential instances of case and number syncretism identified by the analyzer were explicitly annotated. Singular-only forms were encoded as "1" for singular and "0" for plural, while forms compatible with both numbers were marked as "1" for each value. For example, the form *posibnyky* 'tutorials, handbooks', which can occur as plural nominative, accusative, or vocative, was assigned a value of "1" for all compatible case–number combinations and "0" elsewhere. Out of the full dataset of 48,836 Ukrainian noun instances, 22,484 were classified as non-syncretic, i.e., forms that can be unambiguously assigned to a single combination of case, number, and gender.

Syncretism is a central feature of Ukrainian (Press and Pugh 2015b; Comrie et al. 1996), where multiple grammatical categories are often expressed by identical forms. However, in order to ensure a clear and interpretable mapping between morphological categories and distributional representations, the present analysis focuses exclusively on non-syncretic forms.

This restriction makes it possible to isolate the contribution of individual grammatical categories (case and number) without ambiguity. At the same time, syncretism is understood as an important structural property of the system and is taken into account in the interpretation of the results, particularly with regard to the degree of dispersion and overlap observed in the Ukrainian data.

So, the primary quantitative analysis focuses on non-syncretic forms. The non-syncretic noun set was filtered to include only those forms for which a corresponding 300-dimensional vector representation was available in the pre-trained fastText models obtained from https://embeddings.sketchengine.eu/. These vectors were trained on large-scale Ukrainian web corpora provided by Sketch Engine and are not restricted to GRAC, ensuring broad coverage of lexical items and contextual variation.

The analysis targets noun word forms and their behavior in semantic vector space. Vector space models are not explicitly trained on grammatical features (case, number, gender), but analyses of semantic vector spaces often show considerable structure by case, number, and gender, as will become apparent below.

In what follows, our analyses focus on increasingly specific aspects of the distributional structure of Ukrainian nouns. First, the full set of morphologically non-syncretic noun forms was analyzed in order to assess the global organization of the embedding space. In a second step, the dataset was restricted to nouns with both singular and plural forms in our dataset, with no restrictions on case.

In a third step, the analysis was limited further to nouns for which both singular and plural forms are attested within the same case (e.g. *stil* 'table, NOM.SG' vs. *stoly* 'tables, NOM.PL'), enabling direct comparison of number marking effects under identical case conditions. For these matched singular–plural pairs, shift vectors were computed following (Shafaei-Bajestan et al. 2022a). Each shift vector is defined as the difference between the plural and the corresponding singular embedding and represents the change in semantic space when moving from a singular to its corresponding plural while holding grammatical case constant.

To analyze the structure within the high-dimensional semantic space of the collected noun data, we employed t-Distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton 2008), an unsupervised clustering method. Important for our analyses is that t-SNE re-presents the high-dimensional word embeddings as points in a two-dimensional space, in such a way that clusters in the original space become visible in 2-D scatterplots. To interpret these clusters, color-coding based on linguistic features was added by hand to aid interpretation. We utilized the default settings of the Rtsne package in R. Our specific goal was to see if nouns used in different grammatical roles (cases) or number forms (singular/plural) would form distinct clusters in t-SNE maps, as observed in previous studies for Russian (Chuang et al. 2022) and Finnish (Nikolaev et al., n.d.).

For evaluating how well linguistic features can be predicted from embeddings, we complemented the t-SNE method with Linear Discriminant Analysis (LDA), as t-SNE may not be able to find proper clusters for features that are widely distributed in the semantic space and that explain only small proportions of the variance (see, e.g., Stupak and Baayen 2023).

Across all stages of the analysis, Genitive and Accusative noun forms show substantial overlap in the embedding space (Figures 1, 2, 3). To identify small "islands" of closely related words within this overlap in Figure 1, we applied DBSCAN—a clustering method that groups nearby points while treating isolated points as noise—specifically to the subset of word forms labeled as Genitive- or Accusative-specific (i.e., gent_non_sync = TRUE or accs_non_sync = TRUE). Clustering was performed in the two-dimensional space defined by the coordinates X1 and X2. DBSCAN detects clusters as locally dense regions: a point is assigned to a cluster if it lies within a radius eps of a core point, where a core point has at least minPts neighbors within that radius; points that do not satisfy this density criterion are treated as noise (cluster label 0). In our implementation, we set eps = 0.5 and minPts = 5, so that a cluster is formed only when at least five points occur within a neighborhood of radius 0.5 in the (X1, X2) space. The resulting cluster labels were then added back to the dataframe for subsequent inspection and visualization.

The list of Ukrainian nouns used in the subsequent analysis was constructed based on the clusters detected by DBSCAN. After applying DBSCAN to the subset of Genitive- and Accusative-specific word forms, we retained only those items that were assigned to a non-zero DBSCAN cluster. This procedure yielded a set of 1,537 Ukrainian noun forms, which served as the basis for the comparative analysis of Ukrainian and Russian noun forms in the Genitive and Accusative cases. These nouns were translated using automated tools, and each translation was subsequently manually verified to ensure correspondence with the closest Russian lexical equivalent. If a noun corresponds to a Ukrainian feminine professional title but lacks an equivalent feminine form in Russian, it was excluded from both the Ukrainian and Russian lists.

The data were morphologically annotated for Ukrainian following the same procedure, with the analysis subsequently restricted to non-syncretic forms to ensure precise measurement of grammatical case and number. Specifically, masculine animate nouns exhibiting syncretism between the Genitive and Accusative were excluded. In addition, plural forms were filtered out because, in both languages, plural nominative forms are frequently syncretic with genitive or accusative plurals, leaving no non-syncretic plural forms in the dataset after morphological annotation. Consequently, the comparative analysis includes only singular nouns, with 467 Accusative Singular (ACC.SG) forms and 238 Genitive Singular (GEN.SG) forms.

The comparative analysis relies on pre-trained fastText embeddings. A methodological challenge concerned the dimensionality of the embedding spaces: Ukrainian TenTen embeddings are 300-dimensional, whereas Russian TenTen embeddings are available only in 100 dimensions. To ensure dimensional compatibility, we therefore used 300-dimensional fastText embeddings for Russian, fastText embeddings trained on Common Crawl and Wikipedia data (Bojanowski et al. 2017).

The comparative perspective on Ukrainian and Russian introduced in Section 3 is based on previously established distributional patterns for Russian (Nikolaev et al. n.d.) and serves to contextualize the present analysis of Ukrainian data. By contrast, the analysis of Genitive–Accusative alternation in noun embeddings presented in Section 4.2 constitutes the original contribution of the study.

3. Visualizing How Ukrainian Nouns are Used: Cases and Number

The data was collected with the primary goal of determining whether nouns used in different grammatical roles (cases) or number forms (singular/plural) would form distinct groups on these maps.

3.1 Distribution of Ukrainian nouns in Case and Number

To visualize the distribution of forms in vector space, we use t-SNE (t-distributed Stochastic Neighbor Embedding), a dimensionality reduction technique that projects high-dimensional vectors onto a two-dimensional plane. Importantly, the x- and y-axes in Figure 1 do not have an independent linguistic interpretation. Instead, what matters is the relative position of points:

- points that are close together represent forms that occur in similar contexts,
- points that are far apart represent forms that occur in different contexts.

Clusters in the plot indicate that a group of forms shares similar distributional behavior. If forms belonging to the same grammatical category (e.g., case) form a tight cluster, this suggests that the category is distributionally coherent. If they are scattered, this indicates weaker coherence and greater interaction with other categories.

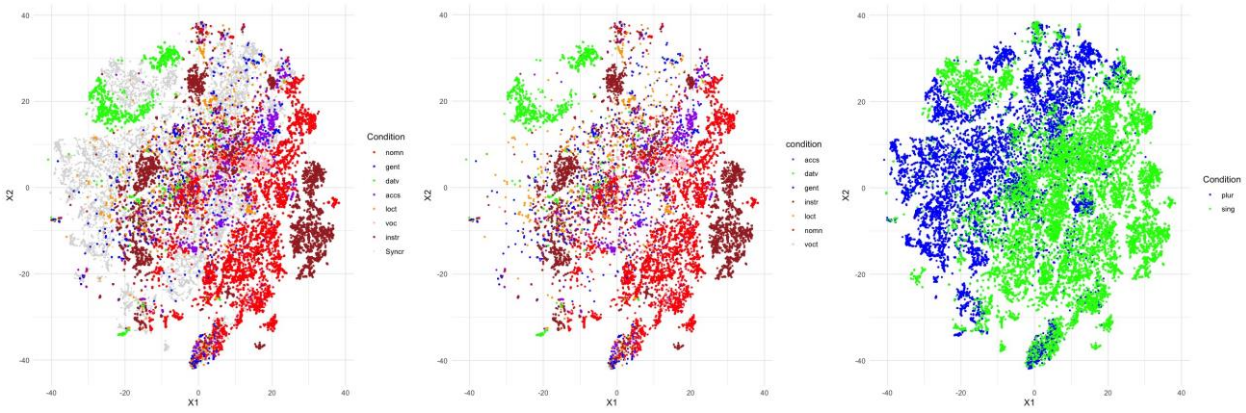


Figure 1: Morphological Landscape of Ukrainian Noun Forms: Syncretism vs. Distinct Realizations. This figure categorizes the dataset by its morphological clarity across case and number. The left panel illustrates the density of syncretic forms (in grey), where case–number realizations overlap, creating potential ambiguity in the vector space. The middle and right panels isolate non-syncretic (distinct) forms to compare the salience of different grammatical categories. A comparison between the middle panel (case distribution) and the right panel (number distribution) reveals that grammatical number provides a more discrete categorical organization than case. This structural clarity is bolstered in Ukrainian by specific morphological markers, such as those found in feminine professional nouns, which reduce syncretic "noise" and allow for more precise vector clustering of number over case

As shown in the left panel of Figure 1, widespread syncretism appears as mixed grey areas, confirming its status as an inherent structural property of the Ukrainian nominal paradigm (Press and Pugh 2015b; Comrie et al. 1996). Many forms simultaneously realize two, three, or more case values. A comparison of the left panel with the middle panel (non-syncretic forms across cases) and the right panel (non-syncretic forms across number) further indicates that syncretism is more pervasive within the domain of case than within number. This distributional pattern is consistent with the formal overlap summarized in Table 1, which shows that syncretism in Ukrainian noun inflection primarily arises from extensive

case-internal overlap of endings within number paradigms, particularly in the singular, whereas number distinctions remain comparatively more stable due to the greater formal uniformity of plural paradigms.

Because this study relies on word-based embeddings, syncretic forms are set aside for future work, and the analysis focuses on non-syncretic forms. Restricting the dataset to non-syncretic forms thus allows for a more precise isolation of the contributions of grammatical case in the middle panel and grammatical number shown in the right panel of the Figure 1.

The middle panel shows the distribution of the same noun forms colored by grammatical case. Here, only partial clustering is observed. Noun forms associated with the Instrumental (dark red), Nominative (red), and Dative (green) cases exhibit some degree of spatial concentration, whereas the remaining cases largely overlap in the central regions of the space. This overlap complicates a straightforward interpretation of case-based separation.

Importantly, the apparent weakness of case-based clustering is partly attributable to the interaction between case and number. The distribution of forms across cases is uneven with respect to number: for example, the Dative is sparsely represented in the singular, while the Nominative and Genitive are underrepresented in the plural. As a result, case distinctions are to some extent subsumed under the more prominent number-based structuring of the space.

The right panel reveals a clear large-scale structuring of the embedding space by grammatical number. Singular and plural forms of animate nouns occupy largely distinct regions of the semantic space, with relatively limited overlap. In contrast, the central area—where some overlap occurs—is predominantly occupied by inanimate nouns. While the separation is not categorical, the distribution suggests that singular and plural noun forms tend to occur in systematically different distributional contexts, which is reflected in their vector representations.

The way the semantic space of Ukrainian is structured differs from the semantic spaces of Russian (Chuang et al. 2022; Zinova et al. 2024) and Finnish (Nikolaev et al. n.d.). For these languages, forms cluster first by case, and within case, by number. The above-mentioned studies present evidence that the change from singular to plural systematically differs by case (see also Shafaei-Bajestan et al. 2022b, for conditional plurality in English).

To clarify how grammatical number and case structure the semantic space of Ukrainian, we created a subset of non-syncretic forms that contains all singular-plural pairs sharing the same case, such as *stil* ‘table’ singular Nominative vs. *stoly* ‘tables’ plural Nominative. This restriction excludes singulare tantum and plurale tantum nouns allowing us to isolate the effect of grammatical number while preserving the internal case structure. Importantly, this restriction does not substantially alter the overall organization of the semantic space observed in Figure 1.

The left panel of Figure 2 reveals a striking separation between singular and plural forms, with relatively little overlap in central regions of the embedding space. This occurs for nouns that denote quantified entities (e.g., *cina* 'price', *mil'jon* 'million') and collective nouns (e.g., *sojuz* 'unity', *pidpryjemstvo* 'enterprise'). These patterns show that semantic collectivity can partially override grammatical number, producing regions where singular and plural nouns converge.

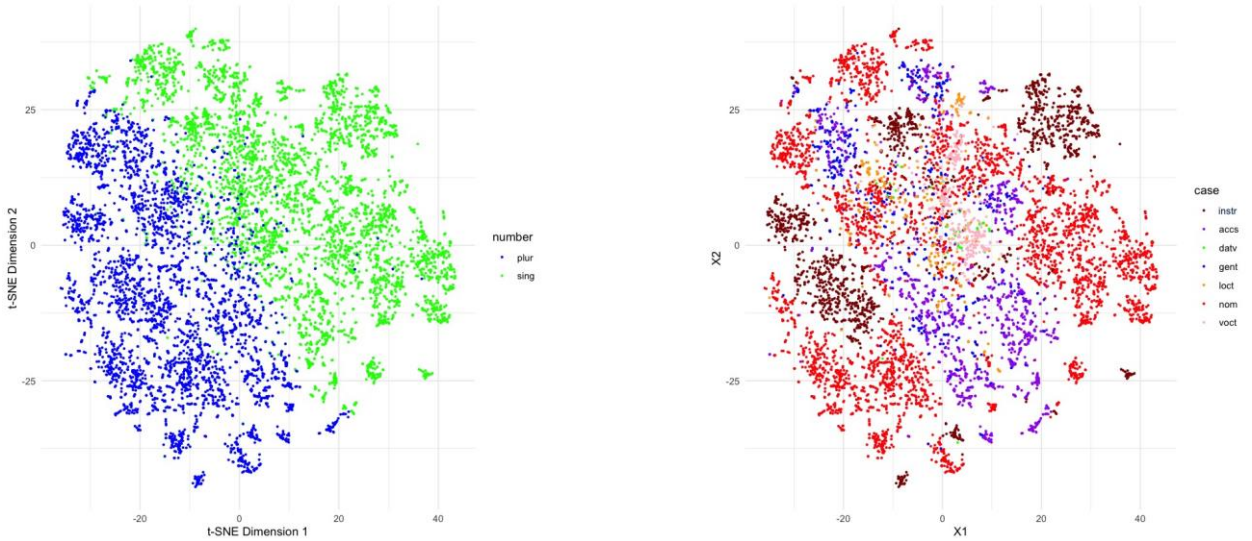


Figure 2: Distribution of Ukrainian non-syncretic nouns restricted to those with both singular and plural forms in the dataset, without case restrictions. The left panel demonstrates a clear, binary bifurcation of the vector space, confirming that grammatical number is the dominant organizing principle for Ukrainian nouns. In contrast, the right panel reveals significant intra-cluster diffusion among grammatical cases. The lack of distinct case boundaries—even in this filtered, non-syncretic dataset — empirically illustrates that case alternations (such as the Accusative-Genitive overlap) are deeply embedded in the distribution of the language, whereas the distinction between singular and plural remains a robust, primary structural divide

Within each number region, smaller clusters are present. The right panel of Figure 2 clarifies that these smaller clusters correspond to individual grammatical cases. For the Nominative (red) and Instrumental (brown) forms, several distinct clusters are present both within the region of singulars and within the region of plurals. The Genitive (blue) and Accusative (purple) forms show considerable overlap, and also form several subclusters. There are very few dative pairs (green), they cluster with the locative forms (pink) in the center of the t-SNE map.

There are also zones of partial overlap—particularly in coordinates approximately ranging from X1: –20 to 20 and X2: –5 to 30—where case boundaries are less distinct. Examples include singular–plural pairs such as *material* 'material', Nom.SING. and *materialamy* INS.PL., *podija* 'event', NOM.SING. and *podijax* LOC.PL., or *sojuza* 'unity' GEN.SING. and *sojuzamy* INS.PL. Notably, shared areas are especially pronounced between the Accusative (purple) and Genitive (blue) cases (see Section 4).

3.2 Shift Vectors for case and number

Although the overall t-SNE maps of Ukrainian, Russian and Finnish differ, it is possible that the semantics of plurality interacts with case. We therefore calculated the shift vector that, when added to the singular vector, results in the corresponding plural vector (Figure 3).

These vectors capture the direction and magnitude of change associated with a given grammatical contrast. If vectors corresponding to a particular contrast (e.g., number) are similar across many words, this suggests that the contrast is systematically encoded in the semantic space. By comparing the consistency of vectors associated with case and number, we can assess which category plays a more prominent role in structuring the distributional space.

The shift vectors in Figure 3 represent the distributional change associated with plural marking while holding case constant. This result provides further evidence for the possibility that plural inflection is not governed by a deterministic, unitary semantic operation. In addition, shift vectors tend to vary from word to word. On the one hand, this could be argued to be measurement noise in the word embeddings, and hence non-deterministic.

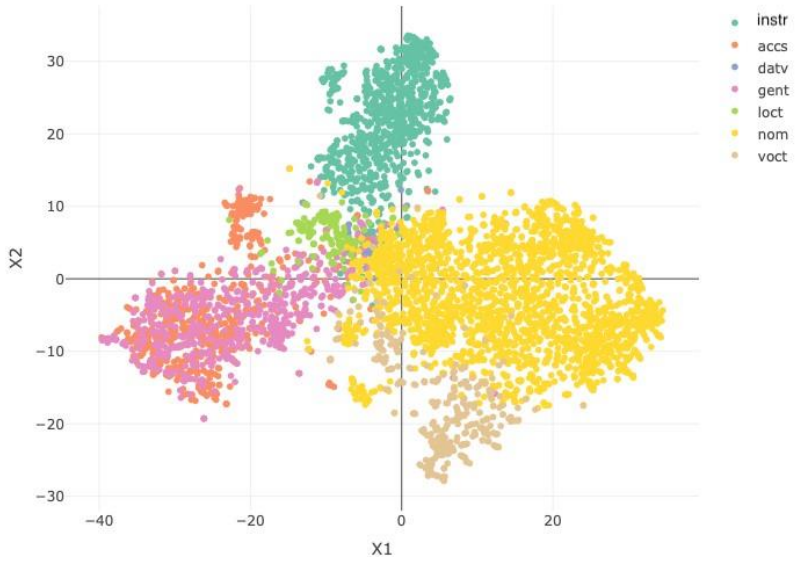


Figure 3: Distribution of Ukrainian shift vectors for number shows considerable clustering by case

Figure 3 shows that the Ukrainian shift vectors for number, when case is held constant, also show by-case clustering. However, the distribution of Ukrainian shift vectors also reveals overlapping areas near the center of the plot, and in addition particularly notable overlap between the Accusative and Genitive cases in the lower left of the t-SNE shift map. In

contrast, Russian shift vectors exhibit more overlap between the Nominative and Accusative cases (see Chuang et al. 2022; Zinova et al. 2024).

A clear difference between Ukrainian and Russian concerns the degree of overlap between the Accusative and Genitive cases: in Ukrainian, these cases show substantial convergence in the embedding space (Figure 1, 2, 3), whereas previous studies report a more distinct separation in Russian (cf. Chuang et al. 2022; Zinova et al. 2024).

3.3 Genitive-Accusative Overlap in the t-SNE map

Figure 4 visualizes the semantic vectors of morphologically non-syncretic Ukrainian nouns in the t-SNE map, with the data restricted to the datapoints in the center area of the middle t-SNE panel in Figure 1, the area where there is considerable overlap of Accusative and Genitive forms. The three panels in Figure 4 all present the same datapoints. In the left and center panels, color coding is used to distinguish between GEN (blue) and ACC (green). The center panel uses different plot symbols for the semantic classes. The right panel color-codes for animacy (black: inanimate, red: animate).

In the left panel, the genitive forms are distributed somewhat more to the upper left, and the accusative forms more to the lower right. Nevertheless, there is considerable overlap.

The center panel uses different plot symbols for the semantic classes that are available in the GRAC corpus (Starko 2020, 2021). Annotation for semantic class is incomplete in this resource, which is still under construction. As a consequence, the diversification across semantic classes has to be interpreted with caution. The classes of animals and professions are represented by many types (646 for the animals, 432 for the professions), but other classes, such as food, furniture, and plant are represented by only a handful of types (1, 2, and 11 respectively). We assigned a word to the animate class if it had at least one sense with an animate interpretation. Thus, since *čornoburka* is animate in the sense of ‘fox’, although inanimate in the sense of ‘fur’, it was categorized as animate. The center panel of Figure 4 suggests that the overlap of genitive and accusative forms in the t-SNE map is not bound to specific semantic classes. For instance, there are areas in the t-SNE map where animals (upward triangles) differentiate between case, but other areas where there is substantial overlap.

The right panel of Figure 4 presents the data points with color highlighting for animacy. Comparison of the left and right panels show that there is considerable overlap between genitive case and animacy on the one hand, and accusative case and inanimacy on the other hand. Within the set of words with genitive case, 361 are inanimate and 339 are animate. Conversely, within the set of words with accusative case, 137 are inanimate and 740 are animate ($p < 0.0001$, chi-squared test).

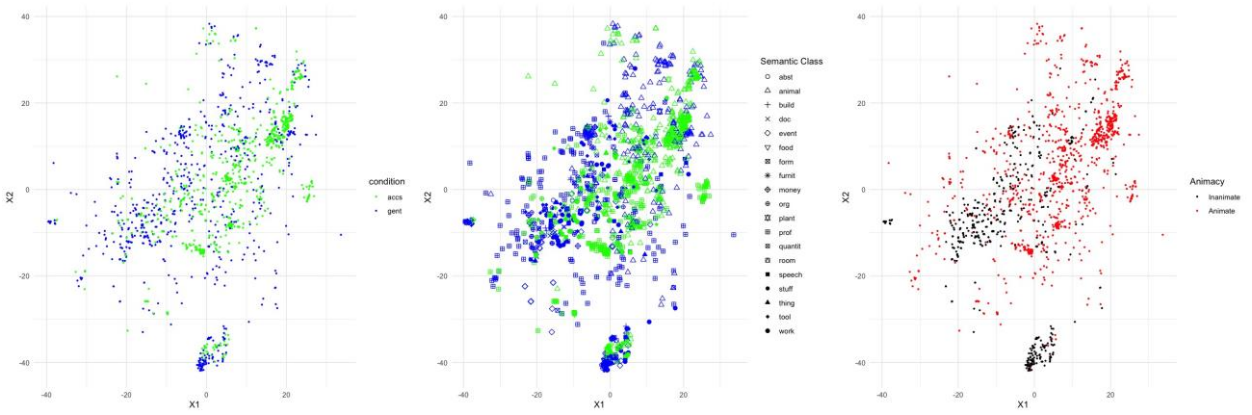


Figure 4: Distribution of Ukrainian nouns in Genitive (in blue) and Accusative (in green) cases. The left plot illustrates the distribution of Ukrainian nouns in genitive and accusative cases. The middle plot presents their distribution in Genitive and Accusative cases within semantic classes. The right panel highlights the distribution of animate and inanimate nouns within genitive and accusative cases

In summary, for the area of overlap for accusative and genitive forms in the t-SNE map Figure 4, the animate nouns tend to favor accusative case, while inanimate nouns favor genitive case. There is no reliable evidence that semantic classes are at issue, and since animacy provides only a trend, the question remains why there is considerable overlap. In the next section, we consider the role of syntactic factors.

4. The Genitive and Accusative Alternation in Ukrainian and Russian

4.1 Theoretical Overview of Genitive and Accusative Alternation in Ukrainian and Russian

Variation between these two cases has been extensively documented in descriptive and theoretical studies in Ukrainian (Bezpal'ko et al. 1962; Vyxovanec' 1987; Vyxovanec' & Horodens'ka 2004; Pljušč 2018) as well as in Russian (Timberlake 1993; Bjeljajev 1980). Although the Accusative is generally regarded as the default case for marking direct objects in affirmative clauses, the Genitive appears in a range of semantically and pragmatically conditioned environments.

One major domain of alternation is negation. Historically, both Genitive and Accusative marking of direct objects have been attested since Old East Slavic (Bezpal'ko et al. 1962). In traditional Ukrainian usage, the Genitive was strongly preferred under sentential negation. However, from the late twentieth century onward, a gradual expansion of Accusative marking in negated contexts has been observed, particularly in scientific,

technical, and colloquial registers (Chaika et al. 2024). As noted by Vyxovanec' & Horodens'ka (2004), speaker intention and communicative focus often play a decisive role. Some scholars argue that the genitive is dominant in negative constructions in Ukrainian (Pljušč 2018; Kryshevich 2010; Kolibaba 2014) and is used more widely than in Russian (Kryshevich 2010) but large-scale quantitative evidence is lacking.

By contrast, Russian Genitive in negative constructions is sensitive to definiteness, scope, aspect, verb class, and information structure (Timberlake 1993; Padučeva 2006; Brown and Franks 1995; Kagan 2012). Furthermore, finding (Krasovitsky et al. 2011) demonstrates that the relative weight of constraints on the GN in Russian has shifted diachronically: while verb aspect played a more prominent role in earlier stages of the alternation, noun abstractness has become the most influential factor in present-day Russian.

Partitive constructions provide another domain of alternation in both languages. In Ukrainian, the Genitive typically signals indefiniteness or partial affectedness, whereas the Accusative tends to mark total affectedness or referential specificity. For example, contrasts such as *ne vzjav salo* 'did not take the (whole) lard' versus *ne vzjav sala* 'did not take any lard' illustrate how case marking interacts with quantity and interpretation (Vyxovanec' 1987).

Similarly, in Russian, the Genitive is preferred in partitive contexts, particularly with mass nouns or when the object is non-specific, while the Accusative indicates a fully affected, definite entity (Timberlake 1993; Pesetsky 1982).

Importantly, in Ukrainian, the Genitive and Accusative can optionally appear with direct objects in affirmative contexts as well. For example, both ACC *Ja napysav lysta* and GEN *Ja napysav lyst* 'I wrote a letter' are acceptable, with subtle distinctions in definiteness or aspectual nuance.

In Russian, by contrast, such constructions typically require the Accusative *Ja napisal pis'mo* and the Genitive without a preposition is rarely used outside negation or partitive contexts.

Taken together, these observations suggest that while both languages exhibit Genitive–Accusative alternation, Ukrainian maintains a comparatively symmetrical and semantically diversified system, whereas Russian displays a more asymmetrical and structurally conditioned distribution. If distributional semantic representations are sensitive to such grammatical organization, we may expect these differences to manifest in the geometry of the embedding space. In particular, a structurally competitive alternation, as in Ukrainian, may yield stronger and more systematic distributional signals, whereas a more restricted and asymmetrical alternation, as in Russian, may result in a more diffuse encoding of case information.

4.2 Comparative Visualization of Ukrainian and Russian Counterparts in Accusative and Genitive Cases

Building on the theoretical overview of Genitive–Accusative alternation in Ukrainian and Russian, we turned to a direct comparison of usage patterns. In what follows, we present a comparative visualization of Ukrainian and Russian counterparts in the Accusative and Genitive cases.

Figure 5 presents t-SNE maps for Ukrainian (left panel) and Russian (center panel) singular nouns in the Genitive (red) and Accusative (green) cases.

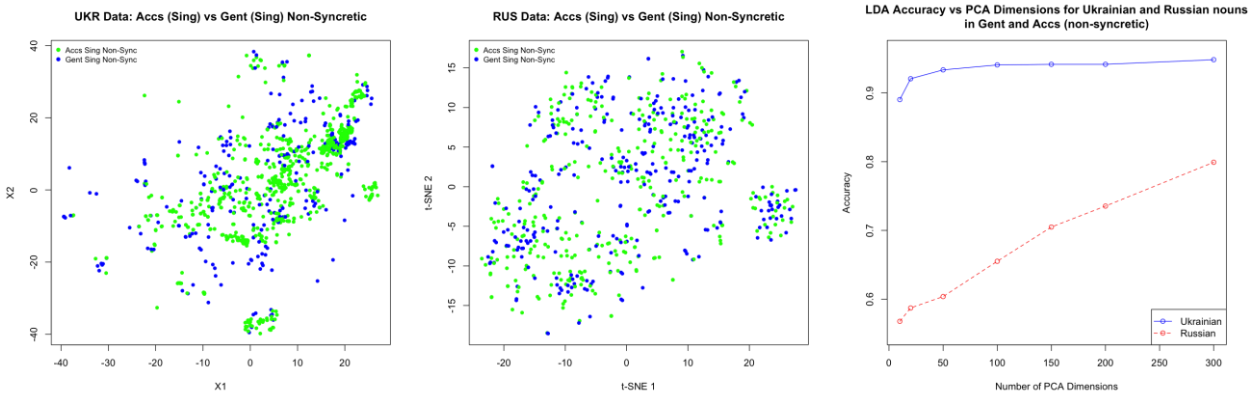


Figure 5: Distribution of Ukrainian and Russian non-syncretic translation equivalents of singular nouns in Genitive (in blue) and Accusative (in green). Left panel: Ukrainian, center panel: Russian, right panel: LDA classification accuracies for increasing numbers of principal components of the embedding spaces

In Ukrainian (Figure 5, left panel), although the two cases overlap substantially, a weak spatial tendency can be observed: Genitive forms are more often located toward the left of the semantic space, whereas Accusative forms tend to appear toward the right. In addition, Accusative forms show locally denser clusters, suggesting greater internal cohesion in their distributional profiles.

In contrast, the Russian t-SNE map shows no systematic spatial separation between Genitive and Accusative forms (Figure 5, middle panel). The two cases are heavily interspersed across the semantic space, and no consistent clustering or directional bias is visible. Moreover, Russian noun embeddings appear more broadly dispersed than their Ukrainian counterparts, indicating greater variability in contextual usage.

While t-SNE reveals some structure in the Ukrainian data, visual inspection alone does not provide a decisive basis for evaluating the separability of the two cases—particularly given that t-SNE primarily captures dominant dimensions of variation and may obscure finer-grained contrasts. To complement the t-SNE analysis, we applied Linear Discriminant Analysis (LDA), treating embedding vectors as predictors and grammatical case as the target category. Using leave-one-out cross-validation, the model achieved 95% accuracy

for Ukrainian and 80% for Russian. This suggests that case distinctions are more clearly reflected in the Ukrainian distributional space.

To assess the extent to which case is represented in higher dimensions of variation, we orthogonalized the semantic spaces using principal components analysis, and evaluated LDA classification accuracy for increasing numbers of retained principal components (10, 20, 50, 100, 150, 200, and all 300). The results are shown in the right panel of Figure 5.

As shown in the right panel of Figure 5, Ukrainian nouns reached 89% accuracy using just the first 10 principal components, whereas for Russian, classification accuracy was 57%. For Ukrainian, learning accuracy asymptotes around 100 PCA dimensions, whereas for Russian, accuracy for 100 dimensions is only 70%. This analysis shows that in Ukrainian, case is represented on dimensions that capture substantial amounts of variance, whereas for Russian, many principal components that explain only small amounts of variance are required to obtain the best possible prediction accuracy (which only reaches 80% when the full 300-dimensional space is taken into account). This indicates a more diffuse distribution of case information in Russian as compared to Ukrainian.

5. General discussion

This study integrates quantitative classification, dimensionality reduction, and visual exploration to examine how grammatical number and case are encoded in distributional embeddings of Ukrainian nouns. Our primary aim was to analyze Ukrainian noun forms across number and case and to compare these patterns with those observed in Russian noun embeddings (Chuang et al. 2022; Zinova et al. 2024). In addition, we examined genitive and accusative forms of identical lexical nouns in both languages to assess how case alternation manifests in embedding space.

A key insight from the t-SNE visualization of Ukrainian noun embeddings is the prominent separation by grammatical number (singular vs. plural), which stands in contrast to previous findings for Russian (Chuang et al. 2022; Zinova et al. 2024), where case emerged as the primary organizing principle. In Ukrainian, case forms secondary sub-clusters within number-based regions in semantic space. The dominance of number in the t-SNE map suggests that number constitutes a more stable and salient organizing dimension in Ukrainian embeddings than case. In Russian, by contrast, distributional patterns captured by the embeddings indicate a stronger role for case and a comparatively weaker contribution of number.

We suggest that this structural clarity is a "distributional footprint" of the broader morphological system. The productive use of feminine professional titles in Ukrainian (e.g., *inspektorka*) creates a high volume of singular and plural non-syncretic data in the training

corpus. Furthermore, the normative alignment of paucal constructions (2, 3, 4) with the Nominative Plural in Ukrainian acts as a categorical reinforcer. By contrast, the Russian requirement for the Genitive Singular in these same contexts creates a "distributional drag"—a latent signal that pulls plural contexts toward the singular-case cluster, reinforcing Case as the dominant organizing signal even for non-syncretic lexemes.

Surprisingly, however, not only in Russian but also in Ukrainian, the shift vectors associated with plural formation cluster by case. Unlike in Russian, the Ukrainian shift-vector clusters show considerable overlap between Genitive and Accusative forms. This pattern aligns with the observed overlap of Genitive and Accusative noun embeddings in Ukrainian more generally and is consistent with well-documented accounts of Genitive–Accusative alternation in Ukrainian grammar, particularly in contexts involving negation, partitivity, and the increasingly productive use of the Genitive as a “second Accusative”.

Notably, this overlap is not restricted to specific semantic noun classes but is broadly distributed across the lexicon, suggesting that it reflects a systemic property of Ukrainian case usage rather than isolated lexical effects. The observed distribution is likely shaped by the interaction of inherited prescriptive norms, more recent re-evaluations of traditional Ukrainian usage patterns, and fluid contemporary language practice.

A direct cross-linguistic comparison of Genitive and Accusative noun embeddings in Ukrainian and their Russian translational equivalents revealed further differences. While t-SNE visualizations show considerable intermingling of the two cases in both languages, Linear Discriminant Analysis (LDA) yields substantially higher classification accuracy for Ukrainian (95%) than for Russian (80%). The fact that Ukrainian reaches peak accuracy with significantly fewer principal components (25 vs. over 300 in Russian) indicates that the semantic and syntactic boundaries of these cases are more compactly encoded in Ukrainian. While both languages exhibit an expansion of the Genitive into traditional Accusative contexts (Vyxovanec’, I. R. & Horodens’ka 2004; Bjeljajev 1980), the Ukrainian system appears to maintain sharper functional differentiation.

These results suggest that, despite somewhat similar surface-level visual overlap in t-SNE plots, the semantic and syntactic distinctions between the Genitive and Accusative cases are more clearly and compactly represented in Ukrainian embeddings. This observation resonates with grammatical analyses noting key differences in the nature of Genitive–Accusative alternation across the two languages. While both exhibit an expansion of Genitive usage into contexts traditionally associated with the Accusative (Vyxovanec’, I. R. & Horodens’ka 2004; Bjeljajev 1980), Russian imposes stronger constraints — for instance, restricting Genitive usage to abstract or idiomatic expressions (Čujkova 2021). These constraints may be leading to greater contextual overlap and more diffuse encoding of case distinctions in Russian embeddings. In contrast, Ukrainian appears to exhibit

sharper functional differentiation, which is reflected in the greater discriminability of its case forms in the embedding space.

Taken together, these observations suggest that the differences observed in vector space are not arbitrary, but reflect deeper properties of the morphological systems. Rather than being determined solely by abstract grammatical categories, the distributional organization of case and number appears to be shaped by the interaction of syncretism, paradigmatic structure, and usage patterns. This perspective supports a dynamic view of grammatical categories as emerging from the interplay of form, function, and frequency in actual language use, rather than as fixed, pre-defined entities.

References

Arkadiev, P. M. 2008. The Russian "new vocative": Synchrony and diachrony. *Russian Linguistics* 32(1). 1–22. https://doi.org/10.1007/s11185-008-9020-1

Bjeljajev, Ju. I. 1980. Rodovyj či znaxidnyj? *Kul'tura Slova* 19. 80–83.

Bezpal'ko, O. P., Bojčuk, M. K., & M. A. Žovtobrjux. 1962. *Istoryčna hramatyka ukrajins'koji movy*. Kyiv: Naukova dumka.

Bezpojasko, O. K. 1991. *Imenni hramatyčni katehoriji (funkcional'nyj analiz)*. Kyiv: Naukova dumka.

Bojanowski, P., Grave, E., Joulin, A. & T. Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5. 135–146. https://doi.org/10.1162/tacl_a_00051.

Breu, W. 2020. Paucal. In Marc L. Greenberg (ed.), *Encyclopedia of Slavic Languages and Linguistics Online*. Leiden: Brill. https://doi.org/10.1163/2589-6229_ESLO_COM_031985.

Brown, S. and S. Franks. 1995. Asymmetries in the scope of Russian negation. *Journal of Slavic Linguistics* 3(2). 239–287.

Chaika, L., Lehka, N., Tamm, A. & N. Vaiss. 2024. Ukrainian Aspect and Object Case in ukTenTen. The Partitive Genitive of Perfective Verbs and Mass Nouns. In E. Glaser, P. Sleeman, T. Strobel, & A. Tamm (eds.), *Partitive Constructions and Partitive Elements within and across Language Borders in Europe*. Venice: Edizioni Ca' Foscari, 173–224. https://doi.org/10.30687/978-88-6969-795-1/008.

Chuang, Y., Brown, D., Evans, R. & R. H. Baayen. 2022. Paradigm gaps are associated with weird "distributional semantics" properties: Russian defective nouns and their case and number paradigms. *The Mental Lexicon* 17(3). 395–421. https://doi.org/10.1075/ml.22013.chu.

Comrie, B., Stone, G. & M. Polinsky. 1996. *The Russian Language in the Twentieth Century*. Oxford: Oxford University Press.

Corbett, G. G. 1983. Resolution rules: Agreement in person, number, and gender. In Gazdar, G., Pullum, G. K. & I.A. Sag (eds.), *Order, Concord, and Constituency*, 175–206. Dordrecht: Foris.

Corbett, G. G. 1991. *Gender*. Cambridge: Cambridge University Press.

Čujkova, O. 2021. K voprosu o (ne)sočetaemosti roditel'nogo partitivnogo i nesoveršennogo vida v russkom jazyke: korpusnoe issledovanie. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue"* 20, 162–178, Moscow.

Franks, S. & R. House. 1982. Genitive of theme in Russian. In *Proceedings of CLS 18*, 156–168.

Horpynyč, V. O. 2004. *Morfologija ukrajinskoji movy: pidručnyk*. Kyiv: Akademija.

Isačenko, A. V. 1960. *Grammatičeskij stroj russkogo jazyka v sopostavlenii so slovackim*. Bratislava: Slovenská akadémia vied.

Kagan, O. 2012. *Semantics of Genitive Objects in Russian: A Study of Genitive of Negation and Intentional Genitive Case*. Dordrecht: Springer.

Kirey-Sitnikova, Y. 2021. Prospects and challenges of gender-fair language in Russian. *Russian Linguistics* 45(2). 123–145. https://doi.org/10.1007/s11185-021-09241-6.

Kolibaba, L. 2014. Variability of case endings of nouns against the background of morphological norms of the ukrainian literary language at the end of the 20th – beginning of the 21st century. *Ukrainian Language* 1. 86–96. Available at: http://nbuv.gov.ua/UJRN/Ukrm_2014_1_10.

Korobov, M. 2015. Morphological analyzer and generator for Russian and Ukrainian languages. In M. Yu. Khachay, N. Konstantinova, A. Panchenko, D. I. Ignatov, & V. G. Labunets (eds.), *Analysis of Images, Social Networks and Texts: 4th International Conference, AIST 2015*, 320–332. Cham: Springer. https://doi.org/10.1007/978-3-319-26123-2_31.

Krasovitsky, A., Baerman, M., Brown, D. & G. G. Corbett. 2011. Changing semantic factors in case selection: Russian evidence from the last two centuries. *Morphology* 21(2). 573–592. https://doi.org/10.1007/s11525-010-9177-y.

Kryshevich, Olesya. 2010. *The genitive of negation in Ukrainian.* M.Phil. thesis. University of Cambridge.

Van der Maaten, L. & G. Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research* 9. 2579–2605. https://doi.org/10.5555/2623330.2623334.

Nesset, T., Piperski, A. & S. Sokolova. 2022. Russian feminitives: what can corpus data tell us? *Russian Linguistics* 46. 95–113. https://doi.org/10.1007/s11185-022-09253-w.

Nikolaev, A., Chuang, Y.-Y. & R.H. Baayen. n.d. A generating model for finnish nominal inflection using distributional semantics. Manuscript under review. *The Mental Lexicon*.

Padučeva, E. V. 1985. *Vyskazyvanie i ego sootnesennost' s dejstvitel'nost'ju*. Moscow: Nauka.

Padučeva, E. V. 2004. *Dinamičeskaja semantika*. Moscow: Jazyki slavjanskoj kul'tury.

Padučeva, E. V. 2006. Genitiv dopolnenija v otricatel'nom predloženii. *Voprosy jazykoznanija* 6. 21–43.
Pesetsky, D. M. 1982. *Paths and categories*. Ph.D. dissertation. Massachusetts Institute of Technology.
Pljušč, M. J. 2018. *Katehoriji sub''ekta i ob''kta v strukturi prostoho rečennja*. Kyiv: Vydavnyctvo NPU imeni M.P. Drahomanova.
Plungian, V. A. 2011. *Vvedenie v grammatičeskuju semantiku*. Moscow: RGGU.
Press, I. & S. Pugh. 2015a. *Ukrainian: A comprehensive grammar*. London: Routledge.
Press, I. & S. Pugh. 2015b. *Ukrainian: A Comprehensive Grammar*. London: Routledge, 2nd edn.
Shafaei-Bajestan, E., Moradipour-Tari, M., Uhrig, P. & R.H. Baayen. 2022a. Semantic properties of English nominal pluralization: Insights from word embeddings. *arXiv.* arxiv.org/abs/2203.15424v1.
Shafaei-Bajestan, E., Uhrig, P. & R.H. Baayen. 2022b. Making sense of spoken plurals. *The Mental Lexicon* 17(3). 337–367.
Shevelov, George Y. 1993. Ukrainian. In Bernard Comrie & Greville G. Corbett (eds.), *The Slavonic Languages*, 947–998. London: Routledge.
Shvedova, M. & A. Lukashevskyi. 2025. Case choice in Ukrainian vocative expressions: A study of parliamentary transcripts (1990–2024) annotated with universal dependencies. In *Grammar and Corpora: 10th International Conference, 26–28 June 2025. Book of Abstracts*, 106–108, Riga: University of Latvia Press.
Shvedova, M., von Waldenfels, R., Yarygin, S., Rysin, A., Starko, V., Nikolajenko, T., et al. 2017. *GRAC: General Regionally Annotated Corpus of Ukrainian. Electronic resource*. Kyiv, Lviv, Jena. Available at https://uacorpus.org/en
Starko, V. 2020. Semantic Annotation for Ukrainian: Categorization Scheme, Principles, and Tools. In *Proceedings of the Conference on Semantic Annotation*.
Starko, V. 2021. Implementing Semantic Annotation in a Ukrainian Corpus. In *Proceedings of the Conference on Semantic Annotation*.
Stupak, I. & R.H. Baayen. 2023. An inquiry into the semantic transparency and productivity of German particle verbs and derivational affixation. *The Mental Lexicon* 17(3). 422–457. https://doi.org/10.1075/ml.22012.stu.
Stupak, I. V. & R.H. Baayen. n.d. How Much Semantics Does Grammatical Number Carry in Ukrainian: A Distributional Semantic Approach. Manuscript under review. *Slavistic Studies Today*.
Švedova, N. J., Lopatin, V. V., Kuz'mina, N. J. et al. 1980. *Russkaja grammatika*. Vol. 1. Moscow: Nauka.
Testelec, Ja. G. 2001. The structure of Russian clause. In *The Languages of the World*. Munich: LINCOM Europa.
Timberlake, A. 1993. Russian. In B. Comrie & G. G. Corbett (eds.), *The Slavonic Languages*, 827–886. London: Routledge.

Titov, E. 2016. Number agreement mismatches in Russian numeral phrases. In Lenertová, D., Meyer, R., Šimík, R., and Szucsich, L. (eds.), *Advances in Formal Slavic Linguistics,* 401–426. Berlin: Language Science Press.

Tyščenko, K. 2000. *Metateorija movoznavstva*. Kyiv: Osnovy.

Vinogradov, V. V. 1952. *Russkij jazyk. Grammatičeskoe učenie o slove*. Moscow: Učpedgiz.

Vyxovanec', I. R. 1987. *Systema vidminkiv ukrajins'koji movy*. Kyiv: Naukova dumka.

Vyxovanec', I. R. & K.H. Horodens'ka. 2004. *Teoretyčna morfolohija ukrajins'koji movy: Akademična hramatyka ukrajins'koji movy*. Kyiv: Pulsary.

Zinova, Y., Yablokova, A., & R. van de Vijver. 2024. Interaction of Semantics and Morphology in Russian Word Vectors. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, 120–128.