Stupak, Inna V., and R. Harald Baayen

How Much Semantics Does Grammatical Number Carry in Ukrainian: A Distributional Semantic Approach

Abstract

This study explores the semantic depth of grammatical number in Ukrainian through a distributional semantic approach. In Ukrainian, the number category is considered a morphological category with a semantic dominance (Vykhovanecj, 2004). This view suggests that some forms are semantically motivated, while others express number only formally. By utilizing the semantically annotated General Regionally Annotated Corpus of Ukrainian (GRAC) and word embeddings from the TenTen Corpus for Ukrainian, this research analyzes the distributional representations of grammatical number to uncover its semantic nuances. The observed patterns in the usage of singular and plural forms in Ukrainian nouns align with the heterogenity of the number category. As such, the distribution of singular and plural forms does not strictly follow grammatical patterns. Instead, semantic factors play a significant role in determining the choice between singular and plural forms in Ukrainian nouns. In Ukrainian, a language with a rich morphological system, the semantics of grammatical number extends beyond simple quantification, touching on nuances of meaning, cultural usage, and syntactic dependencies.

Key words: Ukrainian, word2Vec, corpus study, grammatical number, semantics, t-SNE, DBSCAN, LDA.

Introduction

The category of number in Ukrainian reflects a multifaceted linguistic phenomenon, approached from diverse perspectives by scholars. This aspects gain even greater significance, particular given that the linguistic scholars perceive the category of number in Ukrainian as a classificational (Zaliznjak, 2022; Miloslavskyj, 1986), derivational (Vinogradov, 1947; Bulakhovsky, 1948) and morphological category with a semantic core (Vykhovanecj, 2004; Bezpojasko, 1991; Bondarko, 2022).The latter approach suggests that while certain forms are semantically motivated, others serve purely grammatical functions. The interaction between grammatical number and contextual usage in Ukrainian often results in semantic differentiation, where singular and plural forms diverge in meaning. For instance, the singular noun *spyrt* [alcohol] refers to a substance, whereas its plural counterpart *spyrty* denotes various types of the substance.

In Ukrainian, number can be expressed through various means:

- **morphologically:** formal marking of nouns, pronouns, and associated words to indicate whether they are singular (*brat* [brother]) or plural (*braty* [brothers]). Singularia tantum (*moloko* [milk]) are both morphologically and semantically singular, whereas pluralia tantum (*nozhytsi* [scissors]) are semantically singular but morphologically plural.
- **lexically:** worda can have intrinsic plural semantics and at the same time can be used with both singular and plural morphological number. For instance, *kompleks* [complex] denotes a group of buildings, and several group of buildings can be refer to complexes in the plural. Similar observations hold for collective nouns such as

gorodyna [garden vegetables]. Lexical plurals have existed in the Ukrainian language throughout its history and in modern Ukrainian are limited to specific semantic categories, which are contuniously attracting new members (Vykhovanecj, 2004).

- **contextually:** the interpretation of number is often influenced by the surrounding context, and the forms used need not strictly follow the morphological norms:
    1. Na te j ščuka, ščob karas ne drimav [That's why the pike exists, so that the carp doesn't sleep.]
    2. U nas hosti – mij brat pryjixav [We have guests – my brother has arrived.]

Depending on the context the plural form can take on the meaning of the singular (as in sentence 2), and conversely, the singular form, morphologically denoting a single entity, can actually refer to sets with multiple members (as in sentence 1). In sentence (1) the singular forms of *ščuka* [pike] and *karas'* [carp] convey the generalized collective meaning of the plural. On the other hand, in the sentence (2) the plural form of *gosti* [guests] is used to refer to one person *brat* [brat]. These examples illustrate how the semantics of morphological number in Ukrainian can vary. Depending on semantics and context singular and plural forms can be interchangeable.

Recently, Shafaei-Bajestan et al. (2022) show that the meaning of English plural nouns can also be shaped by the semantic classes to which these nouns belong. The shift in semantic space from singular to plural is conditional on the meaning of the base word. However, for Finnish (Nikolaev et al., 2022) and Russian (Chuang et al., 2022), plural shift vectors vary systematically with case. When the semantics of inflected words are approximated using high-dimensional vectors from distributional semantics, the representation of plurality emerges as significantly richer than what one might expect from basic features like plural inflection alone.

The focus of the present study is on the distributional semantics of Ukrainian nouns, with the aim of unraveling the interplay between the semantic nuances of these nouns and the corresponding morphological variation. More specifically, our study addresses the following research questions: *Are the different uses of singular and plural nouns as described in standard grammars reflected in embeddings ? Can semantic vectors for Ukrainian capture the morphosemantic characteristics of the number category?*

1.2. Data Collection

In a first step of the pipeline, the corpus data used in this study is taken from the General Regionally Annotated Corpus of Ukrainian (GRAC, version 16) (Shvedova et al., 2017). By employing semantic tags as outlined in the works (Starko, 2020, 2021), our study encompassed the examination of 82,151 nouns across 21 semantic classes. The dataset was cross-checked using embeddings from the TenTen Corpus Family (https://embeddings.sketchengine.eu/static/models/uktenten20_rft3.lemma.vec).

The obtained data was analyzed using the Morphological Analyzer and Generator for Russian and Ukrainian Languages (pymorphy2) (Korobov, 2015) to determine the number, case, gender, and animacy/inanimacy of each noun. The morphological analysis revealed that 27.0% of the forms were non-syncretic, while 73.0% exhibited

syncretism. For instance, the form *šašeli* [a beetle or beetles] can appear in the singular in nominative, genitive, and locative cases, and in the plural in nominative and vocative.

Given the widespread syncretism in Ukrainian noun inflection, all instances of syncretism were explicitly identified. Non-syncretic noun forms in the singular were marked as '1', while those in the plural were marked as '0.' For syncretic forms that corresponded to both singular and plural, both singular and plural were marked as '1' for each.

The data obtained was processed using t-SNE (t-Distributed Stochastic Neighbor Embedding, Maaten and Hinton, 2008), an unsupervised clustering method commonly used in data visualization. The primary goal of t-SNE is to capture the relationships between data points by assessing their similarities or dissimilarities. In this study, we employed t-SNE to visualize potential clusters within the high-dimensional semantic space and used DBSCAN to identify meaningful groups, as well as to extract and analyze nouns lying near cluster boundaries. For the clustering analysis, we utilized the default settings of the Rtsne package in R. Additionally, we implemented Linear Discriminant Analysis (LDA) analysis to assess the accuracy of number prediction for both non-syncretic and syncretic forms.

1.3. Exploring Clustering Patterns in Ukrainian Nouns Using t-SNE: Insights from Number, Semantics, Case and Animacy

We conducted three analyses. First, we applied t-SNE (Maaten & Hinton, 2008) to a large set of nouns, including many low-frequency words, and observed some clustering by semantic class, some clustering by case, gender, animacy/inanimacy and a more global but probabilistic separation of singular and plural forms (Subsection 1.3.1). Second, we narrowed our dataset down to those nouns that have a singular and a corresponding plural form in their semantic class, that helps us to exclude singularia tantum, pluralia tantum and those that do not have their number counterpart in our data. As a result the distribution of Ukrainian nouns across the paradigmatic features reveals clear separation especially in number. Third, following Chuang et al. (2023) we considered the Ukrainian shift vectors for number (Subsection 1.3.2).

1.3.1. t-SNE analysis.

First, we applied t-SNE (Maaten & Hinton, 2008) to a large set of nouns, including many low-frequency words (Figure 1, top panel). The top-right plot clearly illustrates the systematic nature of case syncretism in Ukrainian nouns, with syncretic forms, marked in purple, distributed across the plot. Case-syncretic forms account for 27.8% of the entire dataset of Ukrainian nouns. Nouns exhibiting number syncretism (top-left plot) comprise approximately 5.2% of the analyzed data. The most common occurrence is case-number syncretism, which appears in 40.0% of cases.

Given that syncretism prevails in the majority of our data, concerns naturally arise regarding its potential influence on the results. To avoid this, we refined our dataset, concentrating only on non-syncretic forms (Figure1, bottom plots).

Upon examination, the left plot reveals a more global but probabilistic separation of singular (green) and plural (blue) forms. However, we observe some overlap between singular and plural forms, though it is less pronounced than when considering all nouns, including syncretic forms (Figure 1, top left plot). Interestingly, the distribution pattern of Ukrainian nouns in both singular and plural forms closely resembles that of Maltese nouns for singular and sound plurals (Nieder et al., 2023).

Overall, the clustering patterns of Ukrainian nouns vary depending on semantic classes and within each semantic class across number (Figure 1, bottom middle panel). The yellow class pertains to "animal" (33.1% of all data), and the pink class to "profession" (59.3%). These are the two classes that have been extensively annotated in GRAC, which explains why 19 of the semantic classes are underrepresented.

The two lower left plots of Figure 1 illustrates the distribution of Ukrainian nouns by number within each of 21 semantic classes, where singular forms are marked in green and plural forms are marked in blue.

Returning to the lower left and center panels of Figure 1, there are areas close to the general border between singulars and plurals where singulars and plurals are found together. This overlap may stem from the bivalent morphological representation of grammatical number following paucal numerals. In Ukrainian, morphological differentiation for paucal numbers occurs in specific genitive singular and nominative plural forms when preceded by numerals such as 2, 3, and 4. In the genitive singular, this differentiation is particularly evident with neuter nouns of the 4th declension type and those that lose their *-yn* suffix during pluralization (e.g., *2, 3, 4 gromad'janyna* [2, 3, 4 citizen] (Gorpynych, 2004). Semantically, nouns of the 4th declension often denote diminutives (e.g., *košen'ja* [kitten]) or collectives (e.g., *mal'ja* [baby], referring collectively to girls, boys, or both).

A similar pattern is observed in the semantic class of "money," where nouns like *gryvn'ja* [UA currency] are used in the genitive singular after paucal numerals. The alternation between genitive singular and nominative plural forms underscores the morphological flexibility of Ukrainian nouns and likely contributes to the observed fuzzy border between singulars and plurals in the bottom left panel of Figure 1.

Nouns belonging to the semantic classes of "tool," "work," "organization," "stuff," and "speech" also tend to cluster in the fuzzy border area between singulars and plurals. This clustering can be attributed to the fact that many of these nouns represent collective, material, or abstract concepts. For instance, *Berkut* [the name of a military organization] from the "organization" class and *zbroja* [weapon] from the "tool" class inherently possess plural-like characteristics, as they denote groups or multiple entities by default. These inherent plural tendencies influence their semantic and grammatical behavior, contributing to their alignment with plural forms.

By their nature, these nouns express groups, collections, or multiples, inherently carrying a plural meaning. Consequently, their paradigms are incomplete due to subject-logical content they represent (Vykhovanecj, 2004).

The bottom right plot (Figure 1) shows how the case forms cluster in the tSNE space. Whereas Chuang et al. (2023) and Nikolaev et al. (2023) observed clusters by case, and within case, clustering by number, for the present Ukrainian data, we observe primary differentiation by number, and within the number regions further differentiation by case. Nevertheless, when the shift vectors for the non-syncratic singulars and plurals

are calculated, they show strong clustering by case, just as in Russian and Finnish (see Figure 2).

1.3.2. Shift vectors for number.

Following Shafaei-Bajestan et al. (2024), we calculated shift vectors for plural number by subtracting the singular vector from the plural vector for a given case. The clustering of shift vectors in Figure 2 (left plot) shows that there is considerable overlap between the shift vectors for the genitive and the accusative cases, although both cases also have areas where there is little overlap. On the one hand, both cases share the common role of indicating relationships between nouns: semantic category of quality, causality, direct object, and partitive. However, genitive and accusative differ in their primary functions, so accusative designates the subject toward the action is directed, the direction and time. The genitive case serves the function of indicating the possession or its absence, quantity, and metaphorical location in reference to people (e.g. *Ja u doktora* [I am at the doctor's office]). It seems likely that it is these similarities and diferences in semantics that drive the patterning of the two cases in Figure 2 (left plot).

We also inspected whether there is some clustering of shift vectors by semantic class (Figure 2, right plot). To do so, we compiled 5,420 nouns with non-syncretic singular and plural forms across 21 semantic classes. A t-SNE projection of these shift vectors onto a two-dimensional space reveals distinct some semantic clustering, although less extensively compared to what was documented for English nouns by Shafaei-Bajestan et al. (2024), as shown in Figure 2 (right plot), which presents the shift vectors for number for a semantic class. Shift vectors for professions (pink) and animals (yellow), the most numerous classes in our data, are partly disjunct.

As Corbett's hierarchy of animacy (2000), assigns greater importance to pluralization in human and animate nouns, while attribution less importance to plural marking in inanimate nouns, the fact that we observe some differentiation in the shift vectors for words for animate beings (professions and animals) suggests that in Ukrainian, shift vectors differentiate to some extent at least for the core semantic classes for plurality. Unfortunately, the small numbers of words belonging to other semantic classes, due to the semantic database still being under development, make it impossible to provide solid evidence for this possibility: Data for the other semantic classes are sparse, and only a few small clusters distinct from those of the professions and animals are visible, e.g., for speech nouns and for work nouns.

The general lack of clear clustering observed for many semantic types likely arises from shared conceptual functions, when nouns refer to abstract or collective ideas rather than discrete, countable entities. For instance, nouns such as *doslidzhenn'ja* [research] (from the documents class) or *instytut* [institute] (from the organizations class) inherently convey multiplicity or collective action, blurring the distinction between singular and plural forms.

1.3.3. Linear discriminant analysis.

We used linear discriminant analysis (LDA) to assess to what extent number can be predicted from the embeddings. These analyses are carried out on datasets with non-

syncretic forms only. Accuracies are reported for models set up with leave-one-out cross-validation. The LDA analyses are important in that they complement the t-SNE analyses, which due to having to condense a high-dimensional space into a 2-dimensional space, may not be able to show more subtle similarity structure (see also Stupak & Baayen, 2023).

1.3.3.1. LDA analysis for number. The accuracy for prediction of non-syncretic forms focuses on the proportion of correctly classified non-syncretic singular and plural forms. The overall accuracy (84.5 %) is obtained by the ratio of the total count of correctly classified instances to the total number of instances in that category. The confusion matrix can also be used to assess accuracy separately for singulars (93.0%) and plurals (76.0%).

We also investigated prediction accuracy for number within the set of animate nouns as well as within the set of inanimate nouns. For the animate nouns accuracy for singulars was 95.0% and accuracy for plurals was 86.0%. For the inanimate nouns accuracy for singular was 93.1%, and for plurals 84.1%. Thus, accuracy decreases by two percent when changing from animate to inanimate nouns, irrespective of number.

Finally, we inspected classificaton accuracy for the nouns of the profession semantic class (63.7% and 67.6%, respectively) and for the nouns in the animal class (28.8% and 25.1%, respectively). Overall, the classification accuracy was higher for the profession nouns than for the animal nouns. This difference could be due to words for animals being used more often in the singular even when denoting plural ensembles (as in (1)).

1.3.3.2. LDA analysis of shift vectors. We also used LDA to investigate whether it is possible to predict the semantic class of a noun from its number shift vector. Of the nouns in the profession class, 92.2% were correctly classified. For the nouns of the animal class, 70.3% were correctly classified. This analysis fits well with the tSNE clustering analysis in that it shows that the semantics of pluralisation vary systematically not only with case but also with semantic class, thus offering a partial replication of the results reported for English by Shafaei-Bajestan et al. (2024).

The animal class demonstrates significantly lower classification accuracy compared to the profession class. The nouns in the animal category were frequently misclassified not only as profession nouns, but also as memebers of minority classes such as those labeled as money nouns, quantity nouns, tool nouns or work nouns. Our hypothesis is that the animal nouns may enjoy broader metaphorical usage, as in the following examples:

(3) *Akula biznesu* [Business shark'] → denotes a ruthless, ambitious entrepreneur.
(4) *Kury grošej ne kl'ujut'* [The chickens don't peck the money] → describes someone is extremely rich;
(5) *Yak oseledci v bočci* [Like herrings in a barrel] → refers to overcrowding or excessive quantity.

1.4. General Discussion

The patterns observed in the usage of singular and plural forms among Ukrainian nouns reflect the complexity and heterogenity of the number category in this language. Our results indicate that the choice between singular and plural forms is not only governed by classical grammatical rules, but is also by case and by words' semantic class.

However, in the tSNE plots there are areas where singulars and plurals overlap, even when case and semantic class are taken into account. One reason that this overlap exists may be that some of the grammatical uses of number are not visible to analysis that makes use of form specific embeddings. Words such as *kompleks* [complex] or *pidpryjemstvo* [enterprise] can denote plurality even though the form is that of the singular. At the same time, the plural of for instance *kompleksy* [complexes] emphasizes the discreteness and distinction between complexes, whereas the singular *kompleks* [complex] retains a more generalized meaning, capable of referring to either a group of buildings or a single system as a whole:

(6) Ce sučasnyj medyčnyj kompleks [This is a modern medical complex].

(7) Ci kompleksy majut' rizne pryznačenn'ja [These complexes have different purposes].

Such subtle differences are likely beyond the scope of word specific embeddings such as word2vec or fasttext.

Another reason for the observed overlaps between singulars and plurals may be the use of paucal number in Ukrainian. Nouns preceded by numerals 2, 3, or 4 are realised with singular number with the 4th declension class, and even though notionally they represent plurals. Here too we may be running into the limitations that come with words specific embeddings that are unaware of context of use.

For future research we plan to make use of contextualised embeddings such as those developed by Peters et al. (2018) and Devlin et al. (2019). These models can be used to generate word embeddings that are dynamically adjusted to the contexts in which word tokens are used.

The hope is that these contextualised embeddings will be able to capture the subtle differences in use of singulars and plurals of *kompleks* [complex] and of paucal number.

Acknowledgements

This project has received funding through the MSCA4Ukraine project, which is funded 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.

REFERENCES

Baayen, R. Harald, Christina Burani, and Robert Schreuder. 'Effects of Semantic Markedness in the Processing of Regular Nominal Singulars and Plurals in Italian'. *Yearbook of Morphology*, (1996), 13–33.

Bybee, Joan. *Morphology: A Study of the Relation Between Meaning and Form*. (Amsterdam: John Benjamins, 1985).

Chierchia, Gennaro. *Mass vs. Count: Where Do We Stand? Outline of a Theory of Semantic Variation*. (Cambridge, UK: Cambridge University Press, 2021).

Chuang, Yu-Yin, Brown, David, Evans, Roger, and Baayen, R. Harald. 'Paradigm Gaps Are Associated with Weird 'Distributional Semantics' Properties: Russian Defective Nouns and Their Case and Number Paradigms'. *The Mental Lexicon* (*2022).*

Corbett, Greville G. *Number*. (Cambridge: Cambridge University Press, 2000).

Greenberg, Joseph H. *Universals of Language.* (Cambridge, MA: MIT Press, 1966).

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. 'BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding'. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*. 4171–4186.

Doetjes, Jenny. Quantity Systems and the Count/Mass Distinction. *(Countability in Natural Language, 2021).* 52–84.

Ghorpynych, Vasyl O. *Morfologhija Ukrajinsjkoji Movy: Pidruchnyk*. (Kyiv: Akademija, 2004).

Gulgowski, Piotr, and Blaszczak, Joanna. 'Conceptual Representation of Lexical and Grammatical Number: Evidence from SNARC and Size Congruity Effect in the Processing of Polish Nouns'. *Formal Approaches to Number in Slavic and Beyond* (2021), 5:29.

Harbour, Daniel. 'Paucity, Abundance, and the Theory of Number'. *Language, 90(1)* (2014), 185–229.

Ingo, Rune. *Suomen Kielen Pluratiivit Eli Monikkosanat, Numeeris-Semanttinen Tutkimus II.* (Vaasa: Vaasan Yliopisto, 1998).

Landman, Fred. 'Count Nouns–Mass Nouns, Neat Nouns–Mess Nouns'. *Baltic International Yearbook of Cognition, Logic and Communication, 6(1), 12* (2011).

Maaten, Laurens van der, and Hinton, Geoffrey. 'Visualizing Data Using t-SNE'. *Journal of Machine Learning Research, 9(Nov)* (2008), 2579–2605.

Meljchuk, Yurij. *Kurs Obshhej Morfologhyy, T. II.* (Moskva–Vena: [Publisher Not Specified], 1998).

Nieder, Jonas, Chuang, Yu-Yin, van de Vijver, Ruben, and Baayen, R. Harald. 'A Discriminative Lexicon Approach to Word Comprehension, Production, and Processing: Maltese Plurals'. *Language, 99(2)* (2023), 242–274.

Niemi, Jussi, Marja Nenonen, and Esa Penttilä. 'Number as a Marker of Idiomaticity'. In Timo Haukioja (ed.), *Proceedings of the XVIth Scandinavian Conference of Linguistics, Turku/Åbo, November 14–16, 1996.* (Turku: Åbo Akademis Tryckeri, 1998), 293–304.

Nikolaev, Anton, Chuang, Yu-Yin, and Baayen, R. Harald. 'A Generating Model for Finnish Nominal Inflection Using Distributional Semantics'. *The Mental Lexicon, 17(1)* (2023).

Peters, Matthew, Neumann, Mark, Iyyer, Mohit, Gardner, Matt, Clark, Christopher, Lee, Kenton, and Zettlemoyer, Luke. 'Deep Contextualized Word Representations'. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*. 2227–2237.

Shafaei-Bajestan, Elnaz & Moradipour-Tari, Masoumeh & Uhrig, Peter & Baayen, Harald. 'The pluralization palette: unveiling semantic clusters in English nominal pluralization through distributional semantics'. *Morphology* 34, 369–413 (2024).

Shvedova, Maryna, von Waldenfels, Ruprecht, Yarygin, Stepan, Rysin, Andriy, Starko, Volodymyr, Nikolajenko, Tetiana, et al. '*GRAC: General Regionally Annotated Corpus of Ukrainian'.* (Electronic Resource: Kyiv, Lviv, Jena, 2022).

Stupak, Inna V., and R. Harald Baayen. 'An inquiry into the semantic transparency and productivity of German particle verbs and derivational affixation'. *The Mental Lexicon* 17.3 (2022), 422-457.

Tiersma, Peter Meijes. 'Local and General Markedness'. *Language, 58(4)* (1982), 832–849.

Vykhovanecj, Hryhorij. *Teoretychna Morfologhija Ukrajinsjkoji Movy*. (Kyiv: Universytetsjke Vydavnyctvo Puljsary, 2004).

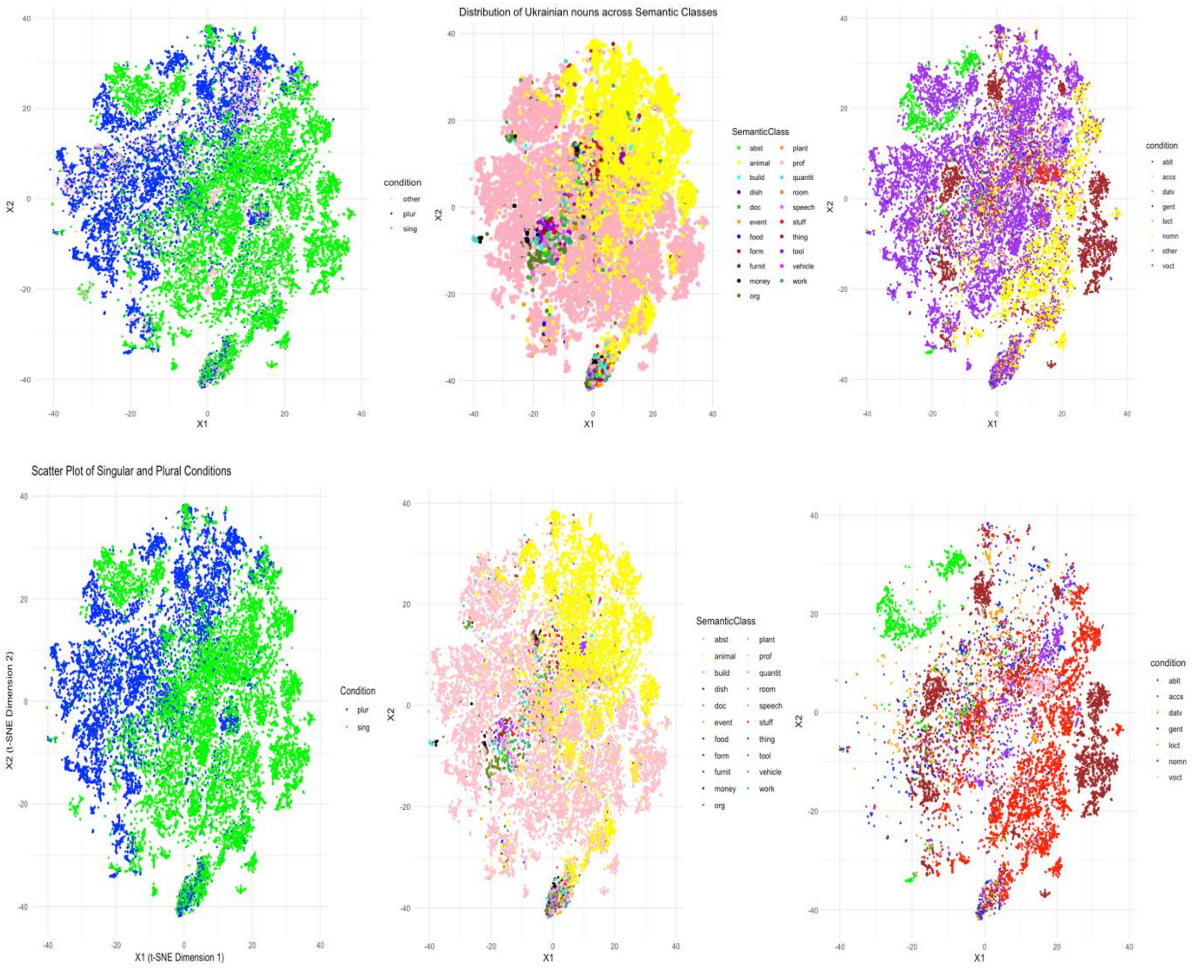

Figure 1: Distribution of the Ukrainian nouns in number, semantic classes, and case. The upper plots illustrate the distribution of Ukrainian nouns based on entire dataset, including low frequency words and syncretic forms. The bottom plots present the more visible clustering of nouns including only non-syncretic forms.

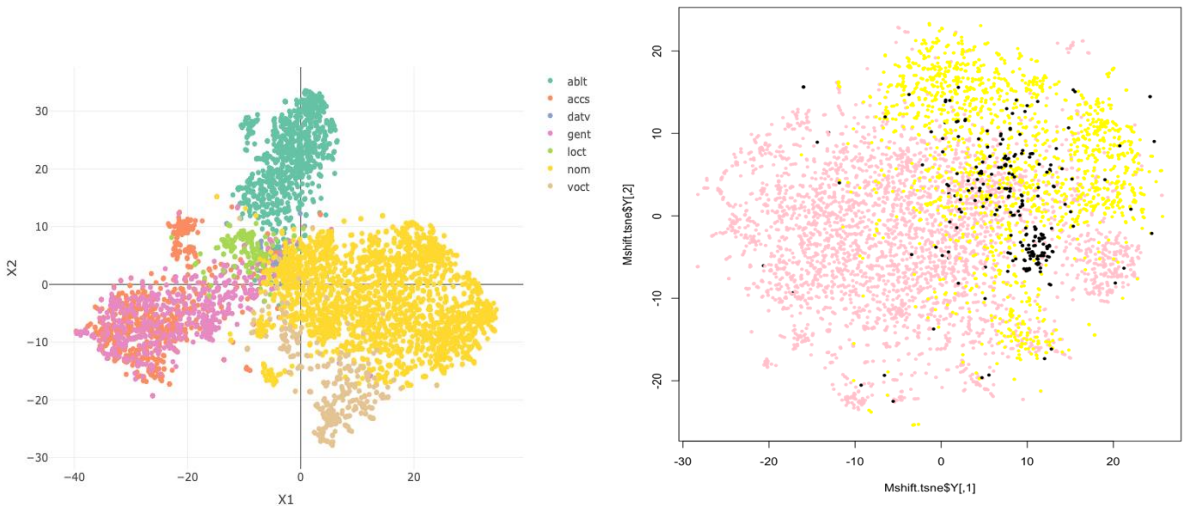

Figure 2: The left plot illustrates the distribution of Ukrainian shift vectors for number, showing notable clustering by case. The dataset comprises all available pairs of singulars and plurals sharing a case. The right plot visualizes the distribution of Ukrainian shift vectors for number within a semantic class. The dataset for this plot comprises all available pairs of singulars and plurals sharing the same semantic class. The 'animal' semantic class is highlighted in yellow, while the 'profession' semantic class is marked in pink. All other semantic classes are represented in black. Ukrainian shift vectors show considerable clustering both by case and by semantic class.