Elections go *bananas*: A First Large-scale Multilingual Study of Pluralia Tantum using LLMs

Elena Spaziani
Sapienza, University of Rome
elena.spaziani@uniroma1.it

Kamyar Zeinalipour
University of Siena
kamyar.zeinalipour2@unisi.it

Pierluigi Cassotti
University of Gothenburg
pierluigi.cassotti@gu.se

Nina Tahmasebi
University of Gothenburg
nina.tahmasebi@gu.se

Abstract

In this paper, we study the expansion of *pluralia tantum*, i.e., *defective* nouns which lack a singular form, like *scissors*. We base our work on an annotation framework specifically developed for the study of lexicalization of *pluralia tantum*, namely *Lexicalization profiles*. On a corresponding hand-annotated testset, we show that the OpenAI and DeepSeek models provide useful annotators for semantic, syntactic and sense categories, with accuracy ranging from 51% to 89%, averaged across all feature groups and languages. Next, we turn to a large-scale investigation of *pluralia tantum*. Using dictionaries, we extract candidate words for Italian, Russian and English and keep those for which the *changing* ratio of singular and plural form is evident in a corresponding reference corpus. We use an LLM to annotate each instance from the reference corpora according to the annotation framework. We show that the large amount of automatically annotated sentences for each feature can be used to perform in-depth linguistic analysis. Focusing on the correlation between an annotated feature and the grammatical form (singular vs. plural), patterns of morpho-semantic change are noted.

1 Introduction

Grammatical number involves an opposition between at least two grammatical values, usually singular and plural (Booij, 2012; Corbett, 2000). An example is the English noun *cat* - *cats* where the singular form refers to a single item (one cat), while the plural form refers to several items (more than one cat). However, this apparent straightforward connection between *form* and *meaning* does not always hold. Consider nouns like *oats* or *trousers* that appear only (or dominantly) in the plural form but are used also in singular meaning: *Please put on a pair of trousers*. We call this category of nouns *pluralia tantum* and consider them to display a defective paradigm of forms (Matthews, 2007).

These nouns resist straightforward morphological analysis, often exhibiting mismatches between number marking (i.e., form) and semantic reference (i.e., meaning). Far from being linguistic curiosities, such phenomena challenge core assumptions about grammatical number and lexical paradigms. Recent research has underscored the value of studying irregularities not only for linguistic theory but also as benchmarks for linguistic competence in Large Language Models (LLMs). For instance, Weissweiler et al. (2024) show that models still falter on argument structure constructions that depart from canonical usage, something *pluralia tantum* can be said to do as well.

Thus far, from a diachronic perspective and using grammatical profiling, Kutuzov et al. (2021) found that the category of number shows a correlation with semantic change cross-linguistically. Other than that, large-scale computational studies or the semantic evolution of *pluralia tantum* has not yet been subject to computational approaches.

In this work, we use an existing annotation framework and first examine the **performance of two families of LLMs in syntactic and semantic comprehension tasks** needed to classify grammatical number irregularities. Upon establishing that both OpenAI and DeepSeek models perform well, we subsequently employ the DeepSeek Reasoner model in a **large-scale diachronic analysis of *pluralia tantum*** in Italian, Russian, and English[1]. The three typologically different languages serve to provide empirically grounded insights into the nature and evolution of the processes around *pluralia tantum*. This paper presents the **first computational, large-scale study on *pluralia tantum* and their evolution across three centuries**.

[1] The languages are chosen because they span different morphological systems and could be annotated by the first author who is a linguist with vast experience in the study of *pluralia tantum*.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6547–6570

1.1 Pluralia tantum

The complex phenomenon of *defectiveness*, where plural and singular forms do not follow the standard relation, is a spectrum composed of different, coherent conditions which exist cross-linguistically (Corbett, 2000, 2018; Koptjevskaja-Tamm and Wälchli, 2001). Here, we consider three conditions[2] with respect to the plural forms:

PD plural dominant noun: if a noun has a plural form that is **more frequent** than its singular counterpart in a given sense (EN pl. *eyelashes* - sg. *eyelash*);

sPT semantic *plurale tantum*, if a noun has a plural form that **corresponds formally** to its singular counterpart but is **semantically isolated** (EN pl. *bananas* 'crazy' - sg. *banana* 'fruit' or pl. *facility* 'equipment' - sg. *facilities* 'ease'). Such plurals are often listed in dictionaries as specific senses of a polysemous lexeme;

mPT morphological *plurale tantum*, if a noun in the plural has **no singular counterpart** and is lemmatized in the plural form in the dictionaries. Historically: a) either because the original paradigm became defective (RU pl. сани 'sledge'), i.e. it lost the singular form over time; b) or because of a semantic detachment from the original paradigm, i.e. the singular form still exists but the two forms are not semantically related anymore (e.g., RU pl. выборы$_2$ 'elections', derived from: RU pl. выборы$_1$ 'choices' - sg. выбор$_1$ 'choice').

The conditions above are sorted by an ascending degree of plural dominance over the singular, reflecting a diachronic process whereby the plural stops denoting several items, and gradually replaces the singular, as in *scissors*.

According to Braun (1930), this process occurs in four consequential stages which involve: 1) the frequency of occurrence of the plural form; 2) the semantic process of *lexicalization*, i.e. the creation and conventionalization of a new meaning in a given word. The latter occurs mostly due to a metonymic shift between the singular and the plural (Blank, 1997; Koch, 2001; Degtjarev, 2014; Šemaeva, 2014) or metaphoric extension on analogy with other plural dominant nouns (Šemaeva, 2016). However, so far, the type of defective expansion that leads to the extreme condition where a word loses its singular form, mPT, has not been determined nor investigated on a large scale and across different languages. For example, it is unclear whether it involves a transition either from the PD condition or from the sPT condition. In this paper, we set out to test this hypothesis.

We chose languages from three different language branches of the Indo-European family: Italian (Romance), Russian (Slavic) and English (Germanic). These belong to different linguistic typologies based on their morphological systems (Bloomfield, 1984; Comrie, 1981; Sapir, 1921).

1.2 The process:

Dictionary selection (Sec. 2) For each language, we choose a reference dictionary and extract all words that have been labeled (in different ways) to fall into one of the three defective conditions. This results in between 800–5 000 candidate words for each language.

Corpus data extraction (Sec. 3) For each language, we calculate the frequency of each candidate word. We then randomly choose up to 3 000 contexts for each word. The data is preprocessed and contexts above a minimum length are kept resulting in, on average, 1 592/978/1 284 contexts for each word in IT/RU/EN.

Frequency filtering (Sec. 4) Our aim is to observe words for which the changing degree of plural dominance takes place in the reference corpus. Thus, we filter out words with no observable change in corpus frequency between the singular and plural form. Remaining are 70/60/241 words for IT/RU/EN.

Annotations (Sec. 5) First, we evaluate different LLMs (from the GPT and DeepSeek families) on a human annotated testset following an in-depth annotation guideline developed for *pluralia tantum*. Next, we annotate all the contexts of the final target words, $\simeq$ 1.5M prompts, using the DeepSeek Reasoner.

Analysis (Sec. 6) We analyze the set of target words based on the annotations and find regular morpho-semantic patterns across the PT conditions and languages. The same annotated data are split based on time and for individual senses offering an in-depth view of each word.

[2] More conditions exist (Allan, 1980; Corbett, 2018) but are not addressed in this paper, e.g., syntactic *pluralia tantum*: when a noun has no morphological plural but it is used with a plural agreement i.e., we say that "the *police* **are** hard on crime" even if *police* formally has no plural inflection.

2 Dictionary selection

Dictionaries usually identify and distinguish *pluralia tantum* to indicate their typical usage. We therefore employed available dictionaries for all three languages, where the three conditions – plural dominant (PD), semantic *pluralia tantum* (SPT) and morphological *pluralia tantum* (mPT) – are distinguishable on the basis of lexicographical labels.

We used De Mauro's *Grande dizionario italiano dell'uso* (2007) for Italian, the digital version of *Tol'kovyj slovar' russkogo jazyka* by Ožegov and Švedova (1992)[3] for Russian, and the *Oxford English Dictionary* OED (1989) for English. We then extracted candidate terms for each criterion and language based on the following; full details on the extraction can be found in Appx. A:

PD are lemmatized in the singular form. Therefore, we retrieved instances by any labels indicating their frequency, for example "spec. al pl" 'especially in the plural'.

sPT are usually specific senses of a polysemous lexeme, also lemmatized in the singular form. We extracted these on the basis of labels denoting their semantic specificity using labels like *almost always in plural*

mPT are lemmatized in the plural form by all the dictionaries; therefore, the plural ending and/or grammatical metadata at the entry level were considered in the extraction.

After extracting lists of words for each condition, set phrases lexicalized in the plural, such as *registration* ***documents***, were excluded, along with proper nouns denoting a group of people or a biological class, such as *Acalepha*. Finally, a singular form was associated to each of the extracted plurals (e.g., pl. *eyelashes* - sg. *eyelash*).

3 Corpus data extraction

For each singular-plural pair, linguistic contexts were extracted from the corpora in Table 1.

For Italian, we first used a fast tokenizer to retrieve all occurrences of the target words[4] by searching for both their singular and plural forms. We then applied SpaCy[5] to filter the results, retaining only those instances tagged as nouns and annotating their morphological number (i.e., singular/plural). For English, since the corpus is already tokenized, lemmatized, and POS-tagged, we directly searched for occurrences of the target words that appear as nouns, considering both singular and plural forms.

Language/Corpus	Time span	N. Tokens	Unique Words
Italian - La Stampa Corpus	1910-2005	3.56B	71M
Russian - Russian National Corpus	1750-2022	48.16M	8.91M
English - The Times Archive	1785-2013	2.37B	264.6M

Table 1: Corpus for each language, total and unique number of tokens are provided for the entire corpus.

Russian proved special in multiple ways. Firstly, concordance lines were manually retrieved from the Main section of The Russian National Corpus – a balanced and annotated collection of texts. Three centuries were selected, ranging from 1700 to 2022. The extraction of concordance lines was possible through the corpus' query system, based on the morphological features encoded in the metadata. However, the manual process had to account for several factors that necessitated further processing:

1. Because some of these plurals have a certain lexicographical autonomy, each form was queried both for a singular and a plural corresponding lemma. Double lemmatization led to the elimination of any duplicate contexts.
2. The sampling included all the available contexts. However, many semantic *pluralia tantum* are part of a very polysemous lexeme, which involves more occurrences. For 31 of such cases (frequency $\geq 50\,000$ occurrences), the sampling was reduced to the selection of just one occurrence per text.[6]
3. The selected time span includes texts characterized by the old orthography, which underwent major changes after the Bolshevik Revolution. Thus, all contexts were automatically converted into modern orthography.[7]

Furthermore, the morphological irregularity of the selected Russian nouns posed several challenges when annotating with lemma and grammatical number. We therefore used two lemmatizers alternately [8] and a subsequent manual intervention

[3] https://www.booksite.ru/fulltext/1/001/001/328/index.htm
[4] NLTK Word Tokenizer [5] it_core_news_sm
[6] For example, ворота, служба, право, действие.
[7] https://github.com/dhhse/prereform2modern. [8] UDPipe (Straka, 2018) and Pymorphy (Korobov, 2015)

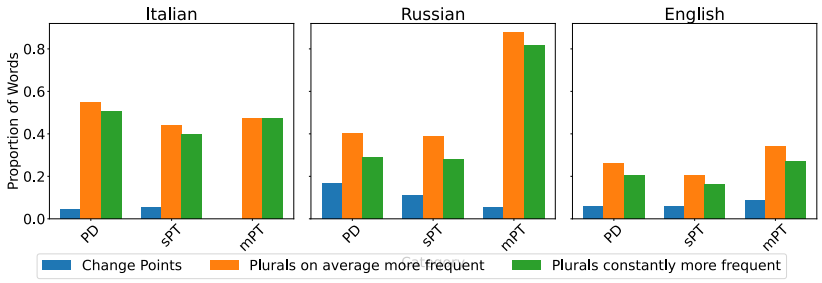


Figure 1: Blue bars represent the proportion of words that exhibit a change point the corpus. Orange refers to plurals being more prominent after the change point, while green involves plurals that are constantly more prominent.

for particularly idiosyncratic nouns. [9]

4 Frequency filtering

In this analysis, we are interested in the 'birth' of defectiveness, that is, we are concerned with those words for which the ratio between plural and singular changes over time, or in which the plural form was almost non-existent and then became significantly more frequent. Therefore, we excluded all words that were stable, i,e., for which we could not find evidence of changing proportions between plural and singular forms.

As a baseline, we used the knowledge that, generally, singular forms are significantly more frequent than their corresponding plural forms (Greenberg, 1963). Haspelmath and Karjus (2017) show that in the Russian National Corpus, 60 million singular forms are found, while only 29 million plural forms are available. According to the BNC, 14.52M tokens are attested in the singular form, while the total occurrence of plural forms is 4.95M. Thus, larger deviations from these ratios (between 2:1 – 3:1) were considered as indications of defectiveness. For instance, the noun *facility*, in Figure 2, shows a significantly different singular:plural ratio than the expected 2:1 or 3:1 and it is thus suitable for the study. The filtering was done by first creating time series from the log ratio of the smoothed relative frequency of the singular and the plural and then applying change point detection. Details on the filtering procedure can be found in Appx. B.

4.1 Results: Frequency

The filtering yielded a list of candidate words experiencing a change in frequency ratio over time, i.e. a change point. Within the respective reference

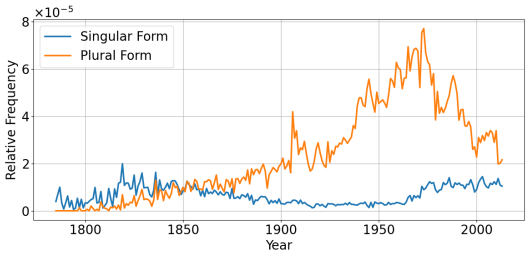


Figure 2: Frequency of *facility-facilities*.

corpora, 70/60/241 words for Italian, Russian, and English respectively.

The results in Fig. 1 show that there is roughly an even distribution among the conditions in Italian and English (where PD is slightly higher for the former, while mPT is slightly higher in the latter). For Russian, the mPT condition stands out. Firstly, there are about double as many mPTs as in each of the other conditions. Secondly, quite a large number of the mPTs, about half, have plurals that are constantly more frequent (green bar) throughout the dataset. This indicates that about half of the *pluralia tantum* had already formed before 1750.

However, Russian shows a higher percentage of words that exhibit a change point (blue bars). English follows with a slightly less prominent percentage, with mPT being the most dynamic condition.

Italian exhibits the lowest percentage of change point in the time series, indicating the highest change in the sPT condition. Importantly, we find an absence of changing words in the condition of mPT, most likely due to the fact that many of the mPT in Italian are Latin loanwords already assimilated as *pluralia tantum* prior to the historical period under analysis. These results align with the general view of *pluralia tantum* as "parasitic" (Koptjevskaja-Tamm and Wälchli, 2001, p. 631) on the degree of grammaticalization of the grammatical number (Wiltschko, 2008) and on the typologi-

[9] Most of them belong to the morphological *pluralia tantum* set: e.g., квасцы; космы; козлы; кранты; крестины; кудри; латы; литавры; макароны; махры; ножницы.

cal nature of the individual language (Bloomfield, 1984).

Russian, as characterized by a rich morphological system, shows more prominent patterns in the mPTs compared to Italian and English, both fusional with analytical tendencies (Sapir, 1921). A complete overview of the different conditions before and after filtering is illustrated in Fig. 8 (Appx. B).

5 Annotations

For our work we used an annotation framework inspired by the *Behavioral Profiles* method from Divjak and Gries 2006; Gries 2010, which involves detailed lexical item annotation in concordance lines. We adopted a similar method called *Lexicalization Profiles*.[10] This annotation framework was designed specifically for examining the evolution of *pluralia tantum* in previous research. However, given its extensive breadth, manual annotation using this framework is highly time-consuming. Consequently, two nouns per language manually annotated served as gold standards for evaluation, allowing the process to be extended to all nouns with the use of LLMs. The following sections illustrate the details of the framework and the evaluation process preceding the full-scale annotation.

5.1 Our annotation framework and prompts

Our detailed annotation framework consists of **52** features for Italian and English, and a total set of **58** features for Russian. The features belong to categories that separate different linguistic information: 1) *sense inventory*; 2) *sense categories*; 3) *semantic categories*; 4) *colligation L1* (left); 5) *colligation R1* (right); 6) *diaphasic preference*; 7) *text theme*; and 8) *morphological categories* (for Russian only). See Table 2 for an example sentence annotated according to the framework.

sense inventory	semantic category	sense category	colligation L1	colligation R1	diaphasic preference	text theme
1.1	concrete; inanim.	primary	adjective	pronoun	neutral	daily life

Table 2: Example of annotation for the sentence: "1878 … ate off that gentleman's hand some banana which he gave him."

After testing a single prompt across all categories, we divided the categories listed above into three thematically motivated groups, as this approach proved to be more efficient.

Group 1 – $\mathbf{P_{SENSE}}$ consists of the *sense inventory*, i.e. an inventory of senses given by a dictionary, and the *sense categories*, i.e. a fixed set of sense relations (metonymy, metaphor and taxonomy). The aim of this prompt is to introduce the model to a double mapping between the sense codes and the sense relations given a specific usage for the considered form.

Group 2 – $\mathbf{P_{SEM}}$ involves *semantic categories*. Exclusively for Russian, *morphological categories* are included. Both describe linguistic properties: semantic categories involve the features of abstractness, concreteness and animacy; morphological features are grammatical case (missing from Italian and English).

Group 3 – $\mathbf{P_{DIST}}$ concerns the *colligational categories* (co-occurring POS for the left and right context), and distributional properties such as *diaphasic preference* (register) and *text theme*.

We subsequently designed one prompt for each group. An illustrative excerpt of each prompt for Russian is reported below, while the full version is available in our GitHub repository. [11]

Prompt 1 – $\mathbf{P_{SENSE}}$

```
"You are a language model specialized in
    in-depth linguistic annotation of Russian
    texts - both historical and modern. Your
    task is to analyze any occurrence of the
    word [] found in a given excerpt, which
    could belong to any period in a time span
    of four centuries (1700-2025), and
    provide a detailed annotation using the
    two categories described below. Each
    annotation should be based on careful
    contextual analysis of a given excerpt
    and on a double mapping involving: 1.
    assigning a sense from a hierarchical
    sense inventory that best matches the
    meaning of the target word in the given
    context; 2. Labeling the sense relation
    (metonymic, metaphoric etc.) between the
    contextual use of the word and the
    selected sense from the inventory..."
```

Prompt 2 – $\mathbf{P_{SEM}}$

```
"You are a language model specialized in
    in-depth linguistic annotation of Russian
    texts - both historical and modern. Your
    task is to analyze any occurrence of the
    word [] found in a given excerpt, which
    could belong to any period in a time span
    of four centuries (1700-2025), and
    provide a detailed annotation using the
    two categories described below. Each
    annotation should be based on careful
    contextual analysis and in line with a
    standardized inventory of semantic and
    morphological features.
```

[10] *Lexicalization Profiles* stems from the first author of this paper (Spaziani, Forthcoming).

[11] https://github.com/ChangeIsKey/pluralia-tantum-eacl2026.

```
3. Semantic Categories
Determine whether the referent of the
     word [] is abstract or concrete, and
     whether it is animate or inanimate in
     the given context. Possible values
     for Concreteness: abstract, concrete
     Possible values for Animacy: animate,
     inanimate...
4. Morphological Categories
Identify the grammatical case of the word
     [] using syntactic cues and word
     endings..."
```

Prompt 3 – $\mathbf{P_{DIST}}$

```
"You are a language model specialized in
     in-depth linguistic annotation of Russian
     texts - both historical and modern. Your
     task is to analyze any occurrence of the
     word [] found in a given excerpt, which
     could belong to any period in a time span
     of four centuries (1700-2025), and
     provide a detailed annotation using the
     three categories described below. Each
     annotation should be based on careful
     contextual analysis and in line with a
     standardized inventory of features
     belonging to the collocates of the word,
     to the text theme and register.
5. Distributional Categories (Colligation
     Left/Right)
     Label the part-of-speech (POS) of the
          immediately adjacent words:
     - L1 POS: the POS for the word
          immediately preceding the word []
     - R1 POS: the POS for the word
          immediately following the word
          []..."
```

5.2 Evaluation of the automatic annotation

Before scaling up to millions of instances, we benchmarked several instruction-tuned language models on their ability to reproduce *human* lexicalization profiles. We evaluated eight candidate models from the GPT and DeepSeek families[12] in both **zero-shot** and **three-shot** variants. For each usage we ran three distinct Chain-of-Thoughts prompts according to Group 1–3 above. For the three-shot variant, three in-context examples for *each* language were used, which were taken from *different* target words than those used in the evaluation.

We evaluated against $\simeq$ 600 manually annotated usages for the six words, (IT-*urna*, IT-*condoglianza*, RU- кошка, RU-переговор, EN-*banana*, EN-*hostility*) and measured agreement as the mean *Jaccard similarity* between each predicted label set and its gold counterpart.

Assessing the models' predictions separately for each category, distinct patterns can be noted. We refer to Appx. C for further details. Overall, *semantic* and *morphological categories* exhibit high agreement with gold annotations. *Colligational categories*, encoded as immediate left and right syntagmatic context (L1/R1), are also captured reliably. In contrast, features related to the *sense inventory* and *sense relations* prove substantially more challenging for the LLMs, resulting in markedly lower levels of agreement. This result aligns with previously documented limitations of such models in WSD tasks (see, e.g., Kibria et al. 2024).

To select our model, we averaged performance across all categories for each language under both zero-shot and three-shot prompting. Results in Table 3 demonstrate that 1) three-shot prompting yields minimal or no performance improvement, and 2) the OpenAI O3-model achieves the best overall performance, followed closely by the DeepSeek Reasoner. Therefore, we adopted a zero-shot strategy and selected the second-best model, DeepSeek Reasoner, as the large number of prompts ($\sim$ 1.5M) and their length made the top-performing model prohibitively expensive.

6 Analysis of automatically annotated nouns

Our aim is to explore how linguistic features relate to grammatical number, that is, how correlated each annotated feature is to the singular vs. plural form of the word across conditions of *pluralia tantum*.

For each word, we identified the presence or absence of a certain linguistic feature and the grammatical number (singular vs. plural). Specifically, for each usage we encoded singular as 1 and plural as 0, and, in parallel, we built a binary vector, marking whether the linguistic feature was present (1) or absent (0). Thus, we obtained one vector for grammatical number and 52 vectors (or 58 in the case of Russian), one for each feature.

We then used Kendall's Tau, a rank-based correlation measure, to assess the relationship between the grammatical number vector and each of the linguistic features. As a result, 15/25/30 features, for IT/RU/EN respectively, were retained, while hidden features were excluded due to their weak correlation (< 0.3).

A positive correlation (red) indicates that a feature tends to occur more often when the word is used in the singular, while a negative correlation (blue) suggests it is more associated with plural use. A correlation close to zero means that the feature does not have a preference for singular or plural.

In Figure 3 we have plotted the correlation between some of the features (rows) and the three

[12] `chatgpt-4o-latest`, `deepseek-chat`, `deepseek-reasoner`, `gpt-4.1-2025-04-14`, `gpt-4o`, `gpt-4o-mini`, `o3`, `o3-mini-2025-01-31`.

Model	Italian			Russian			English			Overall		
	0	3	Δ	0	3	Δ	0	3	Δ	0	3	Δ
o3	**0.786**	**0.674**	-0.112	**0.791**	**0.792**	+0.001	0.782	0.774	-0.008	**0.787**	**0.749**	-0.038
deepseek-re	0.737	0.650	-0.087	0.773	0.756	-0.017	**0.791**	**0.798**	+0.007	0.767	0.736	-0.031
o3-mini	0.699	0.617	-0.082	0.716	0.741	+0.025	0.743	0.794	+0.051	0.719	0.718	-0.001
gpt-4.1	0.682	0.577	-0.105	0.724	0.730	+0.006	0.722	0.748	+0.026	0.710	0.687	-0.023
chatgpt-4o	0.678	0.604	-0.074	0.706	0.730	+0.024	0.724	0.750	+0.026	0.703	0.696	-0.007
gpt-4o	0.649	0.561	-0.088	0.726	0.740	+0.014	0.713	0.733	+0.020	0.697	0.673	-0.024
deepseek-chat	0.637	0.571	-0.066	0.726	0.719	-0.007	0.700	0.723	+0.023	0.690	0.681	-0.009
gpt-4o-mini	0.492	0.477	-0.015	0.598	0.598	+0.000	0.629	0.618	-0.011	0.574	0.566	-0.008

Table 3: 0-shot vs. 3-shot Jaccard scores (Δ = 3-shot – 0-shot) for the average performance of each language. Positive Δ indicates an improvement from the use of few-shot prompting.

defective conditions (mPT, sPT, PD) for each language. A more detailed breakdown by individual nouns (see Figures 14, 15, 18 in Appx. E) shows that some words in the different conditions exhibit higher correlation values between the features and the grammatical form. Both types of plotting allow us to analyze the general patterns across languages and conditions, as well as in-depth patterns for specific words. Accordingly, we will discuss: 1) cross-linguistic differences; and 2) two case studies of lexicalization in the plural form of *rifiuto* and *fondello*. We refer to Appx. E for full-size and more detailed plots.

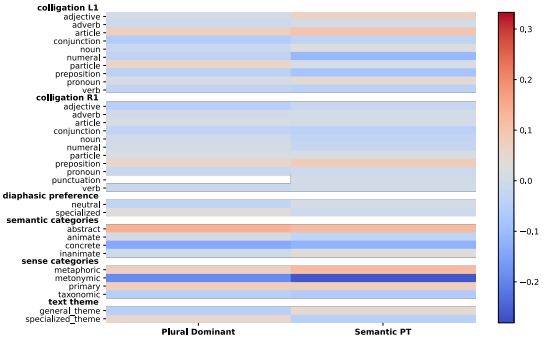
6.1 Cross-linguistic Conditions

Figure 3 enables a comparison of correlation trends across the three languages for each defective condition. Within the sense categories of the three languages, the *metonymic* relation is primarily correlated with the plural form (blue rows) with the exception of Russian mPT; while metaphorical relations are more often correlated with singular forms (red row). The *abstract* category is also correlated predominantly with the singular form (red rows) for Russian and Italian. For English, the correlation is strong with plural for *article* in colligation L1 and *noun* in colligation R1. Apart from this, colligations (right or left) are not strongly correlated with grammatical number for any language or condition.

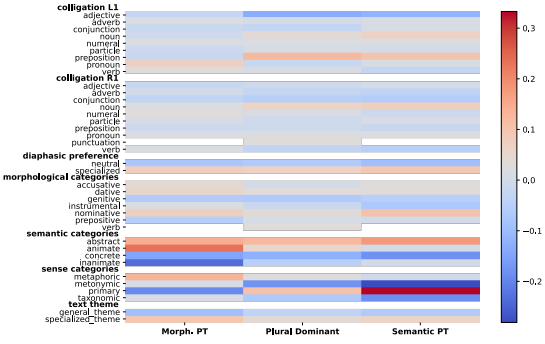
In Sec. 1.1 we stated that sPTs were expected to have a stronger correlation between the primary sense and singular forms, since they are specific senses of nouns whose primary sense is singular dominant. This observation is confirmed only for Russian sPTs, as they are almost exclusively correlated with the singular form (see Fig. 3b). On the other hand, in the Italian and English sPTs, primary senses are equally often correlated with plurals, with English having a slightly higher correlation value with the singular form.

Considering the heatmaps for individual nouns (see Fig. 14-18-15), we can see that for each language, a few words stand out among the defective conditions. These words display a divergence between singular and plural forms across the features, particularly in the metonymic or metaphoric relations, which constitute the primary mechanisms of lexicalization. The general trend is that plural forms correlate more strongly with metonymic and concrete senses, while singular forms correlates with primary and abstract senses, with some exceptions (Italian PD *fondello* 'bottom', Russian mPT ладушка 'beloved' or English mPT *overhead*). Moreover, plural forms occur more frequently in general-themed domains, whereas singular forms correlate with specialized domains, alluding to the predominance of plural in standard discourse. We can notice such patterns in the Italian sPT *addominale* 'abdominal', and the PDs *fondello* 'bottom' and *rifiuto* 'refusal'. For Russian, отход 'waste', the sPTs испарение 'evaporation' and лишение 'deprivation' emerge clearly in the correlation with metonymy. For English there are many words that show such divergence; e.g. we can notice the sPT *antiquity* and the PD *pressing*.

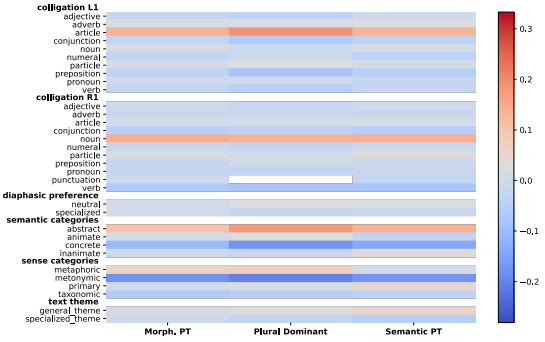
Beyond these observations, distinct morpho-semantic patterns can be identified. Both italian *rifiuto* 'waste' and the Russian equivalent отход 'waste' fall into a cross-linguistic semantic category



(a) Italian



(b) Russian



(c) English

Figure 3: Heatmaps for the correlation of all individual features averaged across each PT condition.

of *pluralia tantum* (Gardelle and Acquaviva, 2023; Koptjevskaja-Tamm and Wälchli, 2001) which concerns 'heterogeneous substances', i.e. nouns denoting substances with perceptually identifiable sub-parts. From a morphological perspective, the outstanding words seem to be mostly nouns deriving from a verb (e.g. IT *rifiuto* 'refusal', RU отход 'waste', EN *pressing*) or nouns deriving from adjectives (EN *iniquity*, *antiquity*, both with a latinate root). In Italian, the latter are especially present as a byproduct of ellipsis, a very productive word formation mechanism (Thornton, 2004). For instance, we can note the deadjectival noun *addominale* 'abdominal (muscle)', which is lexicalized as 'body part' only in the plural form.

Finally, the type of defective expansion can be hypothesized by observing the development of both the singular and the plural form. In the following paragraphs, two case studies will be examined. Specifically, individual heatmaps showing feature correlation by decade and sense-level frequencies will be considered for the analysis of the Italian PDs *rifiuti* and *fondelli*.

6.2 Case study: Rifiuti

According to the dictionary (De Mauro, 2007), *rifiuto* 'refusal' has a different meaning in the plural, that is 'waste'(2 spec. al pl., ciò che si butta via perché inutilizzabile). Indeed, the frequency of the individual senses for this word illustrated in Figure 4 shows that sense 2 'waste' exists almost exclusively in the plural form, confirming its lexicographical status as plural dominant.

To verify the semantic evolution of this sense, we can compare Figures 4 and 5, monitoring the sense relation with the sense frequency of *rifiuto*. As illustrated in Fig. 4, around 1930, sense 2 emerges; while Fig. 5 reveals a high correlation between metonymy and the plural in the same decade. This observation suggests that the plural dominant sense could indeed be the byproduct of a metonymic shift. The sense-level frequency proves insightful for comparing the evolution of the singular and the plural forms. In Fig. 4, we observe that each retains a predominant sense, distinct from the other, with the primary sense of the singular form being visibly less frequent, but somewhat stable. Moreover, in Fig. 5, it is interesting to note that there is a clear distinction between the singular and plural forms. The singular form is highly correlated with the feature of abstractness, while the plural form is conversely highly correlated with concreteness. This could suggest a potential defective expansion of the plural via semantic detachment, i.e. a transition from PD to sPT (see 1.1).

6.3 Case study: Fondelli

As for *fondello*, the sense specialized in the plural is 'glutes' (6. scherz., spec. al pl., culo, sedere: prendere un calcio nei fondelli). The dictionary's division is coherent with the frequency of the individual senses, as sense 6 'glutes' exists almost exclusively in the plural form. Comparing Figures 6 and 7, we can map the sense relation to the derivation of sense 6. As seen in Fig. 6, around 1970, sense 6 emerges and, simultaneously, the

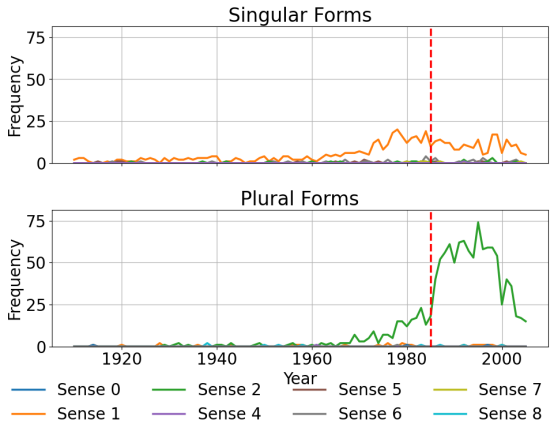


Figure 4: Frequency of individual senses of *rifiuto*.

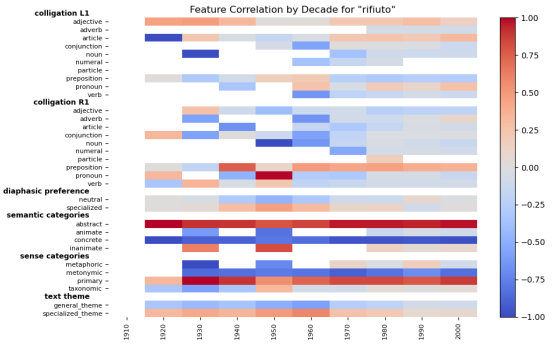


Figure 5: Feature correlation for *rifiuto* across decades.

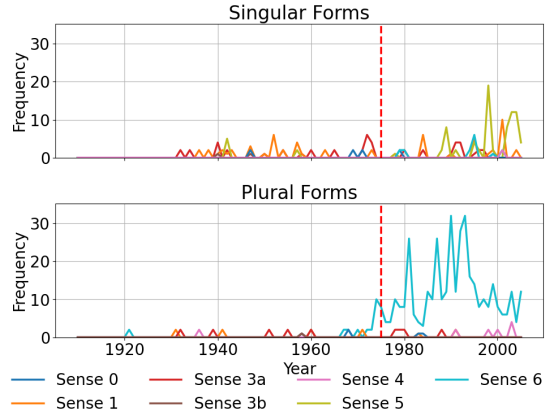


Figure 6: Frequency of individual senses of *fondello*.

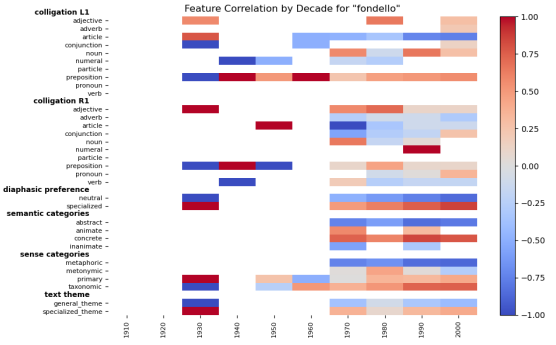


Figure 7: Feature correlation for *fondello* across decades.

correlation with the plural is particularly strong in the feature indicating a metaphorical relation. The metaphorical extension would motivate the lexicalization of the plural form, following the model proposed by Šemaeva (2016), based on analogy with plural dominant nouns. In this case, the primary sense in the plural form *'bottoms' establishes a similarity relation with 'glutes', a plural dominant noun denoting a bipartite body part.

With respect to its defective expansion pattern, from Fig. 6 it is clear that sense 6 is more frequent than sense 1 in the singular form, exhibiting an evident dominance over the whole paradigm. This could be evidence for a possible defective expansion, which eventually leads to the loss of the singular form, following the model of сани 'sledge', i.e. a transition from PD to mPT (see Sec. 1.1).

7 Conclusions

In this paper, we took the first steps towards a large-scale study of defective nouns using computational tools. Our results first reveal that current-day LLMs can be effectively used without the need for fine-tuning. While they can annotate well in certain categories, sense relations are still challenging for the models.

The computational annotations further allowed us to conduct the first linguistic analysis of *pluralia tantum* at large-scale for three different languages, whose results should be interpreted with caution in light of the model's performance. Nonetheless, the findings align well with several existing linguistic studies while shedding new light on the cross-linguistic phenomenon of defectiveness. Overall, we conclude that there is evidence of semantic isolation of the plural with respect to the singular form, in terms of sense frequency as well as correlation to different linguistic features, but that colligation features are less revealing than semantic features.

Considering both the sense-level frequency and the feature correlation in the case studies of *fondello* and *rifiuto*, a potential model for defectiveness can be grasped: their senses in the plural could transition into either morphological *pluralia tantum* category or semantic *pluralia tantum*.

Limitations

After conducting the experiments with three-shot settings, we observed a slight improvement in the overall performance of the models. However, the performance for Italian decreased across all models (see Appx. C). While this decline may be due to the specific examples chosen, it also suggests a language-specific sensitivity. Given Italian's high degree of context-dependent polysemy, providing few-shot examples from unrelated words likely acted as a distraction rather than a guide, introducing noise that misled the models during the annotation task. Important, however, is that we chose the examples to include in the three-shot prompts randomly from words that showed lexicalized plurals, using the same procedure for all three languages. Thus, one of our limitations was the inability to test multiple sets of 3-shot examples, which would have allowed us to better understand how shot selection affects model performance. On the other hand, this suggests that zero-shot prompting might be a safer and more reliable approach for this type of annotation task, given the out-of-the-box quality of the models currently.

We began finding candidate words using dictionary resources. Here, a single dictionary was used for each language and of course, could be complemented by more dictionaries, for example, modern or slang dictionaries, where new candidate words are found. Next, we used a single corpus for each language. For Russian, the corpus is balanced across genre and text type. The other corpora are newspaper archives, and thus more limited in variation. This can lead to the fact that some words do not have frequencies for some of their senses, for example because the sense is domain-specific (e.g., mostly found in legal texts not available in our corpora). A larger variety of sources would offer better evidence for such senses.

For our analysis, we employed computational models to scale up the annotation to multiple words across multiple senses. In contrast, human annotations for the test instances took around a 0.5–1 week per word (about 100-150 instances). While the annotation quality is imperfect, computational annotation allows for scaling up annotations from single words to all candidate words found in a dictionary for three different languages. We used the annotations in two main ways: firstly, we aggregated the results for each category across all time periods and all senses to gain a high-level overview of lexicalization profiles. In the case study for *fondello* and *rifiuto*, we also split the data according to time periods. However, it remains to split the data also into senses, to allow for individual analysis of senses. Such a split might allow us to faster detect senses that are in the process of defecting, but where the defective sense has not yet become the most prominent.

A major limitation of this study is that the corpora licensing prevents us from sharing the annotated contexts for reuse by other researchers, thereby avoiding time and resource-consuming re-annotations. We are, however, sharing all the code

Acknowledgments

This work has in part been funded by the research program Change is Key! supported by Riksbankens Jubileumsfond (under reference number M21-0021). Additionally, the computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

References

Keith Allan. 1980. Nouns and countability. *Language*.

Andreas Blank. 1997. *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*. Max Niemeyer Verlag.

Leonard Bloomfield. 1984. *Language*. University of Chicago Press.

Geert Booij. 2012. *The Grammar of Words. An Introduction to Linguistic Morphology*. Oxford University Press.

Maximilian Braun. 1930. *Das Kollektivum und das Plurale tantum im Russischen: Ein bedeutungsgeschichtlicher Versuch*. Ph.D. thesis, Universitat Leipzig.

Bernard Comrie. 1981. *Language Universals and Linguistic Typology*. Blackwell.

Greville Corbett. 2000. *Number*. Cambridge University Press.

Greville G. Corbett. 2018. Pluralia tantum nouns and the theory of features: a typology of nouns with non-canonical number properties. *Morphology*.

Tullio De Mauro. 2007. *Grande dizionario italiano dell'uso*. UTET.

Vladimir I. Degtjarev. 2014. *Kategorija čisla v slavjanskich jazykach (istoriko-semantičeskoe issledovanie*. Izdatel'stvo Južnogo Federal'nogo Universiteta.

Dagmar Divjak and Stefan Th. Gries. 2006. Ways of trying in russian: Clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory*.

Laura Gardelle and Paolo Acquaviva. 2023. Pluralia tantum and singularia tantum. *The Wiley Blackwell Companion to Morphology*.

Joseph H. Greenberg. 1963. *Universals of language*. M. I. T. Press.

Stefan Th. Gries. 2010. Behavioral profiles: A fine-grained and quantitative approach in corpus-based lexical semantics. *The Mental Lexicon*.

Martin Haspelmath and Andres Karjus. 2017. Explaining asymmetries in number marking: Singulatives, pluratives and usage frequency. *Linguistics*.

Raihan Kibria, Sheikh Intiser Uddin Dipta, and Muhammad Abdullah Adnan. 2024. On functional competence of LLMs for linguistic disambiguation. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 143–160, Miami, FL, USA. Association for Computational Linguistics.

R. Killick, P. Fearnhead, and I. A. Eckley and. 2012. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

Peter Koch. 2001. Metonymy: Unity in diversity. *Journal of Historical Pragmatics*.

Maria Koptjevskaja-Tamm and B. Wälchli. 2001. The circum-baltic languages: An areal-typological approach. In *The Circum-Baltic Languages: Typology and Contact. Vol. 2*. John Benjamins Publishing.

Mikhail Korobov. 2015. Morphological analyzer and generator for russian and ukrainian languages. In *Analysis of Images, Social Networks and Texts*. Springer International Publishing.

Andrey Kutuzov, Lidia Pivovarova, and Mario Giulianelli. 2021. Grammatical profiling for semantic change detection. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 423–434, Online. Association for Computational Linguistics.

Peter Hugoe Matthews. 2007. *The Concise Oxford Dictionary of Linguistics*. Oford University Press.

Oxford English Dictionary OED. 1989. Oxford english dictionary. *Simpson, Ja & Weiner, Esc*, 3.

Sergej Ožegov and Natalija Švedova. 1992. *Tol'kovyj slovar' russkogo jazyka*. Izdatel'stvo Az'.

Edward Sapir. 1921. *Language, an introduction to the study of speech*. Harcourt, Brace and Company.

Elena Spaziani. Forthcoming. Nizy, koški, peregovory and vybory: Lexicalization profiles. *Studi Slavistici*.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. ACL.

Anna Thornton. 2004. *La formazione delle parole in italiano*, chapter Conversione. Max Niemeyer Verlag.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. 2020. Selective review of offline change point detection methods. *Signal Processing*, 167:107299.

Leonie Weissweiler, Abdullatif Köksal, and Hinrich Schütze. 2024. Hybrid human-llm corpus construction and llm evaluation for rare linguistic phenomena. *arXiv preprint arXiv:2403.06965*.

Martina Wiltschko. 2008. The syntax of non-inflectional plural marking. *Nat Lang Linguistic Theory*.

Elena Šemaeva. 2016. Identificirujuščaja metaforičeskaja model' leksikalizacii form množestvennogo čisla anglijskich i russkich imen suščestvitel'nych. In *Kognitivnye issledovanija jazyka*. Izdatel'skij dom TGU im. G.R. Deržavina.

Elena V. Šemaeva. 2014. *Kognitivnye osnovy leksikalizacii form množestvennogo čisla imeni suščestvitel'nogo: na materiale anglijskogo i russkogo jazykov*. Ph.D. thesis, Belgorodskij Gosudarstvennyj Nacionalnyj Issledovatel'skij Universitet.

A Details on target word extraction based on dictionaries

For Italian, De Mauro's *Grande dizionario italiano dell'uso* (2007) was chosen, as it includes a query system that allows for the selection of various properties associated with the lemmas, including stylistic labels. PD were associated with the lexicographic label "spec. al pl" 'especially in the plural', which denotes senses **mainly** used in the plural. The condition of sPT was associated with the labels "al pl." and "pl.", denoting senses **exclusively** used in the plural form. The mPT are lemmatized in the plural form, so they were selected based on grammatical metadata, e.g., "f.pl." at the entry level.

For Russian, the digital version of *Tol'kovyj slovar' russkogo jazyka* by Ožegov and Švedova (1992) was used. Also here, two alternative labels were found for the PD condition, обычно мн; чаще мн. 'usually / more frequent in the plural', while only the label мн. 'pl.' was used for sPT. The instances of mPT in the Russian dictionary do not

present any grammatical information at the entry level regarding the use in the plural. Therefore, the lemmas were extracted by their endings, which necessitated further manual cleaning, due to the presence of homography.

condition	senses mainly in the plural	senses exclusively in the plural	nouns exclusively in the plural
lexicographical label	обычно мн;чаще мн. 'usually / more frequent in the plural'	мн. 'pl.'	plural ending
example	ЧАС, -а 5. обычно мн.'Hour, 5. usually in the pl.'	КОСТЬ, -и 2.мн. '	СУТКИ, -ток. 'twenty-four hours'

Table 4: Example of a sub-set divison for lexicalized plurals in Russian

For English, the *Oxford English Dictionary* OED (1989) was used. Given the *sui generis* composition of the OED, rich and layered, when denoting the behavior of a sense in the plural form, a broader set of labels compared to the other dictionaries was found. Altogether, seven labels were attributed to the plural dominant condition (*often in plural*; *frequently in plural*; *usually in plural*; *commonly in plural*); three labels were attributed to senses used exclusively in the plural (*Almost always in plural*; *chiefly in plural*; *also in plural*; *in plural* - and its allographs).

B Details on the filtering procedure

To perform the filtering such that we keep pairs for which the change takes place in the corpus, we analyzed how the frequency of the plural and singular forms changes in the corpus for each year. First, we computed the smoothed relative frequency p_w^t for each word w and each year *t*, both for the singular and the plural forms, relying on:

$$p_w^t = \frac{f_t^w + 1}{C^t + |V^t|} \quad (1)$$

Subsequently, we computed the $(\text{odds}(w)_t)$, i.e. the log ratio of the smoothed relative frequency of the singular and plural, respectively:

$$\text{odds}(w)_t = \log \frac{p_{w_s}^t}{p_{w_p}^t} \quad (2)$$

Operationally, $\text{odds}(w)_t$ specifies the probability that the singular form will appear in a text relative to the plural form in the specified year t. We then obtained the time-series by concatenating the $\text{odds}(w)$ values computed for each year: $(\text{odds}(w)_{y_1}, \text{odds}(w)_{y_2}, .., \text{odds}(w)_{y_N})$.

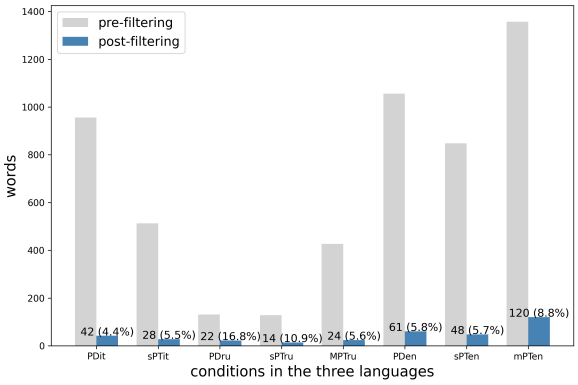


Figure 8: Frequency filtered values per condition in the three languages

Next, we applied change point detection to frequency ratio time series to identify significant shifts in word usage over time.[13] The output of this process consists of both the detected change points and the segments between them. Each time series was split into contiguous segments at the detected change points, and we computed the mean frequency ratio within each segment. This allows us to quantify both the timing and the magnitude of shifts in word usage.

Finally, from the list of target words for each language, we selected those that exhibit an increase in plural usage relative to singular. This selection was based on identifying words for which a change point was detected, such that the average of the ratios in the segment preceding the change point is higher than zero, while the average in the subsequent segment is less than zero. See Figure 8 for an overview.

C Model performance

We evaluated the OpenAI and DeepSeek family of models on the manually annotated test set. The testing was done first by creating a single prompt with all six categories (or seven for Russian) together. We found, however, that such a prompting strategy was less than optimal. Thus, we continued with three separate prompts, one for each thematic group presented in Section 5.1. These prompts were then more detailed in describing the category.

We then tested the models in a zero-shot setting, without any examples from the annotation, as well

[13] We used the Pruned Exact Linear Time (PELT) algorithm (Killick et al., 2012) implementation provided in the ruptures library (Truong et al., 2020), with the Radial Basis Function (RBF) model, which is well-suited for detecting nonlinear changes in time series. A penalty value of pen=10 was used to control the sensitivity of change point detection.

as in a three-shot setting, where we offered three examples from a different word. The results of the best model per category, as well as the overall performance, can be found in Table 5 and 3. The first two tables, Table 5-6, show the performance for individual categories based on the two prompting strategies.

We find that the zero-shot strategy offers competitive results, where the best models often surpass the three-shot prompting. Specifically, The best model for *sense categories* is significantly worse with the three-shot strategy (0.606 compared to 0.663 with the zero-shot), as well as the *text theme* (0.602 vs 0.689). *Morphological categories*, for Russian, are instead improved by offering examples (0.986 vs 0.960).

In Fig. 11 (zero-shot) and 12 (three-shot) we show the performance of each model for each of the categories, for both individual languages as well as the overall performance. Here we can find, e.g., that all models have very high performance for the *semantic categories* for Italian and English, but significantly lower performance for Russian.

If we look at different categories (Table 5), we find that for the *sense inventory*, the DeepSeek models perform best: the reasoner model for English and the Chat model for Italian and Russian. All models cluster tightly around 0.82–0.84 for *semantic categories*, indicating that coarse semantic typing is nearly saturated (see Fig. 11). The lower band for *text theme* (0.55–0.66) confirms that topic detection from a single concordance line is the most subjective part of the profile and is most often misclassified by the models. Importantly, the gold standard allows only *one* theme per context, whereas the same snippet often evokes several overlapping topics. A manual spot-check of 100 disagreements confirmed that the models' alternative themes were almost always plausible, so the lower numbers reflect annotation granularity rather than genuine model weakness.

Task	Italian	Russian	English	Overall
Sense Inventory	deepseek-chat (0.603)	deepseek-chat (0.817)	deepseek-re (0.796)	deepseek-re (0.726)
Sense Categories	o3 (0.629)	o3 (0.689)	o3 (0.670)	o3 (0.663)
Semantic Categories	chatgpt-4o (0.907)	deepseek-re (0.636)	gpt-4o (1.000)	chatgpt-4o (0.840)
Morphological Categories	–	o3 (0.960)	–	o3 (0.960)
Colligation L1	deepseek-re (0.949)	o3 (0.871)	o3 (0.896)	o3 (0.905)
Colligation R1	deepseek-re (0.918)	deepseek-re (0.919)	o3 (0.968)	deepseek-re (0.926)
Diaphasic Preference	o3-mini (0.851)	o3 (0.811)	o3 (0.799)	o3 (0.808)
Text Theme	o3 (0.736)	gpt-4.1 (0.727)	deepseek-re (0.654)	gpt-4.1 (0.689)

Table 5: Best-performing models (Jaccard scores in parentheses) for each linguistic task, ordered according to the specified task hierarchy. (zero-shot)

Task	Italian	Russian	English	Overall
Sense Inventory	deepseek-chat (0.604)	gpt-4.1 (0.827)	deepseek-re (0.823)	deepseek-re (0.715)
Sense Categories	deepseek-re (0.597)	o3 (0.687)	o3-mini (0.605)	deepseek-re (0.606)
Semantic Categories	gpt-4o-mini (0.933)	gpt-4o (0.635)	chatgpt-4o (1.000)	gpt-4o-mini (0.845)
Morphological Categories	–	o3 (0.986)	–	o3 (0.986)
Colligation L1	o3 (0.940)	o3-mini (0.886)	o3 (0.909)	o3 (0.908)
Colligation R1	deepseek-re (0.889)	o3 (0.914)	deepseek-re (0.941)	deepseek-re (0.913)
Diaphasic Preference	deepseek-chat (0.450)	o3 (0.801)	o3 (0.795)	o3 (0.678)
Text Theme	chatgpt-4o (0.463)	gpt-4o (0.686)	deepseek-re (0.676)	chatgpt-4o (0.602)

Table 6: Best-performing models (Jaccard scores in parentheses) for each linguistic task, ordered according to the specified task hierarchy. (3-shots)

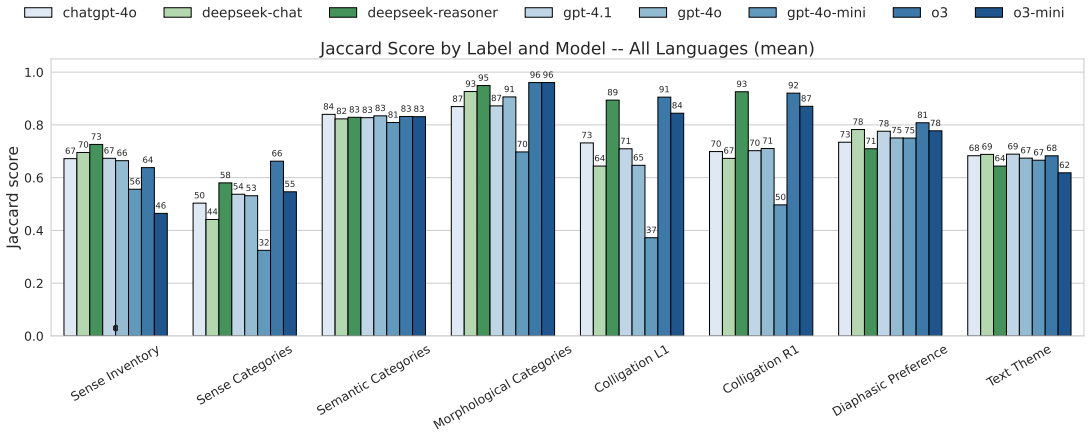


Figure 9: Mean Jaccard scores across different sense-related categories for multiple models, averaged over all languages, in the zero-shot setting.

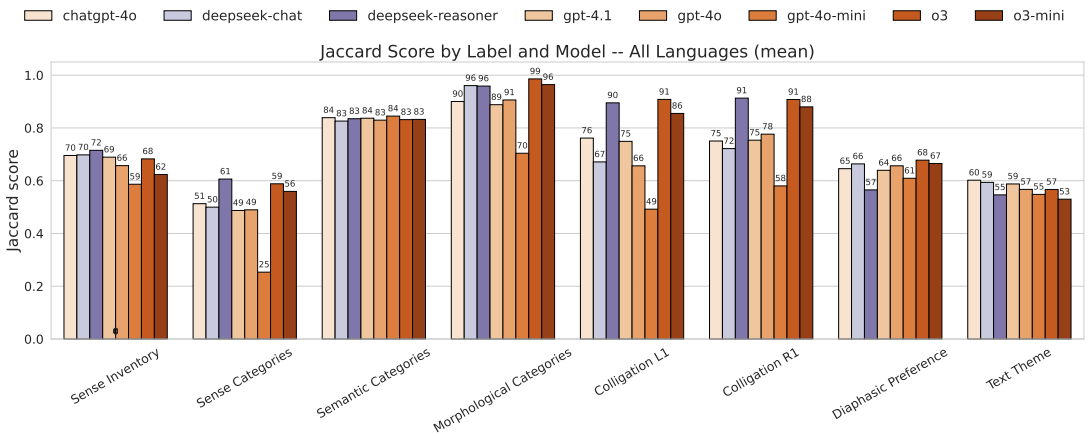


Figure 10: Mean Jaccard scores across different sense-related categories for multiple models, averaged over all languages, in the few-shot setting.

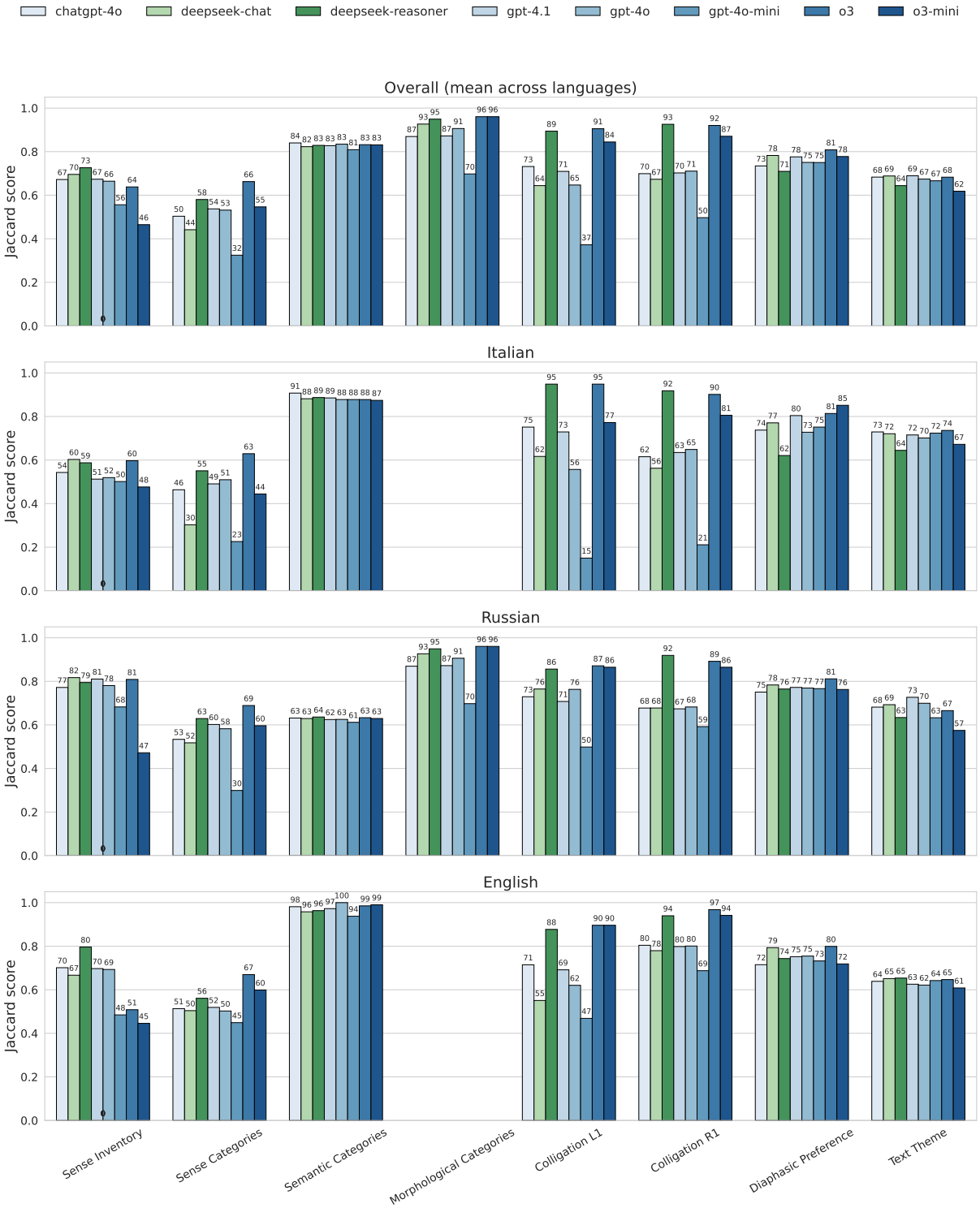


Figure 11: Model performance across different languages and categories in the zero-shot setting.

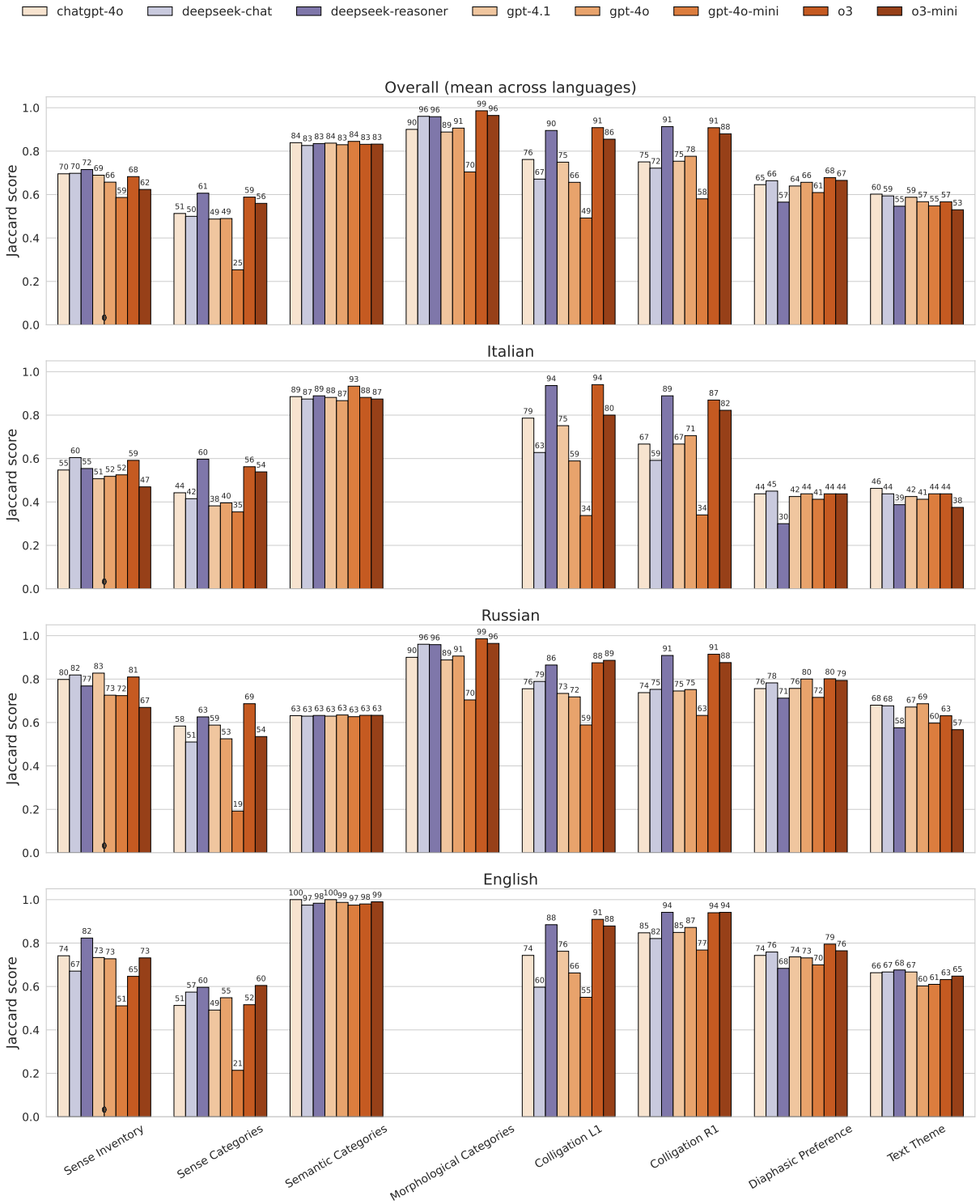


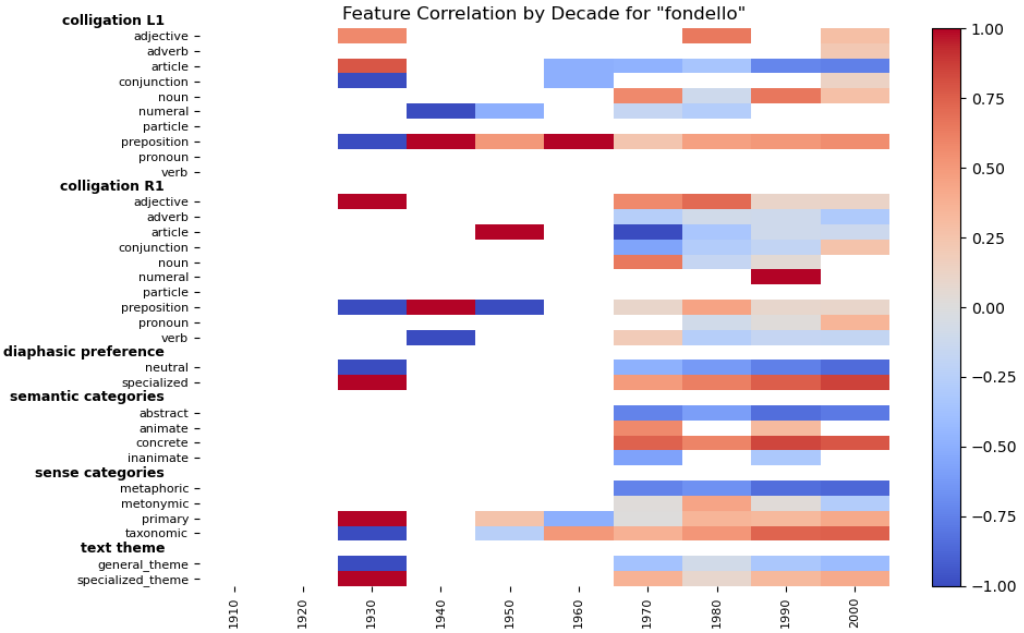
Figure 12: Model performance for individual and all languages. (3-shots)

D Latency and Parallelization

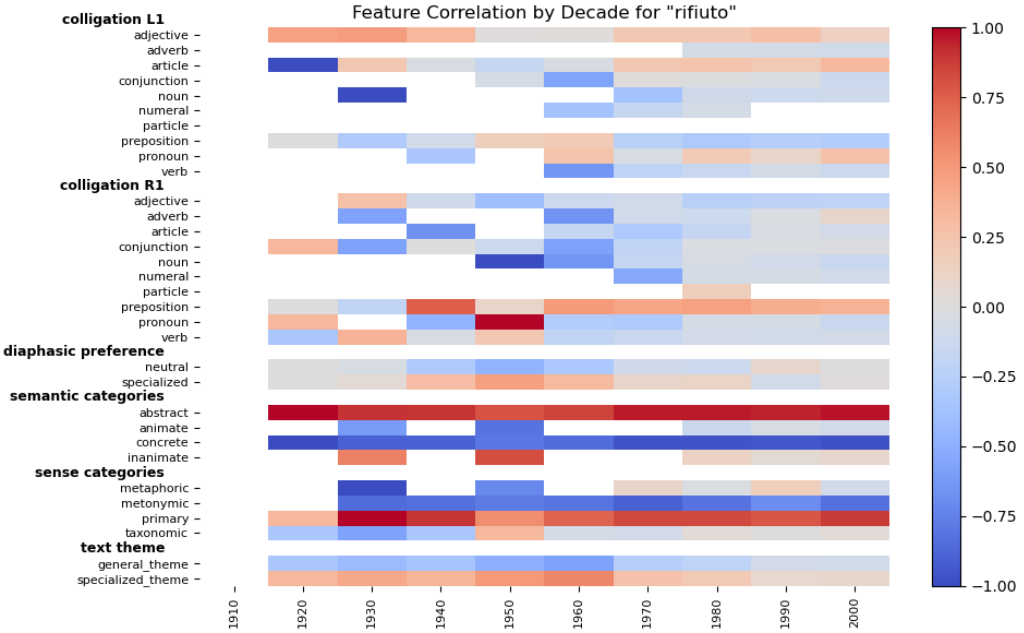
Pilot-scale efficiency study. To obtain an early read on cost–latency trade-offs we first ran a **small set** of 763 multilingual contexts, each with three prompts, for a total of $763 \times 3 = 2{,}289$ sequential calls per model. It should be noted that the context window size varied significantly by prompt type: Prompt 1 (*sense inventory*) necessitated the inclusion of full dictionary entries, which for highly polysemous words in the OED resulted in substantially higher input token counts compared to the syntactic and semantic prompts. Across the eight candidates—*GPT-4.1*, *ChatGPT-4o*, *GPT-4o*, *GPT-4o-mini*, *O3*, *O3-mini*, *DeepSeek-Chat*, and *DeepSeek-Reasoner*—latency varied by over an order of magnitude. The two reasoning-optimized OpenAI models, **O3** and **O3-mini**, were the slowest, requiring 16 ± 1 s per request and thus $\approx 36{,}600$ s (~ 10.2 h) for the whole pilot. All GPT-4o variants and *DeepSeek-Chat* answered in ~ 1 s each, finishing in $\approx 2{,}300$ s (~ 38 min). *DeepSeek-Reasoner* sat between, averaging 6 s per call and completing in $\approx 13{,}700$ s (~ 3.8 h).

Large-scale annotation run. The production evaluation encompassed $494{,}693$ contexts, again with three prompts each, for $1{,}484{,}079$ total API calls. If executed sequentially, *DeepSeek-Chat* would require 1.48×10^6 s (≈ 412 h or 17 days) at 1 s per call, while *DeepSeek-Reasoner* would stretch to 8.90×10^6 s ($\approx 2{,}473$ h or 103 days) at 6 s per call. By distributing requests across a compute cluster with > 400 concurrent workers per model, we cut the wall-clock duration of the entire experiment to roughly one week, transforming what would otherwise have taken months of idle waiting into a manageable engineering task.

E Feature Correlation Plots



(a) *fondello*.



(b) *rifiuto*.

Figure 13: Feature correlations across decades.

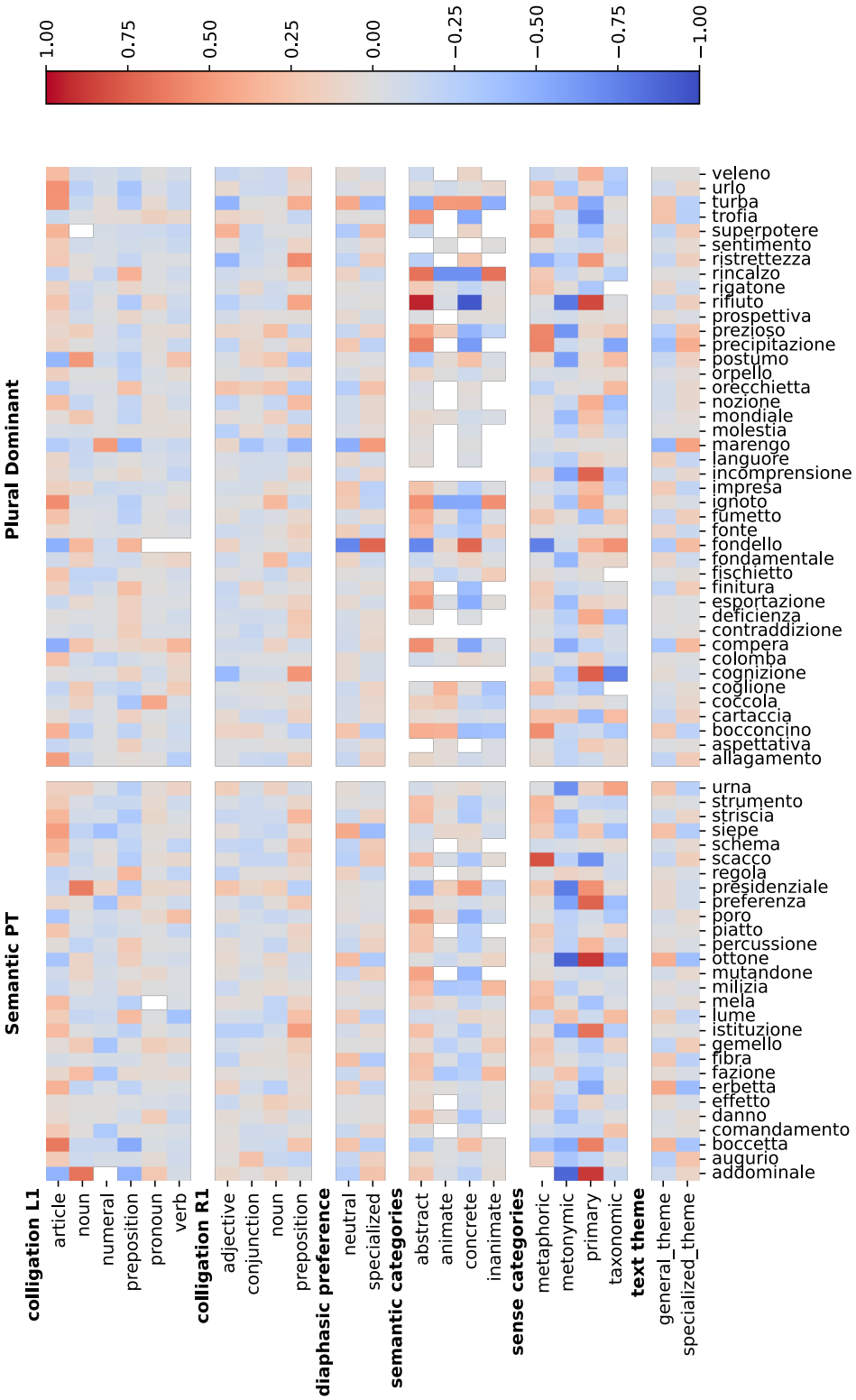


Figure 14: Correlations between grammatical number (plural/singular) and linguistic features in Italian PT conditions. Positive correlation with plural in blue and with singular in red.

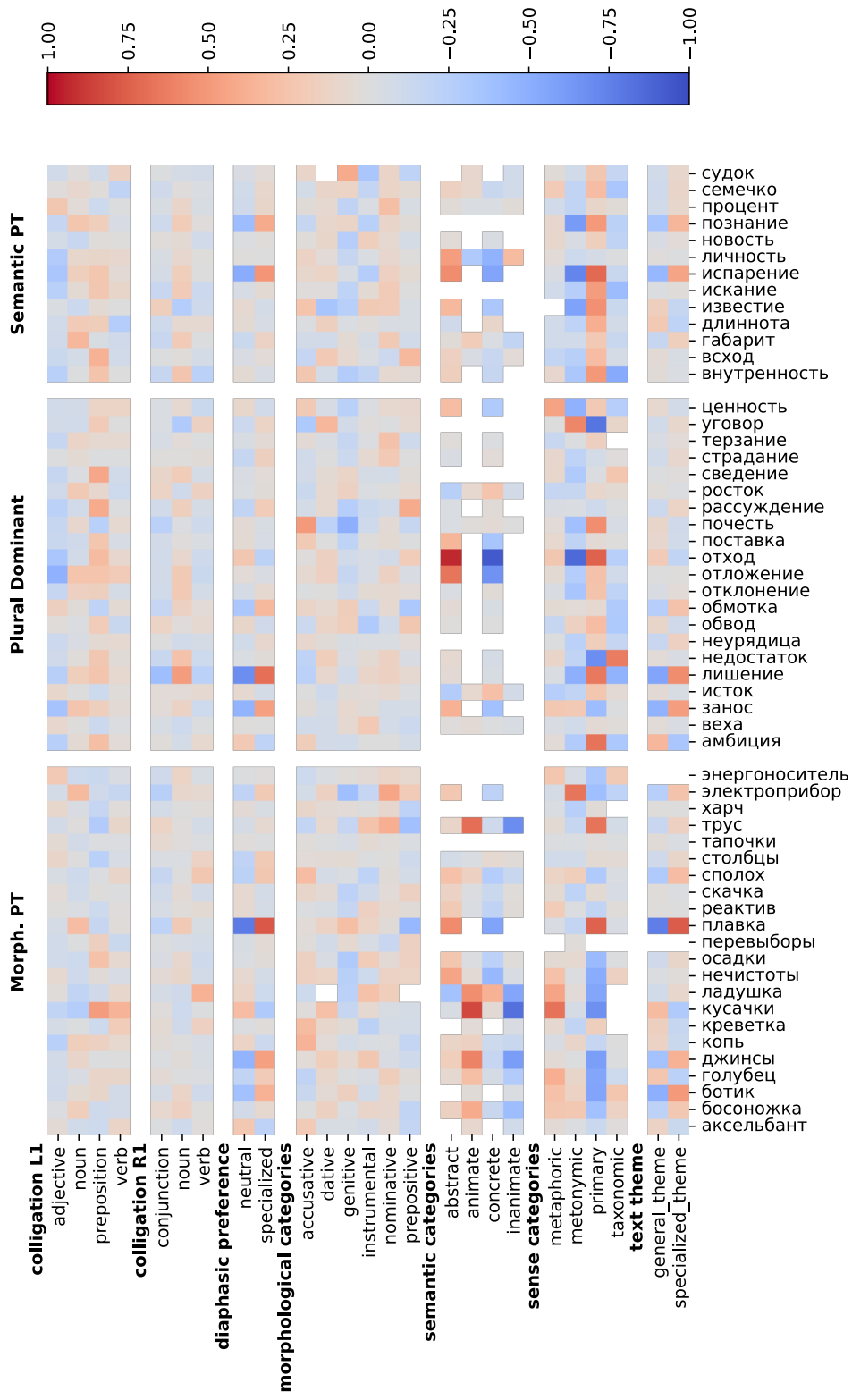


Figure 15: Correlations between grammatical number (plural/singular) and linguistic features in Russian PT conditions. Positive correlation with plural in blue and with singular in red.

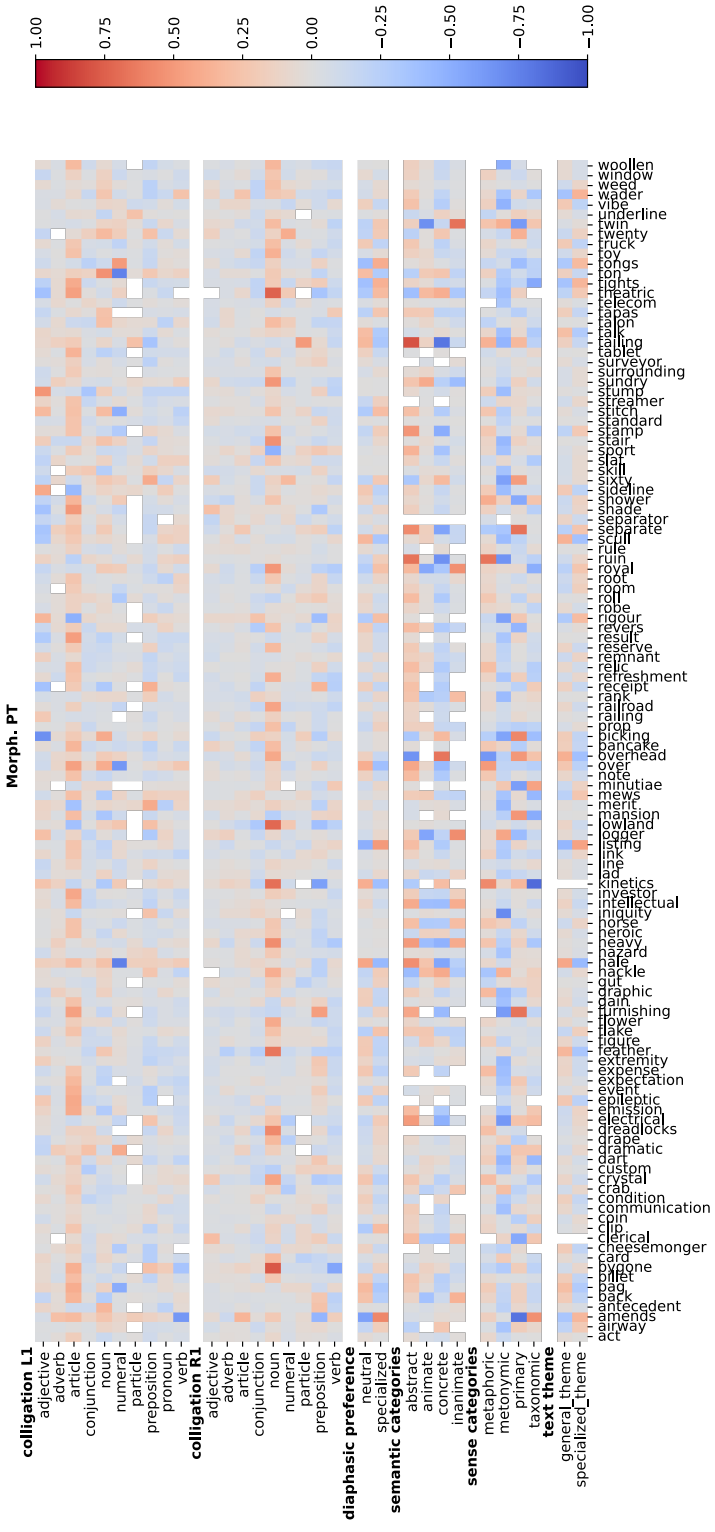


Figure 16: Correlations between grammatical number (plural/singular) and linguistic features in English PT conditions (subsampled for visualization). Positive correlation with plural in blue and with singular in red.

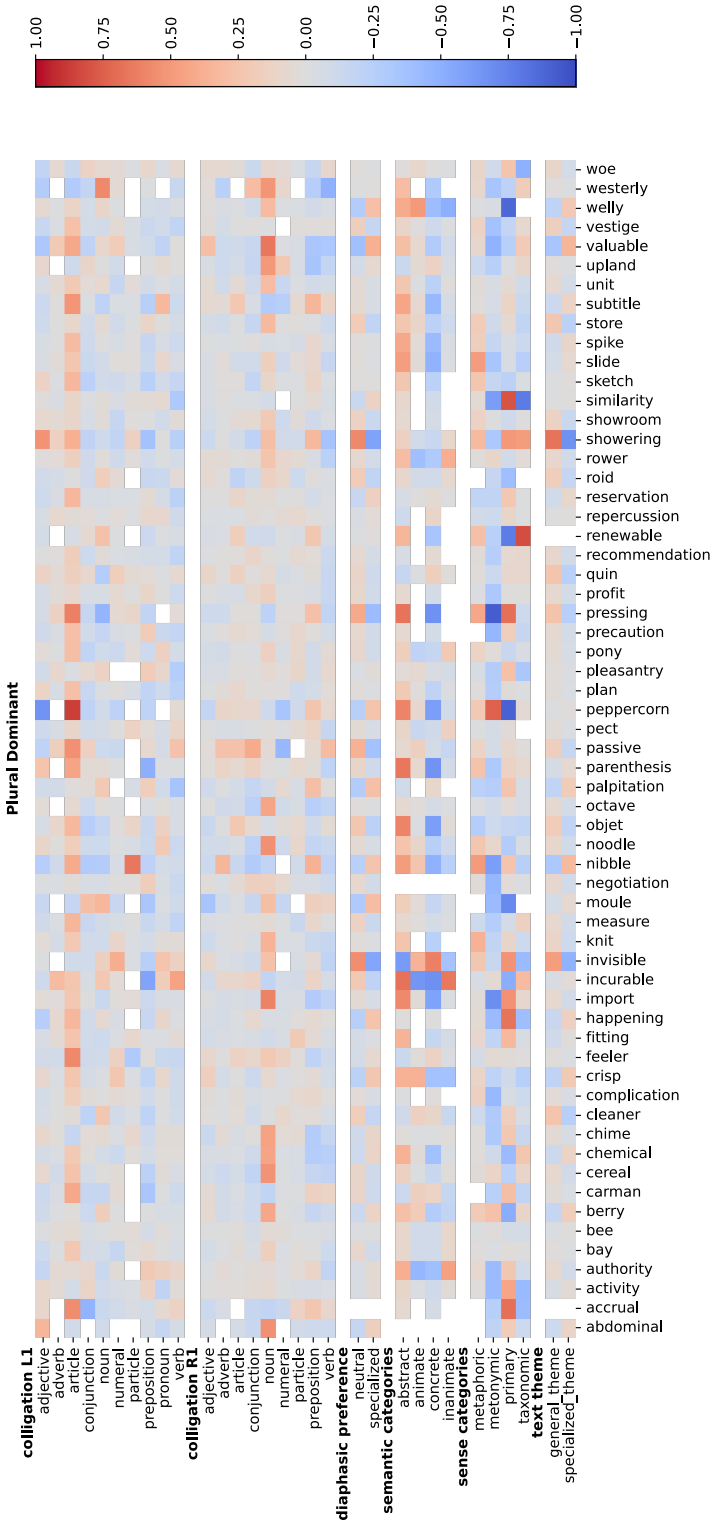


Figure 17: Correlations between grammatical number (plural/singular) and linguistic features in English PT conditions (subsampled for visualization). Positive correlation with plural in blue and with singular in red.

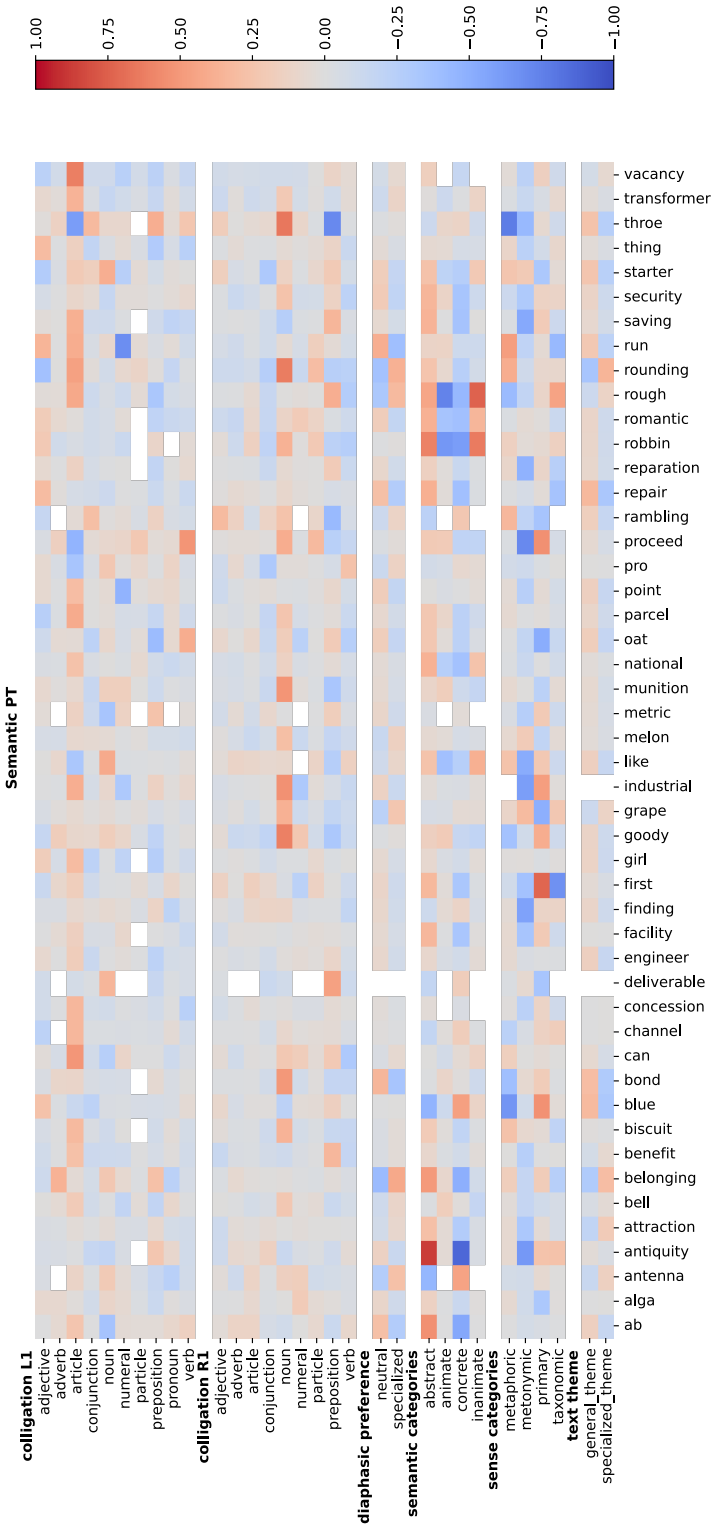


Figure 18: Correlations between grammatical number (plural/singular) and linguistic features in English PT conditions (subsampled for visualization). Positive correlation with plural in blue and with singular in red.