# The Alignment of Form and Function: Corpus-Based Evidence

**Dunstan Brown, Carole Tiberius*, Greville Corbett**

Surrey Morphology Group
University of Surrey
Guildford, Surrey
GU2 7XH
United Kingdom
{d.brown,g.corbett}@surrey.ac.uk

Institute for Dutch Lexicology INL*
Witte Singel / Doelen
Matthias de Vrieshof 2-3
2311 BZ Leiden
The Netherlands
tiberius@inl.nl

## Abstract

This paper analyses constraints on inflectional syncretism and inflectional allomorphy using frequency information. Syncretism arises where one form is associated with more than one function, whereas inflectional allomorphy occurs where there is more than one inflectional class, and a single function is associated with two or more forms. If high frequency is associated with more differentiation on both sides, we expect, on the one hand, that a frequent function will have a high number of forms and, on the other, that a frequent form will have a high number of functions. Our study focusses on Russian nominals, in particular nouns, which exhibit both syncretism and inflectional allomorphy. We find that there is a relationship between frequency and differentiation, but that it is not exceptionless, and that the exceptions can be understood in terms of the use of referrals as default rules.

**Keywords:** inflectional allomorphy, syncretism, function, form, Russian, nominals, frequency, Network Morphology.

## 1 Introduction

Grammatical paradigms define the relationship between the two sides of language, functions and forms[1]. For 'canonical' inflection we expect that a single form has one function, and that a single function has one form. For Russian, for instance, the singular and plural number can be combined with any of six cases[2], yielding 12 combinations of case and number. For these 12 functions we would expect a matching set of 12 forms (Figure 1).

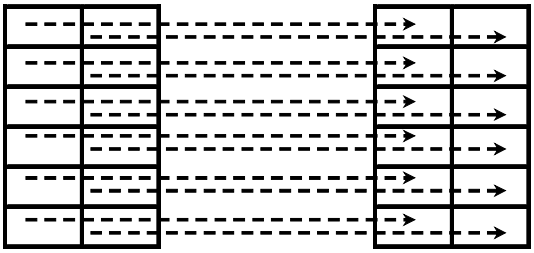


**Figure 1** One-to-one mapping between form and function

However, there are two well-known phenomena which contravene this idealised view: syncretism, where one form is associated with many functions, and inflectional allomorphy, where there is more than one inflectional class, and a single function is therefore associated with two or more forms. Syncretism can be illustrated with the English verb form *hit* which is ambiguous between its function as a past tense (*Mary hit the nail with a hammer*) and as a participle (*Mary was hit by a meteorite*). A simple instance of inflectional allomorphy can be found in Dutch nouns where the plural function corresponds to the inflections *–en* or *–s*. One problem that syncretism poses is that it is difficult to associate an inflection with a particular basic function; the opposite problem is posed by inflectional classes, where it is difficult to associate a basic function with a particular inflection. While there have been a number of proposals regarding constraints on inflectional syncretism and inflectional allomorphy in theoretical linguistics, little has been done on using frequency information to address these two issues. An obvious way of determining a function's basic form is to use frequency information, where that function's most

commonly occurring form is taken as its basic form. Equally, a form's most frequent function could be taken as its basic function. Adopting this perspective, we investigate the relationship between form and function for Russian nominals (nouns and adjectives), where we find instances of both syncretism and inflectional allomorphy.

It is important to note that these phenomena are a matter of degree. Sometimes an inflectional class will share inflections with other inflectional classes, to the extent that there may be no allomorphy if a particular inflection is shared across all classes. Equally, syncretism may occur within a lexical item, within a class, or across more than one class. In the tables below we intentionally abstract away from concrete instances and illustrate the range of possibilities using arbitrary symbols.

| | **Class A** | **Class B** | **Class C** | **Class D** |
|---|---|---|---|---|
| **Cell 1** | a | b | c | d |
| **Cell 2** | e | f | g | g |
| **Cell 3** | h | i | i | i |
| **Cell 4** | k | l | m | n |
| **Cell 5** | o | p | m | n |
| **Cell 6** | q | r | q | n |

**Table 1** Example paradigm with instances of inflectional allomorphy and syncretism

We can interpret a cell in the table as corresponding to a particular function, which is akin to an individual property or property combination, such as nominative singular, within a paradigm (following Carstairs-McCarthy (1996:323). We shall use the term 'paradigm' for the entire set of cells of combinations.[3] The letters in each of the cells are placeholders for actual morphological realisations. For instance, Cell 1 could be the combination nominative singular corresponding to four different forms. In Table 1, Cell 1 shows full allomorphy. It has a different inflection for each of the inflectional classes. For Cell 2, two out of the four inflectional classes share the same inflection. For Cell 3, the same inflection is used in three of the four classes. We also find instances of syncretism in this table. In Class C, Cell 4 and 5 are syncretic, whereas in Class D, three out of the 6 cells are syncretic. However, the situation can be even more complex, as the next table shows.

| | Class A | Class B | Class C | Class D |
|---|---|---|---|---|
| **Cell 1** | a | b | c | d |
| **Cell 2** | e | f | g | g |
| **Cell 3** | h | i | i | i |
| **Cell 4** | k | l | m | n |
| **Cell 5** | o | p | m | n |
| **Cell 6** | q | r | n | n |

**Table 2** Example paradigm with interaction between inflectional allomorphy and syncretism

Here we find interaction between inflectional allomorphy and syncretism in the realisation of Cell 6. In Class C and D, Cell 6 is realised by the form *n* which is syncretic with Cell 4 and 5 in Class D. Thus, on the one hand, there appear to be default realisations for cells, i.e. the form *g* for Cell 2, the form *i* for Cell 3, and the form *n* for Cell 6. On the other hand, when there is syncretism, there appear to be default associations between cells. For example, while cells 4 and 5 contain different forms in class C and class D, they are identical (i.e. syncretic) within each class, indicating a systematic association between these cells. It is hard to envisage tackling relationships such as these without recourse to a hierarchical model.

## 2 The Russian Data

### 2.1 Network Morphology analysis

Given the considerations above, we therefore use a formal theoretical treatment of Russian morphology developed within the Network Morphology framework (Corbett and Fraser 1993; Brown 1998a). Network Morphology is a linguistically motivated framework structuring morphological information in a default inheritance model. This means that the morphology can be represented as a hierarchy in which information is pushed as far up as it can go, capturing as many generalisations as possible. The term *default* means that information can be overridden, i.e. information specified under a particular class in the hierarchy takes precedence over what is inherited.

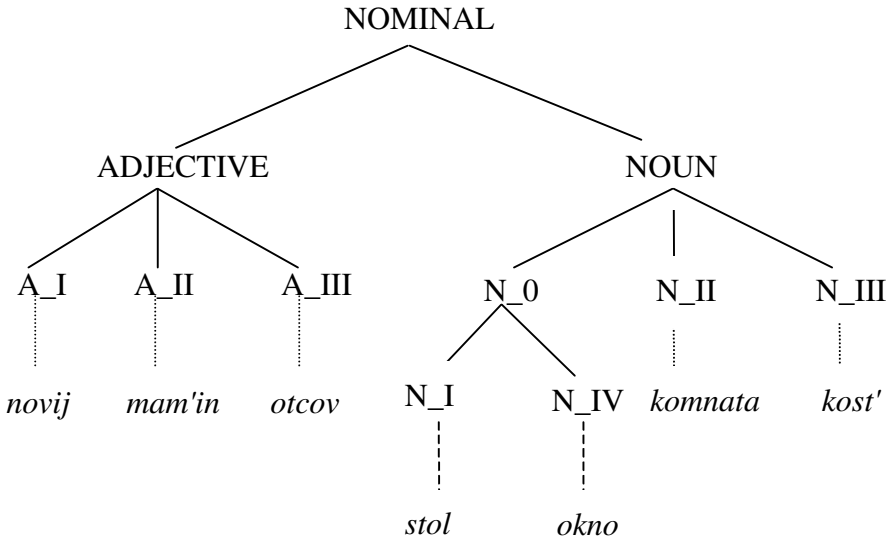


**Figure 2** Hierarchical structure of Russian nominal morphology[4]

Figure 2 represents a hierarchical structure for Russian nominal morphology using the Network Morphology framework (Corbett and Fraser 1993) with example lexemes at the bottom of the hierarchy. At the top of the hierarchy we find information associated with all nominals (such as the inflections for the dative, instrumental and locative plural which are shared between nouns and adjectives) and that information is propagated to others by inheritance, and at the bottom we find information which is unique to particular instances. In our hierarchy, node N_I (representing Noun Class I) and node N_IV (representing Noun Class IV) both inherit from node N_0, which represents the shared inflections for the genitive, dative, instrumental and locative singular of the classes N_I and N_IV. As this is a default inheritance network, information can be overridden lower down in the hierarchy. For example, the value for locative singular which is *stem+e* in three of the four inflectional noun classes can be stated at the NOUN node and its value is overridden for Class III (N_III) by *stem+i*.

Two points should be stressed. First, the original analysis was carried out with the goal of contributing to morphological theory, a goal which was achieved (see comments in Stump 2001:275-6). Second, in order to demonstrate that the analysis was valid, a substantial fragment of Russian, sufficient to include all instances of irregularity was implemented in the lexical

knowledge representation language DATR (Evans and Gazdar 1996) and is available at the DATR archive from the DATR webpages (http://www.datr.org).

## 2.2 Inflectional allomorphy in Russian nominals

Our formal theoretical analysis distinguishes four noun classes and three adjective classes. The forms for the major noun classes are shown in Table 3.[5]

| | **I** | **IV** | **II** | **III** |
|---|---|---|---|---|
| **Singular** | zavod 'factory' | delo 'thing' | komnat-a 'room' | kost´ 'bone' |
| Nom | zavod | del-o | komnat-a | kost´ |
| Acc | zavod | del-o | komnat-u | kost´ |
| Gen | zavod-a | del-a | komnat-i | kost´-i |
| Dat | zavod-u | del-u | komnat-e | kost´-i |
| Instr | zavod-om | del-om | komnat-oj | kost´-ju |
| Loc | zavod-e | del-e | komnat-e | kost´-i |
| **Plural** | | | | |
| Nom | zavod-i | del-a | komnat-i | kost-i |
| Acc | zavod-i | del-a | komnat-i | kost´-i |
| Gen | zavod-ov | del- | komnat | kost´-ej |
| Dat | zavod-am | del-am | komnat-am | kost´-am |
| Instr | zavod-am´i | del-am´i | komnat-am´i | kost´-am´i |
| Loc | zavod-ax | del-ax | komnat-ax | kost´-ax |

**Table 3** Forms for major noun classes in Russian

This table shows that it may be difficult to associate a basic function with a basic inflection. For instance, the locative singular can be realised as *stem+e* or *stem+i.* However, we see that *stem+e* is the realisation of locative singular for three out of the four classes. What is not shown in Table 3 is how frequent each of these classes is. Looking at classes alone we could argue that *stem+e* is the basic form for the function locative singular. The next question is whether locative singular is the basic function for *stem+e*. If *stem+e* were restricted to the locative singular, then the answer would be trivial. However, *stem+e* can also be the realisation of dative singular for nouns of Class II. As it is restricted in this function to Class II nouns only, it is reasonable to conclude that locative singular is the basic function for *stem+e*, as the form has the locative singular function in three classes, whereas it has the dative singular function only in one. However, our goal is to determine whether this argumentation, which is based on inflectional classes, matches with textual frequency.

## 2.3 Syncretism in Russian nominals

Russian nominals have two number values (singular and plural), six cases (nominative, accusative, genitive, dative, instrumental, and locative), and three genders (masculine, feminine and neuter) which can be combined yielding 12 combinations of case and number and 36 combinations of case, number and gender. Both nouns and adjectives have number and case, however gender is an inflectional category only for adjectives. (In our investigation we do not consider separately the two minor cases of nouns, the second locative (Brown forthcoming) and the second genitive. They are treated as part of locative and genitive.) Despite the figures for possible combinations of case, number and gender, a typical Russian noun does not have more than 10 forms (Table 3) and a typical adjective does not have more than 14 forms.

For example, a Class III noun such as *kost´* ('bone') uses the same form for its genitive, dative and locative singular. Russian also has syncretism related to animacy. In the singular, masculine animate nouns which belong to Class I form their accusative on the basis of the genitive form. For classes IV and III, which are associated with neuter and feminine genders respectively, there is always nominative/accusative syncretism. Class II has a separate form for the accusative. In the plural, the situation is more straightforward: any animate noun forms its accusative on the basis of the genitive, and any inanimate noun forms its accusative on the basis of the nominative. In Table 3 only examples of inanimate nouns are given.

Animacy related syncretism is illustrated by the examples below which are taken from the Russian Standard Corpus (Sitchinava 2001; Sharoff 2006). The form *art´istov* is syncretic between genitive and accusative plural. It functions as an accusative in the first example and as a genitive in the second example.

(1) Артистов прошу оставаться на месте!

| | | | | |
|---|---|---|---|---|
| **art´ist-ov** | proš-u | ostavats´a | na | mest-e! |
| artist-ACC.PL | ask-1SG | remain | on | place-LOC.SG |

I ask the performers to remain where they are!

(2) [...] тщеславие посредственных артистов [...]

| | | |
|---|---|---|
| tščeslavijo | posredstvenn-ix | **art´ist-ov** |
| vanity | mediocre-GEN.PL | artist-GEN.PL |

the vanity of mediocre performers

As in the above examples, most morphological syncretisms can be readily disambiguated from the syntactic context, but our purpose is to demonstrate that a morphologically complex language such as Russian still leaves much work to syntax.

There are different ways of analysing syncretism. One way is underspecification, where the form in question is treated as not specified for any of the syncretised functions in the theoretical analysis from which the morphological model can be derived. Another way is referrals (Zwicky 1985; Stump 2001: 212-41), where the form is associated with a basic function, and other cells in the paradigm refer to the cell with this basic function.

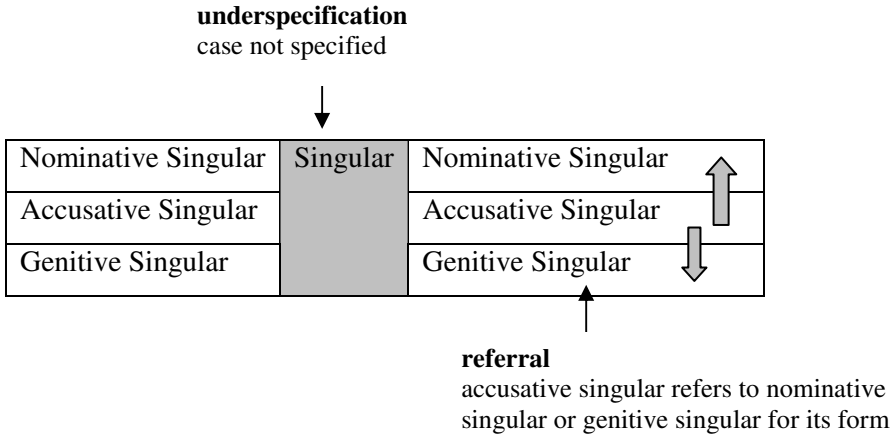


**Figure 3** Illustration of Underspecification and Referrals

Referrals are therefore asymmetrical in their nature, whereas underspecification is not. There is evidence that at least both types of analysis are required (Stump 2001: 212-41; Baerman, Brown and Corbett 2005: 133-170), and it can be argued that a kind of underspecified referral is required for analysing syncretisms in Slovene and Dalabon (Evans, Brown & Corbett 2001: 216; Baerman, Brown and Corbett 2005: 186-204). Therefore we cannot dispense with one at the expense of the other. It is worthwhile examining whether the theoretical asymmetry of referrals can be observed in language use.

In the Network Morphology analysis, syncretisms within paradigms are treated as asymmetrical, in that a particular form is considered to have one function as basic. For example, in Class II dative and locative singular are syncretic, *stem+e.* In the formal model, the locative is assumed to be the basic function as in three of the four inflectional classes the value for locative singular is *stem+e*. If this paradigmatic asymmetry is reflected in frequency distributions, then we expect one function to be more important, i.e. the referred-to cell occurs more frequently than the cell which refers to it. Thus in the case of the dative/locative singular syncretism in Class II, we expect the locative to be more frequent than the dative.

## 3 Corpus Data

For our study, we used data from the 1.5 million word Russian Standard Corpus (Sitchinava 2001; Sharoff 2006), which is fully tagged. The corpus was split into two parts (500,000 word forms (or tokens) and 1 million word forms (tokens)) which allowed us to check our results for consistency. From this data two spreadsheets were automatically created containing frequency information for the different functions of Russian nominals. The lexemes (or types) recorded in the dataset are those represented by word forms occurring in total at least five times. Lexemes occurring less than five times were excluded to avoid large standard errors in the estimates which occur when observed numbers in each category are small (Corbett, Hippisley, Brown and Marriott 2001:208). The resulting datasets contain 8762 noun lexemes (types) (285895 word

forms (tokens)), 3683 adjective lexemes (types) (86033 word forms (tokens) without the comparatives) in total.

## 4 Alignment of Form and Function

As the aim of this paper is to see whether frequency allows us to determine a basic exponent in instances of syncretism, we are going to align forms and functions for Russian nominals based on the Network Morphology analysis described above. Before we start we need to make clear what we mean by form and function. We define forms abstractly as unique realisations within a paradigm. Following Carstairs-McCarthy (1996) we take a paradigm to be an entire set of features or feature combinations. The definition of a function is less clear-cut. We distinguish two different approaches depending on how functions are counted.

### 4.1.1 Method 1

In the first method, we start from a paradigm table and each cell of the paradigm counts as a function. Thus, if we take the Russian noun paradigm, we get 12 functions, as illustrated below.

| | |
|---|---|
| NominativeSingular | NominativePlural |
| AccusativeSingular | AccusativePlural |
| GenitiveSingular | GenitivePlural |
| DativeSingular | DativePlural |
| InstrumentalSingular | InstrumentalPlural |
| LocativeSingular | LocativePlural |

**Table 4** Paradigm cells as functions

Given the various examples of inflectional allomorphy, there are potentially more forms than functions (see Table 3 for nouns). In fact, we find 14 different forms[6]. For instance, the function nominative singular for nouns can be realised by *stem+ø*, *stem+o*, and *stem+a* (Table 5). Do note that although nominative singular is realised as the bare stem, *stem+ø,* in Class I and III, this only counts as one possible form for the function nominative singular.

| | I | IV | II | III |
|---|---|---|---|---|
| **Nom Sing** | stem+ø | stem+o | stem+a | stem+ø |
| | zavod | del**o** | komnat**a** | kost´ |

**Table 5** Possible forms for the function nominative singular for nouns

For adjectives, we get 36 functions, as gender plays a role in the singular. For instance, nominative singular masculine counts as one function.

### 4.1.2 Method 2

In the second approach, we take a function to be a value of a morphosyntactic feature, i.e. number or case for nouns, and number, case or gender for adjectives. For example, the form *stem+ej* can be the realisation of an accusative plural and a genitive plural, and as such has three functions, i.e. plural, accusative and genitive.

## 4.2 Analysis

For each of the methods, the analysis involves a three-step process. First we analyse functions as sets of forms as is illustrated in Figure 4 for method 1. Second, we analyse forms as sets of functions as is illustrated by the picture in Figure 5 and finally, we align the results for form and function (Figure 6).

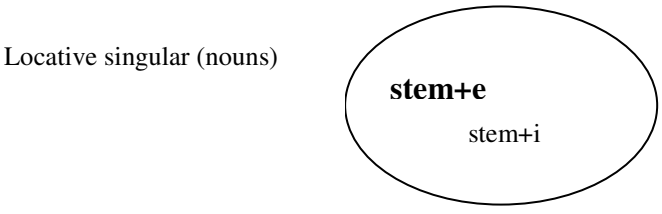


**Figure 4** Functions as sets of forms

Within nouns, the locative singular can be realised as *stem+e* or *stem+i*. The occurrence of *stem+e* functioning as a locative singular is, however, more frequent than the occurrence of *stem+i* functioning as a locative singular. In Figure 4 this is indicated by using bold, and a larger font, for the more frequent occurrence.

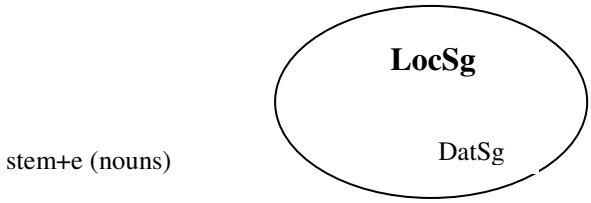


**Figure 5** Forms as sets of functions

Within nouns, *stem+e* can function as a locative singular or as a dative singular. However, it occurs more frequently as a locative singular than as a dative singular indicated by the larger bold font. In the last step we align these two sets, i.e functions as sets of forms and forms as sets of functions. This results in the picture below for the above example. The locative singular function for nouns is most often realised as a *stem+e*, and the form *stem+e* is most often used as a locative singular.

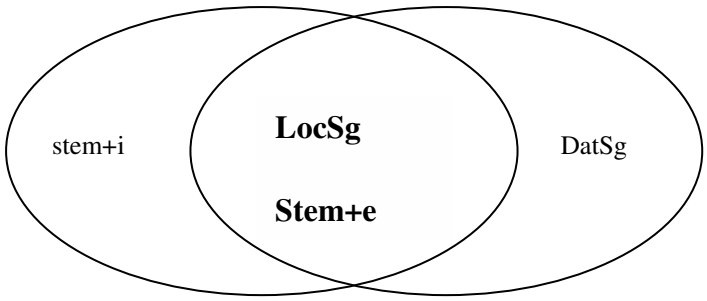


**Figure 6** Aligning form and function

That frequency is associated with greater differentiation is a fact that has already been noted. For example, Mańczak (1966: 84) attributes this to a general law, which associates greater use with greater differentiation. What is not always made clear is that there are two different expectations associated with claims of this type. On the one hand, we can consider the question from the point of view of functions and the associated number of forms, predicting that the higher a function's frequency the greater number of forms it will have associated with it. On the other hand, we can consider the question from the point of view of forms and associated number of functions, predicting that the higher a form's frequency the greater number of functions it will have associated with it.[7] If we apply the expectation that greater frequency means greater differentation to both sides, we might expect that an infrequent function should have a small number

of forms, and that an infrequent form should have a small number of functions. However, this expectation does not seem to be fulfilled by our results.

## 5 Analysis and Results

In this section, we present the analysis and results for the nouns using the dataset created on the basis of 1 million words from the corpus. A consistency check of our results on the second part of the corpus will be discussed in Section 6. The analysis of the long form adjectives did not provide any interesting differences from the noun results. Within adjectives, most functions are associated with at most one or two forms and therefore the adjective data do not provide evidence either way for the alignment of form and function. As such they will not be discussed in this paper. Full details of the analysis of adjectives can be found on our website (http://www.surrey.ac.uk/LIS/SMG/PiU/).

### 5.1 Method 1

#### 5.1.1 Functions to forms

In Table 6 we compare the frequency of a particular function with the potential number of forms which can realise it for Russian nouns. We give two separate columns. One column takes into account indeclinables in determining the potential number of forms. (In effect, this means counting *stem+ø* as a possible form for each function.) Another column excludes indeclinables in the count of potential forms. This means that the number of potential forms will be one less than in the column which includes indeclinables, unless the realisation *stem+ø* is one of the possible forms for declinable nouns, in which case the counts in each column will match. The nominative singular and genitive plural, for example, can already be realised by *stem+ø* for declinable nouns, and so exclusion of indeclinables does not affect the count of potential forms for them, whereas it does for genitive singular, for which declinable nouns must use an affix.

| Function | Frequency | N° of forms | N° of forms without indecl |
|---|---|---|---|
| **NomSg** | 40878 | 3 | 3 |
| **AccSg** | 35258 | 4 | 4 |
| **GenSg** | 31934 | 4 | 3 |
| **LocSg** | 16383 | 4 | 3 |
| **InstrSg** | 14937 | 4 | 3 |
| **NomPl** | 14269 | 4 | 3 |
| **GenPl** | 13812 | 3 | 3 |
| **AccPl** | 11651 | 5 | 5 |
| **DatSg** | 7907 | 4 | 3 |
| **LocPl** | 5698 | 2 | 1 |
| **InstrPl** | 4265 | 2 | 1 |
| **DatPl** | 2413 | 2 | 1 |

**Table 6** Functions to forms for Russian nouns

Note that it would not make a difference if second genitive and second locative were treated separately (recall that we have included them with genitive and locative). They are very low in frequency. There are 1114 second locatives in the data and 347 second genitives. By adding them to the count of genitive and locative, the total frequency of those goes up slightly and both functions can be realised by an extra form, but this does not affect the overall frequency order of the functions in the table.

If there is a relationship between frequency and greater differentiation, then we expect the number of forms to decrease going down the table. In the count which includes indeclinables, the nominative singular and the genitive plural do not fit with this expectation, because indeclinables are counted as *stem+ϕ*, and this is already a possible realisation of nominative singular and genitive plural. So while counting indeclinables does not increase the number of forms for the nominative singular and genitive plural, it does, however, add one extra form for all the other functions. Once indeclinables are excluded the picture alters, and two clearcut instances are left where decreasing frequency fails to fit with decreasing differentiation of form: the accusative singular and accusative plural.

The table shows that accusative singular and accusative plural stand out because they can be realised by a greater number of forms than would be expected on the basis of their frequency. This is an instance of syncretism interacting with inflectional allomorphy. As we noted earlier,

Russian has animacy-related syncretism. In the singular, masculine animate nouns, which belong to Class I, form their accusative on the basis of the genitive form. For classes IV and III, which are associated with neuter and feminine genders respectively, there is always nominative/accusative syncretism in that the accusative takes over its form from the nominative. Class II has a separate form for the accusative. In the plural, the situation is more straightforward: any animate noun forms its accusative on the basis of the genitive, and any inanimate noun forms its accusative on the basis of the nominative. Thus, depending on animacy the accusative takes over its form from the nominative or the genitive. Because the accusative singular is based on either the genitive or the nominative, it has more forms than the more frequent nominative singular. This lack of correspondence between function frequency and number of forms appears to be associated with referral-based syncretisms, where one paradigm cell is referred to another for its form. With the accusative plural the effect is even more apparent. It should be noted that, for nouns overall, the accusative plural is less frequent than both the nominative and genitive plural, but for all noun lexemes in the plural the animacy rule applies, and so the number of forms that the accusative plural may have is the sum of the number of forms for the nominative and genitive plural together. These facts do not fit with the general claim that higher frequency means greater differentiation. A typical intuitive assumption concerning the relationship between low frequency and less differentation is that it would be more taxing on memory to learn many forms for a function which occurs infrequently. Equally, however, this argumentation could be applied in support of referrals. If we assumed that in order to learn a referral-based system it is only necessary to acquire the rule which says that the form of the accusative is the same as the genitive (if animate) or nominative (if inanimate), then this is possibly less taxing on memory than learning all of the inflections as directly associated with the accusative. Furthermore, we associate greater regularity with lower frequency. Referral-based syncretisms are therefore interesting when viewed from this perspective, as they can create unexpected form effects for less frequent functions (i.e. greater differentiation), while at the same time conforming with the expectation that lower frequency and greater regularity go together. It further also suggests that the

relationship between greater regularity and lower frequency may be of greater importance than the association between low differentiation and low frequency. This is, of course, a matter for psycholinguistic investigation.

### 5.1.2 Forms to functions

We now take the forms of Russian nouns and determine the number of functions they can realise. The results are given in Table 7.

| **Form** | **Frequency** | **N$^{o}$ of functions** |
|---|---|---|
| **stem+ø** | 47409 | 12 (4) |
| **stem+a** | 37410 | 5 |
| **stem+i** | 33914 | 5 |
| **stem+u** | 17801 | 4 |
| **stem+e** | 16095 | 3 |
| **stem+o** | 10384 | 2 |
| **stem+om** | 9041 | 1 |
| **stem+ov** | 7361 | 2 |
| **stem+ax** | 5683 | 1 |
| **stem+ami** | 4262 | 1 |
| **stem+oj** | 4229 | 1 |
| **stem+am** | 2412 | 1 |
| **stem+ej** | 1754 | 2 |
| **stem+ju** | 1650 | 1 |

**Table 7** Forms to functions for Russian nouns

We see that there is an association between high frequency of a realisation/form and the number of functions which it may fulfil. However, the relationship does not involve a straightforward decrease as the forms become less frequent. The high number of functions that can be realised by the form *stem+ø* has already been explained and is due to the fact that indeclinables are analysed as *stem+ø* for each function. If we take the indeclinables out, the number of functions that *stem+ø* can realise goes down to 4. In this case, *stem+ø* no longer fits the pattern, as it has fewer functions but is more frequent than *stem+a*. More noticeably, it is again the realisations

which have some involvement with the animacy rule which do not fit the pattern: *stem+ø, stem+a, stem+i, stem+ov, stem+ej*.

### 5.1.3 Alignment of form and function

We now aim to see whether the results for forms and functions align by cross-tabulating the forms (columns) against the functions (rows). If function *x* is the most frequent function of form *y*, and form *y* is the most frequent form of function *x*, then we can treat them as being aligned on a frequency basis. We argue that where the two distributions line up, as in the highlighted cells in Table 8, we can determine a basic exponent in instances of syncretism. For example, *stem+ø* is the most frequent form of the function nominative singular, and nominative singular is the most frequent function of the form *stem+ø*.[8]

| | **ø** | **a** | **i** | **u** | **e** | **o** | **om** | **ov** | **ax** | **ami** | **oj** | **am** | **ej** | **ju** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **NomSg** | 24088 | 11391 | 0 | 0 | 0 | 5399 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 40878 |
| **AccSg** | 16375 | 2502 | 0 | 11396 | 0 | 4985 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 35258 |
| **GenSg** | 197 | 18826 | 12564 | 347 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 31934 |
| **LocSg** | 157 | 0 | 1337 | 1114 | 13775 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 16383 |
| **InstrSg** | 17 | 0 | 0 | 0 | 0 | 0 | 9041 | 0 | 0 | 0 | 4229 | 0 | 0 | 1650 | 14937 |
| **NomPl** | 28 | 2405 | 11743 | 0 | 93 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 14269 |
| **GenPl** | 6076 | 0 | 0 | 0 | 0 | 0 | 0 | 6201 | 0 | 0 | 0 | 0 | 1535 | 0 | 13812 |
| **AccPl** | 380 | 2286 | 7606 | 0 | 0 | 0 | 0 | 1160 | 0 | 0 | 0 | 0 | 219 | 0 | 11651 |
| **DatSg** | 72 | 0 | 664 | 4944 | 2227 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7907 |
| **LocPl** | 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5683 | 0 | 0 | 0 | 0 | 0 | 5698 |
| **InstrPl** | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4262 | 0 | 0 | 0 | 0 | 4265 |
| **DatPl** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2412 | 0 | 0 | 2413 |
| | 47409 | 37410 | 33914 | 17801 | 16095 | 10384 | 9041 | 7361 | 5683 | 4262 | 4229 | 2412 | 1754 | 1650 | 199405 |

**Table 8** Alignment of form and function for nouns

It turns out that there is no alignment of form and function on a frequency basis for accusative singular or accusative plural. The most frequent form associated with accusative singular (*stem+ø*) does not have accusative singular as its most frequent function, nor does the most frequent form associated with accusative plural (*stem+i*) have accusative plural as its most frequent function. Because it is not possible to align form and function for the accusative singular and plural, we cannot assume a basic form for these functions. Nor can we assume accusative

singular as the basic function for any of the forms with which it is associated. Hence, we have a justification on frequency grounds for the asymmetry we have associated with referrals.

However, there is also no alignment for the nominative plural and the dative singular. The dative singular can be explained along similar lines as the accusative case. It is syncretic in Class II and III in the Network Morphology analysis taking over its form from the locative singular.

The nominative plural is a different matter. Although *stem+i* is by far the most frequent realisation of nominative plural, form and function do not align since *stem+i* more frequently realises genitive singular than nominative plural. If we look at the overall frequency of singular versus plural, we see that singular is about 3 times more frequent than the plural, which has an effect on the alignment of form and function for the nominative plural. If we split the above table by number and create a separate table for the plural cases, we find alignment for all functions in the plural except for the accusative.

| | **i** | **ov** | **Ø** | **ax** | **a** | **ami** | **am** | **ej** | **e** | |
|---|---|---|---|---|---|---|---|---|---|---|
| **NomPl** | 11743 | 0 | 28 | 0 | 2405 | 0 | 0 | 0 | 93 | 14269 |
| **GenPl** | 0 | 6201 | 6076 | 0 | 0 | 0 | 0 | 1535 | 0 | 13812 |
| **AccPl** | 7606 | 1160 | 380 | 0 | 2286 | 0 | 0 | 219 | 0 | 11651 |
| **LocPl** | 0 | 0 | 15 | 5683 | 0 | 0 | 0 | 0 | 0 | 5698 |
| **InstrPl** | 0 | 0 | 3 | 0 | 0 | 4262 | 0 | 0 | 0 | 4265 |
| **DatPl** | 0 | 0 | 1 | 0 | 0 | 0 | 2412 | 0 | 0 | 2413 |
| | 19349 | 7361 | 6503 | 5683 | 4691 | 4262 | 2412 | 1754 | 93 | 52108 |

**Table 9** Alignment of form and function for plural nouns

What this step suggests is that the structure of the paradigm may well be important in our consideration of frequency. While the form *stem+i* has genitive singular as its most frequent function, if we were to sum the frequencies of its plural functions (nominative plural and accusative plural), this would be greater than the sum of its singular functions (genitive singular, locative singular and dative singular). In fact, this fits with the original Network Morphology model, where a Category Dependency Constraint determines that case is dependent on number, that is, the number feature may determine the number of case distinctions, but not the other way round (Brown 1998b). The results found in Table 8 and 9 also suggest that we should consider the re-

lationship between form and frequency separately for number and case. We will do this in method 2.

### 5.2 Method 2

In this method, we take a function to be a value of a morphosyntactic feature rather than a combination of morphosyntactic features, i.e. number or case for nouns. In order to avoid counting some values twice in the same table (i.e. nominative singular under nominative, and under singular), we will map forms onto functions and functions onto forms for the features case and number separately.

#### 5.2.1 Functions to forms

The tables below give the number of forms that the different functions can realise. The first table gives the number of forms for the function number, the second for the function case. Although singular is almost three times more frequent than plural, both functions can be realised by 9 forms.

| Function | Frequency | N° of forms |
|---|---|---|
| **Singular** | 147297 | 9 |
| **Plural** | 52108 | 9 |

**Table 10** Functions to forms for number

With regard to the case table, it is important to note that the case functions are given regardless of number, i.e. nominative groups nominative singular and nominative plural. Apart from a higher number of forms than expected for the accusative – for the same reasons as given above - nothing interesting can be noted.

| Function | Frequency | N° of forms | N° of forms without indecl |
|---|---|---|---|
| **Nominative** | 55147 | 5 | 5 |
| **Accusative** | 46909 | 7 | 7 |
| **Genitive** | 45746 | 6 | 6 |
| **Locative** | 22081 | 5 | 4 |
| **Instrumental** | 19202 | 5 | 4 |
| **Dative** | 10320 | 5 | 4 |

**Table 11** Functions to forms for case

We conclude that splitting the functions into separate features does not provide a more useful insight into the data. The number of forms for the function genitive is six rather than five because the form of the second genitive is included in the count. The second locative is also included in the count for locative, and excluding it would also decrease the number of forms by one, with a concomitant decrease in frequency of the locative function. Hence, this method does not provide us with any new clearcut insight into the relationship between function frequency and form differentiation.

### 5.2.2 Forms to functions

We now map the forms onto functions, whereby a function is a value of a morphosyntactic feature. For instance, the form *stem+ju* can realise singular and instrumental which, under the method we are using in this section, counts as 2 functions. The form *stem+e*, on the other hand can be a singular dative, a singular locative, and a plural nominative. This counts as 5 different functions (i.e. the function singular is counted only once).

| **Form** | **Frequency** | **N$^{o}$ of functions** |
|---|---|---|
| **stem+ø** | 47409 | 8 |
| **stem+a** | 37410 | 5 |
| **stem+i** | 33914 | 7 |
| **stem+u** | 17801 | 5 |
| **stem+e** | 16095 | 5 |
| **stem+o** | 10384 | 3 |
| **stem+om** | 9041 | 2 |
| **stem+ov** | 7361 | 3 |
| **stem+ax** | 5683 | 2 |
| **stem+ami** | 4262 | 2 |
| **stem+oj** | 4229 | 2 |
| **stem+am** | 2412 | 2 |
| **stem+ej** | 1754 | 3 |
| **stem+ju** | 1650 | 2 |

**Table 12** Forms to functions for Russian nouns

The resulting table shows a similar pattern as found in Table 7. This suggests again that splitting functions to be values of morphosyntactic features rather than combinations thereof does not provide further insight.

### 5.2.3 Alignment of form and function

In order to see whether our results align for method 2, we cross-tabulate the forms (columns) against the functions (rows) again. Two separate tables are created for number values and case values. The cells where form and function align (i.e. where function *x* is the most frequent function of form *y* and form *y* is the most frequent form of function *x*) are highlighted in grey.

| Form | Singular | Plural | Total |
|---|---|---|---|
| **stem+ø** | 40906 | 6503 | 47409 |
| **stem+a** | 32719 | 4691 | 37410 |
| **stem+i** | 14565 | 19349 | 33914 |
| **stem+u** | 17801 | 0 | 17801 |
| **stem+e** | 16002 | 93 | 16095 |
| **stem+o** | 10384 | 0 | 10384 |
| **stem+om** | 9041 | 0 | 9041 |
| **stem+ov** | 0 | 7361 | 7361 |
| **stem+ax** | 0 | 5683 | 5683 |
| **stem+ami** | 0 | 4262 | 4262 |
| **stem+oj** | 4229 | 0 | 4229 |
| **stem+am** | 0 | 2412 | 2412 |
| **stem+ej** | 0 | 1754 | 1754 |
| **stem+ju** | 1650 | 0 | 1650 |
| | 147297 | 52108 | 199405 |

**Table 13** Alignment of form and number functions

Thus, from Table 13 we conclude that the function singular is most frequently realised by *stem+ø* and that the form *stem+ø* most often realises the function singular and that plural is most frequently realised by *stem+i*, and that *stem+i* most often realises plural.

| | Nom | Acc | Gen | Loc | Instr | Dat | Total |
|---|---|---|---|---|---|---|---|
| **stem+ø** | 24116 | 16755 | 6273 | 172 | 20 | 73 | 47409 |
| **stem+a** | 13796 | 4788 | 18826 | 0 | 0 | 0 | 37410 |
| **stem+i** | 11743 | 7606 | 12564 | 1337 | 0 | 664 | 33914 |
| **stem+u** | 0 | 11396 | 347 | 1114 | 0 | 4944 | 17801 |
| **stem+e** | 93 | 0 | 0 | 13775 | 0 | 2227 | 16095 |
| **stem+o** | 5399 | 4985 | 0 | 0 | 0 | 0 | 10384 |
| **stem+om** | 0 | 0 | 0 | 0 | 9041 | 0 | 9041 |
| **stem+ov** | 0 | 1160 | 6201 | 0 | 0 | 0 | 7361 |
| **stem+ax** | 0 | 0 | 0 | 5683 | 0 | 0 | 5683 |
| **stem+ami** | 0 | 0 | 0 | 0 | 4262 | 0 | 4262 |
| **stem+oj** | 0 | 0 | 0 | 0 | 4229 | 0 | 4229 |
| **stem+am** | 0 | 0 | 0 | 0 | 0 | 2412 | 2412 |
| **stem+ej** | 0 | 219 | 1535 | 0 | 0 | 0 | 1754 |
| **stem+ju** | 0 | 0 | 0 | 0 | 1650 | 0 | 1650 |
| | 55147 | 46909 | 45746 | 22081 | 19202 | 10320 | 199405 |

**Table 14** Alignment of form and case functions

For the case functions (Table 14), we find alignment for all functions except for the accusative and the dative. This is in accordance with our earlier findings using method 1.

## 6 Consistency of results

We checked our results for consistency using a previously unseen part of the Russian Standard Corpus consisting of 500,000 word forms (making up 86490 noun forms and 24916 adjective forms). It is important to note that in this dataset indeclinables were not taken into account. Mapping forms onto functions and functions onto forms, we find similar distributions to the ones obtained for the first part of the corpus. The resulting tables can be found in the appendix.

## 7 Conclusion

In this paper we investigated two linguistic phenomena, syncretism and allomorphy, and their relationship to frequency. We studied this relationship for Russian nominals which exhibit both phenomena.

Previously, it has been noted that higher frequency is associated with greater differentiation (Mańczak 1966). Our data confirm this expectation to a certain extent, a frequent function generally has a high number of forms and a frequent form generally has a high number of functions.

However, the pattern is not exceptionless. In particular, cells in the paradigm which are involved in referral-based syncretism do not fit the pattern. For instance, the accusative has a far greater number of forms than would be expected on the basis of its frequency and forms such as *stem+a* and *stem+ø,* which have some involvement with the animacy rule, show a higher number of functions than expected. These instances do not fit with the general claim about frequency and differentiation. However, there may be a psycholinguistic explanation. If we consider that in order to use the syncretic forms in the Russian nominal system, it is only necessary to remember the rules of referral, then this may be less taxing on memory than learning all the endings as *directly* associated with a function. While accusative itself is still a frequent function, our investigation shows that the claim about frequency and differentiation is not absolutely predictive, and we have suggested a type of rule which can undermine this relationship, namely referrals.

## 8 Acknowledgements

The research reported here is supported by the Economic and Social Research Council (UK) under grant RES-000-23-0082 'Paradigms in Use'. Their support is gratefully acknowledged. We thank Laurie Bauer for comments on a later draft, as well as two anonymous reviewers for their comments.

[1] We define 'forms' abstractly as unique realisations within a paradigm, where a paradigm is an entire set of features or feature combinations (cf. Carstairs-McCarthy 1996). The term 'function' is used loosely in this paper to refer to the value of a morphosyntactic feature or a combination thereof, and is not being used in its strict mathematical sense. We come back to the definition of 'form' and 'function' in Section 4.

[2] We do not consider separately the second locative (Brown forthcoming) and second genitive. They are treated as part of locative and genitive.

[3] There is a variety of ways for referring to grammatical features. An example of a 'property', as used here, is singular, which is a property of the category 'number'. Property is therefore synonymous with 'feature value' in the terminological system where we talk of the feature 'number' and the value of the feature 'singular'.

[4] Forms are given in transcription. We give *i* where the standard Cyrillic orthography has both ы and и, corresponding to the underlying phoneme /i/. (The alternation is conditioned by properties of the preceding consonant.) We give *o*, where the standard orthography has е corresponding to the phoneme /o/ after a palatalised consonant. Palatalisation is represented by the character ´.

[5] We use the following abbreviations: NOM – nominative, ACC – accusative, GEN – genitive, DAT – dative, INSTR – instrumental, LOC – locative.

[6] We do not take stress into account as we obtained similar results with and without stress.

[7] Baayen and Sproat's (1996) investigation of the *–en* suffix in Dutch makes a clear division between form and function, demonstrating that the expected function will be infinitive for hapax legomena (forms which occur exactly once in the corpus), but finite plural for higher frequency verbs.

[8] In this table, indeclinables are included and analysed as *stem* + ∅ for all functions.

## References


Baayen, R. Harald and Richard Sproat (1996). Estimating lexical priors for low-frequency morphologically ambiguous forms. *Computational Linguistics* 22, 155-166.

Baerman, Matthew, Dunstan Brown and Greville G. Corbett (2005). *The Syntax-Morphology Interface: A Study of Syncretism.* Cambridge: Cambridge University Press.

Brown, Dunstan (1998a). From the General to the Exceptional: A Network Morphology Account of Russian Nominal Inflection. PhD thesis, University of Surrey.

Brown, Dunstan (1998b). Defining 'subgender': virile and devirilized nouns in Polish. *Lingua* 104, 187-233.

Brown, Dunstan. forthcoming Peripheral Functions and Overdifferentiation: the Russian Second Locative. To appear in *Russian Linguistics* 31.

Carstairs-McCarthy, Andrew (1996). Paradigmatic Structure: Inflectional Paradigms and Morphological Classes. In Andrew Spencer and Arnold M. Zwicky (Eds.), *Handbook of Morphology* (pp. 322-334), Oxford: Blackwell Publishers.

Corbett, Greville G. and Norman M. Fraser (1993). Network morphology: A DATR account of Russian nominal inflection. *Journal of Linguistics* 29, 113-142.

Corbett, Greville G., Andrew Hippisley, Dunstan Brown and Paul Marriott (2001). Frequency, regularity and the paradigm: a perspective from Russian on a complex relation. In J. Bybee and P. Hopper (Eds.), *Frequency and the Emergence of Linguistic Structure* (pp. 201-226). Amsterdam: John Benjamins.

Evans, Nicholas, Dunstan Brown and Greville G. Corbett (2001). Dalabon pronominal prefixes and the typology of syncretism: a Network Morphology analysis. In G. Booij and J. van Marle (Eds.), *Yearbook of Morphology 2000* (pp. 187-231). Dordrecht: Kluwer.

Evans, Roger and Gerald Gazdar (1996). DATR: A Language for Lexical Knowledge Representation. *Computational Linguistics* 22, 167-216.

Mańczak, Witold (1966). La nature du supplétivisme. *Linguistics* 28, 82-89.

Sharoff, Serge (2006). Methods and tools for development of the Russian Reference Corpus. In A. Wilson, D. Archer, P. Rayson (Eds.), *Corpus Linguistics Around the World* (pp. 167-180). Amsterdam: Rodopi.

Sitchinava, Dmitriy. 2002. K zadače sozdanija korpusov russkogo jazyka, available at: http://www.mccme.ru/ling/mitrius/article.html

Stump, Gregory T. (2001) *Inflectional Morphology*. Cambridge: Cambridge University Press.

Zwicky, Arnold (1985). How to describe inflection. In M. Niepokuj, M. Van Clay, V. Nikiforidou and D. Feder (Eds.), *Proceedings of the eleventh annual meeting of the Berkeley Linguistics Society* (pp. 372-386). Berkeley: Berkeley Linguistics Society.

## Appendix

### Results from consistency-check using method 1.

In order to give an indication of the robustness of our results, we have used method 1 to test against a previously unseen part of the Russian Standard Corpus. We have not included data on the second genitive and second locative singular in the tables. Taking this fact into account, we find similar patterns to the ones observed for the larger dataset.

### Functions to forms

| Function | Frequency | N° of forms without indecl |
|---|---|---|
| **NomSg** | 19742 | 3 |
| **AccSg** | 16378 | 4 |
| **GenSg** | 12923 | 2 |
| **InstrSg** | 7094 | 3 |
| **LocSg** | 6976 | 2 |
| **NomPl** | 5421 | 2 |
| **GenPl** | 5173 | 3 |
| **AccPl** | 4353 | 5 |
| **DatSg** | 3377 | 3 |
| **LocPl** | 2428 | 1 |
| **InstrPl** | 1700 | 1 |
| **DatPl** | 925 | 1 |

### Forms to Functions

| Form | Frequency | N° of functions |
|---|---|---|
| **stem+ø** | 21076 | 4 |
| **stem+a** | 16195 | 5 |
| **stem+i** | 14113 | 5 |
| **stem+u** | 7783 | 2 |
| **stem+e** | 7348 | 2 |
| **stem+o** | 4489 | 2 |
| **stem+om** | 4258 | 1 |
| **stem+ov** | 3101 | 2 |
| **stem+ax** | 2428 | 1 |
| **stem+oj** | 2165 | 1 |
| **stem+ami** | 1700 | 1 |
| **stem+am** | 925 | 1 |
| **stem+ju** | 671 | 1 |
| **stem+ej** | 238 | 2 |

## Alignment of form and function

| | ø | a | i | u | e | o | om | ov | ax | oj | ami | am | ju | ej | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **NomSg** | 11233 | 6205 | 0 | 0 | 0 | 2304 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 19742 |
| **AccSg** | 7459 | 988 | 0 | 5746 | 0 | 2185 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 16378 |
| **GenSg** | 0 | 7851 | 5072 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12923 |
| **InstrSg** | 0 | 0 | 0 | 0 | 0 | 0 | 4258 | 0 | 0 | 2165 | 0 | 0 | 671 | 0 | 7094 |
| **LocSg** | 0 | 0 | 643 | 0 | 6333 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6976 |
| **NomPl** | 0 | 583 | 4838 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5421 |
| **GenPl** | 2235 | 0 | 0 | 0 | 0 | 0 | 0 | 2721 | 0 | 0 | 0 | 0 | 0 | 217 | 5173 |
| **AccPl** | 149 | 568 | 3235 | 0 | 0 | 0 | 0 | 380 | 0 | 0 | 0 | 0 | 0 | 21 | 4353 |
| **DatSg** | 0 | 0 | 325 | 2037 | 1015 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3377 |
| **LocPl** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2428 | 0 | 0 | 0 | 0 | 0 | 2428 |
| **InstrPl** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1700 | 0 | 0 | 0 | 1700 |
| **DatPl** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 925 | 0 | 0 | 925 |