**Number Use in Language:**

**a Quantitative and Typological Investigation**

**Project funded by the ESRC (R000222419)**

**directed by Professor G. G. Corbett, University of Surrey**

**Research Fellow: Dr Andrew Hippisley**

***Dataset deliverable***

**Description document**

**Andrew Hippisley, University of Surrey**

*Abstract*

The researchers Corbett, Hippisley, Brown and Marriott have investigated the relationship between number availability and number use. One of the deliverables promised was a dataset of nouns from the Uppsala corpus encoded for frequency information, case and number features, as well as semantic information, i.e. animacy category. This document contains a description of the dataset column by column, and in some cases a note on methodology. The researchers are grateful for the support of the ESRC (grant no. R00002222419).

**1. Background**

An important contribution to linguistic typology was Smith-Stark's hierarchy of number availability, an extended version of which is given in (1). Nouns with number marking (formally distinguishing singular and plural) typically occupy some top portion. Different languages make the 'split' at different points on the hierarchy (e.g. only Speaker, Addressee, and Kin terms may mark number).

(1)

Speaker > Addressee > Kin > Non-human rational > Human rational >

Human non-rational > Animate > Concrete inanimate > Abstract inanimate

The chief aim of the ESRC project was to investigate to what extent Smith-Stark's hierarchy of number availability impacted on the way number was used. The general methodology was to analyse the way in which the nominals of a one million word Russian corpus distributed their singular and plural forms, and compare that with the nominals' position on the Smith-Stark hierarchy.This task was carried out using the concordance and word list tools in the WordSmith concordance package.

**2. The corpus**

The Uppsala corpus is a set of sub-corpora of various genres, containing in total about 1 million words. It is considered the best Russian corpus available, in terms of scope and design. For information on the Uppsala corpus, see Lönngren (1993) and Maier (1994).

**3. The dataset**

The dataset is in the form of a Microsoft Excel document where case, number (singular and plural), and animacy information about the nouns occurring in the Uppsala corpus are given numerical values, corresponding to case features, animacy features, and frequency. The lexemes recorded in the dataset are those represented by a word form occurring more than five times. The dataset contains around 5440 lexemes, accounting for around 243000 word forms from the entire 1million word corpus.

In oder to present the data as a Microsoft Excel document (diacritics cannot be used), we adopted the following translitertaion system.



**4. Description of the columns in the dataset**

We consider each column in the dataset in turn. The information recorded for some columns involved additional analysis of the corpus (for example affixal homonymy) and in these cases we outline the methodology adopted for retrieving the required information.

*4.1 Columns: Lexeme, Gloss*

The lexemes, the amalgamation of all word-from of a noun lexeme, are arranged in frequency order. They appear with a gloss.

*4.2 Columns: Animacy*

Each lexeme has been recorded for animacy category, based on our extended version of the Smith-Stark hierarchy as shown in (1) above. Animacy has been recorded numerically, and the correspondence between animacy and numerical index is given in Table 1.

|  |  |
| --- | --- |
| Animacy | numeric index |
| Kin | 3 |
| Non-human rational | 4 |
| Human rational | 5 |
| Human non-rational | 6 |
| Animate | 7 |
| Concrete inanimate | 8 |
| Abstract inanimate | 9 |

TABLE 1: Animacy category and its numeric index in the dataset

*4.3 Columns: ‘Frequency’, ‘Sg’, ‘Pl’, ‘Pl/freq’*

Frequency information for each lexeme is recorded in the last four columns. This information is broken down into overall frequency (‘Frequency’), all singular occurrences (‘Sg’), all plural occurrences (‘Pl’), and the proportion of all the occurrences of the lexeme which are plural (‘Pl/freq’).

*4.4 Columns: ‘NomSg’, ‘GenSg’, ‘DatSg’, ‘InstSg’, ‘LocSg’, ‘NomPl’, ‘GenPl’, ‘DatPl’, ‘InstPl’, ‘LocPl’*

In Russian there are two number values (singular and plural) and six cases. These are: nominative, accusative, genitive, dative, instrumental, and locative. As a typical member of Indo-European Russian is a fusional type language where a single ending fuses case and number information. The columns above correspond to the case and number combinations.

*Methodological implication*

The case/number endings fall into a number of paradigms. The main noun classes in Russian are given in Table 2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|   | I | II | III | IV |
|   | stol ‘table’ | karta ‘map’ | kost´ ‘bone’ | okno ‘window |
| SG |   |   |   |   |
| nom | stol | kart-a | kost´ | okn-o |
| acc | stol | kart-u | kost´ | okn-o |
| gen | stol-a | kart-y | kost-i | okn-a |
| dat | stol-u | kart-e | kost-i | okn-u |
| inst | stol-om | kart-oj | kost´-ju | okn-om |
| loc | stol-e | kart-e | kost-i | okn-e |
| PL |   |   |   |   |
| nom | stol-y | kart-y | kost-i | okn-a |
| acc | stol-y | kart-y | kost-i | okn-a |
| gen | stol-ov | kart | kost-ej | okon |
| dat | stol-am | kart-am | kost-jam | okn-am |
| inst | stol-ami | kart-ami | kost-jami | okn-ami |
| loc | stol-ax | kart-ax | kost-jax | okn-ax |

TABLE 2: Russian noun classes

From Table 2 we see that there are four main groups, represented here by *stol*, *karta*, *kost´*, and *okno*. In classes I and IV each case/number combination is marked by a separate form, except for the direct cases. In the other classes their endings tend to merge. For example, in class II the suffix *-y* marks genitive singular and the direct cases in the plural. In class III the merging of endings is widespread. This makes the analysis of the nouns in the corpus a more complex task. Each word form occurrence which does not mark case and number had to be disambiguated by carefully examining the context in which the word form appears.

In addition to homonymy within the lexeme, there is homonymy across word forms, and further analysis had to be done to disambiguate examples of this kind. One actual example from the coprus is the word form *vek-i*. This can either be the nominative plural of *vek-o* ‘eyelid’, or an archaic nominative plural of *vek* ‘century’, and has been disambiguated accordingly.

*4.5 Columns ‘Gen2’ and ‘Loc2’*

In addition to the six cases shown in Table 2, Russian has two sub-cases, the second genitive (a sub-case of the genitive), and a second locative (a sub-case of the locative). The sub-cases occur in the singular paradigm of class I. An example of a noun with a second genitive and a second locative is *glaz* ‘eye’ and its singular paradigm is given in Table 3. Columns ‘Gen2’ and ‘Loc2’ record occurrences of the second genitive and second locative respectfully.

*Methodological implication*

As can be seen from Table 3, the sub-case endings are both in *-u* which in class I is homonymous with the dative singular (Table 2). For nouns known to contain a sub-case we have disambiguated word forms in *-u* by checking the contexts in which they occur. Zaliznjak (1977), a morphological dictionary, has been used as a guide to which nouns contain sub-cases.

|  |  |
| --- | --- |
|   | I |
|   | glaz ‘eye’ |
| SG |   |
| nom | glaz |
| acc | glaz |
| gen | glaz-a |
| gen 2 | glazu |
| dat | glaz-u |
| inst | glaz-om |
| loc | glaz-e |
| loc 2 | glaz-ú |

TABLE 3: sub-cases

*4.6 Columns: ‘AccSg’, ‘Ref to NomSg’, ‘Ref to Nom1Pl’*

Generally the accusative singular is syncretic with the nominative singular. The column ‘AccSg’ records occurrences of the accusative singular of a lexeme where the accusative singular is morphologically distinct from the nominative singular. Morphological distinction for the accusative singular is restricted to class II (inanimate and animate) and class I animate nouns only. Where the accusative singular is not morphologically disambiguated from the nominative singular, the following procedure has been adopted: (i) the ‘AccSg’ column is recorded with a zero value; (ii) the ‘Ref to NomSg’ is given the value 1. In other words, for lexemes with a 1 in the ‘Ref to NomSg’ column, the value in the ‘NomSg’ column corresponds to the occurrences of nominative singular and accusative singular together. If a zero appears in the ‘Ref to NomSg’ column, then any value in the ‘NomSg’ is a record solely of nominative singular occurrences. Similarly in the plural, where the referring column is ‘Ref to Nom1Pl’. All animates have accusative plurals distinct from nominative in the plural, and where this is the case a zero has been recorded in the ‘Ref to Nom1Pl’ column’.

*Methodological implication*

In the singular, in class I accusative case is syncretic with the genitive for animate nouns. All class I animate nouns have been disambiguated for accusative and genitive singular by carefully examining the contexts in which they appear. In the plural, all animates have accuasative / genitive syncretism, and have been disambiguated.

*4.7 Column: ‘InstSg2’*

Some class II nouns have instrumental singular word forms in the optional *-oju* ending, usually in addition to the general -*oj* ending. We have recorded these occurrences separately. Occurrences in the general ending are recorded in the ‘InstSg’ column, and occurrences in the optional ending are recorded in a separate column ‘InstSg2’.

*4.8 Column: ‘Vocative’*

So-called vocatives are restricted to class II animate nouns, and formed by shortening the citation nominative singular to the bare stem. Very few occurrences were found, and they are recorded in the ‘Vocative’ column.

*4.9 Column: ‘Nom2Pl’, ‘Acc2Pl’, ‘Inst2Pl’*

In the plural, some nouns have alternative nominative, accusative and instrumental forms, and occurrences of these have been recorded in the columns ‘Nom2Pl’, ‘Acc2Pl’, ‘Inst2Pl’. The general word forms and the alternative word forms for these nouns are given in the look-up tables below (Tables 4 to 6). Zaliznjak has been used as guide as to what counts as the alternative form. Note that not all cases have been recorded in Zaliznjak. Note also that the general form does not always correspond to the most frequent form. In some cases, an alternative gloss is associated with the alternative word form, and this is also given.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Lexeme | Gloss | General form(‘NomPl’) | Alternative form(‘Nom2Pl’) | Alternative gloss  |
| God | year | gody | goda | - |
| Chelovek | person | ljudi | cheloveki | - |
| Vek | century | veka | veki | (used in expressions) |
| Direktor | director | direktora | direktory | (not in Zaliznjak) |
| Cvet | flower / colour | cvety | cveta | (when ‘colour’) |
| Zub | tooth | zuby | zubja | cog (in machine) |
| Traktor | tractor | traktora | traktory | - |
| Shtorm | gale | shtormy | shtorma | - |
| Zarja | dawn | zari | zori | - |
| Jastreb | hawk | jastreba | jastreby | - |
| Shchenok | puppy | shchenki | shchenjata | - |
| Shtabel´ | stack | shtabelja | shtabeli | - |

TABLE 4: Nominative plural alternatives

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Lexeme | Gloss | General form(‘AccPl’) | Alternative form(‘Acc2Pl’) | Alternative gloss |
| Ptichka | bird / tick | ptichek | ptichki | tick (only) |
| Shchenok | puppy | shchenkov | shchenjat | puppy |

TABLE 5: Accusative plural alternatives

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Lexeme | Gloss | General form(‘GenPl’) | Alternative form(‘Gen2Pl’) | Alternative gloss |
| God | year | let | godov | - |
| Chelovek | person | ljudej | chelovek | (used with numerals) |
| Kurica | hen | kur | kuric | - |
| Korol´ | king | korolej | korolev | (not in Zaliznjak) |
| Prostynja | sheet | prostynej | prostyn´ | - |

TABLE 6: alternative genitive plurals

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Lexeme | Gloss | General form(‘InstPl’) | Alternative form(‘Inst2Pl’) | Alternative gloss |
| Sleza | tear | slezami | slez´mi | - |
| Kost´ | bone | kostjami | kost´mi | - |

TABLE 7: alternative instrumental plurals

**References**

Lönngren, Lennart 1993. *Chastotnyj slovar´ sovremennogo russkogo jazyka*. (=Acta Universitatis Upsaliensis, Studia Slavica Usaliensis 33). Uppsala.

Maier, Ingrid 1994. Review of Lennart Lönngren (ed*.)* ‘Chastotnyj slovar' sovremennogo russkogo jazyka’*. Rusistika Segodnja* 1. 130-6.

Zaliznjak, A. A., 1977.*Grammaticheskij slovar´ russkogo jazyka*. Moscow: Russkij jazyk.