

TYPOLOGICAL STUDIES IN LANGUAGE 45

**FREQUENCY
AND THE EMERGENCE
OF LINGUISTIC
STRUCTURE**

Edited by

**JOAN BYBEE
PAUL HOPPER**

OFFPRINT

JOHN BENJAMINS PUBLISHING COMPANY

This is an offprint from:

Joan Bybee and Paul Hopper (eds)
Frequency and the Emergence of Linguistic Structure.

John Benjamins Publishing Company
Amsterdam/Philadelphia

2001

(Published as Vol. 45 of the series
TYPOLOGICAL STUDIES IN LANGUAGE, ISSN 0167-7373)

ISBN 90 272 2947 3 (Eur.) / 1 58811 027 3 (U.S.) (Hb.)

ISBN 90 272 2948 1 (Eur.) / 1 58811 028 1 (U.S.) (Pb.)

© 2001 – John Benjamins B.V.

No part of this book may be reproduced in any form,
byprint, photoprint, microfilm or any other means, without written
permission from the publisher.

Frequency, regularity and the paradigm: A perspective from Russian on a complex relation*

GREVILLE CORBETT, ANDREW HIPPISEY,
DUNSTAN BROWN, and PAUL MARRIOTT

University of Surrey and National University of Singapore

1. Introduction

The correspondence between irregularity and high frequency is well known (Bybee 1995, Greenberg 1966). What is not always clear is whether the frequency envisaged is based on the lexeme and all its manifestations, including the irregular word form(s), or just the irregular form(s) alone.¹ For example in the case of English *went*, is it this single word-form that is highly frequent compared to other past tense forms, or is it the lexeme *go* that is highly frequent, in both its regular and irregular manifestations, compared to other lexemes? Bybee (1985: 120) suggests the following:

the correlation of irregularity with frequency occurs on two dimensions. The first is the lexical dimension . . . where irregularity correlates with frequent lexical entries. The second is within the paradigm.

To investigate further we have examined frequencies of noun lexemes, and their word forms, in a one million word Russian corpus (the Uppsala corpus), together with information on regularity. So as to more finely locate any correspondence between frequency and regularity, the types of irregularity we considered range from full suppletion to minor irregularity in stress.

2. Claims, hypotheses, and statistical method

In Section 2, we briefly discuss the various claims made about the frequency and irregularity relationship, and outline the hypotheses we test to explore this relation-

ship. At the end of the section we give an overview of the statistical method adopted (see also Appendix 1).

2.1 Claims

The most general claim is that there is a relationship between high frequency and irregularity. This is a claim with which almost any linguist would agree. However, the nature of the relationship is so vague as to be untestable. Once we begin to clarify the claim, we find an interesting range of possible relationships. Our strategy will be to suggest that each may be true, and then look for ways to prove or disprove them. The initial claim that we investigated was that there might be a straightforward linear correlation between regularity and frequency; however the data suggested that in fact it was more appropriate to search for a more complex relationship.

Let us start with irregularity and consider its *extent*, i.e., the distribution of irregular forms within a paradigm. Within a given lexeme it might be that every form could be irregular independently; or else it might be that forms come in groups which are regular or irregular together. We have looked at Russian, specifically at nouns. This word class has two numbers and six cases (presented later in Table 1). We can ask whether irregularity concerns a high level split between singular and plural or whether we should consider individual forms. Of course, we shall try both approaches (this is taken further in Section 2.2).

A second question concerns the *degree* of irregularity. Russian *č'elovek ~ l'ud'-i* 'person ~ people' form an irregular relation, but so do *mést-o ~ mest-á* 'place ~ places'.² In the first example we have different stems (suppletion) and in the second we have the stress unexpectedly on the ending in the plural (where in the singular it is on the stem). Intuitively, the first type of irregularity is more severe than the second. If we believe there is a relationship between frequency and irregularity, then we might claim that it will be sensitive to *degrees* of irregularity. To test this claim we set up a ranking of irregularity, devised of course without reference to frequency (see Section 4).

Turning now to notions of frequency, as hinted at already this can be viewed in two ways. Suppose we have a noun whose plural is irregular. With what precisely do we expect to find a relationship? It is easiest to see the alternatives if we consider a corpus and look at the tests we might apply. We might compare lexemes, one to another, or we could compare regular and irregular forms within lexemes. For the first approach, we could count up how many times each lexeme occurs in the plural. Since we are counting only plurals (without respect to other forms, i.e., the singular) we call this the *absolute frequency* of a lexeme's plural. We can then compare the *absolute frequency* of plural of different lexemes to see if there is a relationship between irregular plurals and their *absolute frequency*. There is, however, a quite

different way to look at the plural (and indeed at any cell or combination of cells in a paradigm), that is we may compare it, within the lexeme, with the other available forms. For a given lexeme, we could count how many times it occurs in the plural as compared to the number of times it occurs in the singular. This is the *relative frequency* of the plural. We can then compare the *relative frequency* of the plural in lexemes where it is irregular with the *relative frequency* in lexemes where it is regular. We consider this question further in the next section.

Since the distinction between *absolute* and *relative frequency* is important, consider a tiny corpus consisting of four lexemes, as in Figure 1.

	Lexeme A	Lexeme B	Lexeme C	Lexeme D
Singular occurrences	10	20	30	40
Plural occurrences	5	5	10	10
Absolute plural frequency	5	5	10	10
Relative plural frequency	0.33	0.2	0.25	0.2

Figure 1. *Absolute and relative frequency*

Lexemes C and D occur in the plural 10 times each. Their *absolute plural frequency* is 10, higher than that of the other two lexemes. But when we turn to *relative plural frequency*, we note that lexeme A has 5 plural occurrences out of a total of 15. Its *relative plural frequency* is therefore 0.33 which is higher than that of any of the other lexemes.

2.2 Terms and hypotheses

We now set out a number of hypotheses to test the relationship between irregularity and frequency. The hypotheses are formulated in such a way that their confirmation or disconfirmation will not only determine whether there is a relationship between irregularity and frequency, but will also answer more specific questions as to what the nature of the relationship is. To determine whether there is a relationship between regularity and frequency we will look for a particular kind of *anomaly* in the corpus. Before looking at the hypotheses, we introduce the terms *plural anomaly* and *cell anomaly*.

2.2.1 Plural anomaly and cell anomaly

The focus of the investigation is specifically on any *anomaly* in the behavior of the plurals in the corpus. Before stating the hypotheses, we need to be clear what we mean by *plural anomaly*. The definition is given in (1):³

(1) *Plural anomaly*

Plural *anomaly* can be in terms of absolute or relative frequency

a. *Absolute plural anomaly*

This is an absolute anomalous frequency of plurals for a given lexeme

b. *Relative plural anomaly*

This is a relative anomalous frequency of plurals for a given lexeme (the proportion of a lexeme's plurals is anomalous)

What we are saying in (1) is that the *anomaly* in the plurals of the corpus can be viewed in two distinct ways. The first is in terms of an anomalous count of plurals for a lexeme compared to the amount one would expect for a typical lexeme of the corpus (*absolute plural anomaly*). In other words, if a lexeme's count of plural word forms is extreme compared to the distribution of counts of plurals, we would have identified an *absolute plural anomaly*. This is an *absolute anomaly* because what is being compared is an absolute number of plurals for a lexeme with the distribution of the absolute number of plurals in the corpus.

The second way of thinking about the *anomaly* is in relative terms. Here the proportion of instances of the lexeme that are plural is examined for an *anomaly*. The distribution of plural proportions can be calculated for the lexemes of the corpus, and if the given lexeme's proportion of plurals is extreme compared to this distribution, we would have identified a *relative plural anomaly*. (This may be seen as a generalisation of Tiersma's (1982) notion of local markedness.)

So far we have thought of *plural anomaly* generally in terms of the plural half of the lexeme's paradigm. We also wish to allow for the possibility of the *anomaly* being due, as it were, to one of the case and number cells. For this we need the idea of another specific kind of *anomaly*, which we will term *cell anomaly*, as defined in (2):

(2) *Cell anomaly*

One specific cell has an extreme proportion compared to the distribution of the proportion of that cell throughout the corpus. This can only be stated in relative terms.

In *cell anomaly* the *anomaly* is that a given lexeme has a significantly higher (or lower) than average proportion of word forms for a given cell.⁴ For example, the lexeme may have a much higher than average proportion of genitive plurals compared to the corpus in general. Note that it is important to define *cell anomaly* in relative terms only, because formulating it in absolute terms might mean that we would be observing *plural* (or *singular*) *anomaly* in disguise. In other words, the cell may be above or below the average simply as a consequence of the singular or plural subparadigm being above or below the average.

2.2.2 Hypotheses to be tested

The relationship between regularity and frequency will therefore be seen in terms of *plural* or *cell anomaly*, as just discussed. We now list four hypotheses which we will test. The four hypotheses are discussed in turn.

(3) Hypothesis 1a

There is a relation between *absolute plural anomaly* and irregularity.

If Hypothesis 1a is confirmed, we will have shown that there is a relation between irregularity and frequency and the data analysis will tell us the nature of this relationship.

Note that if we observed *absolute plural anomaly* in certain groups of lexemes, this might still be because the lexeme as a whole was anomalously frequent. We need a test which tells us whether the frequency relationship is with the general lexeme, or whether it is specifically with the lexeme's irregular forms. Recall our original question in the introduction: is frequency related to the lexeme as a whole or to its irregular word forms? We address this question using Hypothesis 1b:

(4) Hypothesis 1b

There is a relation between *relative plural anomaly* and irregularity.

We also need to test whether there is a stronger relationship with irregularity when we combine *plural anomaly* (either absolute or relative, see (1), with the more specific *cell anomaly* (2). In other words, if a lexeme's plural forms occurred more frequently than average and a particular cell in the plural was proportionally more frequent than average, are we right in expecting the form in question to be even more irregular? We address this question using Hypothesis 2, which allows us to look for a stronger (and more fine-grained) relationship between high frequency and irregularity:

(5) Hypothesis 2

If Hypothesis 1a or Hypothesis 1b is true, there is a stronger relationship between irregularity and the combination of *plural anomaly* and *cell anomaly*.

A particular case and number may occur more frequently than average either due to the lexeme occurring frequently or to the fact that the cell occurs unusually out of proportion to all word forms in the corpus (*absolute frequency* of the cell).

Note that if Hypotheses 1a and 1b were disconfirmed, we would need to find out whether there might be any relationship at all between irregularity and frequency. We would do this by looking at the level of individual case and number cells.

(6) *Hypothesis 3*

There is no relation between irregularity of a plural cell and the *absolute frequency* of that cell.

Hypothesis 3 is independent of Hypotheses 1 and 2. It allows us to find the answer to whether or not irregularity of a single cell by itself has a relationship with *absolute frequency*. Hypothesis 3 is there for completeness. As we shall see, Hypothesis 3 proved to be unnecessary and will play only a minor part in the following discussion.

2.3 *Statistical method*

Using the data extracted from the corpus (see Section 3), we investigated the relationship between irregularity and frequency. This frequency could be in absolute or relative terms.

We extracted subsets of lexemes from the corpus according to the regularity of the lexemes. For all lexemes an appropriate *absolute* or *relative frequency* is calculated. If there were no effect between regularity and frequency then we would expect no statistically significant difference in the measured frequency distributions in the subset and in the full corpus. In order to compare these distributions a simple summary statistic—the median—was chosen. Hence all tests are based on finding statistically significant differences between the median frequency in the subset and in the full corpus. Informal exploratory data analysis was done to investigate the claims. We used box-plots to compare the distributions of frequencies across groups (Daley *et al.* 1995). See for example Figure 2 in Section 5.2 where a box plot is used.

Having formulated the hypotheses in terms of significant differences in median values, it is necessary to use an appropriate statistical test. We decided to use a non-parametric technique, in which we are assuming that the frequency of lexeme use in the corpus is a good representation of their use in the general language. The quantity and quality of the data is sufficiently high that any loss of efficiency in using non-parametric techniques is felt to be unimportant. This small loss is more than compensated for by the simplicity and directness of the non-parametric tests used. For details of the testing procedure see Appendix 1.

3. The data

We tested the hypotheses on the Russian nouns in a corpus. Russian is a good choice for this type of investigation, because noun paradigms have sufficient cells

for us to tease apart the irregularity of the lexeme in its entirety and that of one of its word forms. Also, irregularity in Russian is highly varied, ranging from full suppletion to shift in stress.

3.1 Russian nominal inflection

Russian is an East Slavonic language, part of a branch of Indo-European which has been relatively conservative in terms of inflectional morphology. Nouns distinguish number and case, and fall into different inflectional classes. These inflectional classes share some forms between them, so that it is not self-evident how many inflectional classes should be recognized. The traditional answer is three, but other views are possible, as discussed in Corbett (1982), where he argues for four basic inflectional classes. Our analysis is based on these four classes, as shown in Table 1. There are several partially overlapping reasons for recognising these four as major classes. Each is productive, though the productivity of classes III and IV is dependent on a small number of derivational affixes. Each has a significant number of members (at least several thousand, though again there is some disparity). Each is

Table 1. *Major noun classes of Russian**

		I	II	III	IV
		zakón 'law'	gazéta 'newspaper'	rúkop'is' 'manuscript'	bolóto 'swamp'
Singular	NOM	zakón	gazéta	rúkop'is'	bolóto
	ACC	zakón	gazétu	rúkop'is'	bolóto
	GEN	zakóna	gazéti	rúkop'is'i	bolóta
	DAT	zakónu	gazéte	rúkop'is'i	bolótu
	INST	zakónom	gazétoj	rúkop'is'ju	bolótom
	LOC	zakóne	gazéte	rúkop'is'i	bolóte
Plural	NOM	zakóni	gazéti	rúkop'is'i	bolóta
	ACC	zakóni	gazéti	rúkop'is'i	bolóta
	GEN	zakónov	gazét	rúkop'is'ej	bolót
	DAT	zakónam	gazétam	rúkop'is'am	bolótam
	INST	zakónam'i	gazétam'i	rúkop'is'am'i	bolótam'i
	LOC	zakónax	gazétax	rúkop'is'ax	bolótax

*We use the following abbreviations: NOM—nominative, ACC—accusative, GEN—genitive, DAT—dative, INST—instrumental, LOC—locative.

Notes: (i) forms are given here in phonemic transcription (see note 2). Palatalization (or 'softening') is indicated by 'i'; (ii) there is no overt ending in the nominative/accusative singular in types I and III, nor in the genitive plural of types II and IV.

regular, in that there are mutual predictabilities between certain cells of the paradigm. And, provided these four classes are distinguished, the gender of almost every Russian noun is predictable from information available in the lexicon (Corbett 1982). Thus positing these four classes depends partly on their type frequency, in other words the number of different lexical items, or 'types', found in a dictionary that they apply to (Bybee and Thompson 1997). However, our main interest is the relationship between regularity and token frequency, the number of actual occurrences of a lexical item in running text.

A useful overview of the data can be found in Timberlake (1993: 836–45), and Network Morphology treatments are available in Corbett and Fraser (1993), Brown and Hippisley (1994) and Fraser and Corbett (1995).

3.2 *Irregularity in Russian nouns*

Russian nouns are ideal for our investigation because they provide numerous types of inflectional irregularity, from the most radical to the very minor. Starting with the most radical cases, we find instances of full suppletion—nouns whose singular and plural stems are quite different. Then there are those cases whose stems are clearly related, but they differ in ways which may be unpredictable or only partly predictable. Stem augments may be found in the singular, the plural, or in both. Then, on the other extreme, we find minor inflectional irregularities, such as the use of forms which typically belong with nouns of another class. And finally we find nouns which would be fully regular if we looked only at segmental phonology, but which are prosodically irregular. There are four main types of stress pattern and then there are minor irregular patterns in addition (Brown *et al.* 1996).

3.3 *The corpus and the dataset*

We use the Uppsala corpus, which is a set of Russian sub-corpora of various genres, containing in total about one million words. It is considered the best Russian corpus available, in terms of scope and design. For information on the Uppsala corpus, see Lönnngren (1993) and Maier (1994). The dataset which we created is in the form of a Microsoft Excel document where, in addition to regularity information, case, number (singular and plural), and animacy information about the nouns occurring in the Uppsala corpus are given numerical values, corresponding to irregularity indexes, case features, animacy features, and frequency information.

Since we were interested in estimating proportions in different categories, there would be large standard errors in our estimates where observed numbers in each category are small. Large sampling errors would complicate detailed cluster analysis. For this reason we recorded only those lexemes which occur at least five times.

Given this, the dataset contains around 5440 lexemes, accounting for around 243,000 word forms from the entire one million word corpus.⁵

4. Ranking irregularity

As we have said, our aim is to investigate the relationship between irregularity and frequency; we specifically wish to tease apart the irregularity of a lexeme and that of one of its inflectional forms. It is important to be clear what we mean by irregularity, and what we view as the paradigm of the lexeme. This is clarified in 4.1. In 4.2 we outline a number of principles on which an irregularity ranking is based and in 4.3 we carefully show how lexemes are assigned their rank. Further examples are given in Appendix 2. Finally in 4.4, we look at irregularity as treated in the Natural Morphology theory, as a point of comparison.

4.1 Definitions and assumptions

We briefly state what we mean by regularity, and outline our assumptions about the paradigm of Russian nouns.

4.1.1 Regularity and irregularity

We start by giving a notion of regularity for an inflectional language.

Regularity. We expect a regular noun to have:

- (i) a single (unchanging) stem
- (ii) a fixed stress (whether fixed with respect to the stem or with respect to the word-edge)
- (iii) a consistent set of endings (i.e., a set of endings which predict each other)⁶

Irregularity. We treat each irregular type as a numerical step away from regularity. Suppletion is the most severe type of irregularity. However, even this does not define an end point, since a noun with suppletive stems *and* irregular inflections is more irregular than a noun with suppletive stems but regular inflections. The question is how much structural difference there is between a given irregular noun and the prototypical regular noun, for example *gazéta* 'newspaper'. How much, and how drastic, is the change required to bring an item to a state of regularity? An irregularity type is therefore viewed in terms of distance from the regular type, and the distance itself is viewed in terms of the nature of the adjustment required to 'restore' the item.

'*Structural irregularity*'. We are investigating 'structural irregularity', i.e., irregularity determined by comparing forms according to a set of principles. Since we wish to investigate the relationship with frequency, we must exclude any frequency consideration when determining regularity. Thus a noun with a small deviation from the regular pattern counts as almost regular, even if it is the only noun to behave in that way.

4.1.2 *Assumptions about the paradigm and irregularity*

We treat all cells of the paradigm as equal (though it might be argued that, say, an irregularity in the nominative singular should be treated as more important than a similar irregularity in another cell). More difficult is the number of cells to recognize.

Assumptions about cells and irregularity

We start with a distributional criterion, that is, we determine how many distinctions are justified by the syntax (Comrie 1986, 1991). We accept the traditional view of six cases and two numbers, hence twelve cells in all.⁷ However, if we were simply to assume that each paradigm has twelve cells, this would lead to a counter-intuitive result. The problem is that there are certain cells whose forms must be identical within one or another inflectional class. Consider the contrasting paradigms of *zakón* 'law' and *dom* 'house' in Table 2.

Table 2. *Contrasting paradigms in Russian*

	Singular	Plural		Singular	Plural
NOM	zakón	zakón-i	NOM	dom	dom-á
ACC	zakón	zakón-i	ACC	dom	dom-á
GEN	zakón-a	zakón-ov	GEN	dóm-a	dom-óv
DAT	zakón-u	zakón-am	DAT	dóm-u	dom-ám
INST	zakón-om	zakón-am'í	INST	dóm-om	dom-ám'í
LOC	zakón-e	zakón-ax	LOC	dóm-e	dom-áx

The relevant point is that the accusative plural cannot be a distinct form, in these or any other paradigms; it must be the same as the nominative plural, as here, or as the genitive plural, for animates.⁸ We treat *zakón* 'law' as the regular noun (see Table 1, and the points made in Section 3.1). The noun *dom* 'house' has the irregular form *dom-a* 'houses': the inflection is not predictable given the other forms in its paradigm, and forms in another of the classes established in Section 3.1 are wrongly predicted given the ending. If we count twelve cells, then *dom* 'house' is irregular in two cells. However, if the accusative plural were regular *dom-i* while the nominative plural were irregular *dom-a*, that would actually be much more irreg-

ular. It would be breaking a fundamental pattern which extends to every noun (and adjective and pronoun) in the language, according to which the accusative plural is identical to the nominative or genitive. It therefore appears more logical to treat the nominative and accusative as one cell here. We do this because there is a 'whole word' referral, that is to say, the forms must be absolutely identical, including in respect to stress (see Brown *et al.* 1996). In a similar way, we do not count the accusative singular for nouns of classes I, III, and IV and we treat the dative and locative singular of class II as syncretic. The result is that there is a maximum of ten distinct cells for any given noun.

Assumptions about number and irregularity. There are several instances in which singular and plural are contrasted in Russian. Consider the paradigm in Table 3, from a different inflectional class, class IV (see Table 1), in terms of its stress.

Table 3. *Paradigm with singular/plural split*

	Singular	Plural
NOM	mést-o	mest-á
ACC	mést-o	mest-á
GEN	mést-a	mest
DAT	mést-u	mest-ám
INST	mést-om	mest-ám'i
LOC	mést-e	mest-áx

A perfectly regular paradigm would have the same stress position throughout. In this case, however, we have fixed stem stress in the singular and a different fixed stress in the plural (on the ending). Here it is counter-intuitive to count up cells; the point is that there is a single difference between singular and plural.⁹ We treat this as less irregular than a single cell being 'out of line'. We shall see this same pattern in various types of irregularity at different points below.

4.2 Principles

There are six principles on which the irregularity ranking is based and these will be treated in turn. In the examples, the nominative singular is contrasted with the nominative plural unless otherwise stated, with the gloss given for the singular form only.

Principle 1

Stem irregularity ranks above inflectional irregularity:

stem irregularity > inflectional irregularity

Example: *sosed* ~ *sosed'-i* 'neighbor' > *pleč'-o* ~ *pleč'-i* 'shoulder'

The first example *sosed* displays stem irregularity in that the stem final consonant /d/ alternates with /d'/ in the plural. The second example *pleč'-o* displays inflectional irregularity: the item switches from class IV in the singular to class I in the plural, as seen in the nominative (refer to Table 1). One motivation for this ranking is that inflectional irregularity may be treated as an abstract (featural) difference in stems. John McCarthy (personal communication) suggests other evidence for the primacy of stems: in vowel harmony systems, either the stem may determine the vowel possibilities of the ending or stem and ending may determine each other, but it is never the case that the ending alone determines the properties of the stem.

Principle 2

Segmental irregularity ranks above prosodic irregularity:

segmental irregularity > stress irregularity

Example: *sosed* ~ *sosed'-i* 'neighbor' > *óz'or-o* ~ *oz'ór-a* 'lake'

In the first example the stem final consonant /d/ alternates with palatalized /d'/, a segmental irregularity. In the second example, the alternation concerns stress, from initial to predestinential syllable, a prosodic irregularity.¹⁰ The justification for this principle is that typically there are greater phonological differences available through segmental means than through prosodic means. Note that though stress has a great effect on vowel quality in Russian, such that stress affects the entire word, this effect is 'automatic'.

Principle 3

Within stem irregularity, specifically of the segmental kind, suppletion ranks above irregularity involving augments and augment irregularity, in turn, ranks above simple alternations in the stem:

suppletion > augments > stem alternations

Example: *č'elovek* ~ *l'ud'-i* 'person' > *tatar'in* ~ *tatar-i* 'Tatar' > *sosed* ~ *sosed'-i* 'neighbor'

Note that *tatar'in* represents a noun with an augment in the singular, namely *-in*, but no augment in the plural. This ranking is based on the degree of similarity among the alternates; augments are closer to full suppletion than alternations, which can be accounted for by rules of allomorphy. In terms of stems, this reflects the difference between indexed stems ('morphologically' distinct, see Aronoff 1994), and stem alternants (morpho-phonologically distinct).

Principle 4

Within segmental stem alternations we distinguish motivated alternations (i.e., mobile vowels) and non-motivated alternations. Unmotivated alternations rank higher than motivated:

unmotivated stem alternation > motivated stem alternation

Example: *sosed* ~ *sosed'-i* 'neighbor' > *kn'ižk-a* 'book (diminutive)' ~ *kn'ižek* (genitive plural)

Note that in the second example, a mobile vowel appears in the genitive plural, but given the structure of the lexeme this is where it is expected to occur, and is therefore 'motivated'. This principle is based on phonology; motivated alternates are those which are in accord with a phonological principle of the language (sonority in the case of mobile vowels).

Principle 5

Within unmotivated segmental stem alternations we distinguish two broad classes of alternations, those affecting the segment of the stem adjacent to the inflection and others:

non-adjacent segment alternation > adjacent segment alternation

Example: *č'ort* ~ *č'ert'-i* 'devil' > *sosed* ~ *sosed'-i* 'neighbor'

In the first example, the alternation concerns the vowels /o/ and /e/, appearing within the form, and not at the edge. Note that this example also has alternation of the stem final segment /t/ and /t'/. This ranking is expressed by treating examples such as *č'ort*, which show ablaut, along with stems which have an augment *x* in the singular and an augment *y* in the plural (see Principle 6). The justification for Principle 5 is that the adjacent stem segment is more easily associated with the inflection than is the non-adjacent stem segment.

Principle 6

Finally, we impose a ranking on the various kinds of augmentation. The several possible outcomes outlined in Section 3.2 are ranked with respect to one another as follows:

augment *x* opposing augment *y* >

augment in singular opposing lack of an augment in plural >

augment in plural opposing lack of augment in singular

Example:

koť'onok ~ *koť'at-a* 'kitten' > *tatar'in* ~ *tatar-i* 'Tatar' > *brat* ~ *brat'j-a* 'brother'

Principle 6 could be interpreted in terms of classical notions of markedness (Jakobson 1932). The principles outlined above yield the ranking in (7).¹¹

(7) Irregularity ranking¹²

- suppletion irregularity >
- plurality tantum irregularity >
- stem augments irregularity >
- segmental stem irregularity >
- stress stem irregularity >
- segmental inflectional irregularity >
- stress inflectional irregularity >
- full regularity

4.3 *Natural Morphology and the ranking of irregularity*

It should be mentioned that a number of the principles we propose bear some resemblance to those developed independently (and for an entirely different purpose) within the Natural Morphology theory. Indeed, the notion of a scale to express the nature of word structure can be found in Dressler's (1985: 59) scale of 'phonological naturalness'. Here (morpho)phonological rules are distributed on a scale depending on how closely they match a universal set of phonological processes.¹³ Moreover there are a number of naturalness principles which have an affinity with our structural principles of irregularity. In other words, to some extent Natural Morphology views structural distance from the norm in terms of naturalness. For us, greater structural distance corresponds to greater irregularity; for Natural Morphology it corresponds to greater unnaturalness. Perhaps the most important principle is that of Morphotactic Transparency, i.e., the less one disturbs the perceptual segmentation of stem and ending, the more transparent the item is (Dressler 1985: 316; 1987: 102–10). The hierarchy is given in (8) with least transparent first.

(8) Morphotactic Transparency

- total suppletion >
- partial suppletion >
- modification by MPRs >
- modification by MPRs (morph. boundary intact) >
- modification by PRs (allophonic) >
- no modification

Another principle found in the Natural Morphology literature is that of System Congruity (Wurzel 1987: 65–6; 1989) which states that it is more natural for a given item to follow generalizations in the morphological system than not to do so. Once inflectional classes have been established, the expectation is that nouns will not deviate from the pattern. Closely connected to this principle is the idea of 'implicative paradigm structure conditions' (Wurzel 1987: 76–7). The example

Wurzel gives is that if in Russian a noun in the nominative singular ends in /a/, its genitive singular will end in /i/. Finally, the Principle of Constructional Iconicity states that 'what is more semantically ought to be constructionally more as well' (Mayerthaler 1987: 25–8). For example, SINGULAR is formally unmarked and 'non-featured', and -SINGULAR is formally marked and therefore 'featured'. For further discussion of Naturalness Principles, see Wheeler (1993).

5. Discussion of results

Our results prove interesting. We find relations between frequency and irregularity and a certain degree of correspondence with the irregularity ranking we outlined in Section 4. We also find evidence for a split between prosodic and non-prosodic morphology. Finally we find one intriguing area related to particular cells, where it appears that there might be a relationship between *cell anomaly* and irregularity of the nominative plural. In fact, this turns out to be *plural anomaly* in disguise.

5.1 Absolute plural anomaly

The first of our hypotheses, Hypothesis 1a, is confirmed. There is a relation between *absolute plural anomaly* and irregularity. Below we give eight groups of nouns from the corpus divided up according to our irregularity ranking in (7). In addition, we make a distinction between two stress patterns which divide the singular and plural and would both, therefore, share the same irregularity ranking. These patterns are, according to the classification in Zaliznjak (1977): pattern C (stem stress throughout

Table 4. *Absolute plural anomaly in eight groups of nouns*

	Type of irregularity	Stress pattern	Median plural count	Observed number of types	p-value ¹⁴
Group 1	end stress pl	C	9	64	< 0.001
Group 2	end stress sg	D	5	80	< 0.05
Group 3	stem stress alternation	n/a	22	2	0.25
Group 4	stem alternation	n/a	96	3	< 0.001
Group 5	stem augment in pl	n/a	10	24	< 0.001
Group 6	stem augment in sg	n/a	15	10	< 0.05
Group 7	stem augment in both	n/a	14	14	< 0.05
Group 8	suppletion	n/a	935.5	3	< 0.001

singular, ending stress throughout plural); pattern D (ending stress throughout singular, stem stress throughout plural). The eight groups are given in Table 4.

For each of the groups in Table 4 the median value for plural occurrences is significantly higher than for the corpus as a whole, with the single exception of Group 3 (see p-values in the table).

If we were to rank each group in increasing order according to the median value, we would get the following: Group 2, Group 1, Group 5, Group 7, Group 6, Group 3, Group 4, Group 8. The data do not support irrefutably such an ordering because, despite the fact that the *anomalies* for seven of the groups are significant, the differences between the groups are in some cases insignificant. This also means that the ordering in (7) has not been disproved: the data here could still be consistent with the principled ordering of the Irregularity Ranking, which is an interesting result. Groups 3 and 4 have small sample sizes which means their place in the ordering suggested by Table 4 should be treated with some scepticism.

What is conclusively shown from our investigation is that both singular augments and plural augments are related to *absolute plural anomaly*. This is significant. While we might argue that singular augments mark the unexpected number with *plural anomaly*, this cannot be the case with plural augments, which mark what is the expected number. In other words, it appears that having an augment throughout a particular number (irrespective of whether it is singular or plural) is related to a lexeme having a high *plural anomaly*. We might have expected an augment in the plural to be associated with higher occurrence of singulars than the average for the corpus. The opposite is the case. In sum there is a relationship between frequency and irregularity in absolute terms, but we must now test our Hypothesis 1b in order to see if this is true in relative terms.

5.2 *Relative plural anomaly*

Groups 1–8 were tested for the next of our hypotheses. Evidence for Hypothesis 1b turns out to be not as strong as that for Hypothesis 1a. It involves groups of a specific type. We find evidence for Hypothesis 1b for two groups and, arguably, for a third. The stronger evidence is for group 6 (where there is a stem augment in the singular), and group 5 (where there is a stem augment in the plural). The weaker evidence is for group 4 (where there is a stem alternation). In each case the irregularity is segmental rather than prosodic. The results are given in Table 5.

As the data in Table 5 show, there is some evidence that the frequency of occurrence of the irregular forms, and not just frequency of occurrence of the lexeme as a whole *does* relate to irregularity of the forms in question. However, if the irregularity affecting an entire subparadigm is a prosodic one, there is no evidence for a relationship between this irregularity and high relative frequency. In the box plot in

Table 5. *Relative plural anomaly*

Group	Type of irregularity	Median plural proportion	p-value
1	end stress pl	0.2	0.1
2	end stress sg	0.15	0.54
3	stem stress alternation	0.18	0.54
4	stem alternation	0.68	0.06
5	stem augment in pl	0.36	0.03
6	stem augment in sg	0.82	< 0.001
7	stem augment in both	0.32	0.4
8	suppletion	0.62	0.16

Figure 2 the prosodic groups (Groups 1, 2, and 3) have much lower medians than the others.¹⁵

The median is represented by the white line in the middle of the box; the box itself represents a range of proportions covering the middle 50% of the lexemes in

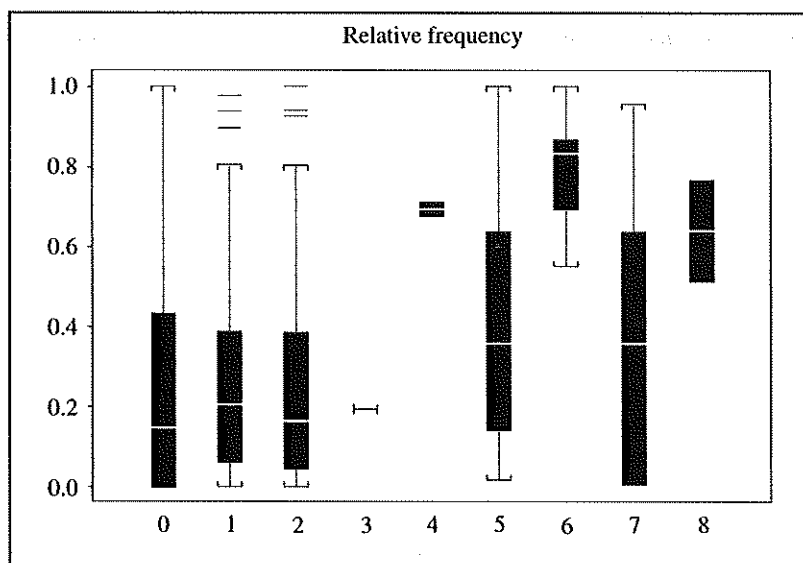


Figure 2. *Irregularity type and relative plural anomaly.*

Key: y axis = proportion of plurals, x axis = irregularity type: 0 = regular; 1 = stress C, 2 = stress D, 3 = stem stress alternation, 4 = stem segment alternation, 5 = stem augment in plural, 6 = stem augment in singular; 7 = different stem augment in singular and plural, 8 = suppletion

the category; the whiskers cover the remaining 50%, except outliers which are indicated separately with horizontal bars (Daley *et al.* 1995). It is an extremely interesting and important result to find the answer to the question we posed at the beginning of this paper is that *relative frequency* of occurrence in the plural appears to be important where non-prosodic irregularity is concerned, but not where prosodic irregularity is concerned. Thus degree of irregularity is important.

Prosodic irregularities involve a high *absolute plural anomaly* but no *relative plural anomaly*. This means that there is a high number of plurals (*absolute plural anomaly*) and also a high number of singulars to match the plurals (no *relative plural anomaly*). Thus the prosodic irregularity may relate to the frequency of occurrence of the lexeme as a whole (the lexeme's plural must be frequent to give the *absolute anomaly*, and its singular must also be frequent, otherwise it would show *relative plural anomaly*). In contrast, the fact that certain non-prosodic irregularities have significant *relative plural anomalies* indicates that these may be related to the frequency of occurrence of a subparadigm, namely the plural.

Hence we have identified an effect which applies to the lexeme as a whole, and one which applies to the plural subparadigm. Having found effects applying to the highest level (the lexeme) and a middle level (the subparadigm), we look to see whether we can find any effect relating to the lowest level, that of the single cell.

5.3 Cell anomaly

Delving deeper into the paradigm, we look to see if frequency of occurrence of individual case and number cells can be related to their irregularity, as we had gauged it according to the ranking. We look at the *absolute frequency* of occurrences for all cells of given lexemes with one individual irregular cell.¹⁶

We look at Hypothesis 2. Having confirmed Hypothesis 1, recall that Hypothesis 2 is testing for a stronger relationship based on cell irregularity and *cell anomaly*. Since we are looking for an effect not caused by high lexeme frequency, we must concentrate on cells which do not have a significantly high lexeme frequency. From Table 6 we see that the nominative plural is the only candidate with which to test the hypothesis: in terms of lexeme frequency those inanimate lexemes with a cell irregularity in the nominative plural are not significantly more frequent than in the corpus in general.¹⁷ From Table 6 note that for other lexemes with observed forms of cell irregularity there is a significant effect, and in these instances instead of *cell anomaly* we may be seeing lexeme frequency in disguise. We therefore concentrate on the nominative plural *cell anomaly* as a good candidate to test Hypothesis 2.

In order to compare like with like, we test for *cell anomaly* in the nominative plural of inanimates only. There are two tests. For the first test we find that in the

Table 6. *Cell irregularity and lexeme frequency*

Irregular cell	Median Lexeme Frequency	p-values
Accusative singular	838	< 0.001
Instrumental singular	158	< 0.01
Nominative plural (animates)	15	0.477
Nominative plural (inanimates)	14	0.34
Genitive plural	50	< 0.05
Instrumental plural	2,771	< 0.01

nominative plural cell there is a significantly higher ($p < 0.01$) proportion of instances of that cell in those lexemes with the cell irregularity. We then need to test if this is due to *relative plural anomaly*. In fact, the increase in the proportion of the nominative plural cell can be explained by the increase in the total plural proportion for those lexemes which have the irregularity. There is no evidence ($p = 0.5$) for an increase in the cell frequency as a proportion of the plural frequency.¹⁸ In sum, this means that having an odd nominative plural cell seems to be connected with having proportionally more plurals than expected. In other words, we are not observing a *cell anomaly*. Thus we found little evidence confirming Hypothesis 2.

As was expected, Hypothesis 3 is disconfirmed: there is a relationship between irregularity and the *absolute frequency* of the cell which is irregular. See Table 7.

Table 7. *Irregularity and absolute frequency of cell*

Irregular cell	Median cell frequency	p-value
Accusative singular	172	< 0.001
Instrumental singular	16	< 0.01
Nominative plural	2	< 0.01
Genitive plural	2	< 0.05
Instrumental plural	99	< 0.001

Given that Hypothesis 1a is true, our result for Hypothesis 3 is not surprising. The relationship simply falls out from the relationship specified in Hypothesis 1. It was formulated to cover for the case where Hypothesis 1a was disproved.

6. Conclusions

Our Hypothesis 1a, that there is a relation between *absolute plural anomaly* and irregularity, is strongly confirmed. More specifically, nouns which have an irregu-

larity involving a split between singular and plural will tend to be nouns which occur frequently in the plural. There is a less dramatic but still significant effect when only stress is involved. The only instance where we did not find a significant effect was where there was a stress alternation involving the stem only. Apart from that, there are some indications of a relation between the degree of irregularity, as postulated in advance independently of frequency, and the degree of *plural anomaly*, with cases of suppletion being an extreme case.

Hypothesis 1b, that there is a relation between *relative plural anomaly* and irregularity was less strongly confirmed. Recall that here we are concerned with the plural forms of a lexeme as a proportion of all occurrences of the lexeme. For those types where we did observe an effect, where the plural was used in proportion to the singular significantly more frequently than found generally through the corpus, the irregularity was always a segmental one—stress irregularity was not sufficient to produce an effect here. Furthermore whether the irregularity concerns the singular or the plural, we still find a high relative plural frequency (for instance nouns with an augment in the singular still have a high plural *relative frequency*).

When we moved down to examine single cells (Hypothesis 2), we found no evidence that irregularity is related to a high *relative frequency* of a specific cell in the paradigm, once the effects discussed under Hypothesis 1a and 1b are factored out. This is an interesting result, since it implies a structuring of lexical items. It suggests that an individual irregular cell does not stand out from its subparadigm (singular or plural) in terms of frequency. (Hypothesis 3 was included to cover outcomes which did not arise and so needs no further discussion here.)

There are three morphological levels which might be relevant for frequency effects. The first is the level of the lexeme as a whole; the second is the level of the subparadigm of the lexeme; and the third is the level of the individual cell. We found no evidence for an effect relating to the third of these levels. We did find evidence for a relation with the other two. This relation may be sensitive to the type of the irregularity. For the relation with the first level, the lexeme as whole, the clearer evidence comes from prosodic irregularities (as we argued in Section 4.2). For the second of these levels, the number subparadigm, there is evidence from non-prosodic types of irregularity (Figure 2). This shows the importance of looking at languages such as Russian with extensive paradigms. In languages where there is just a singular/plural split, with no other category distinction, we could not separate sub-paradigm from individual cell. Given this we see that the relation between irregularity and high frequency is more intricate than we imagined once we pull apart the notion of irregularity on the one hand, in term of a ranking of irregularity, and frequency on the other, in terms of a distinction between *absolute* and *relative frequency*. The conclusion we draw is that there is a relationship, but the relationship

is a complex one which depends on the type of frequency concerned and the degree of irregularity in question.

Appendix 1: Statistics methodology

Testing differences between median values.

In order to test the differences between the median values of two groups, bootstrap testing was used, see Efron and Tibshirani (1993).

Let us suppose that a subset of lexemes S has been extracted from the corpus C according to some linguistic criterion, usually based on regularity. We calculate the median frequency of the distribution of the required frequency. Let us denote this to be $m(S)$ in the subset S and $m(C)$ in the full corpus, C . We need to see if $m(S)$ is significantly different from $m(C)$ assuming the Null Hypothesis that there is no relationship between the extraction criterion (irregularity) and the measure quantity (frequency).

Under this assumption we can evaluate the distribution of $m(S)$ by randomly selecting (with replacement) samples of equal size to S from C , and calculating their median. This procedure is repeated many times and an estimate of the underlying distribution of the median is constructed. This will be the bootstrap distribution of the median under the assumed hypothesis. The actual value of $m(S)$ can then be compared to this bootstrapped distribution to see if it is significantly higher or lower than expected. A p -value can then be directly calculated from the bootstrap distribution. For details of this procedure see Efron and Tibshirani (1993: Ch. 13).

Appendix 2: The Irregularity Scale with examples

The table gives a small number of examples with their irregularity scores.

0	komnata	fully regular
0.1	sad	singular-plural stress difference
0.2	polosá	stress different in one form
0.3	volk	singular-plural stress difference and irregular nominative plural form
0.4	borodá	stress differs in two forms
1.0		
1.1	lič'iko	different inflection (segmental) singular versus plural
1.11	glaz	different inflection (segmental) singular versus plural; and inflectional stress difference singular versus plural

1.2	soldat	irregular inflection (segmental) in one cell (genitive plural)
1.21	dom	irregular inflectional form in one cell (nominative plural); and inflectional stress difference singular versus plural
2.0		
2.1	ó'oro	different stem stress singular versus plural
3.0		
3.1	otec	motivated stem alternation in expected cell
3.201	koleno	unmotivated stem alternation in singular versus plural; different inflection in singular versus plural
3.4	pesn'a	motivated stem alternation in expected cell; plus unmotivated stem alternation in one cell (genitive plural)
4.0		
4.1		
4.10001	nebo	stem augment in plural only; singular versus plural inflectional stress difference
4.2		
4.2004	krestjan'in	stem augment in singular and not plural; irregular inflectional form in two cells (nominative plural and accusative/genitive plural)
4.3001	xoz'ajin	stem augment both in singular and plural; different inflection singular versus plural
4.40002	doč'	stem augment throughout except for one cell (nominative/accusative singular); inflectional stress irregularity in one cell (nominative plural)
5.0		
5.1	vorota	plurale tantum
6.0		
6.1		
6.100002	č'elovek	suppletion singular vs plural; segmental inflectional irregularity in one cell (instrumental plural)
6.2	god	suppletion in one cell (genitive plural)

Notes

* This is a joint paper: Hippiisley contributed the lion's share of the corpus work, Marriott is responsible for the statistical analysis, and other aspects were shared. Corbett and Brown are from the Department of Linguistic and International Studies, and Hippiisley from the Department of Computing, University of Surrey; Marriott is from the Department of Statistics and Applied Probability at the National University of Singapore. The research reported here was supported by the ESRC (grant no. R000222419, and in part grant no. R000237845); we are grateful for this support. A grant from the University of Surrey Research Promotion Initiative is also gratefully acknowledged. Sections of this research were presented at the ESRC seminar series Challenges for Inflectional Description, University of Surrey, June 24, 1998 and at the Linguistics Association of Great Britain, University of Manchester, April 8–10, 1999. We are grateful to those present there and at the conference 'Frequency effects and emergent grammar' for lively discussion. We would like to thank Alan Timberlake for his valuable input to discussions about ranking of irregularity, Harald Clahsen and Andrew Spencer for bringing useful references to our attention, and Maria Polinsky for sharing her judgements on certain Russian forms. Finally we wish to thank Joan Bybee, Paul Hopper, and Catie Berkenfield for their helpful comments.

1. Using Schreuder and Baayen's terms, do we base the frequency on the 'stem-frequency', i.e., the sum of all word forms, or the 'surface frequency', i.e., the sum of one of the word forms (1997)? They were looking at the effect of subjective frequency and visual perception times, and for them it was important to distinguish a number of different counts: 'surface frequency', 'stem frequency', 'morphological family size' (the number of members of a derivational family), and 'cumulative family frequency' (the stem frequency of all members of the derivational family apart from the base). In their work on derivation, they found that this last frequency had surprisingly no effect, whereas morphological family size did.
2. Russian orthography closely follows phonemic representation and the phonemic transcription we use is, therefore, close to standard transliteration, with a few minor points of difference (based on Corbett and Fraser 1993:fn. 2). For an outline of Russian phonology, see Timberlake (1993: 828–32). The main points are summarized as follows:

Consonants

The set of paired palatalized (soft) and unpalatalized (hard) consonants are distinguished by an acute (´) which marks the soft member of the pair. For example, in the minimal pair *l'uk* 'hatchway', and *luk* 'onion' the first form has the soft /l/. Note that consonants are always soft before the phoneme /e/, hence there is no need to mark them with an acute in this context. For example, the locative singular of *zakón* 'law' is represented as *zakóne* since the stem final /n/ is automatically soft.

The velars /g/, /k/, and /x/ are hard except when preceding the /i/ and /e/ phonemes; in these contexts they are automatically softened. We therefore do not use an acute on the velars in these contexts since they are automatically softened. Compare the nominative singular form *ruč'ka* 'handle' with the genitive singular *ruč'ki*, where the /k/ is soft before the -i ending, but not indicated as such. Note that unpaired soft /č/ and /šč/ are redundantly marked with an acute when preceding a vowel, but unpaired soft /j/ is never marked with an acute.

Vowels

We recognize five vowel phonemes (under stress) which are /a/, /e/, /i/, /o/, and /u/. The phoneme /i/, standardly transliterated as 'i', has an allophone [i̯], standardly transliterated as 'y'. The allophone [i̯] is automatically used when following a hard consonant. The correct version of /i/ will therefore be implied by the nature of the preceding consonant.

3. Absolute singular and relative singular *anomaly* are defined analogously.
4. In this instance and throughout the paper we use the word “significant” to mean statistically significant.
5. The basic dataset is available on the world wide web and can be found at <http://surrey.ac.uk/LIS/SMG>, along with a readme file.
6. This notion is akin to Wurzel’s concept of “implicative paradigm structure conditions” (Wurzel 1987: 76–8).
7. For simplicity the few instances of the second genitive in *-u* and the second locative in *-i*, available for class I nouns only, were treated with the corresponding named cases. These can be thought of as ‘sub-cases’ realized by a small minority of nouns in specific contexts. The second locative occurs with the locational prepositions *v* ‘in’ and *na* ‘on’, as for example in *v sneg-i* ‘in the snow’; the second genitive is used in partitive constructions such as *ja ne vip’il čaj-u* ‘I didn’t drink any tea’. See the discussion in Timberlake (1983: 838) for details.
8. In these paradigms the nominative and accusative singular are also identical, but this is not so for all inflectional types in Russian (see Table 1).
9. Where stress is expected on the ending, and yet there is no ending (as in the genitive plural *mest* ‘places’), the stress automatically falls on the last syllable of the stem. Since this is fully automatic we do not count it as an irregularity.
10. We use ‘prosodic’ to cover tone, pitch, and stress. Of these only stress is relevant for Russian.
11. It should be noted that in order to investigate a more fine grained relationship between irregularity and high frequency, the ranking in (7) was converted into a numerical scale to provide for all instances of irregularity, including combinations such as stress and inflectional irregularity. Though nothing significant emerged from the finer-grained rankings, we include the scale with examples in Appendix 2 for completeness.
12. We take suppletion, pluralia tantum, and stem augments to be irregular by definition.
13. For example one such universal process is assimilation, phonologically natural because it “eases articulatory effort by allowing inertia to prevail and smoothing transition from one segment to another” (Dressler 1985: 49). Russian word final devoicing matches this universal process perfectly, and therefore receives the score 1 for phonological naturalness (1985: 59). For full details of all scores, see Dressler (1985: 59–66).
14. The p-value represents the probability that a median value more extreme than that observed could have occurred purely by chance. A value < 0.05 is reasonable evidence that there is a relationship between *anomaly* and irregularity. A value < 0.01 is strong evidence that there is a relationship.
15. Recall that for Hypothesis 1a this does not exclude a relationship between *absolute frequency* and irregularity for these prosodic irregularities.

16. This includes lexemes for which the cell in question is the only irregularity, as well as lexemes for which the cell irregularity is accompanied by a singular-plural irregularity defined independently of that cell irregularity. For example, the lexeme *dom* 'house' has a change in stress between the singular and plural (a singular-plural irregularity). In addition, it has an unexpected nominative plural ending *-a*. This is an irregularity in a single cell which accompanies another irregularity.
17. As discussed in Section 4.1.2 we treat nominative plural and accusative plural of inanimates as one cell. This means that we must restrict our comparison to inanimates only, and exclude animates. When we use the terms 'nominative plural' for inanimates we mean the cell which includes what are, in fact, distributionally accusative plurals.
18. We have also checked the ten inanimates from this group which only have the cell irregularity. There is no significant difference between these and the inanimates as a whole with regard to *cell anomaly* ($p=0.2$).

References

- Aronoff, M. 1994. *Morphology by Itself: Stems and Inflectional Classes*. Cambridge, MA: MIT Press.
- Brown, D., Corbett, G. G., Fraser, N. M., Hippisley, A., and Timberlake, A. 1996. "Russian noun stress and Network Morphology". *Linguistics* 34(1): 53–107.
- Brown, D. and Hippisley, A. 1994. "Conflict in Russian genitive plural assignment: A solution represented in DATR". *Journal of Slavic Languages* 2: 48–76.
- Bybee, J. 1985. *Morphology: a Study of the Relation between Meaning and Form*. Amsterdam/Philadelphia: John Benjamins.
- Bybee, J. 1995. "Regular morphology and the lexicon". *Language and Cognitive Processes* 10(5): 425–55.
- Bybee, J. and Thompson, S. 1997. "Three frequency effects in syntax". *BLS*.
- Comrie, B. 1986. "On delimiting cases". In *Case in Slavic*, R. D. Brecht and J. S. Levine (eds.), 86–106. Columbus, Ohio: Slavica.
- Comrie, B. 1991. "Form and function in identifying cases". In *Paradigms: the Economy of Inflection*, F. Plank (ed.), (Empirical Approaches to Language Typology 9), 41–55. Berlin/New York: Mouton de Gruyter.
- Corbett, G. 1982. "Gender in Russian, an account of gender specification and its relationship to declension". *Russian Linguistics* 6(2): 197–232.
- Corbett, G. G. and Fraser, N. M. 1993. "Network morphology: A DATR account of Russian inflectional morphology". *Journal of Linguistics* 29: 113–42.
- Daley, F., Hand, D., Jones, C., Lunn, D., and McConway, K. 1995. *Elements of Statistics*. London: Addison-Wesley.
- Dressler, W. 1985. *Morphology: the Dynamics of Derivation*. Ann Arbor: Karoma.
- Dressler, W. 1987. "Word-formation as part of natural morphology". In *Leitmotifs in Natural Morphology*, W. Dressler (ed.), 99–126. Amsterdam/ Philadelphia: John Benjamins.
- Efron, B. and Tibshirani, R. J. 1993. *An Introduction to the Bootstrap*. New York/London: Chapman and Hall.

- Fraser, N. and Corbett, G. 1995. "Gender, animacy and declensional class assignment: a unified account for Russian". In *Yearbook of Morphology 1994*. G. Booij and J. van Marle (eds.), 123–50. Dordrecht: Kluwer.
- Greenberg, J. 1966. *Language Universals, with Special Reference to Feature Hierarchies*. The Hague: Mouton.
- Jakobson, R. 1932. "Zur Struktur des russischen Verbums". In *Charisteria Gvilelmo Mathesio qvinqvagenario a discipulis et Circuli Lingvistici Pragensis sodalibus oblata*, 74–84. Prague: Pražký Lingvistický Kroužek. [Reprinted in Jakobson's *Selected Writings II*, 3–15]. The Hague: Mouton.]
- Lönnngren, L. 1993. *Častotnyj slovar' sovremennogo russkogo jazyka*. (Acta Universitatis Upsaliensis, Studia Slavica Upsaliensis 33). University of Uppsala: Uppsala.
- Maier, I. 1994. "Review of Lennart Lönnngren (ed.) *Častotnyj slovar' sovremennogo russkogo jazyka*". *Rusistika Segodnja* 1: 130–6.
- Mayerthaler, W. 1987. "System-independent morphological naturalness". In *Leitmotifs in Natural Morphology*, W. Dressler (ed.), 25–58. Amsterdam/Philadelphia: John Benjamins.
- Schreuder, R. and Baayen, H. 1997. "How complex simplex words can be". *Journal of Memory and Language* 37(1): 118–39.
- Tiersma, P. 1982. "Local and general markedness". *Language* 58(4): 832–49.
- Timberlake, A. 1993. "Russian". In *The Slavonic Languages*, B. Comrie and G. Corbett (eds.), 827–86. London/New York: Routledge.
- Wheeler, M. W. 1993. "On the hierarchy of naturalness principles in inflectional morphology". *Journal of Linguistics* 29: 95–111.
- Wurzel, W. 1987. "System-dependent morphological naturalness in inflection". In *Leitmotifs in Natural Morphology*, W. Dressler (ed.), 59–96. Amsterdam/Philadelphia: John Benjamins.
- Wurzel, W. 1989. *Inflectional Morphology and Naturalness*. Dordrecht: Kluwer.
- Zaliznjak, A. A. 1977. *Grammatičeskij slovar' russkogo jazyka*. Moscow: Russkij jazyk.

Data sources

1. For the Uppsala corpus, see Lönnngren (1993).
2. The basic dataset referred to in note 5 is available at <http://surrey.ac.uk/LIS/SMG>, along with a readme file.

In the series TYPOLOGICAL STUDIES IN LANGUAGE (TSL) the following titles have been published thus far:

18. HAIMAN, John & Sandra A. THOMPSON (eds): *Clause Combining in Grammar and Discourse*. 1988.
19. TRAUGOTT, Elizabeth C. and Bernd HEINE (eds): *Approaches to Grammaticalization, 2 volumes (set)* 1991
20. CROFT, William, Suzanne KEMMER and Keith DENNING (eds): *Studies in Typology and Diachrony. Papers presented to Joseph H. Greenberg on his 75th birthday*. 1990.
21. DOWNING, Pamela, Susan D. LIMA and Michael NOONAN (eds): *The Linguistics of Literacy*. 1992.
22. PAYNE, Doris (ed.): *Pragmatics of Word Order Flexibility*. 1992.
23. KEMMER, Suzanne: *The Middle Voice*. 1993.
24. PERKINS, Revere D.: *Deixis, Grammar, and Culture*. 1992.
25. SVOROU, Soteria: *The Grammar of Space*. 1994.
26. LORD, Carol: *Historical Change in Serial Verb Constructions*. 1993.
27. FOX, Barbara and Paul J. Hopper (eds): *Voice: Form and Function*. 1994.
28. GIVÓN, T. (ed.): *Voice and Inversion*. 1994.
29. KAHREL, Peter and René van den BERG (eds): *Typological Studies in Negation*. 1994.
30. DOWNING, Pamela and Michael NOONAN: *Word Order in Discourse*. 1995.
31. GERNSBACHER, M. A. and T. GIVÓN (eds): *Coherence in Spontaneous Text*. 1995.
32. BYBEE, Joan and Suzanne FLEISCHMAN (eds): *Modality in Grammar and Discourse*. 1995.
33. FOX, Barbara (ed.): *Studies in Anaphora*. 1996.
34. GIVÓN, T. (ed.): *Conversation. Cognitive, communicative and social perspectives*. 1997.
35. GIVÓN, T. (ed.): *Grammatical Relations. A functionalist perspective*. 1997.
36. NEWMAN, John (ed.): *The Linguistics of Giving*. 1998.
37. RAMAT, Anna Giacalone and Paul J. HOPPER (eds): *The Limits of Grammaticalization*. 1998.
38. SIEWIERSKA, Anna and Jae Jung SONG (eds): *Case, Typology and Grammar. In honor of Barry J. Blake*. 1998.
39. PAYNE, Doris L. and Immanuel BARSHI (eds.): *External Possession*. 1999.
40. FRAJZYNGIER, Zygmunt and Traci S. CURL (eds.): *Reflexives. Forms and functions*. 2000.
41. FRAJZYNGIER, Zygmunt and Traci S. CURL (eds): *Reciprocals. Forms and functions*. 2000.
42. DIESSEL, Holger: *Demonstratives. Form, function and grammaticalization*. 1999.
43. GILDEA, Spike (ed.): *Reconstructing Grammar. Comparative Linguistics and Grammaticalization*. n.y.p.
44. VOELTZ, F.K. Erhard and Christa KILLIAN-HATZ (eds.): *Ideophones*. n.y.p.
45. BYBEE, Joan and Paul HOPPER (eds.): *Frequency and the Emergence of Linguistic Structure*. 2001.
46. AIKHENVALD, Alexandra Y., R.M.W. DIXON and Masayuki ONISHI (eds.): *Non-canonical Marking of Subjects and Objects*. n.y.p.
47. BARON, Irene, Michael HERSLUND and Finn SORENSEN (eds.): *Dimensions of Possession*. n.y.p.

A full list of titles published in this series is available from the publisher.

