to appear in *Linguistics*, 2017

Explaining asymmetries in number marking: Singulatives, pluratives and usage frequency

MARTIN HASPELMATH[1] & ANDRES KARJUS[2]
[1]Max Planck Institute for the Science of Human History & [2]University of Tartu
(haspelmath@shh.mpg.de, andres.karjus@ut.ee)

Abstract: This paper claims that cross-linguistic tendencies of number marking asymmetries can be explained with reference to usage frequency: The kinds of nouns which, across languages, tend to show singulative coding (with special marking of the uniplex member of a pair), rather than the more usual plurative coding (with special marking of the multiplex member), are also the kinds of nouns which tend to occur more frequently in multiplex use. We provide cross-linguistic coding evidence from a range of languages from different families and areas, and cross-linguistic corpus evidence from five languages, using large written corpora. Thus, the cross-linguistic pattern of singulative vs. plurative coding is a special instance of the tendency to devote more marking to rarer forms, and can be explained by the grammatical form-frequency correspondence principle.

Keywords: number marking, cross-linguistic tendencies, markedness, corpora, usage-based

1. The claim

In this paper, we propose an explanation for number-marking asymmetries such as those in (1) and (2). In (1), the form denoting a multiplex entity ('days') has an overt marker, and in (2), it is the form denoting a single (uniplex) entity ('pea') that has an overt (singulative) marker *-en*. The other form (the "basic form") has no overt marker.

(1) German

a.	*Tag-Ø*	'day'	(basic form)
b.	*Tag-e*	'days'	(plurative form)

(2) Welsh

a.	*pys-Ø*	'peas'	(basic form)
b.	*pys-en*	'pea'	(singulative form)

We will show that the marking asymmetries seen in these examples follow a cross-linguistic trend, and we claim that the trend can be explained by a parallel cross-linguistic usage trend: Many nouns such as 'day' tend to be used more frequently in a uniplex sense (denoting a single entity), while some nouns such as 'pea' are used more frequently in a multiplex sense (denoting a set of multiple entities). Those that tend to be used more frequently in a uniplex sense, called UNIPLEX-PROMINENT here, tend to show overt marking of the multiplex form (i.e. plurative form), while those that tend to be used more frequently in a multiplex sense, called MULTIPLEX-PROMINENT here, tend to show overt marking of the uniplex form (i.e. singulative form).

The explanatory principle here is Zipfian economy (Zipf 1935; Haspelmath 2008). It has been invoked to account for a wide variety of form asymmetries which correspond to frequency asymmetries (e.g. Greenberg 1966; Croft 2003; Hawkins 2004; Bybee 2007). In earlier work (Haspelmath et al. 2014), the specific principle as applied to grammar has been formulated as in (3).

(3) The grammatical form-frequency correspondence principle
When two minimally different grammatical patterns (i.e. patterns that form an opposition) occur with significantly different frequencies, the less frequent pattern tends to be overtly coded (or coded with more coding material), while the more frequent pattern tends to be zero-coded (or coded with less coding material).

Some further grammatical oppositions for which this principle has been invoked are listed in (4). This is thus a very broadly applicable principle with great explanatory power.

(4) present/future, 3rd person/2nd person, nominative/accusative, active/passive, affirmative/negative, masculine/feminine, attributive adjective/predicative adjective (including copula), positive/comparative, predicative verb/nominalized verb, action word/agent noun

Greenberg (1963) was perhaps the first to observe that the singular-plural overtness contrast is a universal tendency of human languages:

(5) Greenberg's Universal 35 (partial)
There is no language where the plural does not have some nonzero allomorphs, whereas there are languages in which the singular is expressed only by zero.

Thus, the situation in (1) (German *Tag-Ø/Tag-e*) is quite typical of the world's languages. By contrast, the situation in (2) (Welsh *pys-Ø/pys-en*) is unusual, and all languages with such singulative coding also have the ordinary plurative coding for other nouns. Greenberg (1966: 31–32) observed that the coding asymmetry between singular and plural corresponds to a frequency asymmetry (cf. §3 below).

This paper goes beyond Greenberg in showing that the Zipfian frequency-based explanation can account not only for the general trend of (5), but also for the difference between (1) and (2), i.e. between plurative pairs and singulative pairs. We provide corpus evidence from five languages, showing that cross-linguistically, the kinds of nouns that tend to be coded as singulatives (in languages that exhibit overt singulative marking) are more frequent in multiplex use, while the kinds of nouns coded as pluratives are more frequent in uniplex use. In simplified terms, we can say that German *Tag* 'day' has no suffix because the singular is more frequent than the plural, while Welsh *pys-en* 'pea' has a suffix because the singular is less frequent than the plural. For example, in the British National Corpus of English, the frequency of *day/days* is 59298/31542, while the frequency of *pea/peas* is 173/603. The distribution in other languages is presumably quite similar.

Before getting to the details of our story, we need to introduce our terminology for semantic and formal entities (§2), and it will be useful to contrast our frequency-based explanation with an explanation in terms of "markedness" (§3).

2. Basic comparative concepts: Notional and formal

The terms *singular* and *plural* are typically used both in a semantic sense and in the sense of a language-specific formal grammatical category. For this paper, it is crucial to have comparative concepts that clearly pertain to the notional level (*uniplex* and *multiplex*), as well as concepts that clearly refer to kinds of asymmetric marking (*singulative* and *plurative*). Since our goal is limited to explaining cross-linguistic trends, we do not worry about language-specific analysis here. The terms *singular* and *plural* thus play no significant role in this paper.

The notional terms *uniplex* and *multiplex* are used here as in Talmy (1988). Multiplex nominals are nominals that denote entities which are (or can be) conceived of as (internally homogeneous) groups of things (and which therefore are expressed by overt plural forms in many languages). Uniplex nominals denote entities which are conceived of as individuals. Some examples are given in (6). Uniplex nominals are singular nominals in English, and some of them can have the singulative suffix in Welsh. Multiplex nominals are generally plural in English (most often with a plural suffix -*s*), but they can be mass nouns like *hair*, and in Welsh they may be simple root nouns lacking a suffix. (Such simple roots are often called *collective nouns* rather than plural nouns; see Gil (1996) for the wide range of meanings with which this term has been used.)

(6)

UNIPLEX NOMINALS	MULTIPLEX NOMINALS
day-Ø	*day-s*
bee-Ø	*bee-s*
mouse	*mice*
(a) fish	*(many) fish*
(a) hair	*(she has black) hair*
Welsh *pys-en* ‘pea’	*pys* ‘peas’
Welsh *moron-en* ‘carrot’	*moron* ‘carrots’

Nominal meanings which frequently occur in multiplex use (e.g. ‘peas’) can be called *multiplex-prominent meanings*. As we will see below, these occur particularly in the semantic domains of paired body-parts, small animals, fruits/vegetables, and groups of people.

The key formal concepts of this study are *singulative* and *plurative*, or more precisely *basic/plurative pairs* and *singulative/basic pairs*. A basic/plurative pair is a pair of related noun forms where one member is an unmarked (basic) uniplex noun (e.g. *day*), while the other member is a marked multiplex noun (e.g. *day-s*). Since this situation is extremely common in the world’s languages, the great majority of “plural” forms are actually pluratives in this sense.[1] A singulative/basic pair is a pair of noun forms where one member is a marked uniplex nominal (e.g. Welsh *moron-en* ‘carrot’), while the other member is an unmarked multiplex nominal (e.g. Welsh *moron* ‘carrots’). Since this situation is quite rare, few “singulars” are singulatives. A few more examples of basic/plurative pairs (or plurative lexemes) and of singulative/basic pairs (or singulative lexemes) are given in (7).

[1] The term *plurative* (in this sense) is a terminological innovation of this paper. The term has occasionally been used before (e.g. Treis 2014), but apparently mostly for overt plurals that coexist with unmarked multiplex forms which have a singulative counterpart.

(7) BASIC/PLURATIVE PAIRS (*=plurative lexemes*)	SINGULATIVE/BASIC PAIRS (*= singulative lexemes*)	
German	Maltese	
Schuh / Schuh-e	*zarbun-a / zarbun*	'shoe/shoes'
Fisch / Fisch-e	*ħut-a / ħut*	'fish (sg.)/fish (pl.)'
Apfelsine / Apfelsine-n	*laringġ-a / laringġ*	'orange/oranges'
Estonian	Welsh	
tigu / teo-d	*malwod-en / malwod*	'snail/snails'
karv / karva-d	*blew-yn / blew*	'hair/hair(s)'
hernes / herne-d	*pys-en / pys*	'pea/peas'

Singulative lexemes are found in some languages such as Welsh and Maltese, but they are not widespread in the world's languages.[2] Most languages do not have singulative lexemes at all.[3] Nevertheless, we claim in this paper that the occurrence of singulatives is not accidental, but is a manifestation of a cross-linguistic tendency.

3. Markedness explanation vs. frequency explanation

In the literature, form asymmetries of the type seen in (1) and (2) are commonly talked about or explained with reference to a notion of "markedness" (cf. Tiersma 1982; Haspelmath 2006). The contrast between "marked" and "unmarked" values of a grammatical feature was highlighted in a typological context by Greenberg (1966) (see also Croft 2003: 87–101).

The fundamental idea here is that languages exhibit some kind of "markedness matching" (Haspelmath 2008: 6–7), in such a way that marked values of grammatical categories are formally marked (overtly coded), while unmarked values are formally unmarked (zero-coded). A closely related approach is to say that languages tend to give simple expression to semantically simple values of grammatical categories, while semantic complexity is reflected in formal complexity, i.e. overt coding. Thus, Mayerthaler (1981: 25) says that "What is "more" semantically should also be "more" constructionally", and Givón (1991: §2.2) puts it quite similarly: "A larger chunk of information will be given a larger chunk of code". This has also been regarded as a kind of iconicity (*iconicity of complexity* in Haspelmath's (2008: 6) terminology). According to this view, one would say that the singular tends to be zero-coded (cf. Greenberg's Universal 35 in (5) above) because it is semantically unmarked or simple, while the plural is semantically marked or complex.

The frequency explanation, by contrast, would say that the singular tends to be zero-coded because it is more frequent than the plural, and is thus more predictable. Languages generally use more coding for less predictable meanings. This explanation was first proposed by Greenberg, who noted that singular forms tend to

[2] As shown by Cuzzolin (1998), the term *singulative* was coined in the 19th century with reference to Welsh, but it was soon extended to similar phenomena in Semitic languages and elsewhere.

[3] In fact, many languages do not have plural forms, or use plurals only optionally and/or for a restricted set of (mostly animate) nouns (cf. Haspelmath 2005). We see no reason to suspect that this might have an effect on the tendencies noted here, other than that they will not be readily observable in all languages.

be significantly more frequent than plural forms. His corpus counts from four corpora are given in Table 1 (from Greenberg 1966: 32).[4]

Table 1: Relative frequencies of singular, plural and dual forms in four languages

language	sample size	% singular	% plural	% dual
Sanskrit	93,277	70.3	25.1	4.6
Latin (Terence)	8,342	85.2	14.8	
Russian	8,194	77.7	22.3	
French	1,000	74.3	25.7	

These asymmetries can easily be replicated from larger modern corpora. For example, in the Russian National Corpus, there are about 60 million singular nouns and 29 million plural nouns (i.e. about 33%), and in the Eastern Armenian National Corpus, there are 33 million singular nouns and 6 million plural nouns (i.e. about 15%).

An advocate of the markedness explanation could object by saying that the frequency asymmetry is itself due to the markedness asymmetry: The reason the singular is more frequent in discourse is that it is semantically basic or unmarked (cf. Mayerthaler 1981: 136–140; Dressler et al. 2014: 187).

But this view is incompatible with the existence of singulative lexemes. The frequency explanation correctly predicts that if some nouns are different from the majority of nouns in that the multiplex form is more frequent than the uniplex form, then there should be a tendency for the multiplex form to be shorter than the uniplex form. The markedness explanation would have to claim that the plural is unmarked in these nouns, but this would be circular as long as no principled reason is given for why some nouns should have an unmarked singular, while other nouns should have an unmarked plural (cf. Mayerthaler 1981: 51–53).

4. Restating the central hypothesis

Let us now restate our central hypothesis in such a way that it is fully clear how it can be tested. We claim that the coding of uniplex/multiplex pairs of nouns tends to depend on frequency of use, in such a way that

(8) a. uniplex-prominent meanings tend to be expressed by plurative lexemes
b. multiplex-prominent meanings tend to be expressed by singulative lexemes

Recall that a uniplex-prominent meaning (e.g. ‘day’) is a noun meaning whose counterpart nouns tend to be more frequent in uniplex use, and a multiplex-prominent meaning (e.g. ‘pea’) is a noun meaning whose counterpart nouns tend to be more frequent in multiplex use.

The hypothesis in (8) is formulated from the perspective of frequency of occurrence, because uniplex prominence is defined in this way. We can alternatively formulate it from the perspective of the coding asymmetry, by defining PLURATIVE-

[4] Greenberg focused on the correlation between the frequency asymmetries and other asymmetries, not on the explanation, but in a brief passage (Greenberg 1966: 65) he says that the frequency distribution is probably primary with respect to other semantic-grammatical “markedness” phenomena.

PROMINENT MEANINGS as noun meanings that are frequently expressed by plurative lexemes, while SINGULATIVE-PROMINENT MEANINGS are noun meanings that are frequently expressed by singulative lexemes. This leads us to the formulation that

(9) a. plurative-prominent meanings tend to occur frequently in uniplex use
b. singulative-prominent meanings tend to occur frequently in multiplex use

The statements in (8) and (9) are equivalent and differ only in the perspective that is taken. It is important to be aware that the hypothesis is stated as a cross-linguistic tendency, so that no claims about particular forms or particular languages are made. The patterns can be demonstrated (or falsified) only by taking a broadly comparative perspective.

5. Expression tendencies: Singulative-prominent meanings

Let us first look at the coding of uniplex and multiplex meanings, in order to determine which kinds of noun meanings tend to be expressed as singulative lexemes. A fully rigorous method would be to look at a large and representative set of noun meanings (perhaps the 901 noun meanings of the *World Loanword Database,* Haspelmath & Tadmor 2009), at a large and representative set of languages (perhaps 50 languages from different families and regions), and to determine for each noun whether it is a plurative or a singulative noun.

Unfortunately, this method is not practical, because we lack data, and because of an additional problem: While most languages have plurative marking (though it is very often restricted and/or optional, Haspelmath 2005), few languages have singulative marking. In fact, singulative lexemes are attested in substantial numbers only in some Celtic languages, in varieties of Arabic (such as Maltese), in Cushitic languages, and in a few other languages spoken in northeastern Africa.[5]

Thus, instead of a rigorous approach, we adopt an impressionistic approach here. Tables 2 through 6 show a selection of typical singulative nouns from Welsh (Celtic), Maltese (Arabic), Arbore (Cushitic; Ethiopia), Murle (Surmic; South Sudan), and Krongo (Kadugli-Krongo; Sudan). The descriptions on which these tables are based provide a substantial number of singulative and plurative nouns for these languages, but there is no obvious way to compare these systematically. Such a systematic study is a desideratum for the future.

[5] This might at first seem surprising, but the map in Haspelmath (2005) shows that African and European languages are particularly rich in obligatory plural marking, so it is in these areas that we expect the most extensive range of nominal number-matking variation.

Table 2: Typical singulative nouns from Welsh (King 1993: 67–69; see also Stolz 2001)

fruits/vegetables	*madarch*	*maderch-en*	mushrooms
	mwyar	*mwyar-en*	blackberries
	ffa	*ffä-en*	beans
	bresych	*bresych-en*	cauliflower
small animals	*cacwn*	*cacyn-en*	wasps
	clêr	*cler-en*	flies
	hwyaid	*hwyad-en*	ducks
	llygod	*llygod-en*	mice
groups of people	*plant*	*plent-yn*	children
other	*sêr*	*ser-en*	stars
	dillad	*dilled-yn*	clothes
	plu	*plu-en*	feathers[6]

Table 3: Typical singulative nouns from Maltese (Arabic; Mifsud 1996)

paired body-parts	*zarbun*	*zarbun-a*	shoes
	buz	*buz-a*	boots
fruits/vegetables	*amħ*	*amħ-a*	corn
	lewz	*lewz-a*	almonds
	tuffieħ	*tuffieħ-a*	apples
small animals	*dubbien*	*dubbien-a*	flies
	gawwi	*gawwi-a*	swallows
	wizz	*wizz-a*	geese
other	*taraġ*	*taraġ-a*	stairs
	ravyul	*ravyul-a*	ravioli

Table 4: Typical singulative nouns from Arbore (Cushitic; Hayward 1984: 179–183)

paired body-parts	*farró*	*farri-t*	fingers
	ʔedanó	*ʔedan-té*	testicles
	soonó	*soonon̄-té*	nose/nostrils
	moyḍé	*moyḍe-n̄té*	eyebrows
fruits/vegetables	*sáj*	*sayyi-t*	grass
small animals	*kón̄čo*	*kon̄čo-t*	water-snails
	ʔín̄do	*ʔín̄do-t*	grubs
	keḍéy	*keḍe-té*	bees
groups of people	*hamár*	*hamar-tat*	Hamar (ethn.)
other	*húzzuḳ*	*húzzuḳ-an̄té*	stars
	san̄dóy	*san̄doy-té*	graves

[6] For Welsh, Dressler et al. (2014: 187) note that the singulative suffix *-en* is a derivational suffix, and that the form *pluen* can be inflectionally pluralized with the productive suffix *-au* (*pluenn-au*). Similar additional forms are found in some of the other languages, but they are not relevant to the main point that we are making here, which is that pairs such as *plu/pluen* are found in specific semantic classes, and that they correlate with universal frequency asymmetries. Whether the pairs are inflectional or derivational is immaterial (even though our terminology in (8) suggests thinking of them in inflectional terms). We take this as a virtue of our approach, because it is often impossible to tell whether a pattern is inflectional or derivational.

Table 5: Typical singulative nouns from Murle (Surmic; Arensen 1982: 40–44)

paired body-parts	*kɛbɛrɛ*	*kebere-c*	eyes
	zɔɔ	*zoo-c*	feet
	oto	*oto-n*	horns
fruits/vegetables	*ŋadɛɛra*	*ŋadɛɛra-c*	onions
	ŋooru	*ŋooru-woc*	beans
	mɔtɔɔŋ	*motooŋ-toc*	tamarind fruits
small animals	*aguna*	*aguna-c*	black ants
	yɛɛla	*yɛɛla-c*	doves
	kel	*kel-oc*	fleas
groups of people	*codɛ*	*codɛ-n*	twins
	dɔl	*dol-e*	babies
	rotti	*rotti-n*	warriors
other	*lɛtɛ*	*lete-c*	honey
	maam	*maam-oc*	water

Table 6: Typical singulative nouns from Krongo (Kadugli; Reh 1985: 101–126)

paired body-parts	*àaw*	*ǹtìn-àaw*	hair(s)
	íitò	*tìn-íitò*	horns
	màsállíŋ	*tì-màsállíŋ*	ankles
fruits/vegetables	*fólóttó*	*tì-fólóttó*	pods
	tʊ̀lìŋ	*ǹ-tʊ̀lìŋ*	leaves
small animals	*àafúŋ*	*ǹtìn-àafúŋ*	ants
	àasà	*ǹtìn-àasà*	flies
	kwóoyá	*mòtó-kwóoyá*	snails
	òlló	*f-òlló*	wasps
groups of people	*ókkótú*	*b-ókkótú*	twins
other	*màkàaràŋ*	*tì-màkàaràŋ*	clouds
	súlì	*tù-súlì*	eggs
	kwáalà	*mùtú-kwàalà*	dippers

See also Grimm (2012) on Dagaare (Gur).

A clear trend that emerges from these data is that the following semantic classes of nouns tend to be expressed as singulative lexemes:

(10) a. paired body-parts
b. fruits/vegetables
c. small animals that occur in groups
d. groups of people

Not all of these groups are represented in all the languages, but they recur in a way that cannot be accidental. Following (9b), we now need to check whether these kinds of noun meanings do indeed tend to occur frequently in multiplex form.

6. Usage tendencies: The corpus data

In order to check whether it is cross-linguistically the case that singulative-prominent meanings (the meanings in (10a-d)) are highly frequent in multiplex form, we examined large corpora from five languages (English, Estonian, Latvian, Norwegian, Russian). We analyzed the frequencies of 18 lexemes in each language: three lexemes from six (subjective) semantic classes with potentially singulative-prominent meanings, as observed in Section 5. The labels of the classes of concepts are intended to be no more than descriptive.

(11) 18 singulative-prominent noun meanings for our corpus study

paired body-parts	ear, leg, lung
paired items:	glove, shoe, ski
fruits/vegetables:	apple, potato, strawberry
small animals:	bee, pigeon, sheep
people:	child, boy, girl
ethnic groups:	European, American, speaker of (the resp. language)

In addition, we looked at 18 random lexemes in each language (90 in total), hypothesizing that the random lexemes would in general not show the specific usage tendencies as the 18 nouns with the meanings in (11). We sampled the random sets from word lists of nouns of moderately high corpus frequency, in order to avoid behavioral bias from extremely frequent or very rare words. We expect the sets of random nouns to represent the average noun usage in the respective languages (as such, we did not attempt to filter the random sets for potentially multiplex-prominent words). The 18+18 nouns in each of the five languages are given in the Appendix.

The analysis was based on data from written language corpora (mostly media and literature; see the list of references for more details):

(12)	English (British)	BNC (*British National Corpus of English*)
	Estonian	EKK (*Eesti kirjakeele korpus* = Estonian Reference Corpus)
	Latvian	MLVTK (*Mūsdienu latviešu valodas tekstu korpuss* = Modern Latvian Text Corpus)
	Norwegian (Bokmål)	OK (*Oslo-korpuset av taggede norske tekster* = Oslo Corpus of tagged Norwegian Texts)
	Russian	RNC (*Nacional'nyj korpus russkogo jazyka* = = Russian National Corpus)

The choice of languages was motivated by (i) the fact that for each of them, sufficiently large corpora are freely available; (ii) we are at least somewhat familiar with the languages and as such, able to critically evaluate the corpus search results, and (iii) none of the languages can be said to have overt morphological singular marking. All of the involved corpora are already automatically morphologically tagged. Of course, automatic tagging is by no means flawless, which warranted manual counting and filtering in some cases (more on that below). Sub-corpora of texts written no earlier than 1990 were sampled from each corpus to avoid diachronic variation.

Some simplifications were necessary to allow for cross-linguistic comparisons. Only the singulars and plurals of nominative case forms were taken into account (in

Norwegian, only nominative indefinite forms). Latvian and Russian have parallel ethnic terms for the two genders; only the masculine forms in the class “ethnic terms” were considered (a similar distinction is possible in Estonian, but its usage is marginal). In Norwegian, the indefinite singular and plural form is homonymous in ‘ski’, ‘shoe’, ‘child’ and ‘strawberry’, the same holds for the English *sheep*. In the BNC, the proper noun *Apple* is mostly tagged as a common noun, inflating the counts of singular for that concept. To calculate the asymmetry indices for such problematic words, we used small subsamples (40 occurrences each) and manually tagged them for grammatical number based on the context. Noun-noun compounds are very common in English, and the automatically tagged BNC seldom distinguishes multi-word compounds. This leads to inflated counts for the singular forms of nouns which are actually modifiers of the second part of the compound (e.g., searching for *ski* or *strawberry* also yield large volumes of *ski resort* and *strawberry jam*). To avoid such inflation, only nouns not followed by another noun were counted in English. For Russian, a smaller, manually disambiguated subcorpus of the RNC was used to count the forms of the concept ‘speakers of (the respective language)’, as the word is homonymous. The corpus frequency results were furthermore selectively manually checked in an attempt to detect inflated counts caused by homonymy. Naturally, frequent usage in fixed phrases (*I’m all ears; bad apple*, etc.) influence the counts of the involved nouns, as does availability to be used as a mass noun (the probable reason that makes the Estonian ‘potato’ somewhat of an outlier, for example). However, we hope that sufficiently large samples alleviate these problems somewhat, letting the stronger tendencies shine through.

The difference between the counts of singular and plural forms of the nouns was normalized as an “asymmetry index” with a range of -1…1, where negative values indicate dominant singular usage, and positive values dominant plural usage. A ‘0’ means that the counts were equal, and a value of -0.5 or 0.5 means that one of the forms forms occurred twice as often as the other. To put it another way, the value corresponds to (13):

(13) $|x - y| / \max(x, y)$; if $\max(x, y)$ = count of singulars, multiply the result by -1

Statistical significance of the difference in the singular and plural form counts for each individual noun was tested by calculating the cumulative binomial probability for the distributions. The index value for non-significant distributions ($\alpha = 0.05$) was automatically coded as ‘0’, indicating equal distribution. This method would weed out both meaningless differences in small counts and small differences in similar large counts. For example, a distribution of 4 against 2 occurrences would yield a value of 0.5; equally well, a distribution of 90 against 83 is likely to be just chance. However, the samples were mostly fairly large (cf. Table 7), and the majority of the distributions were significantly different.

Table 7: A summary of the sampled corpora used in this study

Language	Corpus	sub-corpus size	mean per million counts (sg+pl)	mean raw counts (sg+pl)	significantly different sg/pl distributions	mean index (random sampled nouns)	mean index (preselected nouns)
English	BNC	73M	50	3479	97%	-0.45	0.44
Estonian	EKK	217M	52	11277	97%	-0.59	0.35
Latvian	MLVTK	4M	69	302	88%	-0.36	0.42
Norwegian	OK	11M	34	323	86%	-0.48	0.56
Russian	NCRL	48M	46	2139	94%	0.03	0.54

7. Results

It is clear from Table 7 that the randomly sampled nouns, on average, tend to occur more in the singular, compared to the nouns representing the predetermined concepts, which occur more in the plural. The detailed distributions of the concepts may be observed below (Figure 1).

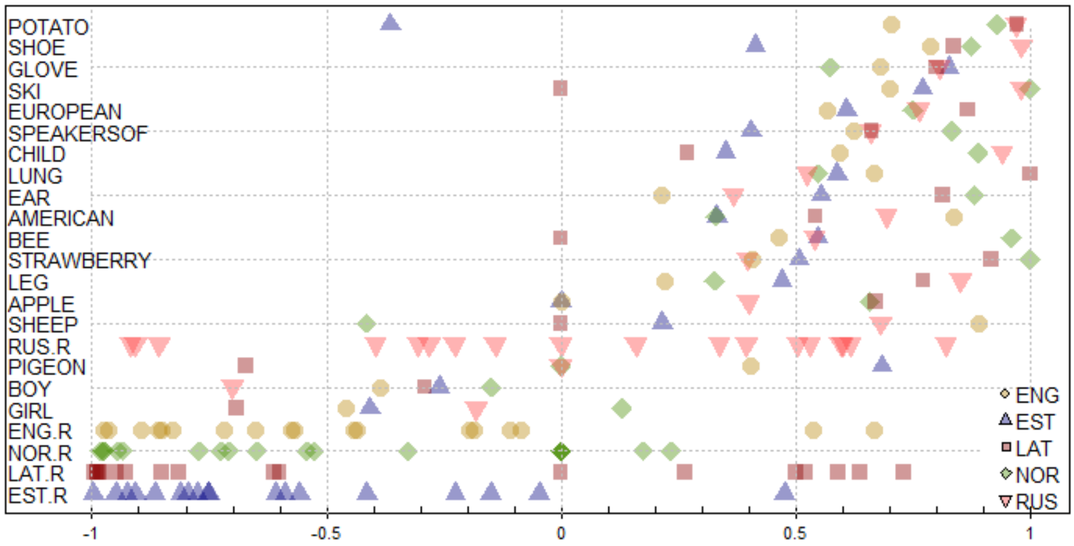

Figure 1: The sample of 180 nouns – 18 preselected nouns and 18 randomly sampled nouns from 5 languages – arranged along the vertical axis by the median asymmetry index value of the concepts. 'R.' marks the random groups. The horizontal axis represents the number asymmetry index, discussed above, so the uniplex-prominent nouns lean to the left, and the multiplex-prominent nouns to the right side of the plot. (Details are given in the Appendix.)

We tested the correlation of the asymmetry index with the semantic concepts using a linear regression model, with the asymmetry index as the response variable and the concept as a factorial predictor. A significant difference in the asymmetry index between the group of random nouns (the intercept of the model) and the rest of the concepts appeared (all concepts: $p < 0.05$, with two exceptions). The model as a whole was found to be significant ($F_{18,\ 161} = 10.4$, $p < 0.001$), with moderately high explanatory power (adjusted $R^2 = 0.48$).[7] The random nouns as a group lean towards

[7] Various model diagnostics (Cook's distance, DFBETA and DFFITS influence statistics, normality of the distribution of residuals, residuals against fitted values, and Levene's test of equality of error variances (Cook 1977; Belsley et al. 1980; Levene 1960) were used to test the validity of the model.

the singular or at least equal distribution in number. The preselected nouns, as hypothesized, gravitate towards the plural, with the two exceptions of the concepts of 'boy' and 'girl', which do not behave significantly differently from the random nouns (i.e., occur more in the singular; but note that 'child' on the other hand is multiplex-prominent in all five languages). The model confirms that the observation that may already be drawn from Figure 1 – that nouns representing the predetermined concepts occur more in the plural, compared to the "general population" of random nouns – is indeed highly likely not due to chance.

In other words, the nouns that belong to such semantic classes which tend to be expressed by nouns with overt singulars (in languages with singulatives) strongly tend to be more frequent in the plural than in the singular, compared to randomly sampled nouns.

8. The explanation

As was made clear earlier, we claim that the tendency for singulative lexemes to be multiplex-prominent (and for plurative lexemes to be uniplex-prominent) is due to a highly general principle of grammatical coding, the grammatical form-frequency correspondence principle (in (3) above), which has a well-known explanation in terms of coding efficiency (Zipf 1935; Fenk-Oczlon 1991; Hawkins 2004; Haspelmath 2008). What is new here is that we apply this principle to singulative and plurative lexemes.

As was noted in Haspelmath et al. (2014) and elsewhere, the correspondence between form and frequency is implemented by diachronic mechanisms which tend to make frequent forms short, because frequent forms are more predictable than rare forms. Ultimately, it is thus predictability that lies at the root of the length difference and the coding asymmetry.

Thus, the causal effect is very indirect (cf. Newmeyer 2014): We cannot say that the *Tag/Tag-e* pattern in Modern German is due to the fact that *Tag* is more frequent than *Tage* in Modern German, and we cannot say that the Welsh *pys-en/pys* pattern is due to the fact that *pys* is more frequent than *pysen* in Welsh.

The causal effect is relatively weak, so it cannot be seen in all languages (many languages lack form distinctions between uniplex and multiplex nouns), and especially the tendency for multiplex-prominent nouns to occur as singulatives is manifested only very rarely. (In most languages, all lexemes join the majority pattern, due to system pressure, cf. Haspelmath 2014.) The explanatory mode can thus be summarized as in (14).

(14) Universal frequency asymmetries (resulting in predictability asymmetries) explain universal form asymmetries, via universal diachronic tendencies.

The latter two indicated a heteroscedasticity problem – owing to the strong uniplex-preference of many of the random nouns – but it was not seen as a severe hindrance for the analysis (and a model using generalized least squares yielded essentially the same results). We also tested a mixed-effects model with different intercepts for the languages, which turned out equally significant, so the simpler model is published here. The models and diagnostics were run and the plot created using functions available in R (version 3.2.2; R Core Team 2015).

It is in this way that corpus data from Norwegian or Russian (or any other language) can be used to explain morphological asymmetries in Maltese and Arbore (or any other language with relevant asymmetries). This presupposes, of course, that frequency distributions are about the same in all languages, i.e. that Maltese or Arbore speakers show roughly the same usage patterns in their speech. While there are of course many cultural differences in language use, and there might be some in this area as well, we are not aware of any suggestions that point in this direction, so we feel that the presupposition is not problematic.

Before concluding, let us briefly address a critical question that readers might have: Couldn't it be that singulative lexemes are conceptualized differently in languages with singulative marking, as "less individualized", or "collective", or "masses"? Linguists have traditionally tended to favour meaning-based explanations over usage-based explanations of grammatical form (cf. Grimm 2012, who tries to explain singulative marking in Dagaare in this way).

Our answer is that we cannot rule out that such meaning differences exist, and if they exist, the semantic explanation would not be incompatible with our usage-based explanation. A meaning-based explanation would have to provide clear criteria for identifying conceptualizations independently of grammatical form (along the lines of Gil's (1996) exemplary discussion). It may well turn out that in some of the languages mentioned above, the basic forms that denote multiplex items (e.g. Maltese *larinġ* 'oranges') are semantically somewhat different from normal (i.e. pluratively coded) plurals (along the lines of the semantic differences between the English singular mass noun *hair* and the plural of the related count noun *hairs*). However, this is a matter for future research.

Whatever the outcome of such studies, a semantic explanation would be compatible with our frequency-based explanation. Note that we define our comparative concept *multiplex* not in terms of 'plural meaning', but in terms of "possible conceptualization". This means that mass nouns such as *sand* can be regarded as multiplex nouns as well. This would fit well with our overall claims, because the corresponding uniplex expression (*grain of sand)* has more formal coding, so one could say that the expression pair *grain of sand / sand* is a kind of singulative/basic pair, like those in (7), with the only difference that the singulative marker is not a grammatical affix, but a noun. The usage-based explanation in terms of frequency of use is thus actually independent of the mass vs. plural meanings of the nouns in question. Crucially in the present context, the semantic explanation does not make the usage-frequency explanation superfluous, because we also want to know which kinds of entities tend to be conceived of as masses. It would seem that it is precisely those that often occur in a multiplex sense, but this is a topic for future research.

Appendix: Frequencies of the 18+18 nouns in each of the five languages
(Estonian, Norwegian, Latvian, Russian, English; pmw = per million words)

lang	type	concept	sg form	pl form	sg count	pl count	total (sg+pl) pmw	asymmetry index
EST	bodyparts	LEG	*jalg*	*jalad*	4528	8561	60.34	0.471
EST	bodyparts	LUNG	*kops*	*kopsud*	179	435	2.83	0.589
EST	bodyparts	EAR	*kõrv*	*kõrvad*	719	1609	10.73	0.553
EST	pairitems	SHOE	*king*	*kingad*	1052	1793	13.12	0.413
EST	pairitems	GLOVE	*kinnas*	*kindad*	135	782	4.23	0.827
EST	pairitems	SKI	*suusk*	*suusad*	508	2212	12.54	0.770
EST	flockanimals	PIGEON	*tuvi*	*tuvid*	73	231	1.40	0.684
EST	flockanimals	BEE	*mesilane*	*mesilased*	218	482	3.23	0.548
EST	flockanimals	SHEEP	*lammas*	*lambad*	740	943	7.76	0.215
EST	ethnic	SPEAKERSOF	*eestlane*	*eestlased*	12263	20607	151.53	0.405
EST	ethnic	AMERICAN	*ameeriklane*	*ameeriklased*	4426	6629	50.96	0.332
EST	ethnic	EUROPEAN	*eurooplane*	*eurooplased*	583	1487	9.54	0.608
EST	children	CHILD	*laps*	*lapsed*	26566	40903	311.03	0.351
EST	children	BOY	*poiss*	*poisid*	17530	12967	140.59	-0.260
EST	children	GIRL	*tüdruk*	*tüdrukud*	13800	8173	101.29	-0.408
EST	fruits	STRAWBERRY	*maasikas*	*maasikad*	298	604	4.16	0.507
EST	fruits	APPLE	*õun*	*õunad*	848	829	7.73	0.000
EST	fruits	POTATO	*kartul*	*kartulid*	1517	964	11.44	-0.365
EST	.random	.R	*erakond*	*erakonnad*	5703	4847	48.63	-0.150
EST	.random	.R	*mark*	*margid*	1177	484	7.66	-0.589
EST	.random	.R	*teadmine*	*teadmised*	3997	3097	32.70	-0.225
EST	.random	.R	*teater*	*teatrid*	8236	764	41.49	-0.907
EST	.random	.R	*toime*	*toimed*	1760	5	8.14	-0.997
EST	.random	.R	*lootus*	*lootused*	6300	2789	41.90	-0.557
EST	.random	.R	*nägu*	*näod*	7918	1782	44.72	-0.775
EST	.random	.R	*klubi*	*klubid*	10107	2493	58.09	-0.753
EST	.random	.R	*järv*	*järved*	1719	351	9.54	-0.796
EST	.random	.R	*vend*	*vennad*	8360	4896	61.11	-0.414
EST	.random	.R	*põhjus*	*põhjused*	17651	4400	101.65	-0.751
EST	.random	.R	*värav*	*väravad*	3403	3245	30.65	-0.046
EST	.random	.R	*töötaja*	*töötajad*	7563	14444	101.45	0.476
EST	.random	.R	*kool*	*koolid*	11913	4660	76.40	-0.609
EST	.random	.R	*jumal*	*jumalad*	10107	509	48.94	-0.950
EST	.random	.R	*treener*	*treenerid*	21834	2915	114.09	-0.866
EST	.random	.R	*süsteem*	*süsteemid*	11833	891	58.66	-0.925
EST	.random	.R	*idee*	*ideed*	14847	2789	81.30	-0.812
NOR	bodyparts	LEG	*fot*	*føtter*	103	153	22.65	0.327
NOR	bodyparts	LUNG	*lunge*	*lunger*	14	31	3.98	0.548
NOR	bodyparts	EAR	*øre*	*ører*	13	110	10.88	0.882
NOR	pairitems	SHOE	*sko*	*sko*	5	40	NA	0.875
NOR	pairitems	GLOVE	*hanske*	*hansker*	20	47	5.93	0.574
NOR	pairitems	SKI	*ski*	*ski*	0	45	NA	1.000
NOR	flockanimals	PIGEON	*due*	*duer*	25	36	5.40	0.000
NOR	flockanimals	BEE	*bie*	*bier*	1	25	2.30	0.960
NOR	flockanimals	SHEEP	*sau*	*sauer*	256	150	35.93	-0.414
NOR	ethnic	SPEAKERSOF	*nordmann*	*nordmenn*	202	1208	124.78	0.833
NOR	ethnic	AMERICAN	*amerikaner*	*amerikanere*	53	79	11.68	0.329
NOR	ethnic	EUROPEAN	*europeer*	*europeere*	11	44	4.87	0.750
NOR	children	CHILD	*barn*	*barn*	5	45	NA	0.889
NOR	children	BOY	*gutt*	*gutter*	763	649	124.96	-0.149
NOR	children	GIRL	*jente*	*jenter*	748	859	142.21	0.129
NOR	fruits	STRAWBERRY	*jordbær*	*jordbær*	0	50	NA	1.000
NOR	fruits	APPLE	*eple*	*epler*	41	120	14.25	0.658
NOR	fruits	POTATO	*potet*	*poteter*	25	351	33.27	0.929

NOR	.random	.R	*angriper*	*angripere*	25	30	5.73	0.000
NOR	.random	.R	*base*	*baser*	118	27	15.10	-0.771
NOR	.random	.R	*belastning*	*belastninger*	142	39	18.85	-0.725
NOR	.random	.R	*garanti*	*garantier*	188	86	28.54	-0.543
NOR	.random	.R	*statssekretær*	*statssekretærer*	256	17	28.44	-0.934
NOR	.random	.R	*stykke*	*stykker*	423	123	56.88	-0.709
NOR	.random	.R	*vik*	*viker*	19	9	2.92	-0.526
NOR	.random	.R	*vitne*	*vitner*	263	319	60.63	0.176
NOR	.random	.R	*lunsj*	*lunsjer*	77	2	8.23	-0.974
NOR	.random	.R	*banker*	*bankere*	77	2	8.23	-0.974
NOR	.random	.R	*artikkel*	*artikler*	235	202	45.52	0.000
NOR	.random	.R	*demonstrasjon*	*demonstrasjoner*	98	82	18.75	0.000
NOR	.random	.R	*offer*	*ofre*	217	283	52.08	0.233
NOR	.random	.R	*virkelighet*	*virkeligheter*	451	8	47.81	-0.982
NOR	.random	.R	*tegning*	*tegninger*	245	264	53.02	0.000
NOR	.random	.R	*storstue*	*storstuer*	73	4	8.02	-0.945
NOR	.random	.R	*søknad*	*søknader*	239	161	41.67	-0.326
NOR	.random	.R	*roman*	*romaner*	376	132	52.92	-0.649
RUS	bodyparts	LEG	*нога*	*ноги*	516	3469	83.32	0.851
RUS	bodyparts	LUNG	*лёгкое*	*лёгкие*	339	711	21.95	0.523
RUS	bodyparts	EAR	*ухо*	*уши*	527	833	28.44	0.367
RUS	pairitems	SHOE	*туфля*	*туфли*	6	323	6.88	0.981
RUS	pairitems	GLOVE	*рукавица*	*рукавицы*	11	57	1.42	0.807
RUS	pairitems	SKI	*лыжа*	*лыжи*	6	297	6.34	0.980
RUS	flockanimals	PIGEON	*голубь*	*голуби*	90	109	4.16	0.000
RUS	flockanimals	BEE	*пчела*	*пчелы*	51	111	3.39	0.541
RUS	flockanimals	SHEEP	*овца*	*овцы*	64	201	5.54	0.682
RUS	ethnic	SPEAKERSOF	*русский*	*русские*	17	50	43.28	0.660
RUS	ethnic	AMERICAN	*англичанин*	*американцы*	320	1046	28.56	0.694
RUS	ethnic	EUROPEAN	*европеец*	*европейцы*	77	324	8.38	0.762
RUS	children	CHILD	*дитя*	*дети*	398	6698	148.37	0.941
RUS	children	BOY	*мальчик*	*мальчики*	1623	486	44.10	-0.701
RUS	children	GIRL	*девочка*	*девочки*	1414	1158	53.78	-0.181
RUS	fruits	STRAWBERRY	*земляника*	*земляники*	53	88	2.95	0.398
RUS	fruits	APPLE	*яблоко*	*яблоки*	224	373	12.48	0.399
RUS	fruits	POTATO	*картофелина*	*картофель*	6	199	4.29	0.970
RUS	.random	.R	*видимость*	*видимости*	299	740	21.72	0.596
RUS	.random	.R	*вор*	*воры*	265	205	9.83	-0.226
RUS	.random	.R	*скатерть*	*скатерти*	93	80	3.62	0.000
RUS	.random	.R	*казарма*	*казармы*	29	162	3.99	0.821
RUS	.random	.R	*набор*	*наборы*	1879	154	42.51	-0.918
RUS	.random	.R	*пауза*	*паузы*	459	319	16.27	-0.305
RUS	.random	.R	*паспорт*	*паспорта*	826	500	27.73	-0.395
RUS	.random	.R	*ребро*	*рёбра*	74	193	5.58	0.617
RUS	.random	.R	*общество*	*общества*	4561	9189	287.50	0.504
RUS	.random	.R	*достоинство*	*достоинства*	810	1221	42.47	0.337
RUS	.random	.R	*действие*	*действия*	3860	8235	252.90	0.531
RUS	.random	.R	*период*	*периоды*	9462	879	216.22	-0.907
RUS	.random	.R	*след*	*следы*	806	959	36.90	0.160
RUS	.random	.R	*сомнение*	*сомнения*	738	1844	53.99	0.600
RUS	.random	.R	*понимание*	*понимания*	1858	1597	72.24	-0.140
RUS	.random	.R	*князь*	*князья*	1024	146	24.46	-0.857
RUS	.random	.R	*клиент*	*клиенты*	654	470	23.50	-0.281
RUS	.random	.R	*дверца*	*дверцы*	52	86	2.89	0.395
LAT	bodyparts	EAR	*auss*	*ausis*	18	97	26.08	0.814
LAT	bodyparts	LUNG	*plauša*	*plaušas*	0	28	6.35	1.000
LAT	bodyparts	LEG	*kāja*	*kājas*	60	263	73.26	0.772
LAT	pairitems	SKI	*slēpe*	*slēpes*	0	1	0.23	0.000
LAT	pairitems	GLOVE	*cimds*	*cimdi*	4	20	5.44	0.800
LAT	pairitems	SHOE	*kurpe*	*kurpe*	8	49	12.93	0.837

LAT	flockanimals	BEE	*bite*	*bites*	9	7	3.63	0.000
LAT	flockanimals	SHEEP	*aita*	*aitas*	4	11	3.40	0.000
LAT	flockanimals	PIGEON	*balodis*	*baloži*	73	24	22.00	-0.671
LAT	ethnic	AMERICAN	*amerikānis*	*amerikāņi*	32	70	23.13	0.543
LAT	ethnic	EUROPEAN	*eiropietis*	*eiropieši*	2	15	3.86	0.867
LAT	ethnic	SPEAKERSOF	*latvietis*	*latvieši*	80	237	71.90	0.662
LAT	children	GIRL	*meitene*	*meitenes*	655	202	194.38	-0.692
LAT	children	BOY	*zēns*	*zēni*	186	132	72.13	-0.290
LAT	children	CHILD	*bērns*	*bērni*	1056	1446	567.48	0.270
LAT	fruits	STRAWBERRY	*zemene*	*zemenes*	4	48	11.79	0.917
LAT	fruits	APPLE	*ābols*	*āboli*	33	100	30.17	0.670
LAT	fruits	POTATO	*kartupelis*	*kartupeļi*	4	142	33.11	0.972
LAT	.random	.R	*izstāde*	*izstāde*	203	4	46.95	-0.980
LAT	.random	.R	*prasība*	*prasība*	145	536	154.46	0.729
LAT	.random	.R	*vieta*	*vieta*	952	16	219.55	-0.983
LAT	.random	.R	*zeme*	*zeme*	359	1	81.65	-0.997
LAT	.random	.R	*gods*	*gods*	89	1	20.41	-0.989
LAT	.random	.R	*krusts*	*krusts*	47	7	12.25	-0.851
LAT	.random	.R	*kaimiņš*	*kaimiņš*	44	121	37.42	0.636
LAT	.random	.R	*stāsts*	*stāsts*	327	127	102.97	-0.612
LAT	.random	.R	*mirklis*	*mirklis*	55	22	17.46	-0.600
LAT	.random	.R	*līnija*	*līnija*	95	1	21.77	-0.989
LAT	.random	.R	*zieds*	*ziedi*	51	102	34.70	0.500
LAT	.random	.R	*koris*	*kori*	81	15	21.77	-0.815
LAT	.random	.R	*dievs*	*dievi*	574	41	139.49	-0.929
LAT	.random	.R	*pacients*	*pacienti*	154	209	82.33	0.263
LAT	.random	.R	*speciālists*	*speciālisti*	190	465	148.56	0.591
LAT	.random	.R	*priekšmets*	*priekšmeti*	121	101	50.35	0.000
LAT	.random	.R	*pakalpojums*	*pakalpojumi*	91	190	63.73	0.521
LAT	.random	.R	*pieprasījums*	*pieprasījumi*	212	11	50.58	-0.948
ENG	bodyparts	LEG	*leg*	*legs*	3257	4174	102.27	0.220
ENG	bodyparts	LUNG	*lung*	*lungs*	189	568	10.42	0.667
ENG	bodyparts	EAR	*ear*	*ears*	1499	1905	46.85	0.213
ENG	pairitems	SHOE	*shoe*	*shoes*	514	2407	40.20	0.786
ENG	pairitems	GLOVE	*glove*	*gloves*	212	660	12.00	0.679
ENG	pairitems	SKI	*ski*	*skis*	52	173	3.10	0.699
ENG	flockanimals	PIGEON	*pigeon*	*pigeons*	165	277	6.08	0.404
ENG	flockanimals	BEE	*bee*	*bees*	251	468	9.89	0.464
ENG	flockanimals	SHEEP	*sheep*	*sheep*	4	36	NA	0.889
ENG	ethnic	SPEAKERSOF	*brit*	*brits*	59	157	2.97	0.624
ENG	ethnic	AMERICAN	*american*	*americans*	309	1887	30.22	0.836
ENG	ethnic	EUROPEAN	*european*	*europeans*	212	491	9.67	0.568
ENG	children	CHILD	*child*	*children*	11475	28253	546.74	0.594
ENG	children	BOY	*boy*	*boys*	7866	4845	174.93	-0.384
ENG	children	GIRL	*girl*	*girls*	9467	5122	200.78	-0.459
ENG	fruits	STRAWBERRY	*strawberry*	*strawberries*	106	179	3.92	0.408
ENG	fruits	APPLE	*apple*	*apples*	18	22	NA	0.000
ENG	fruits	POTATO	*potato*	*potatoes*	294	995	17.74	0.705
ENG	.random	.R	*clearing*	*clearings*	176	30	2.83	-0.830
ENG	.random	.R	*claim*	*claims*	3448	3070	89.70	-0.110
ENG	.random	.R	*headline*	*headlines*	274	592	11.92	0.537
ENG	.random	.R	*representation*	*representations*	2338	1010	46.08	-0.568
ENG	.random	.R	*background*	*backgrounds*	3164	465	49.94	-0.853
ENG	.random	.R	*recorder*	*recorders*	552	193	10.25	-0.650
ENG	.random	.R	*tablet*	*tablets*	222	670	12.28	0.669
ENG	.random	.R	*primary*	*primaries*	216	122	4.65	-0.435
ENG	.random	.R	*batch*	*batches*	372	105	6.56	-0.718
ENG	.random	.R	*noun*	*nouns*	230	128	4.93	-0.443
ENG	.random	.R	*partner*	*partners*	2947	2698	77.69	-0.084
ENG	.random	.R	*governor*	*governors*	1495	1203	37.13	-0.195

ENG	.random	.R	*mistake*	*mistakes*	2490	1056	48.80	-0.576
ENG	.random	.R	*opening*	*openings*	1814	260	28.54	-0.857
ENG	.random	.R	*reconstruction*	*reconstructions*	629	67	9.58	-0.893
ENG	.random	.R	*approval*	*approvals*	2760	95	39.29	-0.966
ENG	.random	.R	*bail*	*bails*	432	11	6.10	-0.975
ENG	.random	.R	*slope*	*slopes*	743	604	18.54	-0.187

References

Arensen, Jonathan E. 1982. *Murle grammar*. (Occasional Papers in the Study of Sudanese Languages, 2). Juba: University of Juba.

Belsley, David A., Edwin Kuh & Roy E. Welsh. 1980. *Regression diagnostics: Identifying influential data and sources of collinearity.* New York: John Wiley & Sons.

Bybee, Joan. 2007. *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Cook, R. Dennis. 1977. Detection of influential observations in linear regression. *Technometrics* 19(1). 15–18.

Croft, William. 2003. *Typology and universals*. 2nd edition. Cambridge: Cambridge University Press.

Cuzzolin, Pierluigi. 1998. Sull'origine del singolativo in celtico, con particolare riferimento al medio gallese. *Archivio Glottologico Italiano* 83(2). 121–149.

Dressler, Wolfgang & Libben, Gary & Korecky-Kröll, Katharina. 2014. Conflicting vs. convergent vs. interdependent motivations in morphology. In: MacWhinney, Brian & Malchukov, Andrej & Moravcsik, Edith (eds.) *Competing motivations in grammar and usage*, 181-196. Oxford: Oxford University Press.

Fenk-Oczlon, Gertraud. 1991. Frequenz und Kognition–Frequenz und Markiertheit. *Folia Linguistica* 25(3–4). 361–394.

Field, A. 2005. *Discovering statistics using SPSS.* London: Sage Publications.

Fox, John. 1997. *Applied regression analysis, linear models, and related methods.* SAGE Publications.

Gil, David. 1996. Maltese “collective nouns”: A typological perspective. *Rivista di Linguistica* 8(1). 53–87.

Givón, T. 1991. Markedness in grammar: Distributional, communicative and cognitive correlates of syntactic structure. *Studies in Language* 15(2). 335-370.

Greenberg, Joseph H. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg (ed.), *Universals of language*, 73–113. Cambridge, MA: MIT Press.

Greenberg, Joseph H. 1966. *Language universals, with special reference to feature hierarchies*. The Hague: Mouton.

Grimm, Scott. 2012. Individuation and inverse number marking in Dagaare. In Diane Massam (ed.), *Count and mass across languages*, 75–98. Oxford: Oxford University Press.

Haspelmath, Martin. 2005. Occurrence of nominal plurality. In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.), *The world atlas of language structures*, 142–145. Oxford: Oxford University Press.

Haspelmath, Martin. 2006. Against markedness (and what to replace it with). *Journal of Linguistics* 42(1). 25–70.

Haspelmath, Martin. 2008. Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics* 19(1). 1–33.

Haspelmath, Martin. 2014. On system pressure competing with economic motivation. In Brian MacWhinney, Andrej L. Malchukov, & Edith A. Moravcsik (eds.), *Competing motivations in grammar and usage*, 197–208. Oxford: Oxford University Press.

Haspelmath, Martin & Uri Tadmor (eds.). 2009. *World Loanword Database.* Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://wold.clld.org/>

Haspelmath, Martin, Andreea Calude, Michael Spagnol, Heiko Narrog & Elif Bamyacı. 2014. Coding causal–noncausal verb alternations: A form–frequency correspondence explanation. *Journal of Linguistics* 50(3). 587–625.

Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Hayward, Dick. 1984. *The Arbore language: A first investigation (including a vocabulary)*. Hamburg: Helmut Buske Verlag.

King, Gareth. 1993. *Modern Welsh: A comprehensive grammar*. London: Routledge.

Kirt, Riin. 2013. *Tasakaalus korpusel põhinevad sagedusloendid ja korpuse sõnavara ning "Eesti keele seletava sõnaraamatu" märksõnaloendi võrdlus.* [Word frequency lists based on the "Balanced Corpus of Estonian" and selective comparison of corpora frequency lists with keywords from the „Explanatory Dictionary of Estonian“]. Tartu: University of Tartu MA thesis.

Levene, Howard. 1960. Robust tests for equality of variances. In Ingram Olkin, Harold Hotelling, et al (eds), *Contributions to probability and statistics: Essays in honor of Harold Hotelling*, 278–292. Stanford: Stanford University Press.

Mayerthaler, Willi. 1981. *Morphologische Natürlichkeit.* Wiesbaden: Akademische Verlagsgesellschaft Athenaion.

Mifsud, Manwel. 1996. The collective in Maltese. *Rivista di Linguistica* 8(1). 29–51.

Newmeyer, Frederick J. 2014. Where do motivations compete? In: MacWhinney, Brian & Malchukov, Andrej & Moravcsik, Edith (eds.) *Competing motivations in grammar and usage*, 299–314. Oxford: Oxford University Press.

Quinn, G. P. and Keough, M. J. 2002. *Experimental designs and data analysis for biologists.* Cambridge: Cambridge University Press.

R Core Team 2015. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Reh, Mechthild. 1985. *Die Krongo-Sprache (nìino mó-dì): Beschreibung, Texte, Wörterverzeichnis*. (Kölner Beiträge zur Afrikanistik, 12). Berlin: Dietrich Reimer.

Stolz, Thomas. 2001. Singulative-collective: Natural Morphology and stable classes in Welsh number inflexion on nouns. *Sprachtypologie und Universalienforschung* 54(1). 52–76.

Talmy, Leonard. 1988. The relation of grammar to cognition: A synopsis. In Brygida Rudzka-Ostyn (ed.), *Topics in cognitive linguistics*, 166–205. Amsterdam: Benjamins.

Tiersma, Peter Meijes. 1982. Local and general markedness. *Language* 58(4). 832–849.

Treis, Yvonne. 2014. Number in Kambaata. In Anne Storch & Gerrit J. Dimmendaal (eds.), *Number – constructions and semantics: Case studies from Africa, Amazonia, India and Oceania*, 112–134. Amsterdam: Benjamins.

Zipf, George Kingsley. 1935. *The psycho-biology of language: An introduction to dynamic philology.* Cambridge, Mass.: M. I. T. Press.

Corpora

BNC = *The British National Corpus* (of English), version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

EKK = *Eesti keele koondkorpus* [Estonian Reference Corpus]. The morphologically tagged version (cf. Kirt 2013) was used in this study. The most current version is available at http://www.cl.ut.ee/korpused.

MLVTK = *Mūsdienu latviešu valodas tekstu korpuss* [Modern Latvian Text Corpus]. Available at http://www.korpuss.lv/ (accessed 01.10.2014)

OK = *Oslo-korpuset av taggede norske tekster (bokmålsdelen)* [The Oslo Corpus of Tagged Norwegian Texts (the *bokmål*-part)]. Available at http://www.tekstlab.uio.no/norsk/bokmaal/

RNC = National Corpus of the Russian Language (НКРЯ = *Национальный корпус русского языка*) Available at http://ruscorpora.ru/ (accessed 23.11.2014).