Appendix for *Zipf's Law and Avoidance of Excessive Synonymy*

D. Yu. Manin, manin@pobox.com

Meaning and frequency

In this Appendix we'll consider some evidence in favor of the hypothesis that word frequency is proportional to the extent of its meaning. Far from being a systematic study, this is rather a methodological sketch. This study was done in Russian, the author's native language. In the English text we'll attempt to provide translations and/or equivalents wherever possible.

Strictly speaking, one could prove the hypothesis only if an explicit measure of meaning extent is proposed. However the frequency hypothesis allows to make some verifiable predictions. Suppose that some "head" word w_0 has a set of partial synonyms and/or hyponyms ("specific" words) $\{w_0^1, \ldots, w_0^n\}$, whose meanings together cover the meaning of w_0 without gaps and overlaps. We can make that judgement even without being able to measure meaning extent. Then, by definition, their total meaning extent is equal to that of w_0. In that case, the frequency hypothesis predicts that the sum total of hyponym frequencies should be close to the frequency of the head word.

One can not expect to find very many such examples in the real language. First, pure hyponyms are not very common; it is more common for words to have intersecting meanings, such as with *плохой*, 'bad, poor', and *худой*, 'skinny; torn, leaky; bad, poor'. Second, only in rare cases one can state confidently that the hyponyms cover the whole meaning of the head word. For example, in the domain of fine arts, *натюрморт* 'still life', *пейзаж* 'landscape', and *портрет* 'portrait' are pure hyponyms of the word *картина* 'picture', but there exist other genres of painting that can't be accounted for with frequency dictionary, since their names are phrases, rather than single words (*жанровая сцена* 'genre painting', *батальное полотно* 'battle-piece').

Nevertheless, examples of this type do exist. Table 1 contains frequencies of the head word *дерево, деревцо* 'tree; also dimin.' and of the specific tree names found in the frequency dictionary (Sharoff, n.d.). We omitted words denoting primarily the fruit or bloom of the corresponding tree, such as *груша* 'pear', *вишня* 'sour cherry', *рябина* 'rowan' or *магнолия* 'magnolia'. To count them correctly, one would have to know the fraction of word instances denoting the tree specifically, and we don't have this data.

From the table one can see that the sum of frequencies of specific tree names is very close to the frequency of the head word (we'll consider the "physicist's error margin" of 20% to be acceptable). Possibly, the word *пальма* 'palm tree' could be removed from the list: it is not clear why it turned out to be the sixth most frequent tree in Russian-language

Table 1: *Tree*.

word	freq./mln	word	freq./mln
дерево 'tree'	224.52	сосна 'pine'	38.07
деревцо 'tree dimin.'	8.08	дуб 'oak'	27.24
		елка 'fir'	26.57
		береза 'birch'	24.36
		тополь 'poplar'	17.75
		пальма 'palm tree'	16.96
		липа 'linden'	13.89
		яблоня 'apple tree'	13.41
		ива 'willow'	7.96
		кедр 'cedar'	7.77
		клен 'maple'	7.53
		осина 'aspen'	6.79
		лиственница 'larch'	6.00
		ель 'fir'	4.84
		орешник 'filbert'	4.84
		вяз 'elm'	3.31
		пихта 'fir'	3.24
		кипарис 'cypress'	3.18
		эвкалипт 'eucalyptus'	2.51
		ольха 'alder'	1.96
		ясень 'ash'	1.90
		ветла 'willow'	1.84
		бук 'beech'	1.78
		платан 'platan'	1.71
sum	**232.60**	sum	**246.82**

texts before *липа* 'linden' и *яблоня* 'apple tree'. However, small changes in the list will not conceptually affect the result.

This is just one example of many. Table 2 contains the frequencies of common flower names. They also sum up very close to the frequency of the word *цветок (цветочек)* 'flower; also dimin.'. (The word *колокольчик* 'small bell; bluebell', frequency 11.08, is omitted here, since primarily it denotes a bell, and not a flower.) Possibly, subtracting the frequencies of figurative meanings of words like *роза* 'rose', would improve the result.

Names of berries also follow this pattern, see table 3. (Here and below, we list in the table captions some words not found in the dictionary, apparently because their frequency is less than one per million.) The difference is somewhat greater in this case, but we should take into account that *малина* 'raspberry' and *клюква* 'cranberry' possess active figurative and idiomatic meanings in Russian (resp., 'a criminal flat' and an approximate equivalent of 'red herring'). Besides, it is not quite clear whether the cherries *вишня* and *черешня* truly belong in this list: first, a considerable number of instances will refer to corresponding trees, not fruits, and second, we are not certain whether the designation *ягода* 'berry' is appropriate for them. For instance, in the classical Dahl's dictionary, the entry for *cherry* starts with "A tree and its frut...", while the entry for *cranberry* or *raspberry* starts with "A

Table 2: *Flower*.

word	freq./mln	word	freq./mln
цветок 'flower'	134.85	роза 'rose'	41.50
цветочек (dimin.)	11.87	мак 'poppy'	27.91
		тюльпан 'tulip'	12
		одуванчик 'dandellion'	11.32
		сирень 'lilac'	9.92
		ромашка 'daisy'	8.63
		лилия 'lily'	7.65
		гвоздика 'carnation'	7.35
		подсолнух 'sunflower'	5.02
		черемуха 'bird cherry'	4.84
		лютик 'buttercup'	4.10
		фиалка 'violet'	4.22
		василек 'cornflower'	3.61
		ландыш 'lily of the valley'	2.94
		хризантема 'chrysanthemum'	2.82
		крокус 'crocus'	2.26
		нарцисс 'daffodil'	2.20
		герань 'geranium'	2.02
		астра 'aster'	1.90
		подснежник 'snowdrop'	1.78
		незабудка 'forget-me-not'	1.65
		гладиолус 'gladiolus'	1.29
		орхидея 'orchid'	1.29
		пион 'peony'	1.22
sum	**146.72**	sum	**169.44**

bush and its berry...". Of course, for the purposes of this work, it is a matter of lexicography, rather than botany.

In all the three examples, we didn't have to face the question of how to prove that the hyponyms indeed cover the head word's meaning without overlaps (an object can't be both a gooseberry and a blueberry) and gaps (each berry has a specific name). However, some subtleties can already be found here. Thus, if "в сорок пять баба ягодка опять" (a proverb; lit.: "at 45 a woman is a berry again") this "berry" is none of the berries we listed. On the other hand, *воровская малина* ('a criminal flat'; lit.: "thieves' raspberry") is not a berry. In this particular case, there is no doubt that such non-literal usage will not appreciably affect the results; what's more important, it is possible, at least in principle, to account for it by studying texts. Below we'll encounter much greater difficulties, which require systematic and more formal approaches.

A somewhat different example is given in table 4, containing a classification of meat produce, which is pretty chaotic from a logician's point of view, but quite common in everyday use. We'll note that although a sausage can contain beef or pork, the meanings of words *колбаса* 'sausage' and *говядина* 'beef' do not intersect (or intersect negligibly). The same can be said about other word pairs in the table. For the non-Russian reader,

Table 3: *Berry*. Not in dictionary: *gooseberry*, *cloudberry*, and *bilberry*.

word	freq./mln	word	freq./mln
ягода 'berry'	25.83	малина 'raspberry'	7.59
ягодка (dimin.)	3.00	вишня 'sour cherry'	6.98
		земляника 'wild strawberry'	5.69
		рябина 'rowan berry'	3.86
		смородина 'currant'	3.98
		клубника 'strawberry'	3.12
		клюква 'cranberry'	2.94
		брусника 'lingonberry'	2.82
		черника 'blueberry'	2.69
		ежевика 'blackberry'	2.08
		черешня 'cherry'	1.47
sum	**28.83**	sum	**43.22**
		without cherries	**34.77**

it should be noted that *мясо* does not have many extended meanings of English *meat*, and means practcally nothing beyond 'the flesh of animals used as food'. But are all the hyponym meanings really contained within the meaning of the word *мясо* 'meat'? For instance, can we say that *паштет* 'paté' $\subset$ *мясо* 'meat' (we will denote the relationships between meanings with mathematical symbols of subset, intersection, and union $\subset, \cap, \cup$)? The evidence in favor of this statement is provided by locutions like *Возьми паштет, тебе надо есть больше мяса* ('Take some paté, you need meat to recover').

So far, we only considered head words from a mid-frequency range (the most frequent, *дерево* 'tree' has a rank of 435). But the supporting data can be found among high-frequency words as well. Table 5 classifies humans by age and gender (the rank of the word *человек* 'human, person' is 33; it is counted together with its plural form, *люди*). As an aside, we note the curious fact that the most frequent words for male and female persons come in exactly opposite order in terms of age: in the order of decreasing frequency we have *старик* 'old man', *мальчик* 'boy', *парень* 'lad, guy', *мужчина* 'man', but *женщина* 'woman', *девушка* 'young woman', *девочка* 'young girl', *старуха* 'old woman'. Also, the net frequency of all the male terms (1377) is practically the same as the net frequency of all the female terms (1339). Frequency is rather uniformly distributed over age groups as well.

There are new difficulties in this case: obviously, there are significant intersections between the meanings of some hyponyms. This is mostly because

мальчик, девочка 'boy, girl' $\subset$ (*ребенок* 'child' $\cup$ *дитя* 'child' $\cup$ *младенец* 'baby')

(a boy or a girl is almost necessarily a child or a baby) [1]. Indeed, the net frequency of the words *ребенок*, *дитя*, *младенец* 'child, baby' is 637.7, and the net frequency of the words *мальчик*, *девочка*, *мальчишка*, *девчонка*, *пацан*, *паренек*, *парнишка*, *мальчонок* 'boy, girl' is 702.94, which is rather close. This provides another example conforming to the frequency

[1] Of course, there are exceptions here, too. Compare a quote from abovementioned Viktor Konetsky: *A fiftyish grocery store saleswoman is universally called "девушка" (girl), even though she has five children. And I once heard older female road workers going for lunch say: "Let's go, girls!"* Such a *girl* is not a *child*.

Table 4: *Meat.* Not in dictionary: *ромштекс* 'rump steak', *шницель* 'schnitzel'.

word	freq./mln	word	freq./mln
мясо 'meat'	84.47	колбаса 'sausage, bologna'	39.48
		котлета 'cutlet'	11.81
		сосиска 'sausage'	9.12
		ветчина 'ham'	6.49
		баранина '(meat of) lamb'	5.88
		свинина 'pork'	5.82
		бифштекс 'steak'	4.96
		говядина 'beef'	4.22
		фарш 'ground meat'	3.12
		паштет 'paté'	3.06
		телятина 'veal'	2.57
		сарделька 'wiener'	1.78
		отбивная 'chop'	1.47
		котлетка 'cutlet (dimin.)'	1.22
sum	**84.47**	sum	**101.00**

Table 5: *Human.*

word	freq./mln	word	freq./mln
человек 'human'	2945.47	ребенок 'child'	593.50
		женщина 'woman'	584.32
		старик 'old man'	313.64
		мальчик 'boy'	290.81
		девушка 'young woman'	286.53
		парень 'lad, guy'	258.74
		мужчина 'man'	252.98
		девочка 'young girl'	191.04
		старуха 'old woman'	105.89
		мальчишка 'boy (derog.)'	92.55
		девица 'girl; virgin'	59.86
		девчонка 'young girl (derog.)'	58.95
		юноша 'young man'	58.09
		старушка 'old woman (dimin.)'	52.21
		старичок 'old man (dimin.)'	40.95
		пацан 'boy (dial., colloq.)'	24.91
		младенец 'baby'	27.18
		паренек 'boy, dimin. of *lad*'	21.73
		парнишка 'boy, dimin. of *lad*'	19.95
		дитя 'child'	17.02
		мальчонок 'boy (dimin.)'	3.00
sum	**2945.47**	sum	**3353.85**
		without neut. terms	**2716.15**

Table 6: *Fish.* Not in dictionary: *краснопёрка 'rudd', салака 'sprat', палтус 'halibut', ставрида 'scad', нотатения, тунец 'tuna', кефаль 'mullet', налим 'burbot', плотва 'roach', севрюга 'sturgeon', пескарь 'gudgeon', мурена 'moray', омуль 'omul'.*

word	freq./mln	word	freq./mln
рыба 'fish'	120.03	сазан 'sazan'	16.47
рыбка (dimin.)	20.02	карась 'crucian'	14.63
		акула 'shark'	10.77
		селедка 'herring'	9.61
		карп 'carp'	9.24
		щука 'pike'	9.06
		сом 'catfish'	8.20
		скат 'ray'	6.98
		судак 'pike perch'	6.06
		лещ 'bream'	5.51
		форель 'trout'	4.53
		окунь 'perch'	4.41
		вобла 'vobla'	2.94
		камбала 'flounder'	2.88
		угорь 'eel'	2.82
		лосось 'salmon'	2.57
		треска 'cod'	2.14
		сельдь 'herring'	2.08
		хек 'hake'	2.02
		семга 'salmon'	1.78
		осетр 'sturgeon'	1.59
		ерш 'ruff'	1.59
		сардина 'sardine'	1.53
		стерлядь 'sterlet'	1.47
		скумбрия 'mackerel'	1.22
		белуга 'beluga'	1.10
		горбуша 'salmon'	1.10
sum	**140.05**	sum	**134.43**

hypothesis. In addition, we can subtract the net frequency of either all neutral terms or all gender-specific terms from the sum of frequencies, which makes the net frequency of the rest of hyponyms very close to the frequency of the head word *человек* 'human'.

The frequency hypothesis works with words of relatively low frequency as well: see tables 6 (*рыба* 'fish') and 7 (*забор* 'fence').

Let us now consider other parts of speech. Two simple examples with adjectives can be found in tables 8 (*старый* 'old') and 9 (*красный* 'red'). A more complicated example is given by the word *большой* 'big, large' shown in table 10. The net frequency of hyponyms significantly (by a quarter) exceeds the frequency of the head word. This is as expected, since some of the hyponyms' meanings definitely intersect: thus, *огромный* and *громадный* are as close to exact synonyms as it gets (cf. Eng. *huge* and *enormous*). However there's a possibility for a deeper and more interesting analysis here.

Table 7: *Fence.* Not in dictionary: *палисад.*

word	freq./mln	word	freq./mln
забор 'fence'	66.72	ограда 'fence'	25.83
		изгородь 'fence, hedge'	10.59
		плетень 'wicker fence'	9.61
		частокол 'stake fence'	5.39
		штакетник 'picket fence'	2.57
		загородка 'fence'	2.20
		тын 'paling'	1.96
sum	**66.72**	sum	**58.15**

Table 8: *Old.* Not in dictionary: *закоснелый, заматерелый, затасканный, зачерствелый, истасканный, подержанный, полинялый, поседелый, потрепанный, старобытный.*

word	freq./mln	word	freq./mln
старый 'old'	528.25	древний 'ancient'	75.60
		пожилой 'elderly'	63.17
		седой 'grey-haired'	62.99
		старинный 'antique'	53.07
		давний 'bygone'	34.71
		бородатый 'bearded; old (of jokes)'	18.67
		немолодой 'not young'	16.34
		многолетний 'longstanding'	11.51
		старомодный 'old-fashioned'	11.51
		престарелый 'very old (of people)'	10.04
		ветхий 'shabby, decrepit'	9.67
		вековой 'age-old'	6.86
		извечный 'primeval'	6.67
		отсталый 'outdated, retrograde'	5.94
		дряхлый 'decrepit'	5.82
		устарелый 'outmoded, outdated'	5.39
		ископаемый 'fossilized'	5.20
		поношенный 'worn, shabby'	4.77
		допотопный 'antediluvian'	4.16
		давнишний 'bygone'	3.55
		застарелый 'inveterate'	3.37
		многовековой 'centuries-old'	3.37
		исконный 'original'	3.06
		заскорузлый 'calloused, backward'	2.69
		закоренелый 'inveterate, ingrained'	1.96
		истертый 'worn'	1.71
		отживший 'obsolete'	1.65
		архаический 'archaic'	1.35
		стародавний 'ancient'	1.35
		обветшалый 'shabby, decrepit'	1.29
		архаичный 'archaic'	1.04
sum	**528.25**	sum	**438.48**

Table 9: *Red.* Not in dictionary: *карминный, рдяный, червленый.*

word	freq./mln	word	freq./mln
красный 'red'	316.64	рыжий 'red-haired; rust-colored'	89.8
		розовый 'rosy, pink'	77.98
		алый 'scarlet'	32.99
		кровавый 'bloody'	32.93
		багровый 'crimson'	22.16
		румяный 'ruddy'	17.2
		малиновый 'crimson'	14.02
		пунцовый 'crimson'	3.55
		бордовый 'vinous'	2.82
		багряный 'crimson (arch., poet.)'	2.63
		коралловый 'coral'	2.57
		морковный 'carrot (adj.)'	2.57
		рубиновый 'ruby (adj.)'	2.2
		пурпурный 'purple'	1.84
		свекольный 'beet (adj.)'	1.04
sum	**316.64**	sum	**306.3**

Consider locutions 1–10.

Is this a raspberry or a strawberry? (1)

*Is this a strawberry or a berry? (2)

Is this a boy or a girl? (3)

Is this a boy or a man? (4)

*Is this a boy or a child? (5)

*Is this a boy or a person? (6)

Do you want pork or paté? (7)

?*Купить свинину или мясо? '≈Do you want pork or meat?' (8)

?*Купить паштет или мясо? '≈Do you want paté or meat?' (9)

*Купить говядину или мясо? '≈Do you want beef or meat?' (10)

Everything is clear with items 1–7: non-intersecting specific words can occur in alternative constructions with each other, but not with the head words. Locutions 8, 9 are possible only if *мясо* 'meat' is used in constrained, specialized (sub)meanings, existing in the vernacular: (*мясо* 'meat')^2 = *говядина* 'beef', (*мясо* 'meat')^3 = *сырое мясо* 'raw meat' (this is proved by the fact that 10 is not possible). These example, therefore, also involve non-intersecting (non-overlapping) meanings. As a first approximation, we will consider this as a criterion of meaning overlap: if two words can participate in an alternative construction of this type, their meanings do not overlap.

To apply this criterion to hyponyms of the word *большой* 'big, large', consider examples 11, 12. Although the semantic difference between them is intuitively obvious, it is not easy to explicate it. There are objects that are both long and wide, as well as objects that are both long and huge, — and still the first example is perfectly valid, while the second

one is impossible. But keeping with the methodological principles of this work, we will not attempt to formulate the difference in semantic terms. On the contrary, we take acceptability of a locution as a linguistic datum, and on this basis draw conclusions about word semantics. That is, we will *define* two words non-overlapping in their meanings, if they can can participate in an alternative construction of the type 1–12.

Is it wide or long? (11)

*Is it long or huge? (12)

Now, accepting the above criterion for non-overlapping meanings, we can select a subset of hyponyms from table 10, which do not overlap and mean roughly 'big/large in a certain dimension or trait'. Almost all remaining hyponyms are in fact emphatic or superlative terms: 'very big/large, regardless of dimension or trait' (only two, *немалый* 'not small' and *изрядный* 'fairly large', are hard to classify). It is easy to make sure that the first group consists of virtually non-overlapping adjectives. Admittedly, in the lower part of the table, the criterion becomes less clear-cut: thus, the question in example 13 is somewhat awkward; however it is meaningful and understandable, in contrast to 12. Of course, there is still some overlap in the meanings; after all, we're dealing with a living language. But it is small enough so that any further corrections will not change the result in any significant way (and may still improve it).

— The king's palace is big!
— Is it spacious or grandiose? (13)

In the 5th column of table 10 we sum up the frequencies of the hyponyms that are specifying the trait or dimension. The net frequency is very close to the frequency of the head word.

The word *маленький* 'small, little' (table 11) is very similar. However we face a new complication here: the main concept is expressed by three words, rather than one: *маленький*, *небольшой* and, possibly, *малый*. It is somewhat similar to the distinction between *small* and *little* in English. Consider the first two adjectives. Both are direct and stilistically neutral antonyms to *большой* 'big, large'. However their meanings are distinct. For example, they are not interchangeable in the common phrases like *маленький мальчик* 'little boy' and *небольшое количество* 'small amount': **небольшой мальчик* and **маленькое количество* are not normative (while the adjective *большой* 'big, large' can modify both nouns). But even when both adjectives are admissible, they mean different things. Thus, *маленькая мышка* '≈a little mouse' means 'small compared to the speaker, as all mice', or, less probably, 'a young mouse', but *небольшая мышка* '≈a small mouse' means 'small compared to other mice, less than usual mouse size'. Even when this distinction is not applicable, there still can be a quantitative difference, as in example 14.

— Этот кусок слишком большой. 'This piece is too big.'
— Отрезать тебе небольшой или маленький? '≈Do you want a smaller one or a small one?' (14)

As a result, we consider the words *маленький* и *небольшой* to have almost non-overlapping meanings. As for the adjective *малый*, in its long form it is used only in

Table 10: *Big/large.*

word	freq./mln	word	freq./mln	trait	emphasis
большой	1630.96	высокий 'tall, high'	310.34	height	-
'big, large'		огромный 'huge'	298.95	-	+
		великий 'great (significant)'	247.90	significance	+
		длинный 'long (space)'	244.05	length	-
		широкий 'wide'	187.31	width	-
		толстый 'thick'	176.12	diameter; thickness	-
		крупный 'large-scale, coarse'	151.74	all dimensions	-
		глубокий 'deep'	135.58	depth	-
		долгий 'long (time)'	132.52	time	-
		значительный 'significant'	60.17	significance	-
		гигантский 'giant'	42.24	-	+
		громадный 'tremendous'	40.77	-	+
		длительный 'prolonged'	35.56	time	-
		просторный 'spacious'	28.03	space	-
		обширный 'vast'	26.20	extent	-
		немалый 'not small'	22.83	-	-
		грандиозный 'grandiose'	18.24	impression, intent	+
		внушительный 'impressive'	13.34	impression	-
		колоссальный 'colossal'	9.79	-	+
		громоздкий 'bulky'	9.73	all dims.; maneuverability	-
		изрядный 'fairly large'	8.75	-	-
		исполинский 'gigantic'	6.37	-	+
		масштабный 'large-scale'	4.16	intent; influence	-
		непомерный 'exorbitant'	3.98	-	+
		объемный 'bulky'	3.43	bulk, volume	-
		объемистый 'voluminous'	3	volume, bulk	-
		большущий 'big (superl.)'	2.14	-	+
		протяженный 'lengthy'	1.47	length	-
sum	**1630.96**	sum	2048.59	**1788.89**	670.38

compound toponyms and scientific nomenclature (cf. *Lesser Antilles*). But in its short form, it has a common and distinctive meaning of 'too small to fit', not covered by adjectives *маленький* and *небольшой*. Indeed, if *туфли малы* '≈shoes are too small', this doesn't necessarily mean that the shoes are small, they still can be size 10. But they are necessarily narrow, short, or tight. This is why the adjective *малый* is also placed in table 11 as a head word, and not as a hyponym.

This argument is based on intuitive judgement about acceptability of certain expressions, which is not a very solid foundation (cf. Wasow & Arnold, 2005). To improve it, one would have to formulate strict criteria of intersection and inclusion for meanings, and then demonstrate that they are satisfied. This is generally beyond the scope of the present essay, but one example of a completely objectivised approach is given below for the word *плохой* 'bad'.

Verbs provide some good examples as well. See tables 12 (*сказать* 'say') and 13

Table 11: *Small*.

word	freq./mln	word	freq./mln	trait	emphasis
маленький	411.52	короткий 'short in length'	202.55	length	-
'small, little'		тонкий 'thin'	144.58	thickness	-
небольшой	180.08	мелкий 'shallow; fine'	125.05	depth; all dims.	-
'not large'		узкий 'narrow'	105.47	width	-
малый	108.71	низкий 'low; short in height'	78.23	height	-
'lesser; too small'		тесный 'tight'	33.18	spaciousness	-
		крохотный 'tiny'	28.4	-	+
		крошечный 'tiny'	24.67	-	+
		незначительный 'insignificant'	20.69	significance	-
		ничтожный 'very insignificant'	19.71	significance	+
		невеликий 'not great'	13.04	significance	-
		миниатюрный 'miniature'	5.26	-	+
		неглубокий 'not deep'	4.77	depth	-
		нешироκий 'not wide'	3.86	width	-
		малюсенький 'small (superl.)'	3.61	-	+
		мизерный 'paltry'	3.61	-	+
		микроскопический 'microscopic'	3.06	-	+
		махонький 'wee'	2.02	-	+
		недлинный 'not long'	1.78	length	-
sum	**700.31**	sum	823.54	**752.91**	90.33

(*думать* 'think') that do not require any comments.

In two other verbs we encounter a complication of a new type: see tables 14 (*подниматься/расти* 'rise/grow') and 15 (*кричать/плакать* 'shout/cry'). The words *подниматься* 'rise, ascend' and *расти* 'grow, increase' have some common sub-meanings, such as *увеличиваться* 'increase in quantity or size' as well as distinct ones, such as *взлетать* 'soar, take off' and *расширяться* 'widen, spread' respectively. For example, a temperature can both rise and grow (in Russian *температура растет* is much more common than English *temperature grows*), these expressions being quite synonymous and meaning the increase in temperature. On the other hand, an elevator can only rise, while a child can only grow (a child can rise up on its toes, but this is a completely different meaning, of course). Apparently, in every or almost every context, the verb *увеличиваться* 'increase' can be replaced with either *подниматься* 'rise' or *расти* 'grow' (this is a statement about Russian verbs, not their approximate equivalents in English), which means that its meaning is a subset of the intersection of their meanings — see Fig. 1.

It turns out that the net frequencies of hyponyms match the head word frequencies in both columns of table 14. This would even allow us to quantify the degree of commonality between the meanings of the two head words. Exactly the same behavior can be observed with words *кричать* 'shout' and *плакать* 'cry'.

Finally, consider two more adjectives, *хороший* 'good' (table 16) and *плохой* 'bad, poor' (table 17). Synonyms (or rather hyponyms) were collected from dictionaries. The former word doesn't cause any difficulties: the net frequency of hyponyms corresponds well

Table 12: *Say.*

word	freq./mln	word	freq./mln
сказать 'say'	3535.97	спросить 'ask (a question)'	934.32
		ответить 'answer'	503.46
		рассказать 'tell'	248.58
		произнести 'pronounce'	178.98
		крикнуть 'shout'	155.97
		попросить 'ask (for a favor)'	154.62
		сообщить 'inform'	148.80
		приказать 'command'	107.18
		велеть 'order'	95.67
		заявить 'state'	86.61
		воскликнуть 'exclaim'	81.66
		проговорить 'utter'	78.35
		возразить 'object'	69.66
		предупредить 'caution'	51.23
		пробормотать 'mutter'	49.52
		прошептать 'whisper'	39.79
		пообещать 'promise'	33.42
		возмутиться 'say indignantly'	27.61
		осведомиться 'inquire'	26.32
		буркнуть 'growl'	25.52
		шепнуть 'whisper'	24.79
		пошутить 'joke'	24.06
		поздороваться 'greet'	22.34
		выразиться 'curse'	22.28
		попрощаться 'say goodbye'	20.51
		скомандовать 'command'	19.59
		проворчать 'growl'	18.79
		рявкнуть 'bark out'	17.14
		выговорить 'utter'	16.22
		прокричать 'shout'	12.67
		высказаться 'express'	12.12
		провозгласить 'announce'	11.94
		гаркнуть 'bawl'	10.04
		молвить 'say (arch., poet.)'	9.67
		промолвить 'say (arch., poet.)'	6.18
		брякнуть 'blurt'	6.18
		пролепетать 'babble'	5.08
		промямлить 'mumble'	4.47
		съязвить 'say sarcastically'	3.67
		вопросить 'inquire'	2.88
		вякнуть 'blather'	2.51
		предостеречь 'warn'	2.45
sum	**3535.97**	sum	**3372.85**

Table 13: *Think.*

word	freq./mln	word	freq./mln
думать 'think'	936.40	считать 'reckon'	396.22
		мечтать 'dream'	83.61
		полагать 'believe'	73.45
		предполагать 'presume'	50.56
		рассуждать 'reason'	38.20
		соображать 'consider'	36.36
		размышлять 'reflect on'	29.75
		воображать 'imagine'	20.69
		мыслить 'concieve'	19.47
		раздумывать 'ponder'	16.53
		прикидывать 'reckon'	11.57
		обдумывать 'think over'	11.14
		вникать 'fathom'	7.53
		помышлять 'dream of'	3.80
		замышлять 'scheme'	2.75
		мнить 'imagine'	2.33
		вдумываться 'ponder'	1.47
		кумекать 'think (low colloq.)'	1.16
sum	**936.40**	sum	**806.59**

Figure 1. *Rise* and *grow* (cf. table 14).

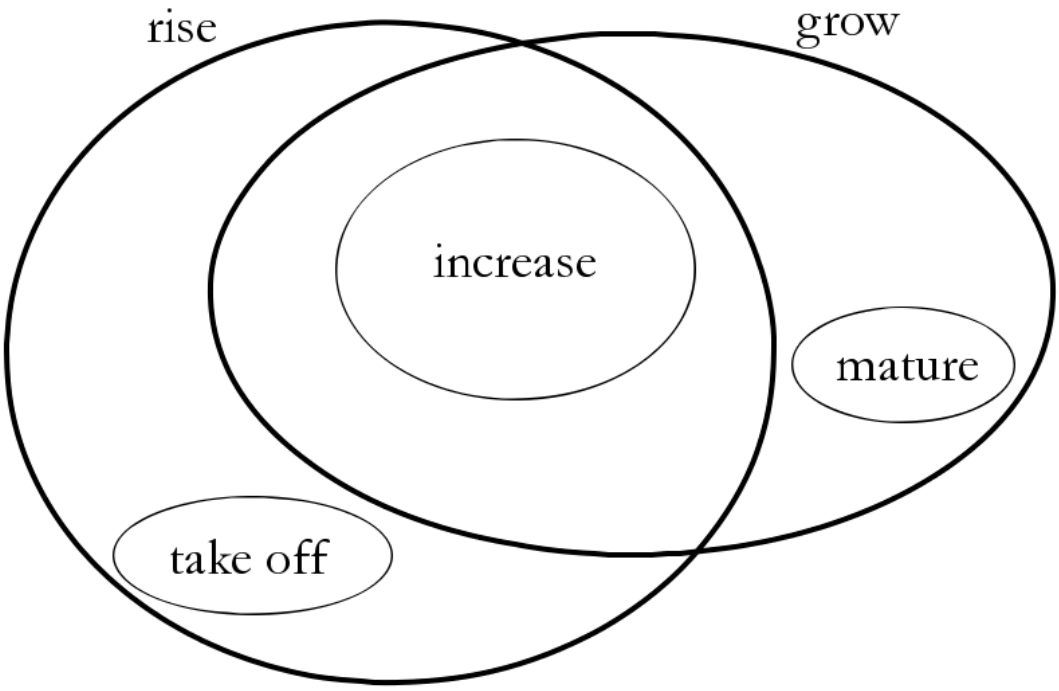

Table 14: *Rise* and *grow*. Translations are very approximate.

word	freq./mln	word	freq./mln
подниматься 'rise'	**102.41**	расти 'grow'	**71.74**
увеличиваться 'increase'	21.24	увеличиваться 'increase'	21.24
вырастать 'grow'	13.04	вырастать 'grow'	13.04
возрастать 'grow'	12.12	возрастать 'grow'	12.12
прибывать 'rise, swell'	12.12	прибывать 'grow, swell'	12.12
взлетать 'soar up, take off'	14.38		
взбираться 'climb'	8.81		
		расширяться 'spread, widen'	8.32
всплывать 'rise to the surface'	6.92		
вздыматься 'heave'	5.51		
подрастать 'grow'	4.77	подрастать 'grow'	4.77
восходить 'rise, ascend'	4.04		
всходить 'rise, ascend'	3.98		
возноситься 'rise, tower'	1.71	возноситься 'rise, tower'	1.71
		взрослеть 'mature'	2.94
		шириться 'expand, widen'	2.69
		совершенствоваться 'improve'	2.69
взвиваться 'soar up, be hoisted'	1.78		
		умножаться 'multiply'	1.16
sum	**108.64**	sum	**82.8**

Table 15: *Shout* and *cry*. Translations are very approximate.

word	freq./mln	word	freq./mln
кричать 'shout'	**220.36**	плакать 'cry, weep'	**120.71**
орать 'yell'	67.64		
шуметь 'make noise'	44.62		
реветь 'roar; cry'	26.99	реветь 'roar; cry'	26.99
		рыдать 'sob'	18.30
выть 'wail'	17.51	выть 'wail'	17.51
визжать 'shriek'	16.04	визжать 'shriek'	16.04
вопить 'bawl'	15.98		
		всхлипывать 'sob'	12.06
надрываться 'bawl'	6.86		
		скулить 'whine'	4.84
галдеть 'clamor'	4.41		
		пищать 'squeak'	4.28
верещать 'chirp, squeal'	3.67		
скандалить 'brawl'	3.67		
голосить 'wail'	3.24		
		хныкать 'whimper'	2.20
горланить 'bawl'	1.84		
гомонить 'shout'	1.35		
sum	**213.82**	sum	**102.22**

with the head word frequency. However, with the adjective *плохой* 'bad' the situation is quite different. Note first of all that the four most frequent synonyms offered by the dictionaries (*худой* 'skinny; leaky; bad', *низкий* 'low, short; base, mean', *дешевый* 'cheap, worthless', *жалкий* 'pitiful; wretched') are not included in the table, because each of them has a primary meaning that does not directly imply badness. Something or somebody can be cheap and good, skinny and good, etc. But even without them, the net frequency of hyponyms is significantly greater than the head word frequency.

Notice though that the hyponyms can be roughly classified into two categories: those denoting more of an objective quality of an object, like *скверный* (cf. Eng. *poor* in its senses unrelated to pitying and lack of wealth), and those denoting more of a subjective feeling towards the object, like *мерзкий* 'loathsome, vile'. The head word itself falls more in the former category. To demonstrate this, consider the expression *плохой вор* 'a bad thief'. Its meaning is 'one who is not good at the art of stealing', in contrast to *мерзкий вор* 'vile thief' = 'one whom I loath because he steals'. Hence, only the frequencies of the hyponyms from the first category (denoting quality) should sum up to the frequency of the head word.

But it is quite difficult to actually classify the words into these two categories. The "subjective" words tend to evolve towards emphatic terms, and further migrate to the "objective" group or close to it. So we need a method that would allow to perform classification without relying on dubious judgements based on the linguistic intuition. To this end, notice that there exist three classes of nouns by their compatibility with the adjectives from table 17. Neutral nouns, like *погода* 'weather' can be equally easy found in noun phrases with both *скверный* 'bad, poor' and *мерзкий* '≈disgusting'. However the nouns carrying distinct negative connotation, such as *предатель* 'traitor' are well compatible with *мерзкий* '≈disgusting', but not with *скверный* 'bad, poor'. On the contrary, nouns with distinct positive connotation have the opposite preference: cf. *скверный поэт* 'bad poet' and ?**мерзкий поэт* 'disgusting poet'. It is possible to find out which of the adjectives in table 17 tend to apply preferentially to positive or negative nouns, by using an Internet search engine.

We considered eight test nouns: negative *гадость* '≈filth', *дрянь* '≈trash', *предатель* 'traitor', *предательство* 'treason' and positive *здоровье* 'health', *врач* 'doctor', *поэт* 'poet', *актер* 'actor'. They were initially selected for maximum contrast in their compatibility with adjectives *скверный* and *мерзкий*. Then we used Russian-specific search engine Yandex (http://www.yandex.ru) to determine the frequencies of noun phrases constructed from each of the adjectives with each of the nouns.

It should be noted here that search engines can't be directly used as replacements for a frequency dictionary. First, they typically report the number of "pages" and "sites", but not the number of word instances. Meanwhile, web pages can be of very different size, and may contain multiple instances of a word or search phrase. Second, search engines trim the results to exclude "similar pages" and avoid duplicates, i.e. texts available in multiple copies or from multiple addresses. It's not clear whether this is correct behavior from the point of view of calculating frequencies. Finally, the corpus with which search engines work, the whole of the Web, is by no means well-balanced according to the criteria of frequency dictionary compilers. So the results from search engines can't be directly compared with the data from frequency dictionaries. But for our purposes we need only relative figures, and we are interested in their qualitative behavior only. The effect we are looking for, if it exists, should be robust enough to withstand the inevitable distortion.

Table 16: *Good.*

word	freq./mln	word	freq./mln
хороший 'good'	853.71	добрый 'good, kind'	201.38
		прекрасный 'splendid, excellent'	143.17
		приятный 'nice'	74.31
		блестящий 'brilliant'	61.33
		замечательный 'remarkable'	60.84
		благородный 'noble'	57.66
		отличный 'excellent'	42.24
		славный 'glorious, nice'	38.44
		великолепный 'magnificent'	34.95
		чудесный 'wonderful'	34.46
		роскошный 'splendid'	27.91
		неплохой 'not bad'	26.38
		чудный 'wonderful'	13.71
		превосходный 'excellent'	13.47
		прелестный 'lovely, delightful'	12.24
		дивный 'charming'	9.79
		благой 'good'	8.88
		безупречный 'impeccable'	8.63
		образцовый 'exemplary'	8.57
		годный 'suitable, valid'	7.96
		путный 'worthwhile'	7.77
		отменный 'excellent'	7.35
		изумительный 'marvellous'	6.79
		восхитительный 'adorable'	6.55
		пригодный 'suitable'	6.49
		добросовестный 'conscientious'	4.53
		удовлетворительный 'satisfactory'	3.86
		доброкачественный 'of good quality'	3.31
		благоустроенный 'well-furnished'	2.08
		похвальный 'laudable'	1.96
		бесподобный 'incomparable'	1.78
sum	**853.71**	sum	**938.79**

Table 17: *Bad*. Some translations are very approximate.

word	freq./mln	word	freq./mln	weight	quality?
плохой 'bad, poor'	102.22	дурной 'bad, mean'	40.40	0.911	+
		противный 'repugnant'	28.34	-0.0584	
		отвратительный 'disgusting'	21.85	-0.439	
		нехороший 'not good'	20.14	0.914	+
		мерзкий 'vile'	13.22	-1.946	
		скверный 'bad, poor'	13.16	0.896	+
		гнусный 'abominable'	12.73	-3.160	
		поганый 'foul'	11.51	-0.330	
		паршивый 'nasty'	10.16	-0.407	
		кошмарный 'nightmarish'	9.30	-0.180	
		негативный 'negative'	7.10	-0.183	
		неважный 'rather bad'	6.00	1.200	+
		омерзительный 'disgusting'	6.00	-0.432	
		гадкий 'repulsive; nasty'	5.33	-0.490	
		хреновый 'bad, poor (colloq.)'	5.14	2.358	+
		никчемный 'worthless'	5.08	0.144	+
		негодный 'worthless'	4.10	0.157	+
		дрянной 'rotten, trashy'	3.92	-0.110	
		никудышный 'worthless'	3.37	3.095	+
		захудалый 'run-down'	2.57	0.347	+
		неприглядный 'unsightly'	2.39	—	
		незавидный 'unenviable'	1.90	-0.161	
		дерьмовый 'shitty'	1.90	0.161	+
		фиговый 'bad, poor (colloq.)'	1.78	0.545	+
		неудовлетворительный 'unsatisfactory'	1.65	-0.077	
		паскудный 'foul, filthy'	1.59	-0.203	
		отвратный 'disgusting'	1.41	-0.165	
		грошовый 'dirt-cheap'	1.35	-0.172	
		бросовый 'worthless, trashy'	1.35	—	
		пакостный 'foul, mean'	1.35	-0.234	
		одиозный 'odious'	1.35	-0.122	
		сволочной 'mean, vile'	1.04	-0.318	
		аховый 'rotten'	0	-0.109	
		дефектный 'defective'	0	-0.179	
		завалящий 'worthless'	0	-0.078	
		мерзостный 'disgusting'	0	-0.270	
		мерзопакостный 'disgusting'	0	-0.302	
		низкопробный 'low-grade'	0	-0.406	
		отталкивающий 'revolting'	0	-0.198	
sum	**102.22**		248.48		**103.64**

Table 18: Compatibility of the hyponyms of *плохой* 'bad, poor' with test nouns on the Web ("the number of pages").

	дрянь	гадость	предатель	предательство	здоровье	врач	актер	поэт
дурной	0	1	0	0	81	66	187	172
противный	140	333	45	0	0	61	35	31
отвратительный	305	5187	0	112	71	45	211	43
нехороший	15	38	11	19	24	282	4	11
мерзкий	627	1354	849	316	7	42	30	16
скверный	4	3	15	3	232	54	250	137
гнусный	156	62	1380	1934	2	1	0	3
поганый	183	27	97	33	87	14	17	32
паршивый	493	32	156	12	27	8	314	38
кошмарный	26	39	0	4	0	7	8	1
негативный	2	0	0	0	0	0	0	0
неважный	0	0	0	0	1589	22	62	141
омерзительный	149	183	10	80	0	5	9	0
гадкий	132	257	140	28	12	3	15	3
хреновый	1	0	0	0	226	166	431	381
никчемный	58?	0	7	0	33	23	38	81
негодный	63	0	6	0	108	39	32	58
дрянной	114	2	0	0	6	1	18	62
никудышный	13	0	0	0	136	146	989	409
захудалый	0	1	0	0	0	8	22	150
неприглядный	0	0	0	0	0	0	0	0
незавидный	0	0	0	0	39	0	0	0
дерьмовый	8	1	51	1	13	14	89	70
фиговый	0	0	0	0	11	33	53	167
неудовлетворительный	0	0	0	0	212	0	0	0
паскудный	4	4	18	4	0	1	0	0
отвратный	28	92	0	0	2	10	12	1
грошовый	0	0	0	0	0	0	6	0
бросовый	0	0	0	0	0	0	0	0
пакостный	34	22	4	3	0	0	0	0
одиозный	0	0	7	0	0	0	28	8
сволочной	99	21	18	0	16	2	0	0
аховый	0	0	0	0	3	0	3	21
дефектный	0	0	0	0	3	0	0	0
завалящий	1	0	0	0	0	26	2	0
мерзостный	41	36	36	0	0	1	0	0
мерзопакостный	96	84	4	2	2	4	0	1
низкопробный	173	29	0	2	0	0	6	1
отталкивающий	0	3	19	0	0	0	1	0
Eigenvector	-0.300	-0.110	-0.418	-0.377	0.206	0.355	0.416	0.489

The frequencies of noun phrases constructed from each of the adjectives a_i with each of the test nouns n_j form a matrix N_{ij} presented in table 18. One can readily see that the rows "мерзкий" and "скверный" clearly separate the test nouns into two groups preferentially compatible with one or the other. Many other rows of the table (e.g. "гнусный" and "неважный") behave in the same way. But there are rows that do not, and that is precisely the reason to consider multiple test words. Thus the adjective *негодный* '≈worthless' is entirely compatible with all the positive test nouns, but also with the negative test noun *дрянь* '≈trash'. The adjectives *неприглядный* 'unsightly' and *бросовый* 'worthless', as it turns out, are not compatible with any of them, so they are excluded from further analysis. Their low frequency can't appreciably change the result anyway.

To recap, we want to classify the rows of table 18 by whether each row is more similar to the row "скверный" (quality of the object) or to the row "мерзкий" (speaker's attitude towards the object). This can be done via a statistical procedure known as Principal component analysis (PCA) or Singular value decomposition (SVD), which has found many different uses in statistical NLP in the recent years (e.g. Pado and Lapata (2007)).

First, each row of table 18 was normalized by subtracting the average and dividing by the standard deviation. As a result, the rows "мерзкий" and "скверный" become almost opposite to each other: positive on positive test nouns and negative on negative ones, or vice versa. Then, the correlation matrix of the table's columns was calculated (size 8×8) and its first eigenvector n_j^1. Finally, the eigenvector's scalar products with the i-th row of the table yields the weight of the corresponding adjective $a_i^1 = \sum_j n_j^1 N_{ij}$.

Mathematically, the result of this procedure is that the product $a_i^1 n_j^1$ provides the best (in terms of mean square) approximation of this kind to the matrix N_{ij}. In other words, each row of the normalized table 18 is approximately proportional to the pattern row n_j^1 multiplied by the weight a_i^1. The pattern row is given at the bottom of table 18. As expected, it correctly classifies test nouns as positive and negative. This means that they actually behave in opposite ways relative to the adjectives of interest. Now we can classify all the adjectives with positive weights $a_i^1 > 0$ as proper hyponyms of the word *плохой* 'bad, poor'. The weights are shown in table 17 (in an arbitrarily normalization). The table shows that the net frequency of these proper hyponyms is very close to the frequency of the head word.

So it can be seen that the frequency hypothesis is confirmed here as well, and this conclusion is not based on any intuitive judgement about word semantics.

We conclude with a brief discussion of some encountered counterexamples. In contrast to the words *дерево* 'tree', *цветок* 'flower', *ягода* 'berry', and *рыба* 'fish', the words *животное* 'animal' and, to a lesser extent, *птица* 'bird' are significantly less frequent than predicted by the net frequency of their hyponyms. The reason is probably that some of the most frequent animal and bird names have very wide connotations, far beyond the notion of 'this or that animal/bird'; e.g. *осел* 'donkey, ass' and *орел* 'eagle' (apparently, a much less loaded word in English than in Russian, where it readily stands for power, grandeur, nobility, both straight and ironic). It is not surpirising then, that the frequency of such words is much greater than had they denoted strictly the corresponding animals. (See also the discussion of the words *собака* 'dog' and *лошадь* 'horse' in Section 2). Among tree and flower names, only a small number are like that, and to a much smaller degree, e.g. *дуб* 'oak' (its Russian figurative meaning as 'a dumb, insensitive person' doesn't seem to have

a counterpart in English) and *роза* 'rose' (which doesn't have any fixed dictionary senses other than the flower, but has an established tradition of metaphoric usage). It is possible, at least in principle, to quantify the last statement by analyzing the actual word usage, and then counterexamples could turn into confirming evidence.

Interesting counterexamples are provided by words *страна* 'country, state', *город* 'city, town', *река* 'river, creek', and *озеро* 'lake'. The net frequency of the nouns *страна* 'country, state', *государство* 'state, nation', *республика* 'republic', and *королевство* 'kingdom' is 705.39 per mln. The net frequency of all the countries of the world found in the dictionary (except the former Soviet republics) is 1206.05, which is about 70% too much. However the first word in the list, *Россия* 'Russia', is four times as frequent as the number two (*Германия* 'Germany'). Its frequency is 358.88 per mln and is responsible for most of the discrepancy. Of course, Russia for Russian speakers is much more than just another country. Most of the rest of the discrepancy can be attributed to the fact that the word *Америка* 'America' denotes two continents and a part of world, in addition to the country.

A very similar situation occurs with the word *город* 'city, town'. Its frequency is 630.59 per mln, while the net frequency of all city names we could find in the dictionary is 1087.18. But here again, *Москва* 'Moscow' (frequency 420.89, 5–6 times more than the next city name) is responsible for the whole discrepancy. "Москва... как много в этом звуке"[2].

On the other hand, the net frequency of all the river names in the dictionary is somewhat less than the frequency of the word *река* 'river' (187.61 vs. 199.36), and that despite the fact that *дон*, *Урал*, and *Амур* are not just river names (a Spanish nobleman title, the Ural mountains, and 'Cupid; love affair' respectively). This same effect is much more pronounced with the word *озеро* 'lake': its frequency is 74.496 while the net frequency of all the lake names in the dictionary is only 21.72. Most probably, this is because only five lake names made it to the dictionary: *Байкал* 'Baikal', *Ладога* 'Ladoga', *Онега* 'Onega', *Виктория* 'Victoria' (some instances are, probably, personal names), *Иссык-Куль* 'Issyk-Kul'. Most lake names either fall below the 1 per mln threshold, or are homonymous with common names or adjectives. The same is true to a lesser degree for river names.

To summarize, we demonstrated on several examples that the hypothesis of word frequency being proportional to the extent of its meaning is supported by available data, while counterexamples are few and tend to have plausible explanations. Of course, a much more thorough and systematic investigation is in order until the hypothesis can be considered proven. We only sketched some promising approaches to such an investigation. But it also should be noted that the examples considered span a wide range of word frequencies, include all three main parts of speech, and involve very common words, not specially hand-picked ones.

References

Pado, S., & Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, *33*(2).

Sharoff, S. (n.d.). *The frequency dictionary for Russian.* Available from `http://www.artint.ru/projects/frqlist/frqlist-en.asp`

Wasow, T., & Arnold, J. (2005). Intuitions in linguistic argumentation. *Lingua*, *115*(11), 1481–1496.

[2]"Moscow... how much the sound embraces", from Pushkin's *Eugene Onegin*