Linguistic Problems on Number Names

Ivan Derzhanski
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
iad58g@gmail.com

Milena Veneva
Independent Researcher
milena.p.veneva@gmail.com

Abstract

This paper presents a contrastive investigation of linguistic problems based on number names in different languages and intended for secondary-school students. We examine the eight problems of this type that have been assigned at the International Linguistics Olympiad throughout the years and compare the phenomena in the number systems featured there with those of the working languages of the Olympiad and other languages known to be familiar to the participants. On the basis of a statistical analysis of the results achieved by the contestants we draw conclusions regarding the ways in which the difficulty of a problem depends on its structure and the kinds of linguistic phenomena featured in it.

1. Introduction

Self-sufficient problems on number names have a long past and a tangible presence at linguistic olympiads for secondary-school students. Of the 75 problems that have been assigned at the individual contest of the first 15 instalments of the International Linguistic Olympiad (IOL) (www.ioling.org/), 8 (11%) are on number names. So are 9 (5%) problems out of 168 at the 11 instalments of the North American Computational Linguistics Olympiad (www.nacloweb.org/) and 3 (15%) out of the 20 problems in (Derzhanski, 2009).

Despite fitting comfortably into the general scheme, these problems are often seen as a maverick category, pertaining to a separate area of linguistics if not to another field of science altogether ('We did not know that linguistics could be mathematical'—from the confession of a team who had had little success with one such problem, and were attributing that to their pre-installed perception of linguistics as a branch of the humanities, unrelated to the exact sciences). And they have a reputation for being fiendishly difficult.

There is some (though not much) truth to the latter. These are the problems on number names that have been used at IOL to date and the languages featured in each, complete with their ISO 639-3 codes, families and countries where spoken:

1. IOL1.#2 (Ivan Derzhanski): Egyptian Arabic (arz: Afro-Asiatic, Egypt);
2. IOL3.#3 (Ivan Derzhanski): Mansi (mns: Uralic, Russian Federation);
3. IOL5.#4 (Ivan Derzhanski): Ndom (nqm: Trans-New Guinea, Indonesia);
4. IOL7.#1 (Evgenia Korovina and Ivan Derzhanski): Sulka (sua: isolate, Papua New Guinea);
5. IOL8.#2 (Ksenia Gilyarova): Drehu (dhv: Austronesian, New Caledonia);
6. IOL10.#2 (Ksenia Gilyarova): Umbu-Ungu (ubu: Trans-New Guinea, Papua New Guinea);

Keywords: number names, numerals, typology, linguistic problems

7. IOL13.#1 (Milena Veneva): Arammba (stk: South-Central Papuan, Papua New Guinea) and Classical Nahuatl (nci: Uto-Aztecan, Aztec Empire);

8. IOL15.#1 (Milena Veneva): Birom (bom: Atlantic-Congo, Nigeria).

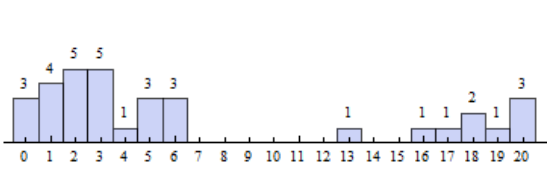
Table 1 presents the average scores for the instalments of IOL where these problems (in **boldface**) appeared, together with their rank within each set (from hardest to easiest *a posteriori*, that is, 1 labels the lowest and 5 the highest average score).[1] One can see that the problem on number

No.	IOL1	IOL3	IOL5	IOL7	IOL8	IOL10	IOL13	IOL15
# 1	14.85: 4	12.91: 5	11.80: 3	**14.77**: 5	15.49: 5	6.41: 2	**3.43**: 1	**7.66**: 3
# 2	**6.88**: 1	11.98: 2	14.17: 4	11.29: 4	**7.38**: 1	**7.69**: 3	5.78: 4	1.68: 1
# 3	11.56: 2	**10.66**: 4	3.43: 1	4.38: 2	14.29: 4	6.29: 1	5.51: 3	10.47: 4
# 4	15.24: 5	11.56: 3	**3.80**: 2	1.33: 1	9.55: 3	8.92: 4	10.43: 5	11.22: 5
# 5	14.06: 3	4.84: 1	14.62: 5	9.28: 3	9.43: 2	9.60: 5	3.57: 2	7.35: 2

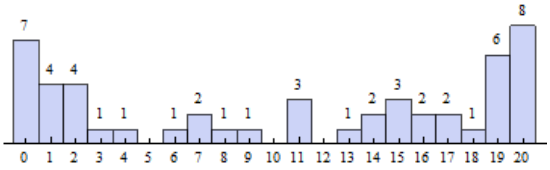
Table 1: IOLs with problems on number names: the average scores for all problems in the sets.

names has turned out to be the hardest one in its set three times out of eight, and the easiest only once.

The histograms in Figures 1, 2, 3, and 4 show the distribution of the points for each of the problems in question. One notices the many occasions when almost half of the participants scored zero. Problem #1 of IOL13 stands out as having been the hardest IOL problem on number names (its average score of 3.43 is the lowest one ever), and at the same time as the only problem of this type for which no solver got full score (20 points).

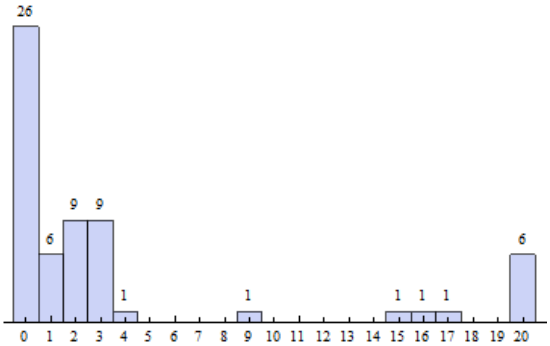


(a) Egyptian Arabic (IOL1.#2)

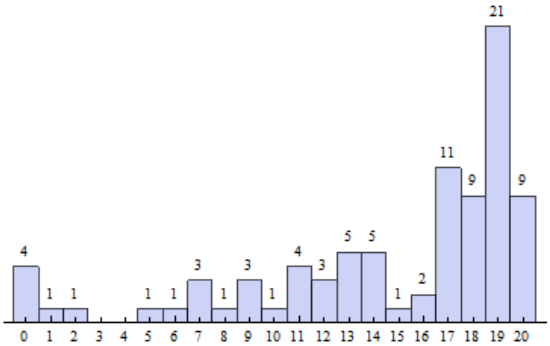


(b) Mansi (IOL3.#3)

Figure 1: The distribution of scores for the first and the second problem.



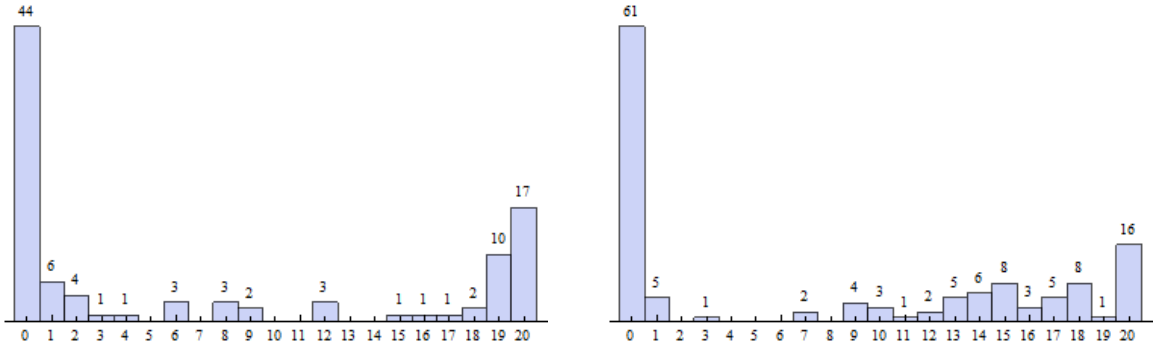
(a) Ndom (IOL5.#4)



(b) Sulka (IOL7.#1)

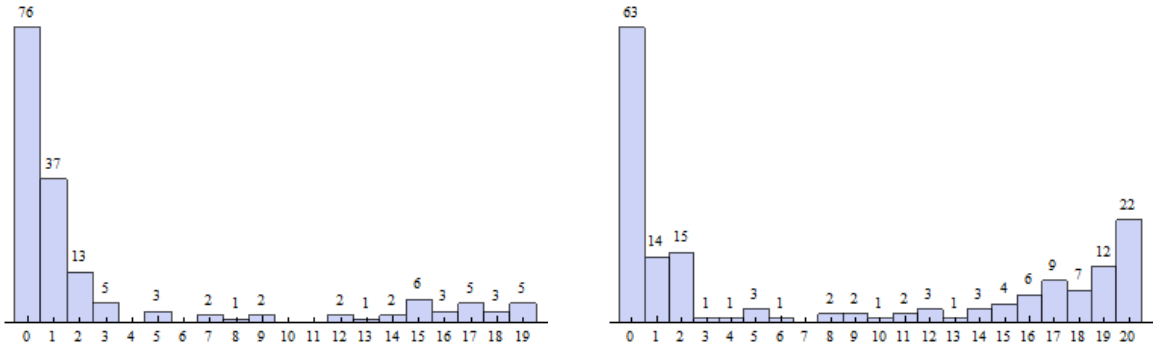
Figure 2: The distribution of scores for the third and the fourth problem.

[1] At IOL every problem carries a maximal score of 20 points; the only exceptions were at IOL1 (2003), where Problems #2 and #3 were worth 25 and 15 points respectively, but in this paper the scores have been normalised in the all-20 system for ease of comparison.



(a) Drehu (IOL8.#2) (b) Umbu-Ungu (IOL10.#2)

Figure 3: The distribution of scores for the fifth and the sixth problem.



(a) Arammba & Nahuatl (IOL13.#1) (b) Birom (IOL15.#1)

Figure 4: The distribution of scores for the seventh and the eighth problem.

The former claim, that about the pronouncedly mathematical character of problems on number names, is harder to defend. Although conspicuously absent in some languages,[2] number names are an integral part of language as numbers are of reality. It is true that languages have developed their systems of number names to different extents (depending on their speakers' need to count, which in turn depends on historical and social circumstances) and in different ways (Comrie, 2013), but the same can be said of any other domain of semantics. Arguably problems on number names are not significantly different from other Rosetta Stone or Chaos and Order[3] problems which illustrate the crosslinguistic variety in the verbalisation of concepts (Derzhanski, 2007), (Bozhanov and Derzhanski, 2013).

In fact, thanks to the universality and the discreteness of numbers, more uniformity can be expected in number names than in many other areas of meaning. Let us see what happens.

2. Numeral Systems

In the most basic scenario languages build number names in accordance with the polynomial formula

$$N = k_n B^n + k_{n-1} B^{n-1} + \cdots + k_1 B^1 + k_0 B^0,$$

[2]The most notorious (though controversial) case of a language with no numerals at all is Pirahã, spoken in Brazil. Some other (mostly Australian) languages seem not to get beyond 'one' and 'two', cf. Warlpiri *jinta* 'one', *jirrama* 'two', *panu* 'many (= three or more); all' (Bittner and Hale, 1995).)

[3]The terms 'Rosetta Stone' and 'Chaos and Order' for problems in which the glosses of the words, phrases or sentences of the unfamiliar language are presented respectively in order or out of order were introduced by Ivan Derzhanski in 2004 and gained currency within IOL's Problem Committee.

where B is the base of the number system (which is 10 most of the time[4]) and $k_0, k_1, \ldots, k_{n-1}, k_n$ are coefficients ($0 \leq k_i < B$). That is, a language tends to have underived names for the numbers from 1 to the base, and to express the other numbers through:

- **addition** of smaller numbers to the base (and its multiples), as in Turkish *on bir* 11 (lit. 'ten [plus] one'); zero terms are usually omitted, though cf. Chinese *bā qiān líng yī* 8001 (lit. 'eight thousand zero one').

- **multiplication** of the base (and its powers) by smaller numbers, as in Bulgarian *pet·deset* 50 (lit. 'five [times] ten'); 1 as a coefficient may or may not be explicit, cf. English *ten* but *one hundred*;

- **exponentiation**, which typically means having new underived words for the square, cube, etc. of the base, such as Italian *cento* $100 = 10^2$, *mille* $1000 = 10^3$.

The many number name systems based on the number 10 and on these three arithmetic operations vary in the details:

- the operations may be expressed by function words, inflexion or simply by juxtaposition, cf. Bulgarian *dvadeset i edno* 21 (lit. '20 and 1'), Hungarian *huszon·egy* 21 (lit. 'on 20 1', *huszon* being a modified superessive case form of *húsz* 20), and English *twenty-one*; not infrequently the sole indication of the operation is the order of its arguments, cf. Chinese *shí·sān* 13 (lit. '10 3'), *sān·shí* 30 (lit. '3 10'), *sān·shí·sān* 33;

- the grammatical expression of the operations may be motivated by the language's gender, number or case system, cf. Czech *sto* 100, *dvě stě* 200, *tři sta* 300, *pět set* 500 with the round number *sto* 'hundred' in different forms (respectively singular, [obsolete] dual, paucal = nominative plural, and partitive = genitive plural) as required by the coefficient, or Russian *dve tysjači* 2000 but *dva milliona* 2 000 000, respectively with a feminine and a masculine form of the coefficient;

- the order of the factors (coefficient and power of the base) in the terms is language-specific, cf. Hawai'ian *kana·kolu* 30, *kana·hā* 40;

- so is the order of the terms in the sum, cf. Malagasy *iraika amby roapolo sy telonjato* 321 (lit. 1 over 20 and 300), German *dreihundert·eins·und·zwanzig* 321 (lit. 300 1 and 20), English *three hundred and twenty-one*;

- there may be a lesser or greater amount of (morpho)phonological change, syncretism or suppletion, cf.
  - Bulgarian *šest·deset* 60, transparently composed of *šest* 6 and *deset* 10,
  - Irish *aon déag* 11 but *dó dhéag* 12 with intervocalic lenition that is very characteristic of the language,
  - Colloquial Bulgarian *šejset* 60 (as above, but somewhat opaque because of the contraction),
  - Hindi *bāwăn* 52 with no discernible relation to either *do* 2 or *păcās* 50, although historically it is compositional: *păcās* and *bāwăn* go back to Sanskrit *pañcāśat* and *dvā·pañcāśat*, respectively (Berger, 1992: 272);
  - Turkish *kırk* 40, altogether unrelated to *dört* 4 and *on* 10;

[4]Of the 196 languages surveyed in (Comrie, 2013), 125 have a decimal number system, and these include at least nine of the world's ten most spoken languages (French being the best-known one that breaks the pattern, though in a very limited way); decimal arithmetic also underlies the expression of numbers by Arabic numerals that are used worldwide.

Quite often a language switches to borrowed numerals beyond a certain threshold, which results in what looks like large-scale suppletion, e. g., Japanese *hitotsu* 1, *futatsu* 2, obsolete *hatachi* 20 (native) but *jū·ichi* 11, *jū·ni* 12, *ni·jū* 20 (Sino-Japanese).

Further, a language may use other operations to derive number names:

- **subtraction** (used for constructing numbers which are just a little smaller than the base or its multiples), as in Latin for numbers whose units are 8 or 9, e.g., *duo·de·viginti* 18 (lit. '2 [missing] from 20'); Hindi for numbers whose units are 9 (except for 89 and 99), e.g., *ŭn·tīs* 29, *tīs* 30; Finnish and Estonian *kahdeksan/kaheksan* 8 and *yhdeksän/üheksa* 9 are transparently related to *kaksi/kaks* 2 and *yksi/üks* 1,[5] though it is debatable whether the second half was originally a negative verb (Suihkonen, 2001) or a word for 10 borrowed from Iranian (Rätsep, 2003: 16);
- **overcounting** (meaning how many units are taken from the next multiple of the base), as in Finnish where the teens are constructed by adding *toista* (the partitive form of *toinen* 'second'),[6] or in Old Turkic, where the numbers in the range 11–89 are expressed as a number of ones from the next decade, e.g., *tört otuz* 24 (lit. 4 [from] 30);
- multiplication by **fractional coefficients**, specifically, by one half (Welsh *hanner cant* 50, lit. $\frac{1}{2}$ 100).

A significant minority of the world's languages use bases other than 10, although very few have had the historical opportunity to construct a full-scale system (that is, one that gets at least to the square of the base) using a single non-10 base. Far more common is the situation where two or more numbers share the duties of the base in different parts of the system. In particular, languages with base-20 systems regularly form the names for 11–19 by adding 1–9 to 10, and often multiply 20 only by 2–4, whereafter a underived word for 100 makes an appearance, as a result of contact with languages with base-10 systems (this is how Basque and Georgian work).

Some languages use different number systems for counting different items, although this tends to mean that there are groups of different sizes used for counting (as in English 10 may be called *five brace* when referring to game birds).

Finally, authors of linguistic problems often use various technical complications in order to make the problem harder or more interesting, such as composing it as a Chaos and Order rather than a Rosetta Stone, including arithmetic equalities with gaps instead of just number names and their corresponding numerical values, or using material from more than one unfamiliar language.

3. The Problems

Let us now go back to the eight problems on number names assigned at IOL to this day. Table 2 summarises the principal features of the languages' number systems and the problems, that is, the answers to the following questions:

1. Is the base of the number system 10, or 20, or 20 with supplementary bases (5, perhaps 10 and then perhaps 15), or something else?
2. Does the base have alternative (suppletive) names?
3. Are there any other numbers that play a base-like part in the number system?
4. Does the language use subtraction, or better, do the numbers just below the base behave – or are they formed – in an unusual way?

[5] Whether their present-day speakers are aware of this is another matter, amd whether it helps them detect the same phenomenon in another language is a third; the results of the marking of the only problem at IOL where a similar thing happens (IOL15.#1 on Birom) do not suggest that the Estonian contestants had an advantage.

[6] In older Finnish the same system worked with larger decades as well, but now this usage is considered archaic beyond 20.

5. Does the language use overcounting?
6. Are all arithmetic operations marked, or only some of them, or none?
7. What is the word order within the polynom (+) and within every summand (×)?
8. Are there any (morpho)phonological changes in the derivation of number names?
9. What other peculiarities of the language or the number system are there that make it harder to crack, ignoring more or less straightforward morphophonological processes?
10. Does the problem present the numeric values of the expressions in an unordered list?
11. Does the problem present equalities or equations in addition to, or instead of, just numbers?
12. What other peculiarities of the problem are there that make it harder to solve?

Also the table restates the average score and the ranking of each problem within its set (where 1 is hardest and 5 is easiest). These two values jointly motivate the ordering of the columns of the table.

problem	13.#1		1.#2	5.#4	8.#2	15.#1	10.#2	3.#3	7.#1		
language	stk	nci	arz	nqm	dhv	bom	ubu	mns	sua		
1: base	other	20+	10	other	20+	other	other	10	20	20	20+
2: other names	yes	yes	no	no	yes	yes	no	no	no		
3: other bases	no	no	no	yes	no	no	yes	no	yes	yes	no
4: subtraction	no	no	no	no	no	yes	no	yes	no		
5: overcounting	no	no	no	no	no	no	yes	yes	no		
6: operations	no	+	×1, ×2	+, ×2	+	+, ×	no	no	+, ×2		
7: word order											
(a): +	↘	↘	——	↘	↗, ↘ (1)	↘	↘	↘	↗		
(b): ×	↗	↗	——	↘	↗	↘	↘	↘	↘		
8: phonology	no	yes	no	no	yes	yes	no	no	no		
9: other	——	——	(2)	(3)	(4)	——	(5)	——	(6)		
10: disorder	no		no	yes	yes	no	no	no	no		
11: equalities	yes		yes	yes	yes	yes	no	no	no		
12: other	(7)		(8)	——	——	——	——	——	——		
score	3.43		6.88	3.80	7.38	7.66	7.69	10.66	14.77		
difficulty within the set	1		1	2	1	3	3	4	5		

Table 2: Linguistic phenomena in the IOL problems on number names.

Notes to Table 2:

(1) In Drehu the compound numbers in the intervals 6–9, 11–14, 16–19 are constructed as a sum of 5, 10, 15 and the remaining augend α, that augend coming first, while the compound numbers past 20 are formed as Γ *nge* Δ, where Γ is a multiple of 20, $1 \leq \Delta \leq 19$.

(2) As a typical Semitic language, Egyptian Arabic has a templatic (non-concatenative) morphology, which many contestants found impenetrable. Also it expresses multiplication by 2 through a dual number form (*tumn* $\frac{1}{8}$, *tumn·ēn* $\frac{2}{8}$, *talat·t itmān* $\frac{3}{8}$).

(3) Ndom operates a base-6 system, but 18 has a special name, which is then used to add smaller numbers to, so that 25 is not **mer an thonith abo sas* $6 \times 4 + 1$ but *tondor abo mer abo sas* $18 + 6 + 1$.

(4) Drehu expresses the numbers 5, 10 and 15 by one set of morphemes when units are added to them and in an entirely different way when that is not the case, e. g., *caa·pi* $5 = 1 \times 5$, *caa·ngömen* $6 = 1 + 5$, *lue·pi* $10 = 2 \times 5$, *lua·ko* $12 = 2 + 10$.

(5) Umbu-Ungu has special (unanalysable) names for all multiples of 4 (the secondary base beside 24, the primary one) up to 32, so that 56 is *tokapu polangipu* $24 + 32$ and 57 is *tokapu talu rurepo·nga telu* $24 \times 2 + 12 \neg 1$ (here '¬' stands for overcounting; $12 \neg 1$ is 9 because it means '1 from the 4 that ends at 12').

(6) The Sulka number systems comes in three varieties (for counting coconuts, breadfruit and everything else), which are all featured in this problem. Some of the nouns have suppletive singular and plural forms (e. g., sg. *tu*, pl. *sngu* 'yam'). There is also a dual number, although in this language it does not preclude the use of a numeral (*a tu a tgiang* '1 yam', *a lo tu a lomin* '2 yams', *o sngu a korlotge* '3 yams'; *lo* is the dual marker).

(7) Besides featuring two completely unrelated number systems from different languages in such a way that the two have to be untangled in parallel, the problem also employs bigger numbers than are usually dealt with in linguistic problems.

(8) The numbers in the problem are actually vulgar fractions.

Ordinal logistic regression (SPSS Inc., 2013) was conducted with the help of IBM SPSS Software (IBM Corp. Released 2017. IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.). This method is used to predict an ordinal dependent variable (response) given a set of independent variables (predictors), which can be factors (categorical predictors) or covariates (continuous predictors). The design of the ordinal regression in SPSS is based on (McCullagh, 1980). The independent variables which have a statistically significant effect on the dependent variable were determined ($p < 0.001$). The following observations and conclusions can be made on the basis of the results:

1. Surprisingly, the extremes of the ranking are the two problems in which there is more than one number system, but Arammba–Nahuatl with its two unrelated languages is hardest and Sulka with its three counting methods is easiest.

2. The base of the number system (10, 20, or other) has no direct bearing on the difficulty of the problem.

3. The existence of an alternative name of the base makes a problem rather more difficult. The same is true for the phonological changes.

4. Neither subtraction nor overcounting make a problem difficult. Nor does the existence of an auxiliary base.

5. However, the explicit marking of the arithmetic operations does increase the difficulty.

6. A problem with different word orders in the polynom and in every summand is difficult.

7. The Chaos and Order format makes a problem harder, but not so much as large integers or vulgar fractions.

8. All problems that involve equalities prove harder than all problems that do not.

4. Conclusions

Our research, based on the results of IOL 1–15, has revealed some ways (mostly unexpected ones) in which the difficulty of a problem on number names depends on its structure and the kinds of linguistic phenomena featured in it. It would be interesting to conduct similar studies on the results of other linguistic olympiads and contests which are old long enough to have accumulated a statistically useful pool of problems on number names, and to compare the findings, which may shed further light on the effect of working languages.

References

Berger, H. (1992). Modern Indo-Aryan. In Gvozdanović, J., Ed., *Indo-European Numerals*, pages 243–287. Mouton de Gruyter, Berlin, New York.

Bittner, M. and Hale, K. (1995). Remarks on Definiteness in Warlpiri. In Bach, E., Jelinek, E., Kratzer, A., and Partee, B. B. H., Eds., *Quantification in Natural Languages*, pages 81–105. Springer Netherlands, Dordrecht.

Bozhanov, B. and Derzhanski, I. (2013). Rosetta Stone Linguistic Problems. *Proceedings of the Fourth Workshop on Teaching Natural Language Processing*, pages 1–8.

Comrie, B. (2013). Numeral Bases. In Dryer, M. S. and Haspelmath, M., Eds., *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig. `http://wals.info/chapter/131`.

Derzhanski, I. (2007). Mathematics in Linguistic Problems. In Dimitrova, L. and Pavlov, R., Eds., *Mathematical and Computational Linguistics. Jubilee International Conference, 6 July 2007, Sofia*, pages 49–52.

Derzhanski, I. (2009). *Linguistic Magic and Mystery*. Union of Bulgarian Mathematicians, Sofia.

McCullagh, P. (1980). Regression Models for Ordinal Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 109–142.

Rätsep, H. (2003). Arvsõnade päritolust eesti keeles. *Oma Keel*, 2:11–18.

SPSS Inc. (2013). *IBM SPSS Statistics V22.0.0 Documentation*. IBM SPSS Statistics Base 22. SPSS Inc., Chicago IL.

Suihkonen, P. (2001). Suomen ja sen sukukielten lukusanoista. *Matematiikkalehti Solmu*, 2.