Glottometrics 48, 2020, 9-16

Concept Realization in Texts

Kateřina Pelegrinová[1], Gabriel Altmann

Abstract. The study proposes a model for the rank-frequencies of individual semantic classes of nouns, verbs, adjectives and adverbs in three English novels analyzed by N. Yesypenko (2009), and compares the ranks using the Kendall-test.

Keywords: *English texts, semantic classes, Menzerath's law, Kendall's test*

In 2009, Nadia Yesypenko wrote a long study concerning concept realization in English texts. She used three novels, namely E. Waugh, *A Handful of Dust*, J. Swift, *Gulliver's Travels* and M. Twain, *The Adventures of Tom Sawyer*. Her aim was to find classes of concepts for nouns, verbs, adjectives, and adverbs. Needless to say, one can find different classifications and one need not be unique. Other researchers can develop other types of classification. Her aim was to find a smaller number of conceptual classes, i.e. semantically more general classes. The procedure is used in all sciences and serves our orientation, analysis, derivation of laws, etc. For nouns she defined 25 classes, for verbs 27, for adjectives 18, and for adverbs 7. Automatically the question arises whether the frequencies in individual texts are similar, i.e. if the ranks are approximately similar. This automatically leads to the question whether one can rank the frequencies in individual texts and find a function/distribution capturing satisfactorily the ranking. Here we shall analyze both problems using Yesypenko's numbers.

In order not to repeat her tables, we merely enumerate the categories she found. For nouns: 1. Appearance/parts of body, 2. Feelings/emotions, 3. Proper names/ nicknames, 4. Establishments/groupings, 5. Deseases/defects, 6. General notions of people/mythical characters, 7. Devices/articles of furniture, 8. Abstract notions, 9. Food/meals, 10. Weight/length/volume, 11. Sound/fragrance/temperature/light, 12. Wildlife/celestial objects, 13. Actions/changes/movement, 14. Time, 15. Clothes, 16. Shape/structure, 17. Speech, 18. Building/premises, 19. Profession, 20. Materials/ liquids, 21. Vehicles, 22, Geographical notions, 23. Weapons, 24. Events/holidays, 25. Other notions.

For verbs: 1. Verbs of motion/removing, 2. Verbs of process, change, development, 3. Verbs of beginning/end of action, 4. Verbs of physical action, 5. Engender verbs, 6. Destroy verbs, 7. Verbs of successful/unsuccessful action implementation, 8. Verbs of attempt, 9. Verbs of sound emission, 10. Verbs of light phenomena, 11. Verbs of temperature phenomena, 12. Verbs of nature phenomena, 13. Verbs of communication, 14. Verbs of moral impact/effect, 15. Verbs of social activity, 16. Position verbs, 17. Verbs of existence, 18. Modality verbs, 19. Verbs of human relations, 20. Verbs of reference, 21. Verbs of emotional psychological impact, 22.Verbs of ownership/loss, 23. Verbs of psychological state, 24. Verbs of perception, 25. Verbs of mental activity, 26. Verbs of subjective assessment, 27. Verbs of emotional psychological state.

For adjectives: 1. Traits of character/emotions, 2. Physical/natural condition, 3. Intellectual capacity, 4. Appearance, 5. Senses, 6. Age/time, 7. Temperature/sound, 8. Shape/size, 9. Flavour, 10. Weight, 11. Degree/intensity, 12. Color, 13. Actions done to the object, 14.

[1] University of Ostrava; Czech Republic, email: pelegrinovak@gmail.com.

Positive evaluation, 15. Evaluation of length/distance/position of the object, 16. Evaluation of value/function of the object, 17. Material, 18. Negative evaluation.

For adverbs: 1. Adverbs of time, 2. Adverbs of repetition and frequency, 3. Adverbs of place and direction, 4. Adverbs of condition and consequence, 5.Adverbs of manner, 6.Adverbs of degree and quantity, 7. Question adverbs.

It is important to adhere to the given text because words may be polysemic and in one context they may have one meaning, in another, another one.

In other publications and other languages one finds different classifications but here we want to test whether the given classification can be modeled and compared. If we rank the frequencies in individual texts, we obtain the results for nouns as presented in Table 1. Here we apply the Menzerathian function defined as

$$y = ax^b \exp(-cx)$$

where *y* are the frequencies and *x* are the ranks. There are, of course, many other functions that would satisfactorily capture the frequencies of ranks but we use here the Menzerathian function because here the nouns (and other parts of speech) are members of some greater classes, and for these purposes the Menzerathian function has been frequently used.

Table 1
Ranking of noun classes

E. Waugh, *A Handful of Dust*				J. Swift, *Gulliver's Travels*			
Class	Rank	Frequency	Menzerath	Class	Rank	Frequency	Menzerath
3	1	281	268.60	18	1	222	214.49
6	2	133	173.32	8	2	154	166.50
14	3	119	127.51	6	3	138	137.47
18	4	118	100.26	12	4	102	116.44
7	5	102	82.13	13	5	102	99.95
8	6	88	69.19	7	6	92	86.57
12	7	70	59.50	1	7	85	75.44
13	8	70	51.99	22	8	68	66.03
17	9	45	46.00	2	9	62	57.98
1	10	45	41.11	14	10	62	51.06
2	11	35	37.06	10	11	54	45.06
22	12	33	33.65	19	12	36	39.83
9	13	30	30.75	21	13	34	35.27
15	14	24	28.25	4	14	32	31.26
21	15	24	26.07	17	15	32	27.74
19	16	21	24.17	9	16	26	24.65
25	17	15	22.48	15	17	20	21.91
10	18	12	20.99	5	18	18	19.49
20	19	12	19.65	20	19	12	17.36
24	20	12	18.45	24	20	10	15.46
4	21	10	17.37	3	21	6	13.78
5	22	8	16.38	25	22	6	12.29
23	23	8	15.49	16	23	4	10.96
11	24	7	14.67	23	24	4	9.78
16	25	3	13.93	11	25	2	8.74
a = 268.6017, b = -0.5532, c = 0.1138, R^2 = 0.9566				a = 238.1384, b = -0.2145, c = 0.1046 R^2 = 0.9851			

M. Twain, *The Adventures of Tom Sawyer*			
Class	Rank	Frequency	Menzerath
3	1	142	140.28
6	2	128	135.93
8	3	122	125.40
18	4	112	113.38
7	5	106	101.38
1	6	100	90.03
14	7	86	79.56
12	8	80	70.06
13	9	70	61.53
17	10	54	53.92
2	11	50	47.17
15	12	28	41.21
10	13	24	35.96
19	14	24	31.34
11	15	22	27.30
20	16	20	23.76
22	17	20	20.66
4	18	18	17.96
9	19	10	15.60
21	20	10	13.55
5	21	6	11.76
25	22	6	10.20
16	23	4	8.85
23	24	2	7.67
24	25	2	6.65
a = 162.9765, b = 0.1710, c = 0.1500 R^2 = 0.9803			

In the verb and other classes, we omit those that do not occur at all, i.e., they have the frequency 0. We obtain the results presented in Tables 2 to 4.

Table 2
Ranking of verb classes

E. Waugh, *A Handful of Dust*				J. Swift, *Gulliver's Travels*			
Class	Rank	Frequency	Menzerath	Class	Rank	Frequency	Menzerath
17	1	271	263.89	17	1	174	178.65
1	2	145	169.62	1	2	148	133.15
13	3	131	126.42	22	3	104	107.37
4	4	113	100.11	4	4	84	89.43
18	5	91	81.96	18	5	70	75.81
25	6	66	68.52	25	6	66	64.99
22	7	58	58.12	24	7	52	56.14
24	8	50	49.83	20	8	50	48.78
14	9	41	43.06	5	9	48	42.56
27	10	39	37.46	23	10	32	37.26
5	11	31	32.75	13	11	30	32.72

3	12	30	28.75	16	12	30	28.79
2	13	24	25.34	2	13	28	25.38
15	14	23	22.40	14	14	24	22.41
20	15	23	19.85	6	15	22	19.82
26	16	18	17.63	27	16	22	17.55
19	17	17	15.69	15	17	22	15.55
16	18	16	13.99	3	18	18	13.80
6	19	15	12.50	19	19	18	12.25
21	20	9	11.18	26	20	8	10.89
23	21	7	10.01	21	21	6	9.68
12	22	4	8.97	8	22	4	8.61
8	23	3	8.05	10	23	2	7.67
9	24	3	7.23	11	24	2	6.83
10	25	3	6.51	7	25		
7	26	2	5.85	9	26		
				12	27		
a = 287.3766, b = 0.5146, c = 0.0853 R^2 = 0.9883				a = 198.2463, b = -0.2740, c = 0.1041 R^2 = 0.9864			

M. Twain, *The Adventures of Tom Sawyer*			
Class	Rank	Frequency	Menzerath
1	1	202	211.72
17	2	184	162.81
4	3	132	132.30
13	4	122	109.96
24	5	78	92.54
25	6	68	78.49
22	7	58	66.93
27	8	52	57.30
16	9	44	49.21
18	10	42	42.36
3	11	36	36.53
5	12	30	31.55
15	13	30	27.29
20	14	30	23.64
2	15	26	20.49
21	16	26	17.77
6	17	20	15.43
14	18	20	13.41
19	19	20	11.65
23	20	12	10.14
8	21	8	8.82
26	22	8	7.68
7	23	6	6.69
9	24	6	5.83
12	25	4	5.08
10	26	2	4.43
a = 241.0447, b = -0.1919, c = 0.1297 R^2 = 0.9802			

Table 3
Ranking of adjective classes

E. Waugh, *A Handful of Dust*				J. Swift, *Gulliver's Travels*			
Class	Rank	Frequency	Menzerath	Class	Rank	Frequency	Menzerath
2	1	122	124.00	8	1	120	114.66
14	2	118	111.24	11	2	99	106.21
11	3	100	96.38	16	3	93	94.00
16	4	68	82.32	13	4	75	81.66
13	5	67	69.77	2	5	66	70.22
1	6	59	58.84	14	6	63	50.00
8	7	48	49.45	15	7	60	51.03
18	8	45	41.46	1	8	57	43.26
15	9	44	34.70	18	9	54	36.59
6	10	42	28.99	6	10	33	30.88
4	11	22	24.19	12	11	18	26.02
12	12	22	20.17	4	12	12	21.90
17	13	19	16.80	17	13	6	18.41
3	14	7	13.99	7	14	3	15.46
7	15	6	11.64				
5	16	2	9.67				
9	17	2	8.04				
a = 150.3096, b = 0.1208, c = 0.1924 R^2 = 0.9692				a = 138.1300, b = 0.15823, c = 0.1862 R^2 = 0.9294			

M. Twain, *The adventures of Tom Sawyer*			
Class	Rank	Frequency	Menzerath
13	1	93	89.00
8	2	87	91.12
16	3	81	85.28
1	4	75	76.94
18	5	69	68.03
14	6	63	59.40
15	7	54	51.40
2	8	48	44.20
6	9	42	37.82
11	10	42	32.23
17	11	21	27.39
7	12	18	23.21
12	13	12	19.63
4	14	12	16.56
a = 107.9052, b = 0.3119 , c = 0.1827, R^2 = 0.9657			

If the resulting theoretical function yields a parabolic course, it is recommended to decrease the parameter *c* because *c* is related to the initial value given by the language and by the text. If we consider the Menzerathian function in its differential form, we obtain

$$\frac{dy}{y} = \left(\frac{b}{x} - c\right) dx$$

where *c* is the language and text constant and *b/x* is the requirement of the author, of the reader, and also of the text type. Needless to say, even if we a priori change another parameter, we can make the Menzerathian function monotonically decreasing. This is the case in the last text (by Twain) concerning adjectives caused by the similarity of the first two frequencies. The deviation from monotonic decreasing can easily be corrected if we fix the parameter *a,* e.g., to a = 120 and perform the fitting iteratively.

Table 4
Ranking of adverbial classes

E. Waugh, *A Handful of Dust*				J. Swift *Gulliver's Travels*				M.Twain, *The Adventures of Tom Sawyer*			
Class	Rank	Frequ	Menz	Class	Rank	Frequ	Menz	Class	Rank	Frequ	Menz
6	1	215	212.22	5	1	135	130.37	5	1	111	113.17
5	2	153	166.40	6	2	111	127.83	6	2	111	112.37
1	3	141	119.52	1	3	108	89.24	1	3	96	90.93
2	4	66	82.82	3	4	63	54.22	3	4	75	67.68
3	5	66	56.28	2	5	18	30.53	2	5	54	48.11
7	6	49	37.77	4	6	3	16.38	4	6	15	33.22
4	7	5	25.13					7	7	12	22.48
a = 334.2977, b = 0.3046, c = 0.4544, R^2 = 0.9508				a = 301.3561, b = 1.1802, c = 0.8378 , R^2 = 0.9265				a = 186.5793, b = 0.7112, c = 0.5000, R^2 = 0.9473			

With adverbs we obtain also a parabolic function in case of the text by Twain. Again, we change a priori the parameter *a* and obtain the monotonically decreasing function. The case of changing a priori the optimization by fixing some parameters must be taken into account with each computation. If one compares the parameters *b* and *c* one can state the increase of *c* in dependence of increase of *b*, however, taking all data, it is not monotonic, though increasing. One can see the course in Figure 1: (c = f(b))

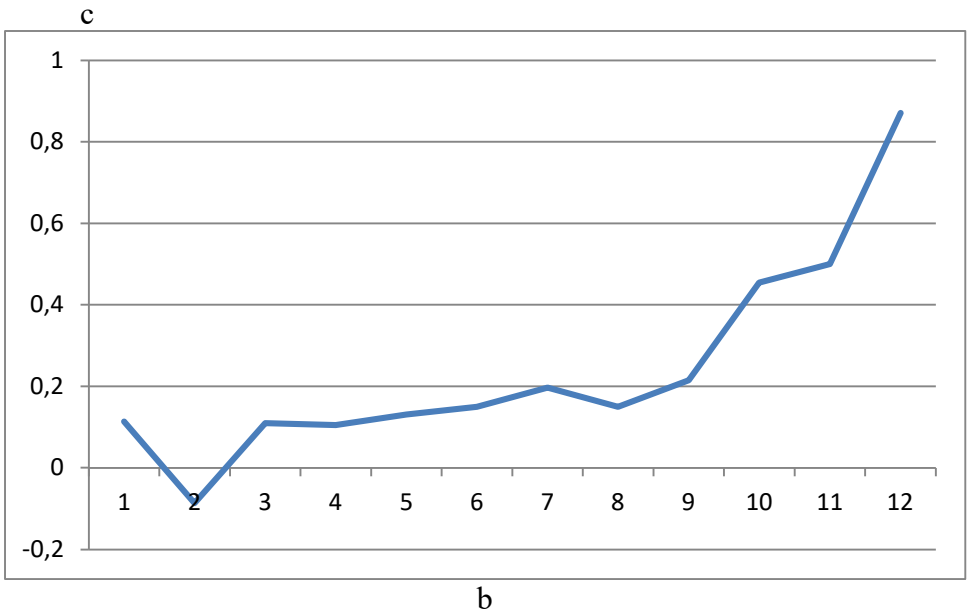

Figure 1. The course of parameter *c* in dependence of *b*

Evidently a number of data is necessary – taking also prepositions, conjunctions, etc. – also in other languages – in order to obtain a background theory.

Usually, one tests the similarity of two distributions by using a chi-square test. But in many cases, the results are not quite reliable because chi-square increases with increasing sample size; it is not adequate if the observed values are smaller than 5 (or 1), hence some classes must be pooled, etc. For this reason we use rather the non-parametric Kendall-test considering the ranks of classes which can easily be stated. For giving an example we use the adverbs in the three English texts and present first Yesypenko's Table 4. There are seven classes in each text. We write the ranks for each text separately. If two (or more) frequencies are equal, we take the mean of the ranks.

Table 5
Yesypenko's Table 4
Frequency of the lexical semantic groups of adverbs

Lexical semantic word group	*A Handful of Dust*	*Gulliver's Travels*	*The Adventures of Tom Sawyer*
Adverbs of time	3	3	3
Adverbs of repetition and frequency	4.5	5	5
Adverbs of place and direction	4.5	4	4
Adverbs of condition and consequence	7	6	6
Adverbs of manner	2	1	1.5
Adverbs of degree and quantity	1	2	1.5
Question adverbs	6	7	7

In order to express our view that the texts are very similar in using adverbials, we test the data using Kendall's test given by the formula

$$W = \frac{12QSR}{m^2(N^3 - N) - m\sum_{j=1}^{m} V_j}$$

where m is the number of texts (here 3.), N is the number of adverbial classes /categories (here 7), T_i is the sum of the i^{th} row (sum of ranks of an adverbial class),

$$QSR = \sum_{i=1}^{N}(T_i - \bar{T})^2 \quad = \sum_{i=1}^{N} T_i^2 - \frac{\left(\sum_{i=1}^{N} T\right)^2}{N}$$

is the square of the deviations of the row sums from their mean. Since we take ties of ranks into consideration (here, they occur only 2 times), we compute for them

$$V_j = \sum_{h=1}^{S_j}(v_k^3 - v_k)$$

where S_j is the number of ties in the given text. One can obtain the chi-square as

$$X^2 = \frac{12 * QSR}{mN(N+1) - \frac{1}{N-1}\sum_{j=1}^{m} V_j},$$

or, having computed *W*, one takes $X^2 = m(N-1)W$, with *N*-1 DF. For the adverbs we obtain X^2 = 12(241)/(3*7*8 – 12/7) = 17,39. The number of degrees of freedom is 7-1 = 6. Since the probability of such a chi-square is near to 0.005, the difference can be considered as significant, i.e. in the texts, the adverb classes are used with different probability.

Doing the same for nouns, verbs and adjectives, we obtain the values

X^2 (nouns) = 59.65 with DF = 24, yielding P < 0.0005
X^2 (verbs) = 325.39 with DF = 26 yielding P << 0.0005
X^2 (adjectives) = 46.1648 with DF = 17 yielding P << 0.0005

As can be seen, the semantic classes are distributed differently in the three English texts. This may be caused by the difference of theme, by the style of the author but also by the choice of the semantic classes by N. Yesypenko, by the attribution of the words to different classes, etc. The belonging to a semantic class is not only a problem of the given word but also the problem of its environment. A number of investigations is necessary in order to obtain clearly interpretable results. Here we wanted merely to show some possible evaluation methods. Needless to say, the similarity of frequencies can be compared also by the usual chi-square test but in that case, the frequencies of the same individual classes must be compared.

References

Pelegrinová, K., Altmann, G. (2017). The study of adverbials in Czech. *Glottometrics 34, 2017, 34–53*

Yesypenko, N. (2009). An integral qualitative-quantitative approach to the study of concept realization in the text. In: Kelih, E., Levickij. V., Altmann, G. (eds.). *Methods of Text Analysis.* Chernivci, ChNU: *308–327.*