By Dorota Lewandowska

Statistical structure of press information lexis (based on the example of news from the 1960s and present day) [1]

Key words: press information, frequency and classification of words, concentration, richness, originality, vocabulary stereotypes

Abstract: This article compares two samples of press information texts, one from the 1960s and the other from the turn of the century. It analyses vocabulary frequency, its concentration, richness, originality and stereotypes. Compared was also data on the frequency of words within different grammar classes. Based on the research, it can be concluded that the language used in press articles has not considerably changed over time. The only differences noted were in the proportions of inflection classes making up lexis resources of the two samples.

This article is a comparative analysis of vocabulary frequency in selected press information texts. It compares two equal size samples (100 000 segments) of texts from short press articles, one sample is from press texts between 1963-1967 and the other between 1998-2002. The first sample was chosen by the authors[2] of the 2nd volume of "Słownictwo współczesnego języka polskiego. Listy frekwencyjne"[3] (SWJP) selected from all dailies published during that time period, the second sample was selected via the same method[4] from among short press information texts[5].

The goal of the analysis was to determine whether any significant changes in the vocabulary used have taken place over the years. In order to do that, a vocabulary frequency analysis was carried out between the two samples, including aspects such as vocabulary concentration, richness, originality and stereotypicality. Secondly, compared was also data on the frequency of words within different flexeme[6] and grammar classes[7].

Vocabulary concentration

Vocabulary concentration is determined by calculating what kind of words have the highest frequency and is carried out in order to decide on the lexis richness of any given text.

Vocabulary concentration is calculated according to Mistrik's formula: $I_{konc.} = \frac{20v_{i>1}}{N}$,

where $v_{i>1}$ is the number of flexemes with a frequency greater than 1 and N – is the length of the text (sample size). It illustrates the amount of thematic vocabulary within a text, interpreted directly proportionally to the degree of text wealth by the value given. A low value

[1] This article is dedicated to Halina Satkiewicz for her birthday and anniversary of her work for UW.

[2] Ida Kurcz, Andrzej Lewicki, Jadwiga Sambor, Jerzy Woronczak.

[3] Warszawa 1974. Data obtained based on analysis of materials included in vol. 2 of SWJP have become also part of "Słownik frekwencyjny polszczyzny współczesnej" (SFPW), published in 1990.

[4] For more on principles and choice method for representative sample see SWJP, vol. 2., p. 2-10.

[5] For more on press information style indicators see SWJP, vol. 2., p. 7-8 and Lewandowska D., "*Drobna wiadomość prasowa" jako informacyjny gatunek dziennikarski*, "Studia Medioznawcze" 2004, nr 2.

[6] The proposed division into flexeme classes is a result of methodological assumptions for broader research on press information language. A flexeme is the basic unit of analysis. It is more narrow that a lexeme as in comparison to it, additional here is the grammar class identity. It means that two forms belong to the same flexeme only if they mean the same thing, have similar morphologic make up and are in the same grammar class. For example, *napiszę, napiszemy, napiszecie* forms are one flexeme but not *napisać* or *napisano*, as they do not have gender or number. Flexeme classes are groups of flexemes, more detailed and more narrowly defined than traditional parts of speech. Hence, there are more of them, 30. (Cf. Przepiórkowski A., *Korpus IPI PAN. Wersja wstępna*, Warszawa 2004., p. 25-26, 29). This article also converts the analysed results to grammar classes (or parts of speech), in bolded letters (ie. VERBS).

[7] Determined based on *f*, frequency of each flexeme and its relation to average frequency of flexeme in the analysed samples.

means poor or uniform text theme, while a wider theme indicates a lower concentration of vocabulary, or a higher $I_{konc.}$[8] value.

This analysis uses a modified version of the above formula, $I_{konc.}= \frac{20v_{i>3}}{N}$, proposed by J. Sambor[9], which takes into account vocabulary f>3[10]. Using this formula seems more appropriate as far as the classification of lexeme frequency classes, according to which words with a frequency of f≤3 are considered very rare.

As mentioned above, a high degree of concentration means poor vocabulary, while a low value means a majority of words with low frequencies, indicating greater vocabulary wealth. Regarding the two samples of interest, vocabulary concentration is as follows, for sample 1 – 15.23 and for sample II – 15.44 which means only a slightly higher vocabulary wealth for press information between 1998-2002.

Vocabulary concentration is supplemented by a percentage evaluation of lexis, with a division into different frequency classes, which will be discussed in the latter part of the article.

Vocabulary wealth

Besides the discussed above vocabulary concentration, text wealth is also measured by the number of different words a text contains. When the two samples were compared, sample II had a slightly greater (by about 8%) vocabulary base[11] for the analysed texts[12].

In order to measure this, an average word frequency per text was calculated. For sample I, the estimated value was 6.6 and for sample II – 6.0. Knowing that vocabulary wealth is inversely proportional to the number of words of high frequency[13], we can deduct that press information texts between 1998-2002 are more diverse than those from the 1960s. Sample II possesses greater vocabulary wealth as it contains more words with frequencies: f=1, f=2 and f=3[14]. The relation is, the more flexemes with lower frequencies, the richer or more original the lexis of a sample[15].

Besides average frequency and total number of words (flexemes) in a given text, other frequency indicators of lexis wealth are:

P. Guiraud's : $\frac{W}{\sqrt{2N}}$, W. Kuraszkiewicz's: $\frac{W}{\sqrt{N}}$ and J. Mistrík's: $\frac{20W}{N}$ formulas.

All of the above formulas determine the relationship of W (lexicon size) to N[16] (sample size), meaning that the interpretation of results obtained, despite the numerical differences, is based on the same assumption – the greater the index number, the greater vocabulary wealth. Table 1 presents the values of the different frequency indicators.

Sample Indicators

---

[8] Cf. Kamińska-Szmaj I., *Różnice leksykalne między stylami polszczyzny pisanej. Analiza statystyczna na materiale słownika frekwencyjnego*, Wrocław 1990, p. 19.

[9] Words with f=2 and f=3 frequencies are not considered thematic (Cf. Sambor J., *Słowa i liczby. Zagadnienia językoznawstwa statystycznego*, Wrocław 1972, p. 223).

[10] Sample I lexicon consists of 3681 flexemes at f>3, sample II lexicon – 3845. The percentage make up of this group in sample I is 23.9% and in sample II – 22.8%.

[11] Materials from 1998-2002.

[12] Number values are for sample I – 15445; for sample II – 16864, including punctuation flexemes: 12 and 15.

[13] The more words with high frequency, the poorer the vocabulary.

[14] Sample I is 76,1% of lexicon, sample II – 77,2%. Of course, from a statistical point of view, these differences are insignificant.

[15] Cf.: Łojek M., op. cit., p. 37-39.

[16] As far as real N value (without punctuation), sample I – 103127, sample II – 102864. W values (without punctuation), sample I – 15433, sample II – 16849.

	Guiraud's	Kuraszkiewicz's	Mistrík's
1963-1967	33.98	48.06	2.99
1998-2002	37.17	52.53	3.27

Table 1: Values of the different frequency indicators for two text samples (I – 1963-1967 materials, II – 1998-2002 materials)

The above values confirm the observations on information text vocabulary wealth, however, due to such slight differences in the values obtained, we cannot state that the variations are significant when it comes to language transformations.

Vocabulary originality

There is a direct relationship between the number of words (here flexemes) with low frequencies and the degree of vocabulary originality (diversity). The deciding factor here are words with frequencies f=1, f=2 and f=3 in the lexicon of the researched text[17].

The indicators used to calculate the degree of vocabulary originality in quantitative linguistics are:

P. Guiraud's: $\frac{W_1}{W}$ and J. Mistrík's: $\frac{20W_1}{N}$, which measure frequency f=1[18] in lexis (Guiraud's formula) or in the text (Mistrík's formula). The greater the value of either indicator, the greater the vocabulary originality. Generally, Guiraud's indicator is considered more objective[19], since its value is not dependent on text length which allows for a comparison of different sample sizes. Below given are the values obtained for both samples.

Sample	indicator	
	Guiraud's	Mistrík's
1963-1967	0.531	1.59
1998-2002	0.534	1.75

Table 2: Values for vocabulary originality for the two text samples (I – 1963-1967 materials, II – 1998-2002 materials).

The above data shows only a slightly greater originality of texts in press articles between 1998-2002. From the statistical linguistics point of view, it cannot be considered that a text more original as far a vocabulary is, at the same time, less stereotypical. This is due to the fact that, according to Mistrik's formula, stereotypicality is calculated by measuring the average frequency of words that are repeated in a text, or with frequency f>1[20]: $I_{stereot.}=\frac{N-W_1}{W_{i>1}}$.

The greater the value, the greater the lexicon stereotypicality. Therefore, the relationship between value and lexicon stereotypicality is directly proportional and inversely proportional to vocabulary wealth. These values for the two samples are as follows: I – 13,1; II – 12.

The above illustrates that, according to Mistrik's indicator, in more modern press texts (sample II) the vocabulary used was somewhat more diverse and original. It should also be noted that the differences between the values obtained are not significantly big.

[17] This was discussed earlier on in the article.
[18] Number of flexemes at f=1 in sample I is 8188, in sample II – 8999.
[19] Cf. m.in. Łojek M., op. cit., p. 39
[20] Number of flexemes at f>1 in sample I – 7245, and In sample II – 7850.

Frequency classes[21]

Exact and objective delineation of a boarderline between common vocabulary (SC) and rare (SR) is not possible since determining such frequency depends on the text length and the accepted convention[22]. For the purposes of this research, Kuraszkiewicz's[23] method was used, later modified by J. Sambor,[24] which calculates average frequency of words appearing in the entire text sample (separately for sample I and for sample II) and indicates different frequency classes for common (and very common) and rare (including very rare) vocabulary, not including punctuation flexemes.

In order to determine the average value of flexeme frequency in a sample ($f_{śr}$), the following formula was applied: $f_{śr} = \frac{N}{W}$, with N being the sample size and W – the number of flexemes in the sample[25].

For both sample groups the average $f_{śr}$ was calculated twice – the sample size was estimated at N=100,000[26], while factual N was 116,804 for sample I and 118,822[27] for sample II, not including punctuation[28]. The number of non-punctuation flexemes for sample I is 15,433 and for sample II – 16,849. Hence, the calculated average, $f_{śr,}$ for sample I is 6.5 (N=100,000) and 6.7 (N=116,804), and for sample II – 5.9 (N=100,000) and 6.1 (N=102864). As a result the following ranges were determined for common and rare vocabulary[29]:

- common vocabulary (SC) – flexemes $f \geq 7$,
- rare vocabulary (SR) – flexemes $f < 7$.

After the calculations were made for both samples, it can be seen that in both samples common flexemes ($f \geq 7$) comprise about 14% of the entire lexicon. In sample II, the SC value is slightly lower (13.2%) than in sample I (14.4%). The details are included in Table 3.

Frequency range	**number of flexemes**		**%**[30]	
	I	**II**	**I**	**II**
SC $f \geq 7$	2,224	2,225	14.4	13.2
SR $f < 7$	13,209	14,624	85.6	86.8
Razem SC+SR	**15,433**	**16,849**	**100**	**100**

Table 3: Non punctuation flexemes divided into frequency ranges, including percentage calculations for sample I (1963-1967) and sample II (1998-2002).

As can be seen, the compared samples differ only slightly regarding the proportions of common and rare vocabulary. This proves a relative stability as far as lexis use within the

[21] The term was used from M. Łojek's article, *Słownictwo homilii w świetle badań statystycznych*, "Poradnik Językowy" 2001, nr 6, p. 32-48.
[22] Cf.: Sambor J., *O słownictwie statystycznie rzadkim. Na materiale derywatów we współczesnej publicystyce polskiej*, Warszawa 1975, p. 9.
[23] See: Kuraszkiewicz W., *Rzeczowniki w „Wizerunku" Mikołaja Reja*, "Pamiętnik Literacki" 1969, nr 4.
[24] See: Sambor J., op. cit., 1975, p.9-10.
[25] Cf: Ibidem.
[26] In accordance with our assumptions, this number of words (excluding punctuation) was considered a representative sample of texts written in Polish. (Cf. SFPW, vol. 1., p. xii).
[27] These sizes reflect the number of signs during automatic morho-syntactic analysis of the samples, including punctuation (not considered in SWJP) plus aglutynants (not considered by SWJP as separate language units, cf. Przepiórkowski A., op. cit., p. 18-21).
[28] In sample I the number of punctuation signs is 13677, in sample – 15959.
[29] Cf. Sambor J., op. cit., 1975, p. 9.
[30] This column contains percentages of common and rare vocabulary within entire samples. 100% is the number of flexemes within the sample, for sample I – 15433, and for sample II – 16849.

different flexeme frequency classes and confirms the earlier observation that modern press information texts have only a slightly more diverse and original vocabulary in comparison to those from the 1960s. The next step, therefore, is to examine the grammar structures of selected lexicon groups giving us more detailed information which will allow us to make final judgments regarding the possible language changes that have occurred over this time period.

In order to do that a numerical comparison of the different flexeme classes was performed for both samples (Table 4).

Flexeme class		**number of flexemes for the given flexeme class**		***f*≥7**		***f*<7**		**% of flexemes with a frequency of *f*≥7 for the given flexeme class**[31]	
		I	**II**	**I**	**II**	**I**	**II**	**I**	**II**
NOUNS	**subst**[32]	7877	8355	1208	1220	6669	7135	15.3	14.6
	depr	1	3	0	0	1	3	0	0
	ger	671	701	48	32	623	669	7.2	4.6
NUMERALS	**num**	49	56	44	45	5	11	89.8	80.4
	numcol	4	13	0	0	4	13	0	0
ADJECTIVES	**adj**	2462	2635	412	368	2050	2267	16.7	14.0
	adja	80	60	6	3	74	57	7.5	5.0
	adjp	5	6	0	0	5	6	0	0
	pact	290	326	16	20	274	306	5.5	6.1
	ppas	612	731	55	38	557	693	9.0	5.2
ADVERBS	**adv**	307	327	50	48	257	279	16.3	14.7
PRONOUNS	**ppron12**	4	3	1	2	3	1	25	66.7
	ppron3	1	1	1	1	0	0	100	100
	siebie	1	1	1	1	0	0	100	100
VERBS	**fin**	762	798	83	85	679	713	10.9	10.7
	bedzie	1	1	1	1	0	0	100	100
	aglt	1	1	1	1	0	0	100	100
	praet	898	1243	131	178	767	1065	14.6	14.3
	impt	4	10	0	1	4	9	0	10
	imps	259	269	13	6	246	263	5.0	2.2
	inf	467	675	14	23	453	652	3.0	3.4
	pcon	182	131	1	0	181	131	0.5	0
	pant	4	1	0	0	4	1	0	0
	winien	2	2	1	1	1	1	50	50
	pred	13	10	6	5	7	5	46.2	50
PREPOSITIONS	**prep**	55	51	37	34	18	17	67.3	66.7
CONJUNCTIONS	**conj**	56	52	34	38	22	14	60.7	73.1
QUBLICS	**qub**	170	182	60	73	110	109	35.3	40.1
FOREIGN	**xxs**	169	27	0	0	169	27	0	0
	xxx	26	178	0	1	26	177	0	0,6
TOTAL		**15433**	**16849**	**2224**	**2225**	**13209**	**14624**	**14.4**	**13.2**

Table 4: Common SC (*f*≥7) and rare SR (*f*<7) flexemes for the different flexeme classes in the entire lexicon for the two sample groups.

[31] 100% is the number of all flexemes within the given flexeme class in the sample. The table only includes flexemes at *f*≥7 in those classes.

[32] The abbreviations are explained in the text below.

The above data shows the variances (at times significant) between the various flexeme groups within the lexicons of the two sample groups. It turns out that in sample II (modern press information) there is a higher percentage of common vocabulary in these classes: active adj. participles (pact) – by 0.6%, non-third party pronouns (ppron12) – by 21.7%, imperatives (impt) – by 21.7%, infinitives (inf) – by 0.4%, predicates (pred) – by 3.8%, conjunctions (conj) – by 12.4%, qublics (qub) – by 4.8% and non-nominal foreign words (xxx) – by 0.6% . In the following classes the numbers are lower: nouns (subst) – by 0.7% and (ger) by 2.6%, cardinal numbers (num) – 9.4%, adjectives (adj) – by 2.7 and (adja) – by 2.5%, passive adj. participles (ppas) – by 3.8%, adverbs (adv) – by 1.6%, verbs (fin) - by 0.2%, pseudo participles (praet) – by 0.3%, impersonals (imps) – by 2.08%, adverbial present participles (pcon) – by 0.5% and prepositions (prep) –by 0.6%. Finally, without change were the following classes: depreciative nouns (depr), collective numerals (numcol), prepositional adjectives (adjp), pronouns (ppron3), *siebie* pronoun, future forms of *byc* (*bedzie*), *byc* aglutynats (agl), adverb participles (pant), *winien* forms and foreign nominals (xxs).

The biggest percentage changes occurred in the pronouns, conjunctions and imperatives categories with the percentage being considerably higher in modern information texts while in the cardinal numbers class observed was the biggest decrease.

Taking into consideration the relations between the flexeme and grammar classes corresponding to traditional parts of speech, it should be noted that the described above relations, in reference to those in Table 5, are as follows: 4 out of 10 classes have increased their common vocabulary percentages. These are: pronouns (by 30%), conjunctions (by 12.4%), qublics (by 4.8%) and foreign words (by 0.5%). The others have decreased, with numerals at the highest percentage drop (cf. Table 5).[33].

class	**number of flexemes for the given flexeme class**		***f*≥7**		***f*<7**		**% % of flexemes with a frequency of *f*≥7 for the given flexeme class**[34]	
	I	**II**	**I**	**II**	**I**	**II**	**I**	**II**
NOUNS	8549	9059	1256	1252	7293	7807	14.7	13.8
NUMERALS	53	69	44	45	9	24	83.0	65.2
ADJECTIVES	3449	3758	489	429	2960	3329	14.2	11.4
ADVERBS	307	327	50	48	257	279	16.3	14.7
PRONOUNS	6	5	3	4	3	1	50	80
VERBS	2593	3141	251	301	2342	2840	9.7	9.6
PREPOSITIONS	55	51	37	34	18	17	67.3	66.7
CONJUNCTIONS	56	52	34	38	22	14	60.7	73.1
QUBLICS	170	182	60	73	110	109	35.3	40.1
FOREIGN	195	205	0	1	195	204	0	0.5
TOTAL	**15433**	**16849**	**2224**	**2225**	**13209**	**14624**	**14.4**	**13.2**

Table 5: Common SC (*f*≥7) and rare SR (*f*<7) flexemes for the different parts of speech classes in the entire lexicon for the two sample groups.

The goal of the analysis for this article was to determine any differences between the lexis in press information texts used in the 1960s (sample I) and at the turn of the century (sample II). The carried out research proved that theories regarding press language changing over the years are false. The statistical analysis performed examined and compared areas such

[33] For more detailed analysis of lexicon changes taking place over the years in press information texts, a further division of vocabulary into very common and very rare lexis would have been appropriate, however, due to volume limitations it was omitted.

[34] 100% is the number of all flexemes within the given flexeme class in the sample. The table only includes flexemes at *f*≥7 in those classes.

as language diversity, originality, stereotypicality and grammar structures used. The results obtained do not illustrate any significant statistical differences as the values obtained as a result of various tests show slight differentiations and regard only some of the aspects examined.

A valuable supplement to this analysis would be research on the topics discussed in press articles then and in modern day. This would, however, require a method of description and classification of vocabulary in its connotation as well as its functioning in recipient consciousness (or the subconscious). In this case, some attention should also be devoted to proper names, their characteristics and their derivatives, being a group of words most influenced by reality and, at the same time, least influenced by author's preferences. This group, more so than others, includes the answers to most basic journalistic questions – who (first and last names, pseudonyms), what (institutions, organizations) are the subjects and objects of described events, and where they take place (geographic names, etc.).

Another imperative issue to analyse would be a characteristic of the stylistic structure of the analysed vocabulary, including such groups of words as professional, emotional and colloquial. However, due to limitations of the method used in selecting such words in the texts[35], this type of analysis was not possible. Although it would have been possible to isolate some of the above mentioned categories, it would only be a partial analysis without guarantee of encompassing the entire sample. In this case, more appropriate would seem to be socio-linguistic research based on context analysis.

The above mentioned additional issues are, according to the author, significant in carrying out not only linguistic analysis of press texts. They would provide us a better picture of how outside language reality has influenced the lexis used in press texts. Due to the above, it would seem appropriate to devote these matters further attention.

[35] A segment was considered the basic unit of text. It is defined as a sequence of signs. Segments are never longer than words, or sequences of signs which are not separators of other words. Traditional separators are spaces and punctuation signs, excluding hyphens, periods (as part of abbreviations) and apostrophes.(Cf: Przepiórkowski A., op. cit., p. 18-21).