

## Pathological Text and its Statistical Parameters: An Automatic Text Processing Oriented Research

X. PIOTROWSKA  
R. PIOTROWSKI  
Y. ROMANOV

*Herzen St. Petersburg Pedagogical University, Russia*

W. PIOTROWSKA

*Pavlov State Medical University, Russia*

### ABSTRACT

*This paper discusses nonlinear "psycholinguistic" behavior as the foundation of the modern natural-language processing, machine translation (MT) and artificial intelligence. For the optimum implementation of the idea, it is necessary to make the best use of the information on the synergetic mechanisms of the normal and pathologic verbal-mental activity (VMA) of human beings. This approach, in addition to other researches, investigates the synergetic transformation of cognition and language production under the effect of mental diseases such as schizophrenia and manic psychosis. We present experimental findings of the pathologic cognitive and linguistic changes that are of principal interest to language engineering in general and machine translation, in particular.*

### INTRODUCTION

The classic, logical-cognitive science with a proof-theoretic approach served as a basis for the NLP research during the second half of the 20<sup>th</sup> century. This approach involves three components:

1. specification of a formal language,
2. a set of proof-theoretic (i.e. syntactic) rules of inferences,
3. mechanistic implementation of the above 2.

This was the theoretical position in cognitive science as well as in the initial period of the MT and AI research (Fodor & Polyshyn 1988: 3-71).

However, the second cognitive revolution in the 80s-90s cast some doubt on the strong logical character of the verbal-mental activity of human beings (Oaksford & Chater 1991: 1-10; Harre 1993: 25ff.). Moreover, in the late 80s, it became clear that all the attempts to obtain coherent and correct MT of a random input text using classical monotonic logical programs, failed (Piotrowski 2000: 235).

That is why modern language engineering started searching for a way to overcome an invisible barrier of rejection between fuzzy natural language and the "rigid" logical artificial language of the linguistic automaton. A key to understanding this problem lies in investigating the synergetic mechanisms of the human VMA and, in particular, its pathology.

It is well known that, in a normal social and physiological state, both the human individual VMA and the language system function in a unified and harmonious way shutting tightly the "windows" through which we can observe the synergetic "language forces" (Altmann & Koehler 1996: 62-65). Where should we look for ways and means that would enable us to observe this esoteric self-regulation mechanism? One such possibility is provided in the situations when a pathological state opens up to us, simplifying what is normally, in the normal state, hidden from us. So, new and detailed information on linguistic synergetics can be obtained by studying the pathology of individual speech. The following aspects are of particular interest:

1. generalized long-term disorders in language competence and performance related to a local damage of the brain (Lurija 1973);
2. medium- and/or long-term disorders in language and speech related to the progress of prodromal mental disease (primarily, schizophrenia) as will be discussed below;
3. short-term disorders in linguistic competence and performance induced in the normal consciousness by the

influence of psychoactive drugs or procedures (Spivak 1997; Roy, Dutta & Majumder 1998).

In this paper, we present some VMA results obtained in the quantitative analysis of the text produced by a number of Russian schizophrenic patients.

#### MATERIALS AND METHODS

The basic material used by pathological linguistics (PL) is an oral or written text produced by the test subject spontaneously or obtained in the form of his/her answers to the experimenter's questions: a text in which one can find objective evidence of VMA-pathology. We will call it pathological text (PT).

Identifying a particular text as a PT is a complicated problem. Let us consider three English passages:

1. When dogs cross the air in a diamond like ideas and the appendix of the meninx tells the time of the alarm program...
2. Cause for treatment novocaine nervous-impotence served for us communication several sick, getting intravenous injection novocaine in connection with hypertonic hypertension, about the fact that after conducted course injections at from them disappeared appearance premature-ejaculation...
3. It being, you know, there all for all insurance policies being, insurance policies being, insurance policies being, a very good investment, death being inevitable.

The first two texts, which constitute a "verbal jumble," most likely reflect disjointed schizophrenic thinking. On the contrary, the third text is grammatically unexceptionable, though it does contain paradoxical metaphors, which can be evaluated as a sign of pathological thinking. On the other hand, it can be perceived as a passage from an avant-garde poem written in blank verse, or as a prosaic fragment reflecting the character's flow of consciousness.

In reality, the first text is a Dadaist poem by Tristan Tzara 'Seven Dada Manifestos' that the poet claims "to have been constructed by talking a newspaper article, cutting out the words,

shaking them in the bag, and then transcribing them one at a time" (Marshall & Gurd 1993: 95). The second fragment is a word-for-word machine translation of a Russian scientific text ("Linguistic and Engineering Studies" 1960: 207), while the third passage was produced by a schizophrenic speaker "with auditory hallucinations and paranoid delusions" (Hoffman & Sledge 1984: 150).

Thus, in spite of the fact that the diagnosis of schizophrenia is significantly dependent upon the characteristics of the patient's language production, it is practically impossible to classify a specific text fragment as a PT without any information on the psychiatric condition of its author. In order to answer the question of whether or not such a text is pathological, it is necessary to reveal deep indicators, often hidden from direct observation and therefore not consciously reproducible as well as not controllable by the patient. It is precisely these indicators that may manifest synergetic disruptions in his/her VMA activity.

If we talk about indicators in regard to expression, then they have to be revealed using a statistical information model. Of primary interest here are such peculiarities of text formation as consistent excessiveness for certain styles and sublanguages, quantum distribution of information throughout the text, and dependence of the multiple meaning of lexical units (word-forms or word-combinations) on their frequencies, i.e., on their positions in a frequency list (FL).

However, it is impossible to perform an elaborate statistic-information experiment on text guessing with mental patients. Therefore, we will turn to statistical analysis of pathological texts, using the classic Zipf-Mandelbrot rank-frequency dependence:

$$(1) f = k/(i + \rho)^\gamma$$

and its parabolic variant

$$(2) f = k/i^{\gamma+g \lg i},$$

where  $i$  is the number (rank) of the lexical unit in the frequency list,  $f$  is the expected relative frequency of the lexical unit occurrence in the text,  $k$ ,  $r$ ,  $\gamma$ , and  $g$  are coefficients depending on the sample size, style, subject, and organization of the text (Hoffmann & Piotrowski 1979: 70-74; Alekseev 1983: 67).

In terms of our task, it is especially interesting to represent:

1. Behavior of parameter  $\gamma = \operatorname{tg} \varphi$  ( $\varphi$  is the slope of Zipf curve to the X-axis of a bi-logarithmic graph, see Fig. 1);
2. Average frequency ( $\bar{F} = V/N$ ) of the lexical units in our text ( $N$  is the sample size measured in text-words,  $V$  is the number of different dictionary words in our sample);
3. Lexical ordering ( $Y = 1 - \lg V/\lg N$ ). This coefficient is obtained by transforming the Herdan type/token duality relation (Herdan 1966: 328-325), which must reflect the Saussurian *langue-parole* antinomy.

Let us start with the  $\gamma$  parameter. With an insufficient sample size  $N$ , the recurrence of the most often used lexical units placed at the top of the FL is comparatively low, then the portion of rare lexical units at the bottom of the list is quite significant. Therefore, the graph of the Zipf dependence descends gradually to the  $x$ -axis, forming an angle  $\varphi < 45^\circ$  with  $\gamma < 1$ . As the sample size increases, the portion of the most often used lexical units rises, and the relative significance of the rare lexical units is reduced. Then the average frequency  $\bar{F}$  of the lexical units increases.

With the same sample size,  $\bar{F}(\text{saturated sample}) > \bar{F}(\text{unsaturated sample})$ . In the course of the saturation of the sample, the angle  $\varphi$  gradually increases, reaching  $45^\circ$  in the so-called ideal Zipf sample, described by dependence (1). Usually, it approaches 100,000 text words. In this case,  $\gamma = 1$ . As the sample is progressively saturated and approaches the general population, the rare lexical units move into the medium-frequency zone. The middle part of the bi-logarithmic graph, the "bulging", shifts to the right and upward, due to which the saturated sample acquires

the form of a convex curve (Fig. 1) described by the parabolic dependence (2). In these cases,  $\gamma > 1$ . The rate of approach to saturation is determined by the topic and the style.

When human life is controlled by a restricted number of repeating situations, the VMA strives for acquiring a rigid order, and the texts are characterized by uniform topics and style, and thus by a restricted dictionary. Even a small set of such texts containing a restricted number of lexical units tends rapidly to saturation, and the formal features of this are reflected by the parabolic Zipf curve, together with large values of  $\gamma$ ,  $\bar{F}$  and  $\gamma$ . Examples are provided by the frequency dictionaries for Russian military documents (Table 1, Fig.2). In particular, in children's speech, with its smaller dictionary, the sample rapidly reaches saturation with  $\gamma = 1.6$  (Piotrowski, Pashkovskii & Piotrowska 1994: 30-31).

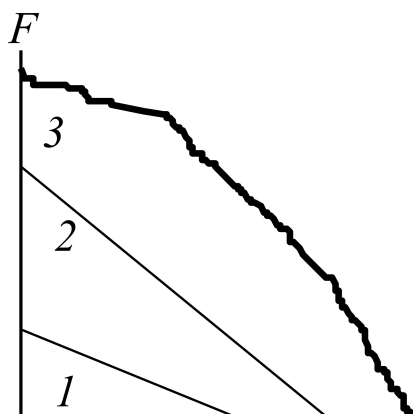


Figure. 1. *Typical bi-logarithmic graphs of the rank-frequency dependence for three types of samples: unsaturated sample, ideal Zipf sample, & saturated sample (cf. texts recorded from patient X).*

Next, we can also consider statistical results obtained with patients showing VMA disorders. Meaning analyses show that for most of these patients their VMA enters the zone of

disequilibrium, with corresponding distinctive characteristics in the texts. Two major types of such texts can be identified.

In the first case, the dominant aspect involves an organizing mechanism that reorders the patient's VMA. Consequently, the text is filled mainly with words and word combinations related to the obsessional topic. The variety of lexical units employed here is restricted, and reiterations are numerous. This naturally leads to rapid saturation. This, in its turn, is reflected in the statistical parameters of the text obtained from patient X (Table 2), who suffered persistent delusion of grandeur. In spite of the restricted length (about 1,000 text words) and the small volume of the word list (162 word forms), this represented a saturated set, as indicated by the unusually large values of  $\gamma$ ,  $\overline{F}$ , and  $\gamma$ .

Table 1. *Nonpathological Russian texts*

| Statistical characteristics | Fiction texts (M. Gorkij "Trilogi") cf. Alekseev, 1996 | Fiction text (N. Hogol "Memoirs of a mad-man") | Texts on electronics (Kalinina 1973) | Frequency List of Russian Words (fiction, scientific and newspaper texts), cf. Zazorina, 1975 | Military documents (Kolguškin 1970) |
|-----------------------------|--------------------------------------------------------|------------------------------------------------|--------------------------------------|-----------------------------------------------------------------------------------------------|-------------------------------------|
| $N$                         | 165927                                                 | 7986                                           | 200388                               | 1056382                                                                                       | 689214                              |
| $r$                         | 1,10                                                   | 0,89                                           | 0,9                                  | 1,0                                                                                           | 1,4                                 |
| $V$                         | 13218                                                  | 831                                            | 21468                                | 39268(words)                                                                                  | 3001                                |
| $\overline{F}$              | 12,55                                                  | 9,6                                            | 9,3                                  | 26,9                                                                                          | 229,7                               |
| $\gamma$                    | 0,789                                                  | 0,759                                          | 0,816                                | 0,762                                                                                         | 0,595                               |

Disorders of the second type are characterized by multiple topics and the absence of a consistent subject. The lexicon of such a text may be varied and chaotic. The history of this condition usually ends in disintegration of the VMA. This is reflected in the texts from patients L, F and K, which represent unsaturated sets having  $\gamma \ll 1$  and small values of  $\gamma$  (Table 2).

Note that the story by Hogol, presenting an imitation of a pathological text, is characterized by parameters that do not differ from the values  $\gamma$ ,  $\overline{F}$  and  $\gamma$  of non-pathological prosaic texts (cf. Table 1). This clearly demonstrates that there is no way to simulate adequately the deep processes of the patient's speech

creativity, i.e. a mentally healthy author cannot imitate deep VMA disorder.

There are thus persistent statistical text characteristics, with variations in certain situations, which not only indicate that the self-organizing mechanisms of VMA exist and can vary as a result of powerful influences such as changes in topic and style (compare military order texts) or else on account of mental disease. These computable characteristics indicate the ways in which VMA is distorted in the period of bifurcation and can also be used to construct dynamic models for linguistic self-regulation and self-organizing.

Table 2. *Russian pathological texts (Paškowskij, Piotrowska & Piotrowski 1994: 59)*

| Statistical characteristics | Patients |       |       |       |       |       |
|-----------------------------|----------|-------|-------|-------|-------|-------|
|                             | X.       | O.    | K.    | L     | F     | P     |
| $N$                         | 993      | 30828 | 14944 | 5526  | 2947  | 3000  |
| $V$                         | 162      | 8750  | 4994  | 2763  | 1016  | 1364  |
| $r$                         | 1,5      | 0,83  | 0,9   | 0,8   | 0,7   | 0,73  |
| $\overline{F}$              | 6,2      | 3,5   | 2,9   | 2,0   | 2,9   | 2,2   |
| $r$                         | 0,263    | 0,122 | 0,115 | 0,081 | 0,133 | 0,102 |

## CONCLUSIONS

The results of statistical investigation of the pathologic speech production make it possible to draw the following conclusions:

- regular or non-regular speech generation and perception depends, first of all, on the state of the communicative-pragmatic operator (CPO), see Piotrowski 1994.
- CPO damage causes some disorders in the semiotic organization of the thesaurus, linguistic competence, and communicative ability of the patient.

Hence, the CPO can be considered a regulator that supports the message generation, its perception and analysis at a certain level. It brings about convergence of various lexical units at this level (Piotrowski 1994: 19-20).

At the same time, we believe that the CPO is a central block of the synergetic mechanism of human verbal-mental activity. It follows that the future of machine translation and other natural-language processing applications depends on the possibility to reproduce in these systems a synergetic model of CPO which will bring closer a text generated by the linguistic automaton and the real-life human speech.

## REFERENCES

- Alekseev, P. M. 1983. *Metodika kvantitativnoj tipologii teksta. Učebnoe posobie*. Leningrad: LGPI im. A.I.Gertsena Press.
- . 1996. *Častotnyj slovar' avtobiografičeskoj trilogii M.Gor'kogo*. St.Petersburg: StPb University Press.
- Altmann, G. & Koehler, R. 1996. "Language forces" and synergetic modeling of language phenomena. In P. Schmidt (Ed.), *Glottometrika 15. Issues in General Linguistic Theory and the Theory of Word Length*. Trier: Wissenschaftlicher Verlag Trier.
- Fine, J. 1995. Towards understanding and studying cohesion in schizophrenic speech. *Applied Psycholinguistics*, 16, 25-41.
- Fodor, J. A. & Polyshyn, Z. W. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28/1.
- Harre R. 1993. Emotion and memory: Second cognitive revolution. *Mind and Language*, 8/1.
- Herdan G. 1966. *The Advanced Theory of Language as Choice and Chance*. Berlin-Heidelberg-N.Y.
- Hoffman, R. & Sledge, W. 1984. A microgenetic model of paragrammatisms produced by a schizophrenic speaker. *Brain and Language*, 21, 147-173.
- Hoffmann, L. & Piotrowski, R. 1979. *Beitraege zur Sprachstatistik*. Leipzig: VEB Verlag Enzyklopaedie.
- Kalinina, E. A. 1973. *Haeufigkeitswoerterbuch der russischen Subsprache der Elektronik*. In P. M. Alexiejew, W. M. Kalinin, R. G. Piotrowski (Eds.), *Sprachstatistik* (pp. 184-191). Muenchen-Salzburg: Wilhelm Fink Verlag.
- Kolguškin, A. N. 1970. *Lingvistika v voennom dele. Razrabotka i ispolzovanie častotnych slovarej voennoj leksiki*. Moscow: Voenizdat.
- Linguistic and Engineering Studies in Automatic Language Translation of Scientific Russian into English. Phase II* (1960). Seattle: University of Washington Press.

- Lurija, A. R. 1973. *Osnovnye problemy nejrolingvistiki*. Moscow: MGU Press.
- Marshall, J. C. & Gurd J. M. 1993. Disordered language in creative writing. In G. Blanken et al. (Eds.). *Linguistic Disorders and Pathologies. An International Handbook* (pp. 92-97). Berlin, N.Y.: Walter de Gruyter.
- Oaksford, M. & Chater, N. 1991. Against logicist cognitive science. *Mind and Language*, 6/1.
- Paškowskij, W. E., Piotrowska, W. R. & Piotrowski, R. G. 1994. *Psychiatricheskaja lingvistika*. St-Petersburg: Nauka.
- Pavlov, I. P. 1949. *Lekcii o rabote bol'sich polušarij golovnogo mozga. Lekcija 18. Polnoe sobranie sočinenij* [Lectures on the Work of the Cerebral Hemispheres. Lecture 18]. *Complete Works*, 4, 316-325. Moscow-Leningrad: AN SSSR Press.
- Piotrowski, R. 1994. Psycholinguistic basis of the linguistic automaton. *International Journal of Psycholinguistics*, 10/27, 15-22.
- . 2000. MT in the former USSR and in the newly independent states (NIS). Prehistory, romantic era, prosaic time. In W. J. Hutchins (Ed.), *Early Years in Machine Translation. Memoires and Biographies of Pioneers*. Amsterdam/Philadelphia: John Benjamins.
- , Pashkowski W. E. & Piotrowska, W. R. 1994. Psychiatric linguistics and automatic text processing. *Automatic Documentation and Mathematical Linguistics*, 28/5, 28-35.
- Roy, P. & Dutta Majumder, D. 1998. Pharmaco-linguistics and Non-equilibrium dynamics. *National Conference on Mathematical Linguistics*. Calcutta: Indian Statistical Institute.
- Spivak, D. 1997. Quantitative measurement in linguistics of altered states of consciousness. In *Proceedings of the XVIth International Congress of Linguists*, Paris, 20-25 Jul.

## ACKNOWLEDGMENT

This research was supported by the Russian Humanitary Scientific Fund (RGNF) under contract № 02-04-00195a.

**X. PIOTROWSKA**

**R. PIOTROWSKI**

**Y. ROMANOV**

LANGUAGE ENGINEERING LABORATORY,  
HERZEN ST. PETERSBURG PEDAGOGICAL UNIVERSITY,  
48 MOYKA, ST.PETERSBURG, 191186, RUSSIA.

E-MAIL: <RP@YR4993.SPB.EDU>

TEL./FAX: +7 (812) 311 2189

**W. PIOTROWSKA**

PAVLOV STATE MEDICAL UNIVERSITY,  
THOREZ AVE. 9, APT. 6,  
ST.PETERSBURG, 194021, RUSSIA.

E-MAIL: <BURM@WPLUS.NET>