Time-Series Analysis in Linguistics:
Application of the ARIMA Method to Cases of
Spoken Polish*

Adam Pawłowski
University of Wrocław

ABSTRACT

In this paper the ARIMA method of time-series analysis is applied to textual data. Four prosodic types of Polish are discussed (children’s verse, complex verse, rhetorical discourse and literary prose). The analysis proves that each of these sorts of prosody conveys rhythmic components. We have also defined and compared formal models of the sequential structure of text for the treated samples.

INTRODUCTION

We can speak of two investigative approaches in the quantitative analysis of language. The analysis of language in the mass is based on the assumption that successive linguistic units in a line of text or speech are statistically independent. For instance, if we admit that the examined unit is the graphical word, the order of words in the text will be considered as irrelevant. The analysis of language in the line rejects the assumption of independence of successive units. In this case, the order of words, syllables, etc., is considered as an important and relevant characteristic to be explored.

The first scholar to have worked out and applied a quantitative method of sequential analysis of text was not a linguist but a mathematician. Seeking examples to illustrate his theory of stochastic processes, Andrej Andrejevič Markov analysed the sequence of vowels and consonants in Russian. The distinction language in the mass vs. language in the line was introduced in the 1960’s by Gustav Herdan. “In the area of language, it is the dimension of time which may have to be taken into consideration. We may deal with language in the mass, or with language in the line.” (Herdan, 1966, p. 423). This was unfortunately a theoretical postulate only – Herdan did not propose any practical research method to be applied in the sequential analysis of verse or language in general.

Although both investigative approaches are of great importance to quantitative linguistics, the majority of studies carried out in this field so far stem from the analysis of language in the mass, based upon the well known and tested methods of descriptive and inferential statistics. In consequence, innumerable papers devoted to

1 In statistics the notion of independence is well defined and from this point of view the above assumption is quite correct. However, we should not forget that out of this context it might lead to strange or humorous statements – indeed, how could one study the style of a literary work assuming that the order of words and phrases is not important?

2 The first important report in linguistic literature dealing with Markov’s ideas, is due to Herdan (1966, pp. 140–153). However, the most exhaustive description of his linguistic works can be found in the monograph of the French scholar Micheline Petruszewycz (1981). The comparison of the Markov method and of the time-series analysis with regard to linguistic applications is presented in Pawłowski (1997, pp. 53–61). It is worth noting here that in spite of having been known for almost a century (since 1910–1915), Markov’s concepts in fact have not had serious applications in stylistics.

* Address correspondence to: A. Pawłowski, ul. Sądowa 10/2, 50-046 Wrocław, Poland.
the analysis in the mass contrast with the few works dealing with sequential analysis of language. This lack of proportion is surprising. After all, if any characteristic of text were to be designated as especially important from the point of view of different schools and theories, it would certainly be the linear structure, opposing natural language to the majority of semiotic codes. A broad survey of studies related to the sequential analysis of text can be found in the monograph by Pawlowski (1997, pp. 49–74).

Goal of the Study

The objective of the present study is a sequential analysis of different prosodic types in Polish by means of the Box and Jenkins ARIMA method of time-series analysis. Our investigation will lead to the formulation of a formal model describing the arrangement of stressed and unstressed syllables in text. If the method confirms the existence of an asymmetry (in qualitative analysis referred to as rhythm of speech), we will say that the text conveys a more or less distinct stochastic process. The material of our study consists of recorded texts representing different prosodic (rhythmical) types of Polish. As a working hypothesis we admit that the ARIMA method will discriminate between these varieties, revealing in each case the particular type of the model as well as a characteristic value of its coefficients. This formal distinction of text samples according to the value of model parameters will also be confronted with the intuitive impression of rhythm in a given text.

Two remarks come to mind here. Since textual data are qualitative (or categorial) and the ARIMA method treats only numerical time-series, we are obliged to convert text into numbers. This of course can be done in many different ways, but we should keep in mind that a given model of the sequential structure of a text is related to one, precise method of quantification. This condition is absolutely necessary for comparison of different samples with regard to a given model.

The second remark concerns the application of the same methods both to prose and verse. As we know, the traditional research object when studying versification is not prose but poetry (cf. Robertson, 1996, pp. 33–34). One of the great advantages of the ARIMA method is the possibility to reveal in a time-series even very weak stochastic processes – just like those we expect to discover in verse samples. It will be shown that the ARIMA method will not only make it possible to compare prose and verse (in this case we expect extremely different models) but also different texts in prose. And this goal cannot be achieved without a reliable mathematical tool, which the ARIMA method certainly is.

DATA AND METHOD

Data

It is very probable that a constant and regular verse introduces an additional order to the distribution of stressed and unstressed syllables in the line of spoken text. In consequence, time-series generated by such texts should convey a strong stochastic process. Conversely, the lack of a constant and regular verse structure, as is the case in prose, is likely to produce a time-series dominated by a random component.

For this reason, four groups of texts were chosen for analysis to represent different systems of versification. Each of these prosodic types is represented by a group of three samples composed of ca. 700 syllables. The exception is poetry for children where shorter series (ca. 250 syllables) were satisfactory. A detailed list of all treated texts (titles, authors, editions, recordings etc.) is to be found in the appendix.

As an example of text with a very strong and regular rhythmic structure, we have chosen the verse for children by Jan Brzechwa, performed by actors of Warsaw theatres. The uncomplicated psychology of children imposes a simple metric structure (here octosyllable with a regular caesura) and requires a very scholastic recitation. As can be seen below, the recited text is a more or less complete series of trochees (Brzechwa, 1965, p. 42):

Każdy piątek i każdy czwarte
Wszelką komonę się wybierze...
I natychmiast przyniesie
Jako udział w świecie.

(“Every bird and every animal should now choose something in his cupboard / and bring it here immediately / as his share in our business”).

The next text to be examined is Beniowski of Juliusz Słowacki. Two characteristics have influenced our choice of this romantic poem written in hendecasyllable (verse of eleven syllables). The first one is of course its regular versification. The other is its unbroken narrative which makes it possible to generate almost any number of homogeneous samples of the desired length (700 syllables). The number of samples is important in an empirical investigation based on inductive reasoning because it validates the results and allows them to be applied over a greater range of material. The two other texts of our corpus have no regular structure of verse. This concerns both the prose of Igor Newsky and the fragments of papal homily. There is no special reason for the choice of the former author – we wanted to examine a contemporary literary prose text and we found that the novel Wżórze Blięknego Snu was adequate as research material.

The factor which determined the choice of the latter was the particular communicative context in which this text appeared. A public person speaking “live” to any large audience should take much care for the perspicuity and persuasibility in his discourse to be understood and to convince. And among all the available rhetorical and stylistic tools the rhythm of speech, i.e., adequate and purposeful distribution of accents in the line of text, plays a considerable role.

In the fragment of the homily below, a sort of culminating point of the previous part of the discourse, the Pope repeatedly lays stress on the first syllable of a word. This unusual fact (the Polish language has a parasytone system of accentuation) is due to the emotive situation and shows that a live performance particularly modifies the prosody of language.

W ewangelicznym przykazaniu miłości
tkwi bowiem najgłębsze źródło
jednoznaczne, niezmiennie
dowodzącego rozwoju każdego człowieka.

(“In the evangelical command of love there is the most profound source of spiritual development of a human being”).

Quantification of Data

Quantification consists in replacing categorial units with adequate numerical values. This is an obligatory stage in all experimental research including the treatment of qualitative data. If the quantified unit is a syllable, the choice of these “adequate” values can be a questionable issue. The notion of accent itself – one of the primary elements of prosody – embraces not only the
dynamic accent (most frequent in Polish), but also tonal accent and quantity. Although the latter are not frequent in Polish, they are present in expressive oral texts like the homilies of John Paul II or recitations of romantic poetry.

Moreover, the technique of quantification we have applied—bearing and coding directly from the tape—is biased by an error of subjectivity and variability in human perception depending on different factors (fatigue, etc.). Another problem is the choice of the numerical scale to be used when substituting numbers for categorial units (three scales are applied in the analysis of data: nominal, ordinal and cardinal). Finally, the last remark has a far more general scope: what we are trying to do is to project a fragment of an extremely complex linguistic reality into the universe of numbers. At the stage of interpretation of our results we may well lose the relationship between these results in numerical form and qualitative notions they should represent.

Taking into consideration the above-mentioned remarks as well as the experience of other scholars, we have found the quantification method which consists of assigning the number 1 to a stressed syllable, 0 to an unstressed syllable, and 0.5 to a syllable bearing a secondary stress to be the best (this accent appears occasionally in words composed of four syllables). The substitution of these values was carried out by a person listening to the tape and then verified by other persons. As we can see, the applied scale is ordinal.

The reason for our decision is that a very similar, binary nominal scale is generally used in the notation of syllabic verse and in prosody. Additionally, we attribute to stressed and unstressed syllables specific meaning—unlike the same type of quantification was used by Bratley and Ross (1981). They noticed that: “A binary ‘either/or’ choice (…) is a reasonable approximation of what happens in speech, and has been used in several major studies of metrics, both traditional and generative” (Bratley & Ross, 1981, pp. 42-43). Also Azar and Kedem (1979), who investigated the sequential structure of Hebrew on the example of biblical texts, substituted zeros and ones respectively for voiced and unvoiced phones. Time-series generated in this way were then submitted to spectral analysis. The results obtained in both cases turned out to be extremely interesting and prove that the application of the ordinal scale in the generation of time-series is admissible.

The final remark concerns our lack of using any technical devices which, at least apparently, would help us determine more precisely stressed and unstressed syllables in the line of text and express the level of stress in precise, physical units. It so happens that since antiquity the phenomenon of accent has been known and perceived as a binary feature, and the lack of technical devices to measure it has never been an obstacle in the study of metrics. The reason for this is that during an ordinary act of communication (when the communication function dominates over emotional and aesthetic ones) the sender uses a binary scale of accentuation in coding his message and the receiver recognizes this manner of coding perfectly. Thus, our method of quantification is not artificial but simply respects a widespread linguistic and communicative norm.

The Notion of Time in Time-series Analysis of Textual Data

The ARIMA method of time-series analysis was worked out for the treatment of data on the axis of real time. However, the notion of time in text analysis has a slightly different sense from that in history or economics. That is why a commentary on some epistemological aspects of that transfer of notions and methodology to the field of linguistics seems advisable, especially as the synchronic analysis of text is concerned.

Although text or speech can be thought of as a series of units, its linear structure is not a simple reflection of the real time which is the basis of diachronic research. (Cf. Bernet (1983), Lebart and Salem (1994), Leech (1953), Lutoslawski (1896), Muller (1979), Salem (1987, 1988), Swadesh (1952, 1953), Yardi (1945)). The epistemological argument which makes it possible to apply the ARIMA method and spectral analysis to the treatment of textual data is the analogy between the linearity of real time and the sequential structure of text. In both cases observed data occur in a sequence, one after another, just like consecutive laps of time, and their order is considered as a relevant feature. This analogy allowed us to introduce to the analysis of text the notion of syntagmatic time which can be substituted for the notion of real time: “Nous allons appeler l’axe linéaire sous jacent à la succession d’unités linguistiques temps syntagmatique (...) et nous allons le substituer au temps réel des événements observés. La pertinence de l’analogie entre ces deux notions est une condition essentielle de l’application de la méthode ARIMA au traitement du langage.” (Pawłowski, 1997, p. 4).

Method

There are two complementary approaches in time-series analysis: time-domain and frequency-domain analysis. In the first case, data are treated directly on the time axis as a sequence of successive values. The estimated model is an additive model and may include such components as deterministic trend, seasonal oscillations and stationary autoregressive or moving-average processes. In the second case, data are regarded as a wave which undergoes spectral decomposition in order to reveal its parameters such as most important harmonic frequencies, their amplitudes and phases. Frequency-domain analysis is based on the Fourier theorem.

Spectral analysis gives results in the treatment of time-series conveying strong deterministic components. However, if these components are feasible and a random element takes over, the analysis in time-domain offers a much clearer image. Although previous research has shown that textual data contain only weak processes (cf. Corduas, 1995; Pawłowski, 1997, pp. 155-158), this need not necessarily be the case here, and so both approaches will be presented and applied.

As was stated, the ARIMA method can be used for the treatment of any series of numbers and reveals mutual relations between its consecutive values. If these relations exist, they can be expressed by means of a mathematical function which in the case of linguistic data becomes a formal model of sequential text structure. Since all types of time-series can be treated with the ARIMA method, there is no limitation on the kind of text to be analysed (in our case we have compared the rhythmic structure of verse and prose). The only real restriction is a thorough and correct quantification with regard to a given feature.

In the Box and Jenkins ARIMA method, a complete time-domain model of a series is defined by means of seven parameters and noted as ARIMA (p,d,q) (P,D,Q). The meanings of these parameters are the following:

- p – order of stationary AR process;
- d – order of non-seasonal differencing;
- q – order of stationary MA process;
- P – order of seasonal AR process;
- D – order of seasonal differencing;
- Q – order of seasonal MA process;
- s – length of seasonality.

Such a great number of parameters may be a complication. Fortunately, the time-series we encounter in the analysis of text (we should rather say text-series) have no trend (d = 0 and D = 0) and quite often no seasonal component (s = 0, so P = 0, Q = 0 and D = 0). A typical model of a stationary series has a form ARIMA (p,0,q) (0,0,1), and is noted as ARMA (p,q). If one of its components is zero, this mixed ARMA model becomes either a simple autoregressive model AR(p) or a moving-average model MA(q). We should however keep in mind that these simple models are always specific cases of the general model ARIMA.

In the ARMA model each element $x_t$ of a stationary time-series is a linear combination of $p$ previous values of the series (AR component) and of $q$ previous random shocks (MA component):

$$x_t = a_1 x_{t-1} + a_2 x_{t-2} + \ldots + a_p x_{t-p} + \epsilon_t - b_1 \epsilon_{t-1} - b_2 \epsilon_{t-2} - \ldots - b_q \epsilon_{t-q}$$

(1)

9 Our description of the ARIMA method is inevitably brief and has no didactic purpose (among the original subjects one finds for instance the problem of forecasting…). A conscientious reader may refer to source texts (Box & Jenkins, 1976), Gotman (1981), McCleary and Hay (1980), Priestley (1981), Chu (1986), and Cryer (1986).

10 This statement is based on the research of Azar and Kedem (1979), Corduas (1995), and Pawłowski (1997).
where
\[ x_t \text{ -- value of the series at the moment } t; \]
\[ a_i \text{ -- coefficient of the autoregressive process;} \]
\[ b_i \text{ -- coefficient of the moving average process;} \]
\[ p \text{ -- order of the autoregressive process;} \]
\[ q \text{ -- order of the moving average process;} \]
\[ e_t \text{ -- noise at the moment } t \text{ (} e_t \text{ normally distributed } N(0,1).) \]

The same expression in operator notation, generally applied in time-series analysis, has the following form:
\[ x_t (1 - a_1 B - a_2 B^2 - \ldots - a_p B^p) = e_t (1 - b_1 B - b_2 B^2 - \ldots - b_q B^q), \]

where \( B \) is the backward-shift operator of order \( n \).

In the treatment of seasonal processes, both stationary and non-stationary, we follow exactly the same reasoning, except that lag corresponds to the length of seasonality and is greater than unity.\(^{13}\) Of course both seasonal and non-seasonal components can be combined in one, complex model.

In order to estimate the above-mentioned parameters we have to determine first the type of the process. In the ARIMA method this can be done with the help of the autocorrelation (ACF) and partial autocorrelation (PACF) functions. ACF (\( \rho_k \)) is one of the fundamental notions in time-series analysis. It is defined as:
\[ \rho_k = \frac{\gamma_k}{\gamma_0}, \]

where \( \gamma_k \) is the autocovariance function (if \( k = 0 \), then \( \gamma_k = \sigma^2 \)); \( k \) is lag, the distance between series elements;
\[^{11} \text{If coefficients } a_i \text{ are zero, a mixed ARMA process becomes a simple moving-average process.} \]
\[^{12} \text{If coefficients } b_i \text{ are zero, a mixed ARMA process becomes a simple autoregressive process.} \]
\[^{13} \text{In economic research, typical periods correspond to calendar seasons like week, month or year and help calculate month-to-month (i.e. January-January, February-February etc.) or day-to-day correlation.} \]

The autocovariance function is defined as:
\[ \gamma_k = E \{ (X_t - \mu) (X_{t+k} - \mu) \} , \]

where \( X_t \) - value of the series in moment \( t \) (all values being equiprobable);
\( \mu \) - mean of the data.

Partial autocorrelation is the correlation between the values of the series without the influence of intermediate (separating) values. In the ARIMA method PACF is the last of the \( \rho^0 \) coefficient in the autoregressive model AR(p).

It can be shown that simple autoregressive and moving-average processes behave in the following way:\(^{14}\)

<table>
<thead>
<tr>
<th>ACF</th>
<th>PACF</th>
</tr>
</thead>
<tbody>
<tr>
<td>process AR(p)</td>
<td>dies out</td>
</tr>
<tr>
<td>process MA(q)</td>
<td>truncates</td>
</tr>
</tbody>
</table>

Table 1.

In our case, the last stage of the analysis is the evaluation of the model. In order to evaluate the quality of the model we compare the variance of the initial time-series with the so-called residual variance, i.e., variance of the residual series, obtained by subtraction of the observed values and of the corresponding values estimated by the model. The percentage of the initial variance explained by the model determines its quality. If the ACF and PACF functions do not indicate clearly the type and order of the stochastic process, we can estimate several models and choose the best one according to this criterion.

One of the basic tools in the frequency-domain analysis is a periodogram, which shows the energy of each component frequency in the series. Out of different formulae of that function we present the one which allows the estimation of spectral power \( \hat{I}(f) \) directly from the values of the series (Gottman, 1981, p. 205).

\[ \hat{I}(f) = \frac{1}{2\pi N} \left( \sum_{j=1}^{N} (x_j \cos 2\pi f j/N + \sum_{j=1}^{N} (\xi_j \sin 2\pi f j/N) \right) . \]

where \( f \) -- frequency; \( \xi_j \) -- successive element of a time-series; \( t \) -- number of the successive observation (time); \( N \) -- number of observations (length of the series).

The first aim of our analysis will be to check (by means of the autocorrelation function) whether the data are random or if they convey some deterministic components. If the series are not random, we will determine the type and order of the processes (we will use for this purpose the ACF and PACF functions), and we will estimate the coefficients \( a_i \) and \( b_i \). If the data contain strong deterministic components, we will also present the frequency model. The results of this mathematical treatment will then undergo a philological interpretation.

RESULTS

Texts with a Strong Rhythmic Component
- Jan Brzechwa, Wiersze dla dzieci

As we could expect, the strongest processes have been found in the verses for children of Jan Brzechwa. We present below an example of numerical treatment of a fragment from the versified novel Opowieść Dzieciom Sowiec.\(^{15}\) The analysis of other samples coming from the same text has proved that their rhythmic structure is similar (cf. Conclusions and Appendix). In consequence, the conclusions we formulate are relevant for the whole cited text of Jan Brzechwa.

The autocorrelation and partial autocorrelation functions suggest that the series is strongly deterministic: ACF doesn’t die out and PACF dies out very slowly.

Once we know the form of the ACF and PACF, we can proceed to the estimation of the model. As we said, for strongly deterministic data the best solution is estimation in frequency-domain.

The periodogram (Figure 3) shows that the sequence of accentuated and non-accentuated syllables can be described by means of two frequencies, \( F = 0.25 \) and \( F = 0.5 \), the energy of the frequency \( F = 0.5 \) being much bigger than that of the frequency \( F = 0.25 \) (the same frequencies dominate in other samples). Such a model fits a repeated sequence of accentuated and non-accentuated syllables where every second and fourth syllable is most often stressed.

Time-domain modelling also gives a faithful image of the process although this model is less

\[^{15} \text{Jan Brzechwa, Dzieciom Sowiec, Nasza Księgarnia, Warszawa, 1965, pp. 34-35 (information on the recording in the Appendix). The length of the sample is 240 syllables.} \]

![Fig. 1. Autocorrelation in a very rhythmic text. Continuous lines on the ACF and PACF graphics show the confidence interval at the 95% level.](image-url)
parsimonious and, intuitively, not as convincing as the periodogram. At least two factors strongly support the application of the seasonal\textsuperscript{16} model in this case: first, a constant verification imposes more regular (thus "seasonal") repetition of accents; second, this solution is indicated by the ACF function. A series of tests with season-

\textsuperscript{16} The notion of "seasonality of text" may, at a first glance, evoke manifold associations whereas it actually indicates nothing other than the economical origin of the ARIMA method. However, responding to that sound reflex of common sense, we cite an appropriate definition: "We define seasonality as any cyclical or periodic fluctuation in a time series that recurs or repeats itself at the same phase of the cycle or period." (McCleary & Hay, 1980, p. 80).

al and non-seasonal models has validated this hypothesis. We present in Table 2 a list of tested models for the examined series:

Assuming as the criterion of choice the percentage of initial variance explained by the model, the best solution is the SARMA(1,1)\textsubscript{s} model for the seasonal lag $s = 4$ or $s = 8$. If $s = 4$ this model has a form:

$$x_t = (1 - 0.99b^4) = (1 - 0.8B^4)e_t.$$  

For the coefficient $a_1$ the standard error on the level of 95% is ± 0.011 and for $b_1$ it is ± 0.044. For the other samples this model explains 57.8% and 68.8% of the original variance. Although the simple AR(8) model gives an apparently good result, it has too many parameters. As we can see, the residual ACF for the chosen seasonal model (Figure 4) has no significant bars.

It is very interesting that we have obtained satisfactory results both with the seasonal lag $d = 8$ (length of the whole verse) and for the lag $d = 4$ (caesura equal to half of the verse). This proves that in spite of the apparent octosyllabic verification, the real rhythmic unit is composed of 4 syllables. Further division of verse into units of 2 syllables leads to a deterioration of model efficiency (SARMA(1,1)\textsubscript{s} explains only 33.33% of the original variance).

Text with a Constant Structure of Verse - Juliusz Słowacki, Beniowski\textsuperscript{17}

Sequential analysis of text with a constant structure of verse has also revealed a very strong stochastic process. The ACF function shows two significant bars for lags 1 and 2 as well as other significant bars which appear in 11 lag cycles. PACF quickly dies out although, just like ACF, it shows significant bars for higher lags.

\textsuperscript{17} Juliusz Słowacki, Beniowski, Państwowy Instytut Wydawniczy, Warszawa, 1974, pp. 97-98 (information on the recording in the Appendix). The length of the sample is 792 syllables. A very similar rhythmical structure has been found in the other samples. This allows generalisation of the result presented.
Such a form of the ACF and PACF functions would suggest in this case a seasonal time-series model. But the number and the value of significant bars in the ACF indicate a strongly deterministic type of process and entitle us to check first a frequency model. Indeed, the periodogram reveals a very interesting form of the spectrum (Figure 7).

We can see in the above graphic that the rhythm of the romantic poem of Juliusz Słowacki written in 11-syllable verse can be described by means of just two harmonic frequencies! Their values $f_1 = 0.364$ and $f_2 = 0.454$, are equal to the periods $T_1 = 2.75$ and $T_2 = 2.2$, i.e. the most frequent intervals (in the nearest integral numbers) between the stressed syllables in the line of text. Exactly the same frequencies dominate in other analyzed fragments of Słowacki. But the observed form of the periodogram is not a consequence of some statistical distribution of intervals (or gaps) between the accentuated syllables. In other words, it is not the number or the proportion of accentuated and non-accentuated syllables in a text which is responsible for the regular rhythm but their sequence – a feature by all means important and, up till now, not examined with adequate formal methods. What seems strange about the form of the periodogram however is that there is no significant frequency for the interval equal to the verse length ($T = 11$).

As was already mentioned, the time-domain analysis leads inevitably to a seasonal model. Assuming the seasonal lag $s = 11$, we have estimated several models, calculating in every case the percentage of the original variance explained (the variance of the original series is $\sigma^2 = 0.24$). We have started with a simple model MA(1), which explains only 33% of the original variance. Also seasonal models SAR(1)_{11} and SMA(1)_{11} leave too much "unfiltered" variance, present in the residual series. For instance, the SAR(1)_{11} model explains only 15% of the original variance and the SMA(1)_{11} model even less (9%). The model which turns out to be satisfactory is a combination of the three previous ones: SARMA(0,1)(1,1)_{11}, explains as much as 49% of the original variance and is, in our opinion, the best time-domain model for the given series. The formula of that model is the following:

$$\begin{align*}
&x_t = (1 - 0.99B^{11})^{-1}(1 - 0.476B)(1 - 0.915B^{11})e_t \\
&\text{For } \phi_{11}, \text{ the standard error on the level of significance 95% is } \pm 0.007, \text{ for } \theta_1 \text{ it is } \pm 0.031 \text{ and for } \theta_{11} \text{ it is } \pm 0.02. \text{ The efficiency of this model is confirmed by the residual ACF which has no significant values;}
\end{align*}$$
Rhetorical Discourse - John Paul II the Homily Pronounced in Wroclaw, (Wroclaw, the 21st of June 1983, Sample length is 754 syllables).

So far we have dealt with texts having a constant and clear versification which strongly influenced their rhetorical layer. In consequence, in both previous cases we have observed very distinct frequency models as well as seasonal time-domain models. Presumably, texts having no constant versification would convey, at best, non-seasonal stochastic processes.

The ACF and PACF functions for the sample from the papal homily (Figure 9 and Figure 10) indicate the presence of some deterministic component but its strength and scope seem quite modest:

We should notice that while the ACF truncates after the second lag, PACF quickly dies out. Such a behaviour of both functions indicates an MA(2) model and suggests time-domain estimation. To dispel doubts, we also present the periodogram of the series (Figure 11). The distribution of power among respective frequencies is uniform and no one of them is significant.

Following the indications of the ACF and PACF functions, we have estimated the MA(2) model which explains 43% of the original variance. Although the MA(1) model explains as much as 38% of the original variance, the remaining residual series is not entirely random. And the models of higher orders do not improve the quality of estimation but only increase the number of parameters. We present below the estimated model MA(2). Standard error for $b_1$ is ± 0.035 and for $b_2$ ± 0.036.

$$X = (1 - 0.795B + 0.264B^2)\varepsilon$$

Autocorrelation of the residual series shows that all deterministic components of the analysed series have been removed by the MA(2) model:

Philological interpretation of this result is not an easy task. Usually, the type and order of the process allow forecasting of the future, unknown values of the series (this is the case when economic or technological data are treated). However, in the case of sequential analysis of text the sense of such an operation seems doubtful. (Pawłowski 1997, pp. 23-24). What we find possible and appropriate instead is to compare different samples of verse, rhetorical discourse and literary prose with regard to the degree of rhythm present. Such a comparison of different estimated models is presented in the last chapter of this study (Conclusions, Table 3). Some of
our conclusions seem out of reach of the traditional approach to versification or prosody.

**Literary Prose** – Igor Newelry, *Wzgórze Błędnego Snu*.\(^\text{18}\)

As in the previous case, we start the analysis of the observed series from the inspection of the ACF and PACF functions. The graphics below (Figure 13 and 14) show that ACF has only one significant bar and PACF quickly dies out. Such a combination indicates that the process in the series is MA(1).

The form of the periodogram (Fig. 15) confirms a strongly random character of the series. Just as was the case with the rhetorical text, the total energy of the spectrum is distributed among the great number of harmonic frequencies and no one of them could be considered as significant.

The choice of a proper model is not difficult. As we have mentioned, ACF and PACF functions clearly indicate MA(1) which explains 30% of the initial variance. However, considering the fact that the ACF for lag 2 is on the limit of the confidence interval (Fig. 13), we have also estimated the MA(2) model. It is less parsimonious, but in return explains a little more of the original variance (31.7%). Finally, we propose the MA(2) model of the following form:

\[ x_t = (1 - 0.658B + 0.165B^2)\epsilon_t. \]  (9)

Standard error for \(b_1\) is \(\pm 0.037\) and for \(b_2\) it is \(\pm 0.038\). The residual ACF shows no significant bars and validates the MA(2) model.

**CONCLUSIONS**

The result of our experiment has proved that the initial hypothesis was correct. It is confirmed that formal models of sequential structure of text based on the notion of syntagmatic time can be created by means of ARIMA modelling and/or spectral analysis. In particular, it was established that the distribution of stressed and unstressed syllables in the line of text is not random but conveys a stochastic process. This process is very distinct in texts having a constant versification. In this case, a simple structure of verse (e.g., children’s verse) generates a stronger process than a complex structure (e.g., romantic poem). The sequence of syllables in texts having no constant versification (e.g., prose) also generates a stochastic process but it is less distinct than the one generated by poetical language.

As far as texts with a constant versification are concerned, both time- and frequency-domain models can be applied. The former are of the

---

\(^{18}\) Igor Newelry, *Wzgórze Błędnego Snu*, Czytelnik Warszawa, 1986, pp. 18-19 (information on the recording in the Appendix). Sample length is 703 syllables. We have obtained similar results with other samples of the novel of Newelry.
seasonal type with the lag of seasonality equal to the length of verse. In the latter, the rhythm of the text can be described with a very limited number of harmonic frequencies (in our case two). If texts under investigation have no constant structure of verse (prose), time-domain analysis is more efficient and leads to simple (thus non-seasonal) moving-average models.

It should be noted here that while the rhythm of poetical language can be quite effectively examined with traditional methods and one might argue whether formal (mathematical) description brings anything new to the matter, the rhythm of prose has remained, so far, beyond the field of interest of both traditional and statistical linguistics. In our opinion, the random character of this phenomenon makes its reliable description impossible without the application of the mathematical procedures of time-series analysis.

As was stated earlier, an effective philological interpretation of our results is not possible if only single samples (models) are considered. Difficulties arise when deciding whether, for instance, a text-series conveying the MA(2) process is more rhythmical than the one with the MA(1) process. Also, the absolute value of model coefficients cannot be directly "translated" into the notions of the qualitative analysis. Reliable conclusions, however, can be drawn from the comparison of several samples. Our comparison was based on the assumption that the percentage of original variance explained by the model testifies not only to the quality of that model but also indicates the degree of rhythm present in the text. In simple words, the more rhythmic and regular a text is, the more distinct and strong is its sequential model and, consequently, a greater percentage of the original variance is explained.

The above-mentioned criterion was applied to compare analysed samples. The table below contains the final result of our experiment. It includes the type of model that fits best to a given group of samples (thus to different prosodic types) and the percentage of the original variance explained by the model in each sample. Although the limited number of samples does not allow far-reaching generalisations, it indicates some important regularities. Texts are in decreasing order according to their degree of rhythm.

It can be noticed in the table above that each prosodic type is represented by a model with a time-domain model and differences between the samples concern only the value of specific coefficients. For verified texts we obtain seasonal models for literary prose moving-average models of the 2nd order.

Now, if we compare the percentage of the original variance explained by the model (directly proportional to the degree of text rhythmisation) we notice that the most rhythmic are simple, verified texts. Slightly less regular is the romantic poem of Słowacki, the rhetorical (non-verified) discourse is less so and the least rhythmic is literary prose. The last place of prose text should not be misleading because it was demonstrated that it does have a rhythm.

While the position and order of verified text was predictable, the positions of prose and of rhetorical discourse were not. The comparison, according to our criterion, of these two types of discourse has shown that in both cases the same model is observed but it is far more distinct in the papal homily. It could mean that a good rhetorical discourse (political, religious, etc.) is more rhythmical than literary prose, even when performed by professional actors. Although the limited number of samples does not allow us to consider this regularity as a general law, such a conclusion seems very probable.

<table>
<thead>
<tr>
<th>Type of the model</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan Brzechwa (simple verse)</td>
<td>SARMA(1,1)</td>
<td>57.8%</td>
<td>62.8%</td>
<td>68.8%</td>
</tr>
<tr>
<td>Juliusz Słowacki (complex verse)</td>
<td>SARMA(0,1)(1,1)</td>
<td>45.0%</td>
<td>48.7%</td>
<td>50.6%</td>
</tr>
<tr>
<td>John-Paul the 2nd (rhetoric discourse)</td>
<td>MA(2)</td>
<td>39.4%</td>
<td>43.1%</td>
<td>39.3%</td>
</tr>
<tr>
<td>Igor Newerly (literary prose)</td>
<td>MA(2)</td>
<td>36.1%</td>
<td>31.3%</td>
<td>32.9%</td>
</tr>
</tbody>
</table>

The evidence is sufficient however to establish that MA(2) is the best linear time-domain model describing the sequence of stressed and un-stressed syllables in Polish. This statement is validated by the fact that if we remove ("filter out") the seasonal component responsible for the 11-syllable cycle in Słowacki’s poem (for that purpose, instead of \( \text{SARMA}(0,1)(1,1) \), we estimate \( \text{SARMA}(1,1)_{11} \)), the remaining residual series will resemble the series generated by the literary prose (we have not discovered the same regularity in Brzechwa’s text). We present below the ACF and PACF functions of the residual series obtained in this way (Fig. 17 and Fig. 18). If we compare these graphics with those of the papal homily and Newerly’s text (Fig. 9–10 and Fig. 13–14), we will notice that they are very similar.
REFERENCES


APPENDIX

Texts Used

Jan Brzechwa:
Performance: Grażyna Barszczewska, Jerzy Bończak, Wiesław Dzewicz, Marek Kendrat, Jerzy Zenuik and others.

Juliusz Słowacki:
Beniowski, ed. Państwowy Instytut Wydawniczy, Warszawa 1974, Samples: first canto verses 1-72, second canto verses 1-72, third canto verses 1-72; Recording: Editions and Recordings by the Polish Association for the Blind, Warsaw 1981.
Performance: Stanisław Zaczuk.

John Paul II:
Homily pronounced by the pope in Wrocklaw on the 21st of June 1983 on the occasion of his second pilgrimage to Poland (private recording).

Igor Newely:
Recording: Editions and Recordings by the Polish Association for the Blind, Warsaw 1988.
Performance: Henryk Machalica.