Marking up Dramatic Text: a Case Study of “7 stories” by Morris Panych

Ivan Bekhta[a], Nataliia Hrytsiv[b] and Anastasiia Matviychuk[b]

[a] Lviv Franko National University, Universytetska Street, 1, Lviv, 79000, Ukraine
[b] Lviv Polytechnic National University, Stepana Bandery Street, 12, Lviv, 79000, Ukraine

Abstract

The paper elucidates the process, challenges and results of using computational linguistics tools (NLP) and pre-computer technique (TEI for personage utterance tagging) in processing dramatic text. As the material for analysis we have chosen the modern play “7 stories” of the Canadian playwright Morris Panych, researched from the viewpoint of statistical indicator’s and textual coefficients. Special attention is paid to statistical parameters of main personages in the play. Results obtained show numeric characteristics of such data: number of meanings (N); maximal meaning (max); minimal meaning (min); range (R); mode (Mo); median (Md); mean ($\dot{X}$); standard deviation (Ϭ); coefficient of variation (ν); standard error ($S\dot{x}$); measurement error (ε).

Keywords [1]

Translation, NLP, quantitative analysis, text mark-up, applied linguistics, drama text, tagging.

1. Introduction

Modern practices and techniques of using markup and NLP tools have proved its efficiency in relation to systematic processing and analysis of texts, which results in generating novel systems and well-elaborated tools for language processing. It is a powerful and promising technology to objectively reconstruct arguments in order to amply exemplify its findings and formulate well-grounded hypotheses.

This study is part of a larger research project on the creativity of Morris Panych and the reception of his writings via translation(s).

Presented in the paper is the preliminary algorithm for developing analytical findings concerning the reasons behind deviation within the aspect of statistical parameters of a source and target texts. Further elaborations are forthcoming.

In focus is the drama “7 stories” of Morris Panych [1] and its Ukrainian translation [2]. We briefly discuss key characteristics of marking-up dramatic text and illustrate the results obtained; we also demonstrate its primary advances.

Based on the aquired evidence, certain considerations are made regarding the usage of quantitative comparable analysis for further comparison of ST and TT statistics and ratio findings.

Thus, linguistic research determines an effective approach to the study of text, using mathematical methods and tools in combination with computer technology, which open new horizons for the linguistic analysis of new broad perspectives.

A complete and comprehensive description of language and speech requires a diligent insight into both qualitative and quantitative features of linguistic objects.

COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine
EMAIL: ivan.bekhta@gmail.com (I. Bekhta); nataliia.m.hrytsiv@lpnu.ua (N. Hrytsiv); anastasija838@gmail.com (A. Matviychuk)
ORCID: 0000-0002-9848-1505 (I. Bekhta); 0000-0001-6660-7161 (N. Hrytsiv); 0000-0001-6056-1799 (A. Matviychuk)

In addition, an approach towards and detailed study dramatic texts, as a unique literary genre, is a separate challenge in current studies, which has special requirements within NLP tools application and text mark-up. Therefore, the study of Morris Panych's playwork "7 Stories" is relevant.

The idea is that modern Canadian drama is the aspect, little studied from numerous viewpoints, i.e. philological, translatological, rhethorical; however, least studied from the angle of mathematical linguistics and statistics.

In order to understand the specifics of dramatic works, the concept of author's style, postmodern literature, to which the work under study belongs, the life path of the author and translator were additionally considered.

The play "7 Stories" by Morris Panych and translated by Ivan Krychfalushiy is an example of postmodern literature that has become a challenge and opposition to the laws of modernism.

2. Method and preparation characteristics

Considering the vast quantities of ST and TT data available today for analysis, as discussed in [3, 4, 5, 6], Natural Language Processing is among most interesting and promising aspects of data science [7, 8, 9, 10, 11, 12, 13].

By default, text data of the original text is difficult to process [14, 15, 16, 17, 18] given the challenge of comparing/contrasting it to the translated drama text [19, 20, 21, 22, 23, 24], the task can be complicated [25, 26, 27, 28], though, incredibly appealing [29, 30, 31, 32, 33].

Within this study project, we opted for exploring the way NLP techniques, especially mark-up possibilities, can advance processing performing/drama text for statistical profiling of ST and TT.

The project outlined in the current paper explores the ddistribution of the number of words in a sentence as well as other numeric characteristics being analyzed collectively and for all the characters of drama under analysis in their contrast with the Ukrainian translation.

2.1. Stages of working with the text document "7 stories" by Morris Panych

A number of actions were performed for statistical analysis. Therefore, the analysis took place in the following stages:

- The books of the original text and the translation were pre-scanned for further manipulations using ABBYY Fine Reader software;
- Afterwards, it was converted from pdf to .docx to make it possible to work with text in terms of mark-up;
- The correct formatting of text was checked and discrepancies between scanned pdf file and text documents were detected; it was normalized in the MS Word editor;

Next, the focus was on:

- Selection of text marking up system according to its features;
- Implementation of proper tags for the original work
- Implementation of proper tags for the translated version;
- Calculated texts results were processed using the Python programming language;
- Afterwards, the results of the statistical parameters, such as *N, max, min, R, Mo, Md, Ẋ, Ϭ, v, Sẋ, ε* were analyzed and described.

The original text and its translation was marked up using the same marking rules.

To recall, the use was made of the XML (eXtensible Markup Language) – a text markup language. It was used to conduct research and implement on the structural level.

The XML language was preffered since it fully determines the logical structure of a document.

The task of the XML language is to ensure certain data: images, texts, and other parts of a Web document; it can be defined and structured regardless of the platform used to recreate them.

Since in the current paper we deal with a dramatic work, text mark up and tag patterns were selected and adjusted for the appropriate analysis of this type of work. Thus, let us now turn our sights to text mark-up system, peculiar to drama text.

2.2. Mark-up pattern

2.2.1. Pattern

Thus, the following text markings were chosen according to the features of the dramatic work:

<chtr>...</chtr> — paired marking, which is used to indicate a solid whole part of the text related to a particular character;

<cnm>...</cnm> — paired marking, which is used to indicate the name of the character with a colon;

<s>...</s> —— paired marking, which is used to denote a sentence in the speech of the character;

<mtr>...</mtr> — paired marking, which is used to mark all author's remarks throughout the text

2.2.2. Example

<mtr>The action of the play takes place outside an apartment building-on the ledge, outside various windows of the seventh storey. As the play progresses, the lights emphasize the time elapsed between early evening and late night. As the play opens, we hear a party in progress from one of the windows, MAN stands on the ledge, in a state of perplexity, contemplating the depths below. He seems disturbed, confused. Then he comes to what seems to be a resolution. He prepares to jump. When he is about to leap, the window next to him flies open. CHARLOTTE appears. She holds a MAN wAllet, which she attempts to throw out the window, RODNEY,charging up from behind, grabs her hand. A window-ledge struggle ensues.</mtr>

<chtr><cnm>CHARLOTTE</cnm>

<s>Let GO of me!!!</s><s> Let GO!!</s></chtr>

<chtr><cnm>RODNEY</cnm>

<mtr>(threatening)</mtr><s> So-help-me-GOD, CHARLOTTE. </s></chtr>

<chtr><cnm>CHARLOTTE</cnm>

<mtr>(daring him)</mtr><s> What??</s><s> WHAT??!! </s></chtr>

<chtr><cnm>RODNEY</cnm>

<s>Give me back my wallet! </s></chtr>

<mtr>*She tries to throw it again. They struggle*. </mtr>

<chtr><cnm>RODNEY</cnm>

<s>What's WRONG with you?</s><s> Are you CRAZY?! </s></chtr>

<chtr><cnm>CHARLOTTE</cnm>

<s>YES! </s><s>YES, I AM!!! </s></chtr>

<chtr><cnm>RODNEY</cnm>

<s>MY GOLD CARD is in there!! </s></chtr>

3. Results

This section of the study presents statistics taken from the calculation of data based on the number of words in a sentence. That is, the unit of measurement in this statistical calculation is the *word*. The findings illustrate the contrast of ST and TT results of statistical parameters, i.e. N, max, min, R, Mo, Md, Ẋ, Ϭ, ν, Sẋ, ε. The schematic representation follows the data of each drama character one by one.

3.1. Analysis of the part of the text that belongs to the drama character of "Charlotte"

Having analysed the distribution of the number of words in a sentence by absolute and relevant frequency, we have obtained such numeric characteristics:

Charlotte: the whole ST data: 1 — 58 (90,62%); 2 — 4 (6,25%); 3 — 1 (1,56%); 4 — 1 (1,56%);.

The data for «Charlotte» presupposes that the absolute frequency of sentence lengths with word number 1 equals to 58; consequently, with word number of 2 equals to 4; with word number 3 equals to 1; with word number of 4 equals to 1.

Talking about translation, the most frequent are sentences with the number of words that equals to 1.

Charlotte: the whole TT data: **1 — 35 (30,97%);** 4 — 17 (15,04%); 5 — 13 (11,50%); 2 — 12 (10,62%); 6 — 12 (10,62%); 3 — 10 (8,85%); 7 — 5 (4,42%); 11 — 3 (2,65%); 9 — 2 (1,77%); 10 — 2 (1,77%); 8 — 1 (0,88%); 12 — 1 (0,88%). The last two are the least frequent.

On the basis of the data above the following calculations are made of number of meanings, maximal meaning, minimal meaning, range, mode, median, mean, standard deviation, coefficient of variation, standard error, measurement error.

Results are presented in Table 1.

Table 1

CHARLOTTE Numeric characteristics of word distribution within the sentence of the drama

Unit	ST	TT
N	64	113
max	4	12
min	1	1
R	3	11
Mo	1	1
Md	2,5	6,5
Ẋ	1,14	3,72
Ϭ	0,25	2,69
v	0,4347	0,7243
Sẋ	0,062	0,2532
ε	0,1065	0,1335

Table 1 proves the following results:

ST data numeric characteristic: Number of meanings (N) — 64; maximal meaning (max) — 4; minimal meaning (min) — 1; range (R) — 3; mode (Mo) — 1; median (Md) — 2,5; mean (Ẋ) — 1,14; standard deviation (Ϭ) — 0,50; coefficient of variation (v) — 0,4347; standard error (Sẋ) — 0,0620; measurement error (ε) — 0,1065.

TT data numeric characteristic: Number of meanings (N) — 113; maximal meaning (max) — 12; minimal meaning (min) — 1; range (R) — 11; mode (Mo) — 1; median (Md) — 6,5; mean (Ẋ) — 3,72; standard deviation (Ϭ) — 2,69; coefficient of variation (v) — 0,7243; standard error (Sẋ) — 0,2532; measurement error (ε) — 0,1335.

3.2. Analysis of the part of the text that belongs to the drama character of "Rodney"

Having analysed the distribution of the number of words in a sentence by absolute and relevant frequency, we have obtained such numeric characteristics:

Rodney: the whole ST data: 1 — 37 (90,24%); 2 — 3 (7,32%); 3 — 1 (2,44%).

The data for «Rodney» presupposes that the absolute frequency of sentence lengths with word number 1 equals to 37; consequently, with word number of 2 equals to 3; with word number 3 equals to 1.

Rodney: the whole TT data: **1 — 16 (21,05%)**; 2 — 13 (17,11%); 3 — 11 (14,47%); 4 — 9 (11,84%); 5 — 9 (11,84%); 6 — 7 (9,21%); 7 — 5 (6,58%); 9 — 3 (3,95%); 8 — 2 (2,63%); 10 — 1 (1,32%).

Based on the data above, the following calculations are made and presented in Table 2.

Table 2
RODNEY Numeric characteristics of word distribution within the sentence of the drama

Unit	ST	TT
N	41	76
max	3	10
min	1	1
R	2	9
Mo	1	1
Md	2	5,5
Ẋ	1,12	3,76
Ϭ	0,39	2,37
ν	0,3519	0,6304
Sẋ	0,0617	0,2721
ε	0,1077	0,1417

Table 2 shows the following results:

ST data numeric characteristic: Number of meanings (N) — 41; maximal meaning (max) — 3; minimal meaning (min) — 1; range (R) — 2; mode (Mo) — 1; median (Md) — 2,0; mean (Ẋ) — 1,12; standard deviation (Ϭ) — 0,39; coefficient of variation (ν) — 0,3519; standard error (Sẋ) — 0,0617; measurement error (ε) — 0,1077.

TT data numeric characteristic: Number of meanings (N) — 76; maximal meaning (max) — 10; minimal meaning (min) — 1; range (R) — 9; mode (Mo) — 1; median (Md) — 5,5; mean (Ẋ) — 3,76; standard deviation (Ϭ) — 2,37; coefficient of variation (ν) — 0,6304; standard error (Sẋ) — 0,2721; measurement error (ε) — 0,1417.

3.3. Analysis of the part of the text that belongs to the drama character of "Man"

By analogue to the previous characters (Charlotte and Rondey) we obtain the results for other characters; here – Man.

Man: the whole ST data: 1 — 228 (87,36%); 2 — 27 (10,34%); 3 — 6 (2,30%).

Thus, the data for «Man» states that the absolute frequency of sentence lengths with word number 1 equals to 228; consequently, with word number of 2 equals to 27; with word number 3 equals to 6.

Man: the whole TT data: **1 — 99 (18,50%);** 3 — 90 (16,82%); 4 — 78 (14,58%); 2 — 61 (11,40%); 5 — 48 (8,97%); 6 — 47 (8,79%); 7 — 37 (6,92%); 8 — 18 (3,36%); 9 — 16 (2,99%); 10 — 9 (1,68%); 11 — 8 (1,50%); 12 — 7 (1,31%); 15 — 4 (0,75%); 13 — 3 (0,56%); 16 — 3 (0,56%); 18 — 2 (0,37%); 14 — 1 (0,19%); 17 — 1 (0,19%); 19 — 1 (0,19%); 23 — 1 (0,19%); 27 — 1 (0,19%).

Next, we have calculated number of meanings, maximal meaning, minimal meaning, range, mode, median, mean, standard deviation, coefficient of variation, standard error, measurement error. The results are demonstrated in Table 3.

Table 3
MAN Numeric characteristics of word distribution within the sentence of the drama

Unit	ST	TT
N	261	535
max	3	27
min	1	1
R	2	26
Mo	1	1

Md	2	11
Ẋ	1,15	4,52
Ϭ	0,42	3,47
ν	0,3619	0,7678
Sẋ	0,0258	0,15
ε	0,0439	0,0651

ST data numeric characteristic: Number of meanings (N) — 261; maximal meaning (max) — 3; minimal meaning (min) — 1; range (R) — 2; mode (Mo) — 1; median (Md) — 2,0; mean (Ẋ) — 1,15; standard deviation (Ϭ) — 0,42; coefficient of variation (ν) — 0,3619; standard error (Sẋ) — 0,0258; measurement error (ε) — 0,0439.

TT data numeric characteristic: Number of meanings (N) — 535; maximal meaning (max) — 27; minimal meaning (min) — 1; range (R) — 26; mode (Mo) — 1; median (Md) — 11,0; mean (Ẋ) — 4,52; standard deviation (Ϭ) — 3,47; coefficient of variation (ν) — 0,7678; standard error (Sẋ) — 0,1500; measurement error (ε) — 0,0651.

3.4. Analysis of the part of the text that belongs to the drama character of "Leonard"

By analogue to the previous characters we obtain the results for the character – Leonard.

Leonard: the whole ST data: 1 — 92 (86,79%); 2 — 12 (11,32%); 3 — 2 (1,89%).

Leonard: the whole TT data: **1 — 30 (14,49%);** 5 — 28 (13,53%); 2 — 27 (13,04%); 3 — 27 (13,04%); 4 — 24 (11,59%); 6 — 23 (11,11%); 8 — 15 (7,25%); 7 — 8 (3,86%); 9 — 5 (2,42%); 10 — 5 (2,42%); 12 — 4 (1,93%); 14 — 3 (1,45%); 13 — 2 (0,97%); 16 — 2 (0,97%); 17 — 2 (0,97%); 11 — 1 (0,48%); 19 — 1 (0,48%).

Table 4

LEONARD Numeric characteristics of word distribution within the sentence of the drama

Unit	ST	TT
N	106	207
max	3	19
min	1	1
R	2	18
Mo	1	1
Md	2	9
Ẋ	1,15	4,94
Ϭ	0,41	3,53
ν	0,3539	0,7148
Sẋ	0,0396	0,2453
ε	0,0674	0,0974

ST data numeric characteristic:

Number of meanings (N) — 106; maximal meaning (max) — 3; minimal meaning (min) — 1; range (R) — 2; mode (Mo) — 1; median (Md) — 2,0; mean (Ẋ) — 1,15; standard deviation (Ϭ) — 0,41; coefficient of variation (ν) — 0,3539; standard error (Sẋ) — 0,0396; measurement error (ε) — 0,0674.

TT data numeric characteristic:

Number of meanings (N) — 207; maximal meaning (max) — 19; minimal meaning (min) — 1; range (R) — 18; mode (Mo) — 1; median (Md) — 9,0; mean (Ẋ) — 4,94; standard deviation (Ϭ) —

3,53; coefficient of variation (ν) — 0,7148; standard error (Sẋ) — 0,2453; measurement error (ε) — 0,0974.

3.5. Analysis of the part of the text that belongs to the drama character of "Jennifer"

Jennifer: the whole ST data: 1 — 21 (84,00%); 2 — 3 (12,00%); 6 — 1 (4,00%);.

Jennifer: the whole TT data: **6 — 5 (19,23%);** 4 — 4 (15,38%); 2 — 3 (11,54%); 3 — 3 (11,54%); 5 — 2 (7,69%); 9 — 2 (7,69%); 1 — 1 (3,85%); 7 — 1 (3,85%); 8 — 1 (3,85%); 10 — 1 (3,85%); 11 — 1 (3,85%); 14 — 1 (3,85%); 15 — 1 (3,85%).

Table 5

JENNIFER Numeric characteristics of word distribution within the sentence of the drama

Unit	ST	TT
N	25	26
max	6	15
min	1	1
R	5	14
Mo	1	6
Md	2	7
Ẋ	1,32	5,96
Ϭ	1,01	3,55
ν	0,7642	0,5948
Sẋ	0,2018	0,6955
ε	0,2996	0,2287

ST data numeric characteristic: Number of meanings (N) — 25; maximal meaning (max) — 6; minimal meaning (min) — 1; range (R) — 5; mode (Mo) — 1; median (Md) — 2,0; mean (Ẋ) — 1,32; standard deviation (Ϭ) — 1,01; coefficient of variation (ν) — 0,7642; standard error (Sẋ) — 0,2018; measurement error (ε) — 0,2996.

TT data numeric characteristic: Number of meanings (N) — 26; maximal meaning (max) — 15; minimal meaning (min) — 1; range (R) — 14; mode (Mo) — 6; median (Md) — 7,0; mean (Ẋ) — 5,96; standard deviation (Ϭ) — 3,55; coefficient of variation (ν) — 0,5948; standard error (Sẋ) — 0,6955; measurement error (ε) — 0,2287.

3.6. Analysis of the part of the text that belongs to the drama character of "Marshall"

Marshall: the whole ST data: 1 — 94 (85,45%); 2 — 15 (13,64%); 4 — 1 (0,91%).

Marshall: the whole TT data: **2 — 31 (15,74%);** 4 — 27 (13,71%); 3 — 26 (13,20%); 5 — 25 (12,69%); 6 — 21 (10,66%); 8 — 16 (8,12%); 7 — 11 (5,58%); 9 — 11 (5,58%); 1 — 9 (4,57%); 10 — 7 (3,55%); 11 — 6 (3,05%); 12 — 2 (1,02%); 16 — 2 (1,02%); 17 — 2 (1,02%); **23 — 1 (0,51%).**

Table 6

MARSHALL Numeric characteristics of word distribution within the sentence of the drama

Unit	ST	TT
N	110	197
max	4	23
min	1	1

R	3	22
Mo	1	2
Md	2	8
Ẋ	1,16	5,39
Ϭ	0,44	3,38
ν	0,376	0,6279
Sẋ	0,0417	0,2409
ε	0,0703	0,0877

ST data numeric characteristic: Number of meanings (N) — 110; maximal meaning (max) — 4; minimal meaning (min) — 1; range (R) — 3; mode (Mo) — 1; median (Md) — 2,0; mean (Ẋ) — 1,16; standard deviation (Ϭ) — 0,44; coefficient of variation (ν) — 0,3760; standard error (Sẋ) — 0,0417; measurement error (ε) — 0,0703.

TT data numeric characteristic: Number of meanings (N) — 197; maximal meaning (max) — 23; minimal meaning (min) — 1; range (R) — 22; mode (Mo) — 2; median (Md) — 8,0; mean (Ẋ) — 5,39; standard deviation (Ϭ) — 3,38; coefficient of variation (ν) — 0,6279; standard error (Sẋ) — 0,2409; measurement error (ε) — 0,0877.

3.7. Analysis of the part of the text that belongs to the drama character of "Joan"

Joan: the whole ST data: 1 — 43 (84,31%); 2 — 7 (13,73%); 3 — 1 (1,96%);.

Joan: the whole TT data: **3 — 16 (16,49%);** 4 — 16 (16,49%); 1 — 13 (13,40%); 5 — 12 (12,37%); 2 — 10 (10,31%); 7 — 10 (10,31%); 6 — 6 (6,19%); 9 — 4 (4,12%); 8 — 3 (3,09%); 12 — 3 (3,09%); 11 — 1 (1,03%); 14 — 1 (1,03%); 17 — 1 (1,03%); 18 — 1 (1,03%).

Table 7

JOAN Numeric characteristics of word distribution within the sentence of the drama

Unit	ST	TT
N	51	97
max	3	18
min	1	1
R	2	17
Mo	1	3
Md	2	7,5
Ẋ	1,18	4,81
Ϭ	0,43	3,35
ν	0,3651	0,6958
Sẋ	0,0602	0,3402
ε	0,1002	0,1385

ST data numeric characteristic: Number of meanings (N) — 51; maximal meaning (max) — 3; minimal meaning (min) — 1; range (R) — 2; mode (Mo) — 1; median (Md) — 2,0; mean (Ẋ) — 1,18; standard deviation (Ϭ) — 0,43; coefficient of variation (ν) — 0,3651; standard error (Sẋ) — 0,0602; measurement error (ε) — 0,1002.

TT data numeric characteristic: Number of meanings (N) — 97; maximal meaning (max) — 18; minimal meaning (min) — 1; range (R) — 17; mode (Mo) — 3; median (Md) — 7,5; mean (Ẋ) — 4,81; standard deviation (Ϭ) — 3,35; coefficient of variation (ν) — 0,6958; standard error (Sẋ) — 0,3402; measurement error (ε) — 0,1385.

3.8. Analysis of the part of the text that belongs to the drama character of "Michael"

Michael: the whole ST data: 1 — 34 (91,89%); 2 — 3 (8,11%);.

Michael: the whole TT data: **4 — 11 (20,00%)**; 5 — 10 (18,18%); 3 — 7 (12,73%); 7 — 7 (12,73%); 6 — 6 (10,91%); 2 — 5 (9,09%); 12 — 3 (5,45%); 1 — 2 (3,64%); 8 — 2 (3,64%); 10 — 2 (3,64%).

Table 8

MICHAEL Numeric characteristics of word distribution within the sentence of the drama

Unit	ST	TT
N	37	55
max	2	12
min	1	1
R	1	11
Mo	1	4
Md	1,5	5,5
Ẋ	1,08	5,16
Ϭ	0,27	2,57
ν	0,2525	0,4979
Sẋ	0,0449	0,3467
ε	0,0814	0,1316

ST data numeric characteristic: Number of meanings (N) — 37; maximal meaning (max) — 2; minimal meaning (min) — 1; range (R) — 1; mode (Mo) — 1; median (Md) — 1,5; mean (Ẋ) — 1,08; standard deviation (Ϭ) — 0,27; coefficient of variation (ν) — 0,2525; standard error (Sẋ) — 0,0449; measurement error (ε) — 0,0814.

TT data numeric characteristic: Number of meanings (N) — 55; maximal meaning (max) — 12; minimal meaning (min) — 1; range (R) — 11; mode (Mo) — 4; median (Md) — 5,5; mean (Ẋ) — 5,16; standard deviation (Ϭ) — 2,57; coefficient of variation (ν) — 0,4979; standard error (Sẋ) — 0,3467; measurement error (ε) — 0,1316.

3.9. Analysis of the part of the text that belongs to the drama character of "Rachel"

Rachel: the whole ST data: 1 — 53 (91,38%); 2 — 5 (8,62%).

Rachel: the whole TT data: **4 — 18 (15,00%);** 5 — 14 (11,67%); 7 — 14 (11,67%); 3 — 12 (10,00%); 2 — 11 (9,17%); 6 — 11 (9,17%); 1 — 10 (8,33%); 8 — 5 (4,17%); 9 — 4 (3,33%); 11 — 4 (3,33%); 10 — 3 (2,50%); 12 — 3 (2,50%); 13 — 3 (2,50%); 14 — 3 (2,50%); 16 — 3 (2,50%); 15 — 1 (0,83%); 20 — 1 (0,83%).

Table 9

RACHEL Numeric characteristics of word distribution within the sentence of the drama

Unit	ST	TT
N	58	120
max	2	20
min	1	1

R	1	19
Mo	1	4
Md	1,5	9
Ẋ	1,09	6,03
Ϭ	0,28	3,94
ν	0,2584	0,6529
Sẋ	0,0369	0,3596
ε	0,0665	0,1168

ST data numeric characteristic: Number of meanings (N) — 58; maximal meaning (max) — 2; minimal meaning (min) — 1; range (R) — 1; mode (Mo) — 1; median (Md) — 1,5; mean (Ẋ) — 1,09; standard deviation (Ϭ) — 0,28; coefficient of variation (ν) — 0,2584; standard error (Sẋ) — 0,0369; measurement error (ε) — 0,0665.

TT data numeric characteristic: Number of meanings (N) — 120; maximal meaning (max) — 20; minimal meaning (min) — 1; range (R) — 19; mode (Mo) — 4; median (Md) — 9,0; mean (Ẋ) — 6,03; standard deviation (Ϭ) — 3,94; coefficient of variation (ν) — 0,6529; standard error (Sẋ) — 0,3596; measurement error (ε) — 0,1168.

3.10. Analysis of the part of the text that belongs to the drama character of "Percy"

Percy: the whole ST data: 1 — 34 (80,95%); 2 — 7 (16,67%); 3 — 1 (2,38%).

Percy: the whole TT data: 6 — 12 (16,44%); 3 — 11 (15,07%); 4 — 10 (13,70%); 5 — 7 (9,59%); 1 — 6 (8,22%); 2 — 5 (6,85%); 7 — 4 (5,48%); 8 — 4 (5,48%); 9 — 3 (4,11%); 11 — 3 (4,11%); 14 — 3 (4,11%); 10 — 2 (2,74%); 12 — 1 (1,37%); 18 — 1 (1,37%); 23 — 1 (1,37%).

Table 10

PERCY Numeric characteristics of word distribution within the sentence of the drama

Unit	ST	TT
N	42	73
max	3	23
min	1	1
R	2	22
Mo	1	6
Md	2	8
Ẋ	1,21	5,9
Ϭ	0,46	4,04
ν	0,3827	0,6839
Sẋ	0,0717	0,4727
ε	0,1158	0,1569

ST data numeric characteristic: Number of meanings (N) — 42; maximal meaning (max) — 3; minimal meaning (min) — 1; range (R) — 2; mode (Mo) — 1; median (Md) — 2,0; mean (Ẋ) — 1,21; standard deviation (Ϭ) — 0,46; coefficient of variation (ν) — 0,3827; standard error (Sẋ) — 0,0717; measurement error (ε) — 0,1158.

TT data numeric characteristic: Number of meanings (N) — 73; maximal meaning (max) — 23; minimal meaning (min) — 1; range (R) — 22; mode (Mo) — 6; median (Md) — 8,0; mean (Ẋ) — 5,90; standard deviation (Ϭ) — 4,04; coefficient of variation (ν) — 0,6839; standard error (Sẋ) — 0,4726; measurement error (ε) — 0,1569.

3.11. Analysis of the part of the text that belongs to the drama character of "Al"

Al: the whole ST data: 1 — 23 (74,19%); 2 — 6 (19,35%); 3 — 2 (6,45%).

Al: the whole TT data: **6 — 11 (18,97%);** 3 — 8 (13,79%); 4 — 7 (12,07%); 5 — 7 (12,07%); 1 — 6 (10,34%); 2 — 5 (8,62%); 7 — 3 (5,17%); 10 — 3 (5,17%); 9 — 2 (3,45%); 12 — 2 (3,45%); 8 — 1 (1,72%); 13 — 1 (1,72%); 14 — 1 (1,72%); 16 — 1 (1,72%).

Table 11

AL Numeric characteristics of word distribution within the sentence of the drama

Unit	ST	TT
N	31	58
max	3	16
min	1	1
R	2	15
Mo	1	6
Md	2	7,5
Ẋ	1,32	5,4
Ϭ	0,59	3,41
ν	0,4457	0,6316
Sẋ	0,1059	0,4476
ε	0,1569	0,1626

ST data numeric characteristic: Number of meanings (N) — 31; maximal meaning (max) — 3; minimal meaning (min) — 1; range (R) — 2; mode (Mo) — 1; median (Md) — 2,0; mean (Ẋ) — 1,32; standard deviation (Ϭ) — 0,59; coefficient of variation (ν) — 0,4457; standard error (Sẋ) — 0,1059; measurement error (ε) — 0,1569.

TT data numeric characteristic: Number of meanings (N) — 58; maximal meaning (max) — 16; minimal meaning (min) — 1; range (R) — 15; mode (Mo) — 6; median (Md) — 7,5; mean (Ẋ) — 5,40; standard deviation (Ϭ) — 3,41; coefficient of variation (ν) — 0,6316; standard error (Sẋ) — 0,4476; measurement error (ε) — 0,1626.

3.12. Analysis of the part of the text that belongs to the drama character of "Nurse Wilson"

Nurse Wilson: the whole ST data: 1 — 42 (87,50%); 2 — 5 (10,42%); 3 — 1 (2,08%);.

Nurse Wilson: the whole TT data: **3 — 10 (13,16%);** 4 — 10 (13,16%); 1 — 9 (11,84%); 5 — 9 (11,84%); 2 — 7 (9,21%); 6 — 6 (7,89%); 7 — 6 (7,89%); 12 — 4 (5,26%); 8 — 3 (3,95%); 9 — 3 (3,95%); 11 — 2 (2,63%); 13 — 2 (2,63%); 18 — 2 (2,63%); 10 — 1 (1,32%); 17 — 1 (1,32%); 23 — 1 (1,32%).

Table 12

NURSE WILSON Numeric characteristics of word distribution within the sentence of the drama

Unit	ST	TT
N	48	76
max	3	23
min	1	22
R	2	3

Mo	1	3
Md	2	8,5
Ẋ	1,15	5,91
Ϭ	0,41	4,48
ν	0,3558	0,7586
Sẋ	0,0588	0,5141
ε	0,1007	0,1705

ST data numeric characteristic: Number of meanings (N) — 48; maximal meaning (max) — 3; minimal meaning (min) — 1; range (R) — 2; mode (Mo) — 1; median (Md) — 2,0; mean (Ẋ) — 1,15; standard deviation (Ϭ) — 0,41; coefficient of variation (ν) — 0,3558; standard error (Sẋ) — 0,0588; measurement error (ε) — 0,1007.

TT data numeric characteristic: Number of meanings (N) — 76; maximal meaning (max) — 23; minimal meaning (min) — 1; range (R) — 22; mode (Mo) — 3; median (Md) — 8,5; mean (Ẋ) — 5,91; standard deviation (Ϭ) — 4,48; coefficient of variation (ν) — 0,7586; standard error (Sẋ) — 0,5141; measurement error (ε) — 0,1705.

3.13. Analysis of the part of the text that belongs to the drama character of "Lilian"

Lilian: the whole ST data: 1 — 68 (91,89%); 2 — 6 (8,11%);.

Lilian: the whole TT data: **2 — 23 (14,94%);** 4 — 19 (12,34%); 3 — 18 (11,69%); 5 — 17 (11,04%); 1 — 14 (9,09%); 6 — 13 (8,44%); 10 — 12 (7,79%); 8 — 10 (6,49%); 7 — 9 (5,84%); 9 — 7 (4,55%); 11 — 3 (1,95%); 13 — 2 (1,30%); 16 — 2 (1,30%); 17 — 2 (1,30%); 12 — 1 (0,65%); 14 — 1 (0,65%); 18 — 1 (0,65%).

Table 13

LILIAN Numeric characteristics of word distribution within the sentence of the drama

Unit	ST	TT
N	74	154
max	2	18
min	1	1
R	1	17
Mo	1	2
Md	1,5	9
Ẋ	1,08	5,51
Ϭ	0,27	3,69
ν	0,2525	0,6704
Sẋ	0,0317	0,2975
ε	0,0575	0,1059

ST data numeric characteristic: Number of meanings (N) — 74; maximal meaning (max) — 2; minimal meaning (min) — 1; range (R) — 1; mode (Mo) — 1; median (Md) — 1,5; mean (Ẋ) — 1,08; standard deviation (Ϭ) — 0,27; coefficient of variation (ν) — 0,2525; standard error (Sẋ) — 0,0317; measurement error (ε) — 0,0575.

TT data numeric characteristic: Number of meanings (N) — 154; maximal meaning (max) — 18; minimal meaning (min) — 1; range (R) — 17; mode (Mo) — 2; median (Md) — 9,0; mean (Ẋ) — 5,51; standard deviation (Ϭ) — 3,69; coefficient of variation (ν) — 0,6704; standard error (Sẋ) — 0,2975; measurement error (ε) — 0,1059.

3.14. Analysis of the part of the text that belongs to the secondary drama characters

3.14.1. Character "One"

One: the whole ST data: 1 — 1 (100,00%). *One: the whole TT data*: 4 — 2 (40,00%); 3 — 1 (20,00%); 5 — 1 (20,00%); 6 — 1 (20,00%).

Table 14

ONE Numeric characteristics of word distribution within the sentence of the drama

Unit	ST	TT
N	1	5
max	1	6
min	1	3
R	0	3
Mo	1	4
Md	1	4,5
Ẋ	1	4,4
Ϭ	0	1,02
ν	0	0,2318
Sẋ	0	0,4561
ε	0	0,2032

ST data numeric characteristic: Number of meanings (N) — 1; maximal meaning (max) — 1; minimal meaning (min) — 1; range (R) — 0; mode (Mo) — 1; median (Md) — 1,0; mean (Ẋ) — 1,00; standard deviation (Ϭ) — 0,00; coefficient of variation (ν) — 0,0000; standard error (Sẋ) — 0,0000; measurement error (ε) — 0,0000.

TT data numeric characteristic: Number of meanings (N) — 5; maximal meaning (max) — 6; minimal meaning (min) — 3; range (R) — 3; mode (Mo) — 4; median (Md) — 4,5; mean (Ẋ) — 4,40; standard deviation (Ϭ) — 1,02; coefficient of variation (ν) — 0,2318; standard error (Sẋ) — 0,4561; measurement error (ε) — 0,2032.

3.14.2. Character "Two"

Two: the whole ST data: 1 — 2 (66,67%); 2 — 1 (33,33%). *Two: the whole TT data*: 4 — 2 (33,33%); 8 — 2 (33,33%); 5 — 1 (16,67%); 10 — 1 (16,67%).

Table 15

TWO Numeric characteristics of word distribution within the sentence of the drama

Unit	ST	TT
N	3	6
max	2	10
min	1	4
R	1	6
Mo	1	4
Md	1,5	6,5
Ẋ	1,33	6,5
Ϭ	0,47	2,29
ν	0,3536	0,3525
Sẋ	0,2722	0,9354
ε	0,4001	0,2821

ST data numeric characteristic: Number of meanings (N) — 3; maximal meaning (max) — 2; minimal meaning (min) — 1; range (R) — 1; mode (Mo) — 1; median (Md) — 1,5; mean (Ẋ) — 1,33; standard deviation (Ϭ) — 0,47; coefficient of variation (ν) — 0,3536; standard error (Sẋ) — 0,2722; measurement error (ε) — 0,4001.

TT data numeric characteristic: Number of meanings (N) — 6; maximal meaning (max) — 10; minimal meaning (min) — 4; range (R) — 6; mode (Mo) — 4; median (Md) — 6,5; mean (Ẋ) — 6,50; standard deviation (Ϭ) — 2,29; coefficient of variation (ν) — 0,3525; standard error (Sẋ) — 0,9354; measurement error (ε) — 0,2821.

3.14.3. Character "Three"

Three: the whole ST data: 1 — 4 (80,00%); 2 — 1 (20,00%).

Three: the whole TT data: **4 — 2 (40,00%)**; 3 — 1 (20,00%); 6 — 1 (20,00%); 8 — 1 (20,00%).

Table 16

THREE Numeric characteristics of word distribution within the sentence of the drama

Unit	ST	TT
N	5	5
max	2	8
min	1	3
R	1	5
Mo	1	4
Md	1,5	5
Ẋ	1,2	5
Ϭ	0,4	1,79
ν	0,3333	0,3578
Sẋ	0,1789	0,8
ε	0,2922	0,3136

ST data numeric characteristic: Number of meanings (N) — 5; maximal meaning (max) — 2; minimal meaning (min) — 1; range (R) — 1; mode (Mo) — 1; median (Md) — 1,5; mean (Ẋ) — 1,20; standard deviation (Ϭ) — 0,40; coefficient of variation (ν) — 0,3333; standard error (Sẋ) — 0,1789; measurement error (ε) — 0,2922.

TT data numeric characteristic: Number of meanings (N) — 5; maximal meaning (max) — 8; minimal meaning (min) — 3; range (R) — 5; mode (Mo) — 4; median (Md) — 5,0; mean (Ẋ) — 5,00; standard deviation (Ϭ) — 1,79; coefficient of variation (ν) — 0,3578; standard error (Sẋ) — 0,8000; measurement error (ε) — 0,3136.

3.14.4. Character "Four"

Four: the whole ST data: 1 — 2 (100,00%).

Four: the whole TT data: **4 — 2 (50,00%)**; 1 — 1 (25,00%); 2 — 1 (25,00%).

Table 17

FOUR Numeric characteristics of word distribution within the sentence of the drama

Unit	ST	TT
N	2	4
max	1	4
min	1	1
R	0	3

Mo	1	4
Md	1	2
Ẋ	1	2,75
Ϭ	0	1,3
ν	0	0,4724
Sẋ	0	0,6495
ε	0	0,4629

ST data numeric characteristic: Number of meanings (N) — 2; maximal meaning (max) — 1; minimal meaning (min) — 1; range (R) — 0; mode (Mo) — 1; median (Md) — 1,0; mean (Ẋ) — 1,00; standard deviation (Ϭ) — 0,00; coefficient of variation (ν) — 0,0000; standard error (Sẋ) — 0,0000; measurement error (ε) — 0,0000.

TT data numeric characteristic: Number of meanings (N) — 4; maximal meaning (max) — 4; minimal meaning (min) — 1; range (R) — 3; mode (Mo) — 4; median (Md) — 2,0; mean (Ẋ) — 2,75; standard deviation (Ϭ) — 1,30; coefficient of variation (ν) — 0,4724; standard error (Sẋ) — 0,6495; measurement error (ε) — 0,4629.

4. Comparative analysis of word distribution in sentences

4.1. Difference in the Number of meanings (N) in ST and TT

Given form the results above that the translated variant statistical parameters data exceeds the original drama in the majority of cases, we now turn our sights to one parameter – Number of meanings (N). We tend to compare the data and find the difference (if present). Our *assumption 1* is that the TT is much longer in terms of word usage within the sentence.

Character’s name	ST	TT	Difference
Charlotte	64	113	+49
Rodney	41	76	+35
Man	**261**	**535**	**+274**
Leonard	106	207	+101
Jennifer	25	26	+1
Marshal	110	197	+87
Joan	51	97	+46
Michael	37	55	+18
Rachel	58	120	+62
Percy	42	73	+31
Al	31	58	+27
Nurse Wilson	48	76	+28
Lilian	74	154	+80
One	1	5	+4
Two	3	6	+3
Three	5	5	0
Four	2	4	+2

Figure 1: Comparative statistics of Number of meanings (N) in ST and TT

To recall, character "Man" is the protagonist and the main character of the play. He is a well-dressed gentleman who is willing to jump off the seventh story.

He has a number of conversations with the residents of the building. He feels lost and compelled to stand on the seventh story of the building. Taking into account the results of Figure 1 we hold *assumption 2* that the translator adds a considerable number of words (274), or, he rather, doubles the ST quantity, due to a number of reasons:

- to explain the original;
- to compensate literary imagery losses;
- to add something from the translator himself, to recreate, so to say, the original;
- due to structural and lexico-gramatical allomorphic features of a language pair.

Whatever reason stands behind this translator’s decision-making, it is a prosperous ground for further Translation Studies analysis.

4.2. Analysis of the whole text

Here we focus on statistical parameters with the defined unit of measurement – a word. The number of words in a drama text utterunces is important due to a couple of reasons:

- the length of lines of the written script;
- chronometry and metrics of the whole drama act;
- pithiness and iconicity of each phrase.

Below are the results on the distribution of the number of words in a TT sentence by absolute and relevant frequency.

The most frequent are sentences in the translated text with the number of words **4 – 259 (14,2%),** 1 – 255(13,98%), 3– 255(13,98%), 2 – 219(12,01%) 5 – 205(11,24%), 6 -182 (9,98), 7-121 (6,63%), 8 – 84(4,61), 9- 62 (3,4%), 10 – 49 (2,69%), 11 – 32 (1,75), 12 – 31 (1,7%), 14 – 14 (0,77%), 13 – 13 (0,71%), 16 – 13(0,71%), 17 – 9 (0,49%), 18 – 7 (0,38%), 15 – 6 (0,33%), 23 – 4 (0,22%), 19 – 2 (0,11%), 20 – 1 (0,05%), 27 – 1 (0,05**%).** The last two results are the least frequent.

In the following Figure 2 we can see a comparison of the number of words in the sentences of the whole TT drama work.

The *x*-axis is the number of sentences, and the *y*-axis is the number of words in a sentence.

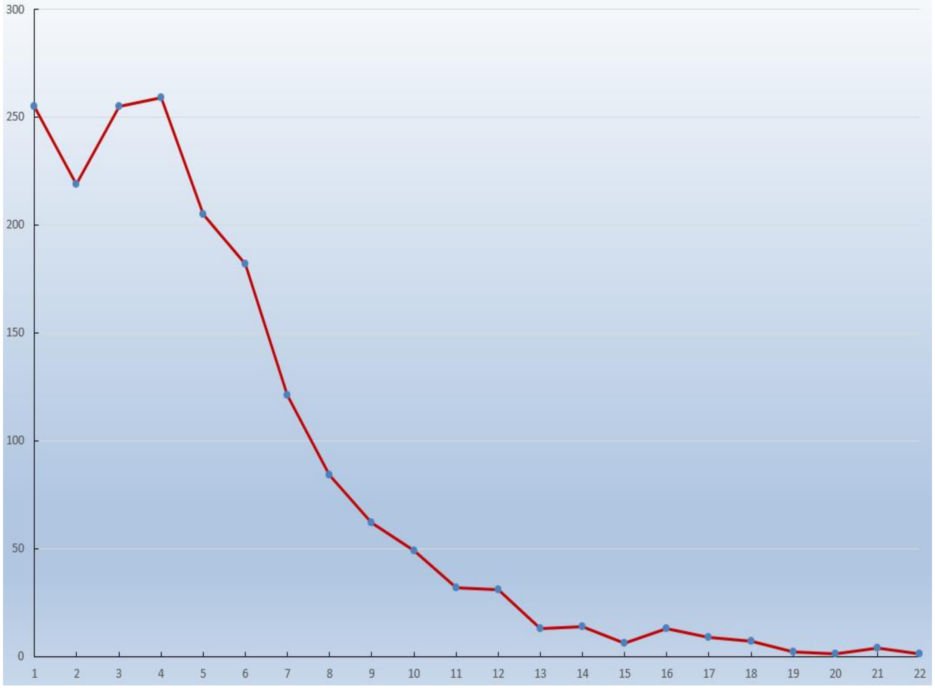

Figure 2: Number of words in sentences of TT

5. Conclusions

The main advances of statistical linguistics have been retrieved in the article. The original Canadian play has been compared with the corresponding translated text in terms of statistical parameters, which has never been done before.

The paper is of practical and applied value; however, the scientific value of the paper is seen as such that the suggested approach and methods will eventually allow formulating and substantiating a plausible scientific hypothesis in the realm of statistical linguistics and translation studies. At this point it is proven that bilingual drama texts are well adoptable for NLP and reveal promising outcomes.

We have verified absolute and relevant distribution, probability measurement, also: N, max, min, R, Mo, Md, Ẋ, Ϭ, ν, Sẋ, ε in the sentences of both texts.

Specifically designed software, which is represented as a combination of XML markup language, Microsoft Excel spreadsheet, and Python programming language, has been used. Results of statistical calculations of the drama "7 stories" by Morris Panych by unit of measure *word* are presented in the corresponding Tables 1 – 17.

Structural recognition provides useful information about the characters of the play, original and translation, namely the length of the sentence in *word* units that will help with further comparisons of ST and TT. The quantitative characteristics of the original play and its Ukrainian translation on the lexical level relying on the linguistic statistical analysis have been clarified: the amount of translated text Numbers of meaning (N) exceeds considerably and demands further analysis. The discrepancy becomes obvious with number of characters (Man, Leonard, Marshal, Lilian)

The correlation of coefficients has been presented in tables and figures to illustrate the material under research.

The prospect of the study is to further explore the problems of translator's meaningful choices which resulted in the declared above data.

6. Acknowledgement

The project has been carried out within the complex academic topic "Application of modern technologies for optimization of information processes in natural language" at Lviv Polytechnic National University. At the initial stage the project underwent the consultancy of Ihor Kulchytskyy, to whom we express our gratitude.

7. References

[1] M. Panych, Seven Stories, Vancouver: Talonbooks, 2013.

[2] M. Panych, 7 istorii, [per. Z anhliiskoi Ivana Krychfalushiia], Brusturiv: Dyskursus, 2014.

[3] S. Laviosa (Ed.), Corpus-based Translation Studies: Theory, Findings, Applications, Rodopy, 2002.

[4] K. H. Chen, and H. H. Chen, Aligning bilingual corpora especially for language pairs from different families. Information Sciences Applications, 1995, 42, pp. 57–81.

[5] J. Munday, A Computer-assisted approach to the Analysis of Translation Shifts, Meta, 1998, XLIII, 4.

[6] F. Zanettin, Parallel corpora in translation studies: Issues in corpus design and analysis. In Intercultural Faultlines. Research Models in Translation Studies I: Textual and Cognitive Aspects, ed. M. Olohan, pp. 105–118. Manchester: St. Jerome, 2000.

[7] J. Allen, Natural Language Understanding. Cummings Publishing Company, Redwood City, 1995.

[8] D. Barnard, et al. "SGML-Based Markup for Literary Texts: Two Problems and Some Solutions." Computers and the Humanities, vol. 22, no. 4, 1988, pp. 265–276. JSTOR, URL: www.jstor.org/stable/30200136. Accessed 28 Feb. 2021.

[9] P. Blackburn, J. Bos, M. Kohlhase, & H. De Nivelle, Inference and computational semantics. In Computing Meaning, Springer Netherlands, 2001, pp. 11–28.

[10] I. Dagan, and O. Glickman, Probabilistic textual entailment: generic applied modeling of language variability. In Proceedings of the PACAL Workshop on Learning Methods for Text Understanding and Mining, Grenoble, France, 2004, pp. 26–29.

[11] R. Dale, H. Moisl, H. Somers (Eds.), Handbook of natural language processing. CRC press, 2000.

[12] M. Dilai, O. Levchenko**,** Discourses Surrounding Feminism in Ukraine: A Sentiment Analysis of Twitter Data 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2018 – Proceedings 2018 | conference-paper doi: 10.1109/STC-CSIT.2018.8526694

[13] A. Hogan, The Web of Data. Springer, 2020.

[14] V. Lytvyn, V. Vysotska, T. Hamon, N. Grabar, N. Sharonova, O. Cherednichenko, O. Kanishcheva (Eds.), Computational Linguistics and Intelligent Systems. Proc. 4thInt. Conf. COLINS 2020. Volume I:Workshop. Lviv, Ukraine, April 23-24, 2020, CEUR-WS.org, online

[15] M. Marcus, B. Santorini, M. Marcinkiewicz, Building a Large Annotated Corpus of English: Penn TreeBank. Computational linguistics: Special Issue on Using Large Corpora, 1993, 19(2), pp. 313–330.

[16] C. Matthews, An Introduction to Natural Language Processing Through Prolog, Routledge: London and New York, 2014.

[17] M. Oakes, Sentence and word alignment in the CARTER project. In Using Corpora for Language Research, ed. J. Thomas, and M. Short, London: Longman, 1996, pp. 211–233.

[18] P. Pavis, Theatre at the Crossroads of Culture, Routledge, 1992.

[19] S. Bassnett, Translating for the Theatre: The Case Against Performability. TTR : traduction, terminologie, rédaction, 1991, 4(1), pp. 99–111. URL: https://doi.org/10.7202/037084ar.

[20] S. Bassnett, Still Trapped in the Labyrinth: Further Reflections on Translation and Theatre, Constructing Cultures: Essays on Literary Translation.-Multilingual Matters, 1998, pp. 90–108.

[21] T.H. Howard-Hill, Modern Textual Theories and the Editing of Plays. The Library, 6th ser., 1989, 11, pp. 89–115.

[22] M. Issacharoff, F. Robin Jones (Eds.), Performing Texts. Philadelphia: University of Pennsylvania Press, 1988.

[23] J. Lavagnino, E. Mylonas, The show must go on: Problems of tagging performance texts. Comput Hum, 1995, pp. 113–121. URL: https://doi.org/10.1007/BF01830705

[24] Corpus-based Language Studies: An Advanced Resource Book, ed. T. McEnery, R. Xiao, Y. Tono, Routledge, 2006.

[25] N. Dershowitz, E. Nissan (Eds.), Language, Culture, Computation: Computing for the Humanities, Law and Narratives. Springer, 2014.

[26] O. Levchenko, O. Tyshchenko and M. Dilai. Associative Verbal Network of the Conceptual Domain БІДА (MISERY) in Ukrainian. Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020). Volume I: Main Conference. URL: http://ceur-ws.org/Vol-2604/Associative Verbal Network of the Conceptual Domain БІДА (MISERY) in Ukrainian

[27] N. Shakhovska, and M. Medykovskyy (Eds), Advances in Intelligent Systems and Computing III: Selected papers from the International Conference on Computer Science and Information Technologies, CSIT 2018, September 11–14 Lviv, Ukraine. Springer: Springer Nature Switzerland, 2019.

[28] C.M. Sperberg-McQueen, Text in the Electronic Age: Textual Study and Text Encoding, with Examples from Medieval texts. Literary and Linguistic Computing, 6 (1991), pp.34–46.

[29] C.M. Sperberg-McQueen, and B. Lou (Eds.), Guidelines for Electronic Text Encoding and Interchange (TEI P3). Chicago and Oxford: Text Encoding Initiative, 1994.

[30] S. Shaheen, and M. Spruit. Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2017, doi:10.1109/dsaa.2017.61.

[31] Topic Modeling in Python with Gensim. Machine Learning Plus, 16 Apr. 2020, URL: www.machinelearningplus.com/nlp/topic-modeling-gensim-python.

[32] K. Aguilar, NLP Techniques with Shakespeare's Plays: Cleaning and Classifying Text with the Bard, 2020. URL: https://medium.com/analytics-vidhya/nlp-techniques-with-shakespeares-plays-d8843ba26a4f.

[33] O. Levchenko, M. Dilai, (2019) Attitudes Toward Feminism in Ukraine: A Sentiment Analysis of Tweets. In: Shakhovska N., Medykovskyy M. (eds) Advances in Intelligent Systems and Computing III. CSIT 2018. Advances in Intelligent Systems and Computing, vol 871. Springer, Cham. doi:10.1007/978-3-030-01069-0_9