*August 2018*

# Integer-based nomenclature for the ecosystem of lexically repetitive expressions in complete works of William Shakespeare

Daniel Devatman HROMADA [a,b,1],
[a] *Berlin University of the Arts, Faculty of Design*
[b] *Einstein Center Digital Future*

**Abstract.** Repetition of morphological or lexical units is an established technique able to reinforce the impact of one's argument upon the audience. Rhetoric tradition has canonized dozens of repetition-involving schemas as figures of speech. Our article shows a way how hitherto ignored repetition-involving schemata can be identified. It shows that certain classes of repetitive figures can be represented in terms of specific sequences of integer numbers and vice versa, how specific sets of integer numbers can be translated into sets of regexes able to match repetition-involving expressions. A "Shakespeare number" S is simply defined as an integer with at least one repeated digit in which no digit bigger than X can occur if ever a digit X had not yet occurred in S's decimal representation. Hence, 121 is a Shakespeare number, while 123 or 211 are not. A set of "entangled numbers" is subsequently defined as a subset of "Shakespeare numbers" with an additional property that all digits which occur in them are repeated at least twice in the decimal representation of the number. Thus, a 1212 is an entangled number while 1211 is not. A complete set E of entangled numbers of maximal length of 10 digits is subsequently generated and every member of E is translated into a regex. Each regex is subsequently exposed to all utterances in all works of William Shakespeare, allowing us to pinpoint 3367 instances of 172 distinct E-schemata. This nomenclature may allow scholars to lead a discussion about schemata which have escaped the attention of classical interpretators.

**Keywords.** repetitive figures of speech, regular expressions, William Shakespeare, back-references, integers, stylometry



[1] Corresponding Author: Daniel Devatman Hromada, Faculty of Design, Berlin University of the Arts, 10823 Schoeneberg, Berlin, Bundesrepublik Deutschland, EU. Mails to: daniel at udk dash berlin dot de

## 1. Exordium

*"A faulty argument repeated twice is already better: repeated twenty times, it is excellent. Our ears adapt to it as to any other music and we applaud it mechanically ... One repeats an argument as one hums a vaudeville: not because it is good, but because it has been often chanted."* [3, note XXIII]

Repetitio mater studiorum, pater oratoriumque. It had been known to ancients that even the clearest reasoning can fail to convince the audience if ever the intended argument is not communicated with sufficient redundancy. And it is well known to moderns that the cheapest yet most efficient way how such redundancy can be attained is by means of repetitive transfer of information [7] from sender to receiver [20].

What's more, in human cognitive systems, repeated information is often amplified [14, p.208-211]. It is therefore little surprising that repetition plays a non-negligable role in the art of persuasion, commonly known as *rhetorics*. Thus, in practically every classical manual, the students of oratory and poetic disciplines are reminded to reassert their arguments; to mould forms which reflect their contents and utter contents which reflect their form; to make appear and reappear certain words and syllables; to repeat certain sounds or reactualize certain movements. Simply stated: to remember the figures by means of which one can reinforce one's influence over one's audience.

Hence, schemata known under names as diverse as *polysyndeton, anaphore, anadiplose, epistrophe, symploche, antanaclasis, paronomasia or even antimetabole* [2]. are traditionally defined in terms of repetition of their components [24]. But there is more, for one should also not forget repetitive figures (RFs) like *alliteration, paregmenon, polyptoton, epizeuxis or even good old psittacism*.

Hence, dozens of RFs are sure to exist but their scholastic nomenclature complicates any further communication with more computational- and NLP- oriented researchers. The objective of this article is to bridge this gap.

## 2. Introduction

In literature studies it is fairly common to speak about so-called "rhyme schemes" like AAAA for monorhymes, ABAB for alternate rhyme, ABBA for enclosed rhymes etc.

It is therefore not much surprising that analogic formalisms - that is, formalisms that involve *alphabetic indices* - have been adopted by scholars aiming to formalize a subgroup of rhetoric figures, known as the group of schemes. For example, [11] use a following formalism:

$$[W]_a ... [W]_b ... [W]_b ... [W]_a$$

to denote the rhetoric figure known as antimetabole. Subsequent studies in automatized chiasm identification and detection pursue a similiar route and often use formulae like ABXBA, ABCBA, ABCXCBA to denote schemata corresponding to utterances such as: "*Drake love loons. Loons love Drake.*", "*All as one. One as all.*" ([12] or "*In prehistoric times women resembled men, and men resembled women.*" ([6])

This being understood, the core idea behind this article is simple to explicate. For what shall be principially elucidated here is truly nothing more than the most basic a formalistic quirk[1] **a notational flip from alphabetic to numeric indices**. Hence, *A*-indices are to be substituted by 1-indices, *B*-indices by 2-indices, *C*-indices by 3-indices et caetera. Hence and henceforth, one is free to use the form 1212 instead of *ABAB*, 1221 instead of *BABA* and 12321 instead of *ABCBA*...

[2]Note that certain RFs included in a so-called "chiasmatic suite" (Reference - this volume) are not only repetition-involving but also fractal-like in a sense that they *embed* other repetitions which include yet other repetitions

Such change of notation may subsequently allow certain scholars to percieve and concieve a set of potentially interesting rhetoric schemata as a potentially infinite subset of the infinite but countable set [2,9,23] of non-positive natural numbers . That is, integers. The main implication of such mapping of a set of surface-based, repetition-involving rhetoric figures onto the set of integers goes as follows: given that the set of integers is enumerable, the set of our integer-based RF-schemata denoting formulae is enumerable as well. And as shell be *shewn*, developping a program which shall enumerate big amounts of such schemata is a fairly trivial enterprise which can fit into dozen lines of code (c.f. listings 1 and 2).

Such program generating such sets, however, was not developped nor is here presented just to accomplish some mathematicians' useless fancy. Rather contrary is the case and our objectives are to be considered more practical than theoretical. For such sets of potentially interesting RF-schemata can be translated - by yet another program (c.f. 4) - into so-called regular expressions ("regexes") which could be subsequently used to match and discover hitherto unknown repetition-based expressions occurent in attested natural language corpora [26,8].

Like that of collected works of William Shakespeare, for example.

## 3. Definitions

### *3.1. Shakespeare number*

A Shakespeare number $S$ is a positive natural number ($S \in N$) whose decimal representation expresses two properties:

- repetitive property: at least one digit occurs twice
- ascending property: $S$ contains no digit $n > 1$ without containing a digit $n-1$ to the left of first occurrence of $n$

In order to see the principle more clearly, table 1 enumerates ten Shakespeare numbers with smallest value.

| S-number | Alphabetic representation | Matchable expression |
|---|---|---|
| 11 | AA | "we split we split " |
| 111 | AAA | "we split we split we split |
| 112 | AAB | "here here sir " |
| 121 | ABA | "to prayers to " |
| 122 | ABB | "trip audrey i attend i attend " |
| 1111 | AAAA | "justice justice justice justice " |
| 1112 | AAAB | "great great great pompey " |
| 1121 | AABA | "here here sir here " |
| 1122 | AABB | "gross gross fat fat " |
| 1123 | AABC | "he he and you " |

**Table 1.** First ten Shakespeare numbers, their corresponding alphabetic representations and arbitrarily chosen Shakespearean expressions which can be subsumed under them.

As a counterexample, let's precise that 22 is not a Shakespeare number because digit 1 does not occur at all and 221 is not a Shakespeare number because 2 occurs with no 1 to its left. These two numbers therefore do not satisfy the ascending property. On the other hand, numbers like 12, 13 or 123 are also not S-numbers because they do not include any repeated digit and therefore do not satisfy the repetition-inclusion constraint.

Listing 1 displays the source code of a routine able to generate the sequence of $S-numbers$ from one to potential infinity. The sequence of first 163553 S-numbers - id est those S-numbers whose value is less than 9999999999 is available at Online Encyclopedia of Integer Sequences [15] under sequence number A273977 [3].

Deeper mathematical and number-theoretical properties of S-numbers are presented in [21].

### *3.2. Entangled number*

| E-number | Alphabetic representation | Matchable expression |
|---|---|---|
| 11 | AA | "we split we split " |
| 111 | AAA | "we split we split we split " |
| 1111 | AAAA | "justice justice justice justice " |
| 1122 | AABB | "gross gross fat fat " |
| 1212 | ABAB | "to prayers to prayers " |
| 1221 | ABBA | "my hearts cheerly cheerly my hearts " |
| 11111 | AAAAA | "so so so so so " |
| 11122 | AAABB | "great great great pompey pompey " |
| 11212 | AABAB | "come come buy come buy " |
| 11221 | AABBA | "high day high day freedom freedom high day " |
| 11222 | AABBB | "o night o night alack alack alack " |
| 12112 | ABAAB | "too vain too too vain " |
| 12121 | ABABA | "come hither come hither come " |
| 12122 | ABABB | "come buy come buy buy " |
| 12211 | ABBAA | "freedom high day high day freedom freedom " |
| 12212 | ABBAB | "on whom it will it will on whom it will " |
| 12221 | ABBBA | "thou canst not hit it hit it hit it thou canst not " |

**Table 2.** All Entangled numbers with no more than 5 digits, their corresponding alphabetic representations and arbitrarily chosen Shakespearean expressions which can be subsumed under them.

A set of entangled numbers is a subset of set of Shakespeare numbers ($E \in S \in N$). $E-numbers$ therefore satisfy repetitive and ascending properties of $S-numbers$. In addition to these does the decimal representation of an entangled number $E$ one additional property:

- closure property: each digit of $E$ occurs at least twice

In order to see the idea more clearly, table 2 enumerates ten Entangled numbers having their digit-length equal to five or less.

As a counter example, let's precise that numbers like 12, 13, 22, or 123 are not E-numbers because they are not even S-numbers. On the other hand, S-numbers like 121 or 1211 are not E-numbers because they contain a digit 2 which is not repeated.

Listing 2 displays the source code of a routine able to verify whether an $S-number$ presented at the input is an $E-number$. The sequence of first 4360 $E-numbers$ - id est those $E-numbers$ whose value is less than 9999999999 is available at Online Encyclopedia of Integer Sequences [15] under sequence number A273978 [4].

[3] https://oeis.org/A273977/b273977.txt
[4] https://oeis.org/A273978/b273978.txt

Deeper mathematical and number-theoretical properties of S-numbers are presented in [21].

## 4. Method

The core idea behind our method can be stated as follows:

**Any $S-$ or $E-$ number is to be "translated" into a backreference-endowed regular expression.**

More concretely, every digit of an S- or E- number can be interpreted as a sort of an element or a "brick". In this article, we work only with one type of bricks, those corresponding to sequences which are between two to twenty-three characters long[5] . More concretely, a first occurence of a novel brick can be represented as a PERL-compatible regular expression:

*(.{2,23})*

However, any subsequent repeated occurence of a digit in the S- or E- number is interpreted not as an occurence of the new brick, but rather as a backreference to the brick which was already denoted by the same digit. Hence, the very first S- number 11 is NOT to be translated into regex */(.{2,23}) (.{2,23})/*. For this would imply existence of two distinct bricks. Rather, the E-number 11 is to be translated into regex:

*(.{2,23})* **\1**

wherein the expression \1 denotes the backreference to the content matched by the regex-brick specified in first parentheses, i.e. brick no.1 .

Hence, the S-number 111 can be easily translated into a regex */(.{2,23}) \1) \1/*, 1111 into a regex */(.{2,23}) \1 \1 \1/* etc.

These, however, are cases which correspond only to repetition of one single brick: 11 for duplication, 111 for triplication, 1111 for quadruplication etc. In order to assure the application of the non-identity principle stating that:

**"Each distinct digit corresponds to distinct content"**

, an additional adjustment is needed in case we want to translate S-numbers containing multiple digits of different kind. That is, S-numbers like 121, 122 or 211.

For if we would not care for the principle of non-identity, a number like 121 could be easily represented as */(.{2,23}) (.{2,23}) \1/* and a number like 122 could be translated into */(.{2,23}) (.{2,23}) \2/*. It could turn out, however, that these regexes would match the very same expressions as other, more simple regexes do as well (e.g. the expression "no no no" could be matched by both */(.{2,23}) \1) \1/*, as well as by */(.{2,23}) (.{2,23}) \1/* or */(.{2,23}) \1 (.{2,23})/*. This is so, because nowhere in such regular expression it is specified that the first brick has to be different from the second brick, or third brick from the second.

Luckily enough, **syntax of PCREs is exhaustive enough to allow us to encode the non-identity constraint into regexes themselves**. This is attained by putting the backreference into a so-called negative lookahead, traditionally expressed by the formula *(?!)*. Hence, by translating the S-number **121** into the regex

[5] Minimal (e.g. 2) and maximal (e.g. 23) brick length are the only parameters of our model and can be, of course, adequately tuned. Sometimes we shall denote this parameter couple with the term **base**. More in discussion.

*(.{2,23}) (?!\1)(.{2,23}) \1*

we can make sure that the content matched by the brick denoted by digit 2 shall be different from the content matched by the brick denoted by digit 1. Thus, an expression "no no no" shall not be matched by such a regex while an expression "no yes no"[6] shall.

Going somewhat further, an S-number 12321 - which could be understood as an instance of chiasmatic ABXBA - is to be translated into regex

*(.{2,23}) (?!\1)(.{2,23}) (?!\1|\2)(.{2,23}) \2 \1*

whereby the disjunctive backreference contained in the negative lookahead (?!\1|\2) assures that the content matched brick no.3 - corresponing to filler X - shall be different from content matched by the brick representing digit 1 as well as the brick representing digit 2.

This being said, the method of translating S- or E- numbers into regexes which do not transgress the non-identity constraint is pretty much straightforward, and is fully and completely described by PERL code given in listing 3.

## 5. Experiment

### *5.1. Corpus*

A digital, unicode-encoded version of Craig's edition of "Complete works of William Shakespeare" [4] has been downloaded from a publicly available Internet source [7]. This corpus contains 17 txt files stored in the sub-folder "comedies", 10 txt files stored in the sub-folder "tragedies" and 10 txt files stored in the sub-folder "historical".

What's more, all utterances are annotated according to the following format:

```
<PERSONA>
        Sentence 1.
        Sentence ...
</PERSONA>
<MIRANDA>
        O, wonder!
        How many goodly creatures are there here!
        How beauteous mankind is!
        O brave new world,
        That has such people in't!
</MIRANDA>
```

That is, a format highly reminiscent of the format of a valid XML document. This format wherein diverse values of the tag $<$ *PERSONA* $>$ denote names of diverse dramatis personae (e.g Miranda, Prospero) , seems to be consistently and stringently followed across all files contained in the corpus. This is advantageous, since it implies that the content present between the opening and closing tag can be understood as a supraphrasal, meaning-encoding monadic unit: a utterance. Verily, this is encouraging.

It is encouraging for both theoretical (1.) as well as for *practical (2.) a reason*:

[6] A cautious reader may now start to observe that non-repeated digits of an S-number in fact correspond to "filler" or "separator" expressions (e.g. "yes") which in many cases fill the space between repeated elements themselves (e.g. "no").

[7] Downloaded from http://www.lexically.net/downloads/corpus_linguistics/ShakespearePlaysPlus.zip. Backup at http://sci.wizzion.com/ShakespearePlaysPlus.zip .

1. school of thought to which our research tends to adhere is principially a constructivist, usage-based linguistic paradigm best manifested in [22]
2. computational complexity of matching backreference-endowed regexes depends supralineary or maybe even non-polynomially [1] from the length of the text being matched

Regarding the practical reason, it could be postulate that our article offers certain evidence for the hypothesis "*backreferenced regex-parsing of Shakespearean utterances is computationally tractable in reasonable time*", whereby the term "reasonable" denotes time scales between miliseconds and minutes. More in discussion.

Regarding the theoretical reason, it is worth making explicit that an implicit leitmotive of Tomasello's theory is a definition stating:

**Utterance is the basic unit of linguistic interaction.**

### *5.2. Processing*

Dramatic pieces are divided into utterances. This is a natural consequence of the fact that dramatic pieces tend to represent scenarios within which diverse dramatis personae interact with each other. It is difficult to see any other litteral genre where division into utterances is as **marked** as in case of drama[8].

And in case of digital version of [4] Shakespeare corpus, such markedness tends to be even more marked.

Therefore, one simply needs to cut the corpus into utterances by interpreting the closing tag of the utterance (e.g. $< /PERSONA >$, $< /MIRANDA >$ etc.) as the utterance separator. Even more concretely, one can simply consider the slash symbol **/** to be the utterance separator. Subsequently, dividing the original dramatic text into utterances is, at least in PERL, as simple as defining the symbol **/** to be the default input separator. That is, in PERLish, by executing following code:

$$\$\backslash = "/";$$

Only two further text-processing steps have been executed during the initialization phase of the experiment hereby presented. Primo, content of each utterance has been put into lowercase. Secundo, non-alphabetic symbols (e.g. dot, comma, exclamation mark etc.) have been replaced by blank spaces. We are aware that such replacement could potentially lead to certain amount of loss of prosody- or pathos- encoding information. However, we consider this step as legitimate because the objective of our experiment was to focus on repetition of **lexical** units.[9]

Pre-processing code once executed, identification of expressions containing diverse types of lexical repetition is as simple as matching each Shakespearean utterance with each regex.

## 6. Results

This section presents results of exposure of Shakespeare's corpus to base=2,23 regular expressions generated out of all entangled numbers with max. length of 10 digits. We focus on $E_{2,23} - numbers$ because their closure property (i.e. "every digit contained in a valid E-number has to occur at least

[8] Plato's dialogues are, of course, set aside as a very particular case. When it comes to film scripts and/or subtitles to other audiovisual media, these are principially understood as a particular subtype of dramatic pieces

[9] Enumerative generation of backreference-involving regexes focusing on repetitions of phonotactic clusters, syllables, phrases or potentially even sememes and prosodies is, in theory, also possible. We prefer, however, not to focus on this topic within the limited scope of this article.

twice") gives an arbitrary $E-number$ ability to match much more rare a gem than just an arbitrary $S-number$.

| Instances | $E_{2,23}-number$ | Example |
|---|---|---|
| 2332 | 11 | "bestir bestir " |
| 525 | 1212 | "to prayers to prayers " |
| 170 | 111 | "ha ha ha " |
| 100 | 123123 | "cover thy head cover thy head " |
| 48 | 12121 | "come hither come hither come " |
| 35 | 1221 | "fond done done fond" |
| 32 | 12341234 | "let him roar again let him roar again " |
| 32 | 1122 | "with her with her hook on hook on " |
| 30 | 1111 | "great great great great " |
| 23 | 121212 | "come on come on come on " |
| 12 | 123231 | "upholds this arm this arm upholds " |
| 12 | 1231231 | "fubbed off and fubbed off and fubbed " |
| 11 | 121233 | "trip audrey trip audrey i attend i attend " |
| 11 | 112323 | "what what what ill luck ill luck " |
| 10 | 123312 | "my hearts cheerly cheerly my hearts " |
| 10 | 11122 | "lady lady lady alas alas " |
| 9 | 121323i | "a lord to a lord a man to a man " |
| 8 | 12321434 | "land rats and water rats land thieves and water thieves " |
| 8 | 11111 | "so so so so so " |
| 7 | 12312312 | "let me see let me see let me " |
| 6 | 11234234 | "on on on to the breach to the breach " |
| 5 | 12123434 | "i thank god i thank god is it true is it true " |
| 5 | 1112323 | "barren barren barren beggars all beggars all " |

**Table 3.** Quantities of utterances present in collected works of William Shakespeare which contain at least five distinct utterances corresponding to an E-number encoding the backreference-encoding regex whose individual brick match expressions not shorter than 2 characters and not longer than 23 characters.

### *6.1. Quantitative*

All in all, 3667 instances of a repetitive expression has been detected in Shakespeare's complete works. These were contained in 2295 distinct utterances and corresponded to 172 distinct $E_{2,32}$ schemata. Among these, 71 matched more than one instance: these schemata could thus potentially correspond to a certain cognitive pattern or a habitus in Shakespeare's mind.

Table 3 contains summary matching frequency information which concerning schemata matching at least five distinct utterances.

Another phenomenon may be found noteworthy by a reader interested in purely quantitative aspects of our research. That is, the relation between the number of digits of a $E-number$ of length $L$ seems to be in a Zipf-like [27] relation to number of occurences of expressions which can be matched by such $E_L$. For example, Shakespeare's dramas seem to contain 2332 duplications ($E = 11$), 170 triplications ($E = 111$), 30 tetraplications ($E = 1111$), 8 pentaplications ($E = 11111$ [10]), two hexaplications ($E = 111111$ [11]), one heptaplication ($E = 1111111$ [12]) and zero octaplications.

[10] E.g. "never never never never never " by Lear in King Lear.

[11] E.g. "kill kill kill kill kill kill " also by king Lear.

[12] E.g. "so so so so so so so " by Shallow in The Second Part of King Henry IV.

It is worth mentioning, however, that generic relation between the length (in digits) of an $E-number\ X$ and the amount of utterances which $X$ matches *seems not to be Zipfian*. This is illustrated by Table 4.

| Digits | Theoretical | Matched |
|---|---|---|
| 2 | 1 | 2332 |
| 3 | 1 | 170 |
| 4 | 4 | 622 |
| 5 | 11 | 91 |
| 6 | 41 | 211 |
| 7 | 162 | 56 |
| 8 | 715 | 86 |
| 9 | 3425 | 67 |

**Table 4.** Schemata corresponding to $E-numbers$ with even number of digits match more frequently than those with odd number of digits.

As indicated by Table 4, an observed preference for repetitive expressions including two, four, six or eight bricks cannot be explained in terms of number-theoretical distribution of $E-numbers$ themselves. For example, there exists eleven $E-numbers$ with five digits and fourty-one $E-numbers$ of length six. However, when exposed to Shakespeare corpus, base(2,23) regexes generated from $E-numbers$ six digits long seem to match 211 utterances while five brick long regexes match only ninety-one of them.

Whether this observed asymmetry is an artefact of our method and our definition of $E-numbers$, or whether it is due to a sort of cognitive bias, a sort of *preference for balanced repetitions* poses us in front of an argument which we do not dare to tackle within the limited scope of the present article.

### *6.2. Qualitative*

It may be said that the longer the E- or S- number is, the more complex a structure, the more cognitively-salient, pathos-filled an entity it potentially represents. For this reason, this subsection principially exposes the reader with few answers to a question:

"**What Shakespearean expressions can be matched with longest possible E-number ?**"

In all following examples, we will use the $base_{2,23}$ E-numbers, i.e. restrict the length of individual bricks to min. 2 and max. 23 characters.

In the realm of comedies[13], one can observe that the regex generated from the number 12343434 pin-points a following utterance from Stephano playing his role in *The Tempest*:

```
Flout (1) 'em (2), and (3) scout 'em (4);
and (3) scout 'em (4),
and (3) flout 'em (4);
Thought is free.
```

while regex generated from number 12343412 identifies Miranda's:

```
All (1) lost (2) to (3) prayers (4),
to (3) prayers (4) all (1) lost (2).
```

[13]Link to the file containing all XXX expressions shall be published in the camera-ready version of the article.

or Caliban's

```
Freedom (1), high (2) day (3) !
high (2) day (3),  freedom (1) !
freedom (1) ! high (2) day (3),
freedom (1) !
```

[14] all appearing in the same play.

Another answer, corresponding to E-number 122133144 is given by Dromio, a personage in Shakespeare's "Comedy of Errors":

```
She is so hot
because (1) the meat is cold (2) ;
The meat is cold (2)
because (1) you come not home (3);
You come not home (3)
because (1) you have no stomach (4);
You have no stomach (4),
having broke your fast;
```

Analyzing the realm of tragedies, one may see Polonius - a character in the Hamlet drama - utter a 11231434231-matchable expression:

```
The best actors in the world,
either for tragedy, comedy,
history, pastoral (1), pastoral (1) - comical (2),
historical (3) - pastoral (1) , tragical (4) - historical (3),
tragical (4) - comical (2) - historical (3) - pastoral (1) ,
scene individable, or poem unlimited:
Seneca cannot be too heavy,
nor Plautus too light. For the law of
writ and the liberty,
these are the only men.
```

or one can hear Hamlet himself pronouncing a following 1231414312-matchable sequence:

```
Let your own discretion be your tutor:
suit the (1) action (2) to (3) the (1) word (4),
the (1) word (4) to (3) the (1) action (2)
```

while Mercutio from the Romeo and Julia narrative states:

```
Come, come, thou art as hot a Jack
in thy mood as any in Italy;
and as soon (1) moved (2) to be (3) moody (4),
and as soon (1) moody (4) to be (3) moved (2).
```

These examples, of course, are just a tip of an iceberg.

[14] It is important to realize that **the very same expression can be matched by multiple regexes**. Hence, an above mentioned Caliban's proclamation can be analyzed not only to match the $base 2,23$ E-number 1232311231, but also analyzed to match E-numbers like 12211121 (if ever "high day" forms only one brick) etc. This is analogic, mutatis mutandi, to sentence having multiple syntactic parses.

Verily, only a tip of an iceberg, because many strongly marked repetitive expressions are also to be found in Shakespear's historical dramata. Among these, dramata eternalizing narratives of Henry IV. and Henry V. tend to top the list. Hence, Gadshill *reasons*

```
will strike (1) sooner (2) than (3) speak (4)
and (5) speak (4) sooner (2) than (3) drink (6)
and (5) drink (6) sooner (2) than (3) pray
and yet i lie for they pray continually
to their saint the commonwealth or
rather not pray to her
but prey on her
```

while Falstaff *emphasizes*:

```
banish peto
banish bardolph
banish poins but for
sweet jack falstaff
kind jack falstaff
true jack falstaff
valiant jack falstaff
and therefore more valiant
being as he is old  jack falstaff
banish (1) not (2) him (3) thy (4) harry s (5) company (6)
banish (1) not (2) him (3) thy (4) harry s (5) company (6)
banish (1) plump jack and
banish all the world
```

It is, however a persona named Shallow which seems to be particulary fond of repetitions, once saying

```
come (1)
on (2)
come (1)
on (2)
come (1)
on (2)
sir (3)
give (4)
me (5)
your (6)
hand (7)
sir (3)
give (4)
me (5)
your (6)
hand (7)
```

and next time saying:

```
where s (1) the roll (2)
where s (1) the roll (2)
where s (1) the roll (2)
```

```
let (3) me (4) see (5)
let (3) me (4) see (5)
let (3) me (4) see (5)
so (6) so (6) so (6) so (6) so (6) so (6) so (6)
yea marry sir ralph mouldy
let them appear as i call
let them do so
let them do so
let me see
where is mouldy
```

Given that Shallow appears in historical dramata, an interesting question could be rightfully posed: Is Shallow's tendency to produce repetitive utterances *en masse* just Shakespeare's *invention* or is it rather a sort of *description* of particular cognitive characteristics of once existing historical personage ?

## 7. Conclusion

Our article presents a way of maping a subset of a set of all possible backreference-endowed regexes onto a set of natural numbers. It indicates that for every **base** of certain kind, the set of regexes-to-be-generated is infinite but enumerable. A set of so-called *Shakespeare numbers* ($S-numbers$) is defined as well as the set of "Entangled numbers". The second being a subset of the first, satisfying one additional constraint:

**Every distinct digit ("symbol") of an entangled number $E_X$ occurs in $E_X$ at least twice.**

We have subsequently generated a list of all such $S-numbers$ (c.f. listing 1) and $E-numbers$ (c.f. listing 2) with at max 10 digits. After which the $E-numbers$ have been translated into backreference-endowed regular expressions whose most elementary units, so-called "bricks", were no shorter than two and no longer than twenty-three characters. In the end, such regexes have been exposed to corpus containing collected works of William Shakespeare.

This approach allowed us to pinpoint 3667 utterances matching at least one among 172 distinct repetitive formulae. We believe that at lease some among these formulae could be of certain interest not only for Shakespearean [16] scholars in particular, but also for wider fields of "digital humanities" [25] or stylometry.

The good news is that the whole matching process is also fairly fast. More concretely, matching all utterances with all $base 2,23$ regexes generated out of all 4360 $E-numbers$ with less than 10 digits lasted 9555 seconds in case Shakespearean comedies, 6607 seconds in case of tragedies and 6900 seconds in case of historical dramata. All this on one single core of an 1.4 GHz CPU.

## 8. Peroratio

Rhetorics undoubtedly belongs among five oldest scientific paradigms ever explicated by scholars of the occidental[15] tradition. Even before Plato noted down discussions between Socrates and Gorgias and Socrates and Parmenides; even before Aristotle projected his point-of-view upon the realm of man, *Athēnaia* had already been venerated.

[15]Note, however, that rhetorics is far from being unknown to Orient as well. Known as Sarasvatī in the sanskrit world, *the goddess embodies knowledge, arts, music, melody, muse, language, rhetoric, eloquence, creative work* ... [19] seems to be active already in vedic or even pre-vedic proto-indo-european times.

Longevity of rhetorics has positive as well as negative sides. Negative, for such lengthy tradition implies potential impediments caused by centuries of terminological and methodological sediments. We are convinced that, similiarly to diverse occult notations of pre-Mendelean chemistry, may alphabetic notation of BABAs and ABBAs be also considered to be such sediments in regards to rhetoric science. Hence, by a trivial act of switching notation from As to ones and Bs to twos, we aspire to do nothing else than to unblock this science from the state of *terminological traffic jam* to somewhat more fluid a state.

Hence and thus, interesting and almost melodical[16] verses of Shakespeare have been pinpointed and juxtaposed side by side to each other. Being unsure of whether such juxtaposition has ever been explored in the depth their merit, we find our qualitative results worthy of not only exploring but also publishing. For who knows, maybe they shall even inspire some potential Shakespeare of the future ?

Quantitative explorations may also turn out to be worthy of further exploration. Three axes of such exploration are immediately visible:

1. "universalia axis": study of language-independent invariants and rhetorical schemata which occur across many distinct languages and/or language groups [12]
2. "ontogenetic axis": exploration of processes by means of which complex eloquency of an individual locutor emerges out of simpler structures, from mind of a child to Shakespeare
3. "historical axis": study of different Digital Humanities resources in order to increase our knowledge about styles, fashions, crossovers and traditions popular during different epoches of human history

In terms of Saussurian linguistics ([5]), one may consider the first axis to be synchronic one while the the second and third can be considered as "diachronic" ones.

One may, for example, extend the work of [12] in domain of "language-independent detection of figures-of-speech" and demonstrate that E-numbers of considerable length match expressions not only Shakespeare, but also in Goethe, Moliere, Milton or others. Or focus on so-called "sacred texts" like Bible, Koran or RgVed where repetitions, indeed, abound. Or pursue a somewhat more psycholinguistic, ontogeny-oriented line of research and study the a corpus like CHILDES [17] in order to explore how complex eloquency emerges out of variations within repetition of complex sequences (another REFs to be given in camera-ready version).

At last but not least, we are convinced that our $S-$ or $E-$ number nomenclatures could be embedded into rhetorical figure ontologies [11,18]. Within such ontologies, antimetaboles could be thus "enriched" with attributes like "12321", "123321", "1234321" etc. ; anadiplosis would be labeled with another set of numbers, antistrophe with yet another, etc. The advantage of such an enrichment is quite easy to see: such enriched elements would become "grounded" [10]. That is - when looking for - or infering the presence of a certain figure of speech $F$ in certain text $T$, one could consult the ontology and see whether $F$ is not labeled with $S_F$ or $E_F$ attributes. If yes, one could simply parse the $T$ with corresponding $S_F$ or $F_E$ regexes. One could thus establish a practical, functional bidirectional bridge between the abstract realm of purely descriptive ontologies and material reality of text corpora which are to be parsed and understood.

And, of course, such nomenclatures - or nomenclatures of a similiar vein - may allow communication between computational and classical scholars in unambigous, precise, yet still concise and sufficiently explanatory terms. This being said, we conclude this article with an expression of hope that the method hereby introduces shall make it possible to spot down, identify, classify and study in deeper level the intricacies of cognitive ecosystems populated with swarms and clusters of hitherto unknown psycholinguistic schemata traditionally known as "figures of speech".

[16] It may be the case that the application of our method upon musical partitures - as stored in MIDI files, for example - shall also yield some worthy insights.

**Listing 1:** *PERL code generating an ascending sequence of Shakespeare numbers. Code hereby transfered to the public domain under license under the mGPL license for general use..*

```
$i=1;
INCREMENT : while ($i++) {
          my %d;
          $d{"0"}=1;
          $r=0;
          for $d (split //,$i) {
                    next INCREMENT if !exists $d{($d-1)};
                    if ($d{$d}) {
                              $r=1;
                    }
                    $d{$d}=true;
          }
          print "$i\n" if $r;
}
```

**Listing 2:** *PERL code checking whether a Shakespeare number given at the input is also an Entangled number. Code hereby transfered to the public domain under the mGPL license..*

```
OUTER: while (<>) {
          my %d;
          $i=$_;
          chop $i;
          for $d (split //,$i) {
                    (exists $d{$d}) ? ($d{$d}++) : ($d{$d}=1);
          }
          for $k (keys %d) {
                    next OUTER if ($d{$k}<2);
          }
          print "$i\n";

}
```

## 9. Acknowledgments

I would like to thank prof. Randy Harris, prof. Michael Ullyot, Ms. Jelena Mitrovic and Ms. Marie Dubremetz for stimulating discussions at 2016 Waterloo Workshop in Computational Rhetoric.

**Listing 3:** *PERL code translating S-numbers into syntactically correct regexes. Code hereby transfered to the public domain under the mGPL license..*

```
my $base='(.{2,23})';
$n=$ARGV[0];
@i = split //,$n;
$re = "";
my %h;
$no = "";
for my $i (@i) {
        $re.=" ";
        if (defined $h{$i}) {
                $re.='\\'.$i;
        } else {
                if ($i>1) {
                        $i>2 ? ($no.='|\\'.($i-1)) : ($no.='\\'.($i-1));
                        $re.='(?!'.$no.')';
                }
                $re.=$base;
                $h{$i}=1;
        }
}
$re.='[ <]';
print "$n translates into $re\n";
```

**Listing 4:** *PERL code for utterance-oriented pre-processing of texts contained in Shakespeare-PlaysPlus corpus. Code hereby transfered to the public domain under the mGPL license..*

```
use open ":encoding(utf-16)";
$/="/"; #consider the slash symbol to be the default input separator
while (<>) {
        $line=lc $_; #lowrecase
        $line=~s/[\r\n\t.,?!:;'"\- ]+/ /g; #remove non-alphabetic chars
        push @{$utterances{$ARGV}},$line; #construct the utterance hash
}
```

## References


[1] Alfred Vaino Aho. Algorithms for finding patterns in strings. *Algorithms and Complexity*, 1:255, 2014.

[2] Georg Cantor. Über eine elementare frage der mannigfaltigkeitslehre. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 1:75–78, 1892.

[3] Gorges Caumont. *Notes morales sur l'homme et sur la societe*. Sandoz&Fischbacher, Paris, 1872.

[4] William James Craig. *The complete works of Wiliam Shakespeare*. Oxford University Press, 1919.

[5] Ferdinand De Saussure. *Cours de linguistique générale: Publié par Charles Bally et Albert Sechehaye avec la collaboration de Albert Riedlinger*. Libraire Payot & Cie, 1916.

[6] Marie Dubremetz and Joakim Nivre. Rhetorical figure detection: the case of chiasmus. *on Computational Linguistics for Literature*, page 23, 2015.

[7] Luciano Floridi. *The philosophy of information*. Oxford University Press, 2011.

[8] Jeffrey EF Friedl. *Mastering regular expressions*. " O'Reilly Media, Inc.", 2002.

[9] Kurt Gödel. Über formal unentscheidbare sätze der principia mathematica und verwandter systeme i. *Monatshefte für mathematik und physik*, 38(1):173–198, 1931.

[10] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.

[11] Randy Harris and Chrysanne DiMarco. Constructing a rhetorical figuration ontology. In *Persuasive Technology and Digital Behaviour Intervention Symposium*, pages 47–52. Citeseer, 2009.

[12] Daniel Devatman Hromada. Initial experiments with multilingual extraction of rhetoric figures by means of perl-compatible regular expressions. In *RANLP Student Research Workshop*, pages 85–90, 2011.

[13] Daniel Devatman Hromada. Extraction of lexical repetitive expressions from complete works of william shakespeare. In *Proceedings of the 14th International Conference on Statistical Analysis of Textual Data*, pages 384–391, 2018.

[14] Daniel Devatman Hromada. *Prolegomena Paedagogica: Intramental Evolution and Ontogeny of Toddlerese*. 2019.

[15] OEIS Foundation Inc. The on-line encyclopedia of integer sequences, 2017. `http://oeis.org`.

[16] Sister Miriam Joseph. *Shakespeare's Use of the Arts of Language*. Paul Dry Books, 2008.

[17] Brian MacWhinney. *The CHILDES project: The database*, volume 2. Psychology Press, 2000.

[18] Miljana Mladenović and Jelena Mitrović. Ontology of rhetorical figures for serbian. In *International Conference on Text, Speech and Dialogue*, pages 386–393. Springer, 2013.

[19] John Muir. *Original Sanskrit texts on the origin and history of the people of India, their religions and institutions*. Trübner & Company, 1873.

[20] Claude E Shannon and Warren Weaver. The mathematical theory of information. 1949.

[21] NJA Sloane and Arndt Joerg. Counting words that are in "standard order", 2016. `https://oeis.org/A278984/a278984.txt`.

[22] Michael Tomasello. *Constructing a language: A usage-based theory of language acquisition*. Harvard university press, 2009.

[23] Alan Mathison Turing. On computable numbers, with an application to the entscheidungsproblem. *J. of Math*, 58(345-363):5, 1936.

[24] Alan Mathison Turing. Rhetorique. *Grand Memento Encyclopedique*, 1:687–689, 1936.

[25] Michael Ullyot. Review essay: Digital humanities projects. *Renaissance Quarterly*, 66(3):937–947, 2013.

[26] Larry Wall and Randal L Schwartz. *Programming perl*. O'Reilly & Associates Sebastopol, CA, 1991.

[27] George Kingsley Zipf. The psycho-biology of language. 1935.