Initial Experiments with Multilingual Extraction of Rhetoric Figures by means of PERL-compatible Regular Expressions

Daniel Devatman Hromada
Lutin Userlab – ChART – Paris 8 – EPHE - Slovak Technical University
hromi@kyberia.sk

Abstract

A language-independent method of figure-of-speech extraction is proposed in order to reinforce rhetoric-oriented considerations in natural language processing studies. The method is based upon a translation of a canonical form of repetition-based figures of speech into the language of PERL-compatible regular expressions. Anadiplosis, anaphora, antimetabole figures were translated into the form exploiting the back-reference properties of PERL-compatible regular expression while epiphora was translated into a formula exploiting recursive properties of this very concise artificial language. These four figures alone matched more than 7000 strings when applied on dramatic and poetic corpora written in English, French, German and Latin. Possible usages varying from stylometric evaluation of translation quality of poetic works to more complex problem of semi-supervised figure of speech induction are briefly discussed.

1 Introduction

During middle ages and before, the discipline of rhetoric composed - along with grammar and logic - a basic component of so-called trivium. Being considered by Platon as the "one single art that governs all speaking" (Plato, trans. 1986) in order to be subsequently defined by Aristotle as "the faculty of observing in any given case the available means of persuasion" (Aristotle, trans. 1954), the basic postulates of rhetoric are still kept alive by those being active in domains as diverse as politics, law, poetry, literary theory (Dubois, 1970) or humanities in general (Perelman & Olbrechts-Tyteca, 1969)

When it comes to more "exact" scientific disciplines like that of informatics or linguistics , rhetoric seems to be somewhat ignored - definitely more than its "grammar" and "logic" trivium counterparts. While contemporary rhetoric disposes with a strong theoretical background - whether in the form of the Rhetorical Structure Theory (Taboada, Mann, & Back, 2006), "computational rhetoric" (Grasso, 2002) or computational models of natural argument (Crosswhite & Fox, 2003); a more practically-oriented engineer has to nonetheless agree with the statement that "the ancient study of persuasion remain understudied and underrepresented in current Natural Language systems" (Harris & DiMarco, 2009) .

The aim of this article is to reduce this "under-representation" gap and in a certain sense augment the momentum of the computational rhetoric not by proposing a complex model of argumentation, but by proposing a simple yet efficient and language-independent method for extraction of certain rhetoric figures (RF) from textual corpora.

RFs, also called "figures of speech", are one of the basic means of persuasion which an orator has to his disposition. Traditionally, they are divided into two categories : tropes - related to deeper, i.e. semantic features of the phrasal constituents under consideration; and schemes - related to layers closer to actual material expression of the proposition, i.e. to the morphology, phonology or prosody of the generated utterance.

The method proposed within this article shall deal only with reduced subset of the latter - that is, with detection of rhetoric schemes anadiplosis, anaphora, antimetabole and epiphora which are based on a repetition or reordering of a given word, phrase or morpheme across multiple subsequent clauses. While such a stylometric approach was currently implemented with encouraging results by (Gawryjolek, 2009), his system is operational only when combined with probabilistic context-free grammar parser adapted to English language, and hence

Proceedings of the Student Research Workshop associated with RANLP 2011, pages 85–90,
Hissar, Bulgaria, 13 September 2011.

dysfunctional when applied upon languages for which such a parser does not exist.

In the following paragraphs of this article we shall present a system of rhetoric figure extraction which tends to be language-independent, i.e. applicable upon a textual corpus written in any language. Ideally, no antecedent knowledge about the grammar of a language is necessary for successful extraction by means of our method, the 1) prescriptive form of the figure-to-be-extracted and 2) the symbol representing phrase and/or clause boundaries is the only information necessary.

More concretely, our proposal is based on a fairly simple translation of a canonical form of a rhetoric figure under question into a computer language, namely into the language of PERL-compatible regular expressions (PCREs). PCREs are, in their essence, simply strings of characters which describe the sets of other strings of characters, i.e. they are a matching form, a template, for many concrete character strings. As many other regular expressions engines, PCREs make this possible by reserving special symbols - "the metacharacters" - for quantifiers and classes. But in addition to these features common to many finite state automata, PCREs offer much more (Wall & Loukides, 2000). These are the reasons why we consider the PCREs to be appealing candidates for a translation of rhetorical figures into a computer-readable symbolic form:

- by implementing "back references" (Friedl, 2006) , PCREs make it possible to refer to *that which was already matched,* hence allowing to construct automata able to match repetitive forms
- by implementing (from PERL version 5.10 on) "recursive matching", PCREs make it possible to match very complex patterns without a need to have recourse to other means, external to PCREs
- since the language of PCREs is very concise, the resulting PCRE describing a rhetorical figure under question is usually a string of few dozens of characters which could be eventually constructed not by means of human intervention, as was the case in this article, but by means of unsupervised genetic programming (Koza, 1992) or other means of grammar induction engine (Solan, Horn, Ruppin, & Edelman, 2005)

Element	Meaning
W	word
...	arbitrary intervening material
< … >	phrase or clause boundaries
Subscripts	identity (same subscripts), nonidentity (different subscripts)

Table 1: part of RF-representation Formalism (RFRF)

2 Method

2.1 PERL-Compatible Rhetoric Figures

Four figures were chosen - namely anadiplosis, anaphora, epiphora and antimetabole – in order to demonstrate the feasibility of the "rhetoric stylometry" approach. We have adopted the Rhetoric Figure Representation Formalism (RFRF) - initially concieved by (Harris & DiMarco, 2009) - and reduced it in order to describe only the four figures of interest. Basic symbols of RFRF and their associated meanings are presented in Table 1.

Since the goal of this article is primarily didactic, i.e. we shall start this *exposé* with very simple anadiplosis involving just one back-reference, and end up our proposal with somewhat more complex recursive PCRE matching epiphorae containing arbitrary number of constituents.

2.1.1 Anadiplosis

Anadiplosis occurs when a clause or phrase starts with the word or phrase that ended the preceding unit. It is formalized by RFRF as :

$$< \ldots \mathbf{W_x} >< \mathbf{W_x} \ldots >$$

We have translated this representation into this PERL-Compatible Rhetoric Figure (PCRF):

/((\w{3,})[.?!,] \2)/sig

The repetition-matching faculty is assured by a backreference to an initial n-gram composed of at least three word characters. Therefore, this PCRE makes it possible to match utterances like the one in Cicero's *De Oratore* :

Sed genus hoc totum orationis in eis causis excellit, in quibus minus potest inflammari animus iudicis acri et vehementi quadam incitatione; non enim semper fortis oratio quaeritur, sed saepe placida, summissa,

lenis, quae maxime commendat ***reos. Reos*** *autem appello non eos modo, qui arguuntur, sed omnis, quorum de re disceptatur; sic enim olim loquebantur.*[1]

This is the simplest possible anadiplosis figure since it matches only string with two occurences of a repeated word. Therefore we label this figure as **anadiplosis{2}**.

2.1.2 Anaphora

Anaphora is a rhetoric figure based upon a repetition of a word or a sequence of words at the beginnings of neighboring clauses. It is formalized by RFRF as :

$$< W_x \ldots >< W_x \ldots >$$

We have translated this representation into the following PCRE form:

/[.?!;,] (([A-Z]\w+) [^.?!;,]+[.?!;] \2 [^.?!;,] +[.?!;,] (\2 [^.?!;,]+[.?!;,])*)/sig

As all RFs presented in this article, this anaphora is also based on back-reference matching. In contrast with anadiplosis where dependency was of very *short-distance* nature, in case of anaphora, the second occurrence of the word can be dozens of characters distant from the initial occurrence. What's more, this RF takes into account possible third repetition of a W_x which makes it possible to match utterances like Cicero's:

Quid *autem subtilius quam crebrae acutaeque sententiae?* ***Quid*** *admirabilius quam res splendore inlustrata verborum?* ***Quid*** *plenius quam omni genere rerum cumulata oratio?*[2]

Since this PCRFs allows us to match anaphorae with two or three occurences of a repeated word, it is seems to be appropriate to label it as **anaphora{2,3}**.

2.1.3 Antimetabole

Antimetabole is a rhetoric figure which occurs when words are repeated in successive clauses in reversed order. In terms of RFRF, one can formalize it as follows:

$$< W_A W_B W_c \ldots W_C W_B W_A >$$

We have translated this representation into following PCRE form:

/((\w{3,}) (.{0,23}) (\w{3,})[^\.!?]{0,23} \4 \3 \ 2)/sig

Differently from previous examples when there was only one element matched and back-referenced, three elements - A, B, C- are determined in initial phases of matching this chiasmatic antimetabole. Subsequently, the order of A & C is switched while B is considered to be identic intervening material intervening between A and C and C and A. Since possible occurrence of other material intervening between ABC and CBA (i.e. ABCxCBA) is also taken into account, this PCRF has successfully matched expressions like:

Alle wie **einer**, **einer** wie **alle.**[3]

2.1.4 Epiphora

Epiphora or epistrophe is a RF defined as "ending a series of phrases or clauses with the same word or words". It is formalized by RFRF as:

$$< \ldots W_x >< \ldots W_x >$$

We have translated this representation into following PCRE form:

/([A-Z][^\.\?!;]+ (\w{2,}+)([\.\?!;] ?[A-Za-z] [^\.\?!;]+ (?:\2|(?-1))*)\2[\.\?!;])/sig

In contrast with anaphora{2,3} figure presented in 2.1.2, the epiphora figure hereby proposed exploits the "recursive matching" properties of latest versions of PCRE (Perl 5.10+) engines. In other words, the expression *(?:\2|(?-1))* match any number of subsequent phrases or clauses which end with W_x and not just three, as was the case in case of epiphora. Hence, a quadruple epiphora :

---

1 *"For vigorous language is not always wanted, but often such as is calm, gentle, mild: this is the kind that most commands the* ***parties****. By '* ***parties*** *' I mean not only persons impeached, but all whose interests are being determined, for that was how people used the term in the old days. "*

2 *"* ***Is there something*** *more subtle than a rapid succession of pointed reflections?* ***Is there something*** *more wonderful than the heating-up of a topic by verbal brilliance,* ***something*** *richer than a discourse cumulating material of every sort? "*

3 *" All as one, one as all. "*

*Je te dis toujou la même **chose**, parce que c'est toujou la même **chose**, et si ce n'était pas toujours la même **chose**, je ne te dirais pas toujou la même **chose**.*[4]

was detected by this recursive PCRF when it was applied upon corpus of Molière's works.

Since the recursive matching allows us to create a sort of "greedy" epiphora, we propose to label it as **epiphora{2,}** in possible future taxonomy of PCRFs.

2.2 Corpora

In order to demonstrate the language-independence of the rhetoric stylometry method hereby proposed, we confronted the matching faculties of initial "PERL Compatible Rhetoric Figures" (PCRF) with the corpora written in diverse languages.

More precisely, we have performed the rhetoric stylometry analysis of 4 corpora written by poets and orators who are often considered as exemplary cases of mastering their respective languages.

For English language, complete works of William Shakespeare had been downloaded from project Gutenberg (Hart, 2000). The same site served us as the source of 40 works of Johann Wolfgang Goethe written in German language. When it comes to original works of Jean-Baptiste Molière, 39 of them where recursively downloaded from French site *toutmoliere.net*. Finally, the basic Latin manual of rhetoric, Cicero's "De Oratore" was extracted from the corpus of Perseus Project (Crane, 1998) in order to demonstrate that PCRF-based approach can yield interesting results when applied even upon corpora written in antique languages.

Corpora from Project Gutenberg was downloaded as pure utf8-encoded text. No filtering of data was performed in order to analyze the data in their rawest possible form. The only exception was the stripping away of possible HTML tags by means of standard HTML::Strip filter.

Before the matching, the totality of the corpus was split into *fragments* whenever frontier **\n[^\w+]** (i.e. new-line followed by at least one non-word character) was detected. Shakespeare's corpus were splitted into 109492 fragments, Goethe's into 46597 fragments , Cicero's into 970 fragments while works of Moliere yielded 6639 fragments.

3 Results

In total, more than 7000 strings were matched by 3 PCRFs within 4 corpora containing in 17 Megabytes of text splitted into more than 163040 textual fragments.

	Anadiplosis{2}	**Anaphora{2,3}**	**Antimetabole{abcXbca}**	**Epiphora{2,}**
Cicero	0.00309	0.2711	0	0.0144
Goethe	0.00242	0.0717	0.0003	0.0042
Molière	0.01129	0.1634	0.000602	0.0210
Shkspr	0.00087	0.008	0.000219	0.008

Table 2: Relative frequencies of occurence of diverse PCRFs within diverse corpora (PCRF per fragment)

As is indicated in Table 2, the instances of anadiplosis, anaphora, antimetabole and epiphora were found in all 4 corpora involved in this study, the only exception being the absence of antimetabole in Cicero. In general, anaphora{2,3} seems to be the most frequent one: number of cases when this PCRFs succeeded to match highly surmounts the other two figures especially in case of Romance language authors – i.e. almost every sixth fragment from Moliere and every fourth from Cicero was matched by anaphore{2,3}.

The only exception to this "dominance of anaphora" seems to be Shakespeare whose complete works yielded exactly the same frequency of epiphora and anaphora occurences.

	Anadiplosis{2}	**Anaphora{2,3}**	**Antimetabole{abcXbca}**	**Epiphora{2,}**
Cicero	20	1	4	19
Goethe	44	3	33	287
Molière	57	1	29	65
Shkspr	7	2	17	64

Table 3: Elapsed time (in seconds) of different PCRF/corpus runs on average PC desktop

As is indicated in Table 3, computational demands of PCRF-based are not high in case of anaphora{2,3}. On the contrary, the recursive epiphora{2,} is much more demanding. As the recursive structure of this PCRF indicates, the speed of matching process is growing non-polynomially with the length of the textual fragment upon which the PCRF is applied and therefore the choice of correct fragment separator

[4] *"I always tell you the same **thing** because it is always the same **thing** and if it wasn't always the same **thing** I would not have been telling you the same **thing.**"*

token (c.f. 2.2) seems to be of utmost importance.

4 Discussion

We propose a language-independent parse-free method of extracting instances of rhetoric figures from natural language corpora by means of PERL-compatible regular expressions. The fact that PCREs implement features like back-references or recursive matching make them good candidates for the detection & extraction of rhetoric figures which cannot be matched by simpler finite state automata or context-free languages.

In order to demonstrate the feasibility of such an approach, we have therefore "translated" the *canonical* definitions of anadiplosis, anaphora and epiphora into four *PERL-compatible rhetoric figures* - namely anadiplosis{2}, anaphora{2,3}, epiphora{2,} and antimetabole{abcXbca} - and applied them upon Latin, English, French and German corpora. All four PCRFs successfully matched some strings in at least three of four corpora, indicating that repetition-based rhetoric figures can possibly belong to the set of *linguistic universalia* (Greenberg, 1957). Anaphora{2,3} surpassed in frequency of occurrences all the other figures, the only exception being Shakespeare in whose case the number of matched epiphorae was equal to the number of matched anaphorae.

We do not pretend that PCRFs presented hereby are the most adequate translations of traditional anadiplosis, anaphora, antimetabole or epiphora into an artificial language. Since PCREs can contain quantifiers and classes, it is evident that for any set of strings – which is one our case the set F of all the occurences of a given figure within its respective corpus – more than one possible regexp could be constructed in order to match all members of the set F. Therefore it may be the case that PCRFs that we have proposed in this "proof of concept" article are not the most specific ones nor the fastest ones.

When it comes to specificity, it may be stated that the closer look upon the extracted data indicates that PCRFs proposed hereby have proposed some "false positives", i.e. have matched strings which are not rhetorical figures (for example an expression "*FIRST LORD. O my sweet lord"* was matched by epiphora{2,} when applied upon Shakespeare's corpus, but is definitely not a rhetoric figure since the substring in capital letters simply denotes the name of dramatic persona pronouncing the following statement and not the clause of the statement itself).

When it comes to speed, it is established that PCREs with unbounded number of back-reference are NP-complete (Aho, 1991) and verily this may be the reason of very high run-times of a recursive epiphora{2,} in contrast to its non-recursive PCRF counterparts. From practical point of view it seems therefore more suitable – especially in case of analysis of huge corpora - to stick to non-recursive PCRFs. The other possible solution how to speed up the parsing – and in certain cases even to prevent the machine to fell into "infinite recursion loop" is the tuning of the "splitting parameter" so that the corpus is split in fragments of such a size that the *NP-complexity of the matching PCRE shall not have observable implications* upon a real run-time of a rhetoric figure detection process.

There are at least three different ways how PCRFs could be possibly useful. Firstly, since PCRFs are very fast and language-independent, they can allow the scholars to extract huge number of instances of rhetoric figures from diverse corpora in order to create an exhaustive compendium of rhetoric figures. For example, the corpus of >7000 strings which were extracted from corpora mentioned in this article (downloadable from http://www.lutin-userlab.fr/rhetoric/) could be easily put to use not only by teachers of language or rhetoric, but possibly also by those who aim to develop a semi-supervised system of *rhetoric figure induction* (c.f. last paragraph). Manual annotation of such a compendium and subsequent tentatives of such a figure of speech induction shall be presented in our forecoming article.

Secondly, the extracted information concerning the quantities of various PCRFs within different corpora could serve as an input element (i.e. a feature) for classifiying or clustering algorithms. PCRFs could therefore facilitate such stylometric tasks like authorship attribution, author name disambiguation or maybe even plagiate detection.

Thirdly, due to their language independence, PCRFs presented hereby can be thought of as a means for evaluation of differences between two different languages, or two different states of the same language. One can for example apply the PCRFs upon two different translations T1 and T2 and see that the distribution of PCRFs within T2 is more similar

to the distribution of PCRFs in the original than the distribution in T2. Therefore, one could possibly state that from rhetoric, stylistic or even poetic standpoint, T1 is more adequate translation of the original text than T2. On the other hand, when we speak about comparing two different states of the same language , we propose to perform PCRF-based analysis not only upon a corpus representing the *l'état de l'art* state of the language - like that of a Shakespeare, for example – but also to compare such a state with more initial states of the language development, as is represented by CHILDES (MacWhinney & Snow, 1985) corpus.

Finally, by considering PCRFs to be a method which could possibly be used as a tool of analysis of the development of language faculties in a human baby, we come closer to its third and somewhat "cognitive" implementation. This implementation - which is the subject of our current research - is based upon a belief that it is not unreasonable to imagine that PCRFs could possibly be constructed not manually, but automatically by means of genetic programming paradigm (Koza, 1992). Given the fact that PCRE-language is one of the most concise programming languages possibles and conceivables, and given the fact that the 1) speed of execution 2) the specifivity 3) the sensitivity could possibly serve as the input parameters of a function evaluating the fitness of a possible PCRF candidate, it is possible that the research initiated by our current proposal could result in a full-fledged and possibly non-supervised method of rhetoric figure induction. In such a way could our PCRFs possibly become something little bit more than just another tool for stylometric analysis of textual corpora – in such a way they could possibly help answering a somewhat more fundamental question: "*What is the essence of figures of speech and how could they be represented within&by an artificial and/or organic symbol-manipulating agent?*"

Acknowledgments

The author wishes to express his gratitude to University Paris8 – St. Denis and Lutin Userlab for support without which the research hereby presented would not be possible, as well as to thank philologues and comparativists of École Pratique des Hautes Études and ÉNS for keeping alive the Tradition within which the Language is considered to be something more than just an object of parsing and POS-tagging.

References

Aho, A. V. (1991). *Algorithms for finding patterns in strings, Handbook of theoretical computer science (vol. A): algorithms and complexity*. MIT Press, Cambridge, MA.

Aristotle. (1954). *Rhetoric*. 1355b.

Crane, G. (1998). The Perseus Project and Beyond: How Building a Digital Library Challenges the Humanities and Technology. *D-Lib Magazine*, 1, 18.

Crosswhite, J., Fox, J., Reed, C., Scaltsas, T., & Stumpf, S. (2003). Computational models of rhetorical argument. *Argumentation Machines—New Frontiers in Argument and Computation*, 175–209.

Dubois, J. (1970). *Rhétorique générale: Par le groupe MY*. Larousse.

Friedl, J. (2006). *Mastering regular expressions*. O′Reilly Media, Inc. Sebastopol, CA, USA.

Gawryjolek, J. (2009). *Automated annotation and visualization of rhetorical figures.*

Grasso, F. (2002). Towards computational rhetoric. *Informal Logic*, *22*(3).

Greenberg, J. H. (1957). The nature and uses of linguistic typologies. *International Journal of American Linguistics*, 23(2), 68–77.

Harris, R., & DiMarco, C. (2009). Constructing a Rhetorical Figuration Ontology. *Persuasive Technology and Digital Behaviour Intervention Symposium*.

Hart, M. (2000). *Project gutenberg*. Project Gutenberg.

Koza, J. R. (1992). *Genetic programming: on the programming of computers by means of natural selection*. The MIT press.

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of child language*, *12*(02), 271-295.

Perelman, C., & Olbrechts-Tyteca, L. (1969). *The new rhetoric: A treatise on argumentation*.

Plato. (1986). *Phaedrus.* 261e.

Solan, Z., Horn, D., Ruppin, E., & Edelman, S. (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences*, *102*(33), 11629.

Taboada, M., Mann, W. C., & Back, L. (2006). Rhetorical Structure Theory. Citeseer.

Wall, L., & Loukides, M. (2000). *Programming perl*. O′Reilly Media, Inc. Sebastopol, CA, USA.