Examples of Formulaity in Narratives and Scientific Communication

Sándor Darányi
Swedish School of Library and Information Science
University of Borås
Borås, Sweden
sandor.daranyi@hb.se

ABSTRACT

The AMICUS project was designed to promote scholarly networking in a topical area, motif recognition in texts, including its automation. Prior to doing so however it is necessary to show the theoretical underpinnings of the research idea. My argument is that evidence from different disciplines amounts to fragmented pieces of a bigger picture. By compiling them like pieces of a puzzle, one can see how the concept of formulaity applies to folklore texts and scholarly communication alike. Regardless of the actual name of the concept (e.g. motif, function, canonical form), what matters is that document parts and whole documents can be characterized by standard sequences of content elements, such formulaic expressions enabling higher-level document indexing and classification by machine learning, plus document retrieval. Information filtering plays a key role in the proposed technology.

1. INTRODUCTION

The identity of higher-order content-bearing elements, i.e., textual units that are typically designated for e.g. document indexing, classification, enrichment, and the like, strongly depends on community perception. An instance of such a prominent yet little investigated content-bearing unit is the *motif*: an element that keeps recurring in an artifact – e.g. in film, music, but also in folklore or scientific texts – by means of which often a narrative theme is conveyed. For example, the victory of the youngest son against all odds is a motif in folktales.

It has been known for almost a hundred years that the oral communication of folklore texts often applies formulaity to help the singer remember his text [37, 45, 46]. Filed under different names, structural and formal investigations of tales [27, 51, 60] and myths, indeed mythologies, have proposed the same approach [36, 40], with or without computer support. Less known is the fact that linguistic evidence points in the same direction: as exemplified by a now famous study in immunology, scientific sublanguages, characteristic of subject areas, may use a formulaic arrangement of content elements in a sequential fashion for the presentation of experiments, results, and their discussion [21].

Motivated by the now widespread use of the concept of a motif in bioinformatics (genomics), where the formulaic, sequential expression of content elements leading to known biochemical consequences enables automated componential filtering of data, below I discuss examples from different fields underlying the AMICUS research proposal.

This paper is structured as follows: Section 2 spells out the research questions and definitions of core concepts. Section 3 brings examples for motif-like structures from tale and myth research. Section 4 points out related classification schemes onto which such structures can be mapped. Section 5 discusses formulaity and scientific communication, with a discussion of the implications in Section 6 and conclusions in Section 7.

2. RESEARCH QUESTIONS AND DEFINITIONS OF CORE CONCEPTS

With their related argumentation outlined below, the research questions of AMICUS are as follows:

1. Given a combination of methodology specified by formal theory, examples, and a test corpus, which combination of tools will be best suited for the automation of motif extraction and consecutive semantic annotation in folklore texts?

2. Given best practice and theoretical predictions, how far does the concept of a motif apply to scientific texts?

The concept of a motif is often rather vaguely used: e.g. in spite of its 286 occurrences in the *Oxford Companion to Fairy Tales*, it was left undefined [66]. For a notable exception in folk narrative research, the reader is advised to Jason's important article [29]. Critizing Stith Thompson's somewhat lax original definition ("The smallest element in a tale having the power to persist in tradition" [59]), she discusses motifs, functions or motuses [51], the 'allo-', and '-etic/-emic' relationships [48], and Dundes's 'motif-eme' and 'allo-motif' [9]. Her definition asks for the following criteria:

- "A literary motif is the *simplest* (not smallest) unit of content which in the work of literature fills a primary formal slot of a literary structure: a 'character' (being/object [noun] and attribute [adjective]); a 'deed' (verb and adverb) fills the slots of 'narrative role' and 'narrative action', respectively. As a whole, [the] motif is of a literary nature; if decomposed, its components are not of a literary nature;

- A motif is *context-free*, i.e. it does not belong to a certain plot (= content type), ethnopoetic genre or ethnos. It 'floats' in the ethnic and literary universe and can be 'used' by any plot, genre, or ethnos in a certain cultural area, or even universally;

- The following qualify as motifs, being the basic units of content in oral and folk narratives: (1) single characters and requisites and their attributes (nouns and adjectives); (2) single deeds and their qualities (verbs and adverbs); (3) spatial (geographic) and temporal marks and their special attributes. At the same time, two groups have been added for practical purposes with respect to the compilation of indices, but they are not motifs: (4) couplings consisting of components of (1) and (2); and (5) formulaity (formulae and formulaic numbers)."

Jason [29] stresses that group (5) lists formal literary elements and is thus of a completely different nature than groups (1)-(3). Its formal elements are not motifs, although, for practical reasons, they were included in Thompson's Motif-Index. When components of groups (1-3) are included in formulae, they comprise motifs in themselves, regardless of their being or not in this particular case part of a formula. (The same is valid for similes, metaphors and attributes which are not detailed in the Motif-Index). Formulae abound in oral literature on all levels and in all genres; consult e.g., [37] and [28] for various formulae in the genre of oral epics and [54] for the genre of fairy tale. Formula tales (see the list in Motifs Z 20-50) belong to the type index [1], and are listed there as AaTh/ATU 2000 ff. [62].

On the other hand, the characteristic sequence of 'Narrative subject-role' - 'Narrative action' - 'Narrative object-role' used by her as the frame in which the above five groups of motifs or motif-like content indicators can be contrasted amounts to the kind of formulaity I regard as important for AMICUS. This content unit, in fact a narrative sentence, is called *motus* by Jason. Motus ('movement' in Latin) is a label for Propp's function, composed of three basic units of content (motifs) with certain relationships between them which form its structure.

3. EXAMPLES FROM TALE AND MYTH RESEARCH

Several kinds of formulaity exist, ranging from short canonical phrases, such as the *epitheton ornans* in Homeric epics, to longer ones used in orally improvised poetry, including canonical sequences of content elements and leading to story grammars [12, 32] or narrative algebra [14, 15, 16]. I will focus on this latter type only.

To recall, according to the the oral-formulaic theory developed by Milman Parry [45, 46] and Albert Lord [37], stock phrases could enable poets to improvise verse called orally improvised poetry. In oral composition, the story itself has no definitive text, but consists of innumerable variants, each improvised by the teller in the act of telling the tale from a mental stockpile of verbal formulas, thematic constructs, and narrative incidents. This improvisation is for the most part subconscious so that texts orally composed will differ substantially from day to day and from teller to teller. The key idea of the theory is that poets have a store of formulas (a formula being 'an expression which is regularly used, under the same metrical

conditions, to express a particular essential idea' [37]), and that by linking these in conventionalised ways, they can rapidly compose verse.

Such linking, however, seems to be pertinent to storytelling in prose as well. Let me give you an example where a chain of motifs characterize a particular tale type about supernatural adversaries:

"300 The Dragon-Slayer. A youth acquires (e.g. by exchange) three wonderful dogs [B421, B312.2]. He comes to a town where people are mourning and learns that once a year a (seven-headed) dragon [B11.2.3.1] demands a virgin as a sacrifice [B11.10, S262]. In the current year, the king's daughter has been chosen to be sacrificed, and the king offers her as a prize to her rescuer [T68.1]. The youth goes to the appointed place. While waiting to fight with the dragon, he falls into a magic sleep [D1975], during which the princess twists a ring (ribbons) into his hair; only one of her falling tears can awaken him [D1978. 2].

Together with his dogs, the youth overcomes the dragon [B11.11, B524.1.1, R111.1.3]. He strikes off the dragon's heads and cuts out the tongues (keeps the teeth) [H105.1]. The youth promises the princess to come back in one year (three years) and goes off.

An impostor (e.g. the coachman) takes the dragon's heads, forces the princess to name him as her rescuer [K1933], and claims her as his reward [K1932]. The princess asks her father to delay the wedding. Just as the princess is about to marry the impostor, the dragon-slayer returns. He sends his dogs to get some food from the king's table and is summoned to the wedding party [H151.2]. There the dragon-slayer proves he was the rescuer by showing the dragon's tongues (teeth) [H83, H105.1]. The impostor is condemned to death, and the dragon-slayer marries the princess" [62].

What matters for my argumentation is that as much as a certain sequence of specific functions amounts to a fairy tale plot [51], here too, it takes a certain linking of consecutive motifs to constitute the specific tale type.

Given this, and the plethora of digitized folklore texts, it is quite striking that although tale motifs as structural plot elements are there in the material, their automatic extraction has not been reported this far. This may be partly due to the question of their identities which is somewhat problematic for motifs, less so for Propp's functions. In other words, we have no proven idea what would amount to a motif in terms of extracts. Educated guesses about the type of construct that can be extruded and behaves as a motif include the following:

1. As constituents of a *narrative macrostructure* (deep structure), anything that meets the following criterion: "Propp's analysis of basic narrative constituents (functions) is based on two abstractions: (i) a classification of the dramatis personae according to their roles and (ii) an evaluation of actions with respect to common effects and according to their positions within the story" [24][1];

2. *Latent variables*: An early attempt identified them with motifs as kinds of concepts [64]. It is reasonably usual to equal latent variables with concepts [8, 63], although this circumvents the known problem of naming latent

---

[1] In contrast to studies in ethnopoetics, [24] does not distinguish between functions and motifs.

variables in general. The idea of concept space goes back at least to Luhn [39], Rocchio [53], Bärtschi [3] and Schäuble [56], but the term, apart from its purely metaphoric use, i.e. "the home of concepts", has at least a philosophical, a thesaurus-oriented, a geometric, plus a word-semantic interpretation and is therefore problematic;

3. *Gross constituent units*: Related research on the canonical formula of myth [35, 40] shows that at least 256 classes (subspaces) can be filtered out based on canonical variables and value configurations leading to group symmetries and symmetry breaking [5, 6, 41]. Such symmetries constitute phases in Markov chain based patterns [7].

In my eyes, the key idea to extracting chains of symbolic content from text in the above sense is the formulaic representation of sentences in Harris [21], bridging the gap between scientific sublanguages and so far unidentified agglomerations of sentences amounting to sequentially linked functions, motifs etc.

4. MAPPING TEXT VARIANTS TO EXISTING CLASSIFICATION SCHEMES

To name but a few opportunities, one can use different domain-specific classification schemes, fragmentary or complete, to test the idea of motif extraction. E.g. it would be important to explore the relationship between concepts describing the life and accomplishments of the hero, unifying different typologies in one overarching concept [4]; or to study the folkloristic underpinnings of classical Greek mythology and their geographic distribution as contrasted with archaeological evidence [2, 31, 33, 43, 44]. Other areas of Proppians applied to plot analysis include creative writing [49, 50], collaborative narrative generation [13, 38, 47], or drama typology [61], among others.

These considerations justify the first research question as presented above.

5. FORMULAITY AND SCIENTIFIC COMMUNICATION

By analogy, how far do the above findings sit well with scientific communication? The very fact that bioinformatics successfully utilizes the concept of a motif advocates for its extended use beyond comparative literature, musicology, or the arts in general, its traditional application domains. We start with a standard application example and continue with another one about canonical content expressions in immunology and other fields [20, 21, 22].

5.1 Motifs in bioinformatics

In bioinformatics oftentimes the task is to compare a protein of unknown structure with its homologues of known 3-D structures. The homologues are modeled based on the idea of motifs. A motif definition is a Hidden Markov Model [52] stating that e.g. in a deoxyribonucleic acid (DNA) sequence, amino acids such as arginine, leucine, cysteine and histidine, follow each other with certain probabilities.

Another definition is as follows: ribonucleic acid (RNA) motifs are directed and ordered stacked arrays of non-Watson-Crick base pairs forming distinctive foldings of the phosphodiester backbones of the interacting RNA strands. They correspond to the 'loops' - hairpin, internal and junction - that intersperse the Watson-Crick two-dimensional helices as seen in two-dimensional representations of RNA structure. RNA motifs mediate the specific interactions that induce the compact folding of complex RNAs. RNA motifs also constitute specific protein or ligand binding sites. A given motif is characterized by all the sequences that fold into essentially identical three-dimensional structures with the same ordered array of isosteric non-Watson-Crick base pairs [34]. In yet another example, shared motifs having a similar 3-D structure in representatives of functionally diverse molecule families are called "molegos" (molecular legos), with e.g. a similar role in substrate binding, that is, functionality. This word based, sequence (motif) to structure (molego) to function method has clear implications for genomic analysis and template based homology modeling, as well as immediate application in recognizing specificity determinants in proteins that share active sites common to many enzymes [57].

5.2 Disciplinary sublanguages

Summing up [11], Zellig Harris proposed a theory of sublanguages that explains why it is possible to process language in specialized textual domains such as those found in genomics and medicine. According to this theory, the languages of technical domains have a structure and regularity which can be observed by examining the corpora of the domains, and which can be delineated so that the structure can be specified in a form suitable for computation. Whereas the theory of general English grammar primarily specifies well-formed syntactic structures only, Harris' sublanguage grammar theory also incorporates domain-specific semantic information and relationships to delineate a language that is more informative than English because it reflects the subject matter and relations of a domain as much as its syntactic structure.

Harris postulated that all occurrences of language are word sequences satisfying certain constraints which express and transmit information. His constraints were dependency relations, paraphrastic reductions, and inequalities of likelihood. Additionally, certain subsets of languages within specialized domains, called sublanguages, do exist that exhibit specialized constraints due to limitations of the words and relations of the subject matter.

In the grammar of a specialized sublanguage, operators and arguments still satisfy the dependency relations of the whole language and paraphrastic reductions still occur, but the vocabulary is limited, only restricted combinations of words occur, and subclasses of words combine in specified ways with other subclasses. In a sublanguage, words form subsets from the larger word classes of the overall language.

Thus, in order to create a sublanguage grammar, the critical task is to discover the subclasses and important relations. For each domain, clustering techniques [26] help to discover a limited number of word classes and sentence types for a large sample of a domain corpus. However, the sentences are in surface forms, and therefore, many reductions have occurred so that the sentences are complex and not necessarily in forms close to the underlying operator-argument forms, making the discovery task more difficult. Here, two general remarks should be indicative:

- Sublanguage analysis reveals formal structures in the sentences of the texts called sublanguage formulae. These

are similar to the formulae of logic, but with certain extensions. The significance of this approach to computational linguistics is that the initial phase of sublanguage analysis establishes a direct relationship between surface sentence forms and their semantic representation (i.e., the formulae). This mapping serves as a basic design for text processing algorithms [30];

- A sublanguage is characterized by a specialized vocabulary, semantic relationships, and in many cases specialized syntax. The purpose of its analysis is to establish classes of objects relevant in the domain, and classes of relations in which the objects participate. The technique groups different arguments of sentences (grammatical subjects or objects) into a class according to their occurrence in the texts with the same operator (main verb, adjective, or preposition). Operators are grouped into classes according to their occurring with the same classes of arguments. When the analysis is carried out on a sample of sufficient size, argument classes are found to correspond to domain objects, and operator classes to domain relations.

Formulae are well-formed expressions made up of an operator class and one or more argument classes, and correspond to the "events" of a domain. Johnson [30] cites Harris *et al.* [21] to bring the following example: let the argument classes include antibody (A), antigen (G), cell (C), tissue (T), and body part (B). Operator classes include inject (J), move (U), and present in (V). Then examples of formulae and the sublanguage sentences they represent are:

G J B "antigen was injected into the foot-pads of rabbits"

A V C "antibody is found in lymphocytes"

G U T "antigen arrives by the lymph stream"

Sublanguage formulae are a compact notation for knowledge representation that employ a number of devices to enrich the basic structure of operator-argument predication. Modifiers can be placed on operator and argument classes as superscripts. On arguments, they function as unary operators or as quantifiers. Modifiers of operators include negation, quantity, aspect, and direction (of movement). Subclasses of operator and argument classes are indicated by subscripts, e.g., cell (C) has subclasses lymphocyte (C_l) and plasma cell (C_z). A rich set of connectives can join pairs of formulae which can be implemented in a fairly straightforward fashion, the choice of implementation obviously depending on the complexity of the sublanguage being processed and on the application that will make use of the data [19].

A key aspect of the sublanguage method is that it is objective, relying on structural features of texts only and not ad hoc semantic judgements. Due to this property sublanguage analysis is repeatable, indeed the resulting sublanguage formulae were the same, regardless of the host language employed by scientists (English-French [21]) vs. English-Korean [22]).

The scientific grounding of Harris's sublanguage theory is well established and has been repeatedly verified by the vast amount of work that has been done in this area. A set of papers on sublanguage processing and research collected by Grishman and Kittridge [17]

includes the domains of lipoprotein kinetics, clinical patient reports, telegraphic Navy messages, and reporting of events in outer space. Additional work pertaining to the sublanguages of pharmacological literature and lipid metabolism is described in Sager [55]. Grishman [18] also mentions sublanguages of weather reports, aircraft repair manuals, scientific articles about pharmacology, hospital radiology reports, and real estate advertisements. Finally, recently the biomolecular domain [11] and social science [23] have been reportedly added to the list of those subject areas in scientific communication showing symptoms of formulaity.

In my eyes, the above sufficiently justify the second research question as well.

6. DISCUSSION

From this interdisciplinary comparison, worth mentioning are the following implications:

- Motif extraction [23] can be used for markup in bioinformatics since markup languages for e.g. chemistry [42], biopolymers [10] or microarray gene expressions do exist [58]. The same role can be assigned to any kind of canonical formula, regardless of their domains;

- One of the possible roadmaps ahead is to check whether sequences of latent variables overlap with sublanguage motifs according to Harris' suggestions. (Since Harris' idea of a controlled vocabulary is based solely on distributional statistics, this would fit the philosophy of latent semantic indexing (LSI) very well.) The issue of vector sequences for document modelling [66] needs also to be considered;

- Another task is to adapt Harris's sublanguage analytical method to studies of formulaic expression in oral and written communication, regardless of the domain;

- It is reasonably clear that practical applications of formulaity in scientific communication include storage of science information in databases, indexing the literature, and identification and resolution of controversy [23]. However, the missing link for the folklorist is, what to do practically with these structures once they have been recognized? What next? As at the other end of the research spectrum document indexing, classification and retrieval are vying for alternatives to mark up their material for subsequent processing by higher-order content indicators, we are dealing with consecutive steps in the same procedural chain: recognized content patterns can be used for information filtering, filtered extracts for markup by language technology and machine learning, and markup for subsequent document processing.

7. CONCLUSION

The AMICUS project was designed to promote scholarly networking in a topical area, the automated motif recognition in folklore and scholarly texts. Literary evidence from different subject fields suggests that in both domains, the formulaic structure of documents is a more or less known phenomenon therefore the automation of their recognition is possible. Such extracts with a sequential structure of content elements of different granularity can

be used for document processing. The methodological toolkit to tackle with issues of recognition to processing includes test collection building, document preprocessing by language technology, information extraction from corpora based on exemplification, and semantic markup by machine learning.

8. ACKNOWLEDGMENTS

9. REFERENCES

[1] Aarne, A. and Thompson, S. 1961. *The Types of the Folktale. A Classification and Bibliography. Second Revision (FFC 184).* Academia Scientiarum Fennica, Helsinki.

[2] Burkert, W. 1979. *Structure and history in Greek mythology and ritual.* University of California Press, Berkeley.

[3] Bärtschi, M.A. 1984. *Term dependence in information retrieval models*. Eidgenössische Technische Hochschule, Zürich.

[4] Campbell, J. 2004. *The hero with a thousand faces*. Princeton University Press, Princeton.

[5] Darányi, S. 2003. Factor analysis and the canonical forumal: where do we go from here? In *Proceedings of the Information Society, Cultural Heritage and Folklore Text Analysis Conference* (Budapest, Hungary, November 24-26, 2003). Department of Information and Knowledge Management, Budapest University of Technology and Economics, Budapest, 55-63.

[6] Darányi, S. 2007. First- and second-order change as symmetry and symmetry breaking in folklore text content evolution: From Heraclitus to Lévi-Strauss. In *Symmetry: Art and Science* Vol. 2-4, C. F. Guerri and D. Nagy, Eds., University of Buenos Aires, Buenos Aires, 162-165.

[7] Darányi, S. 1996. Formal Aspects of Natural Belief Systems, Their Modelling and Evolution: A Semiotic Analysis. *Semiotica* 108 - 1/2, 45-63

[8] Ding, C.H.Q. 2005. A Probabilistic Model for Latent Semantic Indexing, *Journal of the American Society for Information Science and Technology* 56(6), 597–608.

[9] Dundes, A. 1962. From Etic to Emic Units in the Structural Study of Folktales. *Journal of American Folklore* 75, 95–105.

[10] Fenyo, D. (1999). The Biopolymer Markup Language. *Bioinformatics* 15,4, 339-340.

[11] Friedman, C., Kra, P. and Rzhetsky, A. 2002. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics* 35, 222–235.

[12] Garnham, A. 1983. What's wrong with story grammars. *Cognition* 15, 145-154.

[13] Gervás, P., Díaz-Agudo, B., Peinado, F. and Hervás, R. 2005. Story plot generation based on CBR. *Knowledge-Based Systems* 18, 4-5, 235-242.

[14] Griffin, M.2001. An expanded, narrative algebra for mythic spacetime. *Journal of Literary Semantics* 30, 71–82.

[15] Griffin, M.2003. More features of the mythic spacetime algebra. *Journal of Literary Semantics* 32, 49–72.

[16] Griffin, M.2006. Mythic algebra uses: Metaphor, logic, and the semiotic sign. *Semiotica* 158–1/4, 309–318.

[17] Grishman, R. and Kittredge, R. Eds. 1986. *Analyzing language in restricted domains: Sublanguage description and processing.* Lawrence Erlbaum Associates, Hillsdale.

[18] Grishman, R. 2001. Adaptive Information Extraction and Sublanguage Analysis. http://nlp.cs.nyu.edu/publication/papers/grishman- (20-09-10)

[19] Habert, B.and Zweigenbaum, P. 2002. Contextual acquisition of information categories: what has been done and what can be done automatically? In *The Legacy of Zellig Harris: Language and information into the 21st Century, Mathematics and computability of language.* Vol. 2. B.E. Nevin and S.M. Johnson, eds. John Benjamins, Amsterdam, 203-231.

[20] Harris, Z. 1988. *Language and information*. Columbia University Press, New York.

[21] Harris, Z.S., Gottfried, M., Ryckman, T., Mattick, P., Daladier, A., Harris, T.N. and Harris, S. 1989. *The form of information in science: analysis of an immunology sublanguage.* Kluwer, Dordrecht.

[22] Harris, Z.S. 1991. *A theory of language and information: a mathematical approach.* Clarendon Press, Oxford.

[23] Harris, Z. S. 2002. The structure of science information. *Journal of Biomedical Informatics* 35, 215–221.

[24] Hartmann, K., Hartmann, S. and Feustel, M. 2005. Motif definition and classification to structure non-linear plots and to control the narrative flow in interactive dramas. In *Proceedings of the Third International Conference on Virtual Storytelling* (Strasbourg, France, November 30 – December 2, 2005). LNCS 3805, 158-167. Springer, Berlin.

[25] Haverty, P.M. and Weng, Z. 2004. CisML: an XML-based Output Format for Sequence Motif Detection Software. *Bioinformatics Advance Access* published March 4. http://bioinformatics.oxfordjournals.org/content/early/2004/03/04/bioinformatics.bth162.full.pdf (20-09-10)

[26] Hirschman, L. and Grishman, R. 1975. Grammatically-based automatic word class formation. *Information Processing and Management* 11, 39-57.

[27] Jason, H. and Segal, D. Eds. 1977. *Patterns in oral literature*. Mouton, the Hague.

[28] Jason, H. 2000. *Motif, Type and Genre. A Manual for Compilation of Indices & A Bibliography of Indices and Indexing* (FFC 273). Academia Scientiarum Fennica, Helsinki.

[29] Jason, H. 2007. About 'Motifs', 'Motives', 'Motuses', '-Etic/s', '-Emic/s', and 'Allo/s-', and How They Fit Together: An Experiment in Definitions and in Terminology. *Fabula* 48, No 1-2, 85-99.

[30] Johnson, S.B.1989. Review of Harris et al.(1989): The form of information in science: analysis of an immunology

sublanguage, Kluwer, Dordrecht. *Computational Linguistics* 15, 3, 190-192.

[31] Kirk, G.S. 1970. *Myth: its meaning and functions in ancient and other cultures.* University of California Pres, Berkeley.

[32] Lakoff, G.P. 1972. Structural complexity in fairy tales. *The Study of Man, I,* 128-190.

[33] Leach, E. 1973. *Claude Lévi-Strauss.* The Viking Press, New York.

[34] Leontis, N.B. and Westhof, E. 2003. Analysis of RNA motifs. *Current Opinion in Structural Biology* 2003, 13, 300-308.

[35] Lévi-Strauss, C. 1958. The structural study of myth. In *Myth: A Symposium.* T.A. Sebeok, Ed. Indiana University Press, Bloomington, 50-66.

[36] Lévi-Strauss, C. 1964-1971 *Mythologiques I-IV*. Plon, Paris.

[37] Lord, A. 1960. *The singer of tales.* Harvard University Press, Cambridge.

[38] Lönneker, B., Meister, J.C., Gervás, P., Peinado, F. and Mateas, M. 2005. Story generators: models and approaches for the generation of literary artefacts. In *Conference Abstracts of the 17th Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing* (Victoria, BC, Canada, June 2005). Humanities Computing and Media Centre, University of Victoria, 126-133.

[39] Luhn, H.P. 1957. A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development 1, 4, 309-317.

[40] Maranda, P. Ed. 2001. *The double twist: from ethnography to morphodynamics.* University of Toronto Press, Toronto.

[41] Morava, J. 2003. The Klein group and its generalizations in the work of Lévi-Strauss. In *Proceedings of the Information Society, Cultural Heritage and Folklore Text Analysis Conference* (Budapest, Hungary, November 24-26, 2003). Department of Information and Knowledge Management, Budapest University of Technology and Economics, Budapest, 48-54.

[42] Murray-Rust, P., Leach, C., Rzepa, H.S. 1995. Chemical Markup Language. *Abstr. Pap. Am. Chem. Soc.* 210, 40-COMP Part 1.

[43] Nilsson, M.P. 1964. *A history of Greek religion.* W.W. Norton and Company, New York.

[44] Nilsson, M.P. 1972. *The Mycenaean origin of Greek mythology*. University of California Press, Berkeley.

[45] Parry, M. 1930. Studies in the Epic Technique of Oral Verse-Making. I: Homer and Homeric Style. *Harvard Studies in Classical Philology* 41, 73–143.

[46] Parry, M. 1932. Studies in the Epic Technique of Oral Verse-Making. II: The Homeric Language as the Language of an Oral Poetry. *Harvard Studies in Classical Philology* 43, 1–50.

[47] Peinado, F. and Gervás, P. 2006. Evaluation of automatic generation of basic stories. *New Generation Computing* 24, 3, 289-302.

[48] Pike, K.L. 1967. *Language in Relation to a Unified Theory of the Structure of Human Behavior*. Mouton, The Hague.

[49] Polti, G. 1922. *The art of inventing characters*. James Knapp Reeve, Franklin, Oh.

[50] Polti, G. 1924. *The thirty-six dramatic situations.* James Knapp Reeve, Franklin, Oh.

[51] Propp, V.J. 1968. *Morphology of the folktale*. University of Texas Press, Austin.

[52] Rabiner, L.M. and Juang, B.H. 1986. An introduction to Hidden Markov Models. *IEEE ASSP Magazine* 3,1, 4-16.

[53] Rocchio, J.J. 1971. *Relevance feedback in information retrieval*. In *The SMART Retrieval System – Experiments in Automatic Document Processing*. G. Salton, ed. Prentice Hall Inc., Englewood Cliffs, 313–323.

[54] Roshianu, N.1974. *Traditsionnuie formuly skazki (Traditional formulae of the fairy tale).* Moscow.

[55] Sager, N. 1986. Sublanguage: Linguistic Phenomenon, Computational Tool. In *Analyzing language in restricted domains: sublanguage description and processing*. R. Grishman and R. Kittredge, eds. Lawrence Erlbaum Associates., Hillsdale.

[56] Schäuble, P. 1987. Thesaurus based concept spaces. *Proceedings of the 10th annual international ACM SIGIR conference on research and development in information retrieval*, 254–262.

[57] Schein, C.H., Zhou, B., Oezguen, N., Mathura, V.S. and Braun, W. 2005. Molego-based definition of the architecture and specificity of metal-binding sites. *PROTEINS: Structure, Function, and Bioinformatics* 58, 200-210.

[58] Spellman, P. T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M., Swiatek, M., Marks, W.L., Goncalves, J., Markel, S., Iordan, D., Shojatalab, M., Pizarro, A., White, J., Hubley, R., Deutsch, E., Senger, M., Aronow, B. J., Robinson, A., Bassett, D., Stoeckert, C. J., and Brazma, A. 2002. Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biology 3, RESEARCH0046.

[59] Thompson, S. 1946. *The Folktale*. The Dryden Press, New York.

[60] Thompson, S. 1955-1958. *Motif-Index of Folk-Literature 1–6.* Indiana University Press, Bloomington.

[61] Tomaszewski, Z. and Binsted, K. 2007. The limitations of a Propp-based approach to interactive drama. In *Proceedings of the AAAI Fall Symposium on Intelligent Narrative Technologies* (Westin Arlington Gateway, Arlington, Virginia, November 9-11, 2007).

[62] Uther, H. J. 2004. *The Types of International Folktales. A Classification and Bibliography. Based on the System of Antti Aarne and Stith Thompson 1–3 (FFC 284–286).* Academia Scientiarum Fennica, Helsinki.

[63] Vaz Lobo, P. and Martins de Matos, D. 2010. Fairy tale corpus organization using latent semantic mapping and an item-to-item top-n recommendation algorithm. In *Proceedings of LREC 2010* (La Valetta, Malta, May 19-21, 2010). European Language Resources Association, 1472-1475.

[64] Voigt, V., Preminger, M., Ládi, L. and Darányi, S. 1999. Automated motif identification in folklore text corpora. *Folklore* 12, 126-141.

[65] Zipes, J. Ed. 2000. *The Oxford companion to fairy tales*. Oxford University Press, Oxford.

[66] Yamamoto, A. and Agiso, A. 2004. Similarity of documents based on the Vector Sequence Model. In *Intuitive Human Interface 2004,* LNAI 3359, G. Grieser and Y. Tanaka, Eds. Springer, Berlin, 233–242.