DETECTING MULTIPLE MOTIF CO-OCCURRENCES IN THE AARNE-THOMPSON-UTHER TALE TYPE CATALOG: A PRELIMINARY SURVEY

Sándor Darányi[*]

University of Borås. Swedish School of Library and Information Science.

László Forró[**]

Abstract: Catalogs project subject field experience onto a multidimensional map which is then converted to a hierarchical list. In the case of the Aarne-Thompson-Uther Tale Type Catalog (ATU), this subject field is the global pattern of tale content defining tale types as canonical motif sequences. To extract and visualize such a map, we considered ATU as a corpus and analysed two segments of it, "Supernatural adversaries" (types 300-399) in particular and "Tales of magic" (types 300-749) in general. The two corpora were scrutinized for multiple motif co-occurrences and visualized by two-mode clustering of a bag-of-motif co-occurrences matrix. Findings indicate the presence of canonical content units above motif level as well. The organization scheme of folk narratives utilizing motif sequences is reminiscent of nucleotid sequences in the genetic code.

Keywords: tale type; motif space; motif co-occurrence; 2-mode clustering; visualization.

1. INTRODUCTION

"As the history of type and motif indexes shows, the search for principles serving the classification of folk narratives has not yet produced a satisfying system, but indexes have provided scholars with 'many valuable and practical research instruments, many methodical and theoretical by-products', as Vilmos Voigt (Voigt, 1977, p. 570) asserts." (Uther, 2009, p. 11). Uther also states, based on *Acta Ethnographica*, that outlines of a new international classification are now emerging (Uther, 2009, p. 10). Here we continue to show the relevance of automatic text classification for folklore archives (Voigt *et al.*, 1999), with or without machine learning, to such studies.

We depart from the assumption that in catalogs, one meets domain-specific knowledge mapped onto a hierarchical structure. However, by nature such knowledge is also multivariate, i.e. describes many objects of the subject field by many characteristic features, and can be expressed by multivariate classification methods, with or without information visualization.

[*] sandor.daranyi@hb.se

[**] salmonix@gmail.com

Recibido: 22/08/2011; 2ª revisión: 17/01/2012; aceptado: 24/01/2012.

DARÁNYI, S. y FORRÓ, L. Detecting multiple motif co-occurrences in the Aarne-Thompson-Uther tale type catalog: a preliminary survey. *Anales de Documentación*, 2012, vol. 15, nº 1. http://dx.doi.org/10.6018/analesdoc.15.1.134691.

The case we want to test this hypothesis on is the Aarne-Thompson-Uther Tale Type Catalog (ATU), a classification and bibliography of international folk tales (Uther, 2004). In the ATU, tale types are defined as canonical motif sequences such that motif string A constitutes Type X, string B stands for Type Y, etc. Also, it is important to note that types were not conceived in the void, rather they extract the essential characteristic features of a body of tales from all corners of the world, i.e. they are quasi-formal expressions of typical narrative content, mapped from many to one. ATU is an alphanumerical, basically decimal classification scheme describing tale types in seven major chapters (animal tales, tales of magic, religious tales, realistic tales [novelle], tales of the stupid ogre [giant, devil], anecdotes and jokes, and formula tales), with an extensive Appendix discussing discontinued types, changes in previous type numbers, new types, geographical and ethnic terms, a register of motifs exemplified in tale types, bibliography and abbreviations, additional references and a subject index.

The numbering of the types runs from 1 to 2399 (in fact, 2411). Individual type descriptions uniformly come with a number, a title, an abstract-like plot mostly tagged with motifs, known combinations with other types, technical remarks, and references to the most important literature on the type plus its variants in different cultures. At the same time, as the inclusion of some 250 new types in the Appendix indicates, tale typology is a comprehensive and large-scale field of study, but also unfinished business: not all motifs in the Aarne-Thompson Motif Index (AaTh; Thompson, 1955-58) were used to tag the types, difficulties of the definition of a motif imposed limitations on its usability in ATU, and narrative genre related considerations related to classification in general had to be observed.[1] Together with AaTh, ATU is the standard reference tool for librarians and digital curators alike, although other manuals such as Jason (2000) also come handy as means of orientation. However, when using ATU, it is regarded as a matter of fact that its descriptive units, motifs, constitute the highest level of abstraction, and there are no units of content above this. Therefore our research question was, does this assumption hold? If one regards the ATU type descriptions as text, is its content evenly distributed as in the case of a divisive classification with no overlapping categories, or is there granularity (heterogeneity) to it?

As document indexing, document classification, information retrieval and information visualization are closely related research areas with co-dependent methodologies to handle a complex phenomenon, the distribution of semantic content in documents, we regard it important to extend their study from scientific literature to folk narratives as a genre important for digital libraries in any information society.

This paper is organized as follows: Section 2 discusses experiment design, Section 3 and 4 present and discuss the results. Finally, Section 5 offers our conclusions.

2. EXPERIMENT DESIGN

A. Materials and methods

To extract and visualize a map of the respective segments, we considered ATU as a corpus and analysed sub-section “Supernatural adversaries” (types 300-399) in particular and section “Tales of magic” (types 300-749) in general. The two corpora were scrutinized for multiple motif co-occurrences and visualized by the two-mode clustering of a bag-of-motif co-occurrences (BOMC) matrix. After having excluded types not indexed by motifs at all, the first part of the experiment (300-399) worked with 52 tale types defined on the basis of 281 motifs, and the second part (300-745A) with 219 types and 1202 motifs, respectively.

After some early structural exploration by multidimensional scaling (MDS, PROXSCAL algorithm) which yielded inconclusive results (Figure 1), we turned to motif co-occurrence extraction. We augmented a standard lexicographic combination algorithm to compute combinations of tokens with a posting list indexing to calculate frequencies for each token. We applied a frequency threshold to filter out valid but infrequent co-occurrences, assuming that significant occurrences are the more frequent ones (see Figure 2 for the pseudocode). This reduced the combinatorial results to manageable size. We will refer to multiple co-occurrences as *multiplets* below (i.e. duplets, triplets, etc.). Respectively, multiplet-type matrices were constructed for two-mode clustering and visualisation.

For the latter, we used HCE3 (Seo and Shneiderman, 2004). This is a program developed for genome sequence analysis but can be used to text structure analysis as well. For the results presented here, we applied row by row normalisation of data with single linkage (nearest neighbour) clustering using Euclidean distance as a similarity measure.

As an example for items in these corpora, tale type 725 (*Prophecy of Future Sovereignty*) reads as follows: “A clever boy refuses to tell his dream (about his future sovereignty) [M312.0.1, D1812.3.3] to his father and to the king. He is punished and endures various adventures (imprisonment) [L425]. A princess nourishes him in prison. War is to be declared on the emperor if he is not able to solve two riddles and a task. The clever boy solves the riddles and the task, tells the answers to the princess, and is freed from prison. So the boy averts war, marries the princess [H551], and finally receives two kingdoms.” (Uther, 2004, p. 390).

Clearly, the backbone of this type is the motif sequence [M312.0.1/D1812.3.3][L425][H551] where / refers to a forking alternative.

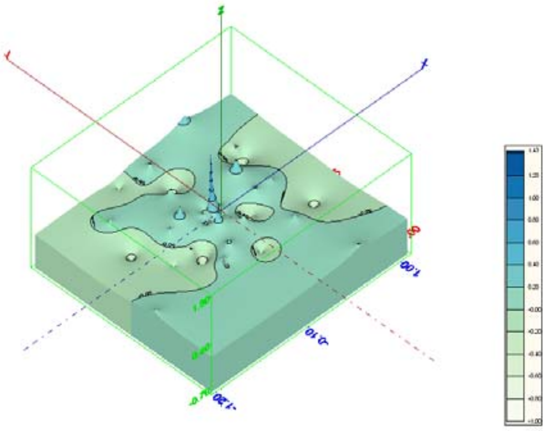


Figure 1. Type clusters as sinks in motif space (Supernatural adversaries segment of ATU).

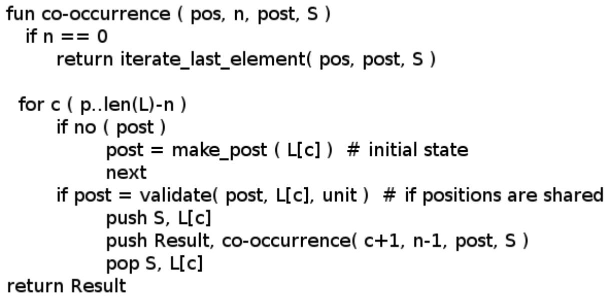
B. Background considerations

Propp’s analysis aimed to find the elementary building blocks of narrative structures. His approach belongs to one of the two major types of structural analysis, also called “syntagmatic” because it results in a sentence-like statement which encapsulates the plot of a fairy tale by a sequence of situations called *functions*.[2] He concluded that the set of Russian fairy tales used for the exemplification of his theory could be described by 31 such functions, and that only canonical sequences of such functions (i.e. a limited set of action types used by another limited set of actors) result in “valid”, i.e. acceptable Russian fairy tales (Propp, 1968).

With typical characters populating these functional situations, organized into higher order plot units called *moves*, this formal toolkit was described by Propp as the morphology of the fairy tale, a vision which can be extended to morphometric studies in general. However without computerized large-scale comparative studies it is an open question if the same functions, or the same number of functions, are sufficient to encode non-Russian and/or non-contemporary fairy tales over cultures. Motifs and motif sequences are not identical with, although related to, Propp’s functions, but the exact nature of dependency of the two atomic approaches is not known for the time being. Nevertheless the two problems beg for a single solution which can handle content unit sequences (and is therefore related to the next frontier of sentence-based information retrieval), charting new territory in folklore studies and therefore of utmost interest to digital libraries.

3. RESULTS

Early on, for Part A of the experiment MDS indicated granularity in motif space (Figure 1), but type clusters – sinks in the motif landscape – were constructed on purely formal grounds, i.e. how many motifs had indexed groups of types. This result was inconclusive to decide about the null hypothesis.



```
fun co-occurrence ( pos, n, post, S )
  if n == 0
      return iterate_last_element( pos, post, S )

 for c ( p..len(L)-n )
      if no ( post )
            post = make_post ( L[c] )  # initial state
            next
      if post = validate( post, L[c], unit )  # if positions are shared
            push S, L[c]
            push Result, co-occurrence( c+1, n-1, post, S )
            pop S, L[c]
return Result
```

Figure 2. The co-occurrence extraction algorithm.

At the same time we found that multiplets occurred among the motifs. Their list for the *Supernatural adversaries* segment of ATU with the respective type numbers is given in Table 1. Such motif strings are displayed by two-mode clustering as horizontal band lines per type which, if the motif co-occurs in several types, form blocks (Figures 3-4).

Further, where they occurred in more than one tale type, all the triplets and quadruplets in Table 1 were also collocated, following the same sequential arrangement (story line).

	Motif numbers	**Types**
1	E341-M241-M241.1	505, 507
2	H1210.1-H1242-K1932	550, 551
3	R155.1-D231-F171.3-F171.1	471
4	B211.1.8-B422-B435.1-F771.4.1	545A, 545B
5	L161-C611-K1933-T68.1	301D
6	Z16-H621.2-H504-F660.1	653
7	Q2–S31-G466-H935	480
8	S31-K1911-K1911.1.2-D688	403, 450

Table 1. Motif triplets (1-2) and quadruplets (3-8) in the "*Supernatural adversaries*" ATU segment.

In Part B of the experiment, dealing with the "Tales of magic" section of ATU, we repeated motif duplet and triplet detection but did not visualize the results. To manage combinatorial explosion, we applied a detection threshold which filtered out co-occurrences below specified levels. Statistics are displayed in Table 2.

Based on the above, the working (null) hypothesis could be rejected because we have found granularity in ATU on two levels, in the pattern of motif co-occurrences and in collocated motif co-occurrences.

Threshold	**Duplets**	**Triplets**
1	4293	618980
2	66	1408
3	4	16

Table 2. Motif duplet and triplet statistics for the "*Tales of magic*" ATU segment.

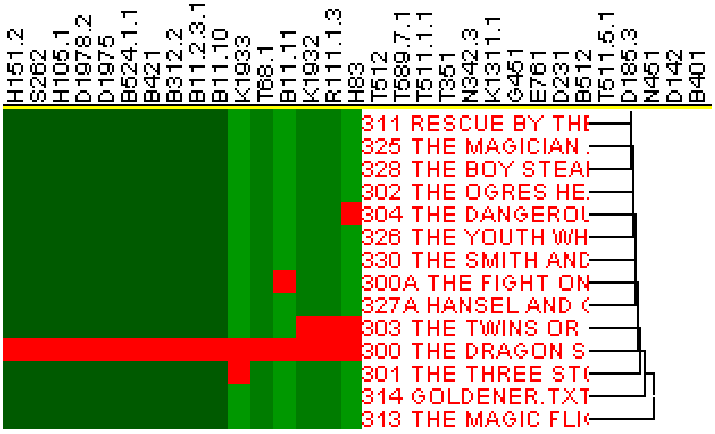


Figure 3. The red block in the middle indicates three co-occurring motifs in tale types 300 and 303.

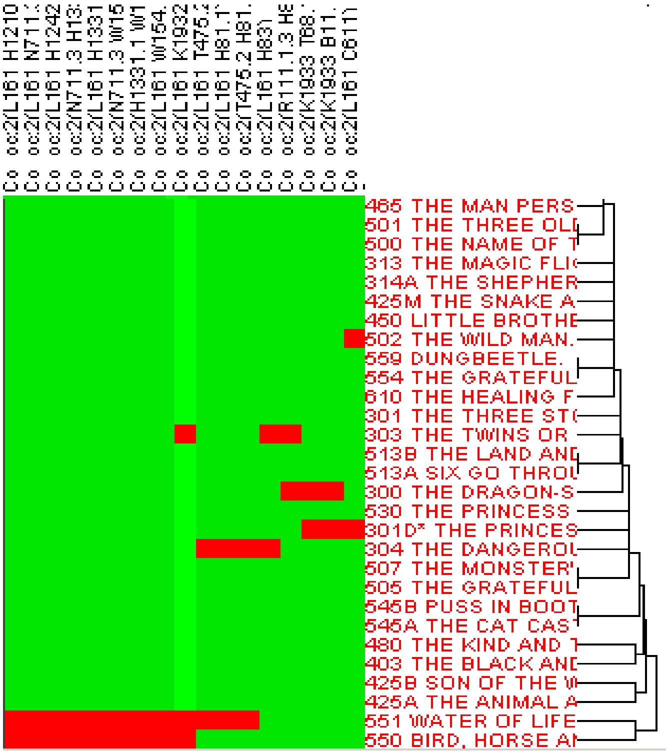


Figure 4. Blocks of motif duplets over tale types.

4. DISCUSSION

By two-mode clustering, the general view one gains is that the structure of ATU is mostly non-overlapping: types defined as motif strings are *almost* unique, their length depends on the number of motifs characteristic for a type, and to compare two tale types equals the matching of their respective motif strings. Whereas on word and sentence level this results in expressions of content similarity, e.g. by mapping tales as clusters of locations into some space and recognising types as their centroids, on merely formal grounds – as was the case with texts being indexed by motif numbers only – the result will reflect formal similarity, e.g. type clusters based on the number of motifs in them. For motifs and types, the striking novelty of multiple co-occurrence analysis was that the motif strings are not *entirely* unique, i.e. some of them have been persistent enough to be reused in different plots.

Apart from being eye-openers as well, these results are interesting for two major reasons. The first broad context is the perception of text variation as an evolutionary process, and the task of mapping evolving semantic content onto structures with both hierarchical and multivariate access. In this frame of thought, the reason why some motif strings have evolved and survived relates to a kind of selection pressure in a cultural historical setting, yet to be modelled. To this end, ATU and AaTh as tools have pioneered and mastered the hierarchical approach to content description but are wanting in terms of being understood as multivariate products at the same time. This is a current deficiency that cannot be overlooked or neglected when it comes to any kind of their overhaul in and for a digital environment.

In other words we need descriptive units of content which can index the source material in its entirety, are both multivariate by nature and fit the hierarchical classification structure, plus are flexible enough to evolve, that is, become more and more enriched variants of the original standard classifications. Indexing by single text words or phrases plus by motifs is clearly not enough to meet this goal – the existence of persistent motif strings in multiple copies underlying several types proves that more than one level of semantic metadata may pertain to the body of tales we want to index.

The other broad context is the parallel between the linguistic and the genetic code as vehicles of information transfer over time. Both use coded transfer mechanisms to transmit their messages, capture instructions to reproduce meaning from form (we regard context as form here); and in both, sequence plays an important role in the coding and decoding process.

Tale types as motif sequences follow the sublanguage approach to content representation, pioneered by Harris (2002). As pointed out by Darányi (2010), this domain-specific practice from the life sciences can be recognized in formal descriptions of narrative content, too. A few similarities between their communication patterns can be considered for methodology import between the two domains:

1. Content is sequential, coded by an alphabet and compiled based on the combinations of its elements, i.e. irrespective of their order on a basic observation level. This holds for nucleotides – the building blocks of nucleic acids such as DNA and RNA – and motifs, the building blocks of tale types alike.
2. On a next level, adding grammar and moving over to permutations, sequences start to play a role. Canonical nucleotide sequences generate secondary and tertiary – in fact spatial – structures such as the famed double helix; canonical motif sequences may contribute to the evolution of tale types, themselves representatives of tale variants in the plenty. Moreover, function sequences develop into fairy tale subtypes as shown by plot analysis (Propp, 1968), and canonical mytheme sequences constitute myths and mythologies (Lévi-Strauss, 1964-71; Maranda, 2001). In a sense, reading and understanding the genetic code and narratives alike demands the mastering of abstract grammars with their equally abstract vocabularies.
3. The concept of motifs is widely used in bioinformatics as well. Motifs in this sense mean primary nucleotide sequences of functional importance for structure generation. Sequential motifs include structural and regulatory motifs, with different functionalities pertaining to them; there may be methodological undercurrents linking the two knowledge domains which would need to be explored in more detail.

Chromosome and story mutations may be more similar than thought previously. Chromosomal mutations produce changes in whole chromosomes (more than one gene), or in the number of chromosomes present, with the major types being (a) deletion – loss of part of a chromosome; (b) duplication – extra copies of a part of a chromosome; (c) inversion – reverse the direction of a part of a chromosome; and (d) translocation – part of a chromosome breaks off and attaches to another one.

Whereas most mutations are neutral and have little or no impact on the functionality of the product, their adding up can dramatically affect the survival rate of the outcome, leading to new genotypes and phenotypes in the course of evolution. In the same vein, deletion and translocation could be standard tools in the narrative building toolkit; inversion is suggested to play a central role in the Bible (Christensen, 2003), and duplication is evident e.g. in the case of the Proppian narrative scheme where complete tale moves may be repeated several times or combined with one another by different embeddings (Propp, 1968). This indicates the need for a theory of text evolution as a series of narrative element recombinations, forming from simple to more complex structures by “mutation mechanisms”.

5. CONCLUSIONS

For a proof-of-concept investigation, we analysed two segments of ATU to find out whether the catalog contained any internal structure as a reflection on overlapping narrative content in the real world. Tale types were indexed by their motifs and the resulting matrices were exposed to two-mode clustering and multiple co-occurrence analysis, respectively. Visualised results were used to highlight those motif combinations which occurred above a frequency threshold and thereby could be regarded as emerging structures in solidification.

Preliminary findings suggested that our line of thought worked because the null hypothesis could be rejected. Due to this, one can consider tale types as strings of single motifs *and* their multiplets, sort of "motif phrases", which is new evidence. In our eyes, the popping up of the latter is proof for text evolution. It is our understanding that two-mode clustering isolated the raw material (i.e. non-collocated sequences) of motif strings acting in their collocated variants as "narrative nucleotides". However the nature of motif collocation will demand more detailed investigation. We are looking forward to applying this technique with cautious optimism. As the AaTh contains about 40.000 motifs (Thompson, 1955-58), this would allow for the prevalence of motif sequences as a new kind of metadata, and enable the use of both single and chained motifs as tags for semantic markup.

On a more general level, any catalog using a faceted classification scheme can benefit from these considerations. Such classification schemes combine elements from different parts of the facet structure to construct a number representing the subject content (often combining two subject elements with linking numbers and geographical and temporal elements) and form of an item rather than drawing upon a list containing each class and its meaning. Such items may include facet sequence, i.e. a primitive grammar, and are therefore language-like, also called indexing languages. Examples include e.g. Ranganathan's Colon Classification (CC), the Library of Congress classification system (LoC), or aspects of the Dewey Decimal Classification (DDC) and the Universal Decimal Classification (UDC).

ACKNOWLEDGEMENT

The first author expresses his gratitude to Prof. Hans-Jörg Uther (University of Duisburg-Essen) for correspondence on the subject.

NOTAS

[1] Hans-Jörg Uther, personal communication.

[2] The other major variant is Lévi-Strauss' "paradigmatic" analysis which disregards such sequences and is looking for patterns in the material.

6. REFERENCES

CHRISTENSEN, D.L. *The unity of the Bible: exploring the beauty and structure of the Bible*. Mahwah, N.J.: Paulist Press, 2003.

DARÁNYI, S. Examples of Formulaity in Narratives and Scientific Communication, in Darányi, S. and Lendvai, P. (eds). Proc. 1st Intl. AMCUS Workshop on Automated

Motif Discovery in Cultural Heritage and Scientific Communication Texts. Vienna, Austria, 2010, p. 29-35.

HARRIS, Z.S. The structure of science information. *Journal of Biomedical Informatics*, 2002, vol. 35, p. 215–221.

JASON, H. *Motif, Type and Genre. A Manual for Compilation of Indices & A Bibliography of Indices and Indexing*. Helsinki: Academia Scientiarum Fennica, 2000.

LEVI-STRAUSS, C. *Mythologiques I-IV*. Paris: Plon, 1964-71.

MARANDA, P. (ed). *The double twist: from ethnography to morphodynamic*. Toronto: University of Toronto Press, 2001.

PROPP, V.J. *Morphology of the folktale*. Austin: University of Texas Press, 1968.

SEO, J. and SHNEIDERMAN, B. *A Rank-by-Feature Framework for Unsupervised Multidimensional Data Exploration Using Low Dimensional Projections*. Proc. IEEE InfoVis2004, Austin, USA, 2004, p. 65-72.

THOMPSON, S. Motif-Index of Folk-Literature 1–6, Indiana University Press, Bloomington. (1955-58).

UTHER, H.J. *The Types of International Folktales: A Classification and Bibliography. Based on the System of Antti Aarne and Stith Thompson*, Part I. Helsinki: Academia Scientiarum Fennica. 2004.

UTHER, H.J. Classifying tales: Remarks to indexes and systems of ordering. *Narodna umjetnost*, 2009, vol. 46, nº 1, p. 15-32.

VOIGT, V.; PREMINGER, M.; Ládi, L. and Darányi, S. Auto-mated motif identification in folklore text corpora. *Folklore*, 1999, vol. 12, p. 126-141.