

# 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

1st International AMICUS Workshop on Automated Motif  
Discovery in Cultural Heritage and Scientific  
Communication Texts  
October 21, 2010 Vienna

# Theoretical underpinnings

- 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

# Formulaity, formulaic

- The theory of **oral-formulaic composition** originated in the scholarly study of epic poetry, being developed in the 2nd quarter of the 20th century. It seeks to explain two related issues:
  - The mechanism whereby some oral poets are able to improvise poetry, and
  - Why orally improvised poetry has the characteristics it has
- The key idea of the theory is that poets have a store of formulae (a formula being “an expression which is regularly used, under the same metrical conditions, to express a particular essential idea”) and that by linking these in conventionalized ways, they can rapidly compose verse
- Milman Parry (1902-1935), Albert Lord (1912-1991): their approach transformed the study of ancient and medieval poetry, and oral poetry in general, with an impact on narratology
- Standard sequences of content elements (formulae) pertain to documents and document parts

# Motif definitions

- Motif (literary) [Jason 2007]: A literary motif is the *simplest* unit of content which matches a primary formal slot of a literary structure: a “character” (being/object [noun] with an attribute [adjective]) plus a “deed” (verb and adverb) fill in a “narrative role” and a “narrative action”
- 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/reused by any plot, genre, or ethnos in a certain cultural area, or even universally
- Motif (my current working hypothesis):
  - Collocation of semantic fields
    - In corpus linguistics, **collocation** defines a sequence of words or terms that co-occur more often than would be expected by chance
    - Where **semantic field** can be represented by the normal form of related terms / a normalized sentence based on abstracting / a latent variable / a cluster of terms

# Parallels

- Motif (bioinformatics): a Hidden Markov Model 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
- Parallels between language and the genetic code (going back to alleged "information exchange" between chemical reactants)
- Same or very similar structural principles at work [Darányi 2000]:
  - Unity of form and content (energy or impact)
  - Replacement by neighbour (chemical reactions, sense relations)
  - Message is sequential
  - Chemical reactions as communication go back to the principle of energy minimization by gradient descent on a potential surface; theoretical fragments from linguistics, computer science, information science point in the same direction, i.e. human communication striving toward more condensed norms, surfing a semantic potential surface [Petitot 2001, Taira et al 2007]

# Sublanguages

- Zellig Harris (1909-1992). A mathematical theory of language and information
- Sublanguage: the language of a restricted domain, particularly a technical domain, in a sense a normal form
- In mathematical terms, "a subset of the sentences of a language forms a sublanguage of that language if it is closed under some operations of the language: e.g., if when two members of a subset are operated on, as by *and* or *because*, the resultant is also a member of that subset" [Harris 1988: 34]
- Suitable for a formulaic, condensed, language-independent description of content (examples below)
- Disciplines where sublanguages have been studied: immunology, lipoprotein kinetics, pharmacology, biomedical domain, clinical patients' radiology reports, telegraphic Navy messages, event reports from outer space, weather reports, aircraft repair manuals, real estate advertisements, social science

# Example 1: Tale types in ATU

- TALES OF MAGIC, SUPERNATURAL ADVERSARIES 300-399
- Tale type 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]. (...)”
- Motifs:
  - B421: *Helpful dog*
  - B312.2: *Helpful animals obtained by exchange*
- Sequential, collocated (syntax of specific content applies)
- Content approached as collocated semantic fields (of normalized terms)
- Leading to motif markup

# Examples 2-3: Lévi-Strauss vs. Harris

$$f_x(a) : f_y(b) :: f_x(b) : f_a^{-1}(y)$$

**A V C "antibody is found in lymphocytes"**

---

Cadmos seeks his sister Europa, ravished by Zeus

Cadmos kills the dragon

The Spartoi kill one another

Oedipus kills his father, Laios

Oedipus kills the Sphinx

Oedipus = *swollen-foot* (?)

Oedipus marries his mother, Jocasta

Eteocles kills his brother, Polynices

Antigone buries her brother, Polynices, despite prohibition

---

Labdacos (Laios' father) = *lame* (?)

Laios (Oedipus' father) = *left-sided* (?)

Canonical Structure of a Discourse						
C	H	V	R	W	K	M
	Proteins	have	optical rotatory power	very sensitive to	the experimental conditions	
wh-	"	"	"	measured under	"	
, particularly	"	"	"	very sensitive to	the wavelength of the light which is used	
	"	"	optical rotation	affected by	a diversity of factors	that... is in many ways an advantage
	several proteins	"	rotatory properties			One of the authors (J.A.S.) has for some time been engaged in a study of...
, wh-	"	"	"	including the effects of	temperature	
,	"	"	"	"	wavelength	
,	"	"	"	"	pH	
and	"	"	"	"	the denaturation reaction	

Information

75

# Decoding the canonical formula of myth

E.g.

*Complete*

a = divine adult male

b = mortal adult male

-a = divine adult female

-b = mortal adult female

x = creates (active voice)

-x = is created (passive voice)

y = destroys (active voice)

-y = is destroyed (passive voice)

*Incomplete, attenuated*

1/a = divine adolescent male

1/b = mortal adolescent male

-1/a = divine adolescent female

-1/b = mortal adolescent female

1/x = heals (active voice)

-1/x = is healed (passive voice)

1/y = wounds (active voice)

-1/y = is wounded (passive voice)

The swapping of function for term values, i.e.  $f_x(a) \rightarrow f_a(x)$  should read as something affecting somebody else vs. the self, therefore:

$f_x(a) : f_y(b) :: f_x(b) : f_a^{-1}(y) \rightarrow$  "divine adult male creates : mortal adult male destroys :: mortal adult male creates : divine adolescent male destroys himself"

# Harris vocabulary ("actor types")

## LIST OF SYMBOLS

### I. Noun Categories:

A - antibody  
 A<sub>p</sub> - protein  
 A<sub>g</sub> - globulin  
 A<sub>q</sub> - plaque  
 A<sub>r</sub> - rosette  
 A<sup>G</sup> - G is referential  
 D<sub>r</sub> - RNA  
 D<sub>d</sub> - DNA

### G - antigen

C - cell  
 C<sub>l</sub> - lymphoid cell  
 C<sub>y</sub> - lymphocyte  
 C<sub>z</sub> - plasma cell  
 C<sup>γ</sup> - cell family  
 C<sub>r</sub> - reticulum cell  
 C<sub>y</sub><sup>b</sup> - lymphoblast  
 C<sub>z</sub><sup>b</sup> - plasmablast  
 C<sub>y</sub><sup>s</sup> - small lymphocyte  
 C<sub>z</sub><sup>m</sup> - immature plasma cell  
 C<sub>b</sub> - hemocytoblast

T - tissue  
 T<sub>n</sub> - lymph node  
 T<sub>m</sub> - Malpighian bodies  
 T<sub>x</sub> - cortex of lymph node  
 T<sub>u</sub> - medullary area of lymph node  
 T<sub>l</sub> - lymph; lymph plasma  
 T<sub>l</sub>' - lymphatic system  
 T<sub>l</sub>' - lymphatic capillaries  
 T<sub>l</sub>" - interstitial fluid  
 T<sub>s</sub> - spleen  
 T<sub>d</sub> - red pulp of spleen  
 T<sub>r</sub> - white pulp of spleen; follicular tissue  
 T<sub>b</sub> - blood; serum  
 T<sub>t</sub> - thymus  
 T<sub>k</sub> - adipose tissue of renal sinus  
 T<sub>p</sub> - retroperitoneal adipose tissue  
 T<sub>h</sub> - thoracic duct  
 T<sub>v</sub> - liver  
 T<sub>c</sub> - muscle  
 T<sub>o</sub> - bone marrow  
 T<sup>c</sup> - packed cells  
 T<sub>n</sub><sup>x</sup> - lymph node extract  
 T<sub>s</sub><sup>s</sup> - splenic cell suspension  
 T<sub>s</sub><sup>u</sup> - splenic tissue culture fluid  
 T<sup>B</sup> - B is referential  
 B - animal, body part or region  
 S - ultrastructure of cell  
 S<sub>n</sub> - nucleus  
 S<sub>c</sub> - cytoplasm  
 T<sub>l</sub><sup>f</sup> - efferent lymph  
 T<sub>l</sub><sup>γ</sup> - afferent lymph  
 S<sub>r</sub><sup>h</sup> - channels of ER  
 S<sub>r</sub><sup>s</sup> - cisternae of ergastoplasm  
 S<sub>r</sub><sup>γ</sup> - parallel or lamellar ER structures

# Harris relations ("action types")

## II. Verb (main operator) categories:

V operators selecting A —  $V_t$  - store ←  
S/C/T

$V_i$  - present in; contain ←

$V_u^f$  - move from

$V_p$  - produce ← ; synthesize

$V_u^t$  - move to

$V_s$  - secrete ←

### W (histological) operators

i) selecting C — T, S — C, S — SC, T — T, T — B

$W_i$  - present in; contain ←

ii) selecting C —, SC —, T —

$W_a$  - active; reaction

$W_c$  - change; develop (intransitive)

$W_d$  - disintegrate

$W_e$  - eccentric (of  $S_n$ )

$W_{e\sim}$  - intact; round (of  $S_n$ )

$W_f$  - inflamed (of  $T_n$ )

$W_g$  - large; enlarged

$W_l$  - procedural, operational terms

$W_m$  - mature; distinct

$W_{m\sim}$  - immature; primitive

$W_n$  - fine and granular (of  $S_c$ )

$W_o$  - mitosis (of C)

$W_p$  - proliferation

$W_r$  - rough (of  $S_r$ )

$W_s$  - stained; pyroniniphilic;  
basophilic; fluorescent

$W_w$  - widened (of  $S_r$  or  $S_c$ )

$W_{w\sim}$  - flattened; narrow (of  $S_r$  or  $S_c$ )

$W_y$  - parallel; lamellar (of  $S_r$ )

$W_{y\sim}$  - random orientation (of  $S_r$ )

# From language to formulae via sublanguage

Table 1  
Formulaic representation of sentences

<p>It seems clear from all the evidence that the cells responsible for the synthesis of antibody shortly after the injection of a second antigenic stimulus are members of a family which arise from some undifferentiated precursor as the direct result of the stimulus.</p>	<p>It seems clear from all the evidence that the cells   are   members of a family WH     antigen   the injection of the second stimulus of    shortly after    antibody   (are) responsible for the synthesis of   (cells) ← which     the stimulus    as the direct result of    (Members of a family)   arise from   some undifferentiated precursor</p>	<p><b>M</b> <math>C^w Y C^w</math> <math>G.J^2: A V_p C</math> <math>G.J^2: C^l Y_c^t C_b</math></p>
<p>The first cells which demonstrably contained antibody and can therefore be assigned to this family are large cells with a thin rim of basophilic cytoplasm and large nuclei whose appearance is indistinguishable from that of other primitive hematogenous cells.</p>	<p>The cells   are   large cells Which     (antigenic stimulus)   (the second injection of)    first (after)   antibody   demonstrably contain   (cells) ← and therefore (which)     (cells)   can be assigned to   this family WH     (large cells) with a thin rim of cytoplasm (which)   (is) basophilic WH     (large cells with) nuclei (which)   (are) large whose     (large cells')   appearance is indistinguishable from that of   other primitive hematogenous cells</p>	<p><math>C^w Y C^w</math> <math>G.J^2: A V^l C</math> <math>C Y C^l</math> <math>C^2 S_b W_s</math> <math>C^2 S_b W_g</math> <math>C^2 Y C_b</math></p>
<p>During the 2 or 3 days after their first appearance they multiply, synthesize antibodies specific for the antigen which stimulated their development, and differentiate through immature to mature plasma cells.</p>	<p>the large cells   multiply,     (antigen)   (was twice injected)    WH ←    antibody specific for the antigen   synthesize   (the large cells) ← which     (antigen)    stimulated    the large cells'   development, and     (the large cells')   differentiate   through immature (plasma cells)   to mature plasma cells during the 2 or 3 days after     (antigenic stimulus)   (a second injection of)    (at a time which was) first (after)    the large cells'   appearance</p>	<p><math>C^2 W_p</math> <math>G^w J^2: A^G V_p C^2</math> <math>G: C^2 W_p</math> <math>C^2 Y_c^t C_z^m C_z^m</math> <math>G.J2: C^2 W_i</math></p>

The middle column is a grammatical transform of the left column. Brackets enclose elementary sublanguage sentences. Material between brackets is the sublanguage conjunction marked by colon. Material before a bracket is a general conjunction (not shown in formulas) to the preceding sentence; **WH** indicates a secondary sentence which has become relative clause or modifier. Vertical bars inside brackets separate the subject, verb, and object. Parentheses indicate zeroed material. ← indicates that the preceding material is to be read in English in reverse order; forward-readable transforms exist but are more complex. The right column gives the formulaic representation of the middle column, obtained directly by writing a sublanguage symbol for each segment between bars or brackets. Superscript **w** on a host letter indicates that the host is carrying a modifier which appears as a secondary sentence, below, introduced by **WH**. Other superscripts indicate a modifier that is written together with the host. Subscripts indicate subclasses of the class marked by the host letter. The sentences are from E.H. Leduc, A.H. Coons, J.M. Connolly, J. Exp. Med. 102, 66 Par. 4 sentences 1-3 (1955) and the analysis is given on pp. 360-361 of Harris, Gottfried, Ryckman, et al. op. cit.

# References

- 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, 1-2, 85-99.
- 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.
- Darányi, S. 2000. Before language: metaphor and metonymy in chemical reactions. *Semiotica* 130, 3-4, 217-241.
- Petitot, J. 2001. A Morphodynamical Schematization of the Canonical Formula for Myths. In Maranda, P. Ed. *The double twist: from ethnography to morphodynamics*. University of Toronto Press, Toronto. 267-311.
- Taira, R.K. *et al.* 2007. A Field Theoretical Approach to Medical Natural Language Processing. *IEEE Transactions on Information Technology in Biomedicine* 11, 4, 364-375.
- Harris, Z. (1988). *Language and Information*. Columbia University Press, New York.