Proceedings of the First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts

Satellite of the Supporting the Digital Humanities conference (SDH-2010)

Edited by Sándor Darányi and Piroska Lendvai

21 October 2010, Vienna, Austria

ALL THE PAPERS AND POSTERS PRESENTED HAVE BEEN SUBJECTED TO AND HAVE PASSED A PEER REVIEWING PROCESS

The AMICUS research networking project is funded by
The Netherlands Organisation for Scientific Research (NWO), The Hague,
under project nr. 236-89-003.

Published by

University of Szeged, Faculty of Arts, Department of Library and Human Information Science
Szeged, Hungary

Publication design: Dániel Mariánovich
AMICUS logo: Mihály Minkó

ISBN 978-963-306-069-8

CONTENTS

Foreword

In cultural heritage objects, digitized or not, content indicators occurring on higher than word level are often called motifs or their equivalent. Their recognition for document classification and retrieval is largely unresolved. Work on identifying rhetorical, narrative and persuasive elements in scientific texts has been progressing, in several, but largely unconnected tracks. The AMICUS project[1] (running between 2009 and 2012) set out to test a possible way to resolve these issues, starting with the identification of Proppian functions in folk tale corpora and adapting the solution to the identification of tale motifs or their functional counterparts.

AMICUS has devoted its first project year to listing the corpora, tools, methods and contacts available to address these issues. The initiators of the project have identified a common need in the processing of texts from both the cultural heritage (CH) and scientific communication (SC) domains: to perform automated, large-scale higher-order text analytics, i.e., to reach an advanced level of text understanding so that structured knowledge can be extracted from unstructured text. The four research groups propose to tackle an important aspect of this complex issue by investigating how linguistic elements convey motifs in texts from the CH and the SC domains. Our shared working hypothesis is that 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 a 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. In bioinformatics, the motif of a gene array study forms the mold of countless articles. In the newly developed area of web sciences, a common rhetorical motif is to refer to the threats of information overload on people. In all of these different fields, insiders are familiar with these motifs, while outsiders are not; motifs constitute a kind of high-level jargon. The modeling of motifs (and in extenso the automatic detection of motifs) is an important, and yet currently missing aspect of the analysis of CH and SC texts beyond the sentence level, on which all of the four teams are focusing their work. Interestingly, as a unit of higher-order content, the concept of motif is widespread in the CH domain (especially in literary and folkloristic genres), but not explicit in the SC domain.

AMICUS aims to establish a scholarly research network with complementary expertise in text analytics, scholarly publishing, network studies and user studies, currently pursuing their own approaches towards the central theme of the modeling of structured knowledge (i.e. networks of knowledge) as seen against a background of different professional user communities with discipline-specific information needs. The current volume holds the papers and posters presented at our first workshop that solicited the presentation of relevant research, emerging from or pertaining to the AMICUS initiative.

The 1st International AMICUS Workshop is a one-day meeting that takes place on the 21 October in Vienna. The workshop is an official satellite of the Supporting the Digital Humanities conference (SDH-2010). The AMICUS workshop aims to overview methods and infrastructure related to motifs, and to facilitate community interaction and cross-fertilisation of research. We would like to express our thanks

[1] http://amicus.uvt.nl

to the Program Committee for providing valuable comments and suggestions on the publications in the poster session:

Kate Byrne, School of Informatics, University of Edinburgh
Thierry Declerck, Language Technology Lab, DFKI, Saarbrücken
Zoltán Hermann, Károli Gáspár University, Budapest
Mihály Hoppál, Ethnographical Research Institute, Budapest
Artem Kozmin, Russian State University for the Humanities, Center of Typological and Semiotics Folklore Studies, Moscow
Ágnes Sándor, Xerox Research Centre Europe, Grenoble
Caroline Sporleder, Cluster of Excellence / Computational Linguistics, Saarland University
Pirkko Suihkonen, Department of Linguistics, University of Helsinki
Vilmos Voigt, Department of Folklore, Loránd Eötvös University, Budapest
Anita de Waard, Elsevier Labs, Burlington & Utrecht University, Utrecht

Sándor Darányi and Piroska Lendvai
Editors

PAPERS READ

Morphology and Morphogenesis of Folktales and Myths

Pierre Maranda
Department of Anthropology
Université Laval
Québec, Canada
pmaranda@videotron.ca

ABSTRACT

This paper draws on a set of previous ones that I develop below. It consists of three parts. Part One begins with (1.1) a prolegomenon on trust, the result of an implicit calculus of transition probabilities, followed by (1.2) a summary consideration of the dynamics of memory and imagination. Part Two deals with analytic units in the processing of folk narratives. Part Three bears on morphology and morphogenesis and presents DiscAn, a computer system that may generate dynamic cognitive maps as a morphogenetic approach..

1. PART ONE

1.1 Prolegomenon: Trust and *p*

First, a trivial but nonetheless basic statement: Trust is the foundation of all productive human relations. It underlies necessarily friendship, exchange, reciprocity. Societies collapse when trust breaks down. Trust in their parents is a given without which children cannot evolve as adept family - and consequently society - members. Trust is slowly built between people that become friends. It is often a long process of trials and sometimes errors, an empirical, pragmatic test of the extent to which one may share one's deepest feelings, thoughts, concerns with another person without fearing reticence, judgment or betrayal. Trust lies at the root of harmonious communal life. In Lévi-Strauss' words, trade or war are the basic vectors that structure relationships between societies. But there is no trade without trust. And, I repeat, trust can emerge and be consolidated only through experience, i.e., both linguistic and extra-linguistic pragmatics.

Be it one's trust in one's older sibling, in one's school teacher or in one's banker or investment broker, one builds it gradually through a subliminal statistical process geared to insuring stability and safety. Actually self-amplifying loops consolidate both trust and distrust (cf. Bayesian Decision Theory - the Bayesian models

First International AMICUS Workshop, October 21, 2010, Vienna, Austria.

of probabilistic coherence and probabilistic inference rules that regulate the levels of confidence and surprise, doubt or disconcertment [22, 31, 33, 52]). What I have called "nets of expectancies of events" [31] provides the basis for an intuitive probability coefficient that affects the degree of plausibility of beliefs. Here, "events" must be understood as being intellectual, pragmatic, emotional, whatever may impinge on the life of a human being. For example, the contemplation by a person of different scripts of eventual decisions and his or her musing ("data mining") about their bearings, rest on scanning expectancies to which are implicitly or explicitly assigned degrees of likelihood and acceptability in given contexts. And different scenarios can be generated by varying purposefully plausible contexts, an operation that will affect expectancies differentially. As I said, that information processing proceeds from past personal knowledge (based on experience) or knowledge acquired from external sources, and it will affect expectancies differentially - and consequently, decisions and other forms of behavior as well. Such mental operations are not alien to some sort of subliminal factor analyses and such statistics as multiple regression where *Y* varies according to the *Y*s taken as inputs, the *X*s being akin to Memory Organization packages (MOps), Memory Organization Processes (MOPs), and Memory Organization Nets (MONs) whereas the *Y*s to ISPs - Imagination Structuring Processes (more on that below).

This is akin to the domain of probabilistic semantic grammars and is more or less systematic practice in some psychographic research and in the strategies of advertising. It is also commonly seen in current verbal interactions. For example, Person A will avoid talking about a given subject to Person B in a given context apprehending that Person B's thought processes might take a turn unattractive to A and lead B to ask specific questions that would increase B's knowledge to the detriment of A. Conversely, A may decide to talk about a given subject so that B will most likely ask specific questions and so that that A will be in a position to alter B's knowledge by his answers. In such cases, the psychographer, the ad-man or Person A all presume to have a relatively adequate knowledge of their targets' appetencies and of the action to be taken to manipulate them. Along the same lines a writer moves on the thin line between what he considers to be a bold statement and what he thinks his readership will accept -- as pointed out by T.S. Eliot in *Notes Towards the Definition of Culture* (chapter "Tradition and the Individual Talent").

1.2 Memory and Imagination

I take the tack that the basic components of cultural universes consist of MOps and ISPs. I have defined those concepts in previous publications [27, 31, 34) of which I rephrase some passages.

1.2.1 Memory

I have suggested that Schank's definition of Memory Organization Packages (MOPs) be renamed hereafter "MOps", because I use MOPs to mean Memory Organization Processes. MOps [25-28, 31, 36, 47-49] should be improved on three scores. Actually the concept of MOps aims at identifying clusters of meanings stored in memory and stemming from personal experiences rather than semantic categories. As such it is context-dependent. As a matter of fact the term MOp re-labels what Hebb called "cell assemblies" in his classic theory of memory (more on that below). MOps constitute sorts of personal vital databases, accretions of past experiences, some of which remain deeply buried in the subconscious, others more or less easily - and more or less serenely - retrieved, or surging in triggering contexts (Marcel Proust's novels masterfully provide and analyze examples of that kind of MOps spurts). Such "packages" are related to what Lévi-Strauss calls "bundles of relations" [19, see Ch. XI and his voluminous *Mythologiques*]. But the MOps approach must be taken three steps further.

(1) To the notion of "packages" should be added that of "processes", MOPs, to better take into account memory dynamics. MOps would remain as paradigmatic units with limited potency, akin to "slots" in Frame Theory [38] whereas MOPs are associative structures that connect clusters of MOps with comput-able probabilities.

(2) The notion of "Memory Organization Nets" (MONs) develops in turn that of "processes" - MOPs by adding to it a coefficient of "appetencies". MONs denote that MOPs do not stand alone but are linked, on that level also with computable transition probabilities, in more or less coherent clusters through appetencies (and thus are of another order than Schank's and other models of graphs of conceptual dependency). "Appetency", a term proposed by Harary in Digraph Theory and developed by Kamp and Hasler [11], means an empirically grounded -- i.e., pragmatic – tendency for an image or concept to "have an appetite for", to attract, or to be attracted by other images or concepts in its gravitational orbit. The concepts of "attractor" and "attraction basins" that I will define below develop and implement *appetency theory* [11, 34, 37] and *resonance theory* [36]. MONs would also take into account, and be related to, "scripts" stored in memory, an example of which is my representation of Proppian functions (Figure 1, below).

(3) Associations between semantically high-loaded terms repeated from generation to generation – like, indeed, between neurones according to *Hebbian learning* – consolidate their directed linkages, i.e., appetencies (for example, mussels and French-fries for Belgians or burka and women for some Muslims, etc.). Some associations have put on so much semantic weight – so much potency – that they eventually acquire a stereotypical force with robust identity functions that "are gut-wrenching". Such appetencies generate "isosemies", that is, ways of giving the same connotations to both terms and behaviors. They polarise either consent and approval, or rejections which can lead to ostracism, even conflicts[1]. Representations and cognitive mappings find in this approach semantic resonances (*Category ART – Adaptive Resonance Theory*; [13, 30, 33, 36, 56-57]) that work in loops and cycles in accordance with the exercise of our memory, our imagination, our mind.

1.2.2 Imagination

Imagination should be considered a significant vector in that approach because MOPs and MONs present only a limited and partial view of mental processes. Indeed, MOPs provide basic data on which imagination works at the same time as it – imagination – feeds-back on those processes for revising MONs to reconfigure them. We must therefore take into account what I call "Imagination Structuring Processes" (ISPs) when investigating how the mind works, both on the personal and cultural planes. Actually, both memory and imagination pertain to "social facts" in that they depend on representation systems, the elements of which have been differentially internalized by the members of a society. In Strecker, Meyer and Tyler's words about Rhetoric Culture [51, see p. 8 – emphasis added], "Rhetoric culture […] seeks to give an accounting of the *imaginative resources that ground our approximations and make our openings and closings*."

The management of MOps is relatively selective. Depending on MOPs reorganized through MONs, i.e. MONs → MOPs → MOps -, some MOps are sent right away to the mental trash can – "Forget it!" - "I don't want to be remembered of that!" But some MOps are relatively un-erasable: they keep haunting personal or collective memories, e.g., the death of one's parent, that of a dear friend, McCarthyism in the USA, September 11 2001, etc. The Propp net shows MONs, and the itinerary between nodes depends on the teller's ISPs. Only a very competent teller will be able to navigate through sequencing the less probable nodes and still build a convincing story.

A summary legend to Figure 1: capital letters refer to broad categories of actions, viz., Propp's "functions". In the network, lower case letters and numerical subscripts stand for more specific functions (subsets of the major ones; for a full description and analysis see Maranda [25, 29, 33].

Transition probabilities from one node to another figure at the arrow heads (1st-order Markov chains):

A = lack or mischief of some sort

B = plan to counter lack/mischief

C = start of counteraction

F = acquisition of magical agent by hero

[1] Some of these appetencies remain unidirectional ("unilateral" in terms of oriented graphs). For example, for cultures where we can find the metaphor "this woman is an angel", "woman" is in appetence of "angelic character". But this metaphor is not reversible to "this angel is a woman". It is therefore a unilateral appetence of "woman" to "angel". A bidirectional appetence ("bidirectional" in terms of oriented graphs), however, is reciprocal, as in the reversible metaphor "youth is the morning of life" compared to "the day is still young".

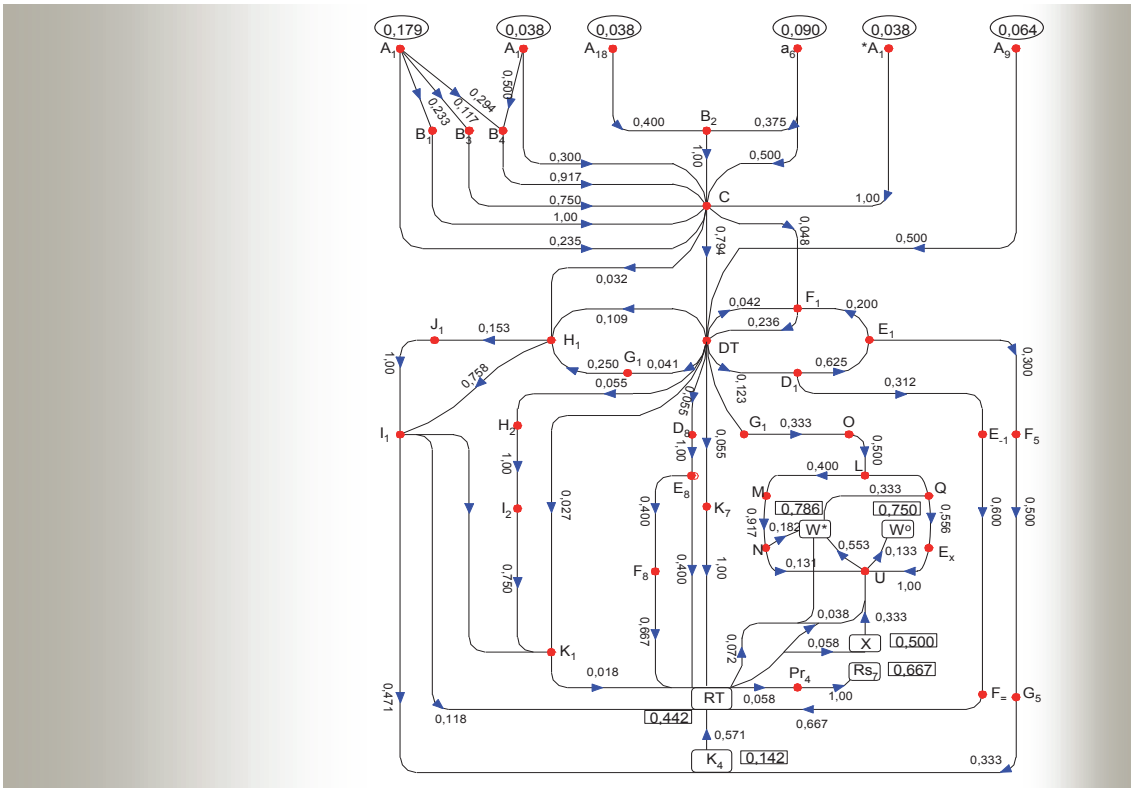

Figure 1. A probabilistic and reticular representation of Propp's algebraic morphology of the folktale[2]

DT = hero departs on mission

H = hero fights villain

J = victory of hero over villain

K = lack liquidated

N = completion of difficult task

RT = return of hero

W* = wedding and accession to the throne

Note that C acts as a condensation node and DT as a diffraction one. A skilled teller will know how to navigate along more or less expected paths to bring his story to an expected or surprising end. An unskilled one will take a direct line to RT and come up with a simple-minded plot the close of which will leave his audience disgruntled.

Each Proppian "function" is actually a culturally pre-stressed Memory Organization Package (MOp) - a sort of *Völkergedank* (Bastian's concept, see below), and the performance of a teller along the Russian fairytale generative network depends on his Imagination Structuring Processes (ISPs).

To sum up, ISPs are based on MOPs and MONs and they review (or even revise) them retrospectively, use them when scanning plausible eventualities before making a decision, or when inventing compensatory and many other kinds of scripts. The arrays of representations MONs make available to their carriers hang together in more or less loose clusters. Through ISPs MOPS can be disconnected and reconnected: the meanings of their constituents can be revamped through creative processes such as metaphors, a prominent Imagination Structuring Process [5, 22-23, 32, 35]. These ISPs are function of pragmatics, hence of percepts and connotations as well as denotations:

((((MOps)*((MOPs))*(MONs)))*(ISPs))).

Now, can those concepts help in the analysis of narratives? I give a positive answer to that question: they can be useful if we operationalize them as "attractors" and "attraction basins". But before that, a brief discussion of analytic units used in narrative research will be appropriate. (Propp [43, p. 176] insisted on the importance

[2] Cf. [7] on Propp's algebraic model.

of situating any folktale in the social context "where it lives". And since action/experience impinges heavily on both memory and imagination, pragmatics comes into play to load MOPs and ISPs with more or less heavy connotations.)

2. PART TWO: ANALYTIC UNITS

2.1 A Summary Review of Analytic Units

Obviously, the accuracy and validity of script analyses depend on the operationality of analytical units. Here a summary review of some such units may be pertinent (see also Darányi's paper in this volume).

We may begin with Adolf Bastian, a foremost ethnographer who did exemplary field work in exotic societies in the 19th century and who founded the Berlin Anthropological Society in 1869. In his books, among them in *Das Beständige in den Menschenrassen und die Spielweite ihrer Veränderlichkeit* [1], he proposed the concepts of *Elementargedanken* ('elementary thoughts') that find their expression as *Völkergedanken* ('ethnic ideas' or 'folk ideas'). Their inventories would describe the basic components of different cultures' worldviews. Such folk ideas include fables, legends, myths, proverbs, sayings, tales, etc.

A few years later, in 1873 Veselovskij [54] used motifs and themes as analytic units. In Propp's words [42, see p. 12]: "Veselovskij means by themes a complex of motifs. 'A theme is a series of motifs. A motif develops into a theme', and 'By theme I mean a subject in which various situations, that is, motifs, move in and out'". Propp [op. cit. p. 13] then discusses Bédier's analytic unit, viz., "elements". As for Aarne's index , it introduces the term "type" to supersede "themes". In keeping with Aarne's terminology, Stith Thompson resorted to motifs in his *Motif Index* and, after Aarne, to types in *The Types of the Folktales*. It is needless to recall here Propp's sharp criticism of the motif, theme, element and type as operational units because of the lack of precision of those terms [42, see Ch. I]. He proposed to use instead the concept of "function": "Function is understood as an act of a character, defined from the point of view of its significance for the course of the action [42, p. 21], i.e., a function is a syntagmatic vector: "Functions constitute the fundamental compon-ent of a tale." He also introduces larger analytical units, namely "spheres of action" [op. cit. p.79] and "moves" [op. cit. p. 92].

On the other hand, in linguistics C.F. Hockett [9] proposed a connectionist "Resonance Theory of Morphology". It implies "recognition units", that postulate what I call a MON (Memory Organization Net). Such recognition units consist of clusters of meaning recognition (cf. "attraction basins" below). Of different sizes, they trigger reverberations in internalized systems in a way somewhat germane to Category Adaptive Resonance Theory.

And more recently, let's return to folkloristics with automated motif identification in folklore text corpora [55, p. 126]. The authors state:

> "We note in passing that the concept of a motif goes back to classics of folklore and literary analysis (see the summary by Würzbach 1998)**.** In our interpretation, a motif is a second-level aggregate of some first-level content criteria, e.g. the motif "Unpromising hero" (see Meletinsky 1958) is a compilation of 'hero', 'son', 'youngest', and the like. In other words, a motif is a broad concept related to those narrower terms which define it.
>
> In library and information science however, it is an established practice to express such broader concepts from more detailed content criteria by automated classification, for example by singular value decomposition (SVD) (Deerwester *et al*. 1990), so that, as a prelude to information retrieval, the results can be used for an advanced type of indexing called latent semantic indexing (LSI) (Lochbaum & Streeter 1989). In short, we wanted to apply LSI in the domain of folklore."

More recently Jean Petitot, in his remarkable paper on Lévi-Strauss's canonical formula for the analysis of myths [41, p. 272], discusses analytical units as follows:

> "(i) *terms*, at the *syntagmatic* level, that is, *actants* in the sense of an actantial syntax (to be distinguished from actors or characters that usually syncretize several actants and support thematic roles),
>
> (ii) semantic *functions*, at the *paradigmatic* level, that depend on codes (in Lévi-Strauss' sense) belonging to deep structures: they are values categorizing the continuous substratum of paradigms into discrete units. […]
>
> The problem, which is an extremely difficult one, lies in holding together the paradigmatic (semantic "codes") and the syntagmatic (actantial interactions) levels. A basic thesis of semio-narrative structuralism is that paradigmatic semantic relations can only be implemented through actantial syntagmatics. Semantic values are "confined", "invested" in the actants and circulate through their interactions.
>
> Three theoreticians have played a crucial role in elucidating these relationships: V. Propp, C. Lévi-Strauss and A.J. Greimas[3]. With his grammar of functions, Propp overly dissociated the narratively dominant actantial syntax from the semantic content. All too often he reduced the latter to simple thematic roles. On the other hand, *by focusing dually on the paradigmatic axis and its projection on the syntagmatic axis, Claude Lévi-Strauss somewhat underestimated the problem of actantial syntax. The synthesis was achieved by Greimassian theory which showed in detail how actantial syntax could handle logico-combinatorial operations on paradigmatic value* [emphasis added]."

But when it comes down to actual analyses of folk narratives, does Greimassian theory really help in defining operational units? In that respect we may want to consider Lévi-Strauss' concept of "mythemes" [19, p. 232.; 21, pp. 168-172]. As culturally pre-stressed components of myths, they are "pregnant" (in René Thom's [53] terminology) in that they contain the seeds of syntagmatic dynamics that will configure narrative structures. This analytic unit might have marked the switch from morphology to morphogenesis but Lévi-Strauss' concept does not meet the requirements specified by Propp, after Veselovskij, viz,

[3]These works are masterpieces of structural epistemology and methodology beginning with the works of Saussure, the Prague Circle, of Husserl's third *Logical Investigation*, Jakobson, Hjelmslev and Brøndal.

that it must be an indivisible narrative unit [42, p. 12]. Mythemes take indeed the form of elementary propositions. Lévi-Strauss wrote [19, p. 233] that his method consists of "breaking down its [a myth's] story into the shortest possible *sentences* [emphasis added]. Thus: "a function is, at a given time, linked to a given subject. Or, to put it otherwise, each gross constituent unit will consist of a *relation.*" And […] "The true constituent units of a myth are not the isolated relations but *bundles of such relations* (cf. MONs) and it is only as bundles that these relations can be put to use and combined so as to produce a meaning". As for "gross constituent units", Darányi (this volume) writes "Related research on the canonical formula of myth (…) shows that at least 256 classes (subspaces) [cf. 45] can be filtered out based on canonical variables and value configurations leading to group symmetries and symmetry breaking (…). Such symmetries constitute phases in Markov chain based patterns (…)."

As an attempt to tackle, in Petitot's words, the "extremely difficult" problem that "lies in holding together the paradigmatic (semantic 'codes') and the syntagmatic (actantial interactions) levels" I will now introduce the notion of attractor and attraction basins as units my team and I use in our analyses of "mentifacts" [12, see Ch. 1].

2.2 Attractors and Attraction Basins

I have written in connection with the structuring of www.oceanie.org [34, English translation in print] as follows:

> "Positioning ourselves beyond linearity we address social facts not as coherent and well articulated structures, but rather as semiospheres [6, 16, 17][4], i.e., as constellations of representations and actions. Relations reverberating onto one another structure the universe of meaning which gives to those sharing it the impression of understanding each other when they communicate[5]. We operationalise the concept of semiosphere by using those of attractors and attraction basins. The latter spread and radiate around the former, "words full of meaning" that "are gut-wrenching", as Oceanians tell us, to whom www.oceanie.org provides a partial overview of their socio-cosmic universe.[6]
>
> Each society and each culture has a repertoire of words full of meaning (*Völkergedanken*) or "carrier categories" [15] which configure them and which they configure through feedback, like women, men, gods, work, sex, etc. The semiotician François Rastier spots about 350 of them in industrial cultures [45][7]. The concepts of attractor and basin can be related to "semantic field of keywords" as in [55].

I will now quote Ott's [39] mathematical definition of "attractor" and "attraction basin" (I have slightly modified the syntax of the first sentence):

> Roughly speaking, an *attractor* of a *dynamical system* is a subset of the state space to which tend, as time increases, orbits originating from typical initial conditions. It is very common for dynamical systems to have more than one attractor. For each such attractor, its *basin of attraction* is the set of initial conditions leading to long-time behavior that approaches that attractor. Thus the qualitative behavior of the long-time motion of a given system can be fundamentally different depending on which basin of attraction the initial condition lies in (e.g., attractors can correspond to *periodic*, *quasiperiodic* or *chaotic* behaviors of different types). Regarding a basin of attraction as a region in the state space, it has been found that the basic topological structure of such regions can vary greatly from system to system.

And, because we have not derived our concept of attractor from Chaos Theory but from the neurosciences, it is pertinent to quote Petitot [41, p. 274]: "Such models where the categorization of a continuous substratum space into sub-domains (values defined by *reciprocal determination* [emphasis added]) is generated by a family of generating potentials, have become widespread in contemporary cognitive sciences. If the categorization process is implemented in a network of formal neurons, the generating potentials become true potentials in the physical sense of the term, i.e., "energy" functions whose minima[8] determine the terms of the categorization."

As for our inspiration by neurosciences [11, 37], I have written elsewhere the following [34 - English version in print].

Developments of the law of Hebb [8] on cellular assemblies opened new research horizons on which artificial intelligence still draws and which motivated our approach to the construction of our notions of "attractors" and of "attraction basins". "A cellular assembly consists of a group of cells connected in a reverberation circuit that is a complex and interconnected loop. When an external trigger excites the cells of the loop, they are mutually excited and the whole circuit goes into reverberation" [2]. For the neuropsychologist Donald Oldings Hebb, the strength of connections (synapses) between the neurons can vary

[4] The semotician Jacques Fontanille [6, p. 296] summarizes Lotman's [16, 17] concept of semiosphere as follows: "The semiosphere is the domain in which the subjects of a culture experience meaning. The semiotic experience in the semiosphere precedes, according to Lotman, the production of speech, because it is one of its conditions. A semiosphere is primarily the domain that allows a culture to define and position itself so as to be able to interact with other cultures ". See also the socio-cosmic approach of the *Erasme* group [3].

[5] The feeling of mutual understanding between speakers requires their semiospheres to overlap, whether they are consonant (for example, political correctness) or dissonant. Hence the recourse to the Theory of Resonance and that of circuits of reverberation which, depending on the degree of sharing of sub-semiospheres, generate friendly communities, more or less closed groups or ghettoes (consonances) and also antagonisms (dissonances) [30, 33, 34, 57].

[6] Stimulated or at least intrigued by our graphs, they use them as sources of inspiration for writings in which they reactivate their cultural roots.

[7] See also [44].

[8] In physics, to minimize energy is the basic variational principle.

diachronically. The synergistic activity of an emitting neuron (pre-synaptic) and of a receiving neuron (post-synaptic) produces a strengthening of the synaptic connection that is designated by the term *Hebbian learning*. Furthermore, the inertia of one of the two neurons causes a weakening of the connection. Hebbian learning also applies to large groups of neurons in the whole of which loops are formed. If a neuron *A* excites a neuron *B* which, in turn, excites a neuron *C* which returns to *A*, then the synapses that connect these neurons become stronger and increase the probability of reiterations of this loop. This positive feedback reinforces the self-stabilisation of neural networks [10, 11, 40].
In terms of representations, this " looping " (or " cycles " in Graph Theory) consolidates the associations of ideas, behaviours and strategies tested in the context of pragmatics [31], such as "empirical deductions" that Lévi-Strauss noticed in the structuring of cultural universes [20]. The same holds as regards policies of matrimonial alliances: systems of generalised exchange expand the economical, social, and political "circuits of reverberation" which the restricted exchange does not allow for [18]. Thus these "cycles" generate semiospheres of variable amplitudes in all societies. Take for example the case of the "priesthood" semiosphere: it includes women in the Anglican church while it does not in the Catholic church; the amplitude of the former's sacerdotal semiosphere is broader than the latter's[9]. The approach we defined can model the connectivity and recursivity that generate the stability in distributive networks in such semiospheres. In other words, models developed in neurosciences and in the Theory of Resonance describe the consolidation of memory and cognitive structures, some relatively inert, others relatively flexible and innovative. Furthermore models of the same type can account for not only neural dynamics but also, extrapolated, for the dynamics of societies: of their representations of the world and of themselves, plus they can also map out vectors of disturbance stemming from discriminatory pragmatics (for example, with regard to women).

2.2.1 Implementation: "Attractors" and "Attraction Basins"

I will now illustrate our use of attractors and attraction basins. I excerpt from www.oceanie,org the attractor 'Ancestors'. to represent the traditional knowledge of Oceanians. The attractors we have selected constitute labile Memory Organization Nets whose dynamics vary from one part of the Pacific to another. The attractors and their basins make it possible to map what recent research calls socio-cosmic bases for geopolitics, on both large and small scales [3]. For us, a cluster of meaning takes shape around a central seme, to which we refer by the term "attractor". As we see below in Figure 3, an "ancestors" attractor occupies the centre of a cluster of meanings. The ideas that revolve around it emanate from it at the same time that they bounce back to it (reflecting barriers in terms of Markov chains [36], "circuits of reverberation" in neurosciences and in Resonance Theory [33, 36]). The basin of this attractor thus unfolds as "nodes" which emanate from it and which, in an inverse motion, consolidate the basin through their convergence on its polysemy. This modelling yields a representation of intersections of vectorial planes within a "culture" where repercussions of components impact each other with various amplitudes that define the scope of the basin that the attractor generates and is generated by[10]. The configurations of such universes of meanings vary of course spatio-temporally.

The hypertext and hypermedia modes also enable users to navigate by diverse and flexible means, making cross-references and associations that give access to information in the form of photos, digital videos, virtual objects that can be manipulated, and sound recordings.

If we click on "Ancestors" in Figure 2, the "attraction basin" associated with this attractor can be seen on the three concentric circles around it (Figure 3). The position of the satellite nodes, which represent other concrete realities in Oceania, corresponds to their semantic distance from the attractor at the centre. The diffraction of some nodes over several vectors illustrates the polysemic scope of cultural symbols.

In Figure 3, we can see that the first circle contains 26 nodes such as Taboos, Masks, Sacrifices, Trances, Cargo Cults, etc; the second and third circles contain fewer nodes - 20 in all. On the second circle we find Dances, Prayers, Funerals and Dreams; and on the third, Chiefs, Costumes, Jewels and others. Some nodes in a basin are connected to an attractor only by relays when others feed to it and are fed by it through direct inputs. For instance there is a direct link connecting "Ancestors" and "Mana" while "Parures" (Ornaments) on the third circle transit through "First fruits" or via "Dances" on the second circle, to make it to the first circle where the immediate associates of " Ancestors " are found.

The length of these "paths" (a term of Graph Theory) – direct or indirect – manifests the degree of semantic proximity between different nodes, displaying their strong connectivity. The set of diffractions and convergences generates the attraction basin of an attractor. And the connected nodes also work as attractors, hence the hypertexts that structure www.oceanie.org.

Oceanians find food for thought and discussion in these representations of connections and the bouncing of nodes on each other. Among them, story tellers as well as writersl use our graphs as sources of inspiration when they produce new narratives.

---

[9] For developments concerning the semiospheres of mentsruation, extra-long penises, widowhood etc., [24, 37, see Ch 1.

[10] We thus reach the concept of transduction of the philosopher Simondon [50, p. 30]: "We mean by transduction a physical, biological, mental, social operation, by which an activity propagates gradually within a domain, by grounding this propagation on a structuring of the domain from place to place: each region of such structures is used in the following region as a principle of constitution ". We shall note the congruity of this theorisation with that of morphogenesis [40], see also the works of François Rastier on semiotics and cognition [45]. As for computer systems moving in some way along the same lines as we do, see *Atlas.ti* (Berlin), *Thinkmap* of *Plumbdesign* (New York) and *Verbatim* [14], which show a similar concern to represent knowledge multidimensionally.

Here, I suggest that the reader click on 'Possession' for videos and also on "Masks" or on "Statues" for 3-D manipulations.

3. PART THREE: FROM MORPHOLOGY TO MORPHOGENESIS?

The preceding section leads to a consideration of what might be a morphogenetic approach to myths and folktales as I have proposed it several years ago in Groningen [31]. I present here a somewhat revised version of some parts of that paper:

1. A myth combines Memory Organization Nets (MONs) and Imagination Structuring Processes (ISPs) through a bricolage of attractors as pre-stressed elements. New combinations reprogram former semiotic processors or alter the weights of semiotic loads, modifying transition probabilities (associative strength);

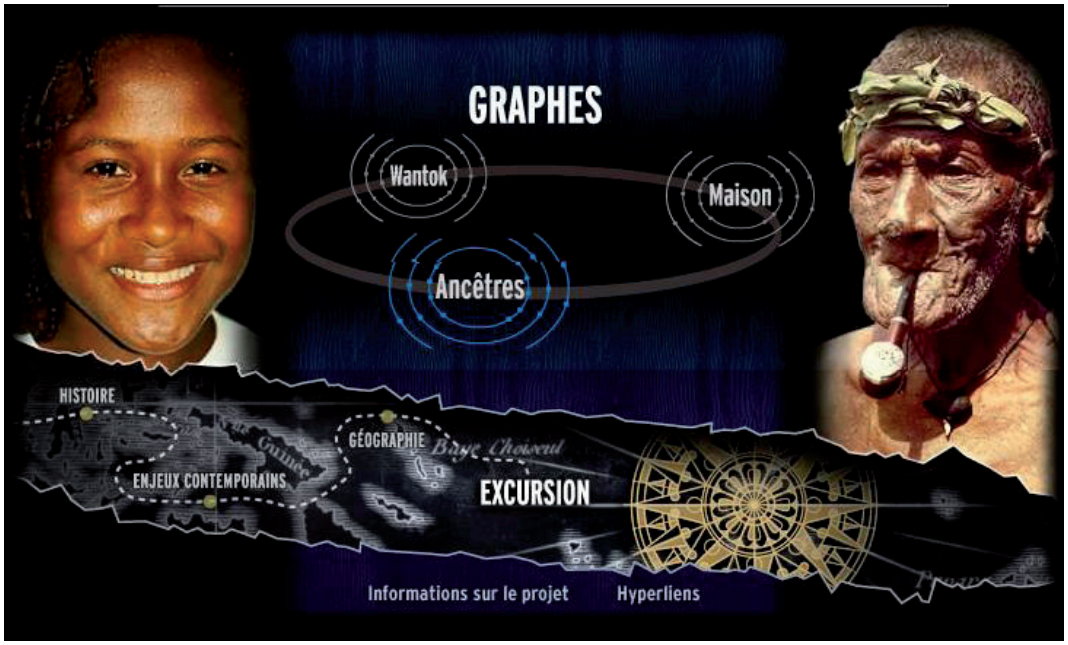

Figure 2. Front page of www.oceanie.org

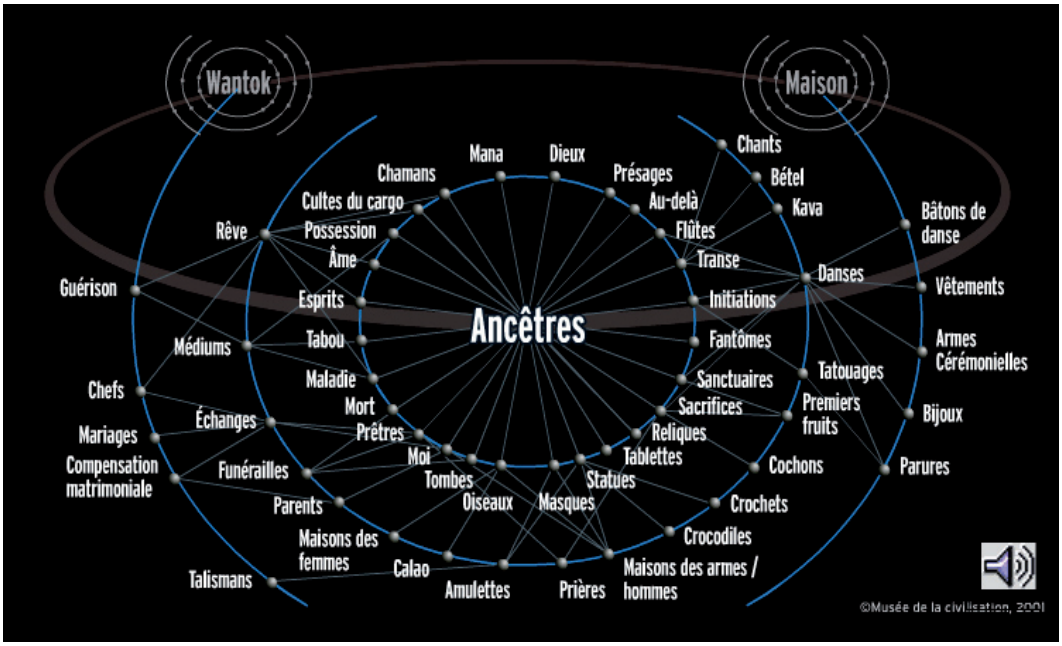

Figure 3. The attractor "Ancestors"

2. MONs and ISPs depend on pragmatics, hence on percepts and connotations;

3. The interplay of MONs and ISPs consolidates or restructures MONs according to transition probabilities;

4. High transition probabilities make for routines (seats of compatibilities, i.e., of inertia); low transition probabilities make for innovation possibilities (breakthroughs); null transition probabilities reflect incompatibilities;

5. Unless modified under stress, routines generate self-amplifying "cycles" in networks and consolidate MONs;

6. Innovations are based on ISPs;

7. Corollary to 1- 6: myths must be flexible or polysemic enough to cope with the interplay of routines and stress in order to allow for innovations, otherwise they collapse;

8. Transition probability computations yield a type of semiotic maps that represent the relative power of myths;

9. Such maps consist of Markovian networks that combine digraphic to probabilistic calculus;

10. DiscAn is a computer system to investigate and map out MONs and ISPs.

3.1 DiscAn: A Computer System for Content and Discourse Analysis

I designed the computer system *DiscAn*: *Discourse Analyser* to implement a dynamic procedure combining the Theory of Oriented Graphs with Markov chains of the first degree [22, 27]. DiscAn is a somewhat archaic computer system in that it still runs only in DOS - a Windows version was begun years ago but was not completed despite demand and a manifestation of interest by Microsoft. I had too much to do with my anthropological field work to carry on its reformatting as well as some improvements. Yet it is still used in different parts of the academic world. Unfortunately the source code has disappeared with my programmer but I can make the DOS version available.

3.1.1 Brief Presentation of DiscAn

DiscAn is a computer system for content and discourse analysis of a particular type. In addition to standard procedures for content analysis of natural languages (which requires ASCII input), DiscAn includes a MARKOVANALYZER for discourse analysis in terms of transition probabilities coupled to a digraphic calculus of node potencies.[11]

3.1.1.1 Thesaurus Building: Tagging a Natural Language Corpus in Terms of MOps as Associative Structures

Such associative sets are mapped onto tags either through listing (bottom-up description) or by rule (top-down taxonomic decisions). DiscAn offers both tagging procedures: through an on-screen interactive mode and through built-in thesauruses. The output is a "normalized" corpus, i.e. one whose lexical diversity has been reduced to semantic categories, i.e., where MOps are indeed indexed as "packages". Here, different types of factor analyses and contingency analysis are most useful tools [4, 46, 55] and the HEROFINDER program I have designed for in-house research some forty years ago. Of course thesauri can be hierarchically structured on as many levels as one wishes to but usually three levels suffice.

3.1.1.2 Definition of ISPs as Script Probabilities

Once the tagging operation is completed, DiscAn processes the normalized data by computing transition probabilities from one tag (node) to another. The algorithm is the one-step Markovian (first-degree chains). Thus the system generates MOPs from sets of MOps through the transition probabilities from MOps to MOps that in turn yield a MON with or without sub-MONs ("cut sets", etc.). The probabilistic linking operation displays all the active paths through the "slots" or MOps, i.e., it shows the "scripts" that have been implemented in the input text. l

3.1.1.3 Computation of MOPs Inner-Degrees (d_-)and Outer-Degrees (d_+) Taking Flows Into Account to Combine Quantitative and Qualitative Analysis

In this phase, DiscAn computes the flows through the net from node to node. This procedure allows for a measure of interplay between MOPs and ISPs, and it integrates to some extent quantitative and qualitative analysis, i.e., flows on the one hand and node input/output ratios on the other. Actually, the transition probabilities from node to node (MOp to other MOps) in a data set are computed concurrently with each MOp's diffraction (outer degree, d_+) and absorption (inner degree, d_-) coefficients. The resulting net (MON) maps a general memory organization system with three gradient facets: (1) inertia nodes, i.e., high- to low-frequency MOps whose d_+/d_- ratios = 1; (2) diffraction (or generating) nodes, i.e., high- to low-frequency MOps whose d_+/d_- ratios > 1; and (3) high- to low-frequency absorption nodes, i.e., MOps whose d_+/d_- ratios < 1. This operation loops back to 3.1.1.1 above, and may lead to a redefinition of lists or rules to specify MOPs and thesaurus revisions or revamping.

3.1.1.4 Dynamic Data Handling Through Digraphic Operations on MOps and MOPs, i.e. Through Script Operations in Which ISPs Act on MONs

To this point, DiscAn provides all the necessary computations to map transition probabilities between MOps, and transforms a set of MOps into a MON (with or without sub-MONs) as stated in 3.1.1.2 and 3.1.1.3. This specifies associative strength between MOps and provides measures of routine weight. High transition probabilities coupled with high frequencies and with d_+/d_- ratios = 1 define strong routines. Such consolidated routines maintain system inertia; they map the fundamental structure of a meta-text. ISPs bear only slightly on such routines which are "taken for granted" and constitute the more or less fuzzy postulates of a knowledge base whose dynamics is one of corroboration.

On the other hand, low transition probabilities coupled with high frequencies and a d_+/d_- ratio < 1 indicate "sinks", i.e., absorption

[11] It would be pertinent here to refer to deterministic or quasi-deterministic stochastic automata.

MOps. Such nodes are semantic dead-ends that may bug down enunciation unless positive regressive imagination processes (RIPs+) can act retroactively on them in order to increase the d_+/d_- ratio.

Then, low transition probabilities coupled with high or low frequencies and a d_+/d_- ratio > 1 define progressive imagination processes (PIPs) that provide for innovations, which is then achieved through a broader relationship system of MOPs, i.e., through an amplified MON whose density becomes higher. New connections will become possible and the whole MON or at least one of its constituent MONs will be revamped – like when a new unifying theory subsumes older and diverging ones.

Reinforcing or weakening d_+/d_- ratios as well as increasing or decreasing flows (frequencies) may result in considerable alterations of a MON. For example, in management as well as in advertising, PIPs are set to scanning MOPs in a MON so that eventual new connections are created and new cognitive habits and/or appetencies are stimulated. Likewise, in psychotherapy, RIPs (Regressive Imagination Processes) are activated so as to provoke abreactions that will neutralize, reorganize or revamp MONs.

Finally, simple algorithms of the same type as those to simulate MON modifications serve to generate new dynamic MONs. For example, the complement of the digraph of a tragedy (Jean Racine's *Andromaque*) through DiscAn generates a comedy structure [29]. In such operations, ISPs act on MONs along "what if" lines. Scripts of different kinds can thus be transformed by inversions, condensations, expansions, derivations, re-combinations and other similar graph-theoretic operations resulting in new knowledge configurations that may even restructure almost completely a knowledge map and its MON base -- as it happens in "scientific revolutions" and other creative texts.

4. CONCLUSION

The question of definition of operational analytical units is, in Petitot's words, "an extremely difficult problem" that "lies in holding together the paradigmatic (semantic "codes") and the syntagmatic (actantial interactions) levels".

Actually, in terms of their constituent units are folk narratives that different from life as paradigmatic sets and syntagms of thoughts and events? Might not paradigmatic sets - attraction basins and their probabilistic vectors - serve as analytic units in both cases? Back to trust: trust rests on inertia, i.e., one must believe that routines will remain sufficiently stable to enable one to lead a life without too much stress, with enough confidence in self and others to dwell in a secure mental and social space. Does not inertia empower existential dynamics that, might I say, would tend to strive toward an asymptotic model - as simplistic as it may be - that assures, through flattening the asperities of incidents, some mental and behavioral ease?

Humans, don't we live our lives as stories we tell ourselves? Aren't we scanning possible scripts - even unplausible ones like in daydreaming - leaving our ISPs working more or less under constraints? Aren't we processing at least occasionally our MOps to reorganize them in new MONs focusing on the best reminiscences we call upon? Aren't we even sometimes contemplating in counterfactual modes of thought paths that might have been open to us but that we never took: "Whom would I be if I had not met that person? What would I be doing now if had not attended this conference? Who would I choose to be if I were cloned?"

Don't weighed attraction vectors steer story tellers in their narratives like they steer our MONs and ISPs in everyday life and in scholarly writings as well [Darányi, in this volume]? After all, isn't life a story that we build as we live it?

5. ACKNOWLEDGMENTS

The research reported in this paper was supported mainly by grants from the Social Science and Humanities Research Council of Canada.

6. REFERENCES

[1] Bastian, A. 1868. *Das Beständige in den Menschenrassen und die Spielweite ihrer Veränderlichkeit*, D. Reimer, Berlin.

[2] Bowles, R. 2000. An Introduction to Cellular Assemblies. http://richardbowles.tripod.com (01-10-10).

[3] Coppet de, D., and Iteanu, A. 1998. *Cosmos and Society in Oceania.* Berg, Oxford.

[4] Darányi, S. 2001. Classical Myths and Transformation: Computer Observation of the Lévi-Strauss Formula at Work, in P. Maranda, Ed. *The Double Twist: From Ethnography to Morphodynamics*. The University of Toronto Press, Toronto, 156-176.

[5] Fernandez, J.W. 1977. The Performance of Ritual Metaphors. In J.D Sapir and J.C. Crocker, Eds. *The Social use of metaphor. Essays in the anthropology of rhetoric,* University of Pennsylvania Press, Philadelphia, 100-131.

[6] Fontanille, J. 2003. *La Sémiosphère. Sémiotique du discours.* Presses de l'Université de Limoges, Limoges.

[7] Gaudreault, R. 2001. Formalization: A Tool for Semiotics. *The American Journal of Semiotics: Rhetorical Semiotics,* 17, 1, 123-133.

[8] Hebb, D.O. 1949. *Organisation of Behavior*, John Wiley and Sons, New York.

[9] Hockett, C.F. 1987 *Refurbishing our foundations: elementary linguistics from an advanced point of view*. J. Benjamins, Amsterdam-Philadelphia.

[10] Jagota, A. 1999. Hopfield Neural Networks and Self-Stabilization. *Chicago Journal of Theoretical Computer Science*, Article 6, http://cjtcs.cs.uchicago.edu/articles/1999/6/contents.html (01-10-10).

[11] Kamp, Y., and Hasler, M. 1990. *Réseaux de neurones récursifs pour mémoires associatives*. Presses polytechniques et universitaires romandes, Lausanne.

[12] Köngäs Maranda, E. K., and Maranda, P. 1971. *Structural Models in Folklore and Transformational Essays.* Mouton, Paris - The Hague.

[13] Leblanc, C. 1994. From Cosmology to Ontology through Resonance: A Chinese Interpretation of Reality. In G. Bibeau and E. Corin, Eds. *Beyond Textuality. Ascetism and*

Violence in Anthropological Interpretation. Mouton de Gruyter, Berlin, 57-78.

[14] Le Roux, D., and Vidal, J. 2000. VERBATIM: Qualitative Data Archiving and Secondary Analysis in a French Company, *FQS Forum: Qualitative Social Research*, vol. 1, 3, http://www.qualitative-research.net/fqs-texte/3-00/3-00rouxvidal-e.pdf (01-10-10).

[15] Létourneau, J. 1992. La Mise en intrigue. Configuration historico-linguistique d'une grève célébrée: Asbestos, *Recherches Sémiotiques/Semiotic Inquiry* 12, 53-57.

[16] Lotman, Y. 1990. *Universe of the Mind: A Semiotic Theory of Culture*. Indiana University Press, Bloomington.

[17] Lotman, Y. 1998. *La Sémiosphère*. Presses de l'Université de Limoges, Limoges.

[18] Lévi-Strauss, C. 1949. *Les structures élémentaires de la parenté*. Presses Universitaires de France, Paris.

[19] Lévi-Strauss, C. 1958. *Anthropologie structurale*. Plon, Paris.

[20] Lévi-Strauss, C. 1971. The Deduction of the Crane. In P. Maranda. and E. Köngäs Maranda, Eds. *Structural Analysis of Oral Tradition.* University of Pennsylvania Press, Philadelphia, 3-21.

[21] Lévi-Strauss, C. 1973. *Anthropologie structurale deux*. Plon, Paris.

[22] Maranda, P. 1979. Myth as a Cognitive Map: A Sketch of the Okanagan Myth Automaton. *UNESCO Conference on Content Analysis*, Pisa, 1975, reprinted in W. Burghardt and K. Holker, Eds. *Textprocessing/Textverarbeitung*. Mouton de Gruyter, Hamburg – Berlin, 253-272.

[23] Maranda, P. 1980. The Dialectic of Metaphor: An Anthropological Essay on Hermeneutics. In S. Suleiman and I. Crosman, Eds. *The Reader in the Text*. Princeton University Press, Princeton, 183-204.

[24] Maranda, P. 1981. Semiotik und Anthropologie, *Zeitschrift für Semiotik* 3, 227-249.

[25] Maranda, P. 1984. Semiography and Artificial Intelligence, *International Semiotic Spectrum* 4, 1-3.

[26] Maranda, P. 1986. L'Imaginaire artificiel: esquisse d'une approche. *Recherches sémiotiques/Semiotic Inquiry* 5, 376-382.

[27] Maranda, P. 1989. *DiscAn: A Computer System for Content and Discourse Analysis*. Laboratoire d'Anthropologie Sociale, Québec.

[28] Maranda, P. 1990. Vers Une Sémiotique de l'intelligence artificielle. *Degrés* 18, 62, a1-a15.

[29] Maranda, P. 1992. Mother Culture Is Watching Us, or Probabilistic Structuralism: The Semiotic Space of Racine's *Andromaque*, In E. Nardocchio, Ed. *Reader Response to Literature: The Empirical Dimensio.,* Mouton-de Gruyter, Berlin, 173-192.

[30] Maranda, P. 1994a. Beyond Postmodernism: Resonant Anthropology. In G. Bibeau and E. Corin, Eds. *Beyond Textuality. Ascetism and Violence in Anthropological Interpretation*. Mouton de Gruyter, Berlin, 329-344.

[31] Maranda, P. 1994b. Imagination Structuring Processes. In L. J. Slikkerveer and B. van Heusden, Eds. *The Expert Sign: Semiotics of Culture. Towards an interface of ethno- and cosmosystems*. DSWO Press, Leiden, 169-186.

[32] Maranda, P. 1997. Metaforas metamorficas: operadores que aplicam cultura ao comportamento. In M. Rector, Ed. *Comunicaçao na era pos-moderna*. Editora Vozes, Petropolis, 116-127.

[33] Maranda, P. 2005. Ethnographie, hypertexte et structuralisme probabiliste (avec commentaires par André Petitat et réponse de P. Maranda). In D. Mercure, Ed. *L'Analyse du social: les modes d'explication*. Les Presses de l'Université Laval, Québec, 183-220.

[34] Maranda, P. 2007.Peuple des eaux, gens des îles: Hypertexte et peuples sans écritures, www.oceanie.org. In B. Reber, Ed. *Humanités numériques I – nouvelles technologies cognitives et épistémologie*. Hermès - Lavoisier, Paris, 215-228. (English translation in press.)

[35] Maranda, P. 2008. Myth and Metamorphic Metaphors. Exchange and Sea-Land Synergy in Malaita, Solomon Islands. In A. Strathern and P. J. Stewart, Eds. *Exchange and Sacrifice*. Carolina Academic Press, Durham, 55-72.

[36] Maranda, P. (in print). Speak, That I Be! Echo Chambers and Rhetoric. In I. Strecker and C. Meyer, Eds. *Rhetoric Culture. Theory and Exemplars*. Studies in Rhetoric Culture I, Berghan Books, Oxford – New York.

[37] Maranda, P. and Nze-Nguema, F. P. 1994. *L'Unité dans la diversité culturelle: Une Geste bantu*. Presses de l'Université Laval et A.C.C.T., Québec – Paris.

[38] Minsky, M.1975. A framework for representing knowledge. In P. Winston, Ed. *The Psychology of Computer Vision,* McGraw-Hill, New York, 211-277.

[39] Ott, E. 2006. Basin of attraction. *Scholarpedia* 1, 8, 1701.

[40] Petitot, J. 1994. Physics of meaning and morphodynamics, *Recherches sémiotiques/Semiotic Inquiry* 14, 387-408.

[41] Petitot, J. 2001. A Morphodynamical Schematization of the Canonical Formula for Myths. In P. Maranda, Ed. *The Double Twist: From Ethnography to Morphodynamics*. The University of Toronto Press, Toronto, 267-311.

[42] Propp, V. 1968. *Morphology of the Folktale.* Port City Press, Baltimore.

[43] Propp, V. 1970. Les Transformations des contes merveilleux. In V. Propp, *Morphologie du conte*. Seuil, Paris, 171-200.

[44] Rastier, F. 1991, *Sémantique et recherches cognitives*. Presses Universitaires de France, Paris.

[45] Rastier, F. 1992. *La Sémantique unifiée*. Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur, Orsay.

[46] Reinert M. 1992. The Methodology 'ALCESTE' and the analysis of a corpus of 304 stories of nightmares of children.

In R. Cipriani and S. Bolasco, Eds. *Ricerca qualitativa e computer. Teorie, metodi e applicazioni*, FrancoAngeli, Milan, 203-223.

[47] Schank, R. C. 1982. *Dynamic Memory.* Cambridge University Press, Cambridge.

[48] Schank, R. C. 1995. *Tell me a story: Narrative and intelligence.* Northwestern University Press, Evanston.

[49] Schank, R., and Abelson, R. 1977. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structure.* Lawrence Erlbaum Associates, Hillsdale.

[50] Simondon G. 1964. L'individu et sa genèse physico-biologique (l'individuation à lalumière des notions de forme et d'information). Presses Universitaires de France, Paris.

[51] Strecker, Y., Meyer, C., and Tyler, S. 2000. *Rhetoric culture. Outline of a project for the study of the interaction of rhetoric and culture.* Institute of Ethnology and African Studies, Johannes Gutenberg University, Mainz.

[52] Talbott W. 2001 "Bayesian Epistemology", *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/epistemology-bayesian (01-10-10).

[53] Thom, R. 1988. *Esquisse d'une sémiophysique*, InterÉditions, Paris.

[54] Veselovskij A. N. 1940. *Istorieskaja poetika*, Leningrad, available to institutions as e-book from De Gruyter Mouton, Berlin.

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

[56] Weenink, D.J. 1992. Introduction to neural nets. *Proceedings of the Institute of Phonetic Sciences*. University of Amsterdam 15, 1-25.

[57] Weenink, D.J. 1997. Category ART: A Variation of Adaptive Resonance Theory Neural Networks. *Proceedings of the Institute of Phonetic Sciences*. University of Amsterdam 21, 117-129. http://www.fon.hum.uva.nl/Proceedings/Proceedings21/DavidWeenink/DavidWeenink.html (01-10-10).

Beyond Reported History: Strikes That Never Happened

Martha van den Hoven
TiCC, Tilburg center for Cognition and Communication
University of Tilburg
Tilburg, The Netherlands
marthov@gmail.com

Antal van den Bosch
TiCC, Tilburg center for Cognition and Communication
University of Tilburg
Tilburg, The Netherlands
Antal.vdnBosch@uvt.nl

Kalliopi Zervanou
TiCC, Tilburg center for Cognition and Communication
University of Tilburg
Tilburg, The Netherlands
K.Zervanou@uvt.nl

ABSTRACT

We present a study on applying text analytics methods to historical text and data to uncover aspects of event structure. First, we associate primary historical resources, newspaper articles, to a secondary resource, a database of labor conflicts in the Netherlands, detecting newspaper stories denoting labor conflicts. For the task of retrieving newspaper articles based on database record information, we construct a query model exploiting the database record fields. We consider labor conflicts as historical events referred to in sequences of newspaper article narratives, of which the climax, i.e., the strike, may or may not have occurred. We analyze documents preceding a strike by considering them as a sub-narrative class of "strike threat" articles, and we then attempt to retrieve articles referring to conflicts which were about to burst into strike, but for some reason never did: strikes that never happened.

1. INTRODUCTION

The advent of the digital information era has started to expand its effect from transforming conventional information media, such as paper archives into digital form, to offering new computational research methods, also to Humanities research areas. In these relatively new domains of application, dedicated research in language and information technologies, for instance the set of tools referred to as text analytics methods, aims to provide solutions for intelligent information storage, access and retrieval.

In historical research, facts and events reported in textual sources play an important role in documenting history. Primary sources of historical information, such as letters, newspapers and brochures, and secondary historical sources, such as biographies and research publications, constitute the principal research material of historical research.

First International AMICUS Workshop, October 21, 2010, Vienna, Austria.

Strikes as an indicator of social and labor unrest are events of interest for social historians. Unfortunately, long-term data collections on events related to labor unrest, such as strikes, exist for very few countries [12]. Much of the current databases are manually compiled by historians, primarily based on investigation of newspaper articles [12, 15]. The development of retrieval and information extraction techniques may contribute by partly automating and speeding up the primary source analysis process. Strikes, if viewed as a series of labor unrest sub-events evolving over a time period and culminating in a strike, may be detected automatically by activity patterns preceding the strike itself. Such an analysis may assist in detecting both conflicts that resulted in a strike, as well as pinpointing conflict periods that were resolved without a strike happening. These resolved conflicts can be characterized as being *counterfactual* [3] strikes, in that the prelude to the non-strike is of the same structure as the prelude to the strike.

In this work, we first attempt to detect associations between secondary historical sources, namely a manually developed strikes database [15] and the respective primary sources: newspaper articles reporting the strikes. Subsequently, we discover patterns of narrated activities preceding a strike, so as not only to automate the detection of eventual strikes, but also to detect conflicts which, for various reasons of potential interest to historians, never resulted in a strike. This work has been carried out within the framework of the HiTiME[1] project [14], a collaboration between the ILK Research Group of Tilburg University[2] and the International Institute of Social History[3] (IISH). The project aims at analyzing and associating social history text documents, so as to improve access and to support historical research.

In this paper, we start by a description of our text data and tools used in our method implementation, followed by a presentation and discussion on the two approaches followed in strikes and labor conflict detection: first the association of primary sources, i.e., the articles, to the strikes reported in a manually constructed database and, second, the detection

[1]http://ilk.uvt.nl/hitime/
[2]http://ilk.uvt.nl/
[3]http://www.iisg.nl/

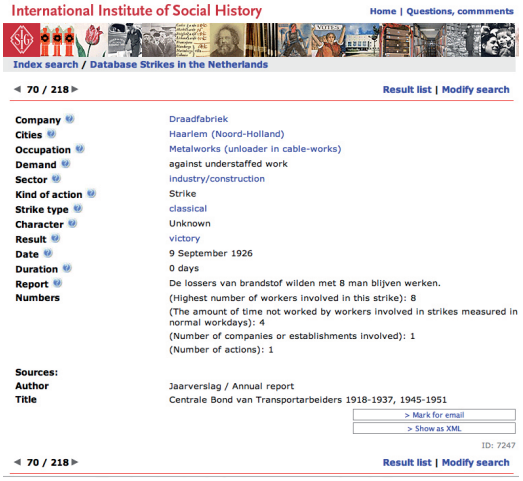

Figure 1: Example of strike record and respective DB fields (online version).

of articles narrating the threat of a strike that has never taken place.

2. DATA AND METHODS DESCRIPTION

In this section, we describe our research data and the methods used in our research.

2.1 Data

2.1.1 The Strikes Database

The strikes database, central to our study has been manually developed for the purposes of historical research [15]. It contains 16,427 records describing strikes, lockouts and other actions in the Netherlands, with the earliest record dating from 1372. The majority of the strikes recorded in the database occurred between the years 1900 and 1940. In this work, we only consider strike actions. A *strike* is defined as a specific type of labor action conforming to the following three criteria [16]:

1. It is undertaken by employees only; students' and farmers' actions are not considered;
2. It involves a temporary interruption of work;
3. It is a collective action, involving the participation of at least two persons.

The strikes database has a relational structure, consisting of 32 linked tables. The most important attribute fields for each record are illustrated in Figure 1, which depicts a screenshot of the online Strikes database search interface on the IISH website[4].

One database field of particular importance is the *report* field. As shown in Figure 1, this field contains free text providing information on strikes that could be otherwise classified, such as the names of negotiators, or agitators involved, and, typically, a short report or summary of the strike. Information such as the names of key persons involved is of great use when searching for strike-related articles, as these names are likely to be mentioned in relevant newspaper articles.

In other instances, the report field provides comments on the content of other database fields, such as the strike duration, or the strike date, for example:

'De duur is een minimum'
(The duration is a minimum)
'Kan ook in 1915 geweest zijn'
(Could have also been in 1915)

In other cases, it may refer to the workers demands, or the strike results, for example:

'Tegen het ontslag van een collega'
(Against the dismissal of a colleague)
'Verloren door onderkruiperij'
(Lost because of strikebreakers)

Our research focuses on the strikes reported within the 1910-1940 period, in which over 50% of all recorded strikes in the database occur.

2.1.2 The Daily Digital Newspaper Collection

The Daily Digital Newspaper Collection has been the result of a digitisation project, initiated in 2006, the *Databank Digitale Dagbladen* project[5] (DDD – Databank of Daily Digital Newspapers). This project is undertaken by the Royal Library of the Netherlands and aims at digitizing and providing online access to eight million pages of daily Dutch newspapers by the year 2011. These newspapers constitute a selection of a representative 8% of all newspaper titles, dating from 1618, when the first newspaper was published in the Netherlands, to 1995 [6]. The online DDD website was officially launched on May 27, 2010, making available one million newspaper pages. The digitized data is stored in XML-format, using Dublin Core standards[6]. The collection has been richly annotated with various types of metadata information, the most relevant to our research being the publication date, the headers and the article type. For the latter, a distinction is made between advertisements, illustrations, social announcements (i.e., births, deaths and marriages), and, finally, the news articles which are of interest for our purposes.

The digitized collection may be accessed by the *Historische Kranten* website[7], where a search form interface allows the users to perform queries on both a limited set of metadata, such as date, article type, title, distribution range and publication area, as well as on the newspaper content by boolean and regular expression term queries. The results are available in pdf and plain text. Additionally, the Royal Library provides an alternative mode of access to the collection through an SRU interface. SRU (Search/Retrieval via

[4]http://zoeken.iisg.nl/search/search?action=transform&xsl=strikes-form.xsl&col=strikes&lang=en

[5]http://kranten.kb.nl/about/
[6]http://dublincore.org/
[7]http://kranten.kb.nl/

URL)[8] is a standard XML-focused search protocol for internet search queries which exploits the HTTP GET method for message transfer [9]. The queries are formed in CQL (Contextual Query Language), a query language designed not only to be intuitive and human readable, but also more powerful than languages, such as Google-like languages [9]. In our experiments, we have opted for the SRU interface, because it provides greater flexibility and more querying options. An example of a typical query in CQL for our purposes would be:

```
http://jsru.kb.nl/sru?query=(staking and haven and
dc.type exact artikel and dc.date >="1924/04/20"
and dc.date<="1924/04/30")&maximumRecords=500
```

This query example would return a single XML file containing pointers (links) to the relevant articles. The pointer to the location of the article is created dynamically by means of article identifier resolution. For example, an element `dc.identifier` containing:

```
http://resolver.kb.nl/resolve?urn=ddd:010001322:
mpeg21:a0051:ocr}
```

would be resolved to the following URL location:

```
http://resources2.kb.nl/010000000/articletext/
010001322/DDD\_010001322\_0051\_articletext.xml
```

This technique allows for the location of the data to be changed, without respective alteration to the metadata referring to it.

Similarly to other digitized data collections, the DDD collection is affected by OCR quality issues. In particular, poor printing, or deterioration of the original paper, often results in erroneous and inconsistent character recognition. This type of problem affects the retrieval of articles based on search for content words, because this search does not return documents containing spelling variations of the search term. CQL supports fuzzy search, nevertheless this functionality has not yet been implemented by the Royal Library in the current version of the interface.

Another issue affecting the retrieval quality is the article segmentation. In the DDD collection the article segmentation has been performed in a semi-automatic manner, whereby the results of the automatic partitioning were inspected and corrected by humans. We observe however that in some cases large articles consisting of several sections including subheaders have been erroneously split into separate article files, whereas in other cases, newspaper sections containing diverse news stories, such as *Allerlei* (All-sorts), or *Uit de provincie* (From the province), are partitioned as a single article. Unlike the problems related to OCR, the issue of article segmentation is not only the result of current limitations in automatic document segmentation, but it also relates to discourse and document structure issues pertaining to human judgement and collection design decisions.

[8]http://www.loc.gov/standards/sru/

The problems related to OCR and article segmentation constitute recurrent issues, common to many digitisation efforts and research challenges for language and text analysis researchers.

2.2 Methods Applied

2.2.1 Memory-Based Learning: TiMBL

The task of identifying articles preceding a strike and denoting some form of labor conflict which may, or may not have resulted in a strike action, may be viewed as a boolean text classification learning task, whereby the articles of interest form one class and all other articles, the other. For this purpose, in our approach we have adopted a memory-based learning approach. In particular, we apply the TiMBL[9] implementation of memory-based learning for text classification. TiMBL is a learner which exploits the k-Nearest Neighbor algorithm to perform supervized classification tasks [2, 1]. This algorithm assigns to each new unclassified instance the majority class of its k most similar known instances, i.e., its "k-nearest neighbors".

TiMBL weighs each instance features according to the information they reveal about the respective instance class. In our experiments we have used the information gain ratio measure. In estimating information gain, one attempts to measure the informative value of a given feature as the difference between the uncertainty about the class in a situation without knowledge of that feature value, and a situation with that knowledge. A problem with information gain is that it tends to overrate features with many values. The information gain ratio measure alleviates this problem by dividing information gain by the entropy of the feature values [1].

2.2.2 Linguistic Analysis: Tadpole

We exploit the proper noun recognition included in the Tadpole part-of-speech tagger for Dutch, to identify expressions denoting proper names, such as organisations, persons and locations, because these expressions are of particular importance for the retrieval of strike related information. Tadpole[10] is a morphosyntactic tagger and dependency parser for Dutch [13], which relies on TiMBL for its POS tagging and parsing classification tasks. Tadpole initially tokenizes the input text and then proceeds to assigning part-of-speech, i.e., grammatical category information, such as noun, verb, etc., to each input text token. Tadpole includes a lemmatizer and a morphological analyser whereby, for a given word surface form, such as *"gesteld"*, its respective affixes are recognized and its lemma *"stellen"* is identified as its canonical, normalized form. Tadpole, finally, includes a dependency parser which creates a graph representing the syntactic dependencies among each sentence constituents.

2.2.3 Boolean and Ranked Retrieval

Boolean retrieval is a classic method of information retrieval. In this type of retrieval, a query term is either found, or not, in a document, thus resulting in the document being considered as relevant, or not, respectively. The query terms may be combined by boolean operators, such as AND, OR, and

[9]http://ilk.uvt.nl/timbl/
[10]http://ilk.uvt.nl/tadpole/

NOT. Operators, such as AND and OR, affect the retrieval results by applying constraints, or relaxing the matching of query terms to documents [7, 11]. For example, a query searching for *'A and B'* will not return documents merely containing A without B, or B without A, thus limiting the set of results. Conversely, a query searching for *'A or B'* will return both the documents where A and B appear individually, as well as those where A and B are in combination. A query formalism can be extended to support regular expression operators, such as Kleene stars and other character range operators, to allow for fuzzy term matching, i.e., query term matching even in cases where the document term differs slightly by one or more characters from the query term. In other approaches, the retrieval system may use dictionaries so as to allow for matching of normalized, e.g., stemmed words, or lemmas, to surface form word variants, or use a relaxed matching function, such as nearness, where search terms have to be found within a certain distance of each other [8].

The principal problem of the boolean type of retrieval lies in that it provides only limited means to express a graded matching of the documents to a given query. In principle, all document results to a query are treated as equally relevant, although in reality some may be more relevant to a given query than others. Relevance cannot be easily defined in an objective manner, since the results to a query may be relevant or irrelevant depending on the particular user information requirements. Thus, relevance cannot be viewed as a strictly boolean concept. Ranked retrieval approaches provide a solution to these problems by estimating the similarity of a document to a given query and assigning higher ranking to documents closely matched to a query and lower ranking to documents less similar to the query terms. The similarity measure used in this work is the cosine similarity. In order to calculate cosine similarity, documents are represented as vectors in a multidimensional space, where each dimension is a term and the number of dimensions is the total number of terms (i.e., words) appearing in the document collection. Cosine similarity measures similarity as the cosine between such document vectors. The similarity value ranges between 0, when the documents have nothing in common, and 1, when the documents are the same. The value of the coordinates in the document vectors can be a simple binary value, 0 if the term is absent in the document, and 1, if the term is present. In our implementation of the cosine similarity, this value is determined by the $tf \cdot idf$ of the term. The $tf \cdot idf$ value is the product of the tf, term frequency, and the idf, inverse document frequency. The term frequency is a count of how often a term occurs in the document in question. The inverse document frequency is the logarithmic value of the total number of documents, divided by the number of documents that the term appears in. The idf of a rare term is high, and the idf of a frequent term is low. Consequently, rare words appearing often in a single document have a high value in this document's vector. The assumption is that such a rare word is meaningful and topical in the context of the document; the document can be assumed to be at least partly about this word.

2.2.4 Evaluation Measures

In this work, we use *Precision* to assess the entire set of our retrieval results and *Mean Average Precision* (MAP) to measure the precision in our ranked retrieval output. Precision is defined as the ratio of the number of correct results divided by the total number of retrieved documents. The Mean Average Precision (MAP) is used to assess the ranking quality of a retrieval method, i.e. whether relevant documents are ranked higher than possibly less relevant, or irrelevant ones, and is estimated as the mean value of the average precision of individual queries. In turn, the average precision of a query is calculated as the sum of all precision values at each rank position of the results. According to Manning et al. [8], if the set of relevant documents for an information need $q_j \epsilon Q$, is $\{d_1, ..., d_m\}$ and R_{jk} is the set of ranked retrieval results from the top result until you get to document dk, then

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{mj} \sum_{k=1}^{mj} Precision(R_{jk})$$

For example, if we have three retrieved documents for a given query, two of which are relevant and ranked at positions 1 and 3 in the ranked results list respectively: The precision at rank 1 is 1/1, precision at rank 2 is not calculated, because that document is not relevant, and the precision at rank 3 is 2/3. Thus, the average precision AP for this query would be:

$$AP = \frac{\frac{1}{1} + \frac{2}{3}}{2} = 0.83$$

In this manner, the MAP of a query result is 1 if all relevant documents are ranked higher than the non-relevant.

3. STRIKE ARTICLES RETRIEVAL

3.1 Boolean Query Modelling

The approach taken to find newspaper articles about strikes is to devise a general query model which can be adapted for each particular strike. For this purpose a boolean querying mechanism is used. As presented in Section 2.1.2, this retrieval method is supported by the Royal Library search interfaces to the newspaper articles collection. For our query model we have to take into consideration the issues discussed in Section 2.2.3, so as to to find the optimal balance between precision and recall by manipulating the boolean query constraints. Our information source for the search query terms is the strikes database records. Our considerations for constructing queries from the database records are as follows.

First, the most general topical term *"staking"* (i.e., strike) itself should be a good identifier. During the period under examination, no synonyms of the word *"staking"* were in use. In the period before our focus period, during the late nineteenth century, when strike activities were becoming common, there was more terminological drift. Foreign loan terms, such as *"strike"* and *"grève"* were used, alongside dialectal terms, such as *"bollejeije"* and *"laveij"* [16] The terms *"staken"* and *"staking"* then became the prevailing terms, at least until after the Second World War, after which new term, such as *"werkonderbreking"* and *"stiptheidsactie"* were introduced to describe new types of labor actions.

Given that newspapers report on current events, the start date and duration of a strike should be important features, as a strike will be a narrative topic for only a limited period, centering around the strike itself. Other salient fields

that our query should include are the occupation sector, the location, and the particular persons, workers' unions and companies involved. The *report* field was POS-tagged (using Tadpole) and all tokens are extracted that receive the proper noun tag. These are included in the query terms. In our query term selection, only terms consisting of more than two characters are used, to avoid matching false positives due to OCR-errors in the newspaper articles. The resulting boolean query is formulated as follows:

```
stak?n* AND (term_1 OR term_2 OR ... OR term_n)
date BETWEEN start_date - 7 AND end_date + 3
```

where any form denoting strike (i.e., *staken, stakend, stakenden, staking,* etc.) should be matched alongside any other term originating from the database record for a given time period ranging from a week before the reported start date to three days past the end date. The size of these pre-strike and post-strike windows were estimated by manual inspection of a few cases of strikes, and were judged to cover the majority of articles on specific strikes; selecting higher values would result in lower precision, without likely compensation at the recall side.

In order to test the effectiveness of the query, we select a particular set of strikes from the database likely to be important enough to be reported in the news. We select strikes from the 1920-1940 period, with 250 or more strikers, and a duration between two and six days—longer strikes were avoided to keep the number of query results low to enable manual evaluation, obviously introducing a bias to short strikes. Our selection produces 27 strikes. For each strike a query was composed as described above, and these queries were executed using the web based interface of the Royal Library. All results were checked manually for their relevance to the strike that was queried. An article is deemed relevant if it is:

- entirely about the strike;
- partly about the strike;
- a news overview article also mentioning the strike;
- about a similar strike in another company concerning the same conflict.

Results are illustrated in Table 1. In this table, we show a sample of individual query results, based on the respective database record and indicated by the record ID, namely the total number of retrieved documents, the relevant documents and the respective precision for each query and the entire (All) query results. The MAP for all queries is 48%. In our query formulation, as described above, we have applied broad and relaxed constraints, thus favouring recall rather than precision. For this reason, we consider our precision results satisfactory. Another issue affecting precision is the query terms themselves. For example, when a frequent term such as *"Amsterdam"* is in the query, the precision is likely to drop. Moreover, using the broad pre-strike period aims at retrieving articles mentioning a threat of strike, negotiations, ultimatums, or other signs of unrest. However, in some spontaneous strike cases where the strike was not foretold in any newspaper article, this extended time period leads to the retrieval of irrelevant articles. In an experiment we performed on this particular category of strikes, further constrains on the time period increased the overall precision to 56%, while decreasing recall. In order to avoid information loss, aim for higher recall, and attempt to solve the consequential loss of precision in other ways.

Table 1: Strike Articles Sample Retrieval Results

ID	Total	Relevant	Precision	Average Precision	Average Precision Extended Query
5876	18	4	0.22	0.16	**0.23**
6333	9	1	0.11	**0.14**	0.14
7843	13	9	0.69	0.51	**0.52**
8080	10	4	0.40	0.33	0.30
8135	15	6	0.40	0.33	**0.47**
8289	11	8	0.73	0.56	**0.58**
8401	4	3	0.75	**1.00**	1.00
8536	23	18	0.78	0.72	0.70
8621	29	17	0.59	**0.65**	**0.69**
...	...	...	...	...	...
All	333	160	0.48		
Mean			0.46	0.46	0.50

3.2 Ranked Retrieval

As discussed in Sec. 2.2.3, a boolean retrieval model considers all document results equally relevant. A ranking method may improve on our average precision if it ranks more relevant documents towards the top. In our implementation of ranked retrieval, we have applied the cosine similarity measure as described in Section 2.2.3. For the similarity estimation, we consider queries as term vectors formed from the content of the database fields, and compare those to the respective document vectors. The ranking results are evaluated using Mean Average Precision measures.

The results are illustrated in Table 1 on the *Average Precision* column. We observe that in some cases indicated by boldface, ranking improves precision. However, in others ranking is ineffective, and the overall MAP is similar to the unranked mean precision, i.e., 46%. We consider that one reason for cosine similarity not resulting in effective ranking is due to the dissimilarity between the articles and the database queries. In order to improve our results, we experimented with query expansion. In particular, we expanded the database record terms in our queries with terms appearing frequently in relevant articles and not as frequently in irrelevant ones. For this purpose, a term frequency list was compiled for all retrieved relevant articles, and one for all retrieved irrelevant articles. These lists contain terms, as tokenized by Tadpole, alongside their respective frequencies. From the list of relevant articles terms we removed:

- words consisting of one or two characters,
- names (as they are present in the strike record),
- words of similar high frequency in both lists, such as stopwords,

- words with low term frequency (less than 15).

The remaining list of 84 terms was added to each strike query document, and the cosine similarity between this document and the retrieved articles was calculated. The results illustrated in Table 1 on the *Extended Query* column show an increase in average precision in the cases indicated by boldface; the overall mean average precision is improved by 4%.

4. STRIKES THAT NEVER HAPPENED

In order to identify conflicts that never resulted in a strike, we first search and analyse the articles which are known to have lead to a strike. For this purpose we have queried the database for all strikes between 1910-1939 which were organized by a union, as these types tend to be organized and announced in advance:

```
year BETWEEN 1910 AND 1939
number of strikers >= 500
character = 'Union'
```

The results were 108 strikes for each of which we identified a preceding article manually, using the following query:

```
(stak?n* OR term_1 OR term_2 OR ... OR term_n)
date BETWEEN start_date - 30 AND start_date
```

Relevance was assessed manually for all retrieved articles. Moreover, if many articles were found in the first days of the search window, an additional search was done for the preceding month. Only articles found relevant were stored.

The task of identifying articles preceding a strike may be viewed as a boolean text classification learning task with a single class, whereby the articles of interest form the positive class and all other articles are negative cases, implicitly. For this task we have used TiMBL, our 108 manually assessed articles as positive examples, and a 100 random articles from the same period as negative. For our article representation model, we have created a frequency list comprising of content words, i.e., those words POS-annotated by Tadpole as nouns, adjectives, or verbs. This list excludes stopwords, forms of the auxiliary verbs *"zijn"* and *"hebben"* (*to be* and *to have*), words of length less than three characters (so as to alleviate OCR errors) and words below frequency of occurrence 5. Contrary to our frequency lists compiled for the *Extended Query* experiment, we do not need to compare frequencies between relevant and non-relevant articles in this case, because TiMBL, as described in Sec. 2.2.1, already weights features according to the information they reveal about the class. The resulting frequency list of 208 words constitutes our classification features. For each article (prelude and non-prelude) all 208 features get a binary value: 0 if the word does not occur in the article, and 1 if the word occurs. TiMBL is then used to calculate the information gain of the different features. In a leave-one-out experiment with TiMBL, using the default IB1 metric, an accuracy of 94.9% was achieved, indicating that articles narrating about the prelude of a strike can to a reasonable extent be distinguished from other random articles—and that the 208 selected features appear a good starting point for further querying in the full newspaper archive. A sample of our top highest scoring terms is listed in Table 2.

Table 2: Top 10 Gain Ratio scoring features (terms) in 'strike prelude' vs. 'other' articles classification task

Gain Ratio	Term	Gloss
0.324	conflict	*conflict*
0.303	werkgevers	*employers*
0.298	mijnwerker	*miner*
0.298	mijnwerkers	*miners*
0.259	rijksbemiddelaar	*state negotiator*
0.258	arbeiders	*workers*
0.254	Schelde	*Schelde* (name)
0.253	loon	*wages*
0.240	mijnwerkersbond	*mining worker's union*
0.239	arbeider	*workers*

For the formation of our query aimed at finding articles reporting on strike preludes, we selected terms from the set of 208 features with gain ratio $GR \geq 0.100$, and excluded all proper names as those are characteristic of a particular strike. We thus compiled a list of 53 query terms. These terms were subsequently combined in an OR-query to search for newspaper articles for each week of the 1910-1939 period. Another query was performed on each week merely to count the total number of articles in that particular week.

In order to select an appropriate period for our experiments, we assume that on weeks of labor conflict there are more articles matching the conflict query. Our results showed that the percentage of candidate conflict articles in the total pool of articles in a week varies from 16.1% to 46.1%, indicating that our query has been very broad. We identified three weeks in which this percentage of conflict articles distinctly peaked compared to their surrounding period, one for each decade: 24–30 December 1912, 11–17 September 1926, and 16–22 April 1938. For these weeks all the articles that scored positive are manually categorized. An article can be labeled as one of four categories:

P (*prelude*): if the article is about a labor conflict that could lead to a strike. An article is categorized as P, when it refers to a named group of workers and employers in conflict. The group may be a profession, or a company, or an industry. Political discussions not referring to a specific conflict are categorized as *Other*.

S (*strike*): if the article refers to a strike happening at the time of writing, or a stike that has ended.

F (*foreign*): if the article is about foreign labor conflicts such as strikes.

O (*other*): all other articles not belonging to any of the above categories.

In cases where an article may be classified into multiple categories, the first applicable is selected. The results of this

Table 3: Strike Prelude Articles Retrieval Results

Week	Total	Prelude	Precision	Average Precision	Prelude, Strike & Foreign	Precision	Average Precision
24 Feb 1912	273	12	0.044	0.116	50	0.183	0.480
11 Sep 1926	133	2	0.015	0.065	6	0.045	0.321
16 Apr 1938	208	1	0.005	0.077	7	0.034	0.489
Mean			0.021	0.086		0.087	0.430

manual assessment is shown in Table 3, where *Total* indicates the total number of articles retrieved, *Prelude* the correct prelude articles, and *Prelude, Strike & Foreign* the prelude articles with the other two relevant categories.

To automatically assess our prelude article retrieval, we exploit our ranked retrieval approach to rank our prelude article retrieval results. However, in this case, we do not consider the query terms in estimating similarity; we rather use the entire set of term features characterizing prelude articles and calculate document cosine similarity to this term feature list. Results evaluation, similarly to our ranked retrieval evaluation for the database queries task, is based on MAP. As illustrated in Table 3, we have estimated MAP for both the relatively low set of strictly prelude articles, (indicated on the left side of the table), as well the MAP taking into consideration the entirety of our correct results indicating conflict, i.e., including strike and foreign article categories (indicated on the right side of the table). Overall, we observe that the results of this automatic process show a great improvement over the unranked precision results for the prelude articles.

5. DISCUSSION

In this work, the first objective was to find newspaper articles related to a particular strike. This has proved to be possible, although the precision has not been too high (0.48 for the 27 strikes the query was tested on). With a ranking system based on the strike record, and extended with terms that often appear in newspaper articles about strikes, a Mean Average Precision of 0.50 could be reached. In our evaluation, MAP is calculated based on binary relevance. Yet, the actual relevance value of the documents may not be so discrete. All documents mentioning the strike in question were deemed relevant, but some of those articles were overview or index articles in which headlines were listed, one of them regarding the strike. Some articles have not been correctly split in the Digital Newspaper archive, so two or more articles with different subjects form only one document. To overcome the shortcoming of MAP which only works on binary relevance judgements, Kishida [5] proposes the generalized average precision measure. This measure does not require binary values; it rather relies on a seven-point sliding scale. The application of a generalized average precision measure could provide a more refined and revealing evaluation of the scoring system, and should be considered in future work, provided that human judges can be recruited to perform the relevance assessment task.

An argument that could be made against efforts to improve precision, is that of serendipity. According to Foster and Ford [4]: *"serendipity would appear to be an important concept of the complex phenomenon that is information seeking"*. Striving for maximal precision sometimes bars the end user from discovering information in strictly non-relevant, but still related articles sharing some of the keywords (and thus perhaps part of the topics) that could lead to interesting findings in their own right.

In our evaluation we have not accounted for recall. In order to determine recall for our document collection, we would have to manually check all newspapers from 1910 until 1940, so as to ascertain that no mention of a strike is overlooked. A more feasible alternative would be to manually check only the newspapers corresponding to the respective date range of the strike query. However, this would still be a laborious and time consuming process. If the resources for this task were available, the estimation of recall could provide us interesting information regarding the performance of the query and the ways by which our recall could be improved.

One issue that is expected to affect recall is that of spelling variations and OCR errors, which are unaccounted for in our current approach. Both issues are relevant when dealing with scanned newspaper articles. Reynaert [10] describes a system which automatically detects and corrects OCR errors. This system, *Parallel Text-Induced Corpus Clean-up*, or *ParTICCL*, works with a set of character confusions that frequently occur in OCR-ed texts, for instance *"in"* is often confused with *"m"*, or *"o"* with *"e"*, and vice versa. These confusions are combined with a lexicon of correct words, so as to detect all existing erroneous word variants, and propose the respective canonical form. This system could be used to find all variants of a word, and use these all in our queries. For example, for the word *"staking"* alone, 199 variants can be found in the 1918 volume of *Het Volk*. These are variants within a Levenshtein distance of 2, so up to two characters in the word are changed. Adding these varieties would expand the queries, and as the method is reported to be rather accurate [10], this query expansion can be expected to increase recall. As it stands, the Royal Library is implementing this systems in the articles, so that this functionality will come with their search facilities automatically.

The second objective of this research, namely the discovery of strikes that never happened, proved to be challenging, but what the study shows is that automatic shortcuts can be devised that can be expected to offer reasonably fast alternatives for what would otherwise be tediously long manual searches. Our limited case study on three weeks of apparently high labor unrest proved successful, but with low precision, leaving some work to the user, who normally would be an expert user (e.g. a researcher of social history) capable of filtering out the relevant cases.

A potential improvement could apply to the term list used for detecting the respective articles. We based this list on contrasting strike prelude articles with randomly selected other articles. Yet, the border around the class that we are looking for could be better defined, if instead of random articles, we seek articles which are similar to the prelude articles in many aspects, yet are not about the prelude of a strike – they might, for instance, mention the same companies in the context of events unrelated to labor unrest. Another potential improvement could result from the refinement of feature selection used in classification, namely the selection of our query terms. In a comparative assessment of different methods for feature reduction in text classification, Yang and Pedersen [17] find that information gain, χ^2 and document frequency are good measures to use for that goal. In our task, querying can be viewed as a form of text classification. We have used the gain ratio (closely related to information gain, especially for binary features) to reduce the number of features, or rather search terms, to query for our articles. When compiling the list of feature terms for TiMBL, term frequency was used to determine which word was used as a feature. It might be an improvement if document frequency were used for this, by determining the number of documents a given word (i.e., term) appears in, rather than its frequency in the entire document collection.

Finally, a qualitative evaluation of our results by expert historians would reveal whether the material we have uncovered was relevant and worth discovering.

6. CONCLUSION

In this work, we have presented a case study that divided into experiments on associating primary historical resources, such as newspaper articles, to secondary resources, such as databases, and, subsequently, experiments on detecting newspaper stories denoting labor unrest. We based our study on a strikes database developed within the context of historical research, and a large collection of digitized newspaper articles from the beginning of the 20^{th} century. We employed two classic retrieval models: boolean and ranked retrieval. The task of detection of newspaper stories denoting labor conflict is viewed in our approach as a single-class classification task. We used for this purpose the TiMBL memory-based approach implementation and its respective information gain ratio functionality to classify our articles and select the most pertinent features for classification. Overall, our results indicate a relatively low precision in both tasks, which can be partly improved by ranking. In our approach we did not optimize precision, as we want to avoid compromising recall, and we assume our retrieval systems to be used by experts who are willing and capable of filtering relevant results also when these constitute less than 5% of the retrieved results—the alternative being to search for these relevant articles in an unlabeled pool of millions of articles.

Future experiments may specifically assess recall and the effects of OCR error to our retrieval. Refinements to our approach may include refinements of feature selection used in classification and fuzzy query matching.

More generally, we aim to connect to work on cross-document summarization, on the narrative of newspaper stories over time, and on real-time tracking of topics in the news, in order to bring hypotheses into our work that would sharpen our definitions of important events in the news, and how to players in these events, as well as the motifs and behavioral patterns in which players can be expected to act in these events.

7. ACKNOWLEDGMENTS

This research was carried out as part of the HiTiME project, funded by the CATCH programme of the Netherlands Organisation for Scientific Research (NWO). The authors wish to thank Sjaak van der Velden, Marien van der Heijden, and Dennis Bos for their expertise and for making available the resources used in this study, and to Matje van de Camp and Steve Hunt for discussions and assistance.

8. REFERENCES

[1] W. Daelemans and A. van den Bosch. *Memory-based language processing.* Cambridge University Press, Cambridge, UK, 2005.

[2] W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch. Timbl: Tilburg memory based learner, version 6.3, reference guide. ILK Technical Report 10-01, University of Tilburg, Tilburg, The Netherlands, 2009.

[3] N. Ferguson. *Virtual history: Alternatives and counterfactuals.* Picador, London, UK, 1997.

[4] A. Foster and N. Ford. Serendipity and information seeking: an empirical study. *Journal of Documentation*, 59(3):321–340, 2003.

[5] K. Kishida. Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments. NII Technical Report nii-2005-014e, National Institute of Informatics, Tokyo, Japan, October 2005.

[6] E. Klijn. Databank of digital daily newspapers: moving from theory to practice. *News from the IFLA Section on Newspapers*, (19):8–9, 2009.

[7] W. Lee and E. Fox. Experimental comparison of schemes for interpreting boolean queries. Technical Report TR-88-27, Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA, 1988.

[8] C. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval.* Cambridge University Press, Cambridge, UK, 2008.

[9] S. McCallum. A look at new information retrieval protocols: SRU, OpenSearch/A9, CQL, and Xquery. In *The World Library and Information Congress: 72nd IFLA General Conference and Council*, Seoul, Korea, 2006.

[10] M. Reynaert. Parallel identification of the spelling variants in corpora. In *Proceedings of the Third Workshop on Analytics For Noisy Unstructured Text Data*, pages 77–84, 2009.

[11] G. Salton, E. Fox, and H. Wu. Extended boolean information retrieval. *Communications of the ACM*, 26(11):1022–1036, 1983.

[12] B. Silver. *Forces of Labor. WorkersŠ Movements and Globalization since 1870.* Cambridge University Press, New York, NY, USA, 2003.

[13] A. van den Bosch, G. Busser, W. Daelemans, and S. Canisius. An efficient memory-based morphosyntactic tagger and parser for dutch. In F. van Eynde, P. Dirix, I. Schuurman, and V. Vandeghinste, editors, *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, pages 99–114, Leuven, Belgium, 2007.

[14] A. van den Bosch, K. Zervanou, M. van de Camp, M. van den Hoven, S. Hunt, and M. van der Heijden. Baseline measurement CATCH-HiTiME, version 1.0. ILK Research Group Technical Report Series no. 10-05, University of Tilburg, Tilburg, The Netherlands, May 2010.

[15] S. van der Velden. *Stakingen in Nederland. Arbeidersstrijd 1830-1995.* Stichting Beheer IISG/NIWI, Amsterdam, The Netherlands, 2000.

[16] S. van der Velden. *Werknemers in actie. Twee eeuwen stakingen, bedrijfsbezettingen en andere acties in Nederland.* Aksant, Amsterdam, The Netherlands, 2004.

[17] Y. Yang and J. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning*, pages 412–420, 1997.

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].

First International AMICUS Workshop, October 21, 2010, Vienna, Austria.

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

I am grateful to Piroska Lendvai (Hungarian Academy of Sciences, Research Institute of Linguistics, Budapest) and Pierre Maranda (Université Laval, Québec) for their comments on the draft of this paper.

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.

The Story of Science: A Syntagmatic/Paradigmatic Analysis of Scientific Text

Anita de Waard
Elsevier Labs, Burlington, VT, USA
Utrecht Institute of Linguistics, Utrecht, The Netherlands
a.dewaard@elsevier.com

ABSTRACT

Following Latour (1987), Latour and Woolgar (1979), Bazerman (1988), and others, we have proposed a model for scientific texts as being 'stories, that persuade with data' (de Waard et al., 2006; de Waard et al., 2009; de Waard and PanderMaat, 2009). The persuasive component (how claims are formulated and recognized) is being developed in a number of projects (de Waard et al., 2009); the data component is the object of a great deal of study on various platforms (see overview in de Waard, 2010). In this paper, we wish to comment on the narrative component: how a scientific paper is similar to a fairy tale, and how techniques developed to parse and access fairy tales could be used to improve access to scientific knowledge. This short paper has three parts: first, we provide a brief introduction on fairy-tale structure analysis; next, we offer a small overview of how scientific text can be similarly analyzed, and third, we discuss some ways we could use tools and technologies developed within digital humanities for improving access to scientific knowledge.

1. The Structure of Fairy Tales: Propp vs. Lévi-Strauss

The analysis of narrative structures in folktales has developed in two directions: one, following Propp, is a syntagmatic analysis of story structure, where the chronological order of events unfolding in folktales is described (Propp, 1968). Building on from Propp's analysis, work in the sixties and seventies by e.g. Thorndyke (1975) and Rumelhart (1977) focused on defining a 'story grammar' or 'story schema'– the 'systematic assignment of constituent structure' to stories. The main goal here is 'to look at a story and to identify the goals, subgoals, the various attempts to achieve the goals, and the various methods that have been employed' to tell the story (Rumelhart, 1977).

An orthogonal method of analyzing stories follows the work of Lévi-Strauss, and focuses on a so-called paradigmatic analysis of a story, where 'groups of relations between actors and events are sought throughout the text' (Lévi-Strauss, 1955). The focus here is to find and group together elements of mythology, also called 'mythemes', which occur in different myths and folk-tales.

The two different views can, and have been (Lévi-Strauss, 1955) represented by a two-dimensional coordinate system, where the syntagmatic analysis occurs in (narrative) time, and the other is a 'paradigmatic grouping', by grouping together different events related to e.g. marriage, murder, etc.

First International AMICUS Workshop, October 21, 2010, Vienna, Austria.

2. The structure of scientific papers: Up and down and side-to-side

2.1 Syntagmatic analysis

We can look at scientific texts in a similar way. To begin with, the Introduction-Method-Results-and-Discussion (IMRaD) structure, or more correctly, the Introduction-Experiments-Discussion structure, closely mimics the basic Story grammar elements of Setting-Episodes-Resolution. If we call the main research question the protagonist of the scientific story, a paper describes its adventures: from being born (through a lack of understanding of current theory) to facing challenges (as different experiments test and explore various characteristics of the research question) to its eventual destiny: usually, in a scientific paper, a part resolution, part transformation of the research question. Like Episodes, Experiments are often mini-stories in themselves, and start with a subgoal, then describe some development, and end with a small resolution, that leads to the next episode. A parallel between Rumelhart's (1980) story grammar and the structure of a scientific paper is given in Table 1.

Table 1: Comparing story grammar elements with syntagmactic components of scientific text

Setting	Time	Introduction	Background
	Characters		Objects of study
	Location		Setup
Theme	Goal		Research question
	Attempt		Hypothesis
Episode 1	Subgoal	Experiment 1	Subquestion
			Subhypothesis
	Attempt		Method
	Outcome		Data
			Results
			Implications
Episode 2	Subgoal	Experiment 2	Subquestion
			Subhypothesis
	Attempt		Method
	Outcome		Data
			Results
			Implications
Resolution	Outcome	Discussion	Results
	Event		Next steps

Table 2. Discourse analysis of (Louiseau, 2009) showing segmented text, verb form and semantic verb class (for details of the analyes see de Waard and Pander Maat, 2009 and 2010).

Elementary Discourse Unit	Segment type	Verb tense	Verb Class
Though D3 receptor antagonists can enhance cognitive function,	Fact	Present	Cause and effect
their sites of action remain unexplored.	Problem	Present	Cognition
This issue was addressed	Reg-Result	Past Perfect	Discourse verb
employing a model of social recognition in rats,	Method	Gerund	Procedure
and the actions of D3 antagonists were compared to D1 agonists	Method	Past Perfect	Procedure
that likewise possess pro-cognitive properties.	Fact	Present	Properties
Infusion of the highly selective D3 antagonists, S33084 and SB277,011 (0.04-2.5 µg/side), into the frontal cortex (FCX) dose-dependently reversed the deficit in recognition induced by a delay.	Result	Past	Cause and Effect
By contrast, the preferential D2 antagonist, L741,626 (0.63-5.0) had no effect.	Result	Past	Cause and effect
The action of S33084 was regionally specific	Result	Past	Change and Growth
inasmuch as its injection into the nucleus accumbens or striatum was ineffective.	Result	Past	Cause and effect
A similar increase of recognition was obtained upon injection of the D1 agonist, SKF81297 (0.04-0.63), into the FCX	Result	Past Perfect	Procedure
though it was also active (0.63) in the nucleus accumbens.	Result	Past	Cause and effect
These data suggest that	Reg-Implication	Present	Interpretation
D3 receptors modulating social recognition are localized in FCX,	Implication	Present	Properties
and underpin their pertinence as targets for antipsychotic agents.	Implication	Present	Interpretation

2.2 Paradigmatic Analysis

It seems that a syntagmatic analysis of scientific text is quite straightforward. So what would a paradigmatic analysis look like? In essence, Lévi-Strauss performs two actions: the first steps involves 'break[ing] down [a] story into the shortest possible sentences, and writing each such sentence on an index card bearing a number corresponding to the unfolding of the story.' That corresponds to a discourse parsing into smallest constituent units, also called elementary discourse units (edu's) (Marcu, 1999). Although index cards are no longer in fashion, I do apply a small taxonomy of segment types to elemental discourse units, and the segment types consist of functions of with specific subjects. See table 2 for an example of a fragment of this analysis.

Lévi-Strauss notes that 'a certain function is predicated to a given subject'; 'Or, to put it otherwise, each gross constituent unit will consist in a relation.' Lévi-Strauss' main claim is now that 'the true constituent units of a myth are not the isolated relations but bundles of such relations.' Is this also the case of scientific text?

In scientific discourse, the predicates studied are first of all verbs: their form (tense, aspect, mood etc.) (de Waard and Pander Maat, 2009), what semantic class they belong to (de Waard and Pander Maat, 2010), and secondly modality markers (e.g., hedging markers (possibly, certainly), modal auxiliaries (might, could) and 'suggests that'-type of constructions). Indeed, the results are similar to Lévi-Strauss' conclusions: we can group these segment types into broader categories. Specifically, it seems that there are three types of text in scientific papers: one type pertaining to conceptual claims and statements; one type pertaining to experimental methods and findings, and one type that contains the 'connecting text', intra- and intertextual elements of the type 'as we have shown' or 'see figure 2', or 'as Lendvai (2008) has indicated', on the one hand, and regulatory or connective segments such as 'these results suggest that', 'from this, it can be deduced' etc.

An attempt at a paradigmatic representation is given in Table 3, where we differentiate between conceptual and experimental discourse. Support for this split is given by the use of verb tense: there is a good correlation with past tense for experimental discourse, and present (modal or direct) forms for both conceptual discourse, and simple present tense for connecting discourse (de Waard and Pander Maat, 2009). A similar overlap with verb form exists (de Waard and Pander Maat, 2010).

Table 3. The segments from Table 2 ordered 'paradigmatically' by topic: concept/connection/experiment.

Conceptual discourse	Connecting discourse	Experimental discourse
Though D3 receptor antagonists can enhance cognitive function,		
their sites of action remain unexplored.		
	This issue was addressed	
		employing a model of social recognition in rats,
		and the actions of D3 antagonists were compared to D1 agonists
that likewise possess pro-cognitive properties.		
		Infusion of the highly selective D3 antagonists, S33084 and SB277,011 (0.04-2.5 μg/side), into the frontal cortex (FCX) dose-dependently reversed the deficit in recognition induced by a delay.
		By contrast, the preferential D2 antagonist, L741,626 (0.63-5.0) had no effect.
		The action of S33084 was regionally specific
		inasmuch as its injection into the nucleus accumbens or striatum was ineffective.
		A similar increase of recognition was obtained upon injection of the D1 agonist, SKF81297 (0.04-0.63), into the FCX
		though it was also active (0.63) in the nucleus accumbens.
	These data suggest that	
D3 receptors modulating social recognition are localized in FCX,		
and underpin their pertinence as targets for antipsychotic agents.		

2.3 Combination/contrast

If we combine these two analyses, we can paint a two-dimensional picture of scientific discourse, where syntagmatic and paradigmatic elements are depicted on orthogonal axes. For scientific text, the same dichotomous analyses can be done: the narrative form of subsequent sections of a paper has a parallel to story grammars or schema's, and the topics can be mapped to paradigmatic categories, as shown above. In Fig. 1, the topical vs. sequential axes represent the story grammar-like components versus the three realms discussed in 2.2.

However, there is a second difference between the Proppian and Lévi-Straussian analyses: Propp looks at supersentitial units of text (as do we in Figure 1), whereas Lévi-Strauss calls us 'break down [a] story into the shortest possible sentences' – that is, to divide it into clauses; which is what I attempt in my segment-type analysis. Figure 2 is a sketch of such a clause-level analysis in terms of discourse flow (x-axis) and in terms of subject type (y-axis), for the text in Table 3. This text is taken from the Introduction, and it is clear the coarser-grained annotation from Figure 1 is an inadequate representation of topics covered in the text: the finer grain is needed (see also Nawaz et al, 2010).

3. How can narrative analysis help scientific understanding?

From the previous, it seems that we can analyze scientific text using insights obtained from narrative analysis. How can this analysis can help improve access to scientific knowledge? Despite the obvious parallels between stories and scientific papers, there is no 'story grammar' defined for scientific papers, as such. Various publishers have different schema's for defining papers: see Table 4 for an overview of 4 publishers' subject headings. There have been past efforts to propose a modular structure for scientific papers (Kircz and Harmsze, 2000), and a proposal for a LaTeX-ased authoring tool for simple narrative scientific text, the abcde format (de Waard and Tel, 2006), but currently, there is no consensus between publishers to obtain a finer-grained, rhetorical structure markup, despite some efforts in this direction (e.g., Groza et al , 2009). Hopefully, utilizing similar steps as followed in defining the Proppian markup standards (Malec, 2004) and corresponding markup (Lendvai et al, 2010 a&b), data standards (Declerck et al., 2010) and ontologies (Peinado et al. 2004) could help establish a robust format for scientific narrative markup. Similarly, experiments in

automated markup of these structured fairytales such as AutoPropp (Malec, 2010) as well as other work in automate story generation (e.g. Callaway and Lester, 2002, and references therein), might help identify core elements in scientific text. It would be very interesting to combine tools focusing on this syntagmatic analysis with the perpendicular task of finding common patterns within scientific text; work on e.g. BioEvent extraction (Naw'az et al, 2010) points in this direction, and collaborations on this front at a more granular level might offer a promising way forward.

Table 4. Article sections and DTDs for several publishers.

Publisher/DTD URL	Sections
Biomed Central http://www.biomedcentral.com/xml/dtdtaggingspec.html	Background Results Discussion Conclusions Methods
Elsevier http://www.elsevier.com/framework_authors/DTDs/ja50_tagbytag5.pdf	Introduction Results Discussion Experimental Procedure
Nature Publishing Group **?**	Background Findings Discussion Methods
National Library of Medicine http://dtd.nlm.nih.gov/articleauthoring/tag-library/	Intro Results Discussion Conclusions Methods Cases Materials Subjects Supplementary-material
Society for Neuroscience **?**	Introduction Result Discussion Materials & Methods

4. References

[1] Bazerman, C. 1988. Shaping written knowledge: the genre and activity of the experimental article in science, Madison, Wisconsin: Univ. of Wisconsin Press, 1988.

[2] Callaway, Charles B., James C. Lester, Narrative prose generation, Artificial Intelligence, Volume 139, Issue 2, August 2002, Pages 213-252.

[3] de Waard, A., and Pandermaat, H. (2010b). A Classification of Research Verbs to Facilitate Discourse Segment Identification in Biological Text, Interdisciplinary Workshop on Verbs. The Identification and Representation of Verb Features, Pisa, Italy, November 4-5 2010.

[4] de Waard, A. (2010). From Proteins to Fairytales: Directions in Semantic Publishing. IEEE Intelligent Systems 25(2): 83-88 (2010)

[5] de Waard, A. and Pandermaat, H. (2010a), Categorizing Epistemic Segment Types in Biology Research Articles. Workshop on Linguistic and Psycholinguistic Approaches to Text Structuring (LPTS 2009), September 21-23 2009. – to be published as a chapter in Linguistic and Psycholinguistic Approaches to Text Structuring, Laure Sarda, Shirley Carter Thomas & Benjamin Fagard (eds), John Benjamins, (planned for 2010).

[6] de Waard, A., Simon Buckingham Shum, Annamaria Carusi, Jack Park, Matthias Samwald and Ágnes Sándor. (2009). Hypotheses, Evidence and Relationships: The HypER Approach for Representing Scientific Knowledge Claims, Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009), co-located with the 8th International Semantic Web Conference (ISWC-2009).

[7] de Waard, A. (2007). A Pragmatic Structure for the Research Article, in: Proceedings ICPW'07: 2nd International Conference on the Pragmatic Web, 22-23 Oct. 2007, Tilburg: NL. (Eds.) Buckingham Shum, S., Lind, M. and Weigand, H. Published in: ACM Digital Library & Open University ePrint 9275.

[8] de Waard, A. and Tel, G., (2006). The ABCDE Format: Enabling Semantic Conference Proceedings, In: Proceedings of the First Workshop on Semantic Wikis, European Semantic Web Conference (ESWC 2006), Budva, Montenegro, 2006.

[9] de Waard, A. Breure, L. Kircz, J.G. Oostendorp, H. van (2006). Modeling Rhetoric in Scientific Publications.

[10] Current Res. in Inf. Sci. and Techn. pp. 352-356, 2006.

[11] Declerck, Thierry Kerstin Eckart, Piroska Lendvai, Laurent Romary, Thomas Zastrow (2010). Towards a Standardized Linguistic Annotation of Fairy Tales, in: LREC 2010, Language Resource and Language Technology Standards – state of the art, emerging needs, and future developments.

[12] Groza, T., Handschuh, S. Clark,T., Buckingham Shum, S. and De Waard, A. (2009). A Short Survey of Discourse Representation Models, ISWC Workshop on Scientific Discourse Representation, 2009.

[13] Kircz J.G. and F.A.P. Harmsze (2000), Modular scenarios in the electronic age. Conferentie Informatiewetenschap 2000. Doelen, Rotterdam 5 april 2000. In: P. van der Vet en P. de Bra (eds.) CS-Report 00-20. Proceedings Conferentie Informatiewetenschap 2000. De Doelen Utrecht (sic), 5 april 2000. pp. 31-43.

[14] Latour, B. (1987). Science in Action, How to Follow Scientists and Engineers through Society, (Cambridge, Ma.: Harvard University Press, 1987)

[15] Latour, B., and Woolgar, S. (1979). Laboratory Life: The Social Construction of Scientific Facts. Beverly Hills: Sage, 1979.

[16] Lendvai, P, T. Declerck, S. Darányi, S. Malec (2010a). Propp revisited: integration of linguistic markup into structured content descriptors of tales. In Digital Humanities 2010. London, United Kingdom, Oxford University Press, July 2010.

[17] Lendvai, P., T. Declerck, S. Darányi, P. Gervás, R. Hervás, S. Malec, and F. Peinado (2010b). Integration of linguistic markup into semantic models of folk narratives: The fairy tale use case. In Proceedings of LREC, 2010.

[18] Lévi-Strauss, Claude. (1955). The Structural Study of Myth. The Journal of American Folklore, Vol. 68, No. 270, Myth: A Symposium (Oct. - Dec.,1955), pp. 428-444

[19] Loiseau, F., Millan, M.J. (2009). Blockade Of Dopa-mine D3 Receptors In Frontal Cortex, But Not In Sub-Cortical Structures, Enhances Social Recogni-tion In Rats. European Neuropsychopharmacology - January 2009 (Vol. 19, Issue 1, Pages 23-33).

[20] Malec, S. A. (2001). Proppian structural analysis and XML modeling. In Proceedings of CLiP, Duisburg, Germany, December 6-9, 2001.

[21] Malec, S.A. (2010). AutoPropp: Toward the Automatic Markup, Classification, and Annotation of Russian Magic Tales, This Workshop (*Amicus Workshop '10*, October 21, 2010, Vienna, Austria)

[22] Marcu, D. (1999) A decision-based approach to rhetorical parsing, Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, p.365-372, June 20-26, 1999, College Park, Maryland.

[23] Nawaz, R., Thompson, P., McNaught, J. Ananiadou, S. (2010). Meta-Knowledge Annotation of Bio-Events. In Proceedings of LREC 2010, pages 2498-2505.

[24] Peinado, Federico, Pablo Gerv´as, Bel´en D´ıaz-Agudo, (2004). A Description Logic Ontology for Fairy Tale Generation, In Fourth Int. Conf. on Language Resources and Evaluation: Workshop on Language Resources for Linguistic Creativity, 2004

[25] Propp, V. J. (1968). Morphology of the folktale. University of Texas Press: Austin, 1968. (Transl. L. Scott and L. A.Wagner).

[26] Rumelhart, D. 1975. Notes on a schema for stories. In D. Bobrow and A. Collins, editors, Representation and Understanding: Studies in Cognitive Science (New York: Academic Press, 1975)

[27] Thorndyke, P. W. 1977. Cognitive Structures in Comprehension and Memory of Narrative Discourse, Cognitive Psychology 9, 77-110 (1977)

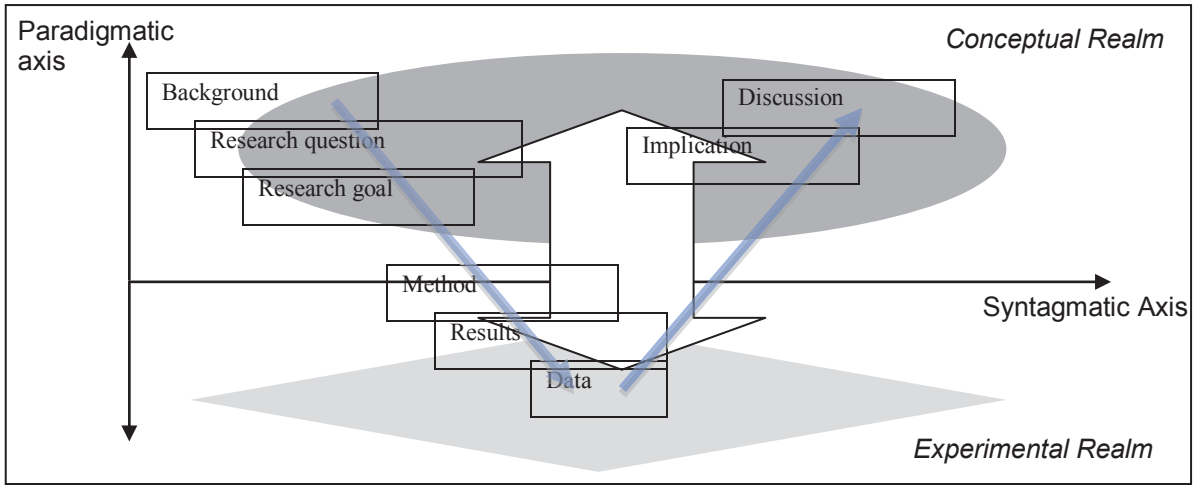

Figure 1. Two axes of analysis: on the x-axis, contiguous structural elements of a research paper; on the y-axis, the main conceptual categories that experimental research covers.

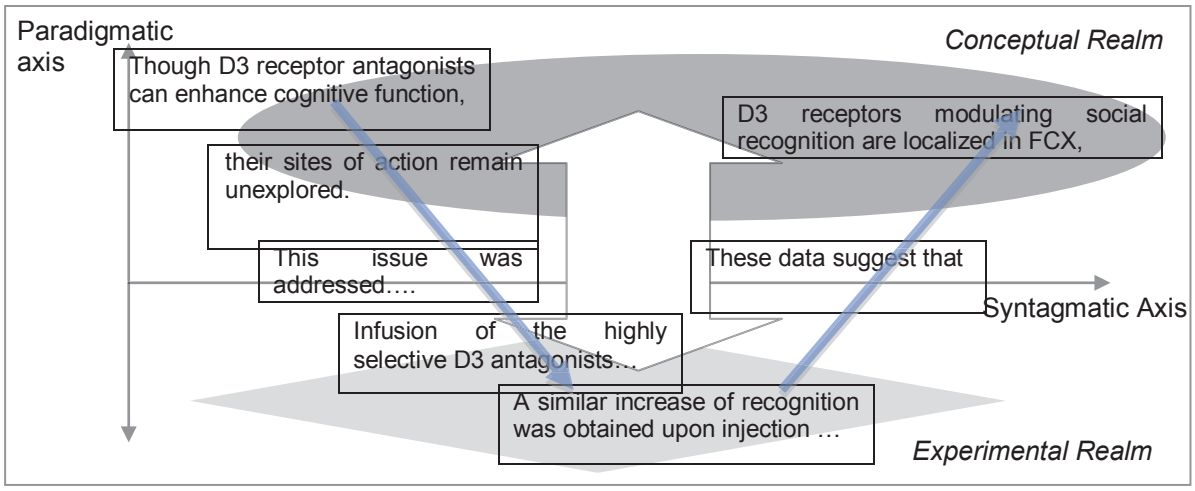

Figure 2. Similar axes as in Figure 1. Note finer-grained textual elements, showing that the move from concept to experiment and back to concept is a recursive pattern that occur at different granularities of discourse.

Improving Search through Event-based Biomedical Text Mining

Sophia Ananiadou
National Centre for Text Mining
University of Manchester
131 Princess Street, Manchester
M1 7DN, UK
+44 (0)161 306 3092
sophia.ananiadou@manchester.ac.uk

Paul Thompson
National Centre for Text Mining
University of Manchester
131 Princess Street, Manchester
M1 7DN, UK
+44 (0)161 306 3091
paul.thompson@manchester.ac.uk

Raheel Nawaz
School of Computer Science
University of Manchester
131 Princess Street, Manchester
M1 7DN, UK
+44 (0)161 306 3091
nawazr@cs.man.ac.uk

ABSTRACT

Recently, there has been a major focus on event extraction for biomedical applications. In this paper, we focus on search, highlight some of the drawbacks of popular search methods, and show how event extraction and associated technologies, e.g., named entity recognition, can help to improve the efficiency of search. We also explore how event extraction can be enhanced through a new type of annotation, i.e. meta-knowledge annotation, which can facilitate the extraction of high-level information relating to the intended interpretation of events, e.g. whether they represent a hypothesis, a claim, a belief, an opinion, a well established fact, a tentative or more confident analysis of experimental results, etc.

1. BACKGROUND

The amount of biomedical literature is increasing at a rapid rate, with the size of PubMed increasing at the rate of approximately 2 papers per minute [17]. As a result, it is becoming increasingly difficult for biologists to locate information relevant to their research contained within textual documents.

The goal of searching the literature is to find relevant pieces of *knowledge* (e.g., biological processes).. Suppose that a biologist is interested in discovering which proteins are *positively regulated* by the protein IL-2. An example of the type of sentence she wishes to locate is the following:

IL-2 activates p21ras proteins in normal human T lymphocytes.

This sentence allows the biologist to discover that *p21ras proteins* are one type of protein to satisfy her query. To locate such sentences of interest using an ordinary search engine, the biologist may enter the search terms *IL-2* and *activates*. Such a query will, however, return a large number of documents. On the one hand, many of the documents are likely to be irrelevant to the user's query. On the other hand, the query also has a high probability of missing documents that are relevant to the user's requirements. The reasons for these problems include:

First International AMICUS Workshop, October 21, 2010, Vienna, Austria.

- **Specificity of the query –** The query will over-generate (i.e. return too many documents) because it cannot convey the biologist's specific requirements. In terms of semantics, the user wishes to retrieve documents conveying sentences that describe positive regulations, where *IL-2* is the instigator of the regulation. Such knowledge is normally expressed according to the syntactic structure of the sentence. In the case of the verb *activate,* for example, the instigator corresponds to the grammatical subject. Hence, the user would only be interested in sentences where *IL-2* is the grammatical subject of *activate.* In ordinary search engines, however, it is not possible to specify how search terms should be related to each other. This is because the search engine simply sees documents as "bags of words" that have no internal structure.
- **Term variation/ambiguity –** Terms in biomedicine are complex; they include an enormous amount of synonyms and different variant term forms are used in the literature [21]. For example, *IL-2* can also appear without a hyphen (*IL 2)*. These terms are themselves acronyms for the longer forms *interleukin 2* and *interleukin-2.* The term *T-cell growth factor* can be considered as a synonym of *interleukin-2*, but this also has its own variant forms (e.g. *TCGF*). Thus, using *IL-2* as a search term without considering its variants would result in many relevant documents being overlooked. For query formulation, we need techniques which include term variation. In addition, many terms and their variants are ambiguous, as they share lexical representations either with common English words (e.g. an, by, cat, can) which also denote gene/protein names, or with other biomedical terms [7]. A further issue concerns terms that can also be ambiguous between biological and general English senses, e.g., the proteins named *cat* and *met.* Using such common words as search terms will return many more irrelevant documents than relevant ones, if only documents containing the protein names are sought.
- **Different ways of expressing knowledge –** The verb *activate* is only one of the ways in which positive regulations can be expressed in texts. Other verbs could also be used, e.g., *stimulate* or *affect*, whilst nouns (nominalised verbs) convey a similar meaning, e.g., *activation, effect, stimulation* could also be substituted. As with term variation, it can be difficult to enumerate all the ways in which a particular type of biological process can be expressed. However, if they are not accounted for in a query, then relevant documents may be missed.

The application of text mining methods[2, 3, 52] can provide solutions to the problems outlined above, and can thus contribute to more efficient and effective search solutions for biologists. Text mining techniques can help to ensure that a greater number of relevant search results are obtained, whilst helping to exclude those results that are irrelevant to the user's query.

The remainder of this paper will focus on a number of these methods, and explain how they can improve search results. We firstly look at named entity recognition (NER)[1], and we will examine search engines which offer advanced search capabilities based on semantic metadata derived from named entities and relations. .

NER is one of the technologies that is required to perform extraction and querying of *events* [4]. Events are structured, semantic representations of pieces of knowledge contained within text. We focus on relations [39] within the biomedical domain, such as descriptions of positive regulation, transcription, gene expression, [6] etc. Through a combination of a number of techniques, such as deep syntactic parsing [26] and NER, event extraction automatically locates events in texts and identifies their individual participants, e.g., the instigator of the event, the location of the event, etc. This allows users to formulate more structured queries that are better related to their actual needs, e.g., the biologist can request the system to return only those documents that mention a positive regulation event, where *IL-2* is the instigator. The system will retrieve only those documents where the specified relationship exists between the search terms.

Following a detailed examination of how events are extracted and can be used in advanced searching, we conclude by examining a new direction of research, i.e., how interpretative information about events can be captured automatically to further enhance event-based searching. For example, an event may represent a generally accepted fact, a hypothesis, an experimental observation, a tentative analysis of experimental results, etc. These different types of information could be important to the biologist, e.g., some biologists may be interested only in retrieving events that correspond to "reliable" pieces of knowledge, rather than hypotheses or hedged interpretations. This is particularly important for maintaining curated databases of biological knowledge [5]. Other biologists may be interested in matching up hypotheses with proven results.

2. NAMED ENTITY RECOGNITION

A large amount of work has been carried out on the automatic recognition of biologically relevant NEs in texts [39]. This activity is important for a number of reasons:

- It can resolve ambiguities between words used in general language and those that represent biomedical entities (e.g. *cat*)
- It can facilitate mapping terms found in texts to entries in curated biological databases, such as UniProt[46] and Entrez Gene. (http://www.ncbi.nlm.nih.gov/gene), or resources such as the BioThesaurus [22, 51]. This can facilitate direct access from search results to detailed information about biological entities found within the database.
- Automatic highlighting of different recognised NEs in retrieved documents can facilitate a quick skimming of the main content of the document.
- NEs map to event participants, e.g. a cell group (filament) participates in a localisation event in angiogenesis. Hence, NER is a necessary pre-processing step in event extraction.

Given the huge amount of variation of terms in biomedical text [38], work has also been carried out on recognising and resolving several types of variations. Although some variations are listed in curated databases, many are missing. However, it is important that even unlisted term variations can be resolved to the entity that they describe via term normalisation [45], to facilitate their linking with the correct biological database entry. Work on term normalisation was recognised as an important task through its incorporation as a BioCreative task [15]; term normalisation also includes the recognition of acronyms [30], and the use of soft-string matching techniques [45] that recognise new variants of known terms.

2.1 Search Engines Incorporating NER

Conventional information retrieval technology, while very good at handling large scale collections, remains at a rough granular level. Semantic metadata generated from named entities (NEs) (e.g. PROTEIN:IL-1, ORGAN:brain) are helpful for increasing granularity of document search. However, conventional information retrieval systems do not allow users to specify in their query the semantic metadata they are interested in. Lack of such functionality restricts users' potential to search and retrieve documents based on their personal and social profiles. Metadata are critical to enhancing user experience of search: for example, they can support improved personalized search. The richer the metadata, and the more they are linked in to other resources of different types (including e.g., experimental data), the better the search experience. NLM's PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) is a primary search facility for biomedical literature. Other search interfaces based on PubMed focus on ranking citations [40], incorporating external web services [10] or using Web 2.0 technologies [27] to enhance user experience in searching. In CiteXplore (2009), text mining results are included in the search based on Whatizit, EBIMed [34, 35] and iHOP [16] such as protein/gene annotations and protein-protein interactions. KLEIO [29] is an intelligent search engine which provides interactive faceted semantic search over MEDLINE based on NEs. Faceted navigation is proposed as a component of a superior interface for searching metadata in a more interactive and flexible manner [12, 13] and has become adopted on several web sites (e.g., http://express.ebay.com). One of the main criticisms of the conventional search systems freely available is that search queries are effective only when well crafted [11]. In KLEIO, the user can select, using an interactive faceted query builder, the types of semantic queries of interest, suggested by the system. KLEIO delivers rapid responses, based on pre-indexed NEs linked with term variants, includes query expansion with dynamic reclassification of results, linking of all NEs with unique identifiers from a variety of databases, and highlighting of the retrieved documents with the NEs identified. It also integrates term normalisation[44] and links to curated databases [45] to facilitate more focussed searches. KLEIO permits operators corresponding to different types of NEs as an integral part of the user's search. Thus, it is possible to specify

PROTEIN:cat, to ensure that only documents containing an instance of the word *cat* that have been recognised as an NE of type *protein* will be returned by the search. This can considerably reduce the number of documents returned: performing a search only for *cat* in KLEIO retrieves 67159 abstracts, whilst the search *PROTEIN:cat* reduces this number to only 195 abstracts. KLEIO also highlights instances of NEs found in documents, identifies and resolves acronyms to their full forms (through the integration of AcroMine [31] and additionally allows searches to be expanded to include variant terms through soft-string matching and database linking.

Whilst these search engines provide improvements and useful functionalities not found in traditional search engines, they still do not facilitate the precise querying of events. Searches carried out in KLEIO work along similar principles to traditional search engines. So, although NE operators can help to refine the scope of the search, it is still not possible to specify relationships between these terms.

3. EVENT EXTRACTION

Although NER and normalization have been helpful for increasing the specificity of document searches in TM systems such as KLEIO (www.nactem.ac.uk/software/kleio/) and for significantly reducing errors compared with simple keyword-based retrieval, other search systems, such as FACTA [15], use co-occurrence statistics for normalized names in text to enhance the discovery of hidden associations among entities. However, textual co-occurrence of entities does not necessarily indicate meaningful relationships. For this, more advanced analytical methods are necessary, namely methods that undertake deeper semantic analysis. To achieve this aim, techniques have been developed that automatically extract biological events[4]. Recognition of events and their participants is often reliant on a structural analysis of the sentence containing the event [26]. As described above, event participants are often organised around either a verb or a nominalised verb (e.g., *activation*). In this case, event participants normally constitute some or all of the words and phrases that are syntactically (i.e., structurally) related to the verb or nominalised verb in question, especially if these phrases constitute or contain NEs. Thus, full or partial syntactic parsing of text, which provides a structural analysis of a sentence, is normally one of the necessary steps of event extraction, in addition to NER.

MEDIE [25] is a search engine for MEDLINE abstracts that combines many of the features found in KLEIO (e.g. recognition of NEs, linking with databases and ontology and subsequent identification of term variants) with a further annotation processing step that involves the structural analysis of the abstracts.. The NLP modules used for the annotation include (but not limited to) a deep syntactic analyzer, an event expression recognizer and a term recognizer. The syntactic analyzer, Enju parser [26], produces a syntactic and semantic analysis of the text, based on the linguistic formalism of HPSG. A relational concept, such as 'protein A activates protein B', can be precisely described as a query which specifies the semantic structure given by the Enju parser as a set of constraints. This is the main strength of MEDIE compared to other publicly available TM modules which use Boolean formulas of keywords or concepts for query formulation. Boolean formulas basically specify co-occurrence of concepts or words as a constraint for retrieval. One can only specify co-occurrence of protein A, protein B and the verb 'to activate' in the same textual unit (usually an abstract) as a constraint, which results in a large number of false positives. Units of retrieval in MEDIE are finer than those in other TM modules. They can be individual sentences in abstracts, or even phrases. MEDIE accepts a search query through an API, in addition to an interactive search UI. The API takes a tuple of <subject, verb, object> as the input, which describes a biological event/relation, such as <p53, activate, beta-4>, and returns a set of articles in which the event/relation is mentioned.

A further feature of MEDIE is that search results are not only limited to those where the value of the *verb* slot corresponds to *activate*. Rather, through reference to the Gene Ontology [5], events centred on other verbs such as *stimulate, induce, augment, enhance*, etc. are also retrieved by the search. Due to the deep parsing technology used, the template to be completed by the user can be considered as an abstract representation of the way that the events actually manifest themselves in the text. For example, *IL-2* would also be identified as the subject of *activate* in a passive sentence such as *p21ras proteins are activated by IL-2*. Additional features include the ability to specify the types of sentences in which the search should be conducted, e.g. *conclusion, method, result,* etc.

3.1 Additional Event Participants and Semantic Representations

Despite its clear advantages over a traditional search engine, MEDIE still presents some limitations. Firstly, events often have more than two participants. In biomedical texts in particular, information corresponding to locations, time, environmental conditions and manner is considered to be highly important to their correct interpretation [43]. It would be useful to be able to identify these types of information separately, in order to allow restrictions to be placed on their values as part of a search, and also to allow them to be displayed as part of the search results.

Secondly, the search template to be filled is closely tied to the syntactic structure of the text. Searching using a higher level semantic representation would, however, be preferable. In the search problem introduced earlier, we wanted to find events where *IL-2* is the instigator of the positive regulation event. Instigators can also be identified in many other types of events, and so we can assign this type of event participant a general *semantic role* label that will be common across many different types of events, i.e., AGENT. Likewise, most events also specify as a participant the thing that is affected by or during the event, e.g., the protein undergoing positive regulation. The semantic role label that is normally used for such participants is THEME.

AGENT and THEME frequently correspond to the grammatical subject and object of verb, respectively. However, this is not always the case, and there is no consistent correspondence between grammatical positions and semantic roles. Thus, using AGENT and THEME rather than subject and object would allow the event search template to be more general and less tied to syntactic structure of the text. A semantic approach is even more desirable if additional participants (e.g. location, environmental conditions, etc.) are specified as part of the search. Several of these participant types are specified through syntactically similar means, i.e., through the use of prepositional or adverbial phrases [47]. Consider the following example:

A promoter has been identified that directs relA gene transcription towards the pyrG gene in a counterclockwise direction on the E. Coli chromosome

In addition to a subject and an object, the verb *directs* occurs with 3 arguments corresponding to prepositional phrases, each of which corresponds to a different semantic role (namely DESTINATION, MANNER and LOCATION). Although the different prepositions can be used to help in distinguishing between different semantic roles, there is not a one-to-one mapping between prepositions and semantic roles. For example, *in* is used in the above example to introduce a MANNER, but it could equally introduce a LOCATION, e.g., *in E. coli.* By allowing search criteria to include semantic role labels such as LOCATION and MANNER, the user could specify semantically precise search criteria without having to worry about their exact form in the text (e.g., which preposition is used, etc.)

Considering the above, the type of semantic representation that would ideally be produced for the positive regulation event in the sentence *IL-2 activates p21ras proteins in normal human T lymphocytes* is as follows:

EVENT_TYPE: *positive_regulation*

AGENT: *IL-2:PROTEIN*

THEME: *p21ras proteins:PROTEIN*

LOCATION: *in normal human T lymphocytes:CELL*

In the above representation, the event has been assigned a semantic type, i.e. *positive_regulation*. This event type is a label selected from a fixed, ontological set of relevant event types. Others would include *binding, gene_expression,* etc. Additionally, each participant of the event has been separately identified and assigned a semantic role. The NEs within each participant have also been identified and assigned appropriate NE types. Such a representation allows structured searches to be performed with the following types of criteria:

- Ontological classes of events as an alternative to specifying particular verbs.
- Specifications of the participants that should be present in the event (in terms of semantic roles)
- Restrictions on the values of particular participants. These restrictions could take the form of actual entities (e.g. *NF-kappa B*), NE classes (e.g. *PROTEIN*), or a combination of both, in a similar way to KLEIO.

The main challenges of producing a system that can produce such a representation of events are the following:

1) How each type of event manifests itself in the text - they are often organised around a particular set of verbs and nominalised verbs.
2) How syntactically related arguments of the verb/nominalised verb map to semantic roles.

Although grammatical parsers such as Enju [24] have reached an appropriately mature level, the same cannot be said for semantic parsers. This means that the mapping between syntactic and semantic representations is not straightforward. This is complicated by the fact that different verbs behave in idiosyncratic ways, with different numbers of syntactic arguments, which can map in different ways to semantic roles.

3.2 Annotated Corpora

The approaches used to map between the syntactic and semantic levels can vary in a number of ways, both in the method used (rule based vs. machine learning approaches) and the types of external resources employed (either lexical or ontological).

Whether a rule-based or machine learning approach is taken, annotated corpora of events are a vital resource for the development of event extraction systems. These corpora provide direct evidence of how events manifest themselves in texts, and as such, they can be used in both the development/training of event extraction systems, as well as in the evaluation of the performance of such systems, by acting as a "gold standard" [14].

The various event corpora that have been produced in the biomedical field generally have in common that they identify an "anchor" expression (e.g., a verb or nominalised verb) around which the event is organised. Event participants are then individually identified and linked to this anchor expression. BioInfer [33] (1100 sentences) concentrates on identifying core participants (i.e., agent or theme type roles), although they are not labelled with semantic roles. Events are, however classified according to an ontology, thus facilitating the discovery of the ways in which different types of events can be expressed in the text

A similar type of event classification is carried out in the GENIA event corpus [19]. This is a larger corpus, consisting of 1000 MEDLINE abstracts, containing over 9000 sentences. In the GENIA corpus, event participants are classified using semantic roles. Although the focus is on identifying the THEME and CAUSE (similar to AGENT) roles, 3 other types of event participants are also identified and labelled, i.e., location, time and experimental methods.

GREC [41] is a smaller corpus of 240 abstracts, but with a richer type of semantic annotation that is focussed on gene regulation and expression events that are described by verbs and nominalised verbs. For each event, all participants (arguments) in the same sentence are identified and assigned a semantic role from a rich set of 13 roles, tailored to biomedical research articles.

Although GREC is a relatively small annotated corpus compared to GENIA, a recent study [23] has shown that combining smaller, richly annotated corpora with larger corpora that are slightly poorer in information content can help to improve the performance of event extraction system. Whilst the benefits of combining disparate sources in machine learning are well known, this idea is especially attractive, given that the production of large, richly annotated corpora can be very time-consuming.

3.3 Lexical Resources as an Aid to Event Extraction: the BioLexicon

Whilst event extraction systems can be trained based on annotated corpora alone, the use of a computational lexical resource with information about the syntactic and semantic behavioural of verbs within the domain can be used to boost the performance of such systems.

Although syntactic parsers can be used to identify the core arguments of a verb, the idiosyncratic behaviours of individual

verbs within the biomedical domain means that determining which of the modifier phrases (i.e., prepositional and adverbial phrases) should be treated as arguments of the verb (and hence as event participants) can be problematic. As mentioned above, such phrases can correspond to vital pieces of information about the event, such as locations, manners, conditions, etc. By accessing information about typical patterns of syntactic behaviour for individual verbs, we can expect that the event extraction system will do a better job of determining which phrases in the sentence correspond to event participants. This is particularly important in the case of sentences that contain multiple events, in order to determine which phrases are participants of which event. In the following example, each underlined verb corresponds to a different event, with different sets of participants:

IHF may inhibit ompF transcription by altering how OmpR interacts with the ompF promoter

After the event participants have been identified, lexical resources can also help by providing verb-specific mappings from the syntactic arguments to appropriate semantic roles.

Extensive computational lexicons have been constructed for use in processing general English texts (e.g. [20, 32, 36]), but these are not suitable for processing biomedical texts for a number of reasons. Firstly, there are many verbs that are domain specific (e.g., *methylate, phospholylate,* etc.). Other verbs appear in both domains (e.g. *activate*), but are likely to have different behaviours. In general, verbs in the general language domain have fewer arguments, largely due to the fact that modifier phrases are often considered to be less tightly associated with the verb than in biomedical texts.

Until recently, an extensive computational lexical resource comparable to those produced for the general English language domain was not available for the biomedical domain. Resources that had been built were either very small [9, 49] or did not contain semantic information [8].

The BioLexicon [37] is a reusable lexical and conceptual resource suitable for advanced biomedical text mining. One of its defining features is to include a wide range of biomedical terms and variant forms, to facilitate accurate NER in a range of biomedical text mining applications. Integration of the BioLexicon within a biomedical search engine would, for example, allow search terms entered in user's queries to be expanded with their known variants, to ensure that a greater number of relevant documents are retrieved

The BioLexicon gathers together terms and their variants from a number of different curated databases and ontologies into a single unified resource. The original database identifiers for each term are preserved, in order to facilitate linking to information in the source databases. A particular innovation of the BioLexicon is the application of text mining methods to recognise new variants of gene and protein names that appear in biomedical abstracts (MEDLINE) but not in existing databases. Genes and proteins tend to exhibit the greatest amount of variation amongst all types of biomedical entities. Application of an NER method was followed by application of the soft-string matching technique to map newly discovered NEs to the most similar existing terms. This method discovered and mapped approximately 70000 new term variants. A further innovation of the BioLexicon is the inclusion of detailed information regarding the behaviour of verbs, which can leverage extraction of events.

In addition to including an extensive repository of biomedical terms and their arguments, the BioLexicon additionally incorporates detailed about typical syntactic and semantic patterns for domain-specific verbs, which is based on observable behaviour extracted from a corpus of biomedical texts [47].

The BioLexicon contains syntactic information (subcategorization frames) for 658 verbs, which were manually selected based on their particular relevance within the biomedical domain. For each verb, grammatical argument patterns (including modifier phrases) were extracted, based on the application of the Enju parser to a domain-specific corpus consisting of both biomedical abstracts on the subject of E. coli and full papers, totalling approximately 6 million words. Although modifier phrases (prepositional phrases and adverbials) are important in biomedical texts, they should only be considered to be arguments of the verb if there is sufficient evidence for this. Thus, a filter was used to ensure that rarely used patterns were not included in the lexicon. As each verb can occur with multiple patterns of syntactic arguments, a total of 1760 syntactic frames were extracted.

Semantic information about verbs was acquired based on a corpus of 677 abstracts that were manually annotated with events by domain experts, using a scheme almost identical to the one used for GREC, with the same 13 types of semantic roles [42]. The only difference is that whilst GREC is annotated with event instances, this second corpus was annotated with the specific purpose of extracting event frames to include within the BioLexicon.

Each extracted event frame was centred on a particular verb or nominalised verb. The subset of these frames that were centred on verbs for which grammatical information had been acquired was selected. This subset consisted of a total of 856 frames, centred on 168 verbs. A manual process was then used to link each argument in the syntactic subcategorisation frame its corresponding argument in the semantic frames. This resulted in 668 linked frames.

4. EVALUATION OF EVENT EXTRACTION

4.1 BioNLP'09 Shared Task

The development of event extraction systems that can reliably extract complex events involving multiple participants is an open research topic. However, the importance placed upon the development of such systems, and the desire of the community to push forward in this area have been demonstrated through the BioNLP'09 shared task [18]. Shared tasks involve teams from the community competing to analyze the same data within a common evaluation framework. They provide standard development and evaluation benchmarks, focusing the attention of the research community on timely issues and acting as a driver for the specification of new tasks and challenges. The BioNLP'09 shared task was the first to focus specifically on event extraction, which was based on protein biology event types.

The shared task evaluated the performance of systems not only in extracting primary event participants (i.e. THEME and CAUSE) but also secondary participants, including the source

and destination of the event. The results of the shared tasked showed that, although simple events can be extracted quite reliably using state of the art methods, more complex events involving multiple participants can currently only be extracted with less than 50% accuracy.

4.2 Evaluating the BioLexicon for Event Extraction

The BioLexicon has been evaluated within a challenging context, namely that of full parsing as part of the UKPubMedCentral (UKPMC) text mining services (http://ukpmc.ac.uk/), to locate and extract facts related to the biology domain. In practice, there are three components in the fact extraction process. Firstly, syntactic arguments of verbs in the texts are located through the application of the Enju parser to the texts. Only those verbs that are included in the BioLexicon are considered as potential textual "anchors" of events. These candidate events are further narrowed down by selecting only those in which an NE relevant to the domain appears in one of the arguments associated with the verb. As a final test, the syntactic argument pattern of the verb should be as predicted in the BioLexicon.

Whilst the primary use of the BioLexicon information in this context is as a filter, it also has a boosting effect on the range of facts to be considered. This is because modifier phrases (e.g., those which begin with prepositions) are explored, which would not be considered without its input. Where these modifier phrases contain recognised named entities, this can provide enough evidence for the extraction of a fact that would not otherwise be recorded. Consider the following example:

The pXPC3 plasmid codes for an XPC cDNA that is truncated by 160 bp from the N terminus compared with the wild-type XPC cDNA

Although the Enju parse result treats *code* as an intransitive verb (i.e. without a grammatical object), the information present in the BioLexicon allows the THEME role to be assigned to the prepositional phrase beginning with *for*.

The method described above has been evaluated through application to a test set of approximately 80,000 documents. Within these documents, only 62.7% of the instances of the verbs match verbal entries in the BioLexicon, thus illustrating its initial filtering effect. A still stronger filter is the requirement that a domain relevant NE should be present in one of the arguments. Applying this constraint results in only 16.9% of the total number of verb instances present in the text collection being extracted as facts. The experimental results also demonstrate, at least to some extent, the boosting effect achieved by using the verbal information in the BioLexicon, in that 9.7% of verb arguments are detected in prepositional modifier phrases, rather than in the arguments initially predicted by the parser output. These preliminary results provide compelling evidence that the BioLexicon can assist in building powerful tools for fact extraction within the biomedical domain.

5. EVENT INTERPRETATION

Although a large amount of work has been carried out on building resources and tools to facilitate extraction of events from biomedical texts, less attention has been paid to the way in which the extracted events should be interpreted. In addition to the event participants themselves, there is frequently additional information (or *meta-knowledge*) present within the context of the event that is vital to its correct interpretation. Examples of meta-knowledge include the type of evidence behind the event (e,g., does it represent a hypothesis, a well-established fact, etc.), whether there is any speculation expressed about the event, whether it is negated, etc.

Meta-knowledge can be expressed in text in a number of different ways. In the majority of cases, this is through the presence of particular "clue" words or phrases, although other features can also come into play, such as the tense of the verb on which the event is centred, or the relative position of the event within the text.

5.1 Expression of Meta-Knowledge

To make the idea of meta-knowledge more concrete, consider Figure 1, which shows a set of eight simple sentences. Two bio-events occur in these sentences. Event E1 represents the expression of an arbitrary gene *X,* whilst event E2 represents the positive regulation of E1 by an arbitrary protein *Y*. Figure 2 shows the typical structured representation of these events.

(S1) *We **found** that Y activates the expression of X*

(S2) *We **examined** the effect of Y on expression of X*

(S3) *These results **suggest** that Y has **no** effect on expression of X*

(S4) *Y is **known** to increase expression of X*

(S5) *Addition of Y **slightly** increased the expression of X*

(S6) *These results **suggest** that Y **might** affect the expression of X*

(S7) ***Significant** expression of X was **observed***

(S8) ***Previous studies** have **shown** that Y activates the expression of X*

Figure 1 – Simple Sentences

EVENT-ID:	E1
EVENT-TYPE:	gene_expression
THEME:	*X :* gene

EVENT-ID:	E2
EVENT-TYPE:	positive_regulation
THEME:	*E1:* event
CAUSE:	*Y*: protein

Figure 2 – Structured Representation of E1 and E2

The event trigger words are underlined in each of the examples. The *expression* event (E1) is always indicated by the nominalised verb *expression*. However, the *positive regulation* event (E2) is expressed in a number of different ways, namely using the verbs *activate*, *increase* and *affect*, or the nominalised verb *effect.* Although each example sentence contains an instance of one or both of the same bio-events (E1 and E2), their interpretations vary according to the sentential context. More importantly, without the annotation of meta-knowledge information, the events extracted from each sentence would be

identical, and the differences in meaning expressed within the sentential context would be lost.

The emboldened words and phrases in the example sentences help to show that the way in which the events should be interpreted can vary considerably. Most of the emboldened words affect the interpretation of the event E2, which is the main event in the sentence. However, in (S7) the interpretation of E1 is altered.

Sentences (S1), (S5), (S7), and (S8) all describe experimental observations. In most of these, the presence of a particular word (i.e., *found*, *shown* and *observed*) marks the E2 positive regulation as being an observation. In (S5), however, it the use of the past tense on the word on which the positive regulation event is centred (i.e., *increased*) that marks it as an observation.

Although all 4 events mentioned above represent observations, each of their interpretations is still slightly different. The different between (S1) and (S8) is the source of the information. The presence of the word *we* in (S1) indicates an observation as part of the current study, whilst in (S8), *previous studies* denotes an observation originally reported outside of the current paper. Thus, in (S1), the positive regulation can be considered as "new" knowledge, but (S8), the knowledge reported is "old". Whether such a difference is important will depend on the task being undertaken by the user. For example, database curators looking only for new knowledge might only be interested in (S1).

In (S5) and (S7) the difference in interpretations concerns event intensity, through the words *slightly* (i.e., low intensity) and *significant* (i.e., high intensity), respectively. The recognition of such information about events may be important, for example, when performing a comparison of different experimental methods. In (S5), the intensity applies to E2, whilst in (S7), the intensity applies to E1, as this is the only event that appears in the sentence.

The positive regulation event is (S4) can be taken as a well-established fact within the field, according to the presence of the word *known*. In a system that is looking for contradictions, events that contradict this well-established fact are potentially more serious than, say, a contradiction of new experimental outcomes (e.g., (S1)), which could later be disputed by other experts within the field.

All the events described above can be seen as reporting factual information. In this respect, (S2) is quite different. The presence of the word *examined* serves to indicate that the positive regulation event is under examination, and so it is not known whether or not it is true.

Sentences (S3) and (S6) should also not be considered as facts. Rather, the presence of the word *suggests* denotes that E2 is being stated as a somewhat tentative analysis of results on the part of the author. In (S6), the author uses the word *might* to increase the amount of speculation about the truth of the event. In (S3), the conclusion is different: the author concludes is that the positive regulation event is unlikely to happen, indicated by the use of the word *no*. Hence, this is a negative event.

From the above sentences, it is possible to isolate at least five important pieces of contextual information which can be regularly identified about events, which somehow modify their default interpretation:

1) What kind of evidence is there for the event, e.g. has it been experimentally observed, inferred from experimental results, is a well established fact, or is it a hypothesis whose truth has yet to be determined?
2) How certain is the author about whether the event is true?
3) Is the event positive, or is it negated (through the use of *no, not* etc.)
4) What is the intensity or magnitude of the event?
5) What is the source of the information contained within the event? Is it reported in the current paper or another paper?

5.2 Meta-Knowledge Annotation of Bio-Events

Existing event annotated corpora within the biomedical domain contain few annotations that relate to their interpretation. Negations are annotated in BioInfer and GENIA. Three different levels of certainty are also annotated for GENIA events. However, negation and speculation clue words are not annotated in these corpora. Negation and speculation were also addressed in one of the subtasks of the BioNLP'09 shared task, but in a fairly basic way. The only requirement was to recognise whether events were negated and/or contained expressions of speculation, without having to identify, e.g. the level of speculation. Only 6 out of the 24 participating teams attempted this task and the highest accuracy was around 25%. This was attributed to the lack of annotated clue phases in the training corpus [18].

More extensive interpretation-focussed annotation has been carried out within the domain at either the sentence level (e.g., [48]) or sentence-fragment level (e.g., [50]). However, these annotations cannot be used straightforwardly to assign interpretations to bio-events. Often, a sentence will contain several bio-events (e.g. both an experimental method *and* the results of applying this method), each of which has a different interpretation. If an expression of speculation is present (e.g. the word *might*), this may affect only certain events in a sentence.

Based on the above, we have designed a multi-dimensional annotation scheme to capture various aspects of meta-knowledge expressed for bio-events [28]. Our scheme is intended to be general enough to allow integration with different bio-event annotation schemes, whilst being detailed enough to capture important subtleties in the nature of the meta-knowledge expressed about the event, which may be important according to the task being undertaken by the biologist.

The annotation task consists of assigning an appropriate value for each dimension, as well as marking the textual evidence for this assignment. This latter part of the task is important to train systems to perform meta-knowledge identification successfully, given the difficulties faced in the negation/speculation part of the BioNLP'09 shared task, where such annotations were not present in the training data.

The advantage of using a multi-dimensional scheme is that the interplay between different values of each dimension can reveal both subtle and substantial differences in the types of meta-knowledge expressed in the surrounding text. This aspect of our scheme is further discussed in section 5.2.1.

Figure 3 provides an overview of the annotation scheme. The boxes with the light-coloured background correspond to

information that is common to most bio-event annotation schemes, whilst the boxes with the darker backgrounds correspond to our proposed meta-knowledge annotation dimensions and their possible values. Below, we provide brief details of each annotation dimension.

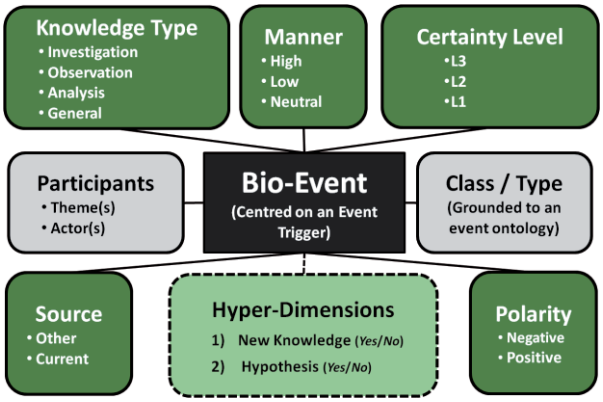

Figure 3 - Bio-Event Annotation

Knowledge Type (KT): Captures the general information content of the event. Each event is classified as either: *Investigation* (enquiries and examinations etc.), *Observation* (direct experimental observations), *Analysis* (inferences, interpretations and conjectures etc.) or *General* (facts, processes, states or methodology)

Certainty Level (CL): Encodes the confidence or certainty level ascribed to the event in the given text. We partition the epistemic scale into three distinct levels: *L3* (no expression of uncertainty), *L2* (high confidence or slight speculation) and *L1* (low confidence or considerable speculation).

Polarity: Identifies negated events. We define negation as the absence or non-existence of an entity or a process.

Manner: Captures information about the rate, level, strength or intensity of the event, using three values: *High* (increase in rate/intensity), *Low* (decrease in rate/intensity) or *Neutral* (no indication of rate/intensity).

Source: Encodes the source of the knowledge being expressed by the event as *Current* (the current document) or *Other* (any other source)

5.2.1 Hyper-Dimensions

A defining feature of our annotation scheme is that additional information (hyper-dimensions) can be inferred by considering combinations of some of the explicitly annotated dimensions. These are as follows:

New Knowledge: A combination of the values of *Source, KT* and *CL* dimensions can be used to isolate those events representing new knowledge. Specifically, new knowledge corresponds to events with a *KT* value of *Observation* or *Analysis* carried out as part of the current study (i.e., *Source=Current*). If *KT=Analysis*, then the event should only be classed as new knowledge if it represents a straightforward interpretation of results (i.e. *CL=L3*), rather than something more speculative.

Hypothesis: Events that represent hypotheses can be isolated by considering *KT* and *CL* values. Events with a *KT* value of *Investigation* can always be assumed to be a hypothesis. However, if the *KT* value is *Analysis,* then only those events with a *CL* value of L1 or L2 should be considered as hypotheses.

5.3 Feasibility and Application

An initial evaluation of the annotation scheme has been performed through the annotation of 70 abstracts randomly chosen from the GENIA Pathway Corpus, containing a total of 2,603 annotated bio-events. Two annotators performed the annotation using a comprehensive set of annotation guidelines developed following a detailed analysis of the various bio-event corpora and the output of an initial case study [28].

The evaluation results have shown high inter-annotator agreement and a sufficient number of annotations along each category in every dimension. The favourable results of this experiment have confirmed the feasibility and soundness of the annotation scheme, and have paved the way for a large scale annotation effort involving multiple independent (i.e. non-author) annotators.

We are currently in the process of creating a large corpus of meta-knowledge enriched bio-events. This corpus will consist of three sub-corpora, which have previously been annotated with different types of bio-events, namely GENIA, GREC and a small corpus of full papers.

6. Conclusion

In this paper, we have described how text mining can help biologists to search and locate relevant information within the literature in a much more effective and efficient manner than is possible using a traditional search engine that performs keyword searches over unstructured documents.

Text mining techniques can be applied to biomedical texts to extract structured, semantically-oriented event representations of the biomedical knowledge contained within the texts. Queries can then be applied to these extracted events, rather than on the unstructured documents. Such queries can themselves be structured, allowing specifications of exactly which search terms should be related to each other, and how.

Extraction of events is a complex process requiring a number of text mining technologies, including NER and deep parsing. NER is important to ensure that only events containing biologically relevant entities are recognised, whilst parsing helps to identify potential event participants through syntactic relations. Annotated corpora of events are important for training systems to recognise events and their participants, as they provide direct evidence of how events manifest themselves in text. Computational lexicons such as the BioLexicon can further enhance performance, in providing detailed information about the idiosyncratic behaviour of verbs on which events are often centred.

Information regarding the intended interpretation of events is also important. Our proposed meta-knowledge annotation scheme for events and ongoing work to produce a large corpus of events annotated according to this scheme will form an important first step in allowing systems to be trained to recognise interpretative information about events from huge repositories.

7. REFERENCES

[1] Ananiadou, S., Friedman, C. and Tsujii, J. (eds.) Special Issue on Named Entity Recognition in Biomedicine. *Journal of Biomedical Informatics*, 37, 6 (2004).

[2] Ananiadou, S., Kell, D. B. and Tsujii, J. Text mining and its potential applications in systems biology. *Trends Biotechnol*, 24, 12 (Dec 2006), 571-579.

[3] Ananiadou, S. and Nenadic, G. *Automatic Terminology Management in Biomedicine.* Artech House Books, 2006.

[4] Ananiadou, S., Pyysalo, S., Tsujii, J. and Kell, D. B. Event extraction for systems biology by text mining the literature. *Trends Biotechnol.*, 28, 7 (Jul 2010), 381-390.

[5] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25 (2000), 25-29.

[6] Bjorne, J., Ginter, F., Pyysalo, S., Tsujii, J. and Salakoski, T. Complex event extraction at PubMed scale. *Bioinformatics*, 26, 12 (Jun 2010), i382-390.

[7] Bodenreider, O., Burgun, A. and Rindflesh, T. Assessing the Consistency of a Biomedical Terminology through Lexical Knowledge. *International Journal of Medical Informatics*, 67, 1-3 (2002), 85-95.

[8] Browne, A. C., Divita, G., Aronson, A. R. and McCray, A. T. UMLS language and vocabulary tools. In *Proceedings of the Proceedings of the AMIA Annual Symposium* (Washington DC, USA, 2003).

[9] Dolbey, A., Ellsworth, M. and Scheffczyk, J. BioFrameNet: A Domain-specific FrameNet Extension with Links to Biomedical Ontologies. In *Proceedings of the Proceedings of KR-MED 2006: Biomedical Ontology in Action* (Baltimore, USA, 2006).

[10] Eaton, A. D. HubMed: a web-based biomedical literature search interface. *Nucleic Acids Res*, 34, Web Server issue (Jul 2006), W745-747.

[11] Falagas, M. E., Pitsouni, E. I., Malietzis, G. A. and Pappas, G. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *Faseb J*, 22, 2 (Feb 2008), 338-342.

[12] Hearst, M. A. *Search User Interfaces*. Cambridge University Press, 2009.

[13] Hearst, M. A., Divoli, A., Guturu, H., Ksikes, A., Nakov, P., Wooldridge, M. A. and Ye, J. BioText Search Engine: beyond abstract search. *Bioinformatics*, 23, 16 (Aug 2007), 2196-2197.

[14] Hirschman, L. and Blaschke, C. *Evaluation of Text Mining in Biology*. Vol 9, Artech House Books, 2006.

[15] Hirschman, L., Colosimo, M., Morgan, A. and Yeh, A. Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinformatics*, 6 Suppl 1 (2005).

[16] Hoffmann, R. and Valencia, A. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, 21 Suppl 2 (Sep 2005), ii252-258.

[17] Hull, D., Pettifer, S. R. and Kell, D. B. Defrosting the digital library: bibliographic tools for the next generation web. *PLoS Comput Biol*, 4, 10 (Oct 2008), e1000204.

[18] Kim, J.-D., Ohta, T., Pyysalo, S., Kano, Y. and Tsujii, J. Overview of BioNLP'09 Shared Task on Event Extraction. In *BioNLP '09: Proceedings of the Workshop on BioNLP* (2009) 1-9.

[19] Kim, J., Ohta, T. and Tsujii, J. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9 (2008), 10.

[20] Kipper-Schuler, K. *VerbNet: a broad-coverage, comprehensive verb lexicon*. PhD thesis, University of Pennsylvania, 2005.

[21] Krauthammer, M. and Nenadic, G. Term Identification in the Biomedical Literature. *Journal of Biomedical Informatics, Special Issue on Named Entity Recognition in* Biomedicine (2004).

[22] Liu, H., Hu, Z. Z., Zhang, J. and Wu, C. BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics*, 22, 1 (Jan 2006), 103-105.

[23] Miwa, M., Saetre, R., Kim, J. D. and Tsujii, J. Event extraction with complex event classification using rich features. *J. Bioinform. Comput. Biol.*, 8, 1 (Feb 2010), 131-146.

[24] Miyao, Y., Ninomiya, T. and Tsujii, J. Corpus-oriented grammar development for acquiring a Head-driven Phrase Structure Grammar from the Penn Treebank. In *Proceedings of the Proceedings of IJCNLP 2004* (2004).

[25] Miyao, Y., Ohta, T., Masuda, K., Tsuruoka, Y., Yoshida, K., Ninomiya, T. and Tsujii, J. Semantic Retrieval for the Accurate Identification of Relational Concepts in Massive Textbases. *Annual Meeting- Association for Computational Linguistics*, 2 (2006), 1017-1024.

[26] Miyao, Y. and Tsujii, J. Feature Forest Models for Probabilistic HPSG Parsing. *Computational Linguistics, MIT Press*, 34(1) (2008), 35-80.

[27] Muin, M. and Fontelo, P. Technical development of PubMed interact: an improved interface for MEDLINE/PubMed searches. *BMC Med Inform Decis Mak*, 6, 36 (2006).

[28] Nawaz, R., Thompson, P., McNaught, J. and Ananiadou, S. Meta-Knowledge Annotation of Bio-Events. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010)* (Malta, 17-23 May, 2010).

[29] Nobata, C., Okazaki, N., Rea, B., Sasaki, Y., Tsuruoka, Y., Tsujii, J. and Ananiadou, S. Kleio: a knowledge-enriched information retrieval system for biology. In *Proceedings of the 31st Annual International ACM SIGIR Conference*, (2008), 787-78.

[30] Okazaki, N., Ananiadou, S. and Tsujii, J. Building a high-quality sense inventory for improved abbreviation disambiguation. *Bioinformatics*, 26, 9 (May 2010), 1246-1253.

[31] Okazaki, N., Ananiadou, S. and Tsujii, J. i. Building a High Quality Sense Inventory for Improved Abbreviation Disambiguation. *Bioinformatics* 26, 9 (2010).

[32] Palmer, M., Gildea, D. and Kingsbury, P. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31, 1 (2005), 71-106.

[33] Pyysalo, S., Ginter, F., Heimonen, J., Bjorne, J., Boberg, J., Jarvinen, J. and Salakoski, T. BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8, 50 (2007).

[34] Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H. and Jimeno, A. Text processing through Web services: calling Whatizit. *Bioinformatics*, 24, 2 (Jan 2008), 296-298.

[35] Rebholz-Schuhmann, D., Kirsch, H., Arregui, M., Gaudan, S., Riethoven, M. and Stoehr, P. EBIMed--text crunching to gather facts for proteins from Medline. *Bioinformatics*, 23, 2 (Jan 2007), e237-244.

[36] Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C. and Scheffczyk, J. *FrameNet II: Extended Theory and Practice*. 2006.

[37] Sasaki, Y., Montemagni, S., Pezik, P., Rebholz-Schuhmann, D., McNaught, J. and Ananiadou, S. BioLexicon: A Lexical Resource for the Biology Domain. In *Proceedings of the Proceedings of the Third International Symposium on Semantic Mining in Biomedine (SMBM 2008)* (Turku, Finland, 2008).

[38] Sasaki, Y., Tsuruoka, Y., McNaught, J. and Ananiadou, S. How to make the most of named entity dictionaries in statistical NER *BMC Bioinformatics*, 9 Suppl 11 (2008).

[39] Sasaki, Y., Tsuruoka, Y., McNaught, J. and Ananiadou, S. How to make the most of NE dictionaries in statistical NER. *BMC bioinformatics*, 9 Suppl 11 (2008).

[40] States, D. J., Ade, A. S., Wright, Z. C., Bookvich, A. V. and Athey, B. D. MiSearch adaptive pubMed search tool. *Bioinformatics*, 25, 7 (Apr 2009), 974-976.

[41] Thompson, P., Iqbal, S. A., McNaught, J. and Ananiadou, S. Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics*, 10, 349 (2009).

[42] Thompson, P., McNaught, J., Ananiadou, S., Montemagni, S., Trabucco, A. and Venturi, G. Building a bio-event annotated corpus for the acquisition of semantic frames from biomedical corpora. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)* 2008.

[43] Tsai, R. T., Chou, W. C., Su, Y. S., Lin, Y. C., Sung, C. L., Dai, H. J., Yeh, I. T., Ku, W., Sung, T. Y. and Hsu, W. L. BIOSMILE: a semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features. *BMC Bioinformatics*, 8, 325 (2007).

[44] Tsuruoka, Y., McNaught, J. and Ananiadou, S. Normalizing biomedical terms by minimizing ambiguity and variability. *BMC Bioinformatics*, 9 Suppl 3 (2008).

[45] Tsuruoka, Y., McNaught, J., Tsujii, J. and Ananiadou, S. Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. *Bioinformatics*, 23, 20 (Oct 2007), 2768-2774.

[46] UniProt Consortium The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, 38, Database issue (2010), D142-D148.

[47] Venturi, G., Montemagni, S., Marchi, S., Sasaki, Y., Thompson, P., Mcnaught, J. and Ananiadou, S. Bootstrapping a Verb Lexicon for Biomedical Information Extraction. In *Proceedings of the Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing* (Mexico City, Mexico, 2009). Springer-Verlag.

[48] Vincze, V., Szarvas, G., Farkas, R., Mora, G. and Csirik, J. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9 Suppl 11 (2008).

[49] Wattarujeekrit, T., Shah, P. K. and Collier, N. PASBio: predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*, 5, 155 (2004).

[50] Wilbur, W. J., Rzhetsky, A. and Shatkay, H. New directions in biomedical text annotations: definitions, guidelines and corpus construction. *BMC Bioinformatics*, 7, 356 (2006).

[51] Wren, J. D., Chang, J. T., Pustejovsky, J., Adar, E., Garner, H. R. and Altman, R. B. Biomedical term mapping databases. *Nucleic Acids Research*, 3, Database Issue: D289-293 (2005).

[52] Zweigenbaum, P., Demner-Fushman, D., Yu, H. and Cohen, K. B. Frontiers of biomedical text mining: current progress. *Brief Bioinform*, 8, 5 (Sep 2007), 358-375.

Corpus Annotation for Narrative Generation Research

A Wish List

Pablo Gervás
Universidad Complutense de Madrid
Madrid, Spain
pgervas@sip.ucm.es

ABSTRACT

This paper outlines the type of annotation of narrative that would be of interest for the purpose of research into narrative generation. This is done by first charting some of the goals of narrative generation research, and then trying to identify the features that might need to be annotated to address those goals.

1. INTRODUCTION

The last decade of NLP research has seen a rise in importance of annotated corpora as research tools. Their applicability extends to providing better understanding of specific phenomena, as evaluation material for natural language processing (NLP) solutions, and as training material for machine learning approaches. In parallel, there has been a rise in interest in narrative as a subject for NLP research, in terms of analysis [5, 21], generation [2, 13, 27, 12] and evaluation [24]. In some cases, this has involved joint work between narratology and computer science departments [14].

It was only to be expected that this situation gave rise to efforts at producing annotated corpora for narrative. In recent times there have been a number of initiatives directed at generating corpora of narrative texts annotated in some form: with descriptive referring expressions [15], emotional information [10, 9], semantical information [6], structural, functional, and emotional aspects connecting discourse segments in a coherent story [16] and an integration of linguistic markup and semantic models of folk narratives [18]. In parallel, there have also been initiatives for the development of specific annotation tools [7], and the application of new methods to narrative corpora, such as the use of analogical story merging for deriving narrative morphologies [8], or the use of latent semantic mapping to organize a fairy tale corpus and apply recommendation algorithms to it [19].

The existence of such a broad range of possible annotations for narrative is indicative of a great potential in terms of possible applications of the resulting corpora, but also of a significant risk. Unless some effort is made to coordinate in some way the various initiatives, the result in a few years time is likely to be a large set of mismatched corpora, each providing a different type of annotation for a different subset of texts. It also brings to light a different problem: each type of annotation will be geared to addressing a particular goal with respect to narrative. Two of the current initiatives ([7] and [18]) involve efforts to integrate different kinds of annotation, which is a positive contribution to the field. As narrative involves such a broad range of phenomena (extending well beyond all those already under research for natural language in general), it is also probable that, to be feasible, particular efforts will need to focus on subsets of the possible annotations.

This paper outlines the type of annotation of narrative that would be of interest for the purpose of research into narrative generation. This is done by first charting some of the goals of narrative generation research, and then trying to identify the features that might need to be annotated to address those goals.

First International AMICUS Workshop, October 21, 2010, Vienna, Austria.

2. RELATED WORK

Since annotated corpora have become established as an emerging tool of recognised importance, there have been numerous efforts to annotate text, both automatically, using NLP tools, and manually. The range of annotation now being carried out extensively covers a large number of phenomena concerning text in general (meaning disambiguation, coreference resolution, semantic roles, temporal expressions, spatial expressions, named entities...). A review of all this work is well beyond the scope of the present paper, but this should not be taken to imply that such work is irrelevant to the annotation of narrative. In fact, the annotation of narrative should begin with as thorough an annotation of the generic linguistic features of the corresponding text as possible. The rest of this section will focus on some concepts of narrative theory that are relevant to the annotation issue, and a brief review of relevant research on narrative generation which will inform the goals that the desired annotation might address.

2.1 Relevant Concepts of Narrative Theory

As Callaway [2] points out "current narrative theories are incapable of serving as the foundation for a comprehensive computational model that informs a decision algorithm for narrative generation". Nevertheless, it is important to consider some concepts of narrative theory that may help to

	English	French	Russian
what	story	histoire	fabula
how	discourse	discours	sjuzet

Table 1: Story and Discourse

stake out the problem. According to many theorists, narrative has two components: what is told (what narrative is: its content, consisting of events, actions, time and location), and the way it is told (how the narrative is told: arrangement, emphasis / de-emphasis, magnification / diminution, of any of the elements of the content). These have been named differently by different researchers, as described in Table 1. There are alternative analyses that postulate different subdivisions. Even between theories that agree on having just two levels of analysis there seem to be many subtleties that cast doubt on whether the same thing is meant by the different words. For instance some authors distinguish between the set of facts that characterise the situation being described (irrespective of whether they are actually mentioned in the discourse) and the set of facts that are mentioned in the discourse. There are also others who distinguish the linear sequence of facts that constitute the discourse and the actual text used to convey these facts linguistically. This lack of agreement on basic terminology presents a serious obstacle for researchers from the computational field trying to address the treatment of stories in any form.

An important number of the narrative-specific aspects of discourse were identified by Genette [11].

Narrative distance can involve *narratised speech* ("He confided in his friend, telling him about his mother's death."), *transposed speech, indirect style* ("He confided to his friend that his mother had passed away."), *transposed speech, free indirect style* ("He confided to his friend: his mother had passed away."), *reported speech* ("He confided to his friend: "My mother passed away.""").

The narrator can play different functions when conveying text: *narrative function* (he just tells), *directing function* (he interrupts the story to comment on its organization), *communication function* (he addresses the text's potential reader in order to establish or maintain contact with him or her), *testimonial function* (he comments on the truth, precision, or sources of the story, or his emotional involvement with it), and *ideological* function (he interrupts his story to introduce instructive comments or general wisdom concerning it).

Narrative distance and function of the narrator conform what is known as narrative mood.

Narrative voice captures whether the narrator is present or absent from the story he tells (and whether or not he is the hero of the story).

The time of narration captures the relation between the time of telling and the time being told about. It covers four kinds: *subsequent narration* (the narrator tells what happened in some past time), *prior narration* (the narrator tells what is going to happen at some future time), *simultaneous narration* (the narrator tells his/her story at the very moment it occurs) and *interpolated narration* (combines prior and simultaneous narration).

Narrative perspective or focalization is the way in which a narrator restricts what he is telling about a particular scene to what might have been perceived by someone present in that scene. There are three kinds of focalization: *zero focalization* (the narrator knows more than the characters), *internal focalization* (the narrator knows as much as the focal character) and *external focalization* (the narrator knows less than the characters).

Narrative voice, the time of the narration and narrative perspective conform narrative instance.

Narrative occurs at more than one level whenever a character in a story starts telling a story. Such nestings of stories within stories determine narrative levels.

Narrative time is characterised in terms of order, speed, and frequency of events.

Order is the relation between the sequencing of events as they actually occurred and their arrangement in the narrative. Any departure from the original chronological order is called anacrony. There are two types of anacrony: *analepsis* (the narrator recounts after the fact an event that took place earlier than the present point in the main story) and *prolepsis* (the narrator anticipates events that will occur after the present point in the main story).

In theatrical representations, the story is told at the same pace as it is happening (on stage). Speed involves introducing differences between the time the story takes to happen and the time taken to tell it. It is described in terms of four narrative movements: *pause* (the event-story is interrupted to make room exclusively for narratorial discourse such as static descriptions), *scene* (narrative time corresponds to the story's time, as in dialogue), *summary* (some part of the event-story is summarized in the narrative, creating an acceleration), and *ellipsis* (the narrative says absolutely nothing about some part of the event-story).

Frequency of events establishes the ratio between the number of times an event happens in the story and the number of times it is mentioned in the narrative. Events may be mentioned more times than they actually happened, or events that happened several times may be be told just once.

These parameters evolve over the course of a narrative, so that different spans of the narrative may have different values for each of these parameters.

2.2 Research on Narrative Generation

Research on narrative generation has a slightly different focus than narrative understanding research or language generation research. As this may impact on the type of annotation that would be most useful, a few relevant ideas are described here.

Narrative generation is an instance of natural language gen-

eration (NLG). As such, it inherits a number of generic issues from this field. Of these, the most significant is the subdivision into tasks [26] of *content determination* (deciding what to say), *discourse planning* (deciding on the order of presentation), *sentence planning* (deciding how to structure each sentence in its context of appearance) and *surface realization* (constructing the surface form text). However, NLG has in the past concentrated mostly on the production of non-narrative texts such as weather reports or instructive manuals. Nonetheless, there have been a number of efforts to develop generators specifically aimed at narrative.

STORYBOOK [2, 3] produced multi-page stories in the Little Red Riding Hood domain. It relied on elaborate natural language generation tasks such as narrative planning, sentence planning, discourse history, lexical choice, revision of drafts, a full-scale lexicon, and the FUF/SURGE surface realiser. Callaway introduces specific terminology for distinguishing what is told (the *fabula*: the sum of the factual content that constitutes the story) and how it is told (the *narrative stream*: the presentational ordering of the text). The differences and the relationships between them are clearly delimited. The narrative stream represents the information given explicitly. The fabula embodies the implicit information that underlies the narrative. Events in the fabula may be recoverable to a different extent from the narrative stream depending on how it has been generated. The fabula is a parallel knowledge structure, where information can be accessed in any order. The narrative stream is linear.

In his system Callaway models Narratorial Point of View and Narrator Modes, which are relevant to the narratological parameters mentioned in section 2.1. He also explains that the system requires a store of background cultural knowledge, including experience with previous stories told by others, ideas about how characters should act or talk, and fundamental knowledge of the world and how objects within it interact. Examples of this include simple facts like "trees have green leaves," "Grandma's house sits on the ground," and "a door serves as a portal between the inside and outside," even though they are never explicitly mentioned in any narrative. This he terms the story ontology.

The overall system involves a narrative planner (which actually generates a fabula and a narrative stream for the story), a narrative organizer (which produces a stream of sentential specifications) and a surface realizer (which converts these into text).

Callaway focuses on stylistics and mechanical realization during the production of text for an input story rather than the generation of fabula or narrative stream. The narrative planner is therefore simply specified and assumed to exist, with the actual implemented system dealing with the fabula and narrative stream as given inputs. Under this approach, a realtime narrative prose generator requires an algorithm that can intelligently combine information found in the narrative stream and fabula and convert it along with stylistic directives into narrative prose.

Lönneker [20] proposes an architecture for advanced NLG systems that handle narratives. In her paper, she reviews Callaway's STORYBOOK as a possible architecture for this task, and concludes that most of the narratological parameters that ought to be considered are in Callaway's proposal left for the narrative planner to decide, as they come already specified in the narrative stream that acts as input. She then proceeds to argue that the most important decisions of a narratologically enhanced system concern the document planner with its content determination and document structuring subtasks. She outlines four tasks that such a narrative planner might address: managing various forms of anacrony, ellipsis, introduction of more than one narrative level, and handling point of view and focalization. To illustrate this point she includes a more detailed analysis of how narrative levels and narrative instance might be implemented.

Montfort's PhD thesis [22] developed the *nn* system for interactive fiction, which was designed to address the issue of narrative variation and therefore covered some of the narratological parameters mentioned in section 2.1. In an interactive fiction system the user controls the main character of a story by introducing simple descriptions of what it should do, and the system responds with descriptions of the outcomes of the character's actions. Within *nn*, the Narrator module provides storytelling functionality, so that the user can ask to be "told" the story of the interaction so far. The Narrator module of *nn* addresses important issues in storytelling that had not been addressed by previous systems: order of presentation in narrative and focalization. Instead of telling events always in chronological order, the *nn* Narrator allows various alternative possibilities: flashbacks, flashforwards, interleaving of events from two different time periods, telling events back to front... It also captures appropriate treatment of tense depending on the relative order of events being told to speech time, reference time, and event time. Focalization is handled by the use of different *focalizer worlds* within the system. Aside from the actual world of the interactive fiction system, *nn* maintains additional separate worlds representing the individual perspectives and beliefs of different characters. These can be used to achieve correct treatment of focalization (telling the story from the point of view of specific characters).

Montfort has developed a new version of his system, now known as *Curveship* [23], which is intended for people to use as an interactive fiction system (as opposed to the research prototype that *nn* was). Curveship can deal with distance, frequency, speed, as well as focalisation and order.

The works of Callaway and Montfort concentrate on the generation of narrative text from an already existing flow of events (the narrative stream received as input in the case of Callaway, and the game flow as dictated by the interactive fiction system in the case of Montfort). There are many story generation systems that focus on the generation of plots, relying on simple template-based solutions for the production of text renderings of these plots. Lönneker reviews some of these, discussing architectural options for a narratology-enhanced generator. The task of inventing a story plot (whether from scratch or to satisfy a given specification or set of input parameters) is slightly different from the task of generating narrative text for a given input flow of events. For a more extensive review of plot generation systems, the interested reader is invited to consult [12].

3. TOWARDS A CORPUS-BASED APPROACH TO NARRATIVE GENERATION

Corpus-based approaches have proved successful for many areas of NLP. Even natural language generation has seen its fill of statistical approaches based on training corpora, from early efforts based on syntactic annotation only, such as for instance [17], to more recent ones [1] which rely on multilevel annotation including semantic structures. However, the specific nature of narrative generation must be taken into account when devising such an approach.

3.1 Layers of Representation of a Story

The discussion presented in section 2.1 has shown that there is a lack of consensus on terminology in this field. However, the analysis of existing systems given in 2.2 allows us to identify a number of concepts that are relevant to a computational analysis of this problem. These concepts arise from differences in the nature of the computational operations and decision processes involved in deriving one from another, either during text understanding or text generation. Although literature on the subject is already overloaded with terminology, I will attempt to give tentative names to them to facilitate the subsequent discussion.

I will consider that the following possible representations can be associated with a given story:

text representation the linguistic realisation of the story

explicit representation the linear sequence of facts mentioned in the story (in some kind of conceptual representation)

underlying selected representation all facts relevant to the story that are mentioned in the explicit representation (the set of facts that are mentioned in the story, but not necessarily organised in a linear sequence and following a chronological partial order not necessarily equivalent to the one in which they appear in the story)

underlying extensive representation all possible facts relevant to the story (including causes, effects, emotional reactions, common knowledge, and generally all the additional material that will be inferred by a reader on reading the story)

The explicit representation corresponds roughly to Callaway's narrative stream. The underlying extensive representation would correspond to the set union of Callaway's fabula and his story ontology, and the underlying selected representation would correspond to the part of the fabula that gets mentioned in the narrative stream.

Some of these distinctions are particularly pertinent for our general goal, as automatic annotation using state of the art NLP tools involves uncovering some of these hidden layers from the given input, which is likely to be the original text of a story. Reviewing some of these in order of their likely application to the text, processes like syntactic parsing, semantic role labelling, named entity recognition, coreference resolution or word sense disambiguation would address part of the conversion from the text representation into the explicit representation. The identification of temporal expressions and temporal relations in a text may be seen as subtasks in the way towards identifying a chronological order for the facts in the story, independent from the order in which they appear in the story. This might produce the underlying selected representation. Finally, the automatic identification of the omitted information that a human reader might reasonably infer would correspond to producing the underlying extensive representation. This is a long sought goal of natural language understanding which seems yet a little far off.

It must be said that these representations are postulated strictly with the intention of addressing the problem from a computational point of view. No claim whatsoever is intended as to their cognitive plausibility. In fact, a large percentage of the problems that we will describe below arise from the fact that the human brain clearly operates in radically different ways. However, a computational analysis of the problem must handle such elements as we can represent and handle in symbolic terms.

Obviously all of these annotations will be useful for research on narrative, as they are common to narrative and text in general. This paper will focus on those aspects that are specific to narrative. These in general constitute aspects that address the relation between the different layers of a story.

3.2 Tasks in Narrative Generation

Narrative generation spans a number of tasks, from the creative activity of coming up with a new convincing story, to the craft of putting together a fluent text that conveys a given set of facts.

In the past, any program that produced the text for a story has been considered a story generator. This included programs that simply concatenated pre-written strings according to a story grammar, programs that used planning to build an underlying representation which was then converted into text using templates, or programs which took an explicit representation (much like STORYBOOK's narrative stream) and used heavy NLG to produce high quality text from it.

The introduction of these definitions allows a more fine-grained classification. For instance, the simplest possible generators would be those that generate directly a text representation, with no conceptual representation equivalent to the explicit representation involved. Systems that generate directly a conceptual representation equivalent to an explicit representation (a conceptual representation that is already already linearised and ordered) would be less complex than those that generate an intermediate conceptual representation which is non linear and ordered differently from the final result). A detailed follow up of this line of work is beyond the scope of the present paper

Having an explicit sketch of the kind of representations that lie behind a story allows identification of some subtasks that might be involved in a complex generator:

invention production of an underlying extensive representation for a story

content determination production of an underlying selected representation from an underlying extensive representation

discourse planning production of an explicit representation from an underlying selected representation

telling production of a text representation from an explicit representation

The subtasks of content determination and discourse planning match those already identified for natural language generation. As in the case for NLG, it is not clear whether they can be carried out independently from one another. Decisions taken while carrying out one task may affect the other. This may indeed be true for all the subtasks that we are considering. Problems encountered during discourse planning or telling a story may be solved by additional processes of invention. This problem is similar to that of the subdivision of natural language generation into subtasks, and it need not be discussed here. In general terms, arguments equivalent to those presented by Reiter [25] could be put forward for and against such a subdivision as a useful abstraction. Additionally, invention may be applied at every layer of representation (inventing directly material for an underlying selected representation, an explicit representation, or even text).

Yet this set of tasks constitutes a useful set of tools for describing narrative generator. Most of the existing storytelling systems could be said to be carrying out invention, usually directly to an explicit representation, sometime to an underlying selected representation, and sometimes to an underlying extensive representation. But then very few of them rely on refined solutions for the content determination, discourse planning or telling subtasks. Important exceptions are STORYBOOK, which focuses exclusively on what we have termed the telling subtask, and nn/Curveship, which applies an interactive solution for the invention task, uses a quite refined solution for content determination and discourse planning (involving explicit treatment of several narrative parameters) and applies a very simple solution for telling.

There is an additional challenge for narrative generation that has not been explored in detail from a computational point of view. This is the craft of telling a set of facts as a story. With respect to our set of subtasks, it would correspond to taking an underlying extensive representation and applying to it processes of content determination, discourse planning, and telling (again, not necessarily in strict succession). This is in fact the kind of narrative generation that would be closest to traditional natural language generation.

However, a number of significant differences arise in this case.

Stories are traditionally understood as "sequences of events": a line of carefully alligned points in time. This is very rarely true. The underlying representation for most stories worth telling usually involves several events taking place simultaneously, in different locations, or even in the same location. If one were to assign a geometrical form to the underlying representation of a story it would hardly be a line. If you think of a story in terms of events, each event takes up a portion of time (duration) and each event can be associated with the portion of space from which it can be perceived (location). Even if you consider only a two dimensional representation for location, an event would have to be represented more as a volume than as a point. Additionally, the durations and/or the locations of different events may overlap. Once you consider a complete story, the geometrical representation for it would not even be a graph, but a set of heavily intersecting volumes of space/time.

The "linear story" mirage arises from the fact that stories are usually told as a sequence (of clauses, sentences, or even images if one considers films as stories). Language or film are by their very nature linear sequences, and stories wanting to be told by such means have to conform to this restriction. Traditional storytelling addresses this problem by drawing lines through the space/time volume that connect elements in a story. Each of these lines constitutes a narrative thread of the story. In this way, the volume is reduced to a graph. The next stage involves deciding how to traverse this graph as a sequence of linear paths through it, which result in the explicit representation of the story. With respect to our subtasks, the twin processes of drawing the threads and traversing the graph constitute instances of discourse planning. They probably do not happen sequentially but as a single complex interdependent operation.

Under this light, narrative parameters (and more specifically the conceptual operations that lead to the different formulations) become fundamental for understanding the storytelling task. For instance, focalization plays a central role in providing a rational way of partitioning the space/time volume into threads defined as what may have been perceived by a given focalizer. Different threads may be traversed by switching from one focalizer to another. This implies that in the underlying representation of a story events should carry additional information concerning their location, so that the focalization decisions can be informed. For correct focalization, decisions should be based not just on absolute concepts such as distance, but rather on whether an event can be perceived by a given focalizer. Many of the other phenomena covered by Gennette's narrative parameters can be seen as important tools for obtaining an appropriate linear sequence to convey the space/time volume of events of a given story. This is a key question for narrative generation research, and one that would clearly benefit enormous from the availability of annotated corpora.

3.3 Story Representations and Narrative Parameters

In the light of these definitions, many of the narrative parameters defined by Genette (described in section 2.1) constitute formal representations of specific relations between these different layers of representation of a story. Others bring into play additional material that we are forced to consider.

Narrative distance captures a number of choices open to the narrator in the way he decides to convey what has happened, when he makes the transition from the explicit representation to the text representation. Whereas natural language understanding is mainly concerned with problems of ambi-

guity (which of many possible representations for a sentence to assign as its interpretation), natural language generation faces problems of choice (which of many possible realizations for a given fact to employ in a certain context). The concept of narrative distance constitutes a very valuable source of information in this respect, and it could be very fruitfully used to inform consistent decisions in this respect across a given document.

Of all 5 functions of the narrator, only the narrative function is directly related to the actual story. This reminds us that narrative may include additional material that may need to be provide on the side for a generator and marked as such during annotation of a corpus. One should consider at what levels of representation such material would have to be included, and whether it should be marked as distinct from the actual story at all levels in which it appears. From the point of view of annotation or analysis, it may be tricky to discern in some cases the extent to which a certain element in the text concerns ideological material, for instance. In contrast, during generation such material would come from a different source to the rest of the content, so the difference is important.

Narrative instance captures the relation between the actual act of narration and the underlying story. Narrative voice will affect the use (in generation) and interpretation (in reading) of personal pronouns in particular spans. The time of the narration will affect the use of tense. These parameters affect the transition between the explicit representation and the text representation.

By changing focalization, however, different fragments of the underlying extensive representation may be selected to be included in the underlying selected representation. This can significantly alter the shape of the explicit representation and the text. It also plays a very significant role in the task of constructing a linear sequence to match a complex underlying story, as discussed above.

Narrative levels introduce the problem of recursion (a story within a story). This is a familiar problem in AI, but one must keep in mind that such recursion may involve a different four-layer representation of each level of embedding. Interactions between different levels may be problematic. This relates to the phenomenon Genette defines as *metalepsis* (characters from the main story appear in the secondary story).

The order parameter of narrative time captures the differences in partial ordering between the underlying selected representation and the explicit representation. This provides the flexibility required to traverse the graph implicit in an underlying selected representation (once threads have been established based on focalization) in a non-sequential way, and yet produce a linear explicit representation.

The speed parameter of narrative time captures the possibility that the underlying extensive representation and the underlying selected representation may differ in granularity. Each narrative move describes a different type of transition from one to another. In a scene, events in the underlying extensive representation get mapped to corresponding events in the underlying selected representation. In a pause, some of the static material (descriptive facts about elements that appear in the story) in the underlying extensive representation gets mapped to the underlying selected representation. In a summary, a set of events in the underlying extensive representation gets mapped to a much smaller set of events in the underlying selected representation. In an ellipsis, parts of the underlying extensive representation get omitted altogether.

Frequency of events captures the possibility that events appearing once in the underlying selected representation may be mentioned more than once in the explicit representation, or that events appearing several times in the underlying selected representation may be mentioned only once in the explicit representation.

3.4 A Corpus for Narrative Generation Research

This overview outlines some of the operations that need to be considered in a narrative generator in view of the selected stance on narrative theory. Read in a different way, it provides some clues as to what information might need to be annotated in a story for the resulting document to be useful from the point of view of narrative generation research.

The ideal case would be a story with the four layers of representation explicit and in which all the relevant narrative parameters fully annotated. As the parameters evolve in the course of a story, and they are determined by particular relations holding between different representation layers, some annotation scheme would have to be devised capable of capturing all this information.

From a corpus of such ideal cases, it would be possible to learn about:

- precise definitions of narrative parameters
- decision procedures for choices in transitions between representations (for generation)
- disambiguation procedures for interpretation (for understanding)

However, the construction of this ideal case would come up with several obstacles, most arising from the fact that only the text representation of a story is generally available.

First, it would be difficult to agree on an appropriate representation formalism for the conceptual knowledge involved in all layers other than the text representation. This is a long standing problem in AI and cognitive science that seems far from being solved.

Second, even if a conceptual formalism is chosen, the inferences involved in producing the underlying extensive representation would have to be carried out by human annotators. Because the range of possible inferences is large, a very low value of inter-annotator agreement is to be expected.

Third, the actual chronology of the underlying facts of a story is usually only vaguely specified in the text. This

problem has been addressed by Mani [21], who introduces *timelines* as an underspecified representation of temporal relations to deal with this problem. This solution allows for an economical representation that reduces disagreement, and it would help in annotating text. Exploiting such a solution to inform a narrative generator may need specific interfacing between whatever representation of time is being used in the input and these timelines.

Fourth, the theoretical definitions of the narrative parameters may not be amenable to precise definition in terms of relation between layers of representation. As the definitions were originally constructed without reference to such representations (which, as mentioned above, are particular to computational approaches to these problems), it is possible that any formal specification may fail to capture their original nuances[1].

Even considering all these obstacles, the task is not impossible. A simpler approach, based on annotating text with as much of this information as can be obtained would go a long way towards informing narrative generation research. The kind of information that might be obtained is quite extensive. As mentioned above, there are a number of NLP tools that could provide a very good starting point towards obtaining a good approximation to what I have called here the explicit representation of a story. Indeed, a large number of these are already integrated in the StoryWorkBench [7], an annotation tool specifically intended for narrative texts. Over this representation it would be interesting to add an additional layer of annotation to distinguish the specific functions of the narrator relevant to each span of the representation. This would help in sorting the specifically narrative material from all other contributions present in the text. As far as I know, this kind of annotation would have to be done manually. To obtain some form of underlying selected representation would require two additional steps: the identification of the temporal and spatial coordinates for the various elements (events but also descriptive material) identified in this explicit representation. Significant progress along these lines, and some interesting lines of future work with interesting potential, are reported in [21]. This kind of annotation would probably only be relevant for those spans of the material that correspond to a narrative function. If something like an explicit representation for a narrative text could be obtained, and annotated with additional information of narrator function, temporal information, and location information, it would in itself constitute a very powerful resource. This point has been forcefully argued by Mani in his recent book [21].

A possible extension would be to include additional annotations reflecting the values for all the remaining narrative parameters for the corresponding spans of text. This would provide a very valuable resource in itself for the study of narratology, possibly informing a more precise definition of these parameters in terms of the texts that they apply to. Such annotations would have to be carried out by hand, and the help of experts on narratology would be required. As mentioned above, a specific annotation scheme would have to be devised, capable of capturing the complex information involved in each case. Once an initial corpus of this kind were available, it might be possible to write tools for automatic annotation, possibly with the application of machine learning techniques to such corpus used as training material.

Many of these narrative parameters in some way encode transition operations between the underlying representation of the story and the explicit representation. If both an approximation to the explicit representation and annotations for the relevant narrative parameters were available, they might constitute a very good starting point towards obtaining an approximation of the underlying representation of the story. By combining the temporal and spatial information in the explicit representation with the information on transitions encoded in the annotations of narrative parameters, it might be possible to obtain a more precise approximation to the underlying representation than would be obtained without the help of these additional annotations. This would directly be a very positive result for narrative understanding. The resulting approximation would probably need to be validated by hand, but once it were available it would constitute a very valuable source for research on narrative generation.

[1]This may also be true of some of the interpretations presented in this paper.

4. CONCLUSIONS

Narrative generation research would benefit significantly from the existence of corpora of narrative texts annotated with the information that is relevant to the field (in addition to annotations already in use for natural language processing). Much like in NLP, only a small subset of the information that would be desirable is likely to be available in the short term, due to the difficulties inherent in the annotation task. However, a number of narrative-specific extensions to existing annotations in NLP have been identified. These concern mostly the annotation of narrative texts with values for Genette's parameters for narrative discourse. Additionally, possible applications of some existing annotations (concerning temporal and spatial information) to narrative-specific issues have been described. These ideas are put forward as suggestions, as a wish list of what it might be useful for narrative generation researchers to find in annotated corpora of narrative. There are indeed many more aspects that would need to be covered for a detailed study of narrative. Issues such as emotion, author goals, intention, reader reaction, and figurative language have not been considered here, but they are probably as relevant to narrative as the more basic issues mentioned here. Nevertheless, even going as far as I did involved already a fair amount of speculation, so detailed discussion beyond the basics should be left till specific advances have been made. Such advances may support or cast doubt on some of the assertion made in this paper.

5. ACKNOWLEDGMENTS

This work has been partially funded by the Spanish Science and Education Ministry under project NOVA (TIN2009-14659-C03-01) and by *Banco Santander Central Hispano* and *Universidad Complutense de Madrid* under the *Creación y Consolidación de Grupos de Investigación* program, Ref. 921332–953.

6. REFERENCES

[1] B. Bohnet, L. Wanner, S. Mill, and A. Burga. Broad coverage multilingual deep sentence generation with a

stochastic multi-level realizer. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, 2010.

[2] C. B. Callaway. *Narrative Prose Generation*. PhD thesis, North Carolina State University, 1999.

[3] C. B. Callaway. Narrative prose generation. *Artificial Intelligence*, 139(2):213–252, 2002.

[4] N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

[5] N. Chambers and D. Jurafsky. A database of narrative schemas. In Calzolari et al. [4].

[6] D. K. Elson and K. R. McKeown. Building a bank of semantically encoded narratives. In Calzolari et al. [4].

[7] M. Finlayson. Collecting semantics in the wild: The story workbench. In *AAAI Fall Symposium on Naturally-Inspired Artificial Intelligence*, pages 46–53. AAAI Press, Menlo Park, CA, 2008.

[8] M. Finlayson. Deriving narrative morphologies via analogical story merging. In *New Frontiers in Analogy Research*, pages 127–136. New Bulgarian University Press, Sofia, 2009.

[9] V. Francisco and P. Gervás. Ontology-supported automated mark up of affective information in texts. *Special Issue of Language Forum on Computational Treatment of Language*, 34(1):23 – 36, 2008.

[10] V. Francisco, R. Hervás, and P. Gervás. Two different approaches to automated mark up of emotions in text. In *AI-2006*, pages 101–114, Cambridge, England, December 2006 2006. Springer Verlag, Springer Verlag.

[11] G. Genette. *Narrative discourse : an essay in method*. Cornell University Press, 1980.

[12] P. Gervás. Computational approaches to storytelling and creativity. *AI Magazine*, 30(3):49–62, 2009.

[13] P. Gervás, B. Díaz-Agudo, F. Peinado, and R. Hervás. Story Plot Generation Based on CBR. *Knowledge-Based Systems. Special Issue: AI-2004*, 18:235–242, 2005.

[14] P. Gervás, B. Loenneker, J. C. Meister, and F. Peinado. Narrative models: Narratology meets artificial intelligence. In *International Conference on Language Resources and Evaluation. Satellite Workshop: Toward Computational Models of Literary Analysis*, pages 44–51, Genova, Italy, 2006.

[15] R. Hervas and M. Finlayson. The prevalence of descriptive referring expressions in news and narrative. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 49–54, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[16] O. Y. Kwong. Constructing an annotated story corpus: Some observations and issues. In Calzolari et al. [4].

[17] I. Langkilde and K. Knight. Generation that exploits corpus-based statistical knowledge. In *ACL-36: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 704–710, Morristown, NJ, USA, 1998. Association for Computational Linguistics.

[18] P. Lendvai, T. Declerck, S. Darányi, P. Gervás, R. Hervás, S. Malec, and F. Peinado. Integration of linguistic markup into semantic models of folk narratives: The fairy tale use case. In Calzolari et al. [4].

[19] P. V. Lobo and D. M. de Matos. Fairy tale corpus organization using latent semantic mapping and an item-to-item top-n recommendation algorithm. In Calzolari et al. [4].

[20] B. Lönneker. Narratological knowledge for natural language generation. In G. Wilcock, K. Jokinen, C. Mellish, and E. Reiter, editors, *Proceedings of the 10th European Workshop on Natural Language Generation*, 2005.

[21] I. Mani. *The Imagined Moment. Time, Narrative, and Computation*. University of Nebraska Press, Lincoln, Nebraska, 2010.

[22] N. Montfort. *Generating narrative variation in interactive fiction*. PhD thesis, University of Pennsylvania, Philadelphia, PA, USA, 2007.

[23] N. Montfort. Curveship: an interactive fiction system for interactive narrating. In *CALC '09: Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, pages 55–62, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

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

[25] E. Reiter. Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible? In *Seventh International Workshop on Natural Language Generation*, pages 163–170, Kennebunkport, Maine, USA, 1994.

[26] E. Reiter and R. Dale. Building applied natural language generation systems. *Nat. Lang. Eng.*, 3(1):57–87, 1997.

[27] M. Theune, N. Slabbers, and F. Hielkema. The automatic generation of narratives. In P. Dirix, I. Schuurman, V. Vandeghinste, and F. V. Eynde, editors, *Proceedings of the 17th Conference on Computational Linguistics in the Netherlands (CLIN-17)*, pages 131–146. LOT Occasional Series 7, LOT, Utrecht, 2007.

An Information Extraction Approach to the Semantic Annotation of Folktales

Thierry Declerck
DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg, 3
66123 Saarbrücken, Germany
declerck@dfki.de

Antonia Scheidel
DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg, 3
66123 Saarbrücken, Germany
Antonia.Scheidel@dfki.de

ABSTRACT

We propose an Information Extraction (IE) approach to the automated semantic annotation of folktales. We introduce and motivate the type of templates that we consider for encoding the (possibly underspecified) information extracted from textual and linguistic units of the tales. Each template is (possibly partially) instantiated on the base of a combined use of linguistic annotation and semantic resources. We opt for an incremental strategy: already instantiated templates can be further specialized on the base of subsequently instantiated templates. Once the full text has been processed, a round of specialization and of merging of the instantiated templates can take place.

1. INTRODUCTION

The work we describe here is part of the projects CLARIN[1] and D-SPIN[2]. While CLARIN is focusing on the establishment of an integrated and interoperable research infrastructure of language resources and technologies that aims at enabling eHumanities research in cooperation with Human Language Technology (HLT), the D-SPIN project, which is the German contribution to CLARIN, is additionally providing for integrated language processing Web services that generate linguistic annotation, which can be concretely used in eHumanities research.

A use case in CLARIN/D-SPIN, conducted in cooperation with the AMICUS Network[3], is investigating the possibilities of an automated processing of folktales that generates annotation that can be exploited by specialists in this specific field of narratives. We propose for this an Information Extraction (IE) strategy, which is applied on linguistically annotated folktales. On the base of such annotation, relevant IE units of a tale are detected. With *IE relevant units* we mean textual and linguistic units out of which basic information related to an event taking place in a particular temporal interval can be extracted and encoded in a corresponding IE template. Great importance is thus given to the recognition of temporal expressions in tales, which provides for a semantic means for text segmentation[4].

The filling – or instantiation – of the templates is done on the base of a combination of linguistic annotation and semantic resources, which will be described in more details in the paper. We adopt an incremental approach: while a template is being generated for each IE unit of the folktale text, this template can remain underspecified and be more specifically instantiated on the base of information detected and extracted in the context of subsequent units. Once the complete text has been processed and the corresponding amount of templates generated and (possibly partly) instantiated, an additional round of template comparisons, filling and merging can start, so that each template is fully specified[5].

The IE task is limited in a first step to providing for an automatic extraction of the characters of a tale, the particular relations existing between them, and the events they are involved in. Beyond this, we aim at establishing profiles of the characters, including their emotional states (if any), and we plan to collect information about all kind of objects mentioned in the tale. The IE templates we implement at this stage can in a sense be considered as giving the basic and generic information about the content of the tale.

First on the top of the instantiated (generic) templates, a classification of the characters in terms of character types and of the events in term of actions that correspond to a specific theory (like the typology of narrative structures suggested by Vladimir Propp in [11]) can be envisaged. We

[1]http://www.clarin.eu/

[2]http://weblicht.sfs.uni-tuebingen.de/

[3]AMICUS – Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts – is a research network on the topic of computational models of motifs in cultural heritage text and in scientific communication. See http://amicus.uvt.nl/

[4]The detection of spatial expressions is also quite important, but we do not use (yet) this information for segmenting a tale.

[5]This IE approach is guided by our actual work in the Monnet project. Monnet (Multilingual ONtologies for NETworked knowledge) is a FP7 R&D project co-funded by the European Commission with Grant No. 248458. See also http://www.monnet-project.eu/. While Monnet deals with e-Government and Business Information use cases, we are here testing the Monnet approach to Ontology-based information extraction when applied to a new domain.

First International AMICUS Workshop, October 21, 2010, Vienna, Austria.

assume that elements of arbitrary theories of narratives[6] can be added to the templates within a specific slot or field. We also assume that our approach can support the recognition of the motifs of a tale, since following Uther "a motif can be a combination of statements about an actor, an object, or an incident - [or of] all three of these elements"[7]. The IE templates contain this information, but we still need to investigate how to compute the relevant "combination".

The paper is organized as follows: we give first a brief description of the kind of linguistic annotation we use. This is followed by an introduction to IE. We present then some of the semantic resources we have consulted for supporting the IE process and the resulting semantic annotation. And finally we expand on the motivation of the IE templates used in our actual work and the incremental approach we follow.

2. LINGUISTIC ANNOTATION

We follow two annotation strategies[8]:

1. Stand off annotation, meaning that the annotation is not added to the text, which we call here *primary data*, but resides in an external data structure that is containing a referential system for pointing back to segments of the text.
2. A multi-layer approach to (linguistic) annotation. From the linguistic point of view, we annotate the tales with the following information:
   - Segmentation in tokens: in EN, GE, FR etc., sequences of characters separated by punctuation or blanks, and the punctuation signs.
   - Morpho-Syntactic properties of tokens. If a token is a verb, specify its person and tense, etc.; if a token is a noun, specify its gender, number, case, etc. Tokens are upgraded to word forms.
   - Constituency: grouping word forms in phrases (nominal phrase, verbal phrase, prepositional phrase, etc.) clauses and sentences.
   - Dependency: grammatical relations between elements of constituents (head elements vs modifiers, etc.) and between constituents (subject, direct object, etc.)
   - Semantic relations at the linguistic level (for example time, space, co-reference etc.)

This annotation strategy (stand-off and multi-layered) is not restricted to the linguistic data, but is valid for all information we want to use for annotating the tales. We give a short and simplified example of the possible linguistic annotation of the tale *The Magic Swan Geese*, which we take from [10]. The annotation is displayed here in an in-line fashion in order to ease readability, and we show only the morpho-syntactic and constituency annotation levels, as they are applied to five tokens of the tale:

```
<wordForms>
<W ID="w11" POS="ART" LEMMA="the"
   MORPH="Sg" tokenID="t11">the</W>
<W ID="w12" POS="NN" LEMMA="daughter"
   MORPH="Sg" tokenID="t12">daughter</W>
<W ID="w13" POS="ADV" LEMMA="soon"
   tokenID="t13">soon</W>
<W ID="w14" POS="ADV" LEMMA="enough"
   tokenID="t14">enough</W>
<W ID="w15" POS="VVFIN" LEMMA="forget"
   MORPH="Past" tokenID="t15">forgot</W>
...
</wordForms>
```

In the morpho-syntactic annotation above, the value of the TokenID of the 12th word is pointing to the original data (*daughter* is the 12th token in the text).

In the constituency annotation level displayed below, words are grouped into syntactic constituents (e.g. the nominal phrase *the daughter*). The span of constituents is marked by the value of the features `from` and `to`, which are pointing to the previous morpho-syntactic annotation layer.

```
<phrases>
<phrase id="p4" from="w11" to="w12" type="NP">
   the daughter</phrase>
<phrase id="p5" from="w13" to="w14" type="ADVP">
   soon enough</phrase>
<phrase id="p6" from="w15" to="w15" type="VG"/>
   forgot</phrase>
<phrase id="p7" from="w16" to="w20" type="REL_COMP">
   what they had told her</phrase>
...
</phrases>
```

On the top of this linguistic annotation, which is described in more details in [9], one can add additional annotation layers, like the various results of IE.

3. INFORMATION EXTRACTION

In this section we give a brief and selective introduction to Information Extraction (IE), presenting the elements that are playing a role for our current research related to the semantic annotation of folktales. We base our introduction on the slides of the lectures *Intelligent Information Extraction* given by Günter Neumann and Feiyu Xu at the ESSLLI Summer School 2004 in Nancy[9]. We just "verbalize" the relevant slides for our purpose, modifying slightly the original text and adding some more extensive explications. The slides contain also references to many classical papers on IE.

As Neumann & Xu state, the goal of IE is to build systems that find and link relevant information from text and to fill up predefined data records/templates with this information. The input to IE is thus twofold: templates that encode the type of information that is of interest for an application, and the textual documents out of which the concrete information can be extracted for filling up (or instantiate) the templates. Core problems of IE are the identification of a general mapping strategy between text fragments and template descriptions and the specification of all possible textual paraphrases

[6]We think here at the analysis of narratives proposed by Greimas in [4] or by Bremond in [3].

[7]This quotation from http://oaks.nvg.org/uther.html

[8]In compliance with ISO recommendations on the annotation of linguistic data, see [6] for details.

[9]See http://www.dfki.de/~neumann/ie-esslli04.html

for a relevant IE natural language expression. As a concrete example, we can consider a template representing a company profile, with respective fields for the name, the legal status, the branch of activity, its address, the number of employee, the name of the members of boards, etc. of a particular company. Natural Language Processing (NLP) of textual documents will be used for helping in finding names of companies and the associated information, and encode this as instances of the template representing a normalized company profile. IE can thus be considered as an interface between natural language processing and domain knowledge[10].

IE is traditionally subdivided in 5 sub-tasks:

- Named Entity (NE) task. Mark into the text each string that represents a person, organization, or location name, or a date or time, or a currency or percentage figure.
- Template Element (TE) task. Extract basic information related to organization, person, and artifact entities, drawing evidence from everywhere in the text (TE consists in generic objects and slots for a given application)
- Template Relation (TR) task. Extract relational information on employee_of, manufacture_of, location_of relations etc. (TR expresses domain independent relationships between entities identified by TE)
- Scenario Template (ST) task. Extract prespecified event information and relate the event information to particular organization, person, or artifact entities (ST identifies domain and task specific entities and relations)
- Co-reference (CO) task. Capture information on co-referring expressions, i.e. all mentions of a given entity, including those marked in NE and TE (Nouns, Noun phrases, Pronouns)

Interesting *shared tasks* in the field of IE were in the past the series of Message Understanding Conferences (MUC)[11]:

- MUC-1 (1987) and MUC-2 (1989) dealing with messages on naval operations
- MUC-3 (1991) and MUC-4 (1992) dealing with news articles on terrorist activity
- MUC-5 (1993) dealing with news articles about joint venture on microelectronics
- MUC-6 (1995) dealing with news articles on management changes
- MUC-7 (1997) dealing with news articles on space vehicle and missile launches

[10]Nowadays ontologies are more and more playing the role of templates in IE, and we use in this case the term of Ontology-based Information Extraction (OBIE).

[11]See also http://www-nlpir.nist.gov/related_projects/muc/

A commonality between all those editions of MUC was that they concentrated on the detection of events and their related arguments. So for example for the detection of events of succession in corporate executive personal, the IE systems had to detect not only the event, but the related position, the name of the persons involved, the reason for the change, the organizations involved (where is the new person coming from, where is the leaving person going to, etc.)

The ACE (Automated Content Extraction) program (1999-2008)[12] was another shared tasks initiative in the broader field of IE. A goal of this program was to develop core information extraction technology by focusing on the detection of specific semantic entities and relations over a very wide range of texts, and so discouraging highly domain- and genre-dependent solutions. ACE stressed the importance of detecting *unique* entities, relations, events and to find *all of their mentions* in documents. The relevance of ACE for our research can be summarized by the following points:

- Syntactic analysis of the text is a vehicle for organizing the information
- Toward the detection of each entity, relation, and event of a specific type
- Recognize all mentions of entities, relations and events, including the resolution of all mentions of the proper entity, relation, or event
- Convert information in human language into structured data, since structured data supports knowledge modeling & analysis
- Extract semantics of communication (this point was particularly missing from MUC)

ACE proposes a way towards the specification of the components of a broader semantic model for the content of different types of text.

- Entities – Individuals in the world
  - Simple entities: singular objects
  - Collective entities: sets of objects of the same type
- Attributes – Timeless unary properties of entities
- Temporal points and intervals
- Relations – Properties that hold of one or more entities over a time interval
- Events – A particular kind of relation among entities implying a change in relation state at the end of the time interval.

[12]See http://www.itl.nist.gov/iad/mig//tests/ace/ for more details.

This type of semantic model is guiding our approach to IE applied to folktales, especially for the first processing step consisting in identifying generic characters, relations, and events. We need therefore to identify what kind of linguistic mentions refer to different components of this model, taking into account here for example the different types of phrases (nominal phrases for entities, prepositional phrases for relations, pronouns for co-reference, etc., naturally dependent on the language in use).

We close our summary of the lectures by Neumann & Xu by stressing that we have to deal with text types that have not been considered till now by the large IE shared tasks campaigns mentioned above. We note for example that in the 10 tales we have been looking at, only very few Named Entities are used. Especially persons are not named very often. And a main type of relation between persons is the one of family relation. Also names of locations are very seldom. And the temporal expressions used are widely underspecified ("one day", "later", "when she came back", "evening" or "winter"). So that compared to the standard IE tasks, we can not normalize extracted temporal expressions to calendar dates and times, but have to confine ourselves to a topological representation of time. One of the consequences for our IE approach is that we take stronger advices from a temporal ontology, as described in [8], and from a family ontology, which is currently under development[13].

As a general strategy for the semantic annotation of folktales, we will first remain at the level of the extraction of entities, relations and events, corresponding roughly to the semantic model of ACE, before trying to further specify the entities, relations and events in terms of a specific theory of folktales or narratives. We identified various semantic resources for guiding our semantic annotation of folktales, and those are briefly described in the next section.

4. SEMANTIC RESOURCES

Besides the temporal and family ontologies mentioned in the former section, we consider the use of FrameNet[14] at the level of the extraction of generic information. For the theory specific annotation we are for sure considering resources in the field of folktales, like the ATU (Aarne-Thompson-Uther) classification system[15] and Vladimir Propp's seminal work *Morphology of the Folktale* (see [11]). And we plan to extend our work for populating the ProppOnto Ontology[16], which we can not present here.

We note that while the ProppOnto ontology or the PftML annotation scheme[17] represent a formalized account of certain aspects of the theory of Propp, we are not aware of any formalization of (parts of) the ATU classification system.

[13]Similar to http://www.owldl.com/ontologies/family.owl but with more complex relations and extended to topics of fairy tales.
[14]http://framenet.icsi.berkeley.edu
[15]http://en.wikipedia.org/wiki/Aarne-Thompson_classification_system
[16]http://www.fdi.ucm.es/profesor/fpeinado/projects/kiids/apps/protopropp/
[17]PftML – Proppian fairy tale Markup language, see http://clover.slavic.pitt.edu/sam/propp/theory/propp.html. See also section 4.4 below for a short discussion of PftML.

4.1 FrameNet

We started to investigate the use of FrameNet (FN)[18] as a semantic resource. FN is dealing with the creation of lexical resources based on frame semantics. FrameNet is available for four languages (English, German, Spanish and Korean), whereas we are aware of developments for Italian as well. The FrameNet consortium developed corpora, annotated with syntactic and grammatical roles information associated to the semantic frames[19].

The motivation behind the use of FN is the ability mark up natural language expressions with relational frame semantics. For example we can annotate the verb "rewarded" (in one version of *Red Little Riding Hood*, the hunter who saved the hero is rewarded with wine), with the semantic Frame Element (FE) `Rewards and punishments`. This Frame Element specifies following core arguments to the `reward.v` lexical unit (the letter *v* staying for the Part-of-Speech `verb`): *Agent*, *Evaluee* and *Reason*. On the base of this frame semantics, we can map natural language expressions to those frame arguments (filling a corresponding template). FN also allows for further non-core arguments that can be associated with the lexical unit: *degree*, *manner*, *instrument*, ..., *place* and *time*. For all of those arguments, the IE engine is trying to find corresponding text segments. FN also provides a list of associated lexical items, with their corresponding Part-of-Speech, which are associated with the same Frame Element `Rewards and punishments`:
`discipline.v`, `punish.v`, `recompense.v`, `reward.v`.

FrameNet proposes a hierarchy of FEs, and for example `Rewards and punishment` is inherited from `Intentionally affect`, which is inherited from `Intentionally act`, which is itself a sub-type of `Event`. Due to this inheritance structure we are able to detect and annotate relevant events in the tales, and also to classify those along the lines of the subclasses of the `Event` Frame Element.

However, we have to note that the examples in the corpus of FN are mostly taken from newspapers, whereas we are dealing with texts belonging to the folktale genre. The question arises on how to enrich – in an automated fashion – FN with new types of annotated examples. Following this direction, our work would support not only intelligent access to folktales but also would also give feedback in a way that enables the enrichment of the lexical semantic resources of FN. We note that a specialized FN for the soccer domain is already available[20], and we will investigate if a similar specialization in the field of folktales can be proposed.

4.2 ATU

The Aarne-Thompson-Uther (ATU) classification system[21] analyzes folktales by motif, such as *Supernatural or enchanted relatives*, *Persecuted heroine* or *Wild and domestic animals*,

[18]See http://framenet.icsi.berkeley.edu and [2]
[19]This corpus resource is a reason why we prefer in our context FN to WordNet (WN, see http://wordnet.princeton.edu/), since in FN syncategorematic information is associated with lexical units and through this with the corresponding semantic frames.
[20]See http://www.kicktionary.de
[21]http://en.wikipedia.org/wiki/Aarne-Thompson_classification_system

but is also a source of vocabulary, since the names of the tales that are categorized under the types reveal some of the typical characters and events that one can encounter in tales, so for example the motif type `Supernatural Opponents`[22]:

```
The Dragon-Slayer, 300
The Three Kidnapped Princesses, 301
The Giant Without A Heart, 302
The Twin Brothers, 303
Seven Sisters, Seven Brothers, 303A
The Trained Hunter, 304
The Twelve Dancing Princesses, 306
The Princess in the Coffin, 307
Rapunzel, 310
Killed by a Giant, 311
The Bluebeard, 312
The Magic Flight, 313
The Golden-Haired, 314
The Treacherous Sister, 315
The Mermaid in the Pond, 316
```

All the nouns, and other lexical units, listed in those titles of tales can be stored in a kind of gazetteer that can guide the IE process, like this is usually done for the recognition of Named Entities. But not only the lexical units are relevant, also the syntactic information is very valuable: so for example the information encoded in prepositional phrases: A princess can be *in* a coffin, someone can be killed *by* a Giant etc. We can relate this syntactic valency information to FEs of FrameNet and so get semantic roles associated with characters mentioned in the titles of the tales. And this again can help in semi-automatically define the templates for the folktale specific Information Extraction task.

A closely related semantic resource is the Thompson Motif Index[23]. Of particular interest for us is the fact that this index offers a kind of specialization of motifs, which can be considered as quite close to a taxonomy, as the example from the *Ogre* Index shows:

```
G500--G599. Ogre defeated
G500. Ogre defeated
G510. Ogre killed, maimed, or captured
G520. Ogre deceived into self-injury
G530. Ogre's relative aids hero
G550. Rescue from ogre
G560. Ogre deceived into releasing prisoner
G570. Ogre overawed
G580. Ogre otherwise subdued
```

In this example one can see again how we could transform this listing – and the vocabulary included in it – into related semantic frames, or even onto a real taxonomy (*being killed* as a sub-class of *being defeated*, formalizing thus Thompson Motif Index and maybe also parts of the ATU system in a semi-automatic manner.

[22]The digits in the listing are the so-called AT number entries

[23]We consulted here an online source: http://www.ruthenia.ru/folklore/thompson/

4.3 Propp's Morphology of the Folktale

This section is widely borrowed from [12], which is describing complementary material to the research presented in this paper.

From the analysis of Alexander Nikolayevich Afanasyev's collection of Russian tales (cf. [1]), Propp identified a number of common components, which we list below:

7 Character types. Propp puts forward the notion that the folktale know no more than seven *dramatis personae*: The villain, the donor, the helper, the princess and her father (sometimes treated as two dramatis personae, resulting in a total of 8), the dispatcher, the hero and the false hero.

31 Functions. At the heart of *Morphology of the Folktale* is the introduction and detailed description of 31 "functions", i.e. (mostly) actions which can be attributed to the dramatis personae of a folktale. According to Propp, every folktale consists of a subset of these 31 functions, arranged in one or more "move". The order of the functions is fixed, with a number of scrupulously defined variations. Functions are frequently divided into sub-functions: In the case of function *A: Villainy*, they range from A^1*: The villain abducts a person* to A^{19}*: The villain declares war.*

150 Elements. In Appendix I of *Morphology of the Folktale*, Propp provides what he calls a "list of all the elements of the fairy tale". The list contains 150 elements, distributed over six tables:

1. The Initial Situation
2. The Preparatory Section
3. The Complication
4. The Donors
5. From the Entry of the Helper to the End of the First Move
6. Beginning of the Second Move

Some of the 150 elements appear alone, others are grouped under a descriptive heading. If these "element clusters" are counted as one, as shown below in Fig. 1, the appendix contains 56 - as they shall tentatively be called in the following - narratemes.

About a third of the narratemes can be mapped directly to functions, such as the aforementioned *30-32. Violation of an interdiction.* Other narratemes can be combined to form an equivalent to a function (together, narratemes *71-77: Donors* and *78: Preparation for the transmission of a magical agent* can presumably be considered as a superset to the information expressed by function *D: First Function of the donor.*

```
30-32.  Violation of an interdiction
      30.  person performing
      31.  form of violation
      32.  motivation
```

Figure 1: Example for a narrateme

For the time being, our approach aims at extending the Proppian classification with a set of semantic relations, on the basis of the FrameNet approach.

4.4 APftML

APftML (Augmented Proppian fairy tale Markup Language)[24] is a markup scheme that combines linguistic, generic and domain-specific (folktales) semantic information. The scheme builds on and extends the mark-up language PftML (Proppian fairy tale Markup Language). PftML has been designed for transforming the grammar-like functions, subfunctions and the rules concerning their combination from *Morphology of the Folktale* into a DTD, allowing for an XML annotation of fairy tales. APftML extends and revises PftML in various ways, two of those being that the augmented scheme does not limit itself to the Proppian functions and the Proppian "information" is integrated in textual and linguistic annotation standards as proposed by TEI (Text Encoding Initiative) and ISO TC37/SC4 on language resources management. APftML, developed in parallel to our IE work, is the annotation scheme that is used for encoding the results of the IE applied to folktales.

5. THE IE TEMPLATES

We present now in an informal way the kind of templates we are using, in our two-level approach to IE applied to folktales.

5.1 The Generic Semantic Roles

We designed the templates so that they contain the information about the *WHs* of the tale, namely *Who*, *WhatObject*, *When*, *Where*, *WhatAction*, *ToWhom*, *Why*, *How*, etc. We are following in this an approach, which is similar to the scheme defined by the MPEG-7 standard for the *structured textual annotation* of multimedia data[25]. In MPEG-7 this annotation has the function to add semantic metadata to the content analysis of images or videos, which very normally remains at the level of physical descriptors (also called *Low-Level Features*). We provide the information about the characters (active or passive), the relations between them, the time and place in which they are mentioned, the actions (or events) in which they are involved etc.

In a first phase, the values that can be given to those descriptors (or slots in the templates) are extracted directly from text, allowing in certain cases for normalization or for establishing equality of information on the basis of basic inferences that can be derived from our family ontology. For illustration we take the first sentence of the tale *The Magic Swan Geese*[26]: "Once upon a time a man and a woman lived with their daughter and small son." The (simplified) corresponding template looks like:

```
When: (T1, past)
Where: Somewhere (P1, inferred:
     someone has to live somewhere)
Who: M1, W1,
     D1, S1
 age(S1) < age(D1), inferred)
WhatAction: Exist((M1,W1, D1, S1)
Updates: Introduction
     characters and relations
     hasChildren(M1,D1 & S1)
     hasChildren(W1, D1 & S1))
Speaker: Narrator
```

In this pseudo-logical representation, we just mark the fact that we have four characters introduced in the tale, and the relations existing between the man (M1) and the children and between the woman (W1) and the children. The family ontology and its associated rules allow us to group the daughter (D1) and the son (S1) under the class `Children`. But nothing allows us to state the M1 and W1 are married.

Within the *WH* features we include temporal information, which is also indicated by the tense of the verb *lived*, local information (giving the global context of the described situation) extracted from text or inferred. The values of *Who* are extracted from text on the basis of the heuristic that indefinite nominal phrases (NPs) are introducing referents, following here broadly theories like [5] or [7], and we use variables for naming those referents. Clearly this approach has to be adapted for languages not using indefinite NPs (or determiners). In the course of the tales, we then consider most of the occurrences of definite NPs as co-referent expressions (for example *the girl*, in "When the girl came back", will be co-referent to *daughter*, mentioned in the first sentence of the tale. Our concrete co-reference algorithm is making here also use of our family ontology, which is stating that both classes *daughter* and *girl* have female gender, and we thus do not rely solely on textual and linguistic clues. This ontology-based resolution of co-reference is already a big step toward a better semantic annotation of folktales, since the user searching for all actions involving *daughters* in tales, will not be forced to formulate her/his query in dependency of the strings that are present in the tale.

We can not consider all definite NPs as co-referring to formerly introduced referents. Examples are like "In the cabin was the old witch Baba Yaga...". In this case we have to deal with a Named Entity, and we consider that such expressions introduce a referent per se. But there are also cases where a character is first introduced by means of a definite NP, so for example the *swan geese* in "In swooped the swan-geese, snatched up the little boy, and flew away with him". Clearly

[24]See [12] and http://www.coli.uni-saarland.de/~ascheidel/APftML.xsd

[25]See http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm for mored details

[26]The English version available under: http://www.fdi.ucm.es/profesor/fpeinado/projects/kiids/apps/protopropp/swan-geese.html

we have to adapt our approach here, and a way for dealing with this case, is to have the *Swan Geese* within a gazetteer for folktales, as we already mentioned in the section 4.2, and so to consider it as kind of Named Entity.

Another issue we have to deal with: In the *Magic Swan Geese* tale the girl sees in the field an oven (introduced in the tale by an indefinite NP, as this is expected by us) and gets involved in a discussion with it. But while she seats in the hut of Baba Yaga, a mouse told her: "She [Baba Yaga] is going to steam you, put you in the oven, ...". Here we can not avoid the co-referencing mechanism to start, since we have in the tale both an indefinite and a definite NP referencing to an *oven*. But due to the fact that in our list of referents, the first oven is included in a template having as information on the location a *field*, we can assume here that we have two different ovens, the second one being located in the hut and not co-referencing to the first one, and so also not to be considered as a character of the tale (for which we assume that there are either introduced by an indefinite NP or by a Named Entity).

Additionally to the *WH* information, we add the features *Speaker* and *Updates*. With updates we mean something very similar as in dynamic predicate logic (see [5]): every utterance is describing a change of information of the interpreter (or reader).[27] At this IE level we could already apply the Proppian theory, and add to the template the information that we have to deal with an `Initial State`, since this would be quite straightforward. We can also postpone this step till we have analyzed the full text and generated all generic templates related to IE relevant units.

For the second sentence: "Dearest daughter," said the mother, "we are going to work. Look after your brother! Don't go out of the yard, be a good girl, and we'll buy you a handkerchief." Here the template looks like:

```
When: (T2, <= T1,  per inference (pi))
Where: H1 = P1 (H(ouse), pi)
Who: Mother = W1 (pi)
ToWhom: D1
WhatAction: talking
About What:
    NextActionOf(W1)
    AdvicesAbout(Brother=S1, pi),
    InterdictionFor(D1, GoOut)), ...
Updates:
    HaveWork (E1)
    WillLeaveHouse(W1)
    Sibling(D1,S1)
    Has(H1, CY1)
    Speaks(M1, D1)
Speaker: Narrator and W(1)
```

The updates are already interesting here. We know that the mother is talking to the daughter, and we can recognize that some commands/interdiction are formulated. At this level we could already apply the Proppian theory, and add to the template that an interdiction has been uttered to a central character of the tale. But we can also postpone this step and first see how often the *girl* is mentioned and in which kind of situation she is involved in the whole tale before marking the *girl* as the *hero* of the story.

Since in this sentence we have to do with a dialog, in which the persons are using the present form of the verbs, we know that we have to deal with another kind of events as if the narrator would be the sole "teller" (or speaker). We see in this particular example that the IE has to take into account two different "worlds": the factual one (the description of events by the narrator) and the description of possible factual worlds, as they are uttered by the participants of the dialog. We have to be careful in this case on extracting from a dialog only the relevant factual information (for example the formulated interdiction).

Such an approach to IE and domain-specific semantic annotation is particularly relevant when one considers all the possible co-reference linkings and more specially the anaphora used in the tale. Let us look again at the second sentence of the *Magic Swan Geese*, which we mentioned in the former section.

The interesting (and complex) point here is the fact that we deal with a dialog, introduced by the narrator. The mother speaks to her daughter, and says "We are going to leave for work". on the basis of the sole string sequence of this sentence, one can not resolve the anaphora *We*. A strategy would be to either consider the whole set of referents, or the two persons involved in this dialog, or just the mother. In the latter case, the co-reference algorithm can subsequently add other entities: if one looks at the next sentence of the tale *The father and mother went off to work*, we can then add the father (M1) to the set of entities denotated by the pronoun *We*. Since it seems to be easier to add referents to a pronoun in the course of the further analysis of the tale, then to remove members out of the set, we go for the minimal solutions, and after the analysis of this segment of the tale only the mother (i.e. the speaker in the first person) is added to the set of referents meant with *We*.

A case in which the attribution of a specific type to a character is definitely better postponed, and not directly attributed is the *snatching up of the boy*. We have in this textual context no full evidence that this is a kidnapping event, and we also can not categorize the Swan Geese as the villain. And in fact first when the girl reaches the hut of Baba Yaga, we can infer that the boy has been kidnapped and the Swan Geese is not the villain, but rather the witch (the Swan Geese can be considered as the Villain's helper, and we can add this information to the first template in which the Swan Geese is introduced as a neutral character.).

5.2 The Assignment of Character Roles and Functions

Just to more precisely motivate our two-level annotation strategy: We do not want to follow only the logic of Propp and assume (or infer) that the person receiving an interdiction, and violating this one, is automatically the hero of the

[27] We have a very pragmatic approach here and we do not consider the formal aspects of such theories, and all the implied philosophical debates. Sentences in a fairy tales are quite straightforward and the interpretation context given by a tale is very small in general.

tale or of the story. We want to have this role assignment also supported by linguistic and semantic evidence. At the lowest level, this can be due to the frequency of the mentioning of a character (supported by a co-reference algorithm in order to make sure that really all the mentions are collected, see again the requirements of the ACE program, described in the section 3). We are currently in the process of writing some rules to allow to map the generic character information to the Proppian descriptors.

6. CONCLUSIONS

We have been presenting the possible components of an Information Extraction approach to the semantic annotation of folktales. We suggest to follow a two-steps procedure, and to adopt the kind of semantic model described by the ACE initiative for defining the templates of IE in a first processing stage. We described for this the type of linguistic annotation and of semantic resources we are using. The second IE processing stage is dealing with the theory specific annotation of folktales, for which we have been consulting the Aarne-Thompson-Uther classification system and Propp's Morphology of the Folktales.

Parallel to our IE approach, an annotation scheme, APftML, has been developed and will be used for annotating the folktales with the results of the IE process. Future work will be dedicated to extending our approach to Ontology-based Information Extraction, allowing to populate existing or future ontologies in the fields of folktale or narratives in general. We will also propose a multilingual extension of our work.

7. ACKNOWLEDGMENTS

The research described in this paper has been partly funded by the European project CLARIN (http://www.clarin.eu/) and the German project D-SPIN (http://weblicht.sfs.uni-tuebingen.de/) for the linguistic and folktale specific aspects, and by the European project MONNET (with Grant No. 248458, http://www.monnet-project.eu) for the IE and Ontology related aspects.

We thank the AMICUS Network (http://amicus.uvt.nl/) for the intensive collaboration within the CLARIN use case on folktales and for the invitation to present our work at the first International AMICUS workshop.

8. REFERENCES

[1] A. Afanas'ev. *Russian fairy tales*. Pantheon Books, New York, 1945.

[2] H. Boas. From theory to practice: Frame semantics and the design of framenet. In S. Langer and D. Schnorbusch, editors, *Semantisches Wissen im Lexikon*, pages 129–160. Narr, Tübingen, 2005.

[3] C. Bremond. *La Logique du Récit*. Editions du Seuil, Paris, 1973.

[4] A. J. Greimas. *Sémantique structurale*. Larousse, Paris, 1966.

[5] J. Groenendijk and M. Stokhof. Dynamic predicate logic. *Linguistics and Philosophy*, 14(1):39–101, 1991.

[6] N. Ide and L. Romary. Representing linguistic corpora and their annotations. In *LREC 2006- The fifth international conference on Language Resources and Evaluation*. ELRA, 2006.

[7] H. Kamp. Discourse representation theory. In J. Verschueren, J.-O. Östman, and J. Blommaert, editors, *Handbook of Pragmatics*, pages 253–257. Benjamins, 1995.

[8] H.-U. Krieger, B. Kiefer, and T. Declerck. A framework for temporal representation and reasoning in business intelligence applications. In K. Hinkelmann, editor, *AI Meets Business Rules and Process Management. Papers from AAAI 2008 Spring Symposium. AAAI 2008 Spring Symposium: AI Meets Business Rules and Process Management, March 26-28, Stanford, CA, United States*, volume SS-08-01 of *Technical Report*, pages 59–70. AAAI Press, 2008.

[9] P. Lendvai, T. Declerck, S. Darányi, P. Gervás, R. Hervás, S. Malec, and F. Peinado. Integration of linguistic markup into semantic models of folk narratives: The fairy tale use case. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[10] P. Lendvai, T. Declerck, S. Darányi, and S. Malec. Propp revisited: Integration of linguistic markup into structured content descriptors of tales. In *Digital Humanities 2010*. Oxford University Press, 7 2010.

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

[12] A. Scheidel and T. Declerck. Apftml – augmented proppian fairy tale markup language. In *Proceedings of the First AMICUS Workshop*, 2010.

AutoPropp:
Toward the Automatic Markup, Classification, and Annotation of Russian Magic Tales

Scott Malec
Software Engineer
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
00 + 1* + (412) 330 7082
malec@andrew.cmu.edu

ABSTRACT

In this paper, I describe the current state of my program of research that uses the R environment and programming language for machine learning and statistics (Feinerer *et al*, 2008) to automatically analyze and/or parse texts in PFTML (Proppian Fairy Tale Markup Language) (Malec, 2005; Lendvai *et al*, 2010a; Lendvai *et al*, 2010b), an XML grammar that can be used for the semantic markup of folk narratives. R can be applied by researchers to manipulate and analyze motifs in tandem with "functions", the sub- and supra-sentential structures as described by Vladimir Propp in his *Morphology of the Folktale* (1928) in his analysis of Russian (magic) tales (Afanas'ev, 1945). These methods may be applied to corpora beyond the domain of folklore, but a limited formulaic corpus provides an interesting use case for these methods. The core idea is that this research program would use a corpus that was annotated in PFTML format as a training set to automate the process of annotating folkloric texts. I describe the conceptual framework, the required software packages, the basic steps of preprocessing text (annotation, stemming, removing whitespace and stopwords, importing data into the R environment) and discuss my progress on this project and the questions that I have encountered along the way.

1. INTRODUCTION

The AutoPropp research program attempts to compute the story structure of Russian magic tales given a training corpus marked up in Proppian Fairy Tale Markup Language and a hidden Markov model using some variant of an Expectation-Maximization algorithm, e.g. Viterbi, Baum-Welch in the R programming language. A hypothetically computed Story structure is conceived as the way in which boundaries of sub- and supra-sentential segments may be inferred by both their semantic content and a probabilistic model of the sequence of those segments, given training data for topic content of particular segments and a hidden Markov model that describes the likely sequence of those segments, respectively. This is a supervised learning technique that applies specifically to Propp's original corpus, that of Afanas'ev, in the original Russian. The laytteris a collection of texts with particular attributes that Propp described in his *Morphology of the Folktale* (1928).

First International AMICUS Workshop, October 21, 2010, Vienna, Austria.

Let me define a few terms first. A *function*, as Propp construes it, is a textual unit that generally moves the story forward, irrespectively of the dramatis personae that one encounters in it or that perform it. A *story grammar* is an abstract hypothetical schema that characterizes how the semantics and sequence of the elements of a particular genre of stories are constructed. Story grammars can be subject to mathematical rigor; as such, AutoPropp attempts to model these mathematical properties with hidden Markov models and distinct topic distribution for each function.

While there is no consensus *vis-a-vis* a grand unified theory of narrative, much less a working lexicon for elements of a general story grammar, yet many story grammars have been proposed that apply to specific corpora, e.g., Colby's grammar of Eskimo tales (Colby 1973). Other scholars across a diverse array of disciplines have performed research in this area, including cognitive science (Rumelhart 1975), story generation (Bringsjord and Ferucci 2000), literary theory, anthropology and elsewhere. I have picked Propp because I am familiar with both his work and his corpus. Another researchers may just as well choose another corpus and its respective generalized characterization. Propp is as good as any for starters.

2. AUTOPROPP: CONCEPTS AND ALGORITHM HIGHLIGHTS IN PSEUDO-CODE

To implement AutoPropp, several sets of data are required: a subset of Afanas'ev's *Russkie nardodye skazki* (1957) with each tale in a separate text file in its own directory, and both a

subset of Afanas'ev marked up in Proppian Fairy Tale Markup Language (PFTML) and an HMM model (to give AutoPropp a kind of roadmap or skeleton key through the terrain of the text) as training data to enable AutoPropp to build a model of the text.

These would be the initial inputs:

PFTML: Afanas'ev's *Russkie nardodye skazki* marked up in PFTML;

HMM_matrix: tab or comma delimited matrix of Proppian functions to create Hidden Markov Model (I have done this using Appendix III in Propp's *Morphology*);

[Note: this would look like A, B, C, Depart, D, E, F, K, Return, W, etc. with each element representing a function and each line representing a "Move"* within a tale in the corpus.]

Raw_input: Afanas'ev's *Russkie nardodye skazki* in plain text in separate text files. It may be useful to perform some preprocessing on this by way of annotating it given sentence structure as defined by sentence punctuation (or NLP annotation if that should be available, as these can give useful information to how the windowing function should proceed).

Allow me to sketch out tentatively what AutoPropp might look like in English:

1. Run trainer ("feed" AutoPropp PFTML and HMM matrix).

   a. Preprocess text within markup and create document text matrices for Proppian function across PFTML and generate semantic characterizations of each Proppian function that is encountered using Latent Semantic Analysis or Latent Dirichlet Analysis.

   b. Create Hidden Markov Model from HMM_Matrix to provide expectations for function sequences for AutoPropp's story parser.

2. Open Raw_Input, preprocess it, and pick text windows from that output starting with the first (or next) Proppian function in the Hidden Markov Model generated from the HMM_Matrix.

   a. Compare candidate passage with result of semantic analysis from that function from PFTML and calculate semantic similarity based on some variant of Bayes-ian co-occurrence or Levenshtein edit distance.

   b. Adjust window parameters left and right in output and calculate output of similarity metric.

   c. Compare variations from parser output by using the Expectation-Maximization (EM) algorithm of log likelihoods based on the Hidden Markov Model and similarity metric:

   *BestParse = Max(p*p*p*etc) + Max(s*s*etc.)*

   Save the maximized parse (and possibly a given percentage of the top contenders) of text windows that have been traversed through the text.

   d. Mark up story according to where function segment boundaries have been inferred, update Hidden Markov Model with new data, etc.

Now let us go through this in pseudocode to gain an additional perspective and to be yet more succinct:

AutoPropp Algorithm

1: Preprocess Raw Input and run HMM trainer.

2: Given PFTML corpus, create topic model of each function using LSA or LDA.

3: **for** each candidate passage s where s ∈ Raw Input, **do**

4: - Pick an s (given punctuation and other cues) and determine topics

5: - Pick a hypothetical function f given HMM model of text

6: - Determine similarity metric to topic given next function in HMM sequence

7: **end for**

8: Calculate for x permutations of s and functions

9: Pick MAX or top y% from MAX

10: Annotate texts from Raw Input accordingly

In Figure 1, the AutoPropp research program has been rendered in terms of a diagram. The character of the various inputs and outputs should be clear by now.

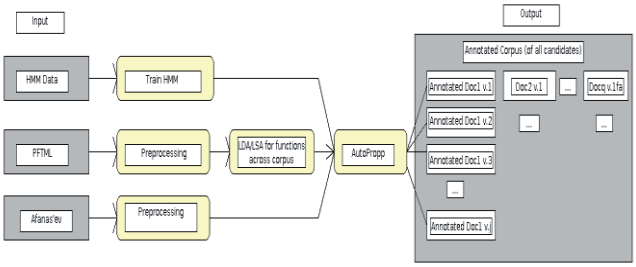

Figure 1: AutoPropp

In the next sub-sections, I will discuss the techniques and sub-components involved in AutoPropp in more depth. For clarity of exposition, the sub-sections below correspond to the sub-sections in section III of this paper, where I outline the packages in R with which I have experimented for implementation purposes.

2.1 Data and Data Preprocessing

2.1.1 XML/PFTML Markup

Approximately 20 tales have been marked up in PFTML. This was painstaking work that I undertook over the course of many excruciating weekends in the late 1990s and early 2000's.

2.1.2 Preprocessing the Data

With unstructured data such as text, one is obliged to preprocess it, i.e., to remove punctuation and stopwords, to remove inflections from it, and the like according to the goals of one's research, the discernment of the individual task, and the power of one's tools to arrive at psychologically compelling results. Many tools exist to make this a fairly painless process. Of course, language is a notoriously complex and shifty substance when it comes to computation. Take, for example, the problem of polysemy – there is no way to distinguish between /bank/ (as in bank of a river) and /bank/ (as in a place where one stores one's money or takes out a mortgage). Natural language processing can be employed to resolve cases of semantic ambiguity, but for a general bag of words of model, the techniques mentioned above satisfy the demands of most projects. The objective, in any case, is to reduce overall variability posed by inflected languages such as English and Russian to a single semantic instantiation of each term.

In the case of AutoPropp, it may ultimately be useful to retain certain types of punctuation, such as semi-colons, commas, and periods, i.e., punctuation that marks clause and sentence boundaries. Punctuation can aid with the efficiency of the story parsing algorithm by providing linguistic delimiters on where the boundaries of hypothetical functions are most likely to be located. This is because it is improbable if not impossible that a boundary of a function (to reiterate, a sub- or supra-sentential block of semantic content that moves a story along without regard to the dramatis personae) would break apart a noun clause (although the same cannot be said for compound sentences).

Specifically, for the purposes of AutoPropp in particular and text data mining projects in general, preprocessing a collection of texts prepares the way to create a document-term matrix, a rectangular matrix which may variably count or weigh instances of each term in each document. This may be reduced to, or decomposed yet further, into constituents by Singular Value Decomposition, where the input matrix is conceived as a product of three matrices – one of terms by the number of terms (U), another one of singular values (Sigmas), and a third one of documents (V). Once having arrived at this, a researcher can perform operations upon the document-term matrix such as to explore the distribution of topics across the corpus using one of the techniques that have been described below, or by others.

2.2 Hidden Markov Models

The central idea of Hidden Markov Models is to evaluate the log likelihood of what the next unknown element might be given what is known about the state of a current, known element. In the case of AutoPropp, this means that we may have an *E* function that there might be a 0.7 probability that it will be followed by some variant of an *F* function. In other words, this technique gives one insight into the probabilistic relations about the sequence of functions (going both forward and backward through a trellis of observed data). The Baum-Welch and Viterbi algorithms are used to calculate efficient routes through new observations and are commonly deployed in such areas as part of speech analysis and speech recognition.

There is a case where tales may have one or more moves. In the Hidden Markov Model that I have developed thus far, I have allowed for this case. In short, a Hidden Markov Model is a way of determining how tales are constructed – and, as such, they can be used to dismantle tales. There may be cases where solutions analogous to those developed in an old fashioned transformational-generative grammar may be applied. For example, the surface structure of a text may be inverted, i.e., the function that AutoPropp expects to see first may be second, due to a turn of phrase.

2.3 The Semantic Characterization of Proppian Functions

A sub-hypothesis in this research is that distinct types of content objects, e.g., distinct Proppian functions, will have distinct topic distributions. AutoPropp will experiment with two popular algorithms that are used for inferring topics from collections of texts: Latent Semantic Analysis (LSA) and Latent Dirichlet Analysis (LDA). As input, LSA takes a document-term matrix and generates a set of topics based on the distribution of terms across those documents. LDA does much the same. In fact, probabalistic LSA (pLSA) is the equivalent of LDA. The results of these methods are notoriously difficult to interpret. As such, these techniques ought to be considered to be stand-ins for coming developments in the highly dynamic area of topic modeling. The lexical field for each topic model could conceivably be expanded upon accurate match between a segment and a function. The success of this approach has been demonstrated in the question / answer subfield of information retrieval (Celikyilmaz *et al*. 2010).

2.4 Bayesian Analysis, Finding Associations and Co-Occurrences

AutoPropp has to have a way to measure the similarity of a segment of text, that is, a way of determining whether a hypothesis that a text segment is a member of a set of a particular Proppian function holds. This could be done with Bayes, with the Weka (http://www.cs.waikato.ac.nz/ml/weka/) packages findAssoc() function[1], or, most simplistically, by calculating the Levenshtein edit distance given semantic matches between the text segment and the topic model for the function to which it may hypothetically belong.

[1] I have turned Appendix III into an .arff file for Waikato's Rweka system. This data can be downloaded, shared, and experimented with in your instance of Weka at your leisure according to the GNU Public Software License from http://www.MalecLabs.com/Propp

3. R Packages Relevant to AutoPropp

In this section, I will go through some of the packages that will play some role in AutoPropp.

3.1 Preprocessing Revisited with R

This list is by no means exhaustive, but it is representative of what one would typically encounter when performing the types of research activities that have been described thus far *vis-a-vis* the AutoPropp project[2].

3.1.1 XML/PFTML Markup

A collection of Russian fairy tales that have been marked up in PFTML can be used to train AutoPropp's model of topics for each Proppian function. This is done by creating a document-term matrix of each function, where instead of a collection of documents per se, we have a collection of texts from each function, i.e., a function-term matrix. Topic model methods such as LSA and LDA, as discussed above, can then be used to generate a list of topic models for that collection of text segments. With XML, the key is to use the methods available in that package to perform the equivalent of XPATH function and penetrate into PFTMLs data model to deliver the content of a leaf, i.e., the content of a text segment.

```
library(XML); # load the XML library
x ← xmlTreeParse("95-Morozko.xml")$doc$children[[1]];
y ← x["Corpus/Folktale/Move/#"];
```

From there, one may drill deeper and venture into more specific areas of content (Proppian functions) within the structure of the PFTML document type.

3.1.2 tm

In the R language, all of the action takes place in the *tm* (text mining) package. Through the methods of the *tm* package, punctuation and stopwords are stripped and the results are stemmed using *Snowball*, another package required by *tm*. *Snowball*, in turn, uses the famous Porter stemmer; stemmers are available for many language of the Eurasian landmass, including the Russian of the Afanas'ev corpus[3].

The tasks that have been mentioned above (removing punctuation, stopwords, etc.) are performed with the following syntax:

```
library("tm"); # load the tm library
setwd("your/working/directory/for/Afanas'ev/");
# set working directory
afancorpus ← Corpus(DirSource(), readerControl =
list(language = "ru"));
afancorpus ← tm_map(afancorpus, removePunctuation); #
remove ; , . ! ?
afancorpus ← tm_map(afancorpus, stripWhitespace); #
remove whitespace, \n
afancorpus ← tm_map(afancorpus, removeWords,
stopwords("russian"));
afancorpus ← tm_map(afancorpus, stemDocument); #
expectations → expect
dtm ← DocumentTermMatrix(afancorpus); # "" ""
findFreqTerms(dtm, 5);
# finds the five most frequent terms in dtm
```

From this point, knowing what we know about the *XML* and *tm* packages, we can create document-term matrices from document collections. These document-term matrices can then be processed into topic models using LSA, LDA, and other techniques.

3.2 HMM

Hidden Markov Models can be used to predict what function is most likely to follow next in a sequence, given the known current state of the AutoPropp Russian magic tale parser. What I have done so far to this end is to create training data from Appendix III of the *Morphology of the Fairy Tale* (1928). Unfortunately, this is not an exhaustive list, but it does give a window into how the syntax in R's HMM package on CRAN can be used to take observations and to turn those observations to train a model.

This is what some of the sample code and data look like:

```
library(HMM);
afan <- initHMM(c("A", "B", "C", "depart", "D", "E", "F",
"G", "o", "L", "H", "M", "Pr", "J", "I", "N", "Rs", "K",
"return", "Q", "Ex", "T", "U", "W", "X", "MOVE"),
c("A", "B", "C", "depart", "D", "E", "F", "G", "o", "L",
"H","M", "Pr", "J", "I", "N", "Rs", "K", "return", "Q", "Ex",
"T", "U", "W", "X",   "MOVE"), startProbs=NULL,
transProbs=NULL, emissionProbs=NULL);
observations <- c(c("A", "B", "depart", "D", "E", "F",
"MOVE"),
        c("A", "depart", "Pr", "Rs", "H", "W"),
        c("A", "B", "D", "E", "F", "MOVE"),
        c("A", "depart", "Pr", "Rs", "H", "I", "W"),
        c("A", "B", "D", "E", "F", "return", "MOVE"),
        c("A", "B", "C", "depart", "D", "E", "F", "return"),
        c("A", "B", "C", "depart", "D", "E", "F", "D", "E",
        "F", "return", "MOVE"),
        c("A", "B", "depart", "D", "E", "F", "D", "E", "F",
        "return"),
        c("A", "D", "E", "F", "M", "N", "W"));
posteriorProbabilities <- posterior(afan, observations);
logForwardProbabilities <- forward(afan, observations);
logBackwardProbabilities <- backward(afan, observations);
print(exp(logBackwardProbabilities));
```

[2] For the latest version of AutoPropp, the data, and related code snippets, visit: http://www.maleclabs.com/Propp.

[3] Should the stemmers included in *tm* prove not to be compatible with one's needs, one can extend R's capabilities by using the *rJava* package to interface other stemmers in other languages, such as these (an aggressive Russian stemmer: http://members.unine.ch/jacques.savoy/clef/RussianAggressiveJava.txt and a light Russian stemmer: http://members.unine.ch/jacques.savoy/clef/RussianLightJava.txt).

```
bw <- baumWelch(afan, observations, maxIterations=100,
delta=1E-9, pseudoCount=0);
print(bw$afan);
afanHMM ← initHMM(afan, observations,
startProbs="posteriorProbabilities");
bwafan <- baumWelch(afanHMM, observations,
maxIterations=100, delta=1E-9,  pseudoCount=0);
```

3.3 Semantic Analysis

The two methods that AutoPropp will experiment with will be variations of LSA and LDA.

3.3.1 lsa

This is a brief synopsis of the *lsa* package according to CRAN: "The basic idea of latent semantic analysis (LSA) is that texts do have a higher order (= latent semantic) structure which, however, is obscured by word usage (e.g. through the use of synonyms or polysemy). By using conceptual indices that are derived statistically via a truncated Singular Value Decomposition (a two-mode factor analysis) over a given document-term matrix, this variability problem can be overcome." Below is some sample code from CRAN for *lsa* which plots a distribution of topics.

```
# make a space, reconstruct original landauerOriginalSpace
= lsa(dtm, dims=dimcalc_raw());
X = as.textmatrix(landauerOriginalSpace);
# X should be equal to dtm (beside rounding errors) all(
(round(X,2) == dtm) == TRUE);
# reduce dimensionality (Y shall be the recalculated 'reduc-
ed' matrix);
landauerSpace = lsa(dtm, dims=2);
Y = as.textmatrix(landauerSpace);
round(Y,2);
# now read in again the landauer sample (but with the
vocabulary of the existing matrix);
pdocs = textmatrix(ldir, vocabulary=rownames(dtm));
# now calc a pseudo SVD on the basis of dtm's SVD
Y2 = fold_in(pdocs, landauerSpace);
round(Y2,2);
# Y and Y2 should be the same (as well as dtm and pdocs
should be equal);
all( (round(Y,2) == round(Y2,2)) == TRUE);
# calc pearson doc2doc correlation;
rawCor = cor(dtm);
lsaCor = cor(Y);
round(rawCor,2);
round(lsaCor,2);
# clean up;
plot(landauerSpace$dk[,1:2]*landauerSpace$sk[1:2],
pch=17, col="darkgreen");
points(landauerSpace$tk[,1:2]*landauerSpace$sk[1:2],
pch=23, col="darkred");
text(landauerSpace$tk[,1:2]*landauerSpace$sk[1:2],rowna
mes(landauerSpace$tk), col="darkred");
text(landauerSpace$dk[,1:2]*landauerSpace$sk[1:2],
rownames(landauerSpace$dk), col="darkgreen");
```

This renders a graph. I have excluded the preprocessing work.

3.3.2 lda and topic models

This is a synopsis of *lda* (a package that implements Latent Dirichlet Analysis) from CRAN: "These functions use a collapsed Gibbs sampler to fit three different models: Latent Dirichlet Allocation (LDA), the mixed-membership stochastic block model (MMSB), and supervised LDA (sLDA). These functions take sparsely represented input documents, perform inference, and return point estimates of the latent parameters using the state at the last iteration of Gibbs sampling."

```
library("tm");
setwd("your/working/directory/for/Afanas'ev/");
txt ← system.file("texts", "txt", package = "tm");
library(tm);
plato ← Corpus(DirSource(), readerControl = list(language
= "eng"));
plato ← tm_map(plato, removePunctuation);
plato ← tm_map(plato, tolower);
plato ← tm_map(plato, stripWhitespace);
plato ← tm_map(plato, removeWords,
stopwords("english"));
plato ← tm_map(plato, stemDocument);
dtm ← DocumentTermMatrix(plato);
dtm ←DocumentTermMatrix(plato, control = list(stemming
= TRUE, stopwords = TRUE, minWordLength = 3,
removeNumbers = TRUE));
dtm ← removeSparseTerms(dtm, 0.99);
p_LDA ← LDA(dtm[1:500,], control = list(alpha = 0.1), k
= 10);
p_CTM ← CTM(dtm[1:500,], k = 10);
post ← posterior(p_LDA, newdata = dtm[-c(1:500),]);
round(post$topics[1:5,], digits = 2);
get_terms(p_LDA, 5);
```

This renders a 5 x 5 matrix of topics from the input.

3.4 bayesm and *RWeka for Co-Occurrence*

One solution would be to use the findAssoc() function to loop through the terms in a text segment in the AutoPropp parser's current window and run the following to determine the level or weight of similarity between the window and what is known about the distribution of terms and topics in a particular function (once library("RWeka"); has been loaded):

```
findAssoc(fDtm, "term_in_window", .5);
# this returns a list of terms from the document term matrix
for function D
# where D is a probability of .5 or greater that there is a co-
occurrence
```

Other techniques would be Levenshtein edit distance and Bayes with the *bayesm* package.

4. Progress

4.1 PFTML

Accurate, consistent data is the most rare and most precious commodity in an endeavor such as this.

4.1.1 The training data set

PFTML is a markup language that describes the structure of Russian magic tales as described by Vladimir Propp in his *Morphology of the Folktale* (1928). Each function is conceived as an XML element and each folktale is conceived as consisting of one or more moves. Each move consists of at least one cardinal function (villainy or lack) which may be preceded by an initial situation (such as an admonition) and is followed by such elements as departures, donor functions, pursuits, rescues, and weddings. Each of these functions may be represented by a topic model and is usually in sequence for the given corpus. The PFTML versions of the Afanas'ev text were OCRed between 1999 and 2001 from a 1957 Soviet imprint. These are available at the URL below.[4]

4.1.2 Afanas'ev

For this corpus, I copied the each text into its own file. As a source, I am using the link in the footnote.[5] Since the training set is in Russian, I am stuck for now using Russian as the target set.

4.2 AutoPropp .01

At this point, AutoPropp performs basic functions such as it preprocesses plain text, creates a document-term matrix, generates elementary topic models using LSA and LDA, and analyzes sets of observations to create a Hidden Markov Model of Proppian functions for a subset of Afanas'ev. This code and data have been included on my website[6].

4.3 Problems to Solve, Questions to Answer and some Proposed Solutions, where Solutions Exist

In this section, I discuss some of the problems that I encountered or questions that came up since I began this project. These will hopefully generate some interesting discourse as I think that they could be alternately perceived as both general and specific enough to provide hours of entertainment and debate among researchers, myself included, in this field.

4.3.1 Size of the Corpus

Is the corpus too small? What is the risk of overfitting the data?

4.3.2 Consistency of the Original Markup

Was I consistent with the original markup? .

4.3.3 ([I]n)fallable Propp

Did Propp get things right? Was Propp consistent with himself in his Appendix III that was used to model the sequence for the Hidden Markov trainer?

4.3.4 Null Data

What if data is not extant for a given function? It may be necessary to focus on a particular function or set of functions (e.g., the "donor" functions D-E-F) for which there is an abundance of data by way of a proof of concept. Divide and conquer may be the way forward.

4.3.5 Noisy Data

How does one cope with textual noise? R makes it easy to remove sparse terms that fall below a given threshold of frequency within a corpus. How should AutoPropp deal with repetition?.

4.3.6 Learning and Scoring

How will AutoPropp "learn" rules for determining the assignment of text segments as members of a Proppian function? Can it learn incrementally?

4.3.7 What Role Might Genetic/Evolutionary Programming Play?

After all, genetic/evolutionary programming (GP) is a search algorithm that uses some fitness function to optimize results. What if the windowing function was "stochasticized"? Might this be a way to turn this entire endeavor from a supervised learning experiment to an unsupervised one?

4.3.8 N-grams

Why not explore n-gram topic models?

4.3.9 WordNet

What about WordNet?

4.3.10 Annotated PFTML (APFTML)

Adding lexico-semantic features to the AutoPropp would help to resolve word and hence topic disambiguation.

5. DISCUSSION

Ultimately, should the proof of concept prove to be successful, a veritable wish list of methods should be included for the user whereby the analyst may activate a feature at will as if it were a black box. At that time, a tender balance will have to be made between elegance of interface and scrappy robustness, but I should not get ahead of myself.

In this paper, I have skirted many difficult issues. For example, what significance does the parsing of Proppian functions, assuming its validity for a moment, have for human communication; how mental story grammars influence the behavior of story tellers in real life, much less pragmatics, the cognitive comprehension (text processing) of stories; how these methods relate to the paradigmatic analyses proposed by Claude Lévi-Strauss and Pierre Maranda, or what the

[4] http://clover.slavic.pitt.edu/sam/propp/theory/propp.html.

[5] http://narodnye-russkie-skazki.gatchina3000.ru/

[6] http://www.maleclabs.com/Propp

implications for this might be for commercial applications such as sentiment analysis.

6. CONCLUSION

I have given what I hope is a useful picture to test, or, conversely, to falsify the epistemic validity of a story grammar, in this case, Propp's. I hereby conclude this paper, but not my research with AutoPropp.

ACKNOWLEDGMENTS

I am indebted to Sándor Darányi and Piroska Lendvai, the organizers of the first international AMICUS workshop for their patience and lively discussion.

REFERENCES

[1] A. Afanas'ev. *Russian fairy tales*. Pantheon Books: New York, 1945. (Transl. Norbert Guterman).

[2] S. Bringsjord and D. Ferrucci. *Artificial Intelligence and Literary Creativity: Inside the Mind of BRUTUS, a Storytelling Machine*. Lawrence Erlbaum Associates. Mahway, New Jersey. 2000.

[3] D. Blei., A. Ng, M. Jordan. Latent Dirichlet Allocation. In *Journal of Machine Learning Research*, Volume 3, 2003, pages 993-1022. DOI= http://jmlr.csail.mit.edu/papers/volume3/blei03a/blei03a.pdf

[4] A. Celikyilmaz, D. Hakkani-Tur, G. Tur, LDA Based Similarity Modeling for Question Answering in *NAACL 2010 – Workshop on Semantic Search*, 2010. DOI=http://www.eecs.berkeley.edu/~asli/asliPublish_files/naaclhlt2010.pdf

[5] B. N. Colby. A Partial Grammar of Eskimo Folktales. *American Anthropology*, Vol. 75, Issue 3, pages 645-662, June 1973.

[6] K. Hornik, and D. Meyer. Text Mining Infrastructure in R. In *Journal of Statistical Software*. March 2008, Volume 25, Issue 5. DOI= http://www.jstatsoft.org/v25/i05/paper

[7] P. Lendvai, T. Declerck, S. Darányi, S. Malec. Propp revisited: integration of linguistic markup into structured content descriptors of tales. In *Digital Humanities 2010*. London, United Kingdom, Oxford University Press, July 2010. DOI= http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/pdf/ab-753.pdf

[8] P. Lendvai, T. Declerck, S. Darányi, P. Gervás, R. Hervás, S. Malec, and F. Peinado. Integration of linguistic markup into semantic models of folk narratives: The fairy tale use case. In *Proceedings of LREC*, 2010. DOI=http://www.dfki.de/web/forschung/iwi/publikationen/renameFileForDownload?filename=Lendvaietal_DH2010_final%5B1%5D.pdf

[9] Maranda, P. and E. K. Maranda. *Structural Models in Folklore and Transformational Essays*. Mouton. The Hague. 1971.

[10] S. A. Malec. Proppian structural analysis and XML modeling. In *Proceedings of CLiP, Duisburg, Germany,* December 6-9, 2001.

[11] V. J. Propp. *Morphology of the folktale*. University of Texas Press: Austin, 1968. (Transl. L. Scott and L. A. Wagner).

[12] Rumelhart, D. On Evaluating Story Grammars in *Cognitive Science*. Volume 4. pages 313-316. 1980.

[13] Rumelhart, D. Notes on a schema for stories in *Representation and Understanding: Studies in Cognitive Science* (Bobrow, D. G. and Collins, A., eds.), pp. 211--236, New York: Academic Press, Inc., 1975.

Harvesting Event Chains in Ritual Descriptions Using Frame Semantics

Anette Frank
Department of Computational Linguistics
Im Neuenheimer Feld 325
69120 Heidelberg
+49-6221-543247
frank@cl.uni-heidelberg.de

Nils Reiter
Department of Computational Linguistics
Im Neuenheimer Feld 325
69120 Heidelberg
+49-6221-543169
reiter@cl.uni-heidelberg.de

ABSTRACT

Led by the observation of similarities and variances in rituals across times and cultures, ritual scientists are discussing an underlying abstract – and possibly universal – structure of rituals, which nevertheless is subject to variation. In an interdisciplinary project, we investigate the use of computational linguistic techniques to make characteristic properties and structures in rituals overt. For this sake, we apply formal and quantitative computational linguistic analysis techniques on textual ritual descriptions. We employ data-driven approaches to detect regularities and variations of rituals, based on semi-automatic semantic annotation of ritual descriptions, thereby addressing this research issue in a novel empirical, quantitative fashion.

Computational linguistics has developed semantic lexica and processing tools for the formal analysis of events and their predicate-argument structure, in terms of semantic roles. Frame semantics (Fillmore et al. 2003), with its concept of scenario frames connected by frame relations and role inheritance, offers a particularly powerful framework for the modeling of complex event sequences. Through the annotation of word senses, we can observe and analyze variations in the selectional characteristics of specific events and their roles across rituals. These structured and normalized representations of event sequences will be used as a basis for identifying recurrent patterns and variations across rituals by quantitative analysis.

In our talk, we present motivations and prospects of this approach to ritual structure research, focusing on two major aspects:
(i) We discuss design decisions for data collection, choices of NLP processing tools and workflows for semantic annotation. We focus on the special characteristics of the textual data and present a number of **domain adaptation** techniques to assess diverse methods for adapting our resources and tools to the novel domain.
(ii) In the second part, we discuss in more detail the semantic annotation layers and how they will contribute to make similarities and differences of (partial) **event sequences** overt. We will present first experiments on computing abstract event chains on the basis of our structured annotations and on detecting recurrent event patterns by statistical analysis.

First International AMICUS Workshop, October 21, 2010, Vienna, Austria.

REFERENCES

[1] Aljoscha Burchardt, Anette Frank, and Manfred Pinkal. 2005. Building Text Meaning Representations from Contextually Related Frames – A Case Study. In *Proceedings of IWCS*, 2005.

[2] Hal Daumé III. 2007. Frustratingly easy domain adaptation. In Proceedings of ACL, pages 256-263.

[3] Hal Daumé III and Daniel Marcu. 2006. Domain Adaptation for Statistical Classifiers. *Journal of Artificial Intelligence Research*, 26:101-126.

[4] Oliver Hellwig. 2009. A chronometric approach to Indian alchemical literature. *Literary and Linguistic Computing*, 24(4):373-383.

[5] Axel Michaels. 2007. "How do you do?" Vorüberlegungen zu einer Grammatik der Rituale. In: *Der Mensch – ein "Animal Symbolicum"? Sprache – Dialog -- Ritual.* Wissenschaftliche Buchgesellschaft, Darmstadt.

[6] Axel Michaels and Niels Gutschow. 2005. Handling Death. The Dynamics of Death Rituals and Ancestor Rituals Among the Newars of Bhaktapur, Nepal, volume 3 of Ethno-Indology. Heidelberg Studies in South Asian Rituals. Harrassowitz Verlag.

[7] Nils Reiter, Oliver Hellwig, Anand Mishra, Anette Frank, and Jens Burkhardt. 2010a. Using NLP methods for the Analysis of Rituals. In *Proceedings of LREC 2010*, Malta.

[8] Nils Reiter, Oliver Hellwig, Anand Mishra, Irina Gossmann, Borayin Maitreya Larios, Julio Cezar Rodrigues, Britta Zeller, and Anette Frank. 2010b. Adapting Standard NLP Tools and Resources to the Processing of Ritual Descriptions. In: Caroline Sporleder and Kalliopi Zervanou, editors, *Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010),* Lisbon, Portugal.

OntoMedia: Telling Stories to Your Computer

Dr. K. Faith Lawrence
Royal Irish Academy
kf03r@ecs.soton.ac.uk

Dr. Michael O. Jewell
Goldsmiths College, University of London
m.jewell@gold.ac.uk

Paul Rissen
British Broadcasting Corporation
paul.rissen@bbc.co.uk

ABSTRACT

This paper describes work that has been undertaken on, and related to, the OntoMedia ontology. Having provided an overview of the ontology and the way in which it can be used to describe the content of fictional and non-fictional narratives, we present a summary of research done using the principles of the OntoMedia ontology to support the description of television drama and an investigation of how TEI and RDFa might be used to automatically generate content descriptions for plays and screenplays which can be used to launch the annotation process.

1. INTRODUCTION

The OntoMedia ontology was developed to allow the semantic annotation of the narrative components of media and the relationships between those components. This annotation allows the computer-assisted search, analysis and exploration of marked up material at a content level which can offer a new perspective into a text or corpus of works. In this paper we present a general introduction to the ontology and how it can be used to support the description of narratives at a conceptual element level.

Finally we describe two investigations that have been undertaken using the OntoMedia Ontology. The first concerns the use of the ontology to support the description of television drama and the second addresses the problem of automatic annotation through the use of TEI and RDFa.

This work has been undertaken by a number of researchers and developers over the last six years although this paper focuses on the work done by Dr K. Faith Lawrence while at the University of Southampton and working at the Royal Irish Academy, Dublin, Dr Michael Jewell while based at the University of Southampton and Goldsmiths College, University of London, and Paul Rissen and his colleagues at the British Broadcasting Corporation.

First International AMICUS Workshop, October 21, 2010, Vienna, Austria.

2. THE ONTOMEDIA (OM) ONTOLOGY

The OntoMedia ontology was first developed at the University of Southampton by Drs K. Faith Lawrence[15], Michael O. Jewell[12] and Mischa Tuffield[23]. The intention was to create an ontology to describe the content of heterogeneous media as a means to expose the narrative elements within the media and the links between those elements (both within and across those narratives).

The decision was made to approach the texts at a very basic level so that the content could be made available to researchers with minimum interpretation imposed at the level of the conceptual model. It was felt that basic concepts as 'Hero'/'Villain' or, even, 'Protagonist'/'Antagonist' were subjective and depended on the viewer's reading of the text. While narrative theory formed a partial basis for the model of events, the terminology of the entities was designed to avoid these labels which were seen as semantically loaded.

Focusing on fiction as the primary use case a number of difficulties were seen as vital to address within the model. These included:

- Potential invalidity of real world assumptions, especially with regards to time and space
- Existence of multiple, cross-linked conceptual universes
- Inexact/Referential details
- Multiple interpretations of any given element or event

Upper-level ontologies such as DOLCE[17], Cyc[16], SUMO[21], the libraries and archives born ABC ontology[14, 5] or the cultural heritage developed CIDOC Conceptual Reference Model[2] all facilitate the description of concepts and events; However these models are fundamentally situated in the "real" world. While these ontologies could deal with some of these problems to a greater or lesser extent, the decision was made to develop an ontology which was specifically designed to deal with these issues. A survey of online authors and readers was carried out within the online media fan community to investigate what metadata was commonly attached to works of fiction shared online and what the preferences were for readers with regards to the type of content

information that they were given in conjunction with the story. This community was chosen due to their practice of providing not just bibliographic but content information as standard when sharing material online[15].

The result was the OntoMedia ontology.

2.1 I Don't Think We're in Kansas Any More

The OntoMedia (OM) ontology was designed on an event-entity basis using a modularized system to allow greater flexibility in usage by making it possible to use or extend only those sections which are relevant in a given situation.

The files in the ontology are divided into three sections:

- **Core**: These files contain the top level classes and properties which are seen as being core to the ontology. This section includes three files:
  - **OntoMedia Expression** which contains the basic classes and properties for describing narrative (*namespace: ome*).
  - **OntoMedia Media** which contains the basic classes and properties for describing the media through which the narrative is expressed and the relationship between the narrative and the media (*namespace: omm*).
  - **OntoMedia Space** which describes locations and geographic areas (*namespace: loc*).
- Extensions: These files contain extensions to the core classes and properties. This section is subdivided into:
  - **OntoMedia Common** which contains those extensions that are likely to be in common use such as those describing beings (*namespace: omb*), traits (*namespace: omt*) and related classes.
  - **OntoMedia Detail** which contains more specific classes such as knowledge and basic humanoid and animal anatomy. These classes while useful are less universal than those concepts described in either the core or common classes.
  - **OntoMedia Events** expands on the classes and properties which relate to events of different types including travel.
  - **OntoMedia Fiction** which contains classes with relate directly to the concept of fiction, fictional characters and the differential between fiction and reality.
- Misc: These files contain classes and properties which are used with the OntoMedia ontology but which do not extend it directly and therefore can be regarded as independent from it.

The most recent version of the ontology is available at `http://code.google.com/p/contextus/`

2.1.1 OntoMedia Entities

An `ome:Entity` is defined as an object or concept. The subclasses of the `Entity` construct (see Fig 1) fall into three categories:

- Those related to objects both physical and abstract, for examples `omb:Being` (physical) and `ome:Item` (physical and abstract).
- Those related to spatial models, for example `loc:Space`
- Those relating to time, for examples `ome:Timeline` and `ome:Occurrence` (see below).

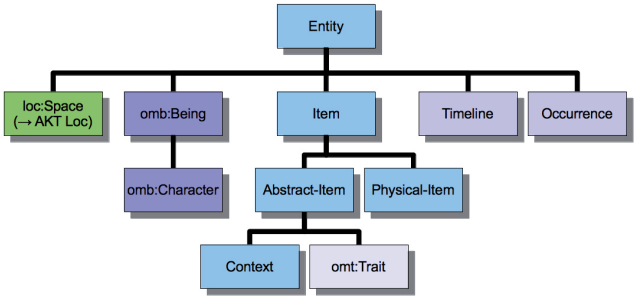

Figure 1: Top Level OntoMedia Entity Classes

It should be noted that the separation between `Physical-Items` and `Abstract-Items` does not carry any implication as to whether a specific instance of an item exists or not but whether items of that class may potentially have physical existence which is not wholy symbolic.

2.1.2 OntoMedia Events

An `Event` describes an interaction between one or more `ome:Entitys` during which zero or more attributes of those `Entitys` are modified or a new `Entity` is created. An `Entity` may have an attribute set to show that it no longer exists but the `Entity` itself is not destroyed.

While not explicitly indicated in the ontology, events (see Fig 2) can be divided into two characteristic types: those which could be described by a persistent change in the metadata, and those in which the event had little or no discernible effect on the underlying information or entities that existed beyond the scope of that event. The latter are regarded as exposition events while the former were narrative events or plot events.

The basic narrative event types were drawn from narrative theory and intended to describe the standard plots that are seen to exist – `ome:Gain`, `ome:Loss` and `ome:Transformation`. Exposition events owe more to visual media where sequences are included for effect (aesthetic, dramatic or otherwise) rather an to further the plot. The `ome:Action` class covers a broad range of event types including `Sex`, `Violence` and `Consumption` (eating/drinking) although events involving sexual activity may be instances of `ome:Social` events in addition to `Sex`. The `ome:Introduction` event models a section of material which exits purely to mark the presence of or otherwise introduce the character which the `ome:Social` class includes general social interactions.

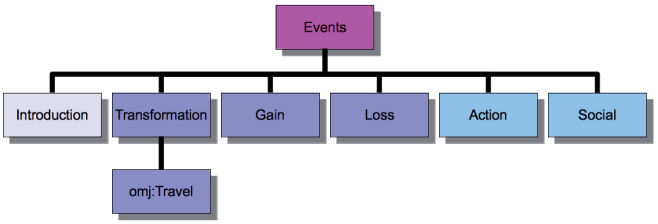

Figure 2: Top Level OntoMedia Event Classes

2.1.3 Building a Narrative

By combining these events and entities it is possible to build up the elements within a narrative. Taking the story of Sleeping Beauty as an example, we consider the pivotal scene in which the Prince, seeing the eponymous heroine, kisses her and wakes her up. This scene involves two characters: Character 1 (The Prince) and Character 2 (The Princess). The first event that we need to model is the kiss. This action could be seen as either a `Sex`, `Social` or combination event and the decision on how to classify it is a matter of interpretation (see below). In this example we have chosen to be guided by the previous versions of the tale and make the event an instance of the `Sex` class. However, since a `Sex` event could be anything from a heated look to an orgy, we have the option of making it explict as to what is involved in this case - the lips of the two characters i.e. a kiss.

The kiss causes the Princess to wake - transforming her from having a state-of-consciousness from Asleep to Awake (see Fig 3).

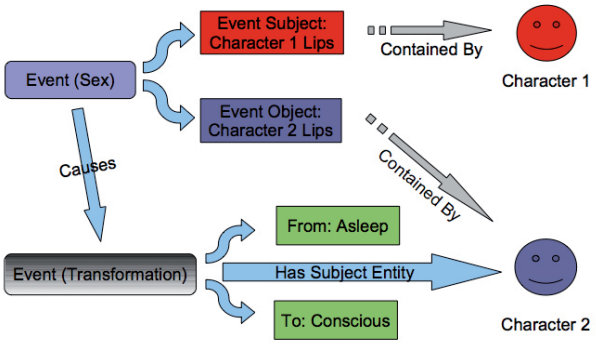

Figure 3: Diagram of Narrative Elements in Sleeping Beauty Scene

Having described the events taking place within a narrative and the entities involved within those events then this information can be analysed. Even with just the basic information about a narrative it is possible to compare amount, location and distribution of character activity to identify the signification characters in the narrative arc, or in a given section of it, whether a character is more active (the subject of events) or passive (the object or spectator to events) or trends in the event type. Fig 4 shows the character-event distribution for the short story 'We Can Remember it for you Wholesale' by Philip K. Dick[1]. In this example, one character – Douglas Quail – not only takes part in significantly more events than any other entity but is the primary entity in all those events. We can conclude from this that the character is the protaganist of the story.

Further research into the patterns of events is intended in the future. If we return to the Sleeping Beauty example, it would be possible to search for any other stories in which a character is awakened with a kiss, a motif that is seen in many stories, or more broadly stories that contain transformative kisses ('kiss' events which cause transformation events). However the question must be raised as to whether it is also possible to analyse the `Entity`s and `Event`s within a collection of narratives and identify common sequences and, if so, whether these match with the tropes that one would expect. Equally we must ask whether categorising the narratives by similarity would result in divisions that mirror the standard genres and what aspects of the narrative give themselves to such classification.

2.2 Time in Fiction: This Watch is Exactly Two Days Slow

Time in the OntoMedia ontology is modelled with the fundamental assumption that time is neither necessarily linear or constant. In fiction time is often more loose and fluid than in reality with events lasting just the right amount of time to fit the dramatic necessity, maybe out of chronological order within the progression of the narrative or may contain characters interacting with, and potentially changing, their own or other characters' past history. Further the exact time and duration of events is rarely specified, and even when specified may be in a calendaring system that doesn't correlate with any currently, or generally, employed on Earth. To deal with these issues the OM model of time takes an equally loose and fluid method of positioning events within temporal parameters and with respect to each other.

An `ome:Occurrence` is a specific instance of an `ome:Event` which occurs within a single `ome:Timeline`. A `Timeline` represents a specific path through a sequence of `Occurrence`s. This sequence can represent anything from the chronological manifestation of events to a representation of the order in which they are presented in the narrative to the view on the happenings in the world as seen by a character or object. `Occurrence`s have a 1:1 mapping on any given `Timeline` and may hold specific temporal information if it is known or just the relative position of that `Occurrence` to others in the sequence. While a given `Occurrence` is a unique event on a given `Timeline` an `Event` may have multiple `Occurrence`s across many timelines. This not only allows for an `Event` to be represented on multiple `Timeline`s without duplication of data but allows for an `Event` to appear multiple times on one `Timeline` (see Fig 5).

Since each `Occurrence` of an `Event` is a separate instance within its own temporal context, as provided by the `Timeline`, it is possible for `Occurrence`s of the same `Event` to have different durations (see Fig 6), be described in different temporal units or appear in different relative orders (see Fig 7).

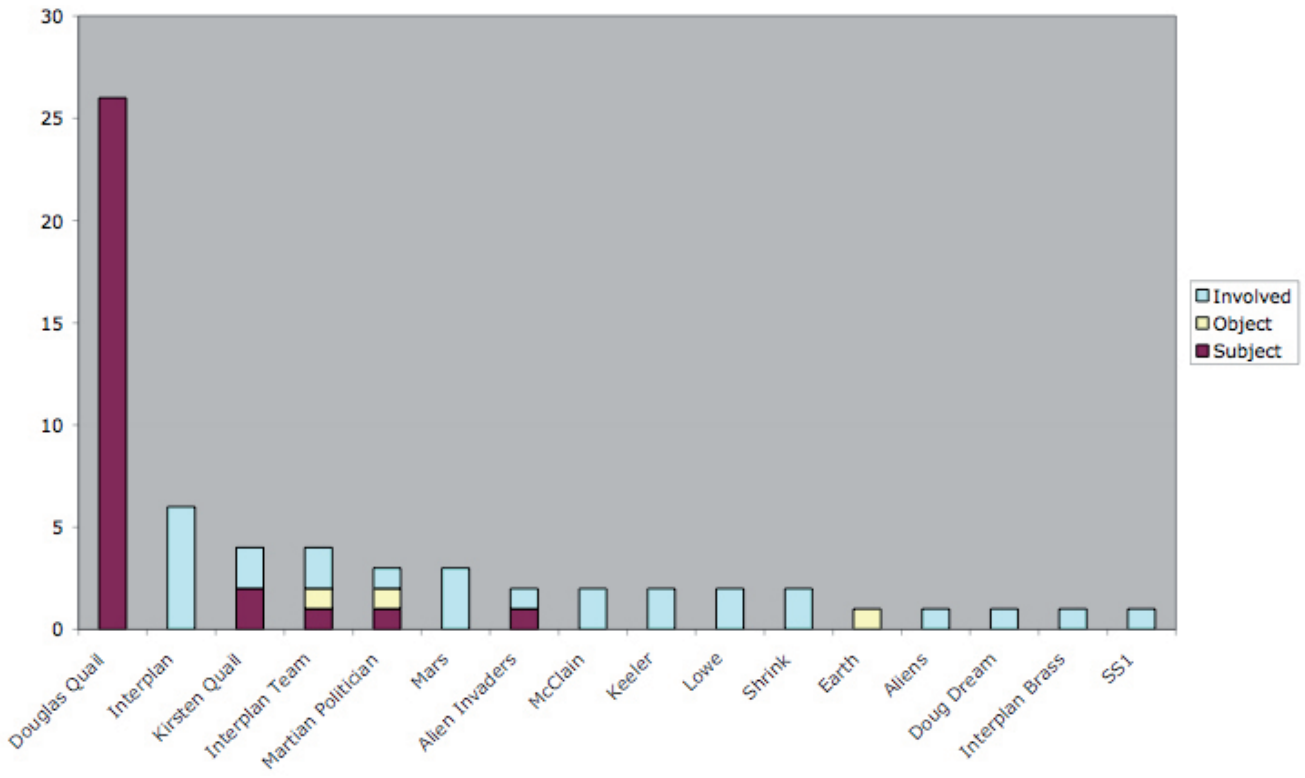

Figure 4: Graph of Character-Event Distribution in 'We Can Remember it for you Wholesale'[1]

A mapping to the Event[1] and Timeline[2] ontologies developed by Yves Raimond and Samer Abdallah concurrently but independently with OntoMedia and uses a similar model is being developed.

2.3 The Eye of the Beholder

Reality, like time, is a concept that is often honoured in fiction more in the breach than in the observance. The idea of multiple universes or realities is a staple in science fiction and fantasy, while dreams and hallucinations are a common theme in all genres. The `ome:Context` class, a subclass of `Abstract-Item`, was created to separate the many different versions of the same `Entity` or `Event` that may exist and to model the existence of many different realities such as fictional universes, dreamscapes, frames of reference or derivations. A `Context` represents a shared truth which we can assume for all the `Entity`s within it. `Context`s can be nested, overlapping or parallel (see Fig. 8). The one reserved `Context` is "Reality" which models what we understand to be the real world.

`Context`s allow us to both contextualise and compare `Entity`s and `Event`s (see Fig. 9). However they also provide a simple way to integrate multiple interpretations of a text rather than necessitating the creation of a canonical description of the events within a text. While the interpretation of the author, where available, can be privileged, this just represents the intended interpretation of the text and may be radically different from how the audience reads and understands the text.

The question of fact, or 'truth', is inextricably linked with that of interpretation. Do we, as the audience, trust the reliability of the narrator? How do we interpret ambiguous text or images? For example, when Hamlet bids Ophelia "get thee to a nunnery", do we, as annotators, interpret this literally – as a reference to a religious establishment – or through the lens of Elizabethan euphemism and, as such, a reference not to the cloister but to a brothel.

This uncertainty can be exacerbated when dealing with visual sources, whether still or moving. While a picture may be held to be worth a thousand words there is no guarantee that it is the same thousand words for every viewer. Indeed, the reverse is more likely to be true and, while the ideal is to provide a neutral annotation of narrative events, the choices that are made by the annotator in how they describe a character or an action will reflect their interpretation of what they see. Two characters kiss – is it an act of friendship, love or violence? A person is seen lying still – are they asleep, unconscious or dead? An event that happened outside the gaze of the audience is alluded to – did it happen or is the speaker lying?

We tested a form based data entry system which allowed users to create basic profiles of characters. For the purposes of the test the details of a non-existent character was described to the volunteers and they entered that information

[1] `http://motools.sf.net/event/event.html`
[2] `http://motools.sourceforge.net/timeline/timeline.html`

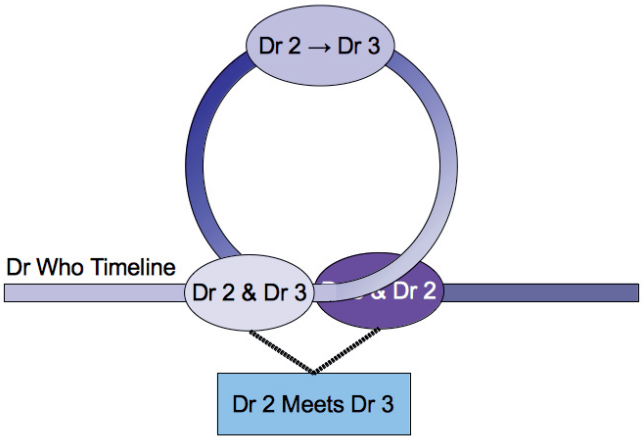

Figure 5: A Diagram of a Loop in The Doctor's Timeline from Doctor Who (BBC) – The Doctor in his Second Incarnation Meets The Doctor in his Third Incarnation, A Single Event that Occurs Twice on The Doctor's Personal Timeline

into the form. The test was intended as a usability experiment on the form, but since all the people taking part were given the same descriptions about the character they were to describe, the data collected also represented multiple interpretations with varying levels of accuracy as one would expect to get from a crowd-sourced annotation of textual content.

For example, Fig 10 shows the `state-of-being` classification, and the level of evidence seen as supporting that interpretation, entered by the ten volunteers when they were informed that it was implied (but never explicitly stated) that the character created for the test was a ghost.

From the data collected is was clear that the majority interpretation was that the character was either 'Dead' or 'Undead' with only two outliers not sharing this viewpoint. While this is only a simple example, further statistical analysis and the use of probabilistic categorisation would allow not only the presentation of the consensus view (or, as in this case, views) but allow the system to compare a given interpretation to others and recommend either complementary views or contradictory ones. In this way the system not only records the multiple interpretations, but also can be used to identify where the ambiguity exists since the presence of the ambiguity itself may indicate a point of interest to researchers. While division of interpretation may highlight an ambiguous portion of the media, and thus a point for discussion of the narrative, it may also highlight the differences in user response. Where this information has been recorded, it can offer valuable information about the way in which material is received and whether it is received in the same way by the different subsections of the audience.

3. FROM CULTURAL HERITAGE TO POP CULTURE AND BACK AGAIN

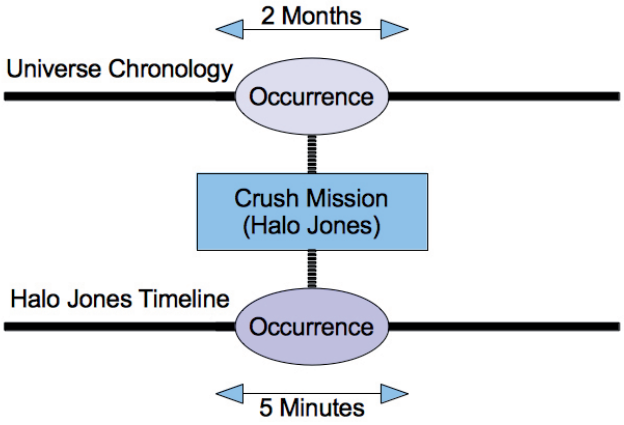

Figure 6: A Diagram Showing One Event Lasting Different Lengths of Time (Example From The Ballad of Halo Jones[18])

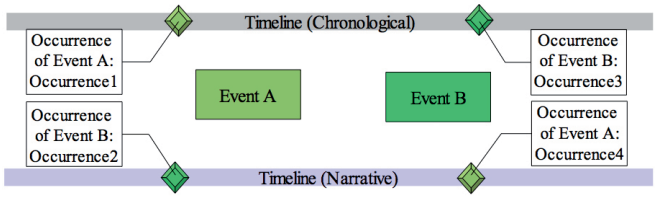

Figure 7: A Diagram Showing Two Events Appearing in Different Orders on Two Separate Timelines

Paul Rissen, based in the BBC's Future Media and Technology department, spearheaded an investigation into how the ontology could be integrated with existing ontologies used within the BBC to allow the description of BBC media content and, through this, improve production, distribution and archiving of such material (Rissen, 2010). Initially the focus was on the BBC's drama output. By describing the elements of the story from key events to characters, locations and objects and creating meaningful links between them it creates the possibility for the user to interact with the material in new ways and follow their own path through the greater narrative rather than following a single, proscribed path.

Fig 11 shows the events of the Doctor Who episode 'Blink' by Steven Moffat[8] within the context of the greater Doctor Who universe. By considering the events as they occur within each characters timeline (see Fig 12) rather than the broadcast narrative the user has the option to view the text from the perspective of a given character. Expanding beyond character-centered entry to a given storyline, the Mythology Engine[3] allowed users to follow the links between people, places and things and view clips of the related sections from the episodes included in the archive (see Fig 13). For example, the story 'Genesis of the Daleks'[20] makes reference to an earlier story 'The Dalek Invasion of Earth'[19]. The events in these two stories are also referenced in 'Journey's End' and 'The Stolen Earth' by Russell T Davies[9, 10] respectively. As Doctor Who is aimed at a primarily young audience who were not alive during the

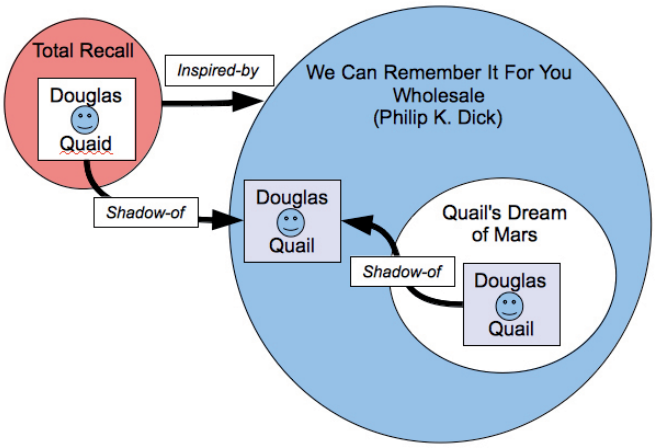

Figure 8: Diagram Showing Examples of Parallel and Nested Contexts (Shown as Circles)

earlier broadcasts this linkage between the modern incarnation of the series and the earlier 'classic' episodes was seen as potentially providing a route into the older material for new viewers.

In addition to making explicit the paths within fictional universes, the same technique allows the elements to be linked to information on real world counterparts, where they exist, thus placing fictional elements within the greater historical or literary context. It was noted that there was a significant spike in google searches for "Pompeii"[3] following the broadcast of an episode set in that location; a trend which has also been noted for other episodes and shows which reference people, places and themes external to the given fictional universe. In the case of the BBC, they have a significant archive of educational material on historical people and events with which to supplement the presentation of their dramatic (and other) content in an interactive, online environment. At the present time these different areas of the BBC's online presence are not integrated at a content level but we have argued[22] that to do so would advance the position of the BBC as information and content provider while supporting and enhancing the user experience.

4. TEI TO RDF

The addition of descriptive metadata to existing material can be problematic; manual entry of data is time and resource intensive, although it can be a successful approach for smaller datasets or where crowd-sourcing can be used. Automating the procedure can significantly increase the efficiency of the system although accuracy can suffer as a result. Jewell[13] investigated whether the information contained in a script marked up with the TEI Performance Texts module[4] could be used to generate OntoMedia RDF. The first stage of this process was to augment the TEI encoded script with RDFa[5] attributes. A conversion programme, `tei2onto`, was

[3] `http://blog.ouseful.info/2009/03/11/the-dr-who-effect-on-google-search-trends/`

[4] `http://www.tei-c.org/P5`

[5] Resource Description Format in Attributes - `http://www.w3.org/TR/rdfa-syntax/`

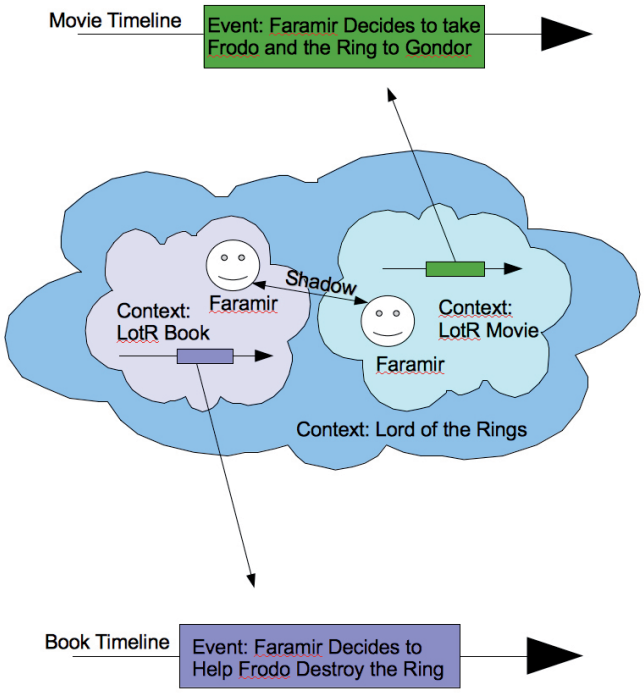

Figure 9: Diagram Showing Different Versions of The Test of Faramir's 'Quality' as Seen in the Book The Two Towers (P. 358, [11]) and The Two Towers Directed by Peter Jackson[7]

developed by Dr Jewell to process the TEI and RDFa and produce valid OM-compliant RDF. This data was ingested into a 4store[4] for storage, search and analysis[6]. The initial test of the procedure was carried out on the original screenplay of 'Dark Star' by John Carpenter[6]. The actor and role TEI elements were used to identify characters with the RDFa information seeding FOAF name data and linking to the DBpedia entry of the person who played the part. The location information was derived from the TEI `stage` elements. By default these generated an instance of the generic `ome:Space` class. RDFa typeof attributes were used to identify specific types of location.

Having created instances of the entities populating the universe described in the screenplay, `tei2onto` identified each element of speech as a social event having the speaker as its primary subject and involving all persons in the immediate location. In addition, the presence of TEI `rs` tags was used to include information about people or places referred to in a given line of dialogue.

The data produced was, by its nature, lacking in detail and did not include any information on non-conversation events (deemed `Social` by default), as it was limited to the information that could be derived from the script. However the experiment was successful in creating a collection of data which could be queried, analysed and linked to other data

[6] `http://contextus.net/datastores`

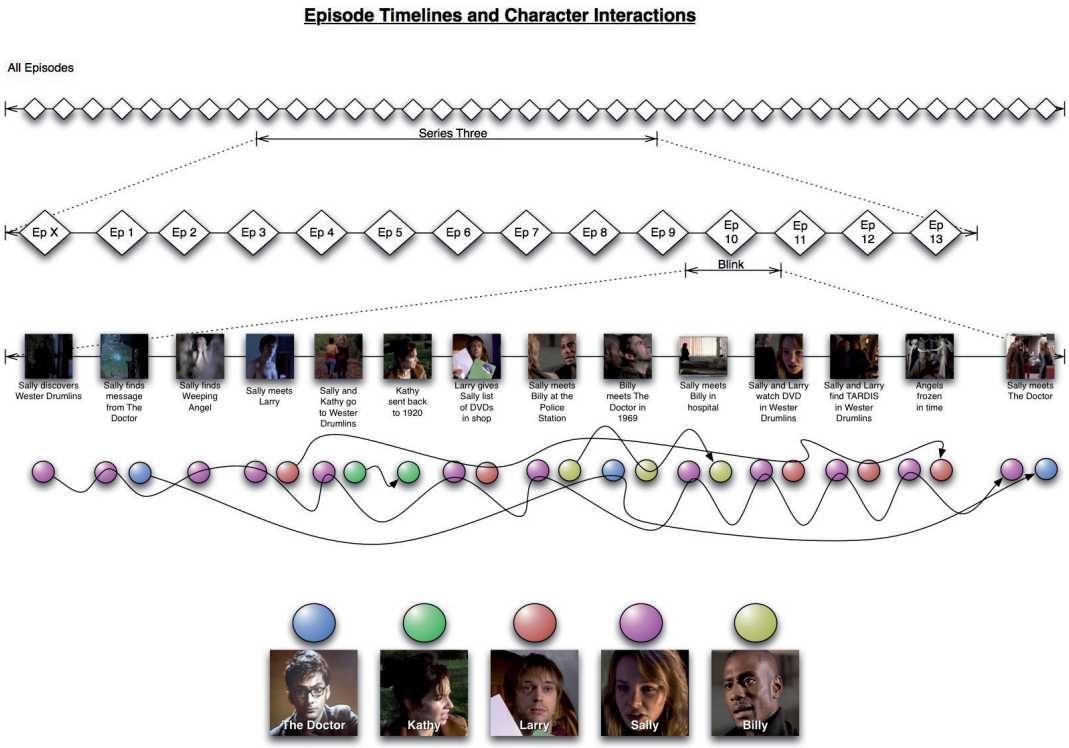

Figure 11: The Timeline for the Doctor Who Episode 'Blink'[8] In Context - Image Created By P. Rissen

Figure 12: The Timelines for the Characters in the Doctor Who Episode 'Blink'[8] - Image Created By P. Rissen

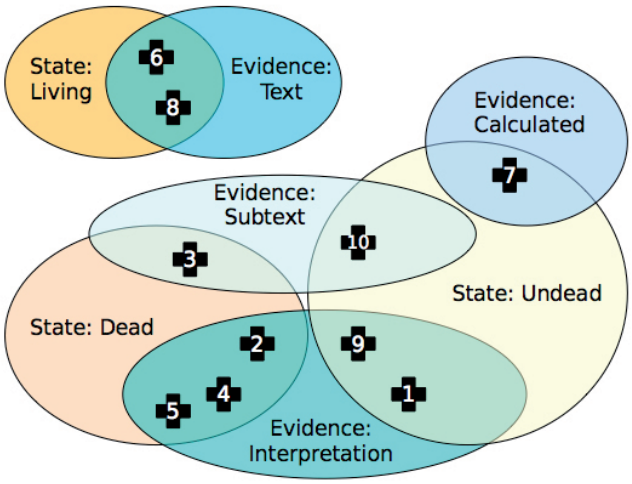

Figure 10: Graph Showing Interpretation of a Given Piece Of Information

Figure 13: A Screenshot of the Mythology Engine Showing the Doctor Who Episode 'Blink'[8]

within the linked data web. The dataset also represented the first pass of information gathering which could then serve as a base for human correction and expansion.

A second version of this experiment is being run with TEI encoded Shakespearean plays available from the Perseus Digital Library[7] acting as the base. In this case there is no actor data being processed as we are concentrating on the play as a narrative rather than in performance. However in addition to the generation of `Social` events, the `tei2onto2` program, expanded by Dr Lawrence, also identifies and creates `Travel` events representing the exits and entrances of the various characters. The scripts were prepared through a combination of XSL transformation to automatically add in the RDFa attributes and a PHP script to identify and encode character names referenced within the text as this information was not included in the Perseus version of the encoding. There is a known problem that this process introduces some inaccuracies into the system since the name recognition within the text is being done by means of basic string matching. Use of more complex natural language processing of the text would lead to a higher rate of accuracy and potentially the identification of characters who are referenced indirectly or by something other than their name. However the level of accuracy was deemed to be acceptable for initial experimentation.

[7] `http://www.perseus.tufts.edu/hopper/`

Having produced the RDF triples the next step in the process is to create a simple method of data checking and correction which will take advantage of the possibilities offered within OntoMedia to model interpretation to allow multiple people to offer corrections which can then be analysed and averaged.

5. CONCLUSION

This paper has presented an overview of some of the work that has been done related to the OntoMedia ontology. More information on the ontology and on these, and other, projects is available at `http://contextus.net/`.

References

[1] P. K. Dick. We Can Remember It for You Wholesale. In *The Preserving Machine*. Ace Books, 1969.

[2] M. Doerr, J. Hunter, and C. Lagoze. Towards a Core Ontology for Information Integration. *Journal of Digital Information*, 4(1), April 2003.

[3] T. Ferne and P. Rissen. The mythology engine - representing stories on the web. BBC Research & Development Blog, BBC, March 2010.

[4] S. Harris, N. Lamb, and N. Shadbolt. 4store: The design and implementation of a clustered rdf store. In *The 5th International Workshop on Scalable Semantic Web Knowledge Base Systems*, October 2009.

[5] J. Hunter. Enhancing the Semantic Interoperability of Multimedia through a Core Ontology. *IEEE Transactions on Circuits and Systems for Video Technology*, January 2003.

[6] IMDB Staff. Dark Star. IMDB, February 1979. http://www.imdb.com/title/tt0069945/ (06/09/2010).

[7] IMDB Staff. Lord of the Rings: The Two Towers. IMDB, December 2002. http://www.imdb.com/title/tt0167261/ (07/09/2010).

[8] IMDB Staff. Doctor Who: Blink. IMDB, June 2007. http://www.imdb.com/title/tt1000252/ (06/09/2010).

[9] IMDB Staff. Doctor Who: The Journey's End. IMDB, July 2008. http://www.imdb.com/title/tt1205438/ (06/09/2010).

[10] IMDB Staff. Doctor Who: The Stolen Earth. IMDB, June 2008. http://www.imdb.com/title/tt1205437/ (06/09/2010).

[11] J. R. R. Tolkien. *Lord of the Rings*. Allen & Unwin, London, 1st edition, 1954 – 5.

[12] M. O. Jewell. *Motivated Music: Automatic Soundtrack Generation for Film*. PhD thesis, Electronics and Computer Science, University of Southampton, 2007.

[13] M. O. Jewell. Semantic screenplays: Preparing tei for linked data. In *Digital Humanities 2010*, June 2010. Paper presented as part of panel: Scanning Between the Lines: The Search for the Semantic Story.

[14] C. Lagoze and J. Hunter. The ABC Ontology and Model. *Journal of Digital Information*, 2(2), November 2001.

[15] K. F. Lawrence. *The Web of Community Trust - Amateur Fiction Online: A Case Study in Community Focused Design for the Semantic Web*. PhD thesis, Electronics and Computer Science, University of Southampton, 2007.

[16] D. B. Lenat, R. V. Guha, K. Pittman, D. Pratt, and M. Shepherd. Cyc: toward programs with common sense. *Commun. ACM*, 33(8):30 – 49, August 1990.

[17] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, and L. Schneider. Wonderweb deliverable d17. the wonderweb library of foundational ontologies and the dolce ontology, 2001.

[18] A. Moore (Illustration by Ian Gibson). *The Complete Ballad Of Halo Jones*. Titan Books, London, 1991.

[19] T. Nation. Doctor Who: The Dalek Invasion of Earth. BBC, 1964. http://www.bbc.co.uk/doctorwho/classic/episodeguide/dalekinvasion/ (06/09/2010).

[20] T. Nation. Doctor Who: Genesis of the Daleks. BBC, 1975. http://www.bbc.co.uk/doctorwho/classic/episodeguide/genesisofdaleks/ (06/09/2010).

[21] I. Niles and A. Pease. Towards a standard upper ontology. In C. Welty and B. Smith, editors, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, Ogunquit, Maine, 2001. ACM.

[22] P. Rissen. Re-imagining the Creative and Cultural Output of the BBC with the Semantic Web. In *Digital Humanities 2010*, June 2010.

[23] M. Tuffield. *Telling Your Story: Autobiographical Metadata and the Semantic Web*. PhD thesis, Electronics and Computer Science, University of Southampton, 2010.

Organic Kinship or Incidental Analogy? Similar Meaning Clusters in and *Correspondances* Between Folklore Texts and Pieces of Poetry

László Z. Karvalics
University of Szeged, Faculty of Arts, Department of Library and Human Information Science
6722 Szeged Egyetem u.2.
Phone: 36-20-5796470
zkl@hung.u-szeged.hu

ABSTRACT

Starting with the spookily accurate coincidences of erotic motifs of a famous poet and the Hungarian folk tradition, we try to portray the hidden framework behind the similarities of different motifs. Using Gell's subtle and holistic model interconnecting the forms and "legacy" of art with the abductive reasoning, we get a point, where lot of "second generation" research questions are raising.

1. INTRODUCTION

> "*No reasonable person could suppose that art-like relations between people and things do not involve at least some form of semiosis*." (Alfred Gell)

The erotic folklore motifs are far more popular to be analized as *iconographical phenomena* than in their *textual forms and versions*. If someone after all prefers texts, the leading "genre" of the examinations is the universe of *folk tales and myths* [1]. However, the majority of this literature is a *simple collection* of "erotic tales", or descriptive *motif indexes and classifications*. It is not so common to "zoom" into the level of specific meaning structures, finding contacts and regular patterns between the erotic motifs and other topics (like Limerov could show, how "erotic motifs occur frequently in the late (komi) mythological texts *associated with the forest theme*" [2].

This one-sidedness is also very common in the world of lyrics and literature. The erotic approach becomes important as an *artistic object* of the writings [3], while the main theoretical challenge is to be able to "reconstruct" the "macrostructure" (like Geoffroy-Menoux interprets Angela Carter's work, who "transformed the children's tales into potent adult fable's", perversely applying "erotic motifs and allusions" into hypertextual hybrid tales form "on the crossroads of paraliterature, oral literature and mainstream literature" [4]. The analytic unit is the text itself, and there is seldom zoom into the "micro" level: *sentences, expressions, phrases, metaphors*. However, these individual *"sub-textual puzzles of meaning"* are also complete *"small worlds"*: *every motif has a holistic explanatory background*.

We make an attempt to take these elementary pieces of erotic motifs underneath a magnifying glass.

First International AMICUS Workshop, October 21, 2010, Vienna, Austria.

2. EROTIC MOTIFS EVERYWHERE

2.1 Representing Impotence in the Poetry of Late Attila József

*I am **drying**, shattering,*
*Growing **old** soon*
Sleeplessly laying down
*On the **desert** ground*

*Fresh and vital **moistures***
*Do not **flushing** humour*
*into my big, **venous meat***
and I mopingly yearn…

This heart-ache is too heavy
To manage it by mind
*Make me **young,** forgiveness,*
Flore's slightly love. (1937)

(Raw translation, L. Z.K.)

One of the biggest 20th century poets, the 32 years old Hungarian Attila József, in a deep and hopeless depression, few months before his tragic death, is holding on to a possible intimacy with his latest Muse, Flore. The key motifs - "*dryness*" versus lack of *"moisture*" and "*flushing*", "*getting old*" versus "*being young*", "*meat*" – are foreshadowing, that all these are not simply about the loneliness, the temporal absence of mating, but about a very explicit manifestation of impotence, whether it is a neuro-anatomical fact or simple role-playing game of a sensitive artist.

One of the preparatory manuscript versions has a far more explicit verse of this "basic message":

It will be hard to enmesh into

*a **hot and fine cuddle** with me.*

*Beautiful Flore strains **dry***

***hunger-weed**[1] to her heart.*

Now let's compare the motifs of Attila József with some Hungarian folk songs and idioms:

Dry hunger-weed doesn't wet

*The **old** girl is not good egg*[2]

or:

*You are **dry bough**,*

*If you'll be **verdant** again*

come back to me.[3]

or:

dry hunger-weed flourishes

*the **old** lady is **on heat***[4]

We can see just the same vocabulary: a kind of "*secondary picture language*" where every special motif is used in a figurative sense, having a strictly defined place in the interlocking system of (symbolic) shadow-meanings.

2.2 The Origins of "Secondary Picture Language"

How to get this stage? In the folk tradition this is a long, but easily detectable process. At the beginning of a story, immediate (esoteric) erotic associations were attached to dozens of everyday, innocent (exoteric) words. As Josef Winthuis explains [5] it was obvious for the gunantuna people in Melanesia and the Australian aborigins to think "vagina" when they were talking about eye, mouth, ear or cave, think "penis", when they were talking about nose, tongue or pike.

The very interesting point is the loss of these archetypic meanings during the civilization process. The contemporary speakers are using lot of folk idioms without knowing anything about their secondary meaning. A studious Hungarian researcher, Béla Bernáth made a giant step ahead to reveal the full system, as a meaning-archeologist in his book [6], deducing the original erotic context and associations in the case of hundreds of well-known folk texts, including nursery rhymes, says, paternosters, love-ditties, figures of speech, songs, old tags, adages, mockering rhymes, etc. He could built a two layer symbol system, with *hundreds* of proven vagina, penis and coitus synonyms as elementary motifs on the lower level, and *plenty of* detectable erotic situations on a higher level, where the speech community were using these motifs in a dynamic way, *(re)combining them constantly in just the same meaning*.

From this aspect we have to modify slightly the basic AMICUS definition of motifs. "*Motifs are complex higher-level patterns that recur in a non-accidental way; they contribute crucially to the function of the story, and have evolved over time, gaining cultural significance along a long path of oral and written transmission and canonization*". The *elementary (erotic) synonyms are also quasi-complex, lower level motifs, with a sophisticated ontology behind them*: all these "simple" motifs[5] are based on the locally and culturally given picture asset of the nature, the human body and the environment, fully overrun with artefacts, objects, types of food and drinks, etc.

The inceptive reason is the (mainly metonymic) similarity of these items with the biological and physical features of sexuality. The homology is constant, (almost) invariant, regular, permanent and persistent, so scanning the full clusters of these motifs provides an excellent possibility to amend and integrate even missing or incomplete pieces of texts.

2.3 The Artificial Version: Picture Creating Power of a Poet

We have to realize, that the logic of motif creation and multiplication is just the same in a full "oeuvre" of a very autonomous world of the poets. To choose an image, to design new verbal constructs and meaning structures is always subjected to strict laws, rules and perceptions, always counting with the decoding and understanding abilities of the readers.

The best example is a well-known Vietnamese poetess, Ho Xuan Huong's magnificent work[6]. Living in the late 18th and the early 19th century, almost her full subtle and witty oeuvre is based on hidden but very coherent and consistent stream of erotic motifs. She applied unique symbols, plays on words, visual tricks using the Chinese pictograms, but first and foremost, she was playing with the pitch-range, using the secondary or tertiary meanings of the syllables, depends on the altitude of the outspeaking voice.

The history of literature calls this solution "*double entendre*"[7] (or *adianoeta* (in rhetoric), and its story has started with Geoffrey Chaucer's The Canterbury Tales in the 12th century[8]. As we could see, Attila József's erotic motifs are using just the same *double entendre*, than the Hungarian folk texts.

Erotic motifs are classical forms of *double entendre* – however, they were not born through individual, accidental and random coinages, like in the folklore, but as a results of conscious composition efforts.

---

[1] hunger-weed = *high grass with hard stalk*

[2] "*Száraz kóró nem nedves, a vén leány nem rendes*"

[3] *"Száraz ág vagy, ha ismét zöldellsz, gyere hozzám"*(Means: I wait for you when you are able and ready for making love again)

[4] "*Száraz kóró virágzik, a vén asszony bogárzik"*

[5] "*Simplicity is complex*" See [10].

[6] About her see: http://en.wikipedia.org/wiki/Ho_Xuan_Huong

[7] French double = *double*, entendre = *to mean, to understand*. Nowadays the French currently refers the phrase with the term „*double sens*", and the original French version became *terminus technicus* in an English speaking world.

[8] The famous Chaucerian "double entendre" is the word "*queynte*", meaning "*domestic duty*" on one hand, and "*queynte*", the early forms of modern English "*cunt*" (vagina) on the other hand.

2.4 The Triple Discovery Hypothesis

Attila József did not know the book of Bernáth Béla. May be he met some direct erotic motifs during his stay in Öcsöd, a small Hungarian village, but it is almost sure, that his cluster of impotence-related motifs is a *sovereign poetic re-discovery of long time evolving folklore motifs*.[9]

The third (re)discovery is taking place in the minds of the readers, the "recipients". In times past the performance and the reception did not come apart: the motifs were well known common goods. Losing the enormous set of these meaning asset, and facing ahead with brand new poetic texts, the understanding becomes investigative: when it is successful, how could it happen?

It is not enough to have almost similar knowledges about the narrow and wide environment, common codes, memory contents, languages and experiences. We need a perfect encounter of our thinking, imagination and picture creating rules with the logic and structure of reality.

2.5 Abductive Thinking: The Common Multiple

This is the momentum which was named *abduction* by Charles S. Peirce, the father of semiotics[10]. The abductive reasoning is the third, non rational, non causal mode of cognition, in contrast the deductive and inductive ways, determining plausibility based on a set of evidence, being able to manage uncertain information very efficiently. "Abductive reasoning use non-sentential representations", and "some abductive inference is better understood as using pictorial or other iconic representations" [10]. The basis of abductive capacity and performance is the right model of the world.

The folk texts are objectivations of collective cognitions, compressing adequate and proved sets of experiences into different genres[11].The poetic neologism and picture creating are also innovative ways of cognition, based on individual experiences and fantasy.

It was Alfred Gell who could apply the notion of abductive reasoning for anthropological research, in his pivotal work [16]. He defined abduction, as *"a case of synthetic inference 'where we find some very curious circumstances, which would be explained by the supposition that it was a case of some general rule, and thereupon adopt that supposition*".

[9] About this common "social manufacturing" of folklore and literature see Susan Stewart's classic monography[7].

[10] See more details about the abduction in Thomas Sebeok's book[8], and a fresh contribution in the context of grounded theory [9].

[11] The researchers has already started to simultaneously apply the abductive reasoning considerations and reviewing folklore materials (mythology, fairy tales, and other folklore texts), see Mira Karjalainen's theses [11].

Approaching Gell's significance, an excellent summary[12] emphasizes, that "Gell criticizes existing 'anthropological' studies of art, for being too preoccupied with aesthetic value and not preoccupied enough with the central anthropological concern of uncovering 'social relationships' specifically the social contexts in which artworks are produced, circulated, and received. Abduction is used as the basis of one gets from art to agency in the sense of a theory of how works of art can inspire a *sensus communis*, or the commonly-held views that a characteristic of a given society because they are shared by everyone in that society". Abduction means "that not only can it operate outside of any pre-existing framework, but moreover, it can actually intimate the existence of a framework... the physical existence of the artwork prompts the viewer to perform an abduction that imbues the artwork with intentionality... art can have the kind of agency that plants the seeds that grow into cultural myths. The power of agency is the power to motivate actions and inspire ultimately the shared understanding that characterizes any given society[13]".

3. "*CORRESPONDANCES*" BETWEEN OTHER REALITY SEGMENTS

3.1 The Message of David M. Guss

David M. Guss is an excellent anthropologist of the Tufts University. Several years ago he tried to translate the famous yekuana mythic text, the Watunna. He had to quickly realize, that to understand the universe of the myth, he has to learn to be able to become abductive. "I had originally come to learn demanded much more than just verbal skills. It required the use of all my senses or, more precisely, a reorientation to the nature of meaning and the manner of its transmission"[12:4].

The yekuana society through the prism of Guss is a laboratory of meanings. „Each activity, whether ritual or material, was determined by the same underlying configuration of symbols. Thus whatever an action's external form or particular function, it was involved in the same dialogue as the rest of the culture, communicating the same essential messages and meanings. It was truly a mutually reflective universe in which every moment was filled with the same possibilities of illumination as any other. To tell a story, therefore, was to weave a basket, just as it was to make a canoe, to prepare barbasco, to build a house, to clear a garden, to give birth, etc".

In a society such as the Yekuana's, it was possible "to see the entire culture refracted through a single object or deed. Every part was a recapitulation of the whole, a synthesis of the intelligible organization of reality that informed every other".

For Guss, to learn the basket-making was the crucial, abductive point. As such, the baskets provided a prism through which the Yekuana universe was reflected. "… cast in a metaphor of endless dualities, the symbols in the baskets, like those elsewhere, confronted the most elemental oppositions between chaos and order, visible and invisible, being and non being. The concept of culture which they presented was not simply one of communication, or what Geertz calls „ a mode of thoughts", but also of transformation, of the constant metamorphosis of reality into a comprehensible and coherent order".

[12] http://en.wikipedia.org/wiki/Alfred_Gell

[13] See more from the „Information community" approach [20].

As both an observer and a basket maker, he "was able to participate in this process of transformation, to experience culture not as the distillation of a set of abstract ideals but as an ongoing act of creation".

Of course, apart from basketry [13] there are lot of other ways of activity (or domains of culture), where we can find just the same connections and homologies. In conclusion, let's look at some very interesting and also motif-based ones.

3.2 Other *Correspondance* Possibilities

Ferrer [18] provides a long enumeration, recycling the results of Java field-work of Geertz [19]: the motifs in *Batik, Arabic calligraphy, wood-carving, jewelry, silverwork, weaponry, music, and the performance arts of dance, theater, and puppet theater* (*wayang kulit*) are interpretable using the same framework. Gell [17] adds the *tattoo* to the list, and gives an example, how a *statue of a goddess*"in some senses actually becomes the goddess in the mind of the beholder; and represents not only the form of the deity but also her intentions (which are adduced from the feeling of her very presence)" [16].

However, it is not a question, that the most serendipitous field is the *music*. Recently the Hungarian physicist and musician, Zoltán Juhász processed the melody strings (musical motifs) of folk songs of enormous quantity [14] by computer, distilling amazing, surprising and wondrous generic relationships between different cultures.

It seems to be attack anywhere the hidden framework presented by Gell, we can find the much-discussed homologies.

4. FURTHER RESEARCH QUESTIONS

This deep "framework" behind the surface of motif flow and manifest motif clusters has to be documented through detailed investigations, particularly being concerned with evolutionary and cross-cultural aspects. We need a methodology to be able to capture the "semantic synaesthesia" and turning it an applicable *pattern-recognition* and *pattern-reconstruction* technology. It is also important to come near to the *stratification of elementary and more complex motifs,* and motivating the fractal-like nature of motifs, discovered by Guss.

5. ACKNOWLEDGMENTS

Many thanks to Piroska Lendvai and Sandor Daranyi, whose sweet-tempered force led me a non-intended shrubbery of anthropology literature, influencing my other research ramblings.

6. REFERENCES

[1] Malotki E. (eds) 1997: The Bedbug's Night Dance and Other Hopi Tales of Sexual Encounters University of Nebraska Press

[2] Limerov, P.F. 2005: Forest myths: a brief overview of ideologies before St.Stefan *Folklore* Vol.30. October http://www.folklore.ee/Folklore/vol30/limerov.pdf

[3] Classen, A. 1994: Love, Sex, and Marriage in Late Medieval German verse. Narratives, Lyric Poetry, and Prose Literature *Orbis Litterarum* Vol.49, No. 2. p 63-83 April 1994

[4] Geoffroy-Menoux, S. 1996: Angela Carter's The Bloody Chamber (1979) : the Double Pastiche of Twice Harnessed Folk-Tales, or How to Re-Write Popular Stories *Paradoxa* Vol.2. No.2. pp. 249-262.

[5] Winthuis, J. 1928: Das Zweigeschlechterwesen bei den Zentralaustraliern und anderen Völkern, Leipzig

[6] Bernáth, B. 1986: A szerelem titkos nyelvén. Erotikus szólások és folklórszövegek magyarázata. (Gondolat, 1986 pp.5-350.) (*On the secret language of love. Explanations of erotic idioms and folklore texts*. In Hungarian)

[7] Stewart, S. 1989: Nonsense: Aspects of Intertextuality in Folklore and Literature The John Hopkins University Press

[8] Sebeok, T- Sebeok, J.U. 1981: "You Know My Method": A Juxtaposition of Charles S. Peirce and Sherlock Holmes In: Sebeok, T: The Play of Musement, Bloomington, Indiana pp. 17-52.

[9] Reichertz, J 2009: Abduction: The Logic of Discovery of Grounded Theory FQS (Forum: Qualitative Social Research Vol.11.No.1. Art 13. http://nbn-resolving.de/urn:nbn:de:0114-fqs1001135

[10] Thaggard, P.- Shelley, C. 1997: Abductive reasoning: Logic, visual thinking and coherence In: In M.-L. Dalla Chiara et al. (Eds.), *Logic and scientific methods*. Dordrecht: Kluwer, pp. 413-427.

[11] Karjalainen, M. 2004: Prisoners of freedom. A study on worldview of contemporary Finnish seamen University of Helsinki Licentiate Thesis https://oa.doria.fi/bitstream/handle/10024/58955/prisoner.pdf?sequence=2

[12] Guss, D.M. 1990: To Weave and Sing: Art, Symbol, and Narrative in the South American Rainforest University of California Press

[13] Reichel-Dolmatoff, G. 1985: Basketry As Metaphor: Arts and Crafts of the Desana Indians of the Northwest Amazon (Occasional Papers Series No. 5) University of California Los Angeles, Fowler p. 1-104.

[14] Juhász Z. 2006: A zene ősnyelve Fríg Kiadó (*The primaeval language of music*. In Hungarian)

[15] Oliveira L.F. et al. 2010: Musical Listening and Abductive Reasoning. Contributions of C.S: Peirce's Philosophy to the Understanding of Musical Meaning *Journal of interdisciplinary music studies* Vol. 4, No.1, pp. 45-70.

[16] Gell, A. 1998: Art and Agency: An Anthropological Theory. Oxford: Clarendon.

[17] Gell, A. 1993: Wrapping in Images: Tattooing in Polynesia. Oxford: Clarendon

[18] Farrer, D.S. 2008 The Healing Arts of the Malay Mystic *Visual Anthropology Review* Vol. 24, No.1, pp. 29–46,

[19] Geertz, C. 1976: The Religion of Java. Chicago: University of Chicago Press

[20] Z. Karvalics, L. 2002: Információközösségek. Kísérlet egy fogalom megragadására. (*Information Communities. Approaching a Notion* In Hungarian) In: Mobilközösség – mobilmegismerés Szerk: Nyíri Kristóf Budapest, MTA Filozófiai Kutatóintézete – Westel pp.19-40.

Granularity Perspectives in Modeling Humanities Concepts

Piroska Lendvai
Research Institute for Linguistics, Budapest, Hungary
Tilburg centre for Creative Computing, Tilburg University, Netherlands
piroska@nytud.hu

ABSTRACT

Our pilot study focuses on the issue of concept granularity, at the intersection of the Digital Humanities and Language Technology disciplines. Based on our own work and on contributions by several authors in the current volume, we would like to initiate the charting of approaches taken from different fields of computational linguistics to the analysis of higher-level content units, typically in folk tale texts, for establishing such units and their composition.

1. INTRODUCTION

The target of several ongoing national and international projects is to establish or agglomerate digital research infrastructures for communities that have so far been less exposed to computational tools for data processing and analysis. Humanities and Social sciences (HSS) disciplines, where primary sources for study are typically text-based (such as literature, law, history, philosophy, religion, ethnography, etc.) are one of these new areas whose traditional research activities are getting enhanced by new and faster methods of text analysis or synthesis. Directly applying already existing tools and frameworks from Human Language Technology (HLT), as well as developing their own, the Humanities community is increasingly engaged in emerging paradigms of the Digital Humanities (DH) discipline. Issues in this field are growing complex, however, as the DH community necessarily mingles with that of HLT.

DH is clearly more than ways and means to enhance the individual researcher's ability to organize and navigate large amounts of data, but is rather a dynamic new field with a focus on exploring possibilities of linguistic and textual analysis, such as semi-automatic semantic annotation, thesauri, concordances (cf. [1]). Most importantly, applying and developing computational approaches for HSS data needs to take into account the legacy of needs and specificities of the types of research questions and research materials in these fields; the survey in [2] aims to provide an overview of these.

First International AMICUS Workshop, October 21, 2010, Vienna, Austria.

It appears that the complexity of typical HSS research issues, together with the modules of possibly applicable HLT solutions might be viewed as a matrix of layers, components, and granularity, in which it is not straightforward to identify correspondences in terms of which HSS and HLT research categories can or could be coupled – the latter being one of the core aims of DH. In recent initiatives, such as by the efforts of the CLARIN project[1], linguistic analysis and processing tools, corpora and repositories holding linguistic information, as well as analysis and synthesis methods of HLT, all lying at the basis of such facilities, are being brought to HSS specialists. It has been noted that HSS scholars (just like all other non-HLT experts) are often unaware that automatic linguistic analysis is a prerequisite of higher-level content processing.

Based on the materials presented at the First International Workshop of the AMICUS project[2] that focuses on automated motif discovery in cultural heritage and scientific communication texts, we aim to report on the approaches and viewpoints of HLT specialists with experience in DH, on the issue of compositionality appearing in their suggested approaches and models of Humanities concepts. The role of the authors of these studies, being creators of DH resources – as opposed to users and curators of these – is to plot the mechanisms that produce structures that can be understood as 'motifs', to map them to computational procedures, resources, or tools, connecting the HSS and HLT disciplines.

Such a pioneering venture has several challenging aspects, out of which we propose to pay attention to but a few ones:

- A certain concept or notion for one scientific field might be fundamental, and thus regarded as atomic, whereas in another field (serving a different purpose) it might or should be modeled as *compositional*;
- Cases when it is not yet possible or desirable to make some notion specific, i.e. explicit, resulting in *underspecification;*
- The mingling of pragmatic and semantic features, and modeling possibilities for *interfacing* these two traditional linguistic levels.

The goals of the authors of papers in this paper are different (analysis, generation, automated markup), but common underlying aims seem to be arriving at "psychologically compelling results" [Malec, this volume], or detecting as much structure as possible from raw texts in order to achieve content analysis (e.g. [Declerck and Scheidel, this volume]) or synthesis

[1] http://www.clarin.eu

[2] http://amicus.uvt.nl

(e.g. [Gervás, this volume]), utilizing semantic concepts from both the HSS and the HLT communities.

2. ADDRESSING FOLK TALE TEXTS IN TERMS OF HLT

The goal of the AMICUS project is to distill knowledge about constructs from texts that seem weakly structured to a machine or to non-experts in (Digital) Humanities. It is possible for skilled humans to infer semantic, or even subtle cognitive structures in these, nevertheless, such mechanisms are even manually difficult to capture by one of the mainstream methodological approaches of natural language processing (NLP), for example, in terms of identifying conceptual boundaries in text (e.g. finding names of heroes in tales, or Proppian functions), assigning category labels to these, and arranging them in a (hierarchical) structure. Once such tasks are solved, automatic annotation is easier to design and enhance by means of picking cues for determining (sequence) similarity, and feeding in structured resources.
To address this gap between HLT and human cognitive processes that produce folklore texts or argumentative presentation of research results, the metaphor AMICUS operates with is that of a *motif*. The project faces the challenge of investigating whether a motif is an underspecified concept or mechanism as such, and whether this property of motifs contributes to or enables devising better HLT models in the HSS fields, or not.

2.1 Semantic and Pragmatic Levels

It turns out from our investigations during the first half-year of the AMICUS project that working with motifs entails examining narration by which they exhibit themselves in text. Just like all types of material carried by spoken or written language, narrative on the higher linguistic levels operates by pragmatics and semantics. Features of the pragmatic level are expressed by properties of discourse, carrying communicative goals, reflected e.g. by stylistic markers and arrangement of the content. The semantic level pertains to content: the kind of entities, facts, and events appearing, but, importantly, also the abstract domain knowledge expressed as categories – such as the notions of 'entity', 'fact', and 'event'.

It appears from, although not stated by, the systematic survey of [Gervás, this volume] that narrative operates by means of both low (i.e. easily detectable by machines) and high level features (i.e. difficult to formalize) on both the semantic and pragmatic levels. An important HLT issue with respect to this, but unclear at the moment, is if it is possible to classify these features and concepts as pertaining to – or, modellable by – exclusively the semantic or the pragmatic level. We note that this is inherent in language, applicable to possibly several, if not most, HLT levels (investigated by us in [3] in the HLT domain of dialogue processing). However, the issue now is whether treating this as a general strategy, i.e. (manifesting itself as) underspecification rather than compositionality, brings one closer to the goals of AMICUS.

The strategy of underspecification (or, opacity) here is that in fact (some) pragmatic and (some) semantic phenomena co-depend on one another. This might entail that separating those by grouping under different linguistic levels and labels is a convenience issue, by now a burdensome legacy, suboptimal in the DH discipline. In other terms, such modeling could pose an obstacle to optimal processing and might risk obtaining valid DH models from and creating valid approaches to Humanities data.

In the following section we will illustrate this line of thought by examples.

2.2 Matrix of correspondences

In Table 1 we attempt to initiate the sketch of a matrix, quoting concepts and the terms used to denote them as suggested in some of the studies published in the current Proceedings, as well as arranging them in a simple structure, reflecting a lightweight and non-systematic hierarchy. The goal is to align related concepts occurring in some of the research papers as belonging to the so-called pragmatic concepts (top section), respectively to the semantic concepts (bottom section). Sometimes these denote similar phenomena or perspectives, expressed by different terms, sometimes they complement one another.

With respect to compositionality, it is clear that the picture is extremely complex, and, judged already by the fact that examples for pragmatic features are only possible to express by means of involving/quoting actual content elements, it seems to be indeed the case that pragmatic and semantic phenomena co-occur in intertwined ways. Events and event sequences are high-level content conglomerates, as we noted above, since the typical elements involved (time, location, action, means, etc.) are themselves possibly composed of multiple (linguistic) entities; the same goes for motifs. For example, a person in history or a character's name in a fairy tale in fact refers to a single entity, but in order to be able to represent e.g. 'Churchill', we need to formalize a range of properties such as first name, gender, etc. in order to disambiguate, deduplicate references, link data, and so on. The same mechanism applies to many other concepts, such as time -- but not necessarily in all genres: fairy tales do not involve explicit dates.

In Table 1 we additionally attempt to signal the difficulty of the required linguistic processing level, which would indicate the feasibility of possibly applicable HLT methodologies: manual or automated. It is common practice that content-related phenomena can be successfully dealt with by gazetteer approaches (i.e., identifying the entities based on a fixed list), such as in Information Extraction (IE). It is to be judged on the basis of actual implemented procedures whether such semantic concepts would/could be subsequently returned to for further specification and/or revision, once the pragmatics involved in expressing them has been identified and separated from content, keeping them underspecified for a while (cf. the approach sketched in [Declerck, this volume]).

In general, since only part of the phenomena listed in Table 1 have been formalized so far, we would like to invite all interested parties to contribute to filling in this, or a similar, matrix, thereby benefitting from input given by representatives of all communities related to analysis of high-level conglomerates in narrative genres. It requires thorough investigation if the sketched matrix would apply to both folk tales (and other literary genres) as well as to scientific communication texts. The table of correspondences in [de Waard, this volume], comparing story grammar elements with syntagmatic components of scientific text, contains (high-level) data categories and ideas that are readily importable.

The matrix could suggest intersections identified between HLT and DH-HSS, and one may understand in more detail which aspect of narration and motif are carried by which formalizable linguistic levels. We also propose to prepare a gold-standard example set marked up by the relevant phenomena.

3. MOTIFS

Motifs are an example of the vehicles of the semantics of communication, for example in folk tales, but possibly in all kinds of discourse. They contain sets of content objects in a predefined way, the segmentation and granularity of which are perhaps roughly equally distributed within a story. It seems that even when some heavy artillery of HLT is utilized, motifs remain troublesome to address, since establishing and computing the segment boundaries of their constituents proves to be complex. The work initiated within the AMICUS network, receiving its initial form in [4], will hopefully alleviate some of the problems related to this. In this section we propose to tackle some further aspects, possibly lesser addressed in previous research of motifs.

It would be interesting to examine which constituent of a motif should change in order to require the update of the motif – which is in fact the moving on to a new motif. Establishing such a criterion might be an indicator for motif segmentation in certain constellations.

Table 1. Pragmatic features (top section) and semantic concepts (bottom section) in approaches to (structural) narrative analysis (cf. Propp) and generation (cf. Gervás), information extraction (cf. Declerck), machine learning (cf. Malec), access to knowledge from scientific papers (cf. de Waard). We indicate the typical linguistic processing level that could be associated with each object in a linguistic processing module.

Analysis feature		Linguistic manifestation e.g.	Example and possible cue	Linguistic analysis level (difficulty: Low, Medium, High)
Mood	Distance	Reported speech	"Once upon a time there lived an old man and an old woman."	Lexical (d:L)
	Voice: Narrator present/absent (IE: 'speaker') Narrator's Function (neutral, directing, ideological…)	Cue words, Grammatical person	"Dear children"	Lexical (d:L) Coreference (d:M)
Time (points and intervals)	Narrative order (Prior, simultaneous, subsequent, interpolated)	Temporal expressions	Event has happened *before*/is going to happen *after* narration	Pragmatic/Semantic interface (d: M/H)
	Presentational ordering	Timeline (underspecified)		(d: M/H)
	Speed/Pace	Event's descriptive granularity	Summary of what happened vs (quoting) a dialogue	Semantic/Cognitive (d:M/H)
	Frequency of mentioning an event	Repetition, Paraphrase, Summary	…	(d:L/M)
Perspective	Focalization: Zero, internal, external	Narrator knows more/equal/less than characters	…	Cognitive (d:H)
	Focalization on individual characters	First person singular	Represent individual perspectives and beliefs of different characters	Depending on underlying representation model (d:L)
Proppian function	textual unit that moves the story forward, irrespective of the dramatis personae	Negation, imperative	"Do not leave the house"	Domain knowledge (d:H)
story	abstract schema of systematic constituent structure	Cue words for Goal, Subgoal, Method…	"the aim of this study is…"	Domain knowledge (d:H)
Entity	IE: - Character - Sets of Characters of same type?	gazetteer, coreference, family relations from ontology	Attributes: age, synonyms?	Semantic/Cognitive (d:M/H)
	- Object - Sets of objects of same type	Gazetteer, cue phrases	apple	Lexical (d:L)
Attr	Profile/adjective	gazetteer	golden	Lexical (d:L)
event	When, Where, WhatAction, who to Whom, Why, How	…	…	Domain knowledge, Pragmatic/Semantic interface (d:H)

We would also like to understand in what ways motifs (not necessarily in the Proppian sense) pertain to the pragmatic level. Clearly, the 'meaning' of a motif is more than the agglomerate of its elements. By a similar mechanism, the units 'youngest' + 'son' would equal to 'Character: hero'. It would be necessary to put into context the underlying formalism of cognitive processes in such cases.

4. INTERDISCIPLINARITY

In DH studies typically some close, goal-oriented research collaboration takes place between the language technology experts and the HSS researchers, where distinct disciplines come together, for example philosophy or sociology, as well as computational linguistics. It is important to ensure that the knowledge of the different communities involved is in such cases as optimally merged as possible in order to facilitate research and to maximize the success of the given project. One crucial aspect in this is that specialists belonging to the different communities need to make sure they understand each other's goals, methods, and, last but not least, terminology. This requirement however necessitates obtaining several insights about a new domain for all parties, which can be realized only by consciously investing time and effort in this process.

There is a clear need at several phases of a research project that representatives of different communities engage in in-depth dialogue with each other in order to gain new knowledge about each other's research foci, priorities, domain concepts, and language use. The establishment and practical implementation of this important layer of collaboration call for the identification of prerequisites that allow for optimally managing and realizing transdisciplinary research, involving simultaneous coordination and evaluation of steps to be taken during what we see as recursive exposure to knowledge from a new domain for both HSS and HLT specialists.

In the current section we would like to raise some issues, which, when answered, could possibly take us closer to our goal of integrating HLT into HSS. For example, it is an intriguing question whether some HLT methods can be readily utilized or adapted for fast modeling of a HSS domain, i.e. for knowledge acquisition and representation, in order for the HLT specialist to obtain a quick overview of it, as an alternative for reading up on a domain from textbooks and publications? Possibly, IE is capable of supplying such a method, semantic technologies likewise.

We also suppose that HLT experts need to be aware of the nature of typical HSS research methodologies. For example, it is not known if the presence of tools for automatically processing data on a certain linguistic level implies a must for applying the tool, even if HSS researchers initially think this information is irrelevant for their project.

We expect such questions to be answered as new research paradigms emerge from cross-fertilization among communities.

5. ACKNOWLEDGMENTS

Our thanks to Thierry Declerck for comments on the draft of this paper.

6. REFERENCES

[1] *HERA Survey on Infrastructural Research Facilities and Practices for the Humanities in Europe*. 2006. DOI= www.ucm.es/info/eurohum/docs/heraonhumanities.pdf

[2] Lendvai, P., Váradi, T., Wynne, M. and Berglund, Y. 2010. *Humanities and Social Sciences Organizations, Initiatives and Projects Report*. CLARIN Deliverable D3C-3.2. DOI= http://www-sk.let.uu.nl/u/D3C-3.2.pdf

[3] Lendvai, P. 2004. *Extracting Information from Spoken User Input. A Machine Learning Approach*.

[4] Declerck, T., Scheidel, A. and Lendvai, P. 2010. *Proppian Content Descriptors in an Augmented Annotation Schema for Fairy Tales*. In: Sporleder, C. and Zervanou, K. (eds.): Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Lisbon, Portugal, IOS Press, European Coordinating Committee for Artificial Intelligence.

POSTERS EXHIBITED

APftML – Augmented Proppian fairy tale Markup Language

Antonia Scheidel
DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg, 3
66123 Saarbrücken, Germany
Antonia.Scheidel@dfki.de

Thierry Declerck
DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg, 3
66123 Saarbrücken, Germany
declerck@dfki.de

ABSTRACT

This poster submission presents the actual state of development of a markup scheme that combines narrative and linguistic information for the fine-grained annotation of folktales. The scheme builds on and extends an existing markup language called PftML (Proppian fairy tale Markup Language) and combines this with textual and linguistic annotation standards as proposed by TEI (Text Encoding Initiative) and ISO TC37/SC4 on language resources management. We call our scheme therefore APftML (Augmented Proppian fairy tale Markup Language). While the poster itself will show detailed examples of the application of the annotation scheme to German versions of "Little Red Riding Hood" and "The Magic Swan Geese", the paper concentrates on describing the resources we have been using, developing and integrating in APftML, which is providing in fact the goal annotation structure of on-going work on the automated semantic annotation of folktales.

1. INTRODUCTION

The work we describe here is part of the projects CLARIN[1] and D-SPIN[2]. While CLARIN is focusing on the establishment of an integrated and interoperable research infrastructure of language resources and technologies that aims at enabling eHumanities research in cooperation with Human Language Technology (HLT), the D-SPIN project, which is the German contribution to CLARIN, is additionally providing for integrated language processing Web services that generate linguistic annotation, which can be concretely used in eHumanities research.

Our particular goal in this context is to integrate linguistic annotation and markup in the fields of folk and fairy tales both in a Markup language and in an automated processing chain. In a first step, which is described in this poster, we propose the combination of standardized linguistic annotation frameworks with a fine-grained annotation scheme that is implemented in accordance with concepts introduced in [3][3].

As a first example. we chose to annotate German version of "Little Red Riding Hood. This annotation exercise is planned to be extended to most of the folktales[4] from the Brothers Grimm's collection, as they are available within the Gutenberg project[5]. In collaboration with the AMICUS project[6] we also propose the annotation of a German version of "The Magic Swan Geese" and will extend this exercise to more tales included in [1][7], also considering versions of the tales in other languages, like English, Hungarian and Russian.

2. THE RESOURCES

Among the sources for our work, besides Propp's "Morphology of the Folktale", we consider Scott Malec's PftML[8], the ProppOnto Ontology[9], FrameNet[10] and the TEI[11] and ISO TC37/SC4[12] standards for textual and linguistic annotation. We concentrate here on the resources described by Vladimir Propp and on the PftML scheme, which we extend into APftML (Augmented Proppian fairy tale Markup Language), using TEI and ISO TC37/SC4 annotation standards.

2.1 "Morphology of the Folktale"

In his study of the Russian folktales, Propp aimed at breaking down those tales to smaller and recurrent narrative units, also called narratemes. We summarize here the main outcomes of his studies:

7 Characters. Propp puts forward the notion that the folktale know no more than seven *dramatis personae*: The villain, the donor, the helper, the princess (or "sought-for per-

[1]http://www.clarin.eu/external/
[2]http://weblicht.sfs.uni-tuebingen.de/englisch/index.shtml

First International AMICUS Workshop, October 21, 2010, Vienna, Austria.

[3]See also http://en.wikipedia.org/wiki/Vladimir_Propp.
[4]some of folktales collected by the Grimms deviate too far from the "magic tale" on which Propp based his theory
[5]See http://www.gutenberg.org/
[6]http://amicus.uvt.nl/
[7]See also http://en.wikipedia.org/wiki/Alexander_Afanasyev
[8]http://clover.slavic.pitt.edu/sam/propp/theory/propp.html
[9]http://www.fdi.ucm.es/profesor/fpeinado/projects/kiids/apps/protopropp/
[10]http://framenet.icsi.berkeley.edu/
[11]http://www.teic.org/index.xml
[12]http://www.tc37sc4.org/

son") and her father (sometimes treated as two characters, resulting in a total of 8), the dispatcher, the hero and the false hero.

31 Functions. At the heart of *Morphology of the Folktale* is the introduction and detailed description of 31 "functions", i.e. (mostly) actions which can be attributed to the dramatis personae of a folktale. According to Propp, every folktale consists of a subset of these 31 functions, arranged in one or more "moves". The order of the functions is fixed, with a number of scrupulously defined variations. Functions are frequently divided into sub-functions: In the case of function *A: Villainy*, they range from *(A1): The villain abducts a person* to *A(19): The villain declares war*[13].

A sequence of all the functions from one folktale is called a "scheme" and can be used as a formal representation of the tale (see Fig. 1 for an example).

$$\gamma^1\beta^1\delta^1A^1C\uparrow\{[DE^neg.Fneg.]^3d^7E^7F^9\}G^4K^1\downarrow \\ [Pr^1D^1E^1F^9 = Rs^4]^3$$

Figure 1: Functional scheme for *The Magic Swan-Geese*

150 Elements. In Appendix I of *Morphology of the Folktale*, Propp provides what he calls a "list of all the elements of the fairy tale". The list contains 150 elements, distributed over six tables:

1. The Initial Situation
2. The Preparatory Section
3. The Complication
4. The Donors
5. From the Entry of the Helper to the End of the First Move
6. Beginning of the Second Move

Some of the 150 elements appear alone, others are grouped under a descriptive heading. If these "element clusters" (as shown in Fig. 2) are counted as one, the appendix contains 56 - as they shall tentatively be called in the following - narratemes[14].

About a third of the narratemes can be mapped directly to functions, such as the aforementioned *30-32. Violation of an interdiction*. Other narratemes can be combined to form an equivalent to a function (together, narratemes *71-77: Donors* and *78: Preparation for the transmission of a magical agent* can presumably be considered as a superset to the information expressed by function *D: First Function of the donor*.

```
30-32.  Violation of an interdiction
       30.  person performing
       31.  form of violation
       32.  motivation
```

Figure 2: Example for a narrateme

Another group of narratemes, however, goes beyond the 31 functions: *70. Journey from home to the donor*, for example, can be seen as filling the gap between the functions *↑: Departure* and *D: First Function of the donor*. The first table (*The Initial Situation*[15]) contains a multitude of narratemes dedicated to the circumstances of the hero's birth and other events/situations which precede the actual adventure.

Furthermore, Table 1 (*The Initial Situation*) includes two "element-clusters"[16] describing the hero and false hero, respectively (see Fig. 3). A closer examination of the appendix reveals such "profiles" for each of the dramatis personae, although sometimes spread over several element clusters.

```
10-15.  The future hero
       10.  nomenclature; sex
       11.  rapid growth
       12.  connection with hearth, ashes
       13.  spiritual qualities
       14.  mischievousness
       15.  other qualities
```

Figure 3: Example for an element cluster serving as profile for a character

In *Morphology of the Folktale*, Propp provides an analysis of "The Magic Swan-Geese", resulting in the scheme shown in Fig. 1 above. It is important to note here that the analysis does not only make use of functions but also of a "list of all the elements of the fairy tale" (given in Appendix I of *Morphology of the Folktale*). For example, Propp annotates the first Donor section from "The Magic Swan-Geese" as shown in the example below:

[13]It is those subfunctions, which introduce "arguments" to the functions and which contain some linguistic material, that led us to think that a link to the FrameNet resources might be very productive. We also think that the limited linguistic material described by Propp, as well as the linguistic information that can be extracted from our fine-grained annotation of the tales, can "feed" the ProppOnto ontologies with some concrete linguistic information to be associated with their classes. This can facilitate or advance the automated semantic annotation of folktales. We are working here on applying the strategies on combining domain ontologies and complex linguistic information described in the MONNET (Multilingual Ontologies for Networked Knowledge) project to the field of folktales. See also [2] or http://cordis.europa.eu/fp7/ict/languagetechnologies/project-monnet_en.html

[14]The comment we made in footnote 13 is valid here too

[15]Propp makes use of the symbol *α: Initial Situation* to refer to everything that happens before the hero's parents announce their departure, but it is not a function as such.

[16]We suspect that the term "narrateme" may not be applicable to them

She ran and ran until she came upon a a stove.
71, 73
"Stove, stove, tell me: where have the geese flown?"
"If you eat my little rye-cake, I'll tell." *76, 78b*
"Oh, we don't even eat cakes made of wheat
in my father's house." E^1 *neg*

where

71 = manner of inclusion into the tale
73 = physical appearance
76 = dialogue with the hero
78b = preparation for the transmission of a magical agent: request
E^1 neg = the hero does not withstand a test (insolent answer)

Table 1: Key to Propp's annotation of *The Magic Swan-Geese*

2.2 PftML

PftML transforms the grammar-like functions, subfunctions and the rules concerning their combination from *Morphology of the Folktale* into a DTD. PftML allows for inline, usually sentence or paragraph-wise XML annotation of fairy tales, as we can see below in the small excerpt of the PftML annotation of *The Magic Swan-Geese* with Proppian functions.

```
<CommandExecution>
<Command subtype="Interdiction">
"Dearest daughter," said the mother, "we are going to
 work. Look after your brother! Don't go out of the yard,
 be a good girl, and we'll buy you a handkerchief."
</Command>
<Execution subtype="Violated">
The father and mother went off to work, and the daughter
soon enough forgot what they had told her. She put her
little brother on the grass under a window and ran into
the yard, where she played and got completely carried
away having fun.
</Execution>
</CommandExecution>
```

The Proppian rules regarding the ways in which functions may be combined are reflected by the DTD design. See, for example, the element `CommandExecution`, which must contain one element of the type `Command` and one `Execution` to make sure that a violation of an interdiction is preceded by the corresponding interdiction. However, this occasionally leads to a lack in flexibility and may bring about unwanted side-effects. Although it is clear from the text that the parents absent themselves from the scene, the tight connection between the interdiction and its violation does not allow in PftML for the function `Absentation subtype="Elders"`, which should have its place between the two, to be marked up.

Also, we have acknowledged before that relying solely on the 31 functions will not allow us to analyze tales to the extent we desire. Seeing that PftML does not go beyond the functions, we will need to find ways to include more information in PftML - or, as the case may be with APftML, to include PftML in an annotation schema affording more detailed markup on various levels.

3. APFTML

Looking at the annotated excerpt from Propp above, we came to two important findings: Firstly, Propp himself clearly did not limit himself to the 31 functions, but used individual "appendix-elements" as he saw fit. Secondly, although only functions will eventually find their way into a folktale's *scheme*, a deeper analysis of the tale will benefit immensely from the more fine-grained analysis (also at the subsentential level) in term of a combination of functions and appendix-elements.

The actual work on APftML[17] is not limited to this extension, but integrates the fairy tale annotation into textual and linguistic annotation standards, like TEI and ISO 37/SC4. For the sake of brevity, we cannot display the full actual annotation here, but give an example of both the TEI and our extension of PftML[18] in the following:

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0"
  xmlns:ht="http://www.w3.org/1999/xhtml">
  <teiHeader>
    ...
    <revisionDesc>
      <change when="2010-06-16">
      Tentative Annotation
      </change>
    </revisionDesc>
  </teiHeader>
  <text>
    <front>
      <docAuthor>
      Alexander Afanasiev</docAuthor>
      <docTitle>
        <titlePart>Die Wilden Schwaene
        </titlePart>
      </docTitle>
    </front>
    <body>
      <p>
        <w xml:id="t1">Es</w>
        <w xml:id="t2">war</w>
        <w xml:id="t3">einmal</w>
        ...
        <?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0"
  xmlns:ht="http://www.w3.org/1999/xhtml">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Die Wilden Schwaene</title>
        <respStmt>
          <resp>collector</resp>
          <persName>Alexander Afanasiev</persName>
        </respStmt>
      </titleStmt>
      <publicationStmt>
```

[17]The schema and an annotation example (The Magic Swan Geese) are available at http://www.coli.uni-saarland.de/~ascheidel/APftML.xsd and http://www.coli.uni-saarland.de/~ascheidel/APftML.xml

[18]To maintain readability, we include redundant information in our example and show the inline equivalent to the future stand-off annotation

```
      <p>http://www.maerchen-sammlung.de/
      Russische%20M%C3%A4rchen_16/
      Wilde-Schwaene_424.html</p>
    </publicationStmt>
    <sourceDesc/>
  </fileDesc>
  <revisionDesc>
    <change when="2010-06-16">Tentative Annotation
    </change>
  </revisionDesc>
</teiHeader>
<text>
  <front>
    <docAuthor>Alexander Afanasiev</docAuthor>
    <docTitle>
      <titlePart>Die Wilden SchwÃd'ne</titlePart>
    </docTitle>
  </front>
  <body>
    <p>
      <w xml:id="t1">Es</w>
      <w xml:id="t2">war</w>
      <w xml:id="t3">einmal</w>
      ...
      <w xml:id="t36">Pass</w>
      <w xml:id="t37">gut</w>
      <w xml:id="t38">auf</w>
      <w xml:id="t39">Dein</w>
      <w xml:id="t40">kleines</w>
      <w xml:id="t41">Bruederchen</w>
      <w xml:id="t42">auf</w>
      <w xml:id="t43">und</w>
      <w xml:id="t44">spielt</w>
      <w xml:id="t45">nur</w>
      <w xml:id="t46">auf</w>
      <w xml:id="t47">dem</w>
      <w xml:id="t48">Hof</w>
      <w xml:id="t49">.</w>
      ...
     <Narrateme>
      <Command subtype="interdiction" id="i0">
        Eines Tages sprach die Mutter: Toechterchen,
        wir gehen jetzt auf die Arbeit.
        Pass gut auf Dein kleines Bruederchen auf
        und spielt nur auf dem Hof. Wir bringen Dir
        auch ein schoenes buntes Tuechlein mit.
      </Command>
      <Agent id="p1">die Mutter</Agent>
      <Patient id="p2">Toechterchen</Patient>
      <Content>Pass gut auf Dein kleines Bruederchen
       auf und spielt nur auf dem Hof.
      </Content>
      <Incentive>Wir bringen Dir auch ein schoenes
       buntes Tuechlein mit.</Incentive>
    </Narrateme>
    <Narrateme>
      <Absentation>Als die Eltern gegangen waren
      </Absentation>
      <Agent id="p0, p1">die Eltern</Agent>
    <Narrateme>
      <CommandExecution subtype="violated"
       commandID="i0">
        setzte das Maedchen das kleine Bruederchen
        ins Gras vor dem Haus und lief auf die Strasse,
        um dort mit den anderen Kindern zu spielen.
      </CommandExecution>
      <Agent id="p2">das Maedchen</Agent>
      <Form> setzte das Maedchen das kleine
       Bruederchen ins Gras vor dem Haus und
       lief auf die Strasse</Form>
      <Motivation>um dort mit den anderen
       Kindern zu spielen</Motivation>
    </Narrateme> ...
```

We plan also to integrate our work within the FrameNet-like approach to the annotation of semantic roles, since we encountered in the appendix of "Morphology of the Folktale" many descriptions that in fact refer to the semantic roles of lexical units, bearing a distinct resemblance to (FrameNet) *frames*. The Proppian function/functional narrateme *Interdiction*, for example has its counterpart in FrameNet, Frame "Deny permission"[19] (see Table 2).

	Proppian "Frame"	FrameNet Frame
Name	Interdiction	Deny_permission
Agent role	person performing	Authority
Patient role	receiver of the interdiction (inferred)	Protagonist
Theme role	contents	Action

Table 2: Comparison of a Proppian "element cluster" and FrameNet Frame in regard to the respective definitions of typical semantic roles.

4. CONCLUSIONS

We described ongoing work in extending and partially redesigning an annotation scheme for fairy tales, which integrates both the full "descriptive" power of Vladimir Propp's work and standards in textual and linguistic annotations, like TEI and ISO TC37/SC4. Examples of this annotation applied to two folk tales will be shown in detail in the poster presentation. As further step in our work, we foresee a multilingual extension, annotating a tale available in different languages (and versions), and an integration of the scheme within more generic semantic resources, like FrameNet and ontologies in the domain of narratives. A test case for the usefulness of our work will lie in the enhanced capability of providing automated comparative studies in the field of folktales.

5. ACKNOWLEDGMENTS

This work has been partially funded by the projects CLARIN & D-SPIN, especially for the linguistic annotation of tales, see http://www.clarin.eu/external/ and http://weblicht.sfs.uni-tuebingen.de/ and by the EU FP7 project MONNET – with grant 248458, especially for the topics related to multilingual ontologies, see http://cordis.europa.eu/fp7/ict/languagetechnologies/project-monnet_en.html

6. REFERENCES

[1] A. Afanas'ev. *Russian fairy tales*. Pantheon Books, New York, 1945.

[2] T. Declerck and P. Lendvai. Towards a standardized linguistic annotation of the textual content of labels in knowledge representation systems. In *LREC 2010- The seventh international conference on Language Resources and Evaluation*. ELRA, 2010.

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

[19] http://framenet.icsi.berkeley.edu/index.php?option=com_wrapper&Itemid=118&frame=Deny_permission

Learning Narrative Morphologies from Annotated Folktales

Mark A. Finlayson
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
32 Vassar St., Cambridge, MA 02139 USA
markaf@mit.edu

ABSTRACT

I describe a research program designed to demonstrate the learning of Proppian morphological functions by computer and test if people are sensitive to the presence of those functions in their cultural narratives. The program has two technical components and three stages. The first component is an annotation tool, the Story Workbench, that allows semi-automatic annotation of natural language text semantics by a lightly-trained annotator; the second component is a pattern-extraction algorithm, Analogical Story Merging. In the first stage, I have annotated 16 of Propp's single-move tales translated into English (21,182 words) for their semantics. In the second stage, in progress, I have performed several proof-of-concept demonstrations of the extraction algorithm, and will soon attempt to extract from the annotated tales actual Proppian morphological functions. I detail three metrics I will use to determine success or failure of this extraction. The final stage, yet to begin, is a recall experiment using at least two cultures to test cultural participant's sensitivity to Proppian functions identified by the technique.

1. INTRODUCTION

The morphological functions introduced by Propp [8] remain a tantalizing window into the cultural information embedded in stories. I describe a research program, the first two-thirds of which is covered by my nearly-completed doctoral dissertation, that is designed to (1) demonstrate the learning, by computer, of Proppian morphological functions from actual folktales, and (2) test via cognitive psychology experiment whether cultural participants are sensitive to Proppian functions identified in the folktales of their culture.

First International AMICUS Workshop, October 21, 2010, Vienna, Austria.

This research program has three stages. First (§2) the semantics of a set of folktales must be represented in a computer-understandable manner. Because natural language processing (NLP) is not yet equal to this task, I have developed a computer application, called the Story Workbench, that allows a lightly-trained annotator to annotate free text for its semantics, while allowing the computer to assist where it can. In this stage, I have annotated 16 of Propp's single-move tales, 21,182 words in English, for 17 different meaning representations. Second (§3), an algorithm is needed to extract the Proppian functions from the annotated folktales. I have developed an algorithm called Analogical Story Merging that has shown promise in extracting Proppian functions. There are a number of possible metrics for measuring the accuracy of the results; I detail three (§3.1). Finally (§4), the validity of the extracted functions must be confirmed by experiments on people. I describe an experimental paradigm in which I will test the sensitivity of cultural participants to Proppian functions automatically extracted from their culture's folktales.

2. SEMANTIC ANNOTATION

To automatically extract Proppian functions from text, we need some way of translating natural language text into computer-understandable representations of meaning. Unfortunately, fully automatic NLP is still far from equal to this task; we must therefore resort to manual (or, at best, semi-automatic) semantic annotation. I have developed an annotation tool called the Story Workbench [4] that facilitates semantic annotation by providing a uniform, extensible, user-friendly platform for semantic annotation. Existing NLP techniques may be integrated into the tool, allowing those techniques to contribute automatically-generated annotations where they are able.

The Story Workbench is a fully functional tool, having been used by 12 different annotators so far to annotate various aspects of the semantics of various texts – for example, we

recently released a corpus of 24,422 words annotated for referring expressions [6]. The Story Workbench currently has 17 implemented representations, the conjunction of which gives fairly reasonable cover of the basic meaning of a narrative. These representations are:

1. Tokens - location of each word token
2. Multi-word Expressions - words that are made up multiple tokens
3. Sentences - location of each sentence
4. Part of Speech Tags - a Penn Treebank tag for each word token and multi-word expression
5. Lemmas - a lemma (i.e., stem, root form) for each word or multi-word expression not already lemmatized
6. Word Senses - a Wordnet sense for each token or multi-word expression
7. Context-Free Grammar Parse - a CFG parse of each sentence
8. Referring Expressions - locations of all expressions that refer to something
9. Referent Attributes - properties (unchanging attributes) of referents referred to in the text
10. Co-reference Relationships - which referring expressions refer to the same referent (co-refer)
11. Time Expressions - location, type, and value of temporal expressions, as defined by TimeML [9]
12. Events - location, features, and type of event mentions, as defined by TimeML
13. Temporal Relationships - event-event, event-time, or time-time temporal relationships, as defined by TimeML
14. Referent Relationships - event-event, event-referent, or referent-referent non-temporal relationships
15. Semantic Roles - predicate features and arguments, as defined in PropBank
16. Mental State - mental state valencies as consequences of actions, as described by Lehnert [7]
17. Proppian Functions - locations of functions as identified by Propp's monograph

Ten trained annotators have annotated 16 of Propp's single move folktales translated into English, a total of 21,182 words. All 17 of the implemented representations have been double-annotated and adjudicated into a gold-standard for each tale. These particular sixteen tales were chosen for the following reasons. First, Propp identified only 46 of the tales he analyzed. Second, I was able to identify extant translations into English for only 31 of Propp's identified tales, even with the help of Russian speakers searching large numbers of translated collections. Third, of those 31 tales, only 16 were single-move. I targeted single move tales because having only one move in a tale simplifies the observed order of Proppian functions; I hypothesized that this would ease learning the functions, and so should form the first attempt. Thus these 16 single-move, English translations of Propp's original tales comprise the initial set to be analyzed.

The first 16 annotations in the list above will form the raw data for the function extraction algorithm. The final representation, Proppian functions, will be used in the second evaluation metric, namely, comparing my extracted functions with Propp's original analysis.

3. LEARNING MORPHOLOGIES

I have developed an algorithm called Analogical Story Merging (ASM) [5] to extract Proppian functions from the annotated folktales. ASM is a variation of the machine learning technique of Bayesian Model Merging [12]. The algorithm begins by constructing an initial model that explicitly encodes each story as one possible output. I do this by first extracting from each the annotation's of each story a sequence of events, shown as D in Figure 1. Each story's event sequence is then incorporated into the initial model, marked as M_0 in the figure, as a single, linear branch of model states. While there are numerous possible orderings, one of the simplest is make the order of states in the model the same as the order in which their associated events occur in the narration of the story (as opposed the order of events in the *story world*).

ASM then searches the space of *state merges*, where two states, each representing an event, are merged into one. To accomplish this, I define both a merge operation over states, and a *prior* probability function to be used when calculating, via Bayes' rule, the posterior probability of the model given the data. The merge operation takes two states and replaces them by a single state, where the merged state inherits the weighted sum of the transitions and emissions of its parents. Because each state in the initial model represents an event in the story, each merged state represents set of all the events of its parents.

The prior is defined such that smaller models are attributed greater probability than larger models, and models that contain merged states representing sets of similar events are given higher probability than otherwise. In ASM the primary calculation of similarity is done via an analogical mapper, an augmented version of the the Structure Mapping Engine [3]. This mapper assesses the similarity between events, taking into account aspects of those event such as their structure (do the number of arguments match?), their classification (is it a *run* or a *love*?), the identities of other events to which the events in question are connected casually or temporally, the consistency of role assignments (is character A in story 1 consistently mapped to character B in story 2?).

The search space for ASM is quite large, being equal in size to Bell's number, B_n, where n is the number of initial states in the model. Bell's number counts the number of unique partitions of a set of n objects [10], and has been shown [2] to be relatively closely bounded above by equation 1.

$$B_n < \left(\frac{0.792n}{ln(n+1)}\right)^n \quad (1)$$

Because the search space is so large, ASM cannot be expected to do an exhaustive search of the state merge space for a set of real stories. Greedy search is required, with efficient pruning of the search space to ensure that the algorithm converges. I have shown that this approach is feasible in two experiments. The first experiment was reported in [5], and was the first proof-of-concept test of the algorithm using summaries of Shakespearian plays. The initial

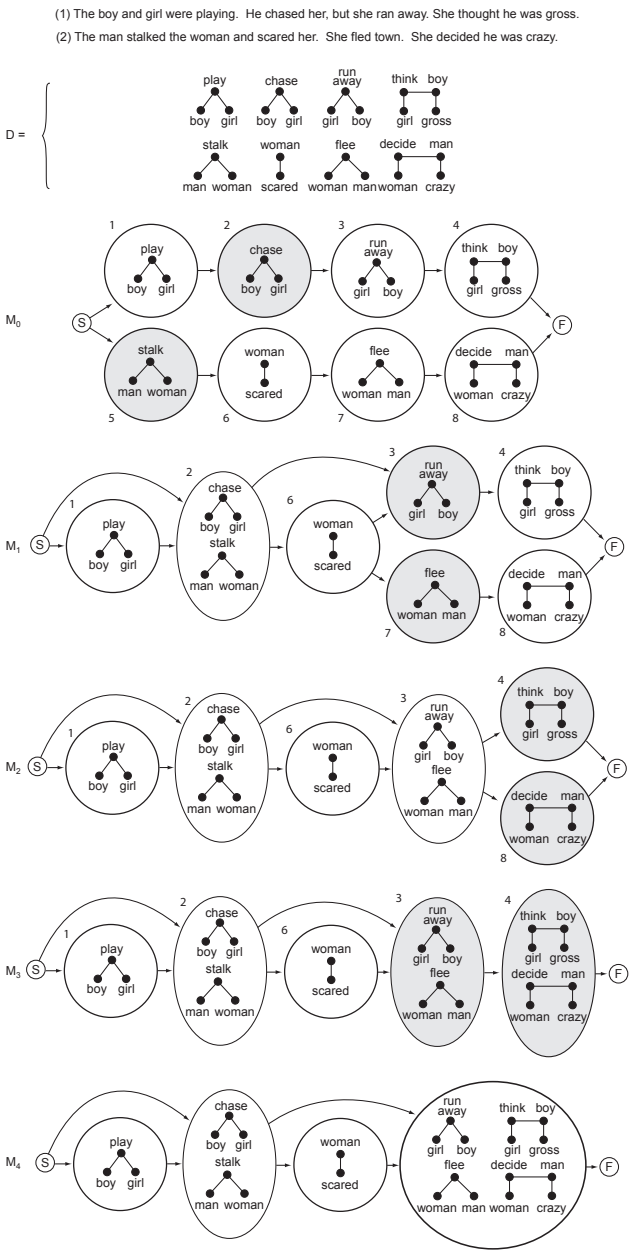

Figure 1: Analogical Story Merging in action. The two stories being merged are written at the top, in (1) and (2). The Story Workbench annotation step produces data structures representing the surface meaning of the story, marked here as D. Each event in each story is then encapsulated in a single state, labeled 1 through 8, in the initial model M_0. ASM searches the space of state merges to find a path to the most probable model, here labeled M_4. From one model to the next, the two states that shaded in the first model are merged together in the second.

model had 48 events across five plays (*Macbeth*, *Hamlet*, *Julius Caesar*, *Othello* and *Taming of the Shrew*) and the search space was pruned by not allowing merges between dissimilar events, but not otherwise optimizing the search. The algorithm converged, and discovered plot similarities that one would expect a human to extract after careful consideration. First, it merged large portions of *Macbeth* and *Hamlet*, the two most similar plays in the set. Second, it merged the ending concluding suicides of *Julius Caesar* and *Othello*, but did not merge these with the (markedly different) suicides of Lady Macbeth and Queen Gertrude. Third, it did not merge the *Taming of the Shrew*, the only comedy in the set, with any of other four tragedies. Numerous other interesting observations may be made, but suffice to say that the algorithm converged on this data and found reasonable patterns.

A second, more recent, experiment has demonstrated that ASM can converge on more complex data. In this experiment, we used four summaries of international conflict situations, written in natural English. These stories were written to illustrate rudimentary plot unit elements (à la Lehnert [7]), in particular, *Revenge* and *Pyrrhic Victory*. After annotation in the Story Workbench, and augmentation of the story graphs with some light commonsense knowledge, each story contained between 34 and 73 states, for a total of 210 states in the initial ASM model. Using a beam search strategy and applying the constraint that all merges in a model must preserve actor mappings across the story, ASM converged and the final graph could be processed to extract the two embedded plot units.

It remains to be seen whether the algorithm, when presented with annotations of real folktales, will be able to extract meaningful functions. Because the extremely large search space induced by 16 folktales of up to 1,800 words each (each folktale potentially containing hundreds of events), I am in the process of augmenting the original ASM implementation to perform efficient, greedy, parallelized beam search, with multiple constraints on valid models, using the 400-node computing cluster available at the MIT Computer Science and Artificial Intelligence Laboratory.

3.1 Evaluation Metrics

I will use at least three metrics to evaluate the output of ASM. The first will be to test the ability of the algorithm to recover patterns purposefully embedded in synthetic data. I will create a synthetic (i.e., artificial) morphology and use it to generate annotations for input into ASM. I will likely start with Propp's own observed morphology over the set of 16 tales that I am analyzing - i.e., including in the morphology only those functions that appear in those 16 tales, and only in those orders. Using this as a skeleton, I will write a generator that outputs, for each Proppian function, a synthetic set of events of the correct semantic character for that function. A set of of synthetic annotations will be generated by this technique, and then fed back into ASM. The functions then discovered by ASM will then be compared with the original synthetic morphology. The measure of success will be an f-measure-like score. The efficiency and reliability of ASM can be evaluated by varying the complexity of the morphology, the number of generated annotations, and the values of the constants in the ASM evaluation functions.

The second metric, perhaps the most interesting, will be to compare with Propp's own analysis the functions that are extracted by ASM when run over the 16 annotated folktales. As we have Propp's original list of functions for these tales, and I will take his analyses as a "gold standard", as it were, to measure the accuracy of the ASM-extracted functions. Beyond the numerical comparison this metric affords, comparing the ASM output with Propp's functions should produce a number of interesting insights. For example, I expect that the annotations I am collecting will not be sufficient to reproduce some of Propp's functions, on account of the wide variation in his level of abstraction. Where ASM breaks down in this case will point to where the abstraction strategy will need to be expanded.

The third metric will be to perform a cross-validation analysis of the set of tales, in which the algorithm is used on different subsets of the 16 tales and the results are compared between the subsets. Such an approach is standard in machine learning studies, and allows testing the sensitivity of the algorithm to variation of input.

4. HUMAN EXPERIMENTS

The true test of this work is whether cultural participants are sensitive to the functions extracted from their own culture's folktales. While there are numerous possible experimental paradigms, in this design we select at least two cultures for study. We will annotate a number of folktales from each culture and extract Proppian functions for each. Using these functions, we will then construct a set of stimuli folktales that are made up primarily of functions from one culture, with the exception of a single function from the other culture. Subjects would then be asked to read these stories and retell them, possibly after a distractor task or delay. Examination of the retold tales should show how subjects treat foreign functions relative to functions from their own culture. If participants preferentially forget or distort foreign functions, we will have fairly clear evidence that people actually detect and extract (and, therefore, probably use) these Proppian functions at some level. There are several possible measures for examining this effect, including reaction time measurements, yes-no judgments of inclusion in the original stimuli (both found in [11]), free-response recall, coded by judges (e.g., [13]), either for a single recall session, or over multiple retellings (such as in a classic study in this area, [1]).

5. ACKNOWLEDGMENTS

This work was supported by the Defense Advanced Research Projects Agency (DARPA) under contract number FA8750-10-1-0076. Thanks to Patrick Winston, Whitman Richards, Ted Gibson, Peter Szolovits, and Josh Tenenbaum for valuable discussion and advice. Thanks to Brett van Zuiden for significant coding and implementation. Thanks to Brett and Nidhi Kulkarni for proof reading.

6. REFERENCES

[1] F. C. Bartlett. Some experiments on the reproduction of folk-stories. *Folklore*, 31(1):30–47, 1920.

[2] D. Berend and T. Tassa. Improved bounds on bell numbers and on moments of sums of random variables. *Probability and Mathematical Statistics*, 30(2), 2010.

[3] B. Falkenhainer, K. D. Forbus, and D. Gentner. The structure-mapping engine. In *Fifth Meeting of the American Association for Artificial Intelligence*, pages 272–277, 1986.

[4] M. A. Finlayson. Collecting semantics in the wild: The story workbench. In J. Beal, P. Bello, N. Cassimatis, M. Coen, and P. H. Winston, editors, *AAAI Fall Symposium on Naturally Inspired Artificial Intelligence*, pages 46–53. AAAI Press.

[5] M. A. Finlayson. *Deriving Narrative Morphologies via Analogical Story Merging*, pages 127–136. New Bulgarian University Press, Sofia, 2009.

[6] R. Hervas and M. A. Finlayson. The prevalence of descriptive referring expressions in news and narrative. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 49–54, 2010.

[7] W. Lehnert. Plot units and narrative summarization. *Cognitive Science*, 4:293–331, 1981.

[8] V. Propp. *Morphology of the Folktale*. Publications of the American Folklore Society, Inc., Bibliographical & Special Series. University of Texas Press, Austin, TX, second edition, 1968.

[9] J. Pustejovsky, J. Castano, R. Ingria, R. Sauri, R. Gaizauskas, A. Setzer, and G. Katz. TimeML: Robust specification of event and temporal expressions in text. In *Proceedings of IWCS-5, the Fifth International Workshop on Computational Semantics*, 2003.

[10] G.-C. Rota. The number of partitions of a set. *The American Mathematical Monthly*, 71(5):498–504, 1964.

[11] C. M. Seifert, R. P. Abelson, G. McKoon, and R. Ratcliff. Memory connections between thematically similar episodes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12(2):220–231, 1986.

[12] A. Stolcke and S. Omohundro. *Inducing probabilistic grammars by Bayesian model merging*, volume 862 of *Lecture Notes in Computer Science*, pages 106–118. Springer, Berlin, 1994.

[13] P. van den Broek, E. P. Lorch, and R. Thurlow. Children's and adults' memory for television stories: The role of causal factors, story-grammar categories, and hierarchical level. *Child Development*, 67(6):3010–3028, 1996.

Event Interpretation: A Step Towards Event-Centred Text Mining

Raheel Nawaz
School of Computer Science
University of Manchester
131 Princess Street, Manchester
M1 7DN, UK
+44 (0)161 306 3091
nawazr@cs.man.ac.uk

Paul Thompson
National Centre for Text Mining
University of Manchester
131 Princess Street, Manchester
M1 7DN, UK
+44 (0)161 306 3091
paul.thompson@manchester.ac.uk

Sophia Ananiadou
National Centre for Text Mining
University of Manchester
131 Princess Street, Manchester
M1 7DN, UK
+44 (0)161 306 3092
sophia.ananiadou@manchester.ac.uk

ABSTRACT

Event-centred text mining facilitates semantic querying of document content, providing greater descriptive power and more focused results than traditional keyword searches. In the biomedical domain, automatic assignment of high-level interpretative information to events, e.g., general information content and level of certainty, is useful for a number of tasks. In this paper we motivate the need for correct interpretation of events and describe a new approach for tackling the problem in the biomedical domain.

1. INTRODUCTION

Event-based text mining approaches constitute a promising alternative to the traditional approaches, mainly based on the bag-of-words principle. Events are template-like, structured representations of pieces of knowledge contained within documents. Our work focuses specifically on bio-events, which are dynamic relations within the biomedical domain. Text mining systems that are able to extract such events automatically can allow much more precise and focussed searches than the traditional keyword-based systems. Event-based searches specify one or more constraints on the events to be retrieved, which are not dependent on the precise wording in the text. These constraints could be in terms of the type of the event (e.g., positive regulation) and/or its participants (e.g., the instigator of the event must be a protein).

Although event-based searching can retrieve many more relevant documents than is possible using traditional keyword searches, they typically do not take into account the *interpretation* of the event. For example, a particular event may represent generally accepted knowledge, experimental observations, hypotheses or analyses of experimental results. For the two latter types of event, the author may express varying degrees of certainty regarding the analysis performed. We term these types of interpretative information *meta-knowledge*.

First International AMICUS Workshop, October 21, 2010, Vienna, Austria.

Without access to meta-knowledge, a large number of extracted bio-events will be treated identically by text mining systems, even though their intended interpretations may vary significantly [5, 9]. This would pose a serious problem for users of the system whose information requirements include the ability to distinguish between certain interpretations. For example, a biologist who wishes to update either an incomplete model of a biological process (e.g., a molecular pathway) [6] or a curated biological database [1] would wish to locate only newly-reported, reliable experimental knowledge. Thus, he would be interested only in experimental observations or confident analyses of results, but not in hypotheses or more tentative analyses.

The work reported here describes a novel annotation scheme that can be applied to bio-events to make explicit the meta-knowledge associated with them. The annotation caters for several different types (or *dimensions*) of meta-knowledge that could be specified about an event. The aim of the annotation is to facilitate the training of text mining systems that can extract automatically not only events and their participants but also meta-knowledge associated with the event.

2. EVENT-CENTRED TEXT MINING

The knowledge expressed by events is normally organised around a particular word (the event trigger), which is typically a verb or noun. Each event has one or more participants which describe different aspects of the event, e.g., what causes the event, what is affected by it, where it took place, etc. Participants can correspond to entities, concepts or other events, and are often labelled with semantic roles such as CAUSE, THEME or LOCATION to aid in their interpretation and to facilitate more precise searching

Typically, bio-events themselves, as well as bio-entities that constitute the event participants, are assigned types/classes from an appropriate taxonomy or ontology (e.g., [1]). Figure 1 illustrates a simple sentence, together with a typical template-style representation of the bio-events contained within it.

Queries for relevant events can be carried out through partial completion of a template that specifies constraints regarding the events to be retrieved, in terms of one or more of the following:

- ontological classes of events e.g. *POSITIVE_REGULATION*.

- specifications of the participants that should be present in the event (in terms of semantic roles).
- restrictions on the values of particular participants, in terms of either actual entities (e.g. *NF-kappa B*) or ontological classes (e.g. *PROTEIN*).

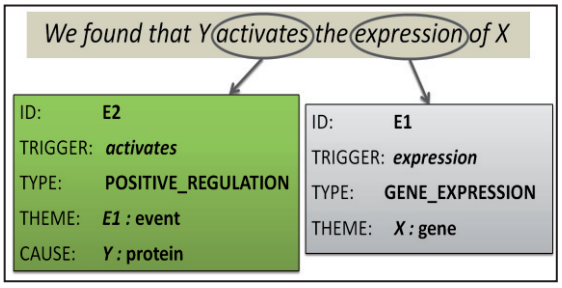

Figure 1. Bio-Event Representation

Searches over events can be more or less specific, depending on the number and nature of the constraints specified.

Event extraction systems are typically trained on collections of texts (corpora) in which events and their participants have been manually annotated by domain experts. Examples include the GENIA Event Corpus [3] and GREC [7]. These corpora allow text mining systems to be trained to recognise and extract events from biomedical texts.

3. INTERPRETATION OF BIO-EVENTS

Existing event annotated corpora within the biomedical domain contain few, if any, annotations that relate to their interpretation. Although more extensive interpretation-focussed annotation has been carried out within the domain at either the sentence level (e.g., [8]) or sentence-fragment level (e.g., [10]), these annotations cannot be used straightforwardly to assign interpretations to bio-events. Often, a sentence will contain several bio-events (e.g. both an experimental method *and* the results of applying this method), each of which has a different interpretation. If an expression of speculation is present (e.g. the word *might*), this may affect only certain events in a sentence.

Our work aims to address this situation through the development of a multi-dimensional annotation scheme that is especially tailored to bio-events. The scheme is intended to be general enough to allow integration with various existing bio-event annotation schemes, whilst being detailed enough to capture important subtleties in the nature of the meta-knowledge expressed about the event.

4. META-KNOWLEDGE ANNOTATION OF BIO-EVENTS

The annotation scheme presented here is a slightly modified version of our original meta-knowledge annotation scheme [5]. Different types of meta-knowledge are encoded through five distinct dimensions (Figure 2), each of which consists of a set of complete and mutually-exclusive categories, i.e., any given bio-event belongs to exactly one category in each dimension. Our chosen set of annotation dimensions has been motivated by the major information needs of biologists, as discussed earlier. The advantage of using multiple dimensions is that the interplay between the assigned values in each dimension can reveal both subtle and substantial differences in the types of meta-knowledge expressed (see section 4.6).

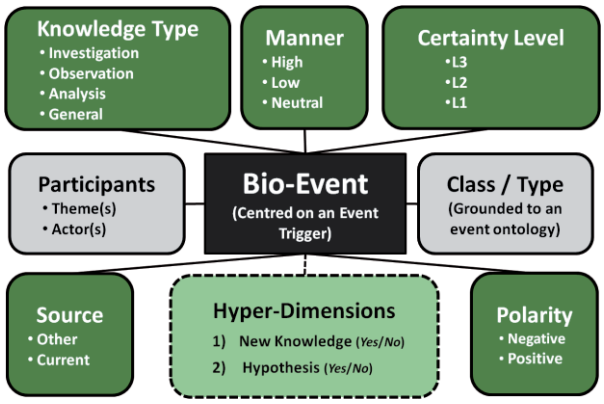

Figure 2. Bio-Event Annotation

Meta-knowledge can be expressed in text in a number of different ways. In the majority of cases, this is through the presence of particular "clue" words or phrases, although other features can also come into play, such as the tense of the verb on which the event is centred, or the relative position of the event within the text.

The annotation task consists of assigning an appropriate value for each dimension, as well as marking the textual evidence for this assignment. In order to minimise the annotation burden, the number of possible categories within each dimension has been kept as small as possible, whilst still respecting important distinctions in meta-knowledge that have been observed during our corpus study. The five meta-knowledge dimensions and their values are described in more detail below.

4.1 Knowledge Type (KT)

This dimension captures the general information content of the event. Each event is classified into one of the following four categories:

Investigation: Enquiries or investigations, which have either already been conducted or are planned for the future, typically marked by lexical clues like *examined*, *investigated* and *studied*, etc.

Observation: Direct observations, often represented by lexical clues like *found* and *observed*, etc. Simple past tense sentences typically also describe observations.

Analysis: Inferences, interpretations, speculations or other types of cognitive analysis, typically expressed by lexical clues like *suggest*, *indicate*, *therefore* and *conclude* etc.

General: Scientific facts, processes, states or methodology. This is the default category for the Knowledge Type dimension.

4.2 Certainty Level (CL)

In scientific text, this dimension is normally only applicable to events whose *KT* corresponds either to *Analysis* or *General*. In the case of *Analysis* events, *CL* encodes confidence in the truth of the event, whilst for *General* events, there is a temporal aspect,

to account for cases where a particular process is explicitly stated to occur most (but not all) of the time, using a marker such as *normally*, or only occasionally, using a marker like *sometimes*. We distinguish three levels of certainty:

L3: No expression of uncertainty or speculation (default category)

L2: High confidence or slight speculation (*Analysis*), event occurs most (but not all) of the time (*General*). Typical lexical markers include *likely* and *probably*. Certain *Analysis* markers also invoke this certainty level, such as *suggest* and *indicate*

L1: Low confidence or considerable speculation (*Analysis*), event occurs infrequently (*General)*; typical lexical markers include *may*, *might* and *perhaps*.

4.3 Source

The source of experimental evidence provides important information for biologists. It can also help in distinguishing new experimental knowledge from previously reported knowledge. Our scheme distinguishes two categories, namely:

Other: The event is attributed to a previous study. In this case, explicit clues (citations or phrases like *previous studies* etc.) are normally present.

Current: The event makes an assertion that can be (explicitly or implicitly) attributed to the current study. This is the default category, and is assigned in the absence of explicit lexical or contextual clues.

4.4 Polarity

This dimension identifies negated events. Although certain bio-event corpora are annotated with this information, it is still missing from others. The indication of whether an event is negated is vital, as the interpretation of a negated event instance is completely opposite to the interpretation of a non-negated (positive) instance of the same event.

We define negation as the absence or non-existence of an entity or a process. Negation is typically expressed by the adverbial *not* and the nominal *no*. However, other lexical devices like negative affixals (*un-* and *in-,* etc.), restrictive verbs (*fail*, *lack*, and *unable*, etc.), restrictive nouns (*exception*, etc.), certain adjectives (*independent*, etc.), and certain adverbs (*without*, etc.) can also be used.

4.5 Manner

This dimension corresponds to indications of the rate, level, strength or intensity of the event described. This can be significant in the correct interpretation of an event. Our scheme distinguishes 3 categories of *Manner,* namely:

High: Typically expressed by adverbs and adjectives like *strongly*, *rapidly* and *high*, etc.

Low: Typically expressed by adverbs and adjectives like *weakly*, *slightly* and *slow*, etc.

Neutral: Default category assigned to all events without an explicit indication of manner.

4.6 Hyper-dimensions

A defining feature of our annotation scheme is that additional information can be inferred by considering combinations of some of the explicitly annotated dimensions. We refer to this additional information as hyper-dimensions of our scheme. At present, we have identified two such hyper-dimensions, as described below.

4.6.1 New Knowledge

A combination of the values of *Source, KT* and *CL* dimensions can be used to isolate those events representing new knowledge. For example, events with the *KT* value of *Observation* may correspond to new knowledge, but only if they represent observations from the current study (i.e., *Source=Current*), rather than observations cited from elsewhere. In a similar way, an *Analysis* drawn from experimental results in the current study could be treated as new knowledge, but generally only if it represents a straightforward interpretation of results (i.e. *CL=L3*), rather than something more speculative.

4.6.2 Hypothesis

Events that represent hypotheses can be isolated by considering their values of *KT* and *CL*. Events with a *KT* value of *Investigation* can always be assumed to be a hypothesis. However, if the *KT* value is *Analysis,* then only those events with a *CL* value of L1 or L2 (speculative inferences made on the basis of results) should be considered as hypothesis, to be matched with more definite experimental evidence when available. A value of L3 in this instance would normally be classed as new knowledge, as explained in the previous section.

5. EVALUATION

An initial evaluation of the annotation scheme has been performed through the annotation of 70 abstracts randomly chosen from the GENIA Pathway Corpus, containing a total of 2,603 annotated bio-events. Two of the authors independently annotated these bio-events with meta-knowledge using a comprehensive set of annotation guidelines developed following a detailed analysis of the various bio-event corpora and the output of an initial case study [5]. The remainder of this section discusses the results of this evaluation experiment in more detail.

Dimension	Cohen's Kappa
KT	0.9017
CL	0.9329
Polarity	0.9059
Manner	0.8944
Source	0.9520

Table 1. Inter-Annotator Agreement

5.1 Inter-Annotator Agreement

The quality of annotation was assessed using Cohen's kappa [2] to calculate inter-annotator agreement. Table 1 shows the agreement figures for each annotation dimension. The highest value of agreement was achieved for the *Source* dimension, whilst the *KT* dimension yielded the lowest agreement value.

Nevertheless, the kappa scores for all annotation dimensions were in the *good* region [4].

5.2 Category Distribution

Knowledge Type: The most prevalent category found in this dimension was *Observation* (45% of events). Only a small fraction of these (4%) was represented by an explicit lexical clue (mostly sensory verbs). In most cases the tense, local context (position within the sentence) or global context (position within the document) were found to be important factors. The second most common category (37% of events) was *General*, of which the majority (64%) were processes or states embedded in noun phrases (such as *c-fos expression*). More than a fifth of the *General* events (22%) expressed known scientific facts, whilst 14% expressed experimental/scientific methods (such as *stimulation* and *incubation* etc.). Explicit lexical clues were found only for facts, but even then in only 1% of cases. *Analysis* was the third most common category of annotated events (16%). Of these events, 44% were deductions (*CL=L3*), whilst the remaining 56% were hedged interpretations (*CL=L1/L2*). All *Analysis* events were marked with explicit lexical clues. The least common category was *Investigation*, comprising 1.5% of all events, all of which were marked with explicit lexical clues.

Certainty Level: *L3* was found to be the most prevalent category, corresponding to 93% of all events. The categories *L2* and *L1* occurred with frequencies of 4.3% and 2.5%, respectively. The relative scarcity of speculative sentences in scientific literature is a well documented phenomenon. Vincze et al. [8] found that less than 18% of sentences occurring in biomedical abstracts are speculative. Similarly, we found that around 20% of corpus events belong to speculative sentences. Since speculative sentences contain non-speculative events as well, the frequency of speculative events is expected to be much less than the frequency of speculative sentences. In accordance with this hypothesis, we found that only 7% of corpus events were expressed with some degree of speculation. We also found that almost all speculated events had explicit lexical clues.

Polarity: Our event-centric view of negation showed just above 3% of the events to be negated. Similarly to speculation, the expected frequency of negated events is lower than the frequency of negated sentences. Another reason for finding fewer negated events is the fact that, in contrast to previous schemes, we draw a distinction between events that are negated and events expressed with *Low* manner. For example, certain words like *limited* and *barely* are often considered as negation clues. However, we consider them as clues for *Low* manner. In all cases, negation was expressed through explicit lexical clues.

Manner: Whilst only a small fraction (4%) of events contains an indication of *Manner*, we found that where present, manner conveys vital information about the event. Our results also revealed that indications of *High* manner are three times more frequent than the indications of *Low* manner. We also noted that both *High* and *Low* manners were always indicated through the use of explicit clues.

Source: Most (99%) of the events were found to be of the *Current* category. This is to be expected, as authors tend to focus on current work in within abstracts. It is envisaged, however, that this dimension will be more useful for analyzing full papers.

Hyper-dimensions: Almost 57% of the events represent *New Knowledge*, and just above 8% represent *Hypotheses*.

6. CONCLUSION AND FUTURE WORK

The recent advent of event-centred text mining approaches mandates the need for correct and consistent interpretation of textual events. We have presented a new approach to address this problem in the domain of biomedical research literature. The cornerstone of our approach is a meta-knowledge annotation scheme that captures the key information required for the correct interpretation of bio-events [5]. An initial evaluation experiment has illustrated high inter-annotator agreement and a sufficient number of annotations along each category in every dimension. The highly favourable results of this experiment have confirmed the feasibility and soundness of the annotation scheme, and have paved the way for a large scale annotation effort involving multiple independent (i.e. non-author) annotators.

We are currently in the process of creating a large corpus of meta-knowledge enriched bio-events. This corpus will consist of three sub-corpora, which have previously been annotated with different types of bio-events, namely GENIA, GREC and a small corpus of full papers.

7. ACKNOWLEDGMENTS

The work described in this paper has been funded by the Biotechnology and Biological Sciences Research Council through grant numbers BBS/B/13640, BB/F006039/1 (ONDEX).

8. REFERENCES

[1] Ashburner, M., Ball, C. A., Blake, J. A. , Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics* 25:25-29.

[2] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 20, 37–46.

[3] Kim, J., T. Ohta and Tsujii, J. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics* 9:10

[4] Krippendorff, K. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills.

[5] Nawaz, R., Thompson, P., McNaught, J. and Ananiadou, S. (2010). Meta-Knowledge Annotation of Bio-Events. In *Proceedings of LREC 2010*, 2498-2507.

[6] Oda, K., Kim, J., Ohta, T., Okanohara, D., Matsuzaki, T., Tateisi, Y. and Tsujii, J. 2008. New challenges for text mining: mapping between text and manually curated pathways. BMC Bioinformatics 9(Suppl 3): S5.

[7] Thompson, P., Iqbal, S., McNaught, J. and Ananiadou, S. 2009. Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics* 10: 349

[8] Vincze, V., Szarvas, G., Farkas, R., Mora, G. and Csirik, J. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics 9(Suppl 11): S9.

[9] Waard, A. de., Shum, B., Carusi, A., Park, J., Samwald M. and Sándor, Á. 2009. Hypotheses, Evidence and Relationships: The HypER Approach for Representing Scientific Knowledge Claims. In *Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse*. Available at: http://oro.open.ac.uk/18563/

[10] Wilbur, W.J., Rzhetsky, A. and Shatkay, H. 2006. New directions in biomedical text annotations: definitions, guidelines and corpus construction. BMC Bioinformatics 7: 356.

Motives and Characters in Folklore Indices and Russian Folktales

Anna Rafaeva
Moscow State University, Moscow, Russia
anna_raf@rambler.ru

ABSTRACT

In this paper, some experiments on automated search for motives and folktale characters will be discussed. First experiments were carried out in 1997-2002, and simple tools for automatic search for some motives and characters in the text of famous Aarne-Thompson index [1] were developed. The second experiment (2002) was an attempt to automatically find one motif in number of folklore indices in Russian. Since 2009 the program SKAZKA-2 (FOLKTALE-2) is developing by author. SKAZKA-2 now contains a corpus of Russian folktales and their descriptions, dictionaries and some tools for automatic and automated proceeding of folktales. Some preliminary results of the work will also be discussed.

1.SKAZKA (FOLKTALE) DATABASE

1.1Description of Program Tool

SKAZKA (FOLKTALE) program tool was developed by author to deal with text of Aarne-Thompson index [1]. For example, the tasks were to find tale types, where *stepdaughter* is heroine, or to count all the tale types, which are present in Russian folklore. The database contains a small part of the index (some types of tales of magic only), so it is working model of possible system.

SKAZKA was made in Starling, powerful linguistic database software, developed by S. A. Starostin (see http://starling.rinet.ru/program.php?lan=en for details). Starling has a lot of useful possibilities, among them are:

- Build-in dictionaries and morphological processor (Russian and English).
- Capability of creating and processing powerful databases specifically suited for linguistic purposes.
- Capability to deal with text fields of variable length.
- Build-in program interpreter for simple user-defined programs.

1.2Archetypical Motives

There is a variety of definitions for motif now. The aim of the first experiment was to search for *archetypical motives (AM)* in the text of [1]. The term *archetypical motif (or plot archetypes)* was suggested by E.M. Meletinsky in [9], examples are *rescue from monster* and *unpromising hero.* Each AM has variants, for example, *unpromising hero* (or heroine) may be youngest son, orphan or stepdaughter. Hero also may be (or become by some reasons) dirty, covered with ash, speechless and so on.

According to Meletinsky, archetypical motives can be found in different folklore and literature genres, from myth to novel. Archetypical motives are changing and developing from myth to folklore and literature, for example, mythical “totemic wife” transforms into “animal bride” (cat, mouse, frog and so on, see type 402 in [1]) in tales of magic. Plots of some myths can be treated as realization of only one AM (for example, *creation*), narrative structure of folktales is more complicated, and *creation* (in the form of *getting or stealing magic objects*) is only one motif among others.

Full description of archetypical motives is impossible or very difficult. We can, however, make partial description of AM system in some genres. Narrative structure of magic folktales can be described by Propp’s formula [12], so we supposed, that archetypical motives could also be described and formalized. So, the formal system of AM in tales of magic was developed to describe narrative structure of folktales.

In this system, each archetypical motif can be treated as predicate with a fixed number of terms (characters, action, condition and object). Every term can be word or expression of natural language (for example, *prince* and *bear* (characters), *fight* (action) in index, *Prince Ivan* (role: hero) in folktale) or predicate itself. So, folktales can be described at abstract level or more concrete when terms are defined or not. The system is described in [13].

So, on the most abstract level archetypical motives are similar to Propp’s functions [12].

1.3Experiment

The first task was to find tale types with given archetypical motives in [1]. The sets of motives, provided by S. Thompson, couldn’t be used directly, but motif definitions were very useful and were treated as text, such as descriptions of tale types. So, search rules were derived for some archetypical motives:

- First, keywords were extracted manually for each archetypical motif or some concrete forms, if possible. Sometimes, the Russian-English and English-Russian dictionaries were used to find possible synonymous for given word (for example, *herd, shepherd* and *rabbit-herd* in [1]).
- For each keyword, its concordances were made automatically.
- After that, the rules themselves were defined.
- The last step was testing and writing additional rules.

First International AMICUS Workshop, October 21, 2010, Vienna, Austria.

1.4Parents in Tales of Magic

According to E. Novik [11], relate terms and parents' relations are very important in Russian magic folktales. Besides, automatic search for characters and even their roles (in Propp's terms) can be easier in index and folklore texts, than that of motives or functions. So, the aim of experiment was to extract information about relate terms and roles (in Propp's terms) of relatives from given tale types (text fields of types 300–681 in database). First, programs to make frequency dictionary (for all text fields in database) and concordances for given word or expression was written. The dictionary was used to extract keywords manually, for example *grandmother, twin, daughter, stepchildren, niece, bride, father-in-law, godson, blood-brothers, relatives* and so on. After that, the table of concordances for all keywords was made automatically; it was used for further analysis. The experiment let me draw some interesting conclusions. For example, the roles of *blood-brothers* are similar to the roles of *twins*: they often act together as hero and helper, while elder brothers of hero are false heroes. More results are described in [14].

2.PROCEEDING FOLKLORE INDICES AS TEXT

The aim of next experiment was to compare indices and to find automatically single motif in descriptions of plots and characters. The motif "yearning after lost bridegroom or husband" (that causes appearance of monster or demon) was chosen, because it was present in tales and legends, so the indices of different genres [3–7] could be compared. All indices are in Russian; they describe East Slavic folktales and Russian and Lithuanian legends. Keywords for search were extracted manually with the use of frequency dictionary for the texts of all indices, which was made automatically. The list of keywords contained nouns, verbs and expressions. The most important results of the work seem to be the following:

- The language of indices is simple enough, so a single list of keywords can be used for automatic search in different indices. Thus, some keywords are special for only one index. The full description of keyword extraction and more detailed conclusions are given in [15].
- No one of indices [3–7] describes motives, so this information is not given directly. Automatic proceeding lets us extract this information from plot descriptions.
- For scholar, the structure of index is very important, while automatic proceeding needs large and detail descriptions. For example, descriptions in [4] are short, so it is difficult to use in for automated motif identification without any additional data. This data is provided by SKAZKA-2 software.

3.SKAZKA-2 SOFTWARE

SKAZKA-2 (FOLKTALE-2) is developing for automatic and automated analysis of Russian folktales. Now it contains a corpus of about 1000 tales, their descriptions, frequency dictionary, some tools to deal with text, for example, to make frequency dictionaries and concordances for given word or expression and the program to construct semantic networks in automated way. The index of East Slavic tale types [4], additional dictionaries and morphological processor should be added later.

Tales are stored in plain test format (txt); the resulting dictionaries and tables of concordances are in text format too, but can be imported into e-table or even some database format for further work. All programs are written in C++ language.

The existing corpus of folktales and simple program tools, however, can be used to carry out a lot of experiments and to conclude, what program tools have to be developed and what data (dictionaries, tale descriptions, indices or semantic nets) is to be added.

3.1Simple Tasks

First task, where SKAZKA-2 was used, was study of numbers in Russian folktales from Afanasjev's textbook [2]. SKAZKA-2 was used to select keywords, find all numbers in corpus and then make the table of concordances. The resulting table was imported into MS Excel for further analysis. Some results of this work were very interesting; for example, 13 is happy or neutral number in Russian tales of magic, as it is often treated as 12 + 1, where the last object is the best one (for example, 12 swan maidens and the last is hero's bride).

An attempt to find motives and Propp's functions in the text of folktales failed, because extracting keywords for search manually was too difficult. For example, there is a strong correlation between motif "Hero of supernatural origin/strength" and appearance of childless couple at the beginning of the tale; but there exist many ways to describe childless couple in text. This result can be explained easily, if we remember, that archetypical motives and Propp's functions are very abstract: if we take word or expression as first level of abstraction, functions and archetypical motives will be third level or even higher. So, it is necessary to find keywords in some automated way (for example with any statistical method) or other elements for automatic search have to be chosen.

3.2Characters and Semantic Network

Study of folktale characters and their formal definition is interesting by itself, and an index of characters can be used for a number of purposes, for example, for automated motif definition. Now an index of characters in the form of semantic network is under constructions. It connects characters, as they are given in folktale, and their roles in fairy tale (the modified system of roles described in [10] is used). To fix all characters in Russian tales of magic it is necessary to add new roles, additional to Propp's ones. For example, *prince* can be *hero, helper* (rarely) or *false hero*, but he can also be *watcher* or *informer*; *cat* can be *helper*, *object of difficult task* or one of *hero's forms*, if hero is magically transformed into cat. Cat can also be one of *witch's helpers* or *servants*. The list of roles is open now and is developing together with network.

To build semantic network in automated way ProSeCa (PROgram for SEmantic Classification) is used. The program is developing to make semantic dictionary of Russian language in the form of directed graph. In this tool, semantic descriptions for every word is performed like chain in digraph, where the beginning vertex is text example (optionally) or word of natural language, and the end vertex is one of user-defined semantic "primitives". ProSeCa is used to store dictionary and to build chains in automated way (see [8]). In next version of ProSeCa not only vertices, but edges too will allow marking to show different relations between characters in Russian folktales.

In SKAZKA-2, semantic network connects text examples, characters, as they appear in text (words or expressions of natural language), abstract descriptions (roles and others) and other characters too. Here some short chains for *bear* are shown. Text examples and their descriptions are omitted.

Mishka (name) – medved' (bear) – wild animal – animal

medved' (bear) – monster (location: forest) – monster

medved' (bear) – animal helper – helper

medvediha (she-bear) – bear – enchantment: transformation into bear

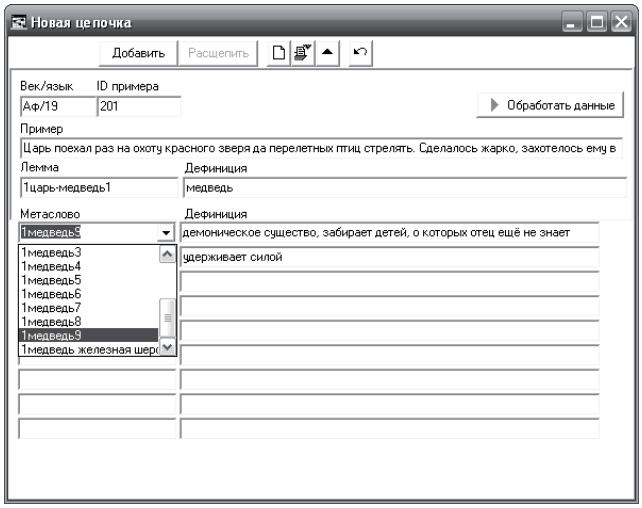

Figure 1. Selecting definition for *tsar-medved'* (bear king)

If given word or semantic description is present in the network, Proseca lets user to select definition for it and then builds the rest of chain automatically (see Figure 1).

Resulting chains can be viewed in text form, where vertices (characters and roles), their definitions in dictionary and examples are placed in different windows (see Figure 2).

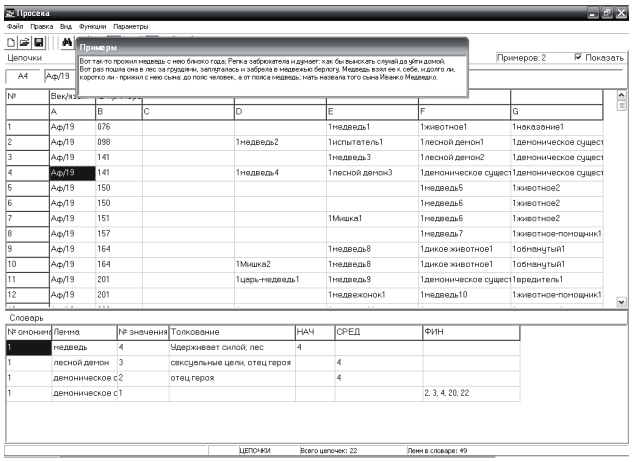

Figure 2. Chains for *bear* with examples and dictionary

Only few characters – *medved'(bear), kot (cat), baba-jaga (witch), tsarevich (prince)* and *lisa (fox)* – are partly described at this time. But even now the network is useful for studies in Russian folktales. Later it will be used not only for automated proceeding in SKAZKA-2, but to make an index of characters in Russian tales of magic too.

4.REFERENCES

[1] Aarne, A. and Thompson, S. 1973. *The Types of the Folktale:* A Classification and Bibliography. Anti Aarne's *Verzeichnis der Märchentypen* (FF Communications 3).Translated and Enlarged by Stith Thompson (FF Communications 184). Academia Scientiarum Fennica, Helsinki.

[2] Afanasjev, A. 1984-1985. Narodnyje russkie skaski A.N. Afanasjeva (Russian folk tales from A.N. Afanasjev). Nauka, Moscow. See also: http://feb-web.ru/feb/skazki/default.asp

[3] Ajvazjan, S. and Jakimova O. 1975. Ukazatel' sujetov russkih bylichek i byval'chin o mifologicheskih personagah (Plot-Index of Legends about Mythical Characters). In: *Pomerantseva E. Mifologicheskije personaji v russkom folklore (Mythical Characters in Russian Folklore).* Moscow. See also: http://ruthenia.ru/folklore/ayvazan1.htm

[4] Barag, L., Berezovsky, I., Kabashnikov, K and Novikov, N. 1979. *Sravnitel\nyj ukfzftel' luzhetov. Vostochoslavjanskfja skazka (Comparative Index of Plots. East Slavic Folktale).* Nauka, Leningrad. See also: http://ruthenia.ru/folklore/sus/index.htm

[5] Gordeeva, N. 1978-1984. Ukazatel' sujetov bylichek i byvalshin Omskoj oblasti (Index of Plots of Legends for Omskij Region). In *Folklore and Postfolklore* (Moscow). DOI= http://ruthenia.ru/folklore/gordeeva1.htm

[6] Kerbelyte, B. 2001. Tipy narodnyh skazanij: Structurno-semanticheskaja klassifikatsija litovskih etiologicheskih, mifologicheskih skazanij i predanij. The Structural-semantic Classification of Lithuanian Aetiological, Mythological and Historical Legends. Sankt-Peterburg. See also: http://ruthenia.ru/folklore/kerbelite1.htm

[7] Kozlova, N. 2000. Vostochnoslavjanskije bylichki o zmee i zmejah. Mificheskij l'ubovnik. Ukazatel' sujetov i teksty. (East Slavic Legends about serpent. Mythical Lover. Index of Plots and Textes. Omsk. See also: http://ruthenia.ru/folklore/kozlova1.htm

[8] Kretov, A. and Rafaeva, A. 2009. Programma semanticheskoj klassifikacii leksiki – ProSeKa: teoreticheskije i prikladnyje aspecty (On semantic classification programm ProSeCa: theoretical and practical aspects). In: Kompjuternaja lingvistica i intellektualnyje tehnologii: Po materialam ejegodnoj Mezhdunarodnoj konferentsii "Dialog'2009". Russian State University for Humanities, Moscow.

[9] Meletinsky, E. 1994 *O literaturnyh arhetipah (About Literary Archetypes)*. Russian State University for Humanities, Moscow.

[10] Meletinsky, E., Nekludov, S., Novik, E. and Segal, D. 2001. Problemy structurnogo opisanija volshebnoj skazki (Problems of structural description of fairy tale). In: *Structura volshebnoj skazki (The Structure of Fairy Tale).* Russian State University for Humanities, Moscow.

[11] Novik, E. 2001. Sistema personagej russkoj volshebnoj skazki (The System of Personages in Russian Fairy Tale). In: *Structura volshebnoj skazki (Structure of Fairy Tale)*. Russian State University for Humanities, Moscow.

[12] Propp, V. 1969. *Morfologija skazki (The Morphology of Folktale)*. Nauka, Moscow.

[13] Rafaeva, A. 1998. *Issledovanie semanticheskih struktur traditsionnyh sujetov i motivov (Research for Semantic Structures of Traditional Plots and Motives)*. Dissertation. Russian State University for Humanities, Moscow.

[14] Rafaeva, A. 2004. Analiz rodstvennykh otnoshenij s pomosh'ju sistemy SKAZKA (Analysis of Relate Terms with SKAZKA Program). In: *Problemy kompjuternoj lingvistiki: Sbornik nauchnykh trudov*. Voronezh State University, Voronezh. See also: http://www.ruthenia.ru/folklore/rafaeva4.htm

[15] Rafaeva, A. 2005. Nekotoryje vozmozhosti kompjuternogo analisa folklornyh ukazatelej (Some Capabilities of Computer-Based Analysis of Folklore Indices). In: Problemy kompjuternoj lingvistiki. Voronezh State University, Voronezh. See alse: http://ruthenia.ru/folklore/rafaeva5.htm

Semantic Processing of a Hungarian Ethnographic Corpus

Miklós Szőts
Applied Logic Laboratory
Hankóczy u. 7.
Budapest, Hungary
szots@all.hu

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

Zoltán Alexin
University of Szeged
Department of Software Engineering
Árpád tér 2., Szeged, Hungary
alexin@inf.u-szeged.hu

Veronika Vincze
University of Szeged
Institute of Informatics
Árpád tér 2., Szeged, Hungary
vinczev@inf.u-szeged.hu

Attila Almási
University of Szeged
Institute of Informatics
Árpád tér 2., Szeged, Hungary
vizipal@gmail.com

ABSTRACT

In this poster, a Hungarian ethnographic database containing linguistic annotation is presented. The corpus contains texts from three domains, namely, folk beliefs, táltos texts and tales. All the possible morphosyntactic analyses assigned to each word and the appropriate one selected from them (based on contextual information) are also marked. Syntactic (dependency) annotation is added semi-automatically to the corpus texts at a second phase of the processing. With the help of these enriched linguistic attributes, the texts can be semantically analyzed and clustered. The research and development team is working on a semantic search tool enabling to browse the texts on the basis of their semantic meaning. The proposed technology may result in a new approach to the ethnographic research and may open a new type of access to the databases.

1. INTRODUCTION

The Applied Logic Laboratory and the University of Szeged (Institute of Informatics and Department of Library and Information Science) are developing a technology to perform meaning-based search for natural language texts. Researchers wish to go beyond the world of simple, *type-in-your-keywords* search engines; and to develop a technology and an integrated search engine which performs genuine content-oriented search in natural language documents (textual data), by adapting and combining existing statistical and symbolic techniques in a novel way in order to exploit the user's semantic competence to a considerably greater extent than traditional search engines do. The kernel of the technology developed during the project is language-independent, while the prototype is going to be developed for patent specifications in English and Hungarian and Hungarian ethnographic texts.

Parallel to the above, a linguistic annotation method and software components are developed. By applying these computational methods and with considerable human efforts a manually checked and corrected ethnographic database is produced. Its importance is twofold: on the one hand, textual databases from the folklore domain have not been (or have hardly been) available in a digitalized version (see e.g. [5] on the difficulties of creating folklore databases). On the other hand, Hungarian folklore texts – to our best knowledge – have not been analyzed with computational linguistic tools.

The processing of the folklore texts follows the traditions of the Szeged Treebank [1]. The format of the ethnographic corpus is TEI XML[1]. This database could serve as the input of further semantic processing, like clustering or thematic classification of texts.

In the following, the ethnographic corpus will be presented, some statistical data on the corpus will be shown, the process of the linguistic annotation will be discussed, and the importance of semantic search in ethnographic texts will be emphasized.

2. TOPICS WITHIN THE CORPUS

The corpus contains texts from three domains: folk beliefs (2704 short texts), táltos texts (432 texts) and folk tales (1185 texts). Beliefs and táltos texts were collected at the beginning of the 20th century and data came from all over the areas of historic Hungary. The original manuscripts can be found in the Museum of Ethnography and they have been published in a volume as well [9]. Tales were collected from the Hungarian Electronic Library (`http://www.mek.hu`) and from various Internet sources, like `http://www.nepmese.hu`.

[1]See `http://www.tei-c.org`.

First International AMICUS Workshop, October 21, 2010, Vienna, Austria.

2.1 Folk beliefs

The topics of folk beliefs involve almost every aspect of ordinary life: main stages of human life (birth, christening, getting a husband, illness, death, otherworld), weather, special days, animals. The short – typically one-sentence-long – beliefs are accompanied with some explanation or short stories. In the collection, simple descriptions and incantations in the form of poems can also been found. Certain beliefs occur in different versions. Some folk traditions are also preserved in the beliefs.

> Ha a menyasszony cipőjét ellopják a lakodalom éjjelén s lekaparva a talpáról a földet felteszik a füstre – ez a házas társak nyugodt életét megrontja.
>
> 'If the bride's shoes get stolen at the night of the wedding, and the soil scratched from their sole is smoked – this will ruin the calm life of the spouses.'

2.2 Táltos texts

The táltos texts tell us about wizards, medicine men and women, táltoses (a mythical figure similar to a shaman, whose task is to heal the body and soul of his/her people) and their abilities and earmarks. The collection comprise short stories in several dialects of Hungarian.

2.3 Tales

The tale collection includes masterpieces of Elek Benedek, who was a famous tale writer. The collection also contains various translated tales from different countries as well as Hungarian folk tales. As for their topic, tales depict stories about kings, queens, princesses, soldiers, rich and poor men, dragons, giants and dwarfs, fairies, ghosts, heaven and hell and many more. Table 1 presents a statistics on the corpus.

Table 1: Statistical data on the corpus

Text type	Number of texts	Number of words
Folk beliefs	2704	65807
Táltos texts	432	44021
Folk tales	1185	1311952
Total	4321	1421780

3. LINGUISTIC ANNOTATION

After the digitalization the texts were segmented into words. Then each word was computationally assigned one or more morphosyntactic codes. The obtained vocabulary was later manually checked and corrected by linguists. Morphosyntactic analysis was based on the categories used in The Concise Dictionary of the Hungarian Language [8]. There were some special words or wordforms that seemed to be problematic from the viewpoint of morphological analysis:

- In the case of regional words, the sense and their parts-of-speech had to be determined. (E.g.: *goroboncás* 'wandering magician', *slájer* 'veil of the bride'.)
- Words and wordforms with substandard orthography: they were paired with their standard forms and the MSD code of those could usually be copied from the Szeged Treebank. If not, they were provided by our linguists. (E.g.: *ígízís* – standard: *igézés* 'spelling', *abbú* – standard: *abból* 'from that'.)
- Sometimes the word with substandard orthography coincided with another existing (standard) word. They needed special attention during the disambiguation for they already had an MSD code (according to the standard orthography, however, in these texts it was typically the substandard form that occurred. Thus, the standard version of those had to be given together with their MSD code(s). (E.g.: *mellül* – standard 'breast-ESSIVE' – substandard: *mellől* 'from it', *aggyá* – standard: 'brain-TRANSLATIVE' – substandard: *adjál* 'give-IMP'.)

```
<s>Ahun gyünnek!
 <choice>
  <orig><w>Ahun</w></orig>
  <reg><w>Ahol
   <ana>
    <msd><lemma>ahol</lemma>
    <mscat>[Pd]</mscat></msd>
   </ana>
   <anav>
    <msd><lemma>ahol</lemma>
    <mscat>[Pr]</mscat></msd>
   </anav>
   <anav>
    <msd><lemma>ahol</lemma>
    <mscat>[Pd]</mscat></msd>
   </anav>
  </w></reg>
 </choice>
 <choice>
  <orig><w>gyünnek</w></orig>
  <reg><w>jönnek
   <ana>
    <msd><lemma>jön</lemma>
    <mscat>[Vmip3p—n]</mscat></msd>
   </ana>
   <anav>
    <msd><lemma>jön</lemma>
    <mscat>[Vmip3p—n]</mscat></msd>
   </anav>
  </w></reg>
  </choice>
 <c>!</c>
</s>
```

Figure 1: Sample XML containing words with substandard orthography

The corpus currently contains morphosyntactic annotation based on the MSD coding system [2]. All the possible morphosyntactic analyses are assigned to each word, and the appropriate one will be selected from them based on contextual information. In Figure 1 a sample XML fragment is shown. The fragment presents how the ethnographic corpus handles substandard morphology taking into account the newest TEI encoding guideline (TEI P5).

Applying <choice> </choice> tags we can retain the original orthography of words and add the regular typing as well. The regular typing is marked up by the <reg> </reg> tags. A morphosyntactic analyses are provided between <anav> </anav> tags. The disambiguated part-of-speech is provided between <ana> </ana> tags.

Syntactic annotation is planned to be carried out by the Malt parser [3]. J. Nivre developed a statistical method for learning dependency parsing. His software needs a learning database in a given format and from this training set it can build a probabilistic model for parsing. An other program can execute the learned model on unknown texts. The conversion of the Szeged Treebank to dependency structure is approaching to the end, which will be used as a training database for the parser. In the fall of 2010 a trained dependency parser for Hungarian will be ready for application.

4. SEMANTIC SEARCH

Generally we use only surface level lexical semantics; however, in certain specific questions deep semantics can be used too – e.g. in the domain of technical texts the quantities should be represented precisely. Note, that the use of deep semantics is generally domain dependent.

The difference between lexical and deep semantics is illustrated by the following example. Let us consider the sentences *The prince got a ring from the magician.* and *The magician gave a ring to the prince.* Clearly, these describe the same situation. In our system both will be represented by a structure like the following one:

event: giving$_1$
- initiator: magician$_1$
- theme: ring$_1$
- recipient: prince$_1$

Here we describe only the connections of the event denoted by *give* (or *get*) with its dependents. In a deep semantic approach we would need to describe the meaning of these words too – e.g. to give the necessary condition and the result of the event.

Verbs *give* and *get* are considered synonyms since they describe the same frame, thus, their case frame is the same. In our planned semantic lexicon the surface and the deep case frames would be represented together; moreover, if there are some constraints for the possible arguments, they are added to this lexicon too.

The representation of one of the case frames of the verb *get* is a structure like the following one:

case_frame$_1$:
- literal: get, got, gotten
- category: verb
- SUBJ: recipient
  agentive
- ACC: theme
- from: initiator
  agentive

It can be seen that the deep case frame is defined by the thematic roles (see e.g. [4]). In fact, they are basically the same; however, there is a difference between the use of thematic roles in linguistic semantics and in our case. We do not intend to develop or borrow some general system of thematic roles for the whole natural language, but to work out systems of role relations for a specific language usage. For example, if the topic is the production of something, the roles initiator, source, result, ingredient, instrument and goal are used.

The most important step of the semantic search is comparing the representation of the query with fragments of a text. The tokens ("meanings") of the representation of the query are compared with the words in the text, and it is tested whether the syntactic relations between the words match the corresponding semantic relations.

The question is when a word in a query matches a word in the text. Our answer is that the word in the text has to have equal or more information content than the word in the query. The most important cases are as follows:

- synonyms; however, our interpretation of synonymy is more permissive that the generally accepted one – we consider two words (with their case frames) as synonyms if they refer to the same thing or eventuality.
- if A is a kind of B, then B in a query matches A – e.g.: *magic spear* in the text is a relevant hit for *magic weapon*.
- if A (or being A) is a necessary condition of B, then B in a query matches A – e.g.: *find* in the text is relevant to the query *look for*.

Our semantic lexicon is based on sets of synonyms – synsets in WordNet terminology; however, the criterion of synonymy is given by our own definition. The synsets are connected by the relation "having more information content". A top ontology is coupled to the system of synsets. It consists of a general top ontology (the same one for every domain) and a top ontology of the domain. The complexities of connecting synsets to the top ontologies are different in different domains. In the case of production it is not complicated since the role relations are defined in the top ontology of the domain. However, in the domain of fairy tales the ontology of Propp's formalism is the domain ontology [6], and to connect the synsets to it will not be simple.

5. GOALS FOR SEMANTIC PROCESSING

The goals for semantic processing can be summarized as follows:

- Working out the linguistic and methodological foundations of model-based semantic search.
- Elaborating visually based methods for the user interface.

From the point of view of the user, the main novel features of our search engine will be as follows:

- The query is not a Boolean combination of terms (keywords), but well-formed sentences of a controlled natural language, like *The princess sleeps for years.*
- The results given by the search engine to be developed contain phrases which may mean the same thing as the query. The relevance of the results depends on the measure of fit between the meaning of the query and some phrases in the document. Going on with our example: if texts like *The princess sleeps for 100 years.* can be found in a document, it will be considered a result of high relevance; however, a tale about a princess who could not sleep because of a pea in her bed will not be given as result.
- The graphical representation of the relevant pieces of text which the engine finds adequate to the query will be shown to the user, so (s)he can easily decide whether (s)he is interested in the document in question.

6. SUMMARY

From the presented processing of folklore texts the semantic search could be the most innovative technology under development. However, other tools like clustering and classification tools can also provide help in information retrieval. Semantic search can be used in different text mining solutions (e.g. information extraction). Researchers have further plans in processing folklore archives, like the semantic tagging of fairy tales, using Propp's morphology [7].

7. ACKNOWLEDGMENTS

This work was supported in part by the Ányos Jedlik Program of the National Office for Research and Technology of the Hungarian government within the framework of the R & D project MASZEKER (Modell Alapú Szemantikus Kereső Rendszer – Model Based Semantic Search System) .

8. REFERENCES

[1] D. Csendes, J. Csirik, T. Gyimóthy, and A. Kocsor. The Szeged Treebank. In *Proceedings of the Eighth International Conference on Text, Speech and Dialogue (TSD 2005)*, pages 123–131, September 2005.

[2] T. Erjavec. *MULTEXT-East morphosyntactic specifications. Version 3.* 2004.

[3] J. Nivre. *Inductive Dependency Parsing.* Springer Netherlands, 2006.

[4] T. Parsons. *Events in the Semantics of English: A Study in Subatomic Semantics.* MIT Press, Cambridge, MA, 1990.

[5] I. Pávai. A néprajzi adatbázis-építés akadályai. *Néprajzi Hírek*, 1(4):86–89, 1996.

[6] F. Peinado, P. Gervas, and B. Diaz-Agudo. *A Description Logic Ontology for Fairy Tale Generation.* http://www.fdi.ucm.es/profesor/fpeinado/publications/2004-peinado-description.pdf, 2004.

[7] V. Propp. *Morphology of the Folktale.* University of Texas Press; 2 edition (June 3, 2010), US, 2010.

[8] F. Pusztai, editor. *Értelmező Kéziszótár.* Akadémiai Kiadó, Budapest, 2006.

[9] K. Verebélyi, editor. *Néphit szövegek.* Magyar Néprajzi Társaság, Budapest, 1998.