Landing Propp in Interaction Space: First Steps Toward Scalable Open Domain Narrative Analysis With Predication-based Semantic Indexing

Scott Malec[1], Sándor Darányi[3], Trevor Cohen[1], and Dominic Widdows[2]

[1] University of Texas, School of Biomedical Informatics at Houston, Texas, USA
[2] Microsoft Bing, Redmond, USA
[3] University of Borås, Sweden

Abstract. In this paper, we explore the possibility of applying high-dimensional vector representations of concept-relation-concept triplets, which have been successfully applied to model a small set of relationship types in the biomedical domain, to the task of modeling folk tales. In doing so, our ultimate aim is to develop representations of narratives through which their underlying structure can be compared. The current paper describes our progress toward this aim, with emphasis on addressing the technical challenges involved in moving from the relatively constrained set of relations that have been extracted from biomedical text to the much larger set of unnormalized relations that have been extracted from the open domain. A toy example using graded vectors demonstrates that our approach will be feasible once more material will be added to the test collection.

Keywords: Distributional Semantics, Vector Symbolic Architectures, Predication-based Semantic Indexing, Folktales

1 Introduction

It is well known that the content of typical documents in knowledge domains can be described by respective sublanguages [1], with examples ranging from biomolecular research [2] over lipoprotein kinetics, clinical patient reports, telegraphic Navy messages, and reporting of events in outer space[3] to social science [4].

Below we depart from the assumption that the same holds for tales as a genre in narrative studies, and will use Vladimir Propp's seminal ideas derived from the Afanas'ev corpus of Russian fairy tales [5]. He stated that any genuine plot thereof is composed from a canonical vocabulary of 31 functions defining action types of the actors in the tale, so that such functions follow each other in sequential order, with omissions permitted. The resulting strings of predefined content elements remind one of DNA and invite the metaphor of narrative genomics [6]. Therefore it is a natural next step to establish a test collection based on Propp and Afanas'ev for explorative research in this area. To this end we have created a workflow that combines natural language processing tools with Predication-based Semantic Indexing (PSI) [7] for the comparison of narratives as function sequences, because PSI has proven successful in biomedical research [8].

As our ultimate goal is to introduce scalable narrative analysis for the safeguarding of cultural heritage by means of quantum-inspired methods, a note is in place here to explain what we mean by interaction space. To wit, a major trend demonstrates that biological and language-based research methodologies overlap to some extent, e.g. PageRank is also usable in a biological context [9, 10]. In these methods, similarity between comparables is routinely modelled on force for visualization [11], i.e. is considered as a kind of quasi-physical interaction. This means that any high-dimensional comparison of objects or features goes back to the same metaphor, including the visual maps to such spaces created by low-dimensional projections, e.g. genetic complementation maps [12], connectomes [13], and interaction maps [14] in biology, or heatmaps displaying semantic content distribution in text collections [15, 16]. Clearly, a generalized model is emerging which takes dependencies among observables, including their semantics, as a property driving interactions between objects and their features – a view that perceives any sort of classification as fundamentally interaction-based.

In what follows we shall briefly discuss the ingredients for a model that demonstrates our thinking.

2 Background

2.1 Sublanguages in a nutshell

To consider folktales as formulaic expressions of a sublanguage, recall that, as Zellig Harris had postulated, all occurrences of a language are word sequences satisfying certain constraints which express and transmit information [1]. His constraints were dependency relations, paraphrastic reductions, and inequalities of likelihood. As he was to find out, certain subsets of languages within specialized domains, called sublanguages, do exist, and they exhibit specialized constraints due to limitations of the words and relations of the subject matter. In the grammar of such 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. These are called *formulae*.

Sublanguage formulae are similar to the ones of logic, but with certain extensions. Because 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.

Such formulae are well-formed expressions and correspond to the ”events” of a domain. Based on [17], 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 and the sublanguage sentences they represent are e.g.:

- G J B “antigen was injected into the foot-pads of rabbit”
- A V C “antibody is found in lymphocyte”
- G U T“antigen arrives by the lymph stream”

As we will see, Proppian function sequences as the backbone of fairy tales display a similar structure.

2.2 Propp’s formulaic method

V.J. Propp’s theory that the canonical form of Russian fairy tales is a compulsory sequence of actions called *functions* and selected from a list of 31 typical activities performed by typical actors was based on a limited sample of cca 50 fairy tales from the Afanas’ev collection, itself comprising cca 600 stories, selected and compiled in the 19th century [5]. Whereas the in-principle applicability of the scheme, with or without modifications, has been extensively debated ever since, researchers have started to look at the reproduction of Propp’s conclusions only lately [19]. Our insight is that his scheme lends itself to semantic markup [20, 21], with subject-verb-object triples underlying Proppian functions suitable for predicate encoding as shown below.

Predication example The following are typical examples of predication from the biomedical domain are the following:

Concept 1	Relation	Concept 2
isoniazid	TREATS	tuberculosis
cell culture	DIAGNOSES	tuberculosis
lung	LOCATION OF	tuberculosis

In comparison, predicates based on Russian fairy tales would look like these:

Concept 1	Relation	Concept 2
Baba Yaga	IS A	donor
Golden apple	IS A	gift
Baba Yaga	LIVES IN	hut on chicken legs
Donor	GIVES	gift (direct object)
Donor	GIVES TO	protagonist (indirect object)

In the resultant corpus, we would know that Baba Yaga has given a magic apple to Ivan Simpleton and therefore he is a protagonist,[4] standing for a donor function. Alternatively, if we know that a magic apple is received by Ivan Simpleton and that it later helps him to overcome obstacles later in the tale, we would characterize Baba Yaga as a donor and Ivan Simpleton as the protagonist.

[4] This is a special case where we need an indirect object, Ivan Simpleton.

Transformation of a Sample Story (Afanasev 96: "Morozko/Jack Frost") Let us look at an example of a donor function from a sample story sample below:

The poor little thing remained there shivering and softly repeating her prayers. Jack Frost came leaping and jumping and casting glances at the lovely maiden. "Maiden, maiden, I am Jack Frost the Ruby-nosed!" he said. "Welcome, Jack Frost! God must have sent you to save my sinful soul." Jack Frost was about to crack her body and freeze her to death, but he was touched by her wise words, pitied her, and tossed her a fur coat.

Here the "maiden", a stepdaughter, is ordered by her stepmother to be left out to the elements in the forest, a plot element representative of Cinderella-like tales (ATU Type 510A) that are called "mat' padcheritsa" or "Zolushka" tales in the Russian folklore tradition. Jack Frost is clearly the donor. "Tossed" would be mapped to the GIVES predicate.

Next we look at PSI as a way to map the above statements as basic constituent units into predication space.

2.3 Predication-based Semantic Indexing

Predication-based Semantic Indexing (PSI) provides the means to efficiently search across tens of millions of concept-relation-concept triplets [7, 22], known as *semantic predications*, extracted from the biomedical literature using a Natural Language Processing (NLP) system called SemRep [23]. SemRep uses the UMLS [24] and MetaMap[25] to map relevant expressions from free text to concepts in a controlled vocabulary, and extracts relationships between these concepts using underspecified syntactic parsing, a set of indicator rules, and constraints present in the UMLS semantic network. PSI derives high-dimensional vector representations of concepts from the predications they occur in, effectively circumventing the combinatorial explosion of possible pathways between concepts by converting the task of traversing individual predications into the task of measuring the similarity between composite concept vectors. Consequently, search time for single, double or triple predicate paths is identical once the relevant concept vectors have been constructed [26]. It can also detect double and triple predicate pathways connecting example pairs of therapeutically related drugs and diseases; and use these inferred pathways to guide search for treatments for other diseases [8]. Further, PSI has been used to mediate semantic search by utilizing high-dimensional vector representations to infer the nature of the relationship between query concepts and other concepts in relevant documents. Inference is accomplished in high-dimensional space using Expansion-by-Analogy, a novel analogical approach to pseudo-relevance feedback, in which the relationships between query concepts and other concepts in documents they occur in guide the query expansion process. The semantic vector based approaches developed show improvements in performance over a baseline bag-of-concepts model, and these are most pronounced on queries that are not conducive to keyword-based search [27]. Therefore, there is reason to believe that the same approach can be used to create predication-based semantic representations of folk narratives.

An appealing feature of predications extracted from the biomedical literature by SemRep is that both concepts and predicates are normalized. Concepts are normalized to discrete ones represented in the UMLS. These are large in number, but the number of predicates is much smaller - SemRep extracts around forty predicate types corresponding to relations in the UMLS semantic network. This constraint made the biomedical domain an ideal domain in which to explore PSI's capabilities as a mediator of analogical reasoning. However, concept-relationship-concept triplets extracted from general domain text by information extraction systems such as ReVerb [28] have a far broader range of relationship types. Such systems leverage syntactic constraints to isolate concept-relationship-concept triplets, without the need for normalization to leverage semantic knowledge resources. But as the relationship types (or predicates) are not normalized, the capacity for analogical and other reasoning is limited.

For example, when applied to the widely-used Touchstone Applied Sciences (TASA) corpus of electronic text, variant forms of the verb "ASK" are extracted as predicates, including CALL_TO_ASK_ABOUT, HAVE_TO_ASK, ASK_LOT_OF, TURN_TO_ASK, GO_ASK and GO_TO_ASK. In order to mediate reasoning with these triples, a system would need to map between these variant expressions. This is particularly important for the current endeavor on account of the importance of predicates in Proppian theory. According to Propp, "predicates give the composition of tales; all subjects, objects, and other parts of the sentence define the theme" [5] (page 113). So a means to map between related predicates is required if we are to recognize structural similarities between tales. To do so, we have developed an iterative variant of PSI that draws on distributional information from the sentences from which these triples were extracted.

When applied to normalized triplets extracted using SemRep, PSI predicates are represented by randomly generated *elemental vectors* ($E(\mathrm{predicate})$). On account of the high-dimensional nature of PSI vector representations, these predicate vectors are guaranteed to be mutually orthogonal, or close-to-orthogonal, with high probability. PSI concept vector representations are generated by superposing vector products generated using reversible vector transformations provided by a set of representational approaches known as Vector Symbolic Architectures [29] (VSAs). These approaches provide the means to generate composite vectors using an operator known as binding ($\otimes$). Various VSA implementations with different binding operators have been described in the literature. Regardless of the specific implementation, binding operators combine two vectors A and B to form a composite vector $C = A \otimes B$, such that C is dissimilar from its component vectors A and B. This transformation can be reversed, such that, for example, $A \otimes B \oslash A \approx B$.

Binding provides the means to train PSI semantic vectors ($S(\mathrm{concept})$) by combining elemental predicate vectors with randomly generated elemental vector representations of concepts ($S(\mathrm{concept})$) . For example, for the predication $\mathrm{isoniazid}$ TREATS $\mathrm{tuberculosis}$, training occurs as follows:

$$S(\mathrm{isoniazid}) \mathrel{+}= E(\mathrm{TREATS}) \otimes E(\mathrm{tuberculosis})$$
$$S(\mathrm{tuberculosis}) \mathrel{+}= E(\mathrm{TREATS\text{-}INV}) \otimes E(\mathrm{isoniazid})$$

In order to represent predicates extracted by ReVerb, we first utilize Random Indexing (RI) [30] to generate representations of terms occurring in a corpus of interest, in our case the TASA corpus. RI term vectors are generated by superposing randomly generated elemental document vectors for the documents a term occurs in, such that terms occurring in similar documents will have similar vector representations after training.

For each predicate, a semantic predicate vector ($S(\text{predicate})$) is generated by superposing the RI term vectors for the terms in the source sentences from which predications involving this predicate were extracted. The iterative superposition of pre-trained RI vectors in this fashion is known as Reflective Random Indexing [31]. PSI is then performed in the usual way, with the exception that semantic predicate vectors are substituted for elemental predicate vectors. In addition, for each predicate a permutation operation is used to generate a vector representation with opposite directionality (-INV). The same permutation is applied across all predicates, such that if $S(\mathrm{P}_1) \approx S(\mathrm{P}_2)$, $S(\mathrm{P}_1 - \mathrm{INV}) \approx S(\mathrm{P}_2 - \mathrm{INV})$.

This substitution enables PSI training and queries to be performed without the need to normalize the broad range of predicate types extracted by ReVerb, as while binding to vector A to each of a pair of almost-orthogonal vectors B and C will result in dissimilar bound products $A \otimes B$ and $A \otimes C$, given a vector $B\prime \approx B$ we would anticipate $A \otimes B$ $\approx A \otimes B\prime$.

3 Initial experiments

3.1 Generating semantic predicate vectors

The overall workflow utilized to generate PSI representations for the Afanas'ev corpus is shown in Figure 1. First (1), semantic term vectors are generated by applying Random Indexing to the TASA corpus (S(term)). Then (2), the semantic term vectors for terms in the sentences from which each ReVerb predicate (from the TASA corpus) was extracted are superposed to generate semantic vector representations of each predicate (S(predicate)). Finally (3), PSI is appled to the predications extracted from the TASA corpus using these pre-trained semantic predicate vectors instead of random predicate vectors. Though not shown in the figure, statistical weighting metrics are applied at various points in the process. Specifically the log-entropy weighting metric [32] is used for the Reflective Random Indexing step, and Inverse Document Frequency (IDF) weighting is applied during the PSI step, such that superposition operations for S(concept) are weighted in accordance with the IDF of other concept in each predication involved in training. All operations were conducted using 4096-dimensional binary vectors using the implementation described in [33], with the Binary Spatter Code [34] as the Vector Symbolic Architecture.

3.2 Searching with semantic predicate vectors

Tables 3.2 and 3.2 show some preliminary results of the sorts of search enabled by this approach. In Table 3.2, we demonstrate queries on the TASA predicates and their respective results and scores. The reader should note that a keyword (such as "AWARD") need not to appear in the predication result itself. It may be the case that these semantic links have been inferred through the semantic relationships as derived training by RRI.

In Table 3.2, we include some results for queries of the following type on the Afanas'ev predication corpus, where we have integrated the RRI trained predicates:

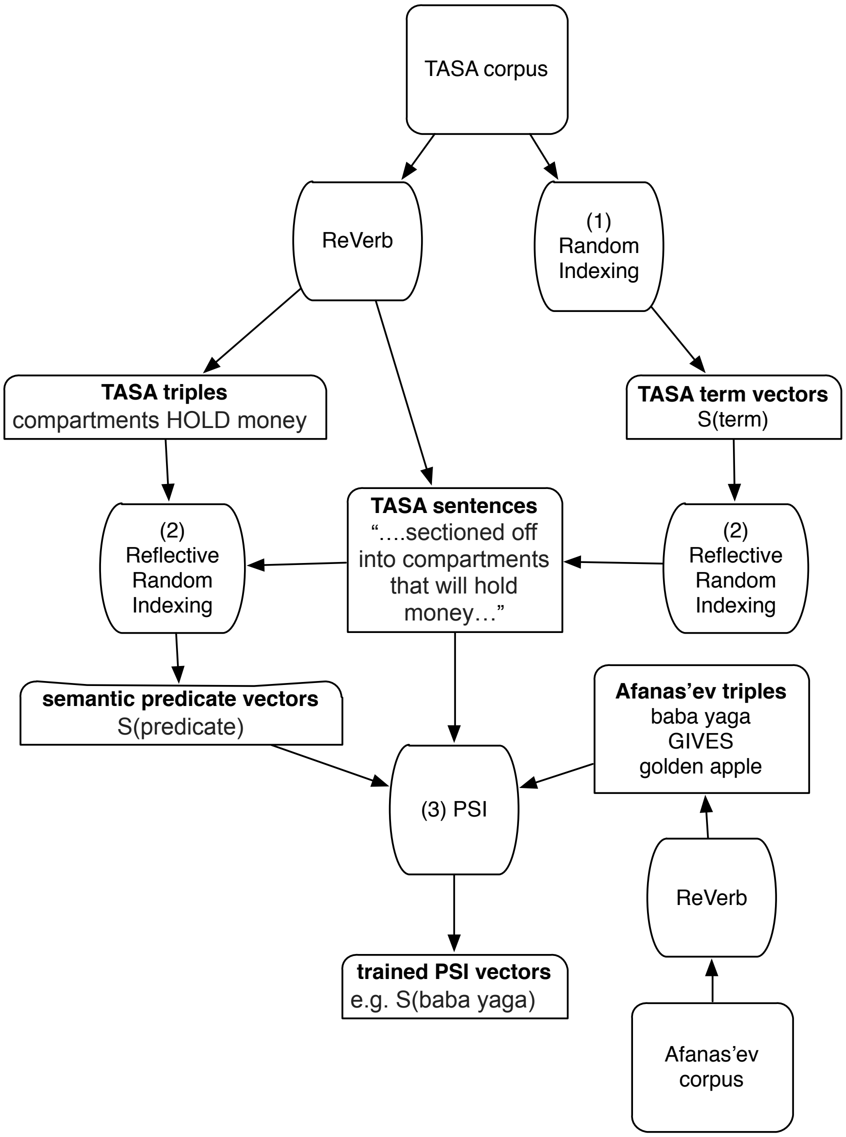

Fig. 1. Generation of PSI vectors.

.

AWARD		PICKET		JUDGE	
NAME-RUNNER-UP_FOR	0.51	RECOGNIZE_RIGHT_OF	0.46	BE_JUDGE_OF	0.52
GIVE_AWARD_TO	0.40	START_LEAD	0.25	DISCUSS_CASE_WITH	0.52
BE_ONESHARE_OF	0.33	SEE_LINE_OF	0.23	THROW_BOOK_AT	0.40

Table 1. Nearest neighbor searches for pre-trained predicate vectors. Scores are $1 - \frac{2}{n}\mathrm{HammingDistance}(\mathrm{x}, \mathrm{y})$ between the binary vectors concerned.

$$\text{Find closest semantic vectors to } E(\text{concept}) \otimes P(\text{PREDICATE})$$

The key finding is that the results obtained by binding to the trained predicate vectors for different, but related, predicates return similar results, demonstrating the feasibility of our proposed approach. With respect to interpretation in relation to the tales from which the underlying predications were drawn, The 'cottage' result appears spurious. However, the little_brother result is interesting from a Proppian perspective. This 'little_brother' is derived from the Magic Swan-Geese where the little_brother in question has been rescued by his big sister after being kidnapped by Swan-Geese. Though this mechanism is in its early stages of development, it shows promise in capturing some of the vagueness and slipperiness of natural language. The semantic clusters that result from RRI trained predicates allow for us to capture notions and general classes of actions that underlie the apparent multiformity that Propp discussed.

PURSUE-INV		RUN_AFTER-INV		CHASE-INV	
bad_reputation	0.44	bad_reputation	0.42	bad_reputation	0.47
cottage	0.31	cottage	0.34	cottage	0.41
little_brother	0.28	direction	0.29	direction	0.32
direction	0.27	little_brother	0.29	little_brother	0.32
wings	0.11	wings	0.15	ivashko	0.13
ivashko	0.10	ivashko	0.10	wings	0.10

Table 2. Nearst neighboring semantic vectors representing ReVerb concepts from the Proppian corpus to the search $E(\text{swan\_geese}) \otimes P(\text{PURSUE-INV})$, $E(\text{swan\_geese}) \otimes P(\text{RUN\_AFTER-INV})$ and $E(\text{swan\_geese}) \otimes P(\text{CHASE-INV})$.

3.3 Using Position within a Story Narrative

For a next step, already paving the way for future experiments with the Propp battery of fairy tales, this section presents a brief and initial exploration of ordering effects in semantic modelling of narratives. Though this has for a long time been important in fields such as dynamic semantics and discourse representation theory [35], it has often been absent from computational work on knowledge bases and knowledge representation.

That is, in ontologies modeled by triple stores, the predications are typically assumed to be static, or true for all times.

This static approach is obviously insufficient for modelling narratives. For example, many folk tales have motifs whereby a promise is made and then fulfiled or broken. Of course, the making of the promise must come before for the breaking or fulfilment of the promise to make sense.

A first step towards modeling these temporal orderings can be made using graded vector representation (see [36], under review separately for this conference). Such a graded model work by binding the vector for each indexed predication with another vector representing its position in the narrative. The position vector is produced by interpolating between demarcator vectors α and ω representing the start and end of a linear scale. Thanks to the sparse similarity properties of vectors in high dimensions (see [37, § 3]), these demarcator vectors can be selected at random with a near-certain guarantee of no interference.

As a test example, consider the following minimal summaries of Shakespeare stories as a generally known shorthand example:

```
<document>
  <setting>Othello is a hero.</setting>
  <marriage>Othello and Desdemona get married.</marriage>
  <confusion>Othello wrongly suspects and kills Desdemona.</confusion>
</document>

<document>
  <setting>Everyone goes to the forest.</setting>
  <confusion>They are tricked into falling in love
             with the wrong people.</confusion>
  <marriage>Hermia marries Lysander. Helena marries Demetrius.</marriage>
</document>

<document>
  <setting>Viola is shipwrecked but useful.</setting>
  <confusion>She disguises herself and woos Olivia
             on behalf of Orsino.</confusion>
  <marriage>Viola marries Orsino. Olivia marries Sebastian.</marriage>
</document>
```

In the absence of ordering information, these stories all have the functional elements of setting, confusion, and marriage. For the purposes of this simple demonstration, we considered only the function elements and not the characters. Thus from a bag-of-elements similarity point of view they are all identical. However, when the position information is bound in with the function elements, a slightly different picture emerges:

```
Structural similarity of Othello with A Midsummer Night's Dream:
0.946345

Structural similarity of Othello with Twelfth Night:
```

```
0.946345

Structural similarity of A Midsummer Night’s Dream with Twelfth Night:
1.0
```

That is, in our simplified example, Othello is seen to be the odd one out. These examples may generalize to some extent: stories with marriages at the end are often happy endings, whereas some stories with marriages earlier in the narrative may still leave room for tragic outcomes.

A larger scale analysis of this sort is underway with the predications extracted from the Afanase’ev folk tales, though the results are too preliminary to report as yet.

4 Conclusion and future research

In this paper, we explored the utility of high-dimensional vector representations as a means to model narrative text drawn from folk tales. The methods used thus far have involved general algebraic structures including vector symbolic architectures and their applications to modelling semantic relations. A new challenge posed in the folk tales domain is that the corpora are far smaller and the vocabulary is much less controlled than in the biomedical domain. To meet this challenge, a key methodological advance in the current work involved the use of trained semantic predicate vectors as a means to map between related verbs. In our experiments, verbs representing similar relationships were clustered together effectively, even though the precise semantic roles of their arguments were not known. Preliminary results suggest PSI searches conducted by generating composite cue vectors from these semantic predicate vectors retrieve similar results when similar verbs are employed, potentially obviating the need to normalize predicates explicitly. Furthermore, the temporal ordering of events within a narrative can also be used to distinguish various patterns of functional constituents and explain the similarities between narrative structures. In future work, we will combine these approaches in order to develop representations that incorporate both the sequence of events, and the semantics of the terms with which they are described. In doing so, we hope to reveal the archetypal structures underlying folk narratives.

Acknowledgments: This research was supported by US National Library of Medicine grant R01 LM011563 and NIH/BD2K supplement R01LM011563-02S1. We are also indebted to Tom Landauer for providing us with the TASA corpus.

References

1. Z. Harris, *The form of information in science: analysis of an immunology sublanguage*. Springer Science & Business Media, 1989, vol. 104.
2. C. Friedman, P. Kra, and A. Rzhetsky, “Two biomedical sublanguages: a description based on the theories of zellig harris,” *Journal of biomedical informatics*, vol. 35, no. 4, pp. 222–235, 2002.
3. R. Grishman and R. Kittredge, *Analyzing language in restricted domains: sublanguage description and processing*. Psychology Press, 2014.

4. Z. S. Harris, “The structure of science information,” *Journal of biomedical informatics*, vol. 35, no. 4, pp. 215–221, 2002.
5. V. Propp, *Morphology of the Folktale*. University of Texas Press, 2010.
6. S. Darányi and L. Forró, “Detecting multiple motif co-occurrences in the aarne-thompson-uther tale type catalog: A preliminary survey,” 2012.
7. T. Cohen, R. Schvaneveldt, and T. Rindflesch, “Predication-based semantic indexing: Permutations as a means to encode predications in semantic space,” *AMIA Annu Symp Proc.*, pp. 114–8, 2009.
8. T. Cohen, D. Widdows, R. W. Schvaneveldt, P. Davies, and T. C. Rindflesch, “Discovering discovery patterns with predication-based semantic indexing,” *Journal of biomedical informatics*, vol. 45, no. 6, pp. 1049–1065, 2012.
9. S. Draghici, P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichita, C. Georgescu, and R. Romero, “A systems biology approach for pathway level analysis,” *Genome research*, vol. 17, no. 10, pp. 1537–1545, 2007.
10. N. Ma, J. Guan, and Y. Zhao, “Bringing pagerank to the citation analysis,” *Information Processing & Management*, vol. 44, no. 2, pp. 800–810, 2008.
11. T. M. Fruchterman and E. M. Reingold, “Graph drawing by force-directed placement,” *Software: Practice and experience*, vol. 21, no. 11, pp. 1129–1164, 1991.
12. S. Benzer, “On the topology of the genetic fine structure,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 45, no. 11, p. 1607, 1959.
13. O. Sporns, “The human connectome: a complex network,” *Annals of the New York Academy of Sciences*, vol. 1224, no. 1, pp. 109–125, 2011.
14. K. Guruharsha, J.-F. Rual, B. Zhai, J. Mintseris, P. Vaidya, N. Vaidya, C. Beekman, C. Wong, D. Y. Rhee, O. Cenaj *et al.*, “A protein complex network of drosophila melanogaster,” *Cell*, vol. 147, no. 3, pp. 690–703, 2011.
15. S. Darányi and P. Wittek, “Demonstrating conceptual dynamics in an evolving text collection,” *Journal of the American Society for Information Science and Technology*, vol. 64, no. 12, pp. 2564–2572, 2013.
16. P. Wittek, S. Darányi, E. Kontopoulos, T. Moysiadis, and I. Kompatsiaris, “Monitoring term drift based on semantic consistency in an evolving vector field,” *arXiv preprint arXiv:1502.01753*, 2015.
17. S. Johnson, “Review of harris et al.: The form of information in science: analysis of an immunology review of harris et al.(1989): The form of information in science: analysis of an immunology review of harris et al.(1989): The form of information in science: analysis of an immunology sublanguage,” *Kluwer, Dordrecht. Computational Linguistics.*, vol. 15, no. 3, pp. 190–192, 1989.
18. T. P. MURPHY, “Propp’s morphology as narrative dna: The 29-function plot genotype of “the robber bridegroom”.”
19. R. Bod, B. Fisseni, A. Kurji, and B. Löwe, “Objectivity and reproducibility of proppian narrative annotations,” in *Proceedings of the Third Workshop on Computational Models of Narrative. Ed. by Mark Alan Finlayson*, 2012, pp. 17–21.
20. S. Malec, “Autopropp: Toward the automatic markup, classification, and annotation of russian magic tales,” in *Proceedings of the First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts*, 2010, pp. 112–115.
21. P. Lendvai, T. Declerck, S. Darányi, and S. Malec, “Propp revisited: Integration of linguistic markup into structured content descriptors of tales,” 2010.
22. D. Widdows and T. Cohen, “Reasoning with vectors: a continuous model for fast robust inference,” *Logic Journal of IGPL*, p. jzu028, Nov. 2014. [Online]. Available: http://jigpal.oxfordjournals.org/content/early/2014/12/03/jigpal.jzu028

23. T. C. Rindflesch and M. Fiszman, "The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text," *Journal of Biomedical Informatics*, vol. 36, pp. 462–477, 2003.
24. O. Bodenreider, "The unified medical language system (UMLS): integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, no. Database Issue, p. D267, 2004.
25. A. R. Aronson and F. Lang, "An overview of MetaMap: historical perspective and recent advances," *Journal of the American Medical Informatics Association*, vol. 17:, pp. 229 –236, May 2010.
26. T. Cohen, D. Widdows, R. Schvaneveldt, and T. Rindflesch, "Discovery at a distance: Farther journeys in predication space." in *Proc First International Workshop on the role of Semantic Web in Literature-Based Discovery (SWLBD2012)*, Philadelphia, PA, Oct. 2012.
27. T. Cohen, D. Widdows, and T. Rindflesch, "Expansion by analogy: A vector-symbolic approach to semantic search," *To Appear In: Proceedings of the 8th International Symposium on Quantum Interactions, Filzbach, Switzerland.*, p. 2014.
28. A. Fader, S. Soderland, and O. Etzioni, "Identifying relations for open information extraction," in *Proceedings of the Conference of Empirical Methods in Natural Language Processing (EMNLP '11)*, Edinburgh, Scotland, UK, July 27-31 2011.
29. R. W. Gayler, "Vector symbolic architectures answer jackendoff's challenges for cognitive neuroscience," in *In Peter Slezak (Ed.), ICCS/ASCS International Conference on Cognitive Science*, Sydney, Australia. University of New South Wales., 2004, pp. 133–138.
30. P. Kanerva, J. Kristofersson, and A. Holst, "Random indexing of text samples for latent semantic analysis," *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, vol. 1036, 2000.
31. T. Cohen, R. Schvaneveldt, and D. Widdows, "Reflective random indexing and indirect inference: A scalable method for discovery of implicit connections," *Journal of Biomedical Informatics*, vol. 43, no. 2, pp. 240–256, Apr. 2010.
32. D. I. Martin and M. W. Berry, "Mathematical Foundations Behind Latent Semantic Analysis," in *Handbook of Latent Semantic Analysis*, 2007.
33. Widdows, D, Cohen, T, and DeVine, L, "Real, Complex, and Binary Semantic Vectors," in *Quantum Interaction*, ser. LNCS, J. R. Busemeyer, F. Dubois, A. Lambert-Mogiliansky, and M. Melucci, Eds. Paris, France: Springer Berlin Heidelberg, 2012, no. 7620.
34. P. Kanerva, "Binary spatter-coding of ordered k-tuples," *Artificial Neural Networks—ICANN 96*, pp. 869–873, 1996.
35. H. Kamp and U. Reyle, *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*. Springer Science & Business Media, 1993, no. 42.
36. D. Widdows and T. Cohen, "Graded semantic vectors: An approach to representing graded quantities in generalized quantum models," in *Under review*, 2015.
37. ——, "Reasoning with vectors: a continuous model for fast robust inference," *Logic Journal of IGPL*, p. jzu028, 2014.