ПИРОН Πυρρων АКАДЕМИЧНО ЕЛЕКТРОННО СПИСАНИЕ ЗА КУЛТУРА И ИЗКУСТВА

CULTURAL CENTER of SU "St. Kliment Ohridski"

ISSN 2367-7031 / www.piron.culturecenter-su.org
Vol. 20 / 2020 / READING MODES IN THE DIGITAL AGE
URL: http://piron.culturecenter-su.org/wp-content/uploads/2020/11/20-2020-Alexander-Popov-Vector.pdf

Vector Semantics and Algorithmically-Assisted Close Reading

Alexander Popov

Abstract: The paper explores a point of convergence between literary theory and computational semantics. Using as its starting point the idea of *deformance* developed within the digital humanities, it seeks to harness a certain kind of *reading machines* to the purposes of literary criticism. These machines are conceptualized as *cognitive models* of some idealized readers and as such can provide insight into the requirements for the productive reading of specific kinds of texts. The theoretical underpinnings of the models are traced to the field of distributional semantics, and the computational – to that of natural language processing. The relevance of these *vector space models* is evaluated empirically via an algorithmically-assisted reading of a text from the genre of science fiction. The task is specifically selected so as to be a good testbed for the capability of the reading machines to actually model different kinds of readers, as predicted by literary theory.

Keywords: digital humanities, computational semantics, distributional semantics, natural language processing, science fiction studies

Alexander Popov, PhD, is a postdoctoral researcher in computational linguistics at the Bulgarian Academy of Sciences. He teaches a course on science fiction at the University of Sofia "St. Kliment Ohridski." His research interests range over computational semantics, literary theory, modern philosophy, utopian studies and ecological criticism.

Reading Machines and Textual Deformance

In *Reading Machines: Toward an Algorithmic Criticism*, Stephen Ramsay outlines a vision of computer-assisted literary studies.[1] Ramsay's programmatic text focuses on the instrumental aspect of computational analysis, but in a way that puts novel digital tools within the same range of supposedly non-algorithmic traditional ones:

> *The "algorithmic criticism" proposed here seeks, in the narrowing forces of constraint embodied and instantiated in the strictures of programming, an analogue to the liberating potentialities of art. It proposes that we create tools—practical, instrumental, verifiable mechanisms—that enable critical engagement, interpretation, conversation, and contemplation. It proposes that we channel the heightened objectivity made possible by the machine into the cultivation of those heightened subjectivities necessary for critical work.*[2]

These *reading machines* are not general AIs that would eventually obviate human critics. Rather, they complement criticism by refocusing it at different scales, making it more precise, and liberating it from the straightjacket of single viewpoints. Their ability to sift deterministically through enormous amounts of text for patterns can provide a more robust foundation for actual critical work. Machine and human readers are not seen as alternatives:

> *[A]lgorithmic criticism attempts to employ the rigid, inexorable, uncompromising logic of algorithmic transformation as the constraint under which critical vision may flourish. The hermeneutic proposed by algorithmic criticism does not oppose the practice of conventional critical reading, but instead attempts to reenvision its logics in extreme and self-conscious forms.*[3]

Key to the argument is the idea of *algorithmic transformation* as distinct from simple measurement, at which automated systems naturally excel. The discovery of "facts" about literary works – such as word frequencies and metric structures – does not yield "the principal objects of study in literary criticism"; such facts are only a prerequisite to "establish webs of interrelation and influence."[4] If the function of algorithmic tools for analysis were merely to perform elaborate counting at humanly impossible scales and with minimal error, these machines would be nothing but extensions to already existent theory.

But if algorithms could "assist the critic in the unfolding of interpretative possibilities,"[5] that would constitute a truly novel contribution, continuous with the

---

[1] Stephen Ramsay, *Reading Machines: Toward and Algorithmic Criticism* (Champaign: University of Illinois Press, 2011).
[2] Ibid, X.
[3] Ibid, 32.
[4] Ibid, 7.
[5] Ibid, 10.

established frameworks of critical reading. Such assistance would change the status of computational tools from the merely observational to the hermeneutic.[6] Ramsay finds this potentiality in textual *deformance* – a combination of different concepts, such as "form," "deform," "perform" – i.e. performative translation of the text in search of a new organizing matrix.[7] A deformance could be a literal rewriting of the analyzed work, a compilation of lists for further research, or a reading through a particular theoretical framework. This implies that computational deformance is not qualitatively different from most critical methods, so long as they all rely on well-defined procedures.[8] We could thus conceive of simple tools that help the critic effect such useful transformations, of "deformance machines"[9]; and since all critical interpretation can be viewed as applying such machine-like procedures, "deformance becomes not just 'the best way' […], or the new way […], but an extreme form of the only way—the way it has always been done."[10] Criticism becomes thinkable in terms of "thinking-with,"[11] of different discourses passing though each other and enacting deformances, or particular *agential cuts.*[12]

In this article I investigate a particular kind of deformance, such that endows the critic with an enlarged capacity for conjecture about the fictional worlds of literature.[13] It is perhaps a truism that language expresses conceptual systems differently depending on the language user. I will present a method for a particular kind of deformance of conceptual systems. The resulting reading machines can be naïvely thought of as *cognitive models*, furnishing the critic with easily accessible perspectives toward the text that might differ significantly from his or her own.

This kind of perspectivization is both easily replicable and hackable,[14] as it relies on transformational principles relatively independent of historical, genre, and discourse differences. It is also mediated through the use of a specific rubric of interaction between the cognitive model and the text – that of the list. Ramsay describes the list as a "paratext that

---

[6] Ibid, 10.
[7] Ibid, 33.
[8] Ibid, 16.
[9] Ibid, 36.
[10] Ibid, 38.
[11] Donna J. Haraway, *Staying With the Trouble: Making Kin in the Chthulucene* (Durham: Duke University Press, 2016).
[12] An agential cut, in the paradigm of agential realism, is a necessary distinction between what is being considered as an object of observation and what is being excluded from such an observation (and therefore masked out). These separations are continuously enacted whenever knowledge is produced. Agential realism is the central topic in Karen Barad, *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning* (Durham: Duke University Press, 2007).
[13] Stephen Ramsay, "Reading Machines," 16.
[14] Ibid, 17.

now stands alongside the other, impressing itself upon it and upon our own sense of what is meaningful."[15] The class of reading machines that I focus on has its origins in the discipline of distributional semantics and has been further developed within modern computational semantics. I briefly introduce a number of key related concepts, by way of a study that has significantly influenced my own: Michael Gavin's "Vector Semantics, William Empson, and the Study of Ambiguity."[16]

Vector Space Models and Literary Analysis

Gavin's survey is an attempt to introduce *vector space models* to humanists and demonstrate how these computational tools are in some ways similar to established techniques of literary criticism. Gavin turns to the work of William Empson as an example of the kind of deformance achievable by such models. Empson performed meticulous analyses of the possible meanings of words in context. He used a comprehensive dictionary in his reading in order to perform an inverted variant of *word sense disambiguation* (WSD) – a core task in contemporary computational linguistics. Whereas typically WSD restricts the possible meanings of a word form, Empson's reading seeks to hold all of them in "productive interpretive juxtaposition."[17] Instead of collapsing the interpretation of a text into a stable and determinate description, meaning is exploded into a multi-dimensional field of possibilities, where every word is able to connect in many ways with all the rest.

This does not simply mean that the number of determinate interpretations of a text is enlarged, even if to an extreme degree. By sustaining all possible senses of all the words in context as active virtualities, the reading mind opens up vistas to the territories which lie in between the sense-points in the interpretive space. Empson views words as "'compacted doctrines that always carried their various senses as latent semantic potential."[18] A word, therefore, not only carries with itself a range of associated concepts but also encodes the potential of these points in concept-space to extend in various directions, to reach out to other word senses, and activate meaning representations lying dormant in between lexicalized meanings. Critical reading emerges from this view as a practice of *word-making*, with the

[15] Ibid, 12.
[16] Michael Gavin, "Vector Semantics, William Empson, and the Study of Ambiguity," *Critical Inquiry*, 44.4 (2018): 641-673, https://doi.org/10.1086/698174.
[17] Ibid, 642.
[18] Ibid, 641.

reader pushing the conceptual and contextual envelope of a word to previously non-apparent reaches. I will argue later on that it is also a practice of *world-making*.

Empson used the dictionary as a model of the lexical universe occupied by poems,[19] but as such lexical resources are static, they can only adumbrate this "latent potential."[20] As Gavin observes, Empson was aware of this problem. According to him, words have:

> *not so much a number of meanings as a body of meaning continuous in several dimensions; a tool-like quality, at once thin, easy to the hand, and weighty, which a mere statement of their variety does not convey. In a sense all words have a body of this sort; none can be reduced to a finite number of points, and if they could the points could not be conveyed by words.*[21]

A word has its own logic of manifesting itself in texts; it is not simply a collection of instructions to be followed. Rather, it is more akin to an instrument that restricts its own possible usages, but at the same time allows for almost infinite variation within those bounds. Words are like objects, or processes, depending on the theoretical preference:

> *Words have bodies and agency, Empson argues. Even a sort of personhood. They occupy an invisible lexical "thoughtspace" where they break apart and recombine to form superstructures, molding opinions and otherwise forging human experience.*[22]

At this point vector space models (VSMs) are introduced by Gavin as a means to transcend the limitations of the dictionary. Whereas lexicographic resources describe words by merely listing their possible senses, VSMs represent them by constructing high-dimensional spaces and situating the words in those spaces. Within them the representation of words varies along the various meaning dimensions, and it usually varies in continuous fashion, i.e. in terms of real number values instead of in terms of categorically distinct symbolic encodings. This makes it possible to identify multiple clusters of concepts and topics that partition the thoughtspace infinitely. Words are situated in "a vast, interconnected space of meaning"[23]; the points in space they occupy are merely starting positions from which to unfold their meaning in context, and their meaning components (i.e. their coordinates in the space) might or might not be verbally expressible.

One way to represent words in numeric format, which can be then read and manipulated by natural language processing software, is to encode them as *one-hot vectors*. Consider the following toy corpus, which we will use to obtain word representations: "Every cyborg loves science and science fiction." It contains a total of seven tokens (not counting the

[19] Ibid, 645.
[20] Ibid, 646.
[21] William Empson, *Seven Types of Ambiguity* (New York, 1966), 48. Quoted in: Michael Gavin, "Vector Semantics," 647.
[22] Michael Gavin, "Vector Semantics", 647.
[23] Ibid, 641.

full stop) and a total of six unique types. One-hot representations of the words would therefore look like this:

"every" = [1,0,0,0,0,0] "cyborg" = [0,1,0,0,0,0] "science" = [0,0,0,1,0,0]

The words are marked by 1-values at unique vector positions, where the schema for picking the 1-slots is usually a simple one: order of appearance in the reference corpus, alphabetical order, etc. Each word in the vocabulary gets to be represented by a point in this six-dimensional space, and each one is orthogonal to the rest, i.e. they are all equidistant. There is no possible concept of word similarity in this space.

A solution to this problem is to encode the semantics of words in the vectors themselves. This line of research points back to the 1950s, when Zellig Harris proposed the distributional hypothesis, often encapsulated anecdotally by a quote from John Firth: "You shall know a word by the company it keeps."[24] In other words, "difference of meaning correlates with difference of distribution."[25] By examining many instances of real-life word usage and tabulating the collocations a word enters into, linguists can compile meaning profiles that are much more flexible. A word is no longer defined by its mere symbolic distinctness, but rather via its myriad context-specific relations to other lexical units.

With the advent of large corpora and methods for their automatic manipulation it became possible to carry out this kind of work via software. This new scalability resulted in the production of truly comprehensive VSMs built around the distribution hypothesis. The goal in constructing these models is to obtain a vector per word in the vocabulary, which defines it in terms of its values. There are many different techniques for obtaining VSMs. Consider the following one. First, we extend our corpus to the following collection of sentences: "Every cyborg loves science and science fiction. A cyborg is both flesh and metal. Metallurgy is a science that studies metal materials." We can define a VSM on the basis of *context co-occurrence*. We define a shared context naively: as a single sentence in the corpus. This time we omit function words for the sake of compactness and thus get the following table of co-occurrences:

[24] John R. Firth, "A Synopsis of Linguistic Theory 1930–1955," *Studies in Linguistic Analysis* (1957): 1-32.
[25] Zellig Harris, "Distributional Structure," *The Structure of Language: Readings in the Philosophy of Language*, ed. Jerry A. Fodor and Jerrold J. Katz (Englewood Cliffs, N.J., 1964), 43.

	cyborg	loves	science	fiction	flesh	metal	metallurgy	studies	materials
cyborg	0	1	2	1	1	1	0	0	0
loves	1	0	2	1	0	0	0	0	0
science	2	2	1	1	0	1	1	1	1
fiction	1	1	2	0	0	0	0	0	0
flesh	1	0	0	0	0	1	0	0	0
metal	1	0	1	0	1	0	1	1	1
metallurgy	0	0	1	0	0	1	0	1	1
studies	0	0	1	0	0	1	1	0	1
materials	0	0	1	0	0	1	1	1	0

Table 1: A co-occurrence matrix showing how often words in a corpus share a context.

This is of course a very simplistic representation based on extremely limited data. But already we can see that the approach gives us word representations that are more meaningful than the one-hot vectors. Were we to use much larger corpora, the procedure would yield significantly more fine-grained and informative representations. There are, however, a number of problems with this approach. With large corpora, the size of the vocabulary also grows large and so do the word vectors. Not only is this computationally inefficient but also still somewhat simplistic. While meaning is now more nuanced and words are complexly interrelated with each other, they are still expressed in terms of other words, a design feature we would need to transcend in order to fulfill Empson's vision.

Modern methods for learning VSMs offer a way out of this prison house of language. One possible approach is to use techniques from linear algebra in order to shrink the dimensionality of the VSM resulting from the co-occurrence matrix. These methods first identify the most expressive dimensions of variation (i.e. degrees of freedom, also called *principal components*) between all word vectors. Roughly, this means that every dimension of the matrix represents an axis of variation, and data points (which in our case are words – encoded as rows and columns in the matrix) are distributed relative to each other along all of these axes. If most words are bundled closely together along one axis, then this particular dimension does not encode a large amount of variance in the data set (i.e. corpus, in our case). For instance, the word "the" likely co-occurs heavily with almost all other words in any large

corpora and therefore does not constitute a particularly informative dimension of variation. Axes which encode the greatest amount of variance are approximately preserved in the algebraic transformations enacted by these methods, whereas low-information dimensions are discarded. This process results in the overall reduction of the dimensionality of the VSM. Thus, instead of size *NxN*, where *N* is the size of the vocabulary, the model becomes *NxK*-big, where *K* is the number of principle components per vector.[26] These *K* dimensions of meaning are ideally independent of each other – one cannot be expressed in terms of the rest, just as in semantics, core semantic features are irreducible to each other and are combined to express more complex meanings.

Another method for obtaining VSMs is through the use of *artificial neural networks* (ANNs), computational architectures loosely inspired by biological neural networks. ANNs are able to learn *approximate* representations of the same kind of dependencies which can be encoded in large co-occurrence matrices. This has been a big breakthrough for natural language processing, since performing operations like principal component analysis on truly large matrices can be computationally intractable. ANNs, and especially *shallow* ANNs,[27] have made possible the training of VSMs on corpora comprising billions of words. VSMs built on data of such scale are able to encode an enormous range of meaning variability (see, for instance, figure 1, which represents in 3D space the relations between the term "king" and its 100 closest neighbors).

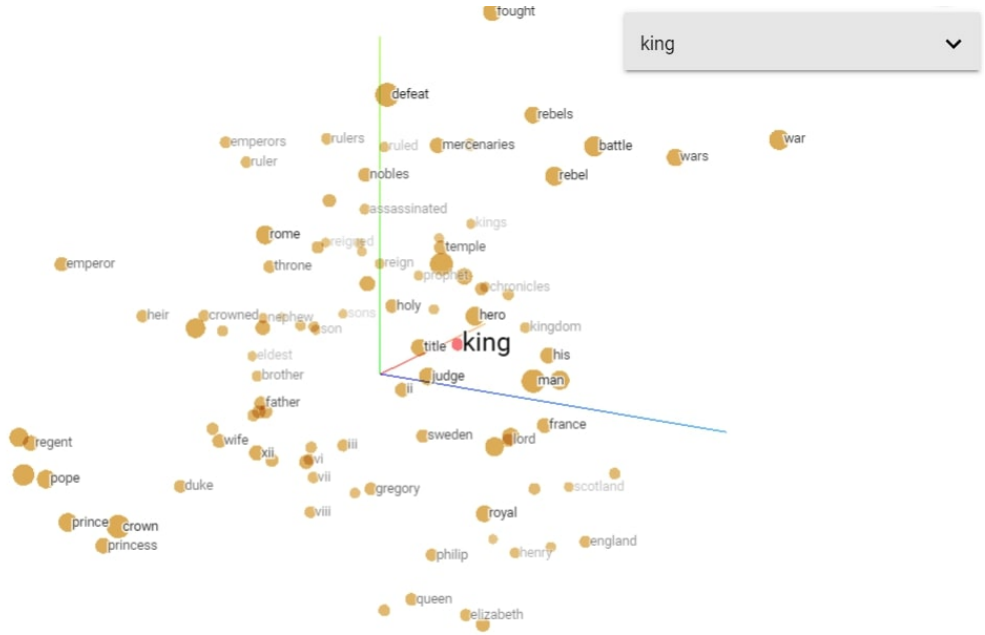

Figure 1: Lexical neighborhood of the word "king" in 3D space. . Source

[26] In practical applications in natural language processing, *K* typically varies in the interval [50:1000], whereas for large corpora, *N* usually stands for at least *tens of thousands* of words.

[27] The most important reference here is: Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean, *Efficient Estimation of Word Representations in Vector Space* (arXiv preprint arXiv:1301.3781, 2013).

It is difficult to say precisely what the dimensions of a VSM encode, since the methods described above produce a black-box representation of meaning: there are no resulting labels that identify what each dimension signifies. One oblique method to extract such information is connected with a truly impressive property of VSM: that simple arithmetical operations on vectors from within the space produce results that mirror what a lexico-semantic theory of meaning would predict. Here is a popular example:

$$\text{Vector}[\text{“king”}] - \text{Vector}[\text{“man”}] + \text{Vector}[\text{“woman”}] \approx \text{Vector}[\text{“queen”}]$$

Taking the representation of the word “king,” subtracting from it that of “man” and adding that of “woman” gives a vector which is very close to the one corresponding to “queen.” Operations of this kind can align word representations with regards to a wide variety of meaning features: from semantic (sex, animateness, color), through morphosyntactic (gender, number, subjecthood), to world knowledge of various kinds (association of terms with genres, discourses, etc.). Comparing the values of the analyzed vectors can give us clues as to where the different features are encoded. The separate dimensions, however, rarely happen to stand in one-to-one relations with semantic features adopted in linguistic theory. While presenting an unfortunate inconvenience to the model interpreter, this tendency is actually predictable from Empson’s observations, which hold that the coordinates of word meaning cannot be described in mere words.

Another implication, even more important, is that the combination of words can result in new meanings within the vector space, and this kind of combinatorics can now be formalized in terms of mathematical operations. Similarity between words now becomes the geometric distance between them. Lexical *word-locations* can be seen as the contours of conceptually expressible *world-volumes*. Textual meaning becomes locatable within the space in-between words:

> *Continuous methods enable us to model not only atoms of meaning such as words, but the space or void in between these words. Whole passages of text are mapped to points of their own in this void, without changing the underlying shape of the space around them.*[28]

The procedures for amalgamating words into complex meaning constructs vary in their effectiveness and complexity, but even a very simple one, such as averaging the sum of the

[28] Dominic Widdows, *Geometry and Meaning* (Stanford: CSLI Publications, 2004), 164. Quoted in: Michael Gavin, “Vector Semantics,” 664.

vectors of all words in a passage, known as calculating their *centroid*, can work reasonably well, as shall be seen later.

Thus VSMs present a method for capturing conceptual systems that are themselves the outgrowth of specific textual canons. In this sense VSMs can be seen as *cognitive models* of abstract readers who have been exposed to controlled collections of texts. As crude as those approximations are to actual human learning, their effectiveness justifies in-depth research on their application to literary criticism. Just as Empson used a dictionary to make explicit the many possible interrelations of words in fictional contexts, VSMs can be used to generate alternative readings that probe different readerly competences.

In his essay on vector semantics, Gavin builds a VSM and applies it as deformance machine to an excerpt from a literary text. He constructs a corpus of around 18 000 publicly available documents dated 1640 to 1699; preprocesses it, extracting 1 751 keywords (i.e. the most prominent of the meaning dimensions of the model); and subsequently takes context windows extending for five words around each occurrence of the keywords.[29] From these windows he calculates co-occurrence counts. At the end he constructs a large matrix of 1 751 columns (the keywords) and 28 235 rows (the collocate terms), with the co-occurrence counts being used to fill in the cells; the matrix is thus a VSM like the ones described here. He uses a mathematical technique called *K-means clustering* in order to obtain the closest neighbors to words in the model, which can then be projected onto a two-dimensional representation of the much higher dimensional VSM.[30]

Gavin uses the semantic space to analyze a popular passage from John Milton's *Paradise Lost*, book 9, in which Satan has infiltrated Heaven as a serpent and, upon seeing Eve, is struck still by her, his evil nature momentarily suspended. Gavin's algorithmically-assisted reading focuses in particular on the ninth line of the excerpt: "That space the Evil one abstracted stood." He first generates visualizations of the lexical neighborhoods of each of the four open-class words[31] in the line: "space," "evil," "abstracted," "stood" (see figures 2-5). Various interesting observations can be made on the basis of the individual word distributions. "Space," for instance, has prominent clusters related to temporal and spatial

[29] A context window is defined as a series of tokens preceding, including, and following the target word (i.e. the word about which statistical information is being collected). For instance, a context window extending for two words in each direction and centered around the word "loves" in the sentence "Every cyborg loves science and science fiction" yields the sequence [Every, cyborg, loves, science, and].
[30] Michael Gavin, "Vector Semantics," 662.
[31] I.e. *content words* like nouns, verbs, adjectives and adverbs, as opposed to *closed-class* or *function words*, like pronouns, prepositions, determiners, etc.

extension; “abstracted” associates with terms related to philosophical disciplines like ontology and epistemology; “evil” is closely related with concepts of intentional doing, etc.

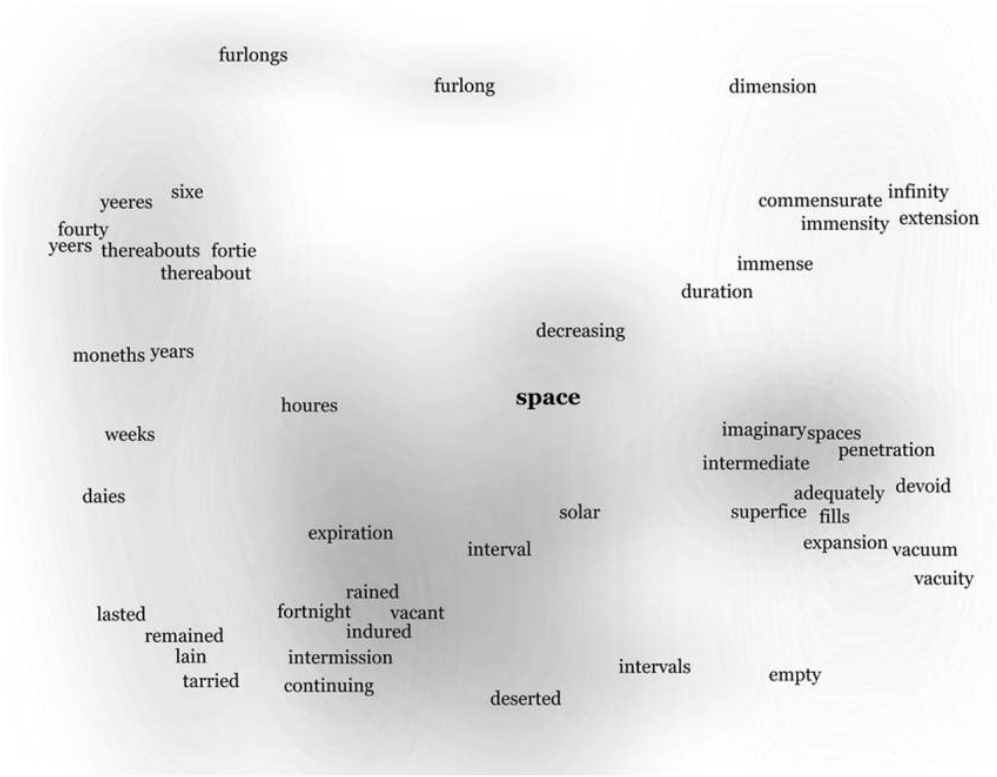

Figure 2: Lexical neighborhood of the word “space.” Source: Gavin, 2018. Reproduced with the author’s permission.

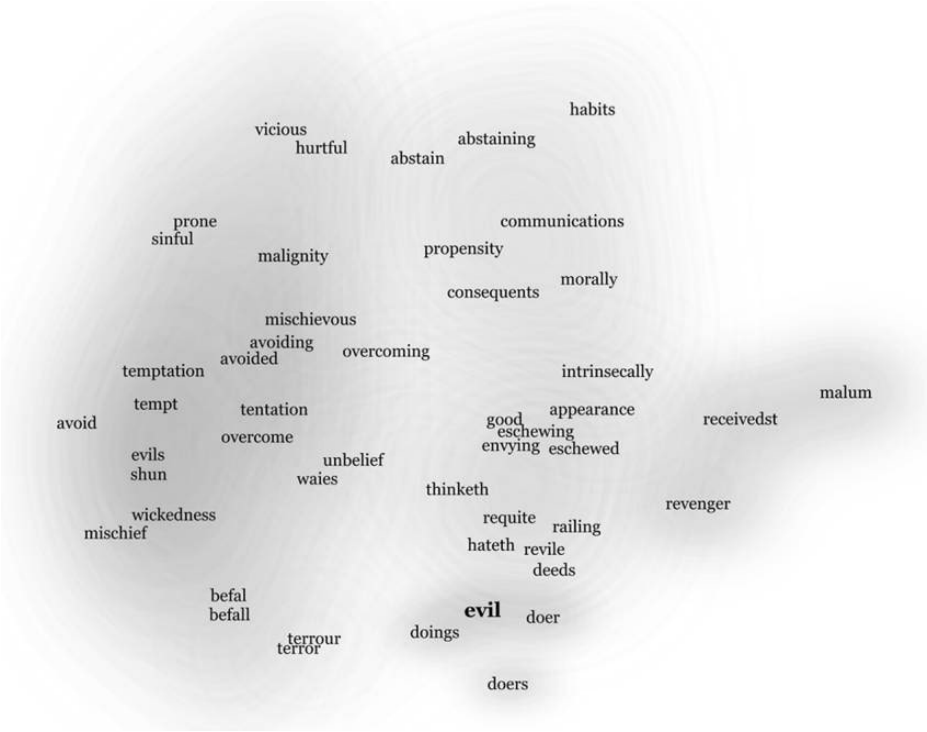

Figure 3: Lexical neighborhood of the word “evil.” Source: Gavin, 2018. Reproduced with the author’s permission.

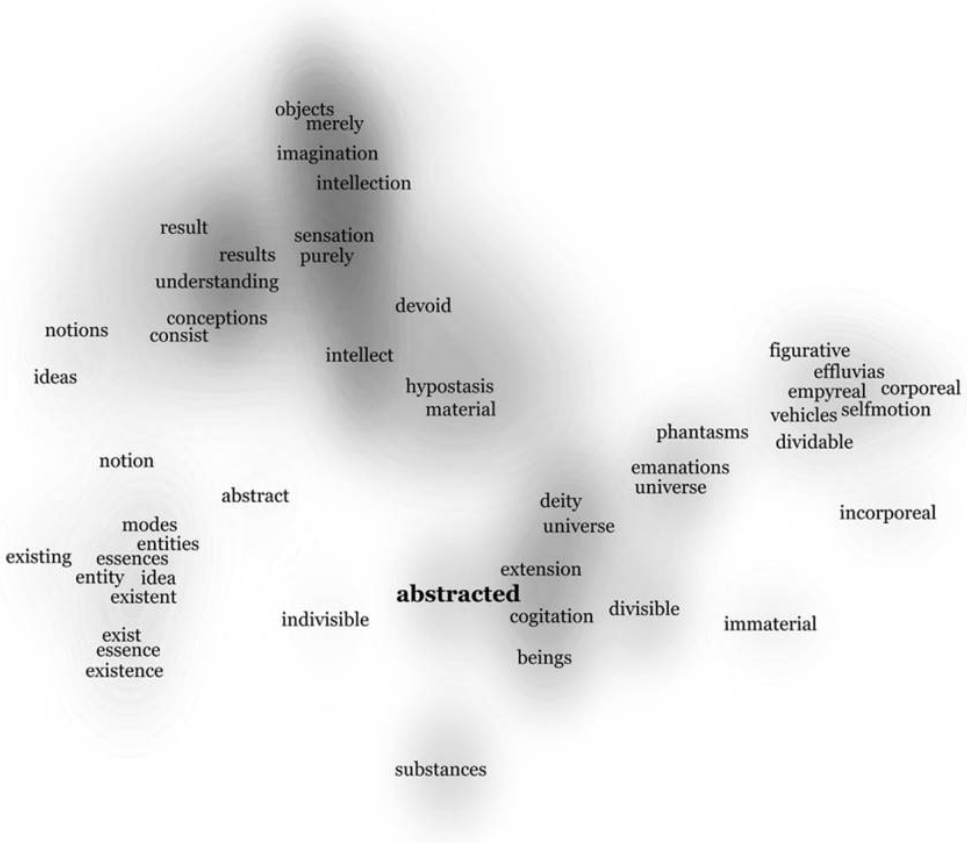

Figure 4: Lexical neighborhood of the word “abstracted.” Source: Gavin, 2018. Reproduced with the author’s permission.

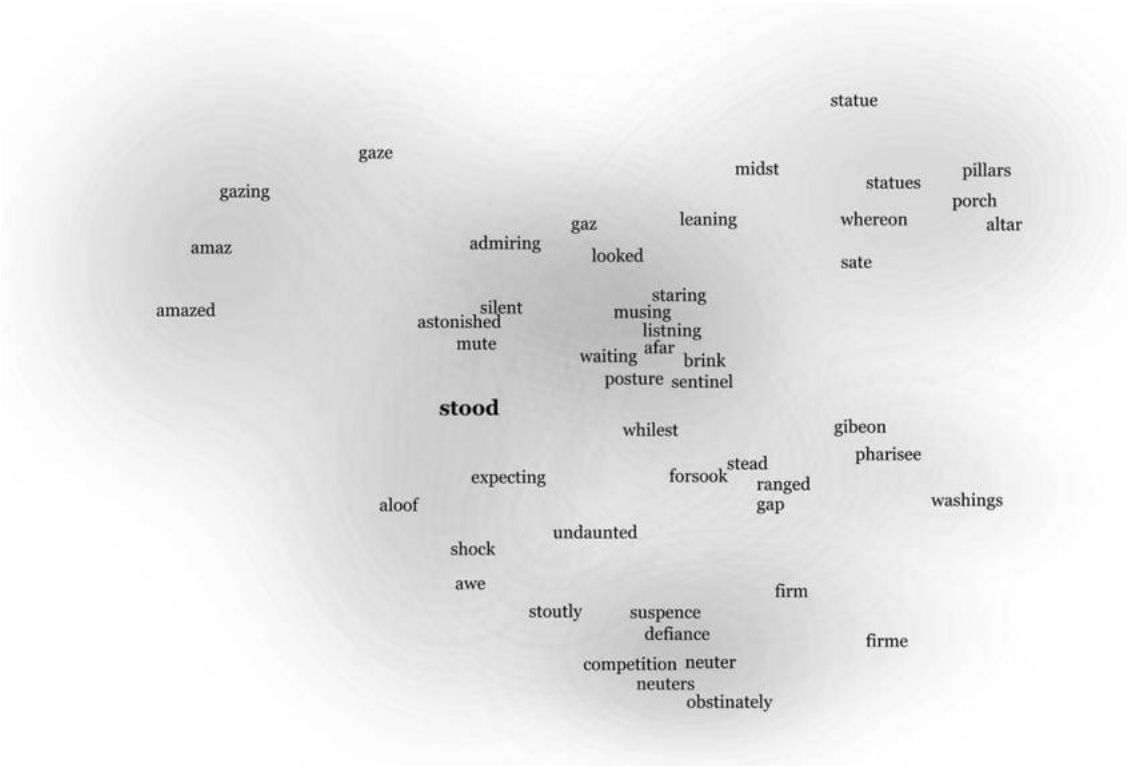

Figure 5: Lexical neighborhood of the word “stood.” Source: Gavin, 2018. Reproduced with the author’s permission.

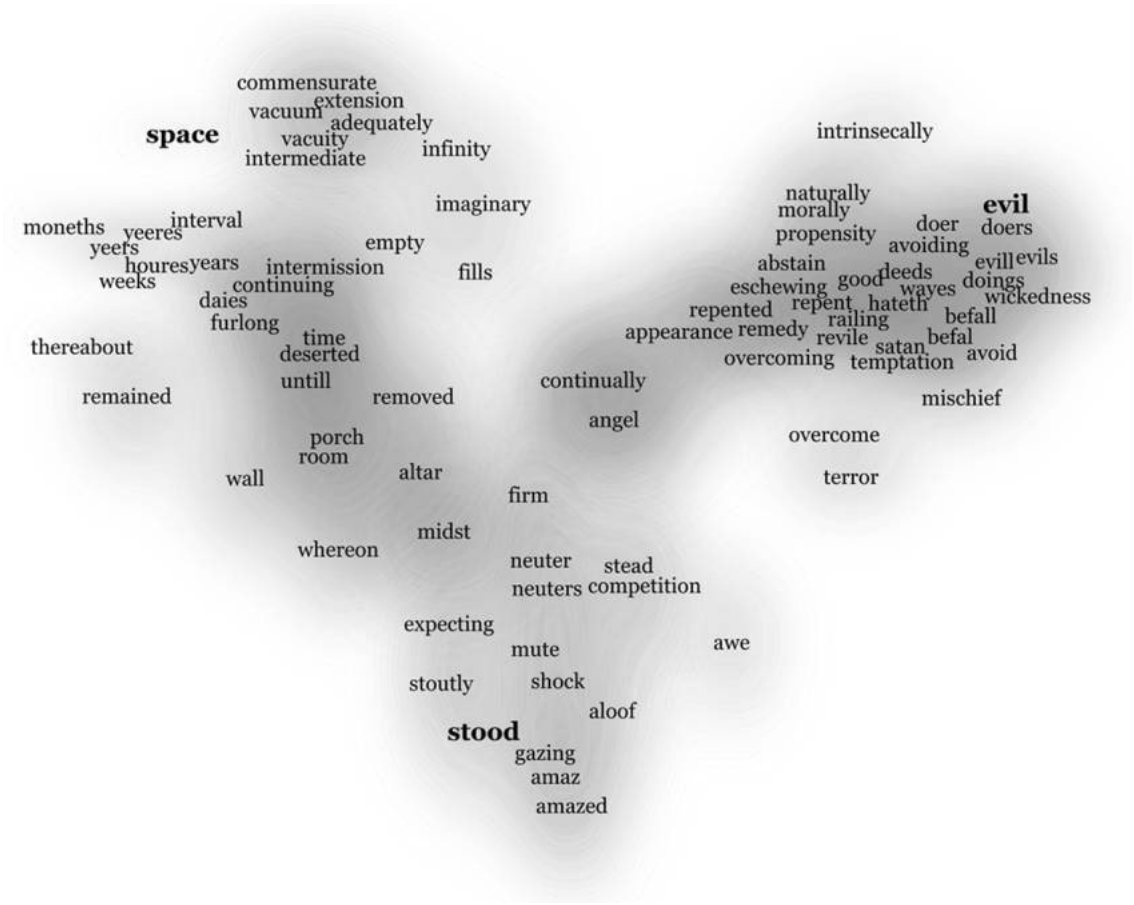

Figure 6: Semantic neighborhood of the vector constructed out of the representations of "space," "evil," "abstracted" and "stood." Source: Gavin, 2018.
Reproduced with the author's permission.

A most striking thing happens, however, when the centroid of the four word vectors is computed and its own semantic neighborhood is visualized, which in a way is equivalent to constructing a meaning representation of the whole line (figure 6). Many of the neighbors of the individual component words fall away and new ones appear. Most notably, the whole lexical neighborhood of "abstracted" and the very word itself are canceled out, possibly due to the relative distance of the term from the other three. But echoes of the abstract aspect seem to reverberate in the subspace. We can find those, for instance, in the activation of "removed," which is in the locality of "space" but also close to the center of the composite triangular shape. Also close to the center and in the margins of a word's semantic neighborhood, this time "stood," are words relatively unrelated to it, like "neuter" and "neuters" – possibly closer in meaning to "abstracted." In the vicinity of "evil," the word "satan" makes an appearance; its association with the primary term is commonsensical, but arguably it is buoyed up too by the underlying influence of "abstracted." This is signaled with particular force by the word

that occupies the very center of the triangle – "angel," a concept intimately connected with that of abstraction, diametrically opposed to, but also constitutive of, that of "evil" (especially in 17th century texts).

Following Gavin's interpretation, the sense of potent agency suggested by the neighbors of "evil" seems to melt away into a place of abstraction in the middle of the plot, the eye in the storm in which evil is stilled into its angelic origin, preoccupied again by problems of morality, penitence, overcoming, and appearance. The combined force of three relatively general words – "stood," "space" and "abstracted" – has pulled the powerful concept of evil to a radical reformulation, to a state of being "Stupidly good, of enmity disarm'd."[32]

Modeling Genre Readers in Terms of Vector Spaces

The final part of this article will illustrate the principles of VSM-based computational deformance of literary texts by presenting several purpose-built tools and applying them to a task that should be a particularly good match for the approach.

The task at hand is to read a passage of fictional text that belongs to a particular genre and to automatically generate associative lists that could aid the reader in achieving a fuller understanding of the passage. The insistence on applying the deformance procedure to a genre text is due to the intuition that genre literature relies on conceptual systems which depart in significant ways from those used in non-genre analogues. The genre chosen here is that of science fiction (SF), which is particularly dependent on the reader's capabilities to imagine a fictional world that can be radically, and/or very subtly, divergent from our normative representations of reality. This makes SF a good testbed for investigating the hypothesis that words do indeed compact parts of worlds inside themselves and that the combination of words must yield world constructions overlaid spectrally in-between the verbal coordinates.

The passage that will be subjected to deformance via VSM comes from the very beginning of Philip K. Dick's novel *Do Androids Dream of Electric Sheep?*:

> *A merry little surge of electricity piped by automatic alarm from the mood organ beside his bed awakened Rick Deckard. Surprised – it always surprised him to find himself awake without prior notice – he rose from the bed, stood up in his multicolored pajamas, and stretched. Now, in her bed, his wife Iran opened her gray, unmerry eyes, blinked, then groaned and shut her eyes again.*[33]

---

[32] John Milton, *Paradise Lost* (London, 1667), tei.it.ox.ac.uk/tcp/Texts-HTML/free/A50/A50919.html#index.xml-body.1_div.9. Quoted in: Michael Gavin, "Vector Semantics", 666.

[33] Philip K. Dick, *Do Androids Dream of Electric Sheep?* (London: Victor Gollancz, 1999), 3.

The novel is primarily concerned with the definition of what it means to be human. The very notion is called into question by the plot: Rick Deckard's quest to "retire" several escaped intelligent androids, who are able to blend very convincingly in human populations. Central to the novel is the technological development of a future society whose emotions and memories have become intimately intertwined with technological materialities. Carl Freedman argues that by problematizing this relation (already nascent in the author's own time in the use of drugs and the rise of television) Dick raises "the question of the *historicity* of feelings."[34] The categories of emotion and technology are therefore put in a dialectical relation, and I would argue that this is made possible by SF's specific reading protocols. The latter require the mental construction of fictional worlds that are significantly at odds with the perceived reality and consequently shift received conceptual systems around new centers, through the inevitable mediation of language.

Before moving on to the computational deformance of the text, it is worth lingering briefly over the specificities of the SF genre that might justify deliberately seeking such a contrastive view of the text that I am about to offer – as read by a non-genre reader and by an experienced SF reader. SF has been famously defined by Darko Suvin as a "literature of cognitive estrangement,"[35] a definition that encodes, "on the molecular level,"[36] the genre's tendency to simultaneously make the familiar new, but also necessarily to render it explainable through *some* logical apparatus. Around the same time Samuel Delany, in his paradigmatic essay *About 5,750 Words,*[37] suggested a simple yet powerful model for the productive reading of SF. He reads one sentence word-by-word, providing rich notation of his own supposed time-lapsed meaning-making, as the lexical units follow one another and establish connections in the gaps between them. "The red sun is high, the blue low," is his famous example.[38] He demonstrates through his reading how the interpretive trajectory of the mind would normally explore well-trodden lexical paths, suggestive of perfectly expected, almost trivial modes of writing. That is, until a moment of lexical, and therefore conceptual, rupture, announced by the reference to *a blue sun*. Suddenly the fictional world shifts light years away, and everything about the narrative is transformed.

---

[34] Carl Freedman, *Critical Theory and Science Fiction* (Middleton: Wesleyan University Press, 2000), 32.
[35] Darko Suvin, *Metamorphoses of Science Fiction: On the Poetics and History of a Literary Genre* (Bern: Peter Lang, 2016[1979]), 15-27.
[36] Carl Freedman, "Critical Theory and Science Fiction," 32.
[37] Samuel R. Delany, *The Jewel-Hinged Jaw: Notes on the Language of Science Fiction* (Middleton: Wesleyan University Press, 2009[1978]), 1-15.
[38] Ibid, 5-7.

In this early essay Delany calls these micro-level generic shifts *changes in the subjunctivity level of the text*: what has not happened yet (SF), what could have happened but has not (alternate history), what could not have happened (fantasy), what could have happened (naturalistic fiction).[39] Later Delany evolved a more sophisticated method of reading SF, whereby he holds two worlds in tension with one another – the 'real' and the SF one – as separate sources of meaning with which to make sense of the text.[40] This interposition of the SF world in the reality-text pair blows up the unidirectional supply line of ready-made meaning from the normative knowledge silos to the textual surfaces; it starts challenging the solidity of the "real," while simultaneously defining itself through differentiation from it. It gives voice to the text, freeing it from its role of a mute servant/doomed rebel. This trivalent discursive matrix allows Delany to explore precisely those voids of potential meaning opened up between words in context. He demonstrates how this could be done through meticulous reading, tabulating all rhetorical forces and genre protocols in play and coaxing out their cognitively estranged fictional referents.

Many authors have since suggested similar views on the SF genre. Damien Broderick argues that SF is uniquely suited to interact with the contemporary episteme "because of the unease with which [it] poises its narrative modality between *artistic* attention to the *subject* and *scientific* attention to the *object*."[41] This double vision "must be learned by apprenticeship" and is always dependent on readerly access to a sort of genre encyclopedia, "a mega-text of imaginary worlds, tropes, tools, lexicons, even grammatical innovations borrowed from other textualities."[42] Adam Roberts hypothesizes that SF is shaped by the dialectic between *magic* and *technology*, or in other words – reading in terms of Catholic or Protestant protocols.[43] And according to John Rieder, the SF genre, from a historical perspective, is not so much a particular selection of texts as it is a *way* of drawing relations between them; it is part of the mass media cultural system.[44]

All of these accounts maintain a shared tenet: that SF constitutes a particular practice of meaning construction that crucially relies on the reader and her competencies. This is of great relevance to the practice of deformance and the application of VSM to criticism. If words are indeed compacted doctrines, as Empson believed, the kind of knowledge about

[39] Ibid, 10-11.
[40] Samuel R. Delany, *The American Shore: Meditations on a Tale of Science Fiction by Thomas M. Disch—"Angouleme"* (Middleton: Wesleyan University Press, 2014[1978]).
[41] Damien Broderick, *Reading By Starlight: Postmodern Science Fiction* (London: Routledge, 1995), xi.
[42] Ibid, xiii.
[43] Adam Roberts, *The History of Science Fiction* (Basingstoke: Palgrave Macmillan, 2005).
[44] John Rieder, *Science Fiction and the Mass Cultural Genre System* (Middleton: Wesleyan University Press, 2017).

their potential extensions in the world must be held collectively and individually by the users of each language. And if a genre has its own doctrines that can be learned only through immersion in the genre megatext and meaning-making procedures, then genre readers inevitably create very different fictional worlds compared to non-genre readers. Next I will emulate different kinds of readers via VSMs and elicit from them particular deformances of the passage by Dick, embodied in lists of semantic neighbors like the ones calculated in the algorithmically-assisted reading of Milton.

The procedure for enacting the critical deformance is simple. We focus on just the first sentence of the quoted passage ("A merry little surge of electricity piped by automatic alarm from the mood organ beside his bed awakened Rick Deckard.") and examine how its critical deformance through semantic neighborhood generation might suggest a particular reading of the whole passage. The sentence is first preprocessed: stop words like "a," "of," "by" are removed; the named entity "Rick Deckard" is replaced with the generic identifier "person"; the remaining words are lemmatized. This yields the following list:

[merry, little, surge, electricity, pipe, automatic, alarm, mood, organ, bed, awaken, person]

Various subsets of this list can be used to calculate semantic neighborhoods based on different criteria. I will focus on just two examples here. The first one is selected on the basis of readerly intuition – the group comprising the words "mood," "electricity," and "organ." The second example groups together all elements of the above list.

I present the results of two experimental setups. The first one aims to compare two kinds of cognitive models. The first model is that of a hypothetical reader who has been exposed to large volumes of natural language, but has had minimal contact with SF literature. The second model is that of a hypothetical reader who has read *nothing but* SF. To model the non-genre reader I have used the GloVe word embeddings[45] – one of the most popular off-the-shelf VSMs, trained on a corpus of about six billion words comprising the contents of Wikipedia and a large collection of newswire texts. The SF reader, here called SF2Vec, I have modeled by training a dedicated VSM using the Word2Vec tool.[46] SF2Vec is trained on a corpus of exclusively science fictional texts gathered from *The Pulp Magazine Archive*[47] – a

[45] Jeffrey Pennington, Richard Socher, and Christopher Manning, "Glove: Global Vectors for Word Representation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (2014): 1532-1543.
[46] Mikolov et. al., "Efficient Estimation," 2013.
[47] https://archive.org/details/pulpmagazinearchive%26tab=about&tab=collection.

selection of pulp magazine issues from the 1950s to the 1970s. The GloVe model is undoubtedly much more powerful, as it has been trained on a much larger data set. Yet, the research hypothesis holds that the lists generated by the genre-oriented reader would provide a more productive deformance.

Table 2 shows a comparison between the lists generated with the non-genre (GloVe) and the SF (SF2Vec) cognitive models. It contains the ten neighbors that are closest to the centroid of the word group, i.e. the word labels for the instantiated vectors that are geometrically closest to the averaged vector from the representations of "mood," "organ," and "electricity."

GloVe	**SF2Vec**
organs	stimulation
electric	stimulus
power	feedback
electrical	electrical
tone	impulses
supply	hormones
energy	stimuli
generating	absorption
functioning	harmonic
generator	brain

Table 2: Closest semantic neighbors to the centroid of the group ["mood," "organ," "electricity"], using GloVe and SF2Vec.

The list produced by the GloVe model is hardly useful. It mostly contains words that are only trivially related to the original terms, since they do not contribute any new information that is meaningfully related to the *combination* of the terms. Clearly, the word "electricity" stands in close relations with a great many of the words in the space and is not related in a particularly interesting way with "mood" and "organ," which seem to have weaker influence on the composition of the immediate neighborhood of the centroid. The SF2Vec model, however, produces a very interesting list. From the combination of merely three words, the SF cognitive model has been able to summon up an association with the brain – indeed an organ, running on electricity and capable of producing moods. The rest of the neighborhood is also certainly indicative of processes potentially connected to the neural system and the brain: stimulation, stimulus, feedback, hormones. The exposure to SF

literature has made this cognitive model more sensitive to the implied meaning of such configurations – in the proper context, which Dick's novel certainly is.

Upon examination of the wider semantic neighborhoods according to the two models (the top 100 associations), a few interesting associations do come up from the non-generic conceptual space: "sense," "sentiment," "feeling," "anxiety"; but those can arguably be attributed to the influence of "mood" alone, and thus might appear to be just as trivial as "electrical." The SF-centric model gives many more relevant suggestions: "reflex," "perception," "sensory," "olfactory," "cortical," "cerebral," "responses," "auditory," "chemicals," "neural," "glandular," "subliminal," "enzyme," "entropic," "glands," "sensitivity," "omnipresent," "matrix," "synapses," "cortex," "stimulated," "adaptive," "chemical," "circuits," "artistry," "adrenal," "interaction," "gene," "interplay," "excitation," "integration," "reactions." Most of those connect to the subject of human emotions as produced by the body and controllable by technology. Feeding the simple triangle of terms into the model has activated an apparently cohesive subspace. An experienced and attentive SF reader would probably make a similar interpretation on his or her own, but the VSM allows us to do this in a replicable and directly observable manner, and more importantly, to do it without having to be a reader of SF at all. The list is a deformance of the original text that brings its world and meaning into focus with minimal human supervision.

GloVe	**SF2Vec**
kind	euphoria
you	vibration
even	sensation
so	bladder
sort	spasm
actually	resonance
enough	throb
something	torrent
get	rhythm
just	stimulation

Table 3: Closest semantic neighbors for the centroid of the group [merry, little, surge, electricity, pipe, automatic, alarm, mood, organ, bed, awaken, person], using GloVe and SF2Vec.

The two models produce markedly different lists when used to calculate the semantic neighborhoods of all the terms in the sentence group (table 3). The usefulness of the non-generic cognitive model completely breaks down on this task. It generates only extremely general words, catalyzing no associative chains. This is easily explainable: the technique used here for calculating the center of the neighborhood is a very simple arithmetic operation. When applied to a larger number of vectors from the space, their centroid is all the more general and non-specific. Therefore, it is all the more surprising that the generic cognitive model does produce meaningful results with this word group. At the very top of the list sits "euphoria," which one could argue is a very good approximation of the feeling Rick Deckard wakes up with. "Vibration," "spasm," "throb," "torrent," "stimulation" – these do hint at an almost spatial and bodily representation of the euphoric sensation; electricity is merry because it literally brings euphoria, but perhaps also because it takes a shape of its own in Deckard's mind; it is like the body of a homunculus put there by technological forces transcending the biological.

Going through the hundred closest neighbors, there are few results retrieved by the GloVe model that are worth mentioning: "sense," "sleep," "mind," "feeling" – although all of them are more or less trivially associated with some of the words within the group. SF2Vec yields a much richer crop: "reflex," "pain," "emotion," "elation," "exhilaration," "receptor," "vitality," "dizziness," "torpor," "odors," "alertness," "emanation," "frenzy," "tranquilizer," "awareness," "illumination," "relaxation." Once again, the SF model has been able to tap into a much more interesting topology of meaning, derived from exposure to systemically changed conceptual/world systems.

One might argue, however, that the two VSMs used are too dissimilar. One is trained entirely on non-fiction texts, the other exclusively on fiction. In order to provide a fairer comparison, I have trained three additional VSMs. The first one (called WikiVec) is based solely on the contents of Wikipedia. The second model (WikiSFVec-0.05) is essentially a copy of the first one that is subsequently trained further on *The Pulp Magazine Corpus*. This means that after the model has been successfully trained on the Wikipedia corpus, its parameters are additionally modified via training iterations over the SF data, with the difference that the learning rate (i.e. the velocity of acquiring new knowledge) is then increased from 0.025 to 0.05. The third model (WikiSFVec-0.075) is built analogously but with a learning rate of 0.075 on the SF corpus, i.e. it learns three times faster than on the Wikipedia data. In very crude terms, we have three readers: one who has acquired its whole world knowledge solely through encyclopedic data; another one who has done that and then

switched entirely to science fiction, paying *twice as much* attention; and a final reader who has done the same as the second one but has actually paid *three times* as much attention to the SF texts. The lists generated from these three models should more fairly indicate how reading SF changes the conceptual system of the reader, as the second and third readers are actually *future versions* of the first one.

WikiVec	**WikiSFVec-0.05**	**WikiSFVec-0.075**
vibration	amplification	output
instrument	vibration	transistor
sound	energy	organs
energy	transformer	internal
heating	amplifier	excitation
fluid	brain	absorption
pipe	electrical	transformer
electric	feedback	external
organs	heating	air-conditioning
apparatus	emotion	memory

Table 4: Closest semantic neighbors to the centroid of the group [“mood”, “organ”, “electricity”], using GloVe and SF2Vec.

Table 4 gives the ten closest semantic neighbors to the centroid of the already familiar group [“mood”, “electricity”, “organ”]. The baseline model trained solely on Wikipedia does in fact generate some meaningful associations, and looking further down through the first hundred neighbors one even comes across results like “emotion” (21st), “machinery” (29th), “consciousness” (52nd), “brain” (57th), “circuitry” (62nd), “feedback” (92nd). This seems to suggest that the GloVe model is not just disadvantaged by its lack of exposure to SF texts, but it is also handicapped by its exposure to newswire texts, i.e. the conceptual system of another genre. Training on a purely encyclopedic resource, however, seems to provide at least a partial skillset for interpreting SF, which implies that the genre reader’s conceptual system is not discontinuous with that of normative reality. The WikiSFVec-0.05 model does seem to behave similarly to its own baseline version, but it also pays more attention to some of the key associated terms like “brain,” “emotion,” and “feedback.” In its list of the first hundred neighbors are also words like “mechanism,” “memory,” “machine,” “chemoreceptors,” “neurotransmitter.” WikiSFVec-0.075 seems to have learned a little too anxiously, as “brain” and “emotion” have dropped out of the first hundred neighbors entirely. On the other hand, it

has associated the word group with very specific terms from the domain implicitly activated by the sentence: "neurons" (47th), "cerebellum" (76th), "parasympathetic" (82nd), "sensor" (83rd), "inhibition" (84th), and even "ego" (100th). The increased pace of learning from SF texts has apparently made the reader prone to construct much more detailed and technical conceptual connections between the input terms.

Our algorithmically-assisted explorations of a short segment of Philip K. Dick's prose have demonstrated that it is perfectly possible to use computational tools to activate whole strata of implicit conceptual tissue lying dormant beneath the surface of a text. One can imagine that with the increasing availability of textual data and interest in the digital humanities, it could at some point become manageable to train and deploy huge flocks of such deformance machines. Those could sift through fiction, generating lists, graph structures, maps, and frames, joining forces to construct even more complex deformances to be perused by the critic. Or perhaps the critic would only sparingly choose to apply these machines against the text – only when uncertain, or a little too disconcertingly certain, learning the new art of contextualizing deformance within some as of yet unimaginable critical practice of the future.

In any case, computational methods not only hold the potential to aid the critic but also emerge as viable tools for testing hypotheses about the human reading mind itself. And the more we know, the stranger it feels. The plurality of imaginable and lived-in worlds becomes almost tangible via the mediation of computationally constructed cognitive models. Reading machines can adumbrate the contours of unexpected material aspects of the fictional world; they act as automatic archaeologists of potential meaning, buried in the historically conditioned ambiguity of language. And even though a critic may need those tools to detect such dormant worlds, it inevitably remains his or her own task to collate the data into a coherent interpretation. This hermeneutic process can only become more revealing when drawing intelligently upon a richer ensemble of oracles, be they human or machinic. As Feyerabend writes: "we need a dreamworld in order to discover the features of the real world… which may actually be just another dream world."[48] If we can construct, even awkwardly and without great precision at first, these dreamworlds inhabited by minds other than our habitual selves, then their powers of deformance may serve us well. But perhaps even more inspiring is the potential of literature to be a guide *within* the dreamworlds themselves: "a line of [fiction] operates essentially like a search query that selects items from

[48] Paul Feyerabend, *Against Method* (London: Verso Books, 2010[1975]), 15. Quoted in: Stephen Ramsay, "Reading Machines," 22.

a database."[49] Perhaps this is a true sign of our cyborgian future – that not only can machines help us better understand literature, but that literature can help us better understand machines.

[49] Michael Gavin, "Vector Semantics," 668. The original phrase is "a line of poetry."

Bibliography

Barad, Karen. *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Durham: Duke University Press, 2007.

Broderick, Damien. *Reading By Starlight: Postmodern Science Fiction*. London: Routledge, 1995.

Delany, Samuel R. *The American Shore: Meditations on a Tale of Science Fiction by Thomas M. Disch—"Angouleme."* Middleton: Wesleyan University Press, 2014 [1978].

Delany, Samuel R. *The Jewel-Hinged Jaw: Notes on the Language of Science Fiction*. Middleton: Wesleyan University Press, 2009 [1978].

Dick, Philip K. *Do Androids Dream of Electric Sheep?* London: Victor Gollancz, 1999.

Feyerabend, Paul. *Against Method*. London: Verso Books, 2010 [1975].

Firth, John R. "A Synopsis of Linguistic Theory 1930–1955." *Studies in Linguistic Analysis* (1957), 1-32.

Freedman, Carl. *Critical Theory and Science Fiction*. Middleton: Wesleyan University Press, 2000.

Gavin, Michael. "Vector Semantics, William Empson, and the Study of Ambiguity." *Critical Inquiry*, 44.4 (2018), 641-673.

Haraway, Donna J. *Staying With the Trouble: Making Kin in the Chthulucene*. Durham: Duke University Press, 2016.

Harris, Zellig. "Distributional Structure." *The Structure of Language: Readings in the Philosophy of Language*, edited by Jerry A. Fodor and Jerrold J. Katz. 33-49. Englewood Cliffs, N.J., 1964.

Mikolov, Tomas, Kai Chen, Greg Corrado and Jeffrey Dean. *Efficient Estimation of Word Representations in Vector Space*. arXiv preprint arXiv:1301.3781, 2013.

Pennington, Jeffrey, Richard Socher and Christopher Manning. "Glove: Global Vectors for Word Representation". *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (2014), 1532-1543.

Ramsay, Stephen. *Reading Machines: Toward an Algorithmic Criticism*. Champaign: University of Illinois Press, 2011.

Rieder, John. *Science Fiction and the Mass Cultural Genre System*. Middleton: Wesleyan University Press, 2017.

Roberts, Adam. *The History of Science Fiction*. Basingstoke: Palgrave Macmillan, 2005.

Suvin, Darko. *Metamorphoses of Science Fiction: On the Poetics and History of a Literary Genre*. Bern: Peter Lang, 2016[1979].

Widdows, Dominic. *Geometry and Meaning*. Stanford: CSLI Publications, 2004.