Spectral Composition of Semantic Spaces

Peter Wittek and Sándor Darányi

Swedish School of Library and Information Science
Göteborg University & University of Borås
Allégatan 1, 50190 Borås, Sweden

Abstract. Spectral theory in mathematics is key to the success of as diverse application domains as quantum mechanics and latent semantic indexing, both relying on eigenvalue decomposition for the localization of their respective entities in observation space. This points at some implicit "energy" inherent in semantics and in need of quantification. We show how the structure of atomic emission spectra, and meaning in concept space, go back to the same compositional principle, plus propose a tentative solution for the computation of term, document and collection "energy" content.

1 Introduction

In quantum mechanics (QM), the spectrum is the set of possible outcomes when one measures the total energy of a system. Solutions to the time-independent Schrödinger wave equation are used to calculate the energy levels and other properties of particles. A non-zero solution of the wave equation is called an eigenenergy state, or simply an eigenstate. The set of eigenvalues $\{E_j\}$ is called the energy spectrum of the particle. This energy spectrum can be mapped to frequencies in the electromagnetic spectrum.

In this paper, we argue that by decomposing a semantic space, one can gain a "semantic spectrum" for each term that makes up the space. This makes sense for the following reason: mapping spectra to the electromagnetic spectrum is a unification effort to match energy and intellectual input stored in documents by modelling semantics on QM. Energy is a metaphor here, lent from machine learning which imitates pattern recognition and pattern naming in cognitive space. We adopted this as our working hypothesis based on [1].

To this end, we ascribe significance to two aspects of the above parallel. Both make the comparison between semantics and QM reasonable. The first is an alleged similarity between them, namely eigendecomposition and related methods leading to meaningful conclusions in both. The

second is the evolving nature of QM and semantic systems, based on interactions among constituents, leading to structuration. The insights we offer in this paper do not rely on extensive quantitative benchmarks. Instead, the paper reports our initial foray into exploring the above metaphor.

This paper is organized as follows. Section 2 discusses core concepts in QM relevant to this treatise. Section 3 gives an overview of semantic spaces in general and Section 4 describes their spectral composition in particular, including their treatment as observables, corpus and term semantic spectra, and indications for future work such as evolving semantics. Section 5 sums up the conclusions.

2 Related concepts in quantum mechanics and spectroscopy

In quantum mechanics, observables are not necessarily bounded, self-adjoint operators and their spectra are the possible outcomes of measurements. The Schrödinger wave equation is an equation that describes how the quantum state of a physical system changes over time. Approximate solutions to the time-independent Schrödinger wave equation are commonly used to calculate the energy levels and other properties of atoms and molecules. From this, the emission spectrum is easy to calculate.

Fig. 1. The emission spectrum of hydrogen

Emission is the process by which two quantum mechanical states of a particle become coupled to each other through a photon, resulting in the production of light. The frequency of light emitted is a function of how far away in energy the two states of the system were from each other, so that energy is conserved: the energy difference between the two states equals the energy carried off by the photon (Figure 1).

Since the emission spectrum is different for every element of the periodic table, it can be used to determine the composition of a material. In general, spectroscopy is the study of the interaction between matter and radiated energy. A subset of spectroscopic methods, called spectrophotometry, deals with visible light, near-ultraviolet, and near-infrared

wavelengths. For the rest of this paper, we limit ourselves to visible spectroscopy, because this approach focuses on the electronic orbitals (i.e., where the electrons can be found), whereas, for instance, infra-red spectroscopy is concerned with the internal motions of the molecule (how the bonds stretch, angles bend, etc.).

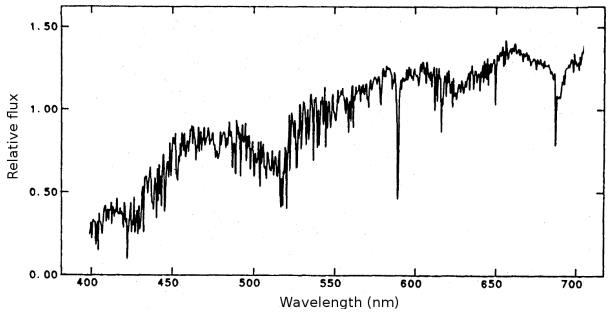


Fig. 2. The visible spectrogram of the red dwarf EQ Vir (figure adapted from [2])

A spectrogram is a spectral representation of an electromagnetic signal that shows the spectral density of the signal. An example is astronomical spectroscopy that studies the radiation from stars and other celestial objects (Figure 2). While discrete emission bands do not show clearly, the intensity of certain wavelengths indicates the composition of the observed object. The emission lines are caused by a transition between quantized energy states and theoretically they look very sharp, they do have a finite width, i.e. they are composed of more than one wavelength of light. This spectral line broadening has many different causes, with the continuum of energy levels called "spectral bands". The bands may overlap. Band spectra are the combinations of many different spectral lines, resulting from rotational, vibrational and electronic transitions.

3 A brief overview of semantic spaces

We regard semantic spaces as algebraic models for representing terms as vectors. The models capture term semantics by a range of mathematical relations and operations. Language technology makes extensive use of semantic spaces. Among the reasons are the following:

– The semantic space methodology makes semantics computable allowing a definition of semantic similarity in mathematical terms. Sparsity

plays a key role in most semantic spaces. A term-document vector space (see below), for instance, is extremely sparse and therefore it is a feasible option for large-scale collections.
- Semantic space models also constitute an entirely descriptive approach to semantic modelling relying on the distributional hypothesis. Previous linguistic or semantic knowledge is not required.
- The geometric metaphor of meaning inherent in a vector space kind of model is intuitively plausible, and is consistent with empirical results from psychological studies. This relates especially to latent semantic indexing (see below) [3]. A link has also been established to Cognitive Science [4].

While there are several semantic space models, we restrict our discussion to the following two major kinds: term-document vector spaces [5] and latent semantic indexing (LSI, [6]); and the hyperspace analogue to language (HAL, [7]).

The coordinates in the vector of a term in a term-document space record the number of occurrences of the term in the document assigned to that particular dimension. Instead of plain term frequencies, more subtle weighting schemes can be applied, depending on the purpose. The result is an $m \times n$ matrix A, where m is the number of terms, and n is the number of documents. This matrix is extremely sparse, with only $1-5\%$ of the entries being non-zero. This helps scalability, but has an adverse impact on modelling semantics. For instance, in measuring similarity with a cosine function between the term vectors, we often end up with a value of zero, because the vectors do not co-occur in any of the documents of the collection, although they are otherwise related. To overcome this problem, LSI applies dimension reduction by singular value decomposition (SVD). The term-document matrix A can be decomposed as $A = U\Sigma V^{\mathrm{T}}$, where U is an $m \times m$ unitary matrix, Σ is an $m \times n$ diagonal matrix with nonnegative real numbers, the singular values, on the diagonal, and V is an $n \times n$ unitary matrix. By truncating the diagonal of Σ, keeping only the k largest singular values, we get the rank-k approximation of A, $A_k = U_k \Sigma_k V_k^{\mathrm{T}}$. This new space, while not sparse, reflects semantic relations better [3]. Apart from LSI, a term co-occurrence matrix is another alternative to overcome the problem of sparsity. It is obtained by multiplying A with its own transpose, A^{T}.

The HAL model considers context only as the terms that immediately surround a given term. HAL computes an $m \times m$ matrix H, where m is the number of terms, using a fixed-width context window that moves incrementally through a corpus of text by one word increment ignoring

punctuation, sentence and paragraph boundaries. All terms within the window are considered as co-occurring with the last word in the window with a strength inversely proportional to the distance between the words. Each row i in the matrix represents accumulated weights of term i with respect to other terms which preceded i in a context window. Similarly, column i represents accumulated weights with terms that appeared after i in a window. Dimension reduction may also be performed on this matrix.

We note in passing that there exists a little recognized constraint of the model in testing: for a match between theories of word semantics and semantic spaces, a semantic space is a statistical model of word meaning observed [8]. For its workings, it has to match a reasonably complex theory of semantics; but whereas Lyons regarded meaning a composite [9], i.e. a many-faceted complex phenomenon, the distributional hypothesis [10] as the sole semantic underpinning of eigenmodels is anything but complex and must be hence deficient. One can use it as long as there is nothing else available but, at the same time, one must not stop looking for a more comprehensive model. It holds in this sense that we look at the validity and some consequences of the semantic collapse model based on quantum collapse, treating semantic deep structure as an eigenvalue spectrum.

4 Spectral composition of semantic spaces

4.1 Semantic spaces as observables

Our line of thought is as follows: in QM, atoms have ground states low on energy, and excited states high on it. Such states are expressed as separate spectral (latent) structures, based on the way they can be identified. By analogy a term should have a "ground state" and may have several "excited states" as well, all in terms of spectra.

In what follows, we regard a semantic space an observable. This being a real or a complex space, its spectrum will be the set of eigenvalues. If we decompose a semantic space we get the so-called concept space or topic model in which terms map to different locations due to their different composition. We identify this latent topic mixture in LSI with the energy eigenstructure in QM. This means that more prevalent hidden topics correspond to higher energy states of atoms and molecules.

Identifying "excited states" of word forms with homonyms, and word sense disambiguation with observation, the above shows resemblance with the quantum collapse of meaning described by [8]. They argue that a sense can be represented as a density matrix which is quite easily derived from summing the HAL matrices of the associated contexts. In addition, a

probability can be ascribed the to a given sense. For example, the density matrix ρ for the meaning of a word can be formalized at the following linear combination:$\rho = p_1\rho_1 + \ldots + p_m\rho_m$, where each i is a basis state representing one of the m senses of the term and the probabilities p_i sum to unity. This is fully in accord with QM whereby a density matrix can be expressed as a weighted combination of density matrices corresponding to basis states. Context is modelled as a projection operator which is applied to a given density matrix corresponding to the state of a word meaning resulting in its 'collapse'. The probability of collapse p is a function of the scalar quantity resulting from matching. The analogy with orthodox QM is the following - a projection operator models a measurement on a quantum particle resulting in a collapse onto a basis state. Spectral decomposition by SVD also allows the description of a word as the sum of eigenstates using the bra-ket terminology [11]. The formal description is similar to the above. Projection operators are defined by singular vectors. These are orthogonal.

The semantic space must be Hermitian to pursue the metaphor of an observable in a quantum system. The sum of a HAL space H and its transpose is a Hermitian matrix [11]. A different approach is to pad the corresponding matrix of a term-document space A with zeros to make an operator map a Hilbert space onto itself, and then use a product with its own transpose as the Hermitian operator [12]. For the rest of the paper, we adopt a similar approach, taking the term co-occurrence matrix AA^{T}, which is a Hermitian operator. For symmetric and Hermitian matrices, the eigenvalues and singular values are obviously closely related. A nonnegative eigenvalue, $\lambda \geq 0$, is also a singular value, $\sigma = \lambda$. The corresponding vectors are equal to each other, $u = v = x$. A negative eigenvalue, $\lambda < 0$, must reverse its sign to become a singular value, $\sigma = |\lambda|$. One of the corresponding singular vectors is the negative of the other, $u = -v = x$. Hence a singular value decomposition and an eigendecomposition coincide.

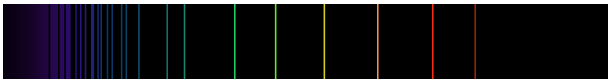
4.2 Semantic spectrum

In a metaphoric sense, words in an eigendecomposition are similar to chemical compounds: as both are composed of doses of latent constituents, the dosimetric view applies to them. The field that studies substances of unknown composition is called spectrometry. By analyzing their spectra, components of such substances can be identified because each chemical component has a unique "fingerprint".

The case of a semantic spectrum is not unlike. We performed the eigendecomposition of the term co-occurrence matrix of the Reuters-

21578 collection. There are many other methods to capture the latent constituents of terms, for instance random indexing [13], latent Dirichlet allocation [14], or spherical k-means [15]. It is an open question which method captures the latent structure best. We use eigendecomposition due to its similarity to spectrometry. The term co-occurrence matrix is a Hermitian operator, hence the eigenvalues are all real-valued. Since the term co-occurrence matrix does not have an underlying physical meaning, we mapped the eigenvalues to the visible spectrum. If 400nm is the lowest visible wavelength and 700nm is the highest, then, assuming that the lowest eigenvalue is approximately zero, and λ_{max} denotes the highest eigenvalue, the mapping is performed by $F(x) = 400 + x\frac{700-400}{\lambda_{max}}$. The resulting spectrum is plotted in Figure 3(a). By this mapping one obtains a visual snapshot of an unknown topic composition.

In other words, by this metaphor we regarded the semantic spectrum of the above test collection as a composite, a sum of spectra of elementary components, which would correspond to individual elements in a chemical compound in spectrophotometry. This representation stresses the similarity of chemical composition of elements to the semantic composition of words.



(a) The spectrum of the Reuters collection

(b) The spectrum of the term *Japan*

(c) The spectrum of the term *courage*

(d) The spectrum of the term *male*

Fig. 3. The spectrum of the collection and of different words. Higher energy states correspond to the right end of the spectrum.

We propose matching spectral components to terms based on their proximity to latent variables. This creates individual, albeit overlapping,

spectra for every term. Having used a 0.05 threshold value of the cosine dissimilarity measure between term vectors and eigenvectors, if the cosine was above this value, we added the corresponding scaled eigenvalue to the term's spectrum. In this regard, term spectra may overlap, and their simple sum will provide the spectrum of the collection. This metaphor does not account for more complex chemical bonds that create the continuous bands as pictured in Figure 2.

By such experimentation, one can end up with interesting interpretation problems. For instance, the term *Japan* (Figure 3(b)) has a high wavelength component, and a number of low wavelengths. This means that by the formula $E_{\text{photon}} = h\nu$, where h is Planck's constant and ν is the frequency (the inverse of wavelength multiplied by the speed of light), the term has one low-energy state which it is likely to take, and a number of other, high-energy states which it takes given an appropriate context. In its low-energy states the term is likely to refer to the country itself, whereas the less frequently encountered contexts may activate one of the four nominal and one verbal senses listed in WordNet. In other words, the term was correctly treated as a homonym by considering its senses as atoms in a molecule.

Another example, the term *courage* does not have a true low-energy state, it takes only higher-energy configurations. Here our tentative suggestion is that eigendecompositon does not distinguish between molecular or atomic electron orbits, hence future research may indicate that such high energy states are typical for terms treated as atoms (Figure 3(c)).

The term *male* can take two fairly low-energy states, but very few higher ones (Figure 3(d)). Since this word has three nominal and three verbal senses in WordNet, it is a reasonable working hypothesis to say that the term was treated as a molecule with six states. We trust that by more experimentation, we will gain better insight into the art of semantic spectrogram interpretation.

4.3 Evolving semantics and considerations for future work

A related aspect of our approach is the quest to formalize corpus dynamics, in line with the recommendations spelled out by [16], also keeping the possible differences between language and quantum interaction systems in mind. We depart from the assumption that two types of dynamics characterize any text document collection: external forces leading to its expansion, and the inherent quality in terms and their agglomerates called their meaning. We offer two observations why this inherent quality may have something to do with the concept of energy (a.k.a. work content):

– Interestingly, spectral theory in mathematics has been key to the success of as diverse application domains as QM and LSI. In other words, both the Schrodinger equation and LSI rely on eigenvalue decomposition for the localization of their respective entities in observation space. This points at some implicit "energy" inherent in semantics and in need of quantification. Another indication of the "energetic" nature of word meaning comes from dynamic semantics where it is regarded as an agent or promoter of change [17, 18]. However, contextual and referential theories of word meaning [10, 19] currently used in applications trying to capture and exploit semantic content focus on the *quantities* of qualities only, and may therefore miss part of the underlying framework;
– The phenomenon of language change and its modelling [20] necessitates a coherent explanation of the dynamics of evolving collections. In line with the above, since any matrix has an eigendecomposition and therefore a latent structure, evolving vector spaces of terms and documents follow directly from variable matrix spectra. However, this has implications for modelling semantics on QM, plus offers an illustration to the problem of assigning an "energetic" nature to word meaning. Namely, whereas Salton's dynamic library model [21], except for mass, already embodied all the key concepts of Newtonian mechanics, it is exactly this missing element which prevents one from constructing time-dependent term and document potential fields, and hence evolving "energy" landscapes. Also, without assuming that terms and documents have specific "masses" and corresponding "energies", it is very difficult to explain how intellectual work can be stored in documents and collections. In other words, unless one comes up with a better solution to the problem of how thinking amounts to work, one must assume that work as the line integral of force needs a language model which utilizes the concepts of distance, velocity, acceleration, mass, force and potential.

The implication is that if we want to be coherent, applying QM for a better understanding of meaning begs for the concept of a term-specific mass. However, such specific values cannot be extracted from an evolving environment, therefore they must reside somewhere else, e.g. in a stable environs such as an ontology, from where they can "charge" entities as their forms with content. This would amount to a challenge to the current view on semantic spaces which strives to explain the presence of all the meaning in vector spaces by term context only, and would resemble a referential model of word semantics instead. A series of semantic spectro-

grams, i.e. snapshots taken of collection content over time could display this evolving latent "energy" structure, and illustrate our point. In such an environment, term "energies" cannot be either constant or specific though, a contradiction to be explored.

In QM, it is the Hamiltonian which typically describes the energy stored in a system. With the above caveat, it is evident that in order to experiment with the dynamic aspect of meaning, one needs to take a look at the Hamiltonian of a collection. Further because in the above experiment, we identified the superposition of term states in the absence of an observer with that of homonyms in need of disambiguation, the same word form with different senses invites the parallel of molecular orbitals, and hence the use of the molecular Hamiltonian. This is the equation representing the energy of the electrons and nuclei in a molecule, a Hermitian operator which, together with its associated Schrödinger equation, plays a central role in computational chemistry and physics for computing properties of molecules and their aggregates.

At the same time it is necessary to point out that, whereas the demonstrated applicability of QM to semantic spaces implies the presence of some force such as lexical attraction [22] or anticipated term mass [23], because of the "energetic" explanation we can calculate with two kinds of attraction between terms only, i.e. one caused by polarity and leading to the Coulomb potential, the other caused by mass and leading to gravitational potential. But whereas there is hope that some aspect of vocabularies can be associated in the future with the role mass plays in physics, we do not know of any attempts to explain vector spaces in terms of polarity such as negative and positive electric charges unless one considers absence and presence in a binary matrix as such. However, then some kind of existential polarity is modelled by the wrong numerical kit, but nevertheless, as the results prove, the metaphor works: the expression could be constructed. Meanwhile, semantics modelled on QM also works, but we do not know why, as according to our current understanding, with this many ill fits between physics and language, it should not. These contradictions call for continued research.

5 Conclusions

Apart from semantic spectrograms bringing closer the idea of mathematical energy, a frequent concept in machine learning and structured prediction [1], our approach has the following attractive implications with their own research potential:

– Studying and eventually composing semantic functions from matrix spectra is a new knowledge area where the mathematical objects used, i.e. functions, have a higher representation capacity than vectors. This surplus can be used for the encoding of different aspects of word and sentence semantics not available by vector representation, and in general opens up new possibilities for knowledge representation;
– This form of semantic content representation provides new opportunities for optical computing, including computation by colours [24];
– Connecting QM and language by the concept of energy, represented in the visual spectrum, has a certain flair which goes beyond the paedagogical usefulness of the metaphor. Namely, considering semantics as a kind of energy and expressing it explicitly as such brings the very idea of intellectual work stored in documents one step closer to measurable reality, of course with all the foreseeable complications such an endeavour might entail.

6 Acknowledgement

This work was partially funded by Amazon Web Services and the large-scale integrating project Sustaining Heritage Access through Multivalent ArchiviNg (SHAMAN) which is co-funded by the European Union (Grant Agreement No. ICT-216736).

References

1. LeCun, Y., Chopra, S., Hadsell, R.: A tutorial on energy-based learning. In: Predicting Structured Data. (2006) 1–59
2. Pettersen, B., Hawley, S.: A spectroscopic survey of red dwarf flare stars. Astronomy and Astrophysics **217** (1989) 187–200
3. Landauer, T., Dumais, S.: A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review **104**(2) (1997) 211–240
4. Gärdenfors, P.: Conceptual spaces: The geometry of thought. The MIT Press (2000)
5. Salton, G., Wong, A., Yang, C.: A vector space model for information retrieval. Journal of the American Society for Information Science **18**(11) (1975) 613–620
6. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science **41**(6) (1990) 391–407
7. Lund, K., Burgess, C.: Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods Instruments and Computers **28** (1996) 203–208
8. Bruza, P., Woods, J.: Quantum collapse in semantic space: interpreting natural language argumentation. In: Proceedings of QI-08, 2nd International Symposium on Quantum Interaction, Oxford, UK (March 2008)

9. Lyons, J.: Introduction to theoretical linguistics. Cambridge University Press, New York, NY, USA (1968)
10. Harris, Z.: Distributional structure. In Harris, Z., ed.: Papers in structural and transformational Linguistics. Formal Linguistics. Humanities Press, New York, NY, USA (1970) 775–794
11. Bruza, P., Cole, R.: Quantum logic of semantic space: An exploratory investigation of context effects in practical reasoning. In Artemov, S., Barringer, H., d'Avila Garcez, A.S., Lamb, L., Woods, J., eds.: We Will Show Them: Essays in Honour of Dov Gabbay. College Publications (2005)
12. Aerts, D., Czachor, M.: Quantum aspects of semantic analysis and symbolic artificial intelligence. Journal of Physics A: Mathematical and General **37** (2004) L123–L132
13. Kanerva, P., Kristofersson, J., Holst, A.: Random indexing of text samples for latent semantic analysis. In: Proceedings of CogSci-00, 22nd Annual Conference of the Cognitive Science Society. Volume 1036., Philadelphia, PA, USA (2000)
14. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. The Journal of Machine Learning Research **3** (2003) 993–1022
15. Dhillon, I., Modha, D.: Concept decompositions for large sparse text data using clustering. Machine learning **42**(1) (2001) 143–175
16. Kitto, K., Bruza, P., Sitbon, L.: Generalising unitary time evolution. In: Proceedings of QI-09, 3rd International Symposium on Quantum Interaction, Saarbruecken, Germany (March 2009) 17–28
17. Beaver, D.: Presupposition and assertion in dynamic semantics. CSLI publications (2001)
18. van Eijck, J., Visser, A.: Dynamic semantics. In Zalta, E.N., ed.: The Stanford Encyclopedia of Philosophy. (2010)
19. Frege, G.: Sense and reference. The Philosophical Review **57**(3) (1948) 209–230
20. Baker, A.: Computational approaches to the study of language change. Language and Linguistics Compass **2**(3) (2008) 289–307
21. Salton, G.: Dynamic information and library processing. (1975)
22. Beeferman, D., Berger, A., Lafferty, J.: A model of lexical attraction and repulsion. In: Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics, Madrid, Spain (July 1997) 373–380
23. Shi, S., Wen, J., Yu, Q., Song, R., Ma, W.: Gravitation-based model for information retrieval. In: Proceedings of SIGIR-05, 28th International Conference on Research and Development in Information Retrieval, Salvador, Brazil (August 2005) 488–495
24. Dorrer, C., Londero, P., Anderson, M., Wallentowitz, S., Walmsley, I.: Computing with interference: all-optical single-query 50-element database search. In: Proceedings of QELS-01, Quantum Electronics and Laser Science Conference. (2001) 149–150