Postprint of: Aerts D., Czachor M., Quantum aspects of semantic analysis and symbolic artificial intelligence, Journal of Physics A: Mathematical and General, Vol. 37, Iss. 12 (2004), L123, DOI: 10.1088/0305-4470/37/12/L01

Quantum Aspects of Semantic Analysis and Symbolic Artificial Intelligence

Diederik Aerts [1] and Marek Czachor [2]
[1] Centrum Leo Apostel (CLEA) and Foundations of the Exact Sciences (FUND)
Vrije Universiteit Brussel, 1050 Brussels, Belgium
[2] Katedra Fizyki Teoretycznej i Metod Matematycznych
Politechnika Gdańska, 80-952 Gdańsk, Poland

Modern approaches to semanic analysis if reformulated as Hilbert-space problems reveal formal structures known from quantum mechanics. Similar situation is found in distributed representations of cognitive structures developed for the purposes of neural networks. We take a closer look at similarites and differences between the above two fields and quantum information theory.

PACS numbers: 03.67.-a, 89.70.+c, 11.30.Pb

I. PROLOGUE

Let us consider an arbitrary text written by means of a 16-letter alphabet, say: a, b, c, . . . , n, o, p. Let us regroup as large part of the text as possible in quadruples belonging to the set $Q = \{\text{aeim, afim, agim}, \dots, \text{dhlm, dhln, dhlo, dhlp}\}$, and formed by strings obtained by picking out a single letter from a row of the matrix

$$\begin{bmatrix} \text{a} & \text{b} & \text{c} & \text{d} \\ \text{e} & \text{f} & \text{g} & \text{h} \\ \text{i} & \text{j} & \text{k} & \text{l} \\ \text{m} & \text{n} & \text{o} & \text{p} \end{bmatrix} \qquad (1)$$

when one moves downwards starting from the first row. Now let us define the functions F and G by $F(\text{a}) = F(\text{d}) = F(\text{e}) = F(\text{h}) = F(\text{i}) = F(\text{l}) = F(\text{m}) = F(\text{p}) = +1$, $F(\text{b}) = F(\text{c}) = F(\text{f}) = F(\text{g}) = F(\text{j}) = F(\text{k}) = F(\text{n}) = F(\text{o}) = -1$, $G(x_1x_2x_3x_4) = F(x_1) + F(x_2) + F(x_3) - F(x_4)$. On each four-character string of the regrouped part of the text we evaluate the value of G and take its average value $\langle G\rangle$.

The above awkward-looking manipulation with the text is an example of a procedure one might find in a paper on quantitative linguistics or semantic analysis. The analysis reveals certain correlational or contextual aspects of the text, the role of the contextuality measure being played by the average $\langle G\rangle$.

To see what kind of a correlation one can capture, let us parametrize the alphabet by primed and unprimed bits $0, 1, 0', 1'$:

a = (00), b = (01), c = (10), d = (11),
e = (00′), f = (01′), g = (10′), h = (11′),
i = (0′0), j = (0′1), k = (1′0), l = (1′1),
m = (0′0′), n = (0′1′), o = (1′0′), p = (1′1′).

After the reparametrization the regrouped text might represent data of an experiment testing the Bell inequality [1] and the function F represents values of the Bell observable for a single pair of measurements. And vice versa, any result of an experiment that tests the Bell inequality can be represented as a text written in a 16-letter alphabet.

The result of the form $|\langle G\rangle| > 2$ reveals a nonclassical probabilistic structure behind the text. This structure is, of course, typical of the *source* of the text, since the text itself may be a simple collection of characters on a computer printout. Actually, we can immediately identify the nonclassical elements disclosed by $|\langle G\rangle| > 2$: The bits 0 and $0'$ (or 1 and $1'$) correspond to *nonorthogonal* vectors, and ordered pairs such as (01) are represented by tensor products. The possibility of hiding information behind nonorthogonal bases is the key idea of quantum cryptography [2, 3] and tensor representations of conjuctions are fundamental to quantum information theory (QIT). The observation of Bell that correlations between symbols in "texts" may reveal the presence of nonorthogonal bases is perhaps the most ingenious ingredient of his famous paper [1].

The idea that some sort of mathematical manipulation with texts, or some apparently artificial mathematical representation of them, may reveal deep structures such as similarity of meaning or other nontrivial correlations, is at the roots of semantic analysis (SA). Still another field where analogies with the Bell inequality example are particularly striking is related to neural-network distributed representations of concepts [4]. The links of such scientific disciplines with quantum mechanics, and QIT in particular, are almost unexplored as yet. The present paper is an attempt of filling up the gap [5].

II. VECTOR MODELS OF TEXTS

Modern approaches to SA typically model words and their meanings by vectors from finite-dimensional vector spaces. The prominent examples of such approaches are Latent Semantic Analysis (LSA) [6, 7], Hyperspace Analogue to Language (HAL) [8], Probabilistic Latent Semantic Analysis (pLSA) [9], Latent Dirichlet Allocation [10], Topic Model [11], or Word Association Space (WAS) [12]. In the present Letter we concentrate on a simplified version of LSA, but we believe the discussion we present can be applied to all vector models of language and concept representation.

SA is typically based on text co-occurence matrices and data-analysis technique employing singular value decomposition (SVD). Various models of SA provide powerful methods of determining similarity of meaning of words and passages by analysis of large text corpora. The procedures are fully automatic and allow to analyze texts by computers without an involvment of any human understanding. For example, what makes LSA quite impressive comes from the experiments with simulation of human performance. LSA-programmed machines were able to pass multiple-choice exams such as Test of English as a Foreign Language (TOEFL) (after training on general English) [13] or, after learning from an introductory psychology textbook, a final exam for psychology students [7].

These and other achievements of LSA raise the question of its relevance for the problem of brain functioning and AI [14]. However, an element we found particularly intriguing and which is the main topic of our paper, is in similarities between LSA and formal structures of QIT.

LSA is essentially a Hilbert space formalism. One represents words by vectors spanning a finite-dimensional space and text passages are represented by linear combinations of such words, with appropriate weights related to frequency of occurence of the words in the text. Similarity of meaning is represented by scalar products between certain word-vectors (beloging to the so-called semantic space).

In QIT, words, also treated as vectors, are being processed by quantum algorithms or encoded/decoded by means of quantum cryptographic protocols. Although one starts to think of quantum programming languages [15, 16, 17], the semantic issues of quantum texts are difficult to formulate. LSA is in this context a natural candidate as a starting point for "quantum linguistics".

Still, LSA has certain conceptual problems of its own. As stressed by many authors, the greatest difficulty of LSA is that it treats a text passage as a "bag of words", a set where order is irrelevant [18]. The difficulty is a serious one since it is intuitively clear that syntax is important for evaluation of text meaning. The sentences "Mary hit John" and "John hit Mary" cannot be distinguished by LSA; "Mary did hit John" and "John did not hit Mary" have practically identical LSA representations because "not" is in LSA a very short vector [14]. What LSA can capture is that the sentences are about violence.

We think that experience from QIT may prove useful here. A basic object in QIT is not a word but a letter. Typically one works with the binary alphabet consisting of 0 and 1 and qubits. Ordering of qubits is obtained by means of the tensor product. Ordering of words can be obtained in the same way, but before we proceed with QIT formalism, let us explain the standard LSA and formulate it in quantum mechanical notation.

III. SEMANTIC ANALYSIS: AN ILLUSTRATION

Let us consider the following passage:

"(s_1) How much wood would a woodchuck chuck if a woodchuck could chuck wood? (s_2) Woodchuck would chuck as much wood as a woodchuck could chuck if a woodchuck could chuck wood. (s_3) Could woodchuck chuck 35 cubic feet of dirt? (s_4) If a woodchuck could chuck wood woodchuck would chuck 700 pounds of wood."

The LSA matrix representation of this text reads

$$\begin{array}{rcccc} & s_1 & s_2 & s_3 & s_4 \\ \text{how} & 1 & 0 & 0 & 0 \\ \text{much} & 1 & 1 & 0 & 0 \\ \text{wood} & 2 & 2 & 0 & 2 \\ \text{would} & 1 & 1 & 0 & 1 \\ \text{a} & 2 & 2 & 0 & 1 \\ \text{woodchuck} & 2 & 3 & 1 & 2 \\ \text{chuck} & 2 & 3 & 1 & 2 \\ \text{if} & 1 & 1 & 0 & 1 \\ \text{could} & 1 & 2 & 1 & 1 \\ \text{35} & 0 & 0 & 1 & 0 \\ \text{cubic} & 0 & 0 & 1 & 0 \\ \text{feet} & 0 & 0 & 1 & 0 \\ \text{of} & 0 & 0 & 1 & 1 \\ \text{dirt} & 0 & 0 & 1 & 0 \\ \text{700} & 0 & 0 & 0 & 1 \\ \text{pounds} & 0 & 0 & 0 & 1 \end{array} \to A_0 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 2 & 2 & 0 & 2 \\ 1 & 1 & 0 & 1 \\ 2 & 2 & 0 & 1 \\ 2 & 3 & 1 & 2 \\ 2 & 3 & 1 & 2 \\ 1 & 1 & 0 & 1 \\ 1 & 2 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

It is usual to pre-process A_0 by multiplying each entry by a function associated with the entropy of an appropriate word evaluated on the basis of an entire text. The question of what kind of a co-occurence matrix should one relate to a text is actually an open one, and is investigated in various alterantives to LSA (HAL, WAS, Topic Model). For simplicity we skip this point.

The text corresponds now to the map $A : \boldsymbol{R}^4 \to \boldsymbol{R}^{16}$, whose SVD (up to numerical roundup errors) is $A_0 =$

$U^\dagger D_0 V$ where

$$U^\dagger = \begin{pmatrix} -0.06 & -0.12 & 0.15 & 0.70 \\ -0.14 & -0.15 & 0.35 & 0.08 \\ -0.40 & -0.22 & -0.26 & 0.23 \\ -0.20 & -0.11 & -0.13 & 0.11 \\ -0.34 & -0.26 & 0.21 & 0.20 \\ -0.50 & 0.11 & 0.04 & -0.20 \\ -0.50 & 0.11 & 0.04 & -0.20 \\ -0.20 & -0.11 & -0.13 & 0.11 \\ -0.30 & 0.23 & 0.17 & -0.32 \\ -0.02 & 0.37 & 0.11 & 0.18 \\ -0.02 & 0.37 & 0.11 & 0.18 \\ -0.02 & 0.37 & 0.11 & 0.18 \\ -0.07 & 0.41 & -0.36 & 0.20 \\ -0.02 & 0.37 & 0.11 & 0.18 \\ -0.05 & 0.04 & -0.48 & 0.02 \\ -0.05 & 0.04 & -0.48 & 0.02 \end{pmatrix}, \quad (2)$$

$$D_0 = \begin{pmatrix} 8.38 & 0 & 0 & 0 \\ 0 & 2.52 & 0 & 0 \\ 0 & 0 & 1.79 & 0 \\ 0 & 0 & 0 & 1.04 \end{pmatrix}, \quad (3)$$

$$V = \begin{pmatrix} -0.52 & -0.67 & -0.17 & -0.48 \\ -0.30 & -0.07 & 0.94 & 0.10 \\ 0.28 & 0.34 & 0.21 & -0.86 \\ 0.73 & -0.64 & 0.18 & 0.02 \end{pmatrix}. \quad (4)$$

The essential step of LSA is the reduction

$$A_0 = U^\dagger D_0 V \mapsto A_1 = U^\dagger D_1 V \quad (5)$$

where $D_1 = PD_0$ and P is a projector commuting with D_0. For example, if

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad (6)$$

then

$$A_1 = \begin{pmatrix} 0.26 & 0.33 & 0.08 & 0.24 \\ 0.61 & 0.78 & 0.19 & 0.56 \\ 1.74 & 2.24 & 0.56 & 1.60 \\ 0.87 & 1.12 & 0.28 & 0.80 \\ 1.48 & 1.90 & 0.48 & 1.36 \\ 2.17 & 2.80 & 0.71 & 2.01 \\ 2.17 & 2.80 & 0.71 & 2.01 \\ 0.87 & 1.12 & 0.28 & 0.80 \\ 1.30 & 1.68 & 0.42 & 1.20 \\ 0.08 & 0.11 & 0.02 & 0.08 \\ 0.08 & 0.11 & 0.02 & 0.08 \\ 0.08 & 0.11 & 0.02 & 0.08 \\ 0.30 & 0.39 & 0.09 & 0.28 \\ 0.08 & 0.11 & 0.02 & 0.08 \\ 0.21 & 0.28 & 0.07 & 0.20 \\ 0.21 & 0.28 & 0.07 & 0.20 \end{pmatrix}. \quad (7)$$

We will not go very deeply into details of how and why a reduced representation, of the type illustrated by A_1, may allow a computer to pass TOEFL not worse than an average non-native speaker who wants to study in the USA, and refer the reader to publications on LSA. For our purposes it is sufficient to know that the rows of A_1 are termed the word-vectors and the space of word-vectors is known as the semantic space. Cosines between two word-vectors (or just their scalar products) are measuring a semantic distance (similarity of meaning) between words within a given set of text corpora represented by A. What is important, SVD can make some entries of A_1 negative and even make some scalar products negative, the latter occuring for antonyms. The coefficients of word-vectors lose, after SVD, the simple link to frequncies of occurences of words.

Of course, the dimensions appearing in real texts investigated by means of LSA are much greater (for example 30473 columns and 60768 rows in the experiment discussed in [13]). Experience shows that the analysis is most efficient if the projector P projects on a subspace of dimension around 300, but what is the meaning of this dimension is yet a subject of speculations [19].

IV. SEMANTIC ANALYSIS IN QUANTUM NOTATION

In our example the matrix $U^\dagger$ is not square but its columns are mutually orthogonal. Taking any 12 orthonormal vectors that are, in addition, orthogonal to the columns of $U^\dagger$ we can replace $U^\dagger$ by a 16×16 unitary matrix $\widetilde{U^\dagger}$ whose first 4 columns coincide with those of $U^\dagger$, and end up with SVD of the form

$$\tilde{A}_k = \Big(A_k, 0\Big) = \tilde{U}^\dagger \begin{pmatrix} D_k & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V & 0 \\ 0 & V^\perp \end{pmatrix} = \tilde{U}^\dagger \tilde{D}_k \tilde{V},$$

$k = 0, 1$, where all the matrices are square and $V^\perp$ is an arbitrary unitary matrix of appropriate dimension. The map $A_k \mapsto \tilde{A}_k$ neither adds nor removes any information from the text; its only objective is to work with text matrices and their SVDs that may be regarded as operators mapping certain Hilbert space $\mathcal{H}$ into itself.

The Hilbert space $\mathcal{H}$ is finite dimensional, but in principle one cannot impose any limitation on the number of words or sentences one wants to take into account. It is therefore natural to treat all the concrete examples as subspaces of an infinite dimensional Hilbert space of all the possible words. Whether sentences or other text units are regarded as collections of words or as new words is a matter of convention. Assume each word of a vocabulary is represented by a basis vector $|n\rangle$, where n is a natural number. The text matrix ($\tilde{A} = \tilde{A}_0$ or $\tilde{A} = \tilde{A}_1$) corresponds to the operator $\hat{A} = \sum_{mn} A_{mn}|m\rangle\langle n|$. The column representing a nth sentence is given by the (unnormalized) vector

$$|s_n\rangle = \hat{A}|n\rangle = \sum_m A_{mn}|m\rangle. \quad (8)$$

For example, the sentence s_2 is in LSA represented by the sentence-vector

$$|s_2\rangle \;=\; |2\rangle + |4\rangle + |8\rangle + 2\Big(|3\rangle + |5\rangle + |9\rangle\Big) + 3\Big(|6\rangle + |7\rangle\Big).$$

After SVD the coefficients of a sentence-vector are typically neither natural nor positive. Let us note that $|s_2\rangle$ is not a word-vector in the sense of LSA, but a sentence-vector: Word-vectors are the *rows* of the text matrix. The rows are obtained from $\hat{A}$ by $\langle w_m| = \langle m|\hat{A}$. The similarity of meaning of, say, "how" and "much" is given by $\cos(\text{how}, \text{much}) = \langle w_1|w_2\rangle / \big(\parallel w_1 \parallel \cdot \parallel w_2 \parallel\big)$. (Recall that LSA gives optimal characterization of meaning if one calculates the scalar product after the reduction $D_0 \mapsto D_1 = PD_0$ with appropriately chosen P; in the example, before reduction $\cos(\text{how}, \text{much}) = 0.707107$ and after the reduction $\cos(\text{how}, \text{much}) = 0.999985$).

Putting this differently, the word-vectors characteristic of a text represented by the operator $\hat{A}$ are given by $|w_m\rangle = A^\dagger|m\rangle$. The matrix representing similarities of meaning between all the possible pairs of words corresponding to the text $\hat{A}$ is thus given by

$$\cos(m\text{th word}, n\text{th word}) = \frac{\langle m|\hat{A}\hat{A}^\dagger|n\rangle}{\sqrt{\langle m|\hat{A}\hat{A}^\dagger|m\rangle}\sqrt{\langle n|\hat{A}\hat{A}^\dagger|n\rangle}}.$$

As we can see, the entire information about mutual relations between words is in LSA encoded in the operator $\rho = \hat{A}\hat{A}^\dagger$. Taking into account (8) and the resolution of unity $\mathbf{1} = \sum_n |n\rangle\langle n|$ we can write

$$\rho = \hat{A}\sum_n |n\rangle\langle n|\hat{A}^\dagger = \sum_n |s_n\rangle\langle s_n| = \sum_n p_n^s|\sigma_n\rangle\langle\sigma_n|, \quad (9)$$

with $p_n^s = \langle s_n|s_n\rangle$ and $\langle\sigma_n|\sigma_n\rangle = 1$. Since in any practical application the number of words is finite, the sum in (9) is finite as well and $\operatorname{Tr}\rho = \parallel A \parallel_{\text{HS}}^2 = \sum_n \lambda_n < \infty$, where λ_n are eigenvalues of $N = \hat{A}^\dagger\hat{A}$, and $\parallel\cdot\parallel_{\text{HS}}$ is the Hilbert-Schmidt norm. For this reason ρ is formally an unnormalized density matrix of the set of sentences.

The operator N plays an essential role in LSA. To see this let us look at the explicit proof of SVD formulated in the quantum notation (physicists will recognize here the so-called Schmidt decomposition). Let $|\lambda_n\rangle$ be a normalized eigenvector of N, i.e. $N|\lambda_n\rangle = \lambda_n|\lambda_n\rangle$. Denoting $|\alpha_n\rangle = \hat{A}|\lambda_n\rangle$ we compute

$$\begin{aligned} \hat{A} &= \sum_{|\alpha_n\rangle\neq 0} |\alpha_n\rangle\langle\lambda_n| \\ &= \sum_{|\alpha_n\rangle\neq 0} \frac{|\alpha_n\rangle}{\parallel\alpha_n\parallel}\sqrt{\lambda_n}\langle\lambda_n| \\ &= \underbrace{\sum_k |\beta_k\rangle\langle k|}_{\tilde{U}^\dagger}\underbrace{\sum_l \sqrt{\lambda_l}|l\rangle\langle l|}_{\tilde{D}}\underbrace{\sum_m |m\rangle\langle\lambda_m|}_{\tilde{V}} \quad (10) \end{aligned}$$

where $|\beta_k\rangle = |\alpha_k\rangle / \parallel \alpha_k \parallel$ if $\lambda_k > 0$, or any other basis vector from the subspace corresponding to $\lambda_k = 0$, if $\lambda_k = 0$. It is clear that the singular values in SVD are given by $\sqrt{\lambda_k}$. The LSA procedure is essentially equivalent to the spectral analysis of N.

Let us finally note that N can be written as

$$N = \hat{A}^\dagger\sum_n |n\rangle\langle n|\hat{A} = \sum_n |w_n\rangle\langle w_n| = \sum_n p_n^w|\omega_n\rangle\langle\omega_n|,$$

with $p_n^w = \langle w_n|w_n\rangle$ and $\langle\omega_n|\omega_n\rangle = 1$, i.e. as an unnormalized density matrix representing a mixture of word-vectors.

V. SUPERSYMMETRY AND DIMENSIONAL REDUCTIONS

The duality between sentence-vectors and word-vectors whose one of the manifestations is the link $\hat{A}\hat{A}^\dagger \leftrightarrow \hat{A}^\dagger\hat{A}$ is well known from supersymmetric theories [20]. In supersymmetric terminology operators $\hat{A}\hat{A}^\dagger$ and $\hat{A}^\dagger\hat{A}$ are known as superpartners.

The dimensional reduction employed in LSA is performed on the spectrum of N. Since one eliminates in this way small eigenvalues, the procedure is analogous to some sort of purification of word-vector density matrices. But we know that one of the standard results of supersymmetric quantum mechanics states that N and ρ are isospectral. The interchange of N and ρ is equivalent to replacing word-vectors by sentence-vectors. Dimensional reduction can be thus performed for both N and ρ, in the latter case the reduction deals with sentence-vector density matrices. Finally, one can combine the two approaches. A "supersymmetric LSA" can be based on supercharges $Q = \begin{pmatrix} 0 & A \\ A^\dagger & 0 \end{pmatrix}$ and the two density matrices taken simultaneously in $H = Q^2 = \rho \oplus N$.

In addition to the above dimensional reductions, two additional reductions are very natural from the viewpoint of our quantum interpretation. Let us note that in addition to the spectrum $\{\lambda_n\}$, we have two sets of "mixing parameters": $\{p_n^s\}$ and $\{p_n^w\}$. The relations between them are the following

$$\begin{aligned} p_n^w &= \langle w_n|w_n\rangle = \langle n|AA^\dagger|n\rangle = \rho_{nn}, & (11) \\ p_n^s &= \langle s_n|s_n\rangle = \langle n|A^\dagger A|n\rangle = N_{nn}. & (12) \end{aligned}$$

Elimination of small diagonal elements ρ_{nn} or N_{nn} is not equivalent to eliminating small eigenvalues of N or ρ. However, after this type of "purification" the resulting operators $\tilde{\rho}$ and $\tilde{N}$ are still positive and, hence, can be factorized as $\tilde{\rho} = BB^\dagger$, $\tilde{N} = C^\dagger C$, leading effectively to two new types of reduction: $A \mapsto B$ and $A \mapsto C$.

VI. FOCK SPACE OF WORDS

As we have seen, LSA can be formulated as a Hilbert space problem. The "bag of words" analysis is performed in $\mathcal{H}$. Ordered sequences of words can, in principle, be constructed in exact analogy to ordered sequences of letters in QIT. Still, there is a subtlety we want to point out.

Consider a phrase, i.e. an ordered n-tuple of words, $(\mathrm{word}_1, \dots, \mathrm{word}_n)$. Quantum physicist's intuition tells us that the natural representation of the sentence is a tensor product of vectors representing the words. The difficulty is this: Which vectors should one choose? The mutually orthogonal basis vectors $|j_1\rangle, \dots, |j_n\rangle$, or rather the associated word-vectors $|w_1\rangle = A^\dagger|j_1\rangle, \dots, |w_n\rangle = A^\dagger|j_n\rangle$?

Whatever representation one chooses, the phrase $(n_1, \dots, n_K)$ will be mapped into

$$|n_1 \dots n_K\rangle = |n_1\rangle \otimes \cdots \otimes |n_K\rangle \in \overbrace{\mathcal{H} \otimes \cdots \otimes \mathcal{H}}^{K} = \mathcal{H}^{\otimes K}.$$

Including the empty word we arrive at the Fock space of all the text passages $\mathcal{H}_F = \oplus_{K=0}^{\infty}\mathcal{H}^{\otimes K}$.

LSA is performed in $\mathcal{H}_F$ in exactly the same way as in $\mathcal{H}$. The structures one can investigate are much richer. Taking as an example G. Stein's phrase "Rose is a rose is a rose is a rose", not only can we work with

$$|s_1\rangle = 4|\mathrm{rose}\rangle + 3|\mathrm{is}\rangle + 3|\mathrm{a}\rangle \in \mathcal{H} \tag{13}$$

but also with vectors revealing the syntactic structures, for example,

$$\begin{aligned} |s_2\rangle &= |\mathrm{rose}\rangle \oplus 3|\mathrm{is}\rangle \otimes |\mathrm{a}\rangle \otimes |\mathrm{rose}\rangle \in \mathcal{H} \oplus \mathcal{H}^{\otimes 3} \subset \mathcal{H}_F, \\ |s_3\rangle &= \Big(|\mathrm{rose}\rangle + 3|\mathrm{is}\rangle\Big) \oplus 3|\mathrm{a}\rangle \otimes |\mathrm{rose}\rangle \in \mathcal{H} \oplus \mathcal{H}^{\otimes 2} \subset \mathcal{H}_F. \end{aligned}$$

The above formulas show a typical feature of Fock spaces, namely superpositions of vectors belonging to different tensor powers. It is very interesting that similar constructions are encountered in convolution-based memory models, such as TODAM [21] or Holographic Reduced Representations (HRRs) [4].

VII. RELATION TO SMOLENSKY'S TENSOR PRODUCT BINDING

Smolensky in [22] proposed tensor products of vectors as a means of solving the so-called binding problem: How to keep track of which features belong to which objects in a formal connectionist model of coding? In the linguistic context of SA the binding problem is equivalent to the problem of representing syntax. Links to quantum structures are particularly striking here, but there are also intriguing logical differences with what one would expect from a QIT perspective.

First, one represents an *activity state* of a network by a vector, and this is very close to what a quantum physicist would do. In comments to his Definition 2.1 Smolensky stresses that the vectors are always written in the same and fixed basis. So formally we do not really need vectors, but n-tuples of numbers are enough. This is against the philosophy of QIT where states are indeed vectors and the same information may be encoded in non-parallel vectors.

The fact that preferred basis is used becomes even more important in models such as TODAM or HRRs where the tensor product is replaced by its "compressed form": convolution or circular convolution. Both operations are defined on n-tuples and not on vectors. Still, one can argue that in quantum measurement theory we do indeed deal with preferred pointer bases [23] and the models such as HRRs may refer to this level of analysis.

A predicate `p(a,b)`, such as `eat(John,fish)`, is represented by the vector $\boldsymbol{r}_1 \otimes \boldsymbol{a} + \boldsymbol{r}_2 \otimes \boldsymbol{b}$ where the vectors $\boldsymbol{r}_k$ represent *roles* and $\boldsymbol{a}$, $\boldsymbol{b}$ are *fillers*. A predicate is, accordingly, given by an *entangled activity state*. A person trained on QIT would expect the vector to mean "role $\boldsymbol{r}_1$ AND filler $\boldsymbol{a}$, OR role $\boldsymbol{r}_2$ AND filler $\boldsymbol{b}$". Of course, the intention of Smolensky was different: The sum is meant to represent the conjuction (AND) and not the alternative (OR). This feature is also characteristic of other neural-network models. Why is it so and is this type of representation crucial for symbolic AI?

The above similarities and differences show that further exploration of possible implications of connectionist models for QIT, and vice versa, may be worth of further studies. We will not pursue these matters further here.

VIII. EFFICIENCY OF TENSOR REPRESENTATIONS

Tensor products are more "economic" than Cartesian powers due to the identifications of the type $\big(\alpha|\psi\rangle\big) \otimes |\phi\rangle = |\psi\rangle \otimes \big(\alpha|\phi\rangle\big) = \alpha\big(|\psi\rangle \otimes |\phi\rangle\big)$ that do not hold in Cartesian products. Thus the Fock space automatically performs a kind of dimensional reduction, which is the main idea of both LSA and distributed representations.

If we are more interested in the issue of binding than in ordering of words then further compression of information is possible if one employs symmetric (bosonic) or antisymmetric (fermionic) Fock spaces. Symmetric tensor powers are closer to convolutions employed in HRRs but, unlike convolutions, are defined on vectors and not n-tuples of numbers.

Let us also note that in binary (or qubinary) representations all tensor powers can be decomposed into irreducible components, exactly in the same way it is performed in 2-spinor calculus [24]. It is known that any irreducible representation corresponds to symmetric spinors and any antisymmetric spinor is a scalar times the singlet (all antisymmetric two-index spinors are proportional to one another). So it is very natural indeed to employ representations based on symmetric operations as the main building blocks of, say, memory models (convolution used in HRRs is also commutative).

All these links are interesting from the point of view of the discussions between Penrose and proponents of classical AI [25]. If brain is a quantum device, as suggested in [26] or, which is a weaker condition, if the conceptual part of the mind entails a formal quantum structure [27, 28], then the presence of tensor structures in SA or AI will not be accidental.

The question of tensor representations of semantic aspects of texts in principle can be settled experimentally. Document retrieval experiments based on quantum logic were already performed [29] and the results are encourageing.

Let us finally make the remark that some authors stress (cf. [30]) that semantic categorizations cannot be modelled by a set logic. Experiments were reported where, for instance, people were willing to accept that chairs are a type of furniture and that carseats are a type of chair, but would then deny that carseats are a type of furniture (for a review cf. [31]). Trying to model the meanings of 'furniture', 'chair', 'carseat' by means of set-theoretical constructions one arrives at contradiction with the inequality $P(A \wedge B \wedge C) \leq P(A \wedge C)$ (cf. also [32]). In QIT this type of contradiction is at the roots of the Bell inequality violation, whose proof is based on set-theoretic constructions while QIT employs tensor structures in Hilbert spaces. Similarly, tests of tensor structures via SA may play an analogous role in AI, quantitative linguistics, or experimental psychology, as the Bell inequality did for hidden-variables theories.

Acknowledgments

This research was supported by Grant G.0339.02 of the Flemish Fund for Scientific Research and the KBN Grant No. PBZ-Min-008/P03/03. We are indebted to T. Plate for his comments on tensor structures in distributed representations.

[1] J. S. Bell, "On the Einstein-Podolsky-Rosen paradox", Physics **1**, 195 (1964).

[2] C. H. Bennett, G. Brassard, in Proceedings of IEEE, International Conference on Computers, Systems and Signal Processing, Bangalore, India (IEEE, New York, 1984) 175.

[3] A. Ekert, Phys. Rev. Lett. **67**, 661 (1991).

[4] T. A. Plate, *Holographic Reduced Representations: Distributed Representations for Cognitive Structures* (CSLI Publications, Stanford, 2003).

[5] This paper is an extended version of the preprint D. Aerts, M. Czachor, "Bag-of-words problem and semantic analysis in Fock space", quant-ph/0309022.

[6] S. Deerwester et al., "Indexing by Latent Semantic Analysis", J. Am. Soc. Information Science **41**, 391 (1990).

[7] T. K. Landauer, P. W. Foltz, and D. Laham, "Introduction to Latent Semantic Analysis", Discourse Processes **25**, 259 (1998)

[8] K. Lund and C. Burgess, "Producing high-dimensional semantic spaces from lexical co-occurence", Behavior Research Methods, Instruments and Computers **28**, 203 (1996).

[9] T. Hofmann, "Probabilistic Latent Semantic Analysis", *Proceedings of Uncertainty in Artificial Intelligence, UAI'99*, Stockholm (1999)

[10] D. M. Blei, A. N. Ng, and Michael I. Jordan, "Latent Dirichlet Allocation", J. Machine Learning Research **3**, 993 (2003).

[11] T. L. Griffiths and M. Steyvers, "Prediction and semantic association", Advances in Neural Information Processing Systems" (2002).

[12] M. Steyvers, R. M. Shiffrin, D. L. Nelson, "Semantics spaces based on free association that predict memory performance", submitted to J. Experimental Psychology.

[13] T. K. Landauer and S. T. Dumais, "A solution of Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge", Psychological Review **104**, 211 (1997).

[14] T. K. Landauer, "On the computational basis of learning and cognition: Arguments from LSA", Psychology of Learning and Motivation **41**, 43, B. H. Ross (Ed.) (Academic Press, 2002).

[15] B. Oemer, *Quantum Programming in QCL*, M.Sc. Thesis, Technical University of Wien (2000).

[16] S. Bettelli, L. Serafini, and T. Calarco, "Towards an architecture for quantum programming", Eur. Phys. J. D **25**, 181 (2003).

[17] P. Seilinger, "Towards a quantum programming language", Mathematical Structures in Computer Science — in print.

[18] T. K. Landauer, D. Laham, and P. W. Foltz, "Learning human-like knowledge by Singular Value Decomposition: A progress report", Advances in Neural Information Processing Systems **10**, 45 (1998).

[19] WAS is a memory vector model where performance comparable to LSA is obtained with only 20-40 dimensions.

[20] F. Cooper, A. Khare, U. Sukhatme, "Supersymmetry and quantum mechanics", Physics Reports **151**, 268 (1995).

[21] B. B. Murdock, "A theory for the storage and retreival of item and associative information", Phychological Review **89**, 627 (1982).

[22] P. Smolensky, "Tensor product variable binding and the representation of symbolic structures in connectionist systems", Artificial Intelligence **46**, 159 (1990).

[23] P. Bush, P. J. Lahti, and P. Mittelstaedt, *The Quantum Theory of Measurement*, Lecture Notes in Physics, vol. m2 (Springer, Berlin, 1991).

[24] R. Penrose, W. Rindler, *Spinors and Space-Time*, vol. 1 (Cambridge University Press, Cambridge, 1984).

[25] See the collection of papers in the electronic journal Psyche **2**, http://psyche.cs.monash.edu.au

[26] R. Penrose, *The Emperror's New Mind* (Oxford University Press, Oxford, 1990); R. Penrose, *Shadows of the Mind*, (Oxford University Press, Oxford, 1994).

[27] L. Gabora and D. Aerts, "Contextualizing concepts using a mathematical generalization of the quantum formalism", J. Exp. Theor. Artificial Intelligence **14**, 327 (2002).

[28] D. Aerts and L. Gabora, "A quantum model for the rep-

resentation of concepts and their combinations", Kybernetes — in print.

[29] D. Widdows, S. Peters, "Word vectors and quantum logic experiments with negation and disjunction", *Proceedings of Mathematics and Language*, R.T. Oehrle & J. Rogers (Eds.) — in print.

[30] J. A. Hampton, "A demonstration of intransivity in natural categories", Cognition **12**, 151 (1982).

[31] J. Hampton, "Conceptual combination", *Knowledge, Concepts, and Categories*, K. Lamberts and D. Shanks (Eds.) (Psychology Press, Hove, 1997) 133.

[32] D. Aerts, S. Aerts, "Applications of quantum statistics in psychological studies of decision process", Found. Sc. **1**, 85 (1994).