

SEMANTIC SPACES

Yuri I. Manin¹, Matilde Marcolli²

¹*Max-Planck-Institut für Mathematik, Bonn, Germany,*

²*California Institute of Technology, Pasadena, USA*

ABSTRACT. Any natural language can be considered as a tool for producing large databases (consisting of texts, written, or discursive). This tool for its description in turn requires other large databases (dictionaries, grammars etc.). Nowadays, the notion of database is associated with computer processing and computer memory. However, a natural language resides also in human brains and functions in human communication, from interpersonal to intergenerational one. We discuss in this survey/research paper mathematical, in particular geometric, constructions, which help to bridge these two worlds. In particular, in this paper we consider the Vector Space Model of semantics based on frequency matrices, as used in Natural Language Processing. We investigate underlying geometries, formulated in terms of Grassmannians, projective spaces, and flag varieties. We formulate the relation between vector space models and semantic spaces based on semic axes in terms of projectability of subvarieties in Grassmannians and projective spaces. We interpret Latent Semantics as a geometric flow on Grassmannians. We also discuss how to formulate Gärdenfors’ notion of “meeting of minds” in our geometric setting.

O INTERIOR DO EXTERIOR DO INTERIOR

Pascal Mercier
“Nachtzug nach Lissabon”

1. Introduction: linguistics, semiotics, and topology

One of the basic “meta-physical” principles of classical physics consisted in the subdivision of informational content of any physical model into two parts:

- a description of the configuration and phase spaces of the studied system;
- a description of the time evolution law (usually a vector field in the phase space).

Some of the recent approaches to semantics of natural languages describe various versions of “spaces of meanings” which we consider as a metaphorical analog of configuration spaces: cf. comprehensive accounts [Gä00], [Gä14]. For Gärdenfors, semantics is (in particular) *meeting of minds*, and the space of meanings is the space where minds meet.

Our initial motivation for undertaking this survey and the research summarised in this paper was our desire to introduce “a time dimension” in this discussion, to see a discourse or reception of a text as a path in the appropriate space of meanings.

In particular, we wanted to use mathematical models in order to bridge the approaches to semantics reviewed in [Gä14], neurolinguistic studies reviewed in [JeLe94], [InLe04], and neurobiological studies of neural mechanism involved in coping with tasks related to orientation in physical space (see [CuIt08], [CuItV-CYo13] an brief survey for mathematicians [Ma15].)

In the remaining part of the introduction we will give a very short list of several approaches to description of “meaning” using geometric/topological representations and/or metaphors.

1.1. Semic axes. In the following it is essential to keep in mind that core “meanings” are generally assigned not to “words” but to “lexemes”. According to [Me16], p. 240, lexeme is “a word taken in one well defined sense – more precisely, a set of all word forms and analytical form phrases that differ only by inflectional significations.”

Example ([Me16], p. 135): lexeme TAKE_(V) includes the following lexical items: *take, takes, took, taking, ... , have taken, has taken, ... , have been taken, ...*

The tag *(V)* here means that our lexeme refers to the word ”take” understood as *a verb* rather than *a noun*.

When one extracts a vocabulary of lexemes from a dictionary of words, one should do “stemming” (extracting roots of words), “tagging” etc., cf. a more detailed description in Sec. 3 of [TuPa10].

We will allow ourselves the use of the term “word” in place of “lexeme” when it cannot lead to a confusion.

The approach to encoding of meaning, or “sense” of lexemes, briefly surveyed in [Gui08], starts with postulating a list of “semes” such as *animate, inanimate, actor, process* etc.

The meaning is specified by listing a subset of semes.

In the respective geometric picture, N senses are represented by basis vectors e_i , $i = 1, \dots, N$, of \mathbf{R}^N , and meanings are represented by (a subset of) vertices of the unit cube $[0, 1]^N$. P. Guiraud actually prefers the “bisemic” description, in which meanings are represented by a subset of vertices of $[-1, 1]^N$. Sign changes of basic coordinates represent the complementarity relations such as in animate/inanimate.

A qualitative weakening of the bisemic model allows meanings to be represented by points in \mathbf{R}^N that are localised near the boundary of the unit cube, but not necessarily coincide with its vertices. A nice illustration is given on p. 59 of [Gä14]. It represents bisemes in a two-dimensional “emotional space” \mathbf{R}^2 whose bisemic axes represent dichotomies *pleasure/displeasure* and *high/low* whereas, say, the quadrant “low pleasure” accommodates lexemes *content, serene, calm, relaxed, sleepy*.

Some of the largest subsets of the space of meanings that can accommodate, say, path of a narrative, might encode notions related to

- senses: vision, hearing, feelings, time, space ...
- some subregions like “far away – near” , “quiet – loud” , “past – future”
- regions related to “me” , to “other people” , “unrelated to humans” , etc.

What is important is that we should construct this semantic space at first in a way maximally independent of the “natural language” we choose, and that it will widen at each stage of construction in order to accommodate new words, sentences, languages etc.

1.2. Semic axes and neural encoding of place field recognition. We want to derive from semantics of a natural language a structure encoding it that would be a space covered by subsets, say, U_i . (Some) non-empty finite intersections should correspond to words or short sentences, paths through this space should correspond to texts.

A nice example of this is provided in [Li], together with a picture representing symbolically two different subsets of semantic space in two possible mutual relationships: (i) inclusion of one in another, and (ii) non-empty intersection without inclusion.

This picture illustrates the difference between usages of words *which* and *that* in the following two sentences:

Correct use of *that*: “*Tiffany likes shoes that are expensive*”.

“The set of things called shoes includes both expensive and inexpensive shoes, so when we say ‘that are expensive,’ we are talking only about a subset of the set of all things called shoes.”

Correct use of *which*: “*Tiffany likes emeralds, which are expensive*”.

“The set of things called emeralds are all expensive, so the clause ‘which are expensive’ talks about the whole set of emeralds. There is no inexpensive subset of emeralds. ‘Which are expensive’ simply gives you additional information about this whole set” ([Li]).

This basic picture representing meanings by domains in the space of meanings and the relationship of intersection/inclusion between the respective domains fits very well the studies aimed to the understanding how brain copes with multiple tasks of orienting and navigating in the world, cf. [CuIt08], [CuItVCYo13], [Yo14], and references therein.

The brain of an animal must be able to reconstruct, say, a map of its environment and its current position in it, using only the action potentials (spikes) of the relevant cell groups. In laboratory experiments it is found that stimuli related to the positions are naturally divided into groups, and with each group a certain type of neural activity is associated. In [CuIt08] and [Yo14], it is postulated that a given domain of stimuli can be modelled via a topological, or metric stimuli space X . Furthermore, brain reaction to a point in X is modelled by spiking activity of certain finite set of neurons NX . The list of subsets of NX consisting of subsets whose neurons can be activated simultaneously, corresponds to a certain covering of X . Thus this covering can be described by a binary code, and relations of intersection/inclusion between domains coincide with the relations of intersection/inclusion between the respective code words. For more details, see [Ma15].

1.3. Meaning–Text model. In the model of semic axes, there is one intrinsic source of incompleteness: as P. Guiraud says ([Gui68], p. 157), the lexical units (corresponding to vertices of $[-1, 1]^N$) “*must in addition be associated in syntagms, each one of them constitutes a ‘sense’. But there again we must setup rules for combinations, for the sense supposes that certain syntagms are permitted, other excluded.*” The difference between *which* and *that* discussed above is precisely an example of such syntagms.

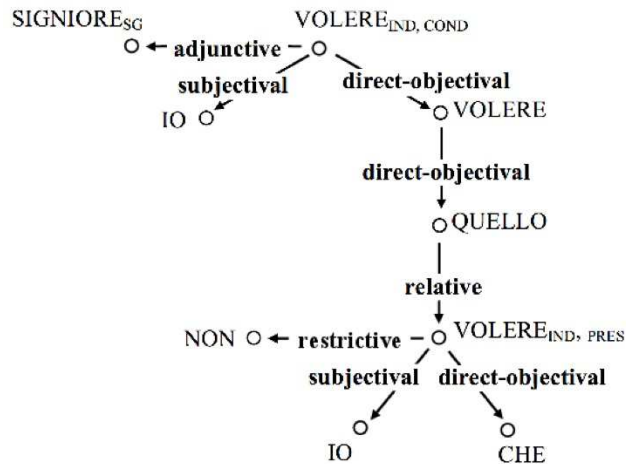
This problem is very systematically addressed in the model “Meaning–Text” which I. Mel’čuk and his collaborators have been developing for several decades: see [Me16] for its most recent summary and further references.

In this model, meaning of a text of language \mathbf{L} “*is exclusively [. . .] linguistic meaning*” that can be extracted “*only on the basis of the mastery of \mathbf{L} , without the*

participation of common sense, encyclopaedic knowledge, logic etc. ([Me16], Sec. 3.2.2).

On the other hand, geometry/topology figures in this model mainly as a tool for producing graphs of various levels of linguistic representation. Each such graph consists of several vertices, certain pairs of which are connected by edges. Moreover, both vertices and edges are additionally marked. For example, on the level of the surface-syntactic structure, a sentence is represented by the graph, whose vertices are marked by lexemes corresponding to the words in this sentence, and by additional information encoding the passage from the lexeme to the word. Edges of this graph are marked by technical terms expressing syntactic relations between the respective pair of words.

Below we illustrate this principle by presenting the surface-syntactic graph of the first line of a sonnet by Michelangelo. We are very grateful to I. Mel'čuk who produced for us this graph and allowed us to reproduce it here.



The Surface-Syntactic Structure of the sentence

Vorrei voler', Signor', quel' ch'io non voglio.

1.4. Neurolinguistic data. There exists a large body of neuroimaging studies of production and perception of spoken language. In the survey [InLe04], the reader will find descriptions of methodology used and results obtained in *“the enterprise of relating the function components of word production, such as lexical selection,*

phonological code retrieval, and syllabification, to regions in a cerebral network” ([InLe04], p. 102.)

An illustration of segment of lexical production network ([JeLe94], p. 826) shows fascinating parallels with the Meaning–Text model.

Due to the vastness of semantic space needed to accommodate all meanings expressible in a natural language, direct comparison with the neural encoding of place field recognition as in [CuIt08] is not yet feasible. However, the development of new methods of studying and collecting databases of results allows us to hope that such comparison will become possible. In this paper, we try to contribute some mathematical tools that may be useful for this endeavor.

1.5. This report. Most of the approaches discussed above directly appeal to the linguistic intuition and communicative experience of scientists, experimenters, and participants in experiments. Information obtained by the respective methods should be considered as local data about semantic space, and/or about short paths in it.

On the other hand, if we want to obtain mathematical models of topology of semantic spaces and of longer routes in such a space, expressed by texts of the size, say, of a chapter in “War and Peace”, we may turn to the statistical natural language processing.

Then, in the first approximation, a text becomes a point in the space of paths in the semantic space, and we discuss here approaches to studying the topology of such spaces appealing mostly to the data about frequencies of lexemes and other text fragments taking in account their linear ordering in the text. Semantics of such fragments as it is represented by dictionaries and experiments is thus put aside to a certain degree, although not fully.

In the main body of this article, we will describe some mathematical tools that can be used for the introduction of “time dimension” in the study of texts. They will refer to the geometry of real projective spaces and real Grassmannians. Passage from texts to the relevant geometry is based here on the Vector Space Models of semantics (VSM) surveyed in [TuPa10], and we will briefly explain this model for further use.

In Section 2 we discuss how the frequency matrix of the VSM approach, that counts occurrences of lexemes in contexts in a given corpus of texts, determines a point in a Grassmannian. We show that, in the case of a large vocabulary of

lexemes and a smaller number of contexts, the condition that the resulting point lies in the positive Grassmannian provides a geometric test for the property that a choice of lexemes gives a good semantic disambiguation of the contexts. A similar condition holds in the case of a small number of lexemes, where one wants to test if a set of contexts would disambiguate the words semantically. This geometric viewpoint takes into account the fact that contexts come with a specific ordering by occurrence in a text.

In Section 3 we discuss other geometric models associated to the frequency matrices of the VSM approach, which also takes into account the specific ordering of contexts in a text. We assign to a text a piecewise geodesic path of points in a projective space. Instead of measuring semantic relatedness in terms of angle distances between the semantic vectors of the frequency matrices, as it is customary to do in Natural Language Processing, we compare the paths in an ambient projective space through a geometric distance function between (geodesic) polygonal curves, which is known to be computable in polynomial time. In a variant of this construction, we also consider assigning to the frequency matrix a point in a flag variety, where the flag corresponds to the span of successive semantic vectors for the successive contexts ordered by occurrence in a text. Again semantic similarity can be measured in terms of the geodesic distance in the flag variety, with respect to its natural metric as a quotient of Lie groups.

In Section 4 we consider the case where lexemes are grouped together according to some semantic axes, either by explicit semantic tagging (supervised learning) or just by grouping together lexemes with similar occurrences in contexts (unsupervised learning). In both cases, we describe the process of passing from frequency matrices for a given corpus of text, computed with respect to a dictionary of lexemes, to density matrices with respect to a semantic dictionary, where identification of lexemes by semantic criteria has already occurred. When we view the frequencies as determining points in Grassmannians, we can view geometrically this operation as a projection between two Grassmannians. The question of whether one can avoid loss of semantic information in this process, when applied to a given collection of texts, is then interpreted in terms of whether the points corresponding to these texts lie on a subvariety of the Grassmannian that can be isomorphically projected to the other Grassmannian. A similar condition arises when we assign to a given text a piecewise geodesic path in a projective space as discussed in Section 3.

In Section 5 we connect the geometric setting described in the previous section with the point of view of persistent topology. According to our previous construc-

tion, a large corpus of texts determines a corresponding set of points in an ambient Grassmannian, where we assume that the same fixed dictionary of lexemes (or semes) is used to analyze all texts in the corpus. We then show that one can identify more refined forms of semantic relatedness between these points. These are topological in nature and arise from constructing Vietoris–Rips simplicial complexes at varying scales, associated to the set of points in the ambient variety and computing their persistent homology. We discuss possible relations to the use of persistent topology in the theory of neural codes.

In Section 6 we show that the Latent Semantics technique for dealing with very sparse frequency matrices in the VSM approach, which identifies lower dimensional subspaces (latent meanings) through singular value decomposition, can be interpreted in terms of the geometry of Grassmannians described in Section 2, as a Riccati flow on the ambient Grassmannian.

In Section 7 we discuss how to implement, in our geometric setting, a model analogous to Gärdenfors’ “meeting of minds”, where common meaning between different users communicating with one another is achieved as via a fixed point problem in a convex semantic space. We suggest that a similar idea can be implemented in our setting if different users come to somewhat different semantic interpretations of a given texts, on the bases of semantic interpretations based on other texts available to them, under the assumption that users have access to different (partially overlapping) corpora of texts. We then describe the procedure of “meeting of minds” as the construction of a geodesic barycenter in the ambient geometric space of the distribution of points obtained by the users, possibly weighted according to some measure of “reliability” of the different corpora used for semantic interpretation.

In Section 8 we discuss how to compensate for the fact that the frequency distribution for words in a dictionary is skewed towards the more frequent and less semantically significant words according to Zipf’s law.

2. Vector Space Models of semantics

2.1. Texts and their processing. A concrete VSM starts with a large corpus of natural language texts and produces from it a matrix of numbers (frequencies). The intermediate steps of this production are subdivided into two groups: *(i) linguistic processing*, and *(ii) statistic processing*.

For us, linguistic processing results in the creation of the relevant *vocabulary of lexemes* where we understand “lexeme” as in 1.3 above. Each text is also represented

as a sequence of the relevant lexemes, although from the description of [TuPa10] it becomes clear that at least some fragments of it are modelled by their surface-syntactic structures in the sense of Meaning-Text model.

We accept this as a reasonable approximation to the procedures described in Sec. 3 of [TuPa10].

Statistic processing, as we mentioned, produces a (normalised) *matrix of frequencies*, see [TuPa10], Sec. 4.

In the typical case called “the term-document matrix” in [TuPa10], rows of the matrix are labelled by lexemes (“terms”), whereas columns are labelled by texts in our collection.

In another typical case called “the word-context matrix” ([TuPa10], Sec. 2.5), the texts, already at the stage of linguistic processing, are represented as a union of “contexts”. Here again the rows of matrix are labelled by lexemes, whereas columns are labelled by contexts.

Finally, matrix entries as we mentioned characterise correlations between the lexemes and text/context. We will treat in more detail some cases below, and address the question of “smoothing”.

2.2. “Time dimension” and other linear orderings. Any vocabulary of lexemes, or contexts, must be in the final count also presented as linearly ordered dictionary. This ordering might be totally irrelevant to the situation under study (as e.g. alphabetic ordering). It can take into account the order of first appearance of the respective lexeme in the text. Finally, it can be a Zipf’s-like ordering according to diminishing frequency rate.

A considerable part of statistical characteristics of a VSM does not depend of the chosen orderings (although the mode of their usage might depend on it). However, for the purpose of our paper this might become essential, and we will pay due attention to it.

2.3. Notation and assumptions. We will consider word-contexts matrices described above in one of two possible extreme subcases.

(A). *Large vocabulary case.* In this setting, we assume that our vocabulary of lexemes is sufficiently large and includes at least all the lexemes that appear in the texts (excluding words with large occurrences in all contexts such as “and” or “the” in an English text that are semantically less informative). Moreover, we assume that the size of the vocabulary is large compared to the number of contexts in the

texts. In this case, one aims at selecting from the large dictionary choices of words that best represent the given contexts semantically.

(B). *Information retrieval case.* In this case we consider a vocabulary that is small compared to the number of contexts, as would be the case with a choice of words used in a query. In this case one aims at selecting among the various contexts in a given corpus those that best match semantically the chosen words in the query.

Let D be our vocabulary, and $M = \#D$ be the number of lexemes in it.

A given text T is then endowed with a set of subtexts called *contexts*: $C(T) = \{c_1, \dots, c_N\}$. Typical examples of contexts are: sentences, paragraphs, or else windows of certain length around each word/lexeme.

2.4. Matrix of frequencies. Following [TuPa10], one produces from these data an $N \times M$ matrix of frequencies $P = P(T)$ with entries p_{ij} . Here p_{ij} is the estimated probability (frequency) of occurrence of the word $w_i \in D$ in the context $c_j \in C(T)$. In the VSM model, one usually considers also the matrix $X = X(T)$ with entries $X = (x_{ij})$,

$$x_{ij} = \max\left\{0, \log\left(\frac{p_{ij}}{p_{i*}p_{*j}}\right)\right\},$$

where $p_{i*} = \sum_j p_{ij}$ is the estimated probability of the word $w_i \in D$ and $p_{*j} = \sum_i p_{ij}$ is the estimated probability of the context $c_j \in C(T)$. The condition that $p_{ij} = p_{i*}p_{*j}$ corresponds to statistical independence of word w_i and context c_j , while $p_{ij} > p_{i*}p_{*j}$ signals the presence of a semantic relation between them.

More precisely, the formula $p_{i*} = \sum_j p_{ij}$ gives the frequency of appearance in the text in the case where contexts do not overlap whereas their union is the whole text.

In the more general case where contexts may overlap one still uses the same matrix but now its entries are the frequencies of appearance across all contexts (or, equivalently, the frequencies of appearance in the text, weighted by some multiplicities that keep track of when a word appears in the intersection of more than one context).

If a word is in the intersection of two adjacent contexts j and $j+1$, then it affects the counting in both p_{ij} and $p_{i,j+1}$, so $\sum_j p_{ij}$ is still the normalization factor.

The typical example of this “overlapping contexts” method is the original Shannon 3-gram model: here one has probabilities (based on frequencies) for occurrences

of 3 words in a row. For example, one can have a word sequence a-b-c-d where the three words a-b-c have a very high probability of occurring together, while the probability of the triple b-c-d is very low. This suggests that it is a-b-c rather than b-c-d that clarifies better the semantic meaning of the words b and c, and that the semantic meaning of d will more likely be clarified by the trigrams that follow like c-d-e and d-e-f, with the following words e,f, rather than by b-c-d.

2.5. Large Vocabulary case. Here we have $N \leq M$. The dictionary D includes (at least) all the (relevant) lexemes that occur in the text T , and the number of contexts in which the text T is subdivided is smaller than the number of words in the dictionary.

The *Statistical Semantics Hypothesis* states that statistical patterns of word usage in texts determine their semantical meaning, and in particular that (parts of) text that have similar vectors in the above frequency matrices also have similar meanings.

Let $r = \text{rank}(P)$ be the rank of the matrix $P(T)$. In the case of large dictionary, we have $r \leq N$. Under the Statistical Semantics Hypothesis, the rank r measures the largest number of words and contexts that the text T disambiguates semantically. Namely, the linear dependence of frequency vectors is interpreted as revealing the presence of underlying semantic relations. When $r = N$, all the contexts in T have a choice of corresponding words that they semantically disambiguate.

In the case where $r = N$, the matrix $P(T)$ of a text T determines a point $p(T)$ in the real Grassmannian $Gr(N, M)$ of N -planes in real Euclidean space \mathbf{R}^M . Similarly, if $\text{rank}(X(T)) = N$, the matrix $X(T)$ determines a point $x(T) \in Gr(N, M)$. For simplicity, we argue about the matrix $P(T)$. When not otherwise stated, the same will apply to $X(T)$. Let $\mathcal{M} = \mathcal{M}(T)$ be the set of subsets of $\{1, \dots, M\}$ of cardinality N , such that the determinant of the corresponding minor is $\Delta_I(P(T)) \neq 0$. The set \mathcal{M} determines a matroid stratum $\mathcal{S}_{\mathcal{M}} \subset Gr(N, M)$, with $P(T) \in \mathcal{M}$.

In the case $N \leq M$, instead of working with a fixed (large) dictionary D for all texts, it is convenient, given a text T , to discard all the words in D that do not appear anywhere in T , as the text does not have any relevance for those words. Thus, we can assume that $D = D(T)$, with $\#D(T) = M(T)$ is the list of words that appear in T (with a suitable stop list). A text T has a linear ordering, which induces an ordering on the set $D(T)$ that lists words in order of apparition in T . We identify $D(T)$ with the set $\{1, \dots, M(T)\}$ using this ordering. Similarly, the set $C(T)$ of contexts is also ordered by how they are ordered in the text T , and we

identify $C(T)$ with $\{1, \dots, N(T)\}$ using this ordering. The order of apparition of words in the text T is relevant to the semantic interpretation of the text, as the first occurrence of a word is the first instance where a semantic interpretation for that word is required.

Consider then the set of subsets $I = \{i_1, \dots, i_N\}$ of $[M] := \{1, \dots, M\}$ with $i_1 < i_2 < \dots < i_N$. These correspond to choices of words w_{i_1}, \dots, w_{i_N} in $D(T)$, such that the order of apparition of these words in the text T is respected, and we consider the frequency vectors $P_{i_k} := (p_{i_k, j})_j$ for the occurrence of the word w_{i_k} in the context c_j . We consider the Gale ordering on these subsets I . Namely, two such subsets $I = \{i_1, \dots, i_N\}$ and $J = \{j_1, \dots, j_N\}$, with $i_1 < i_2 < \dots < i_N$ and $j_1 < j_2 < \dots < j_N$, we have $I \leq_G J$ iff $i_1 \leq j_1, i_2 \leq j_2, \dots, i_N \leq j_N$. The Gale ordering corresponds therefore to the relative position of words w_{i_k} and w_{j_k} in the dictionary $D(T)$ according to first apparition in T .

The original dictionary D also has an ordering, and therefore the smaller dictionary $D(T)$ also has an induced ordering, which is different than the order of apparition in the text T . One then has some permutation $\sigma \in S_M$, such that the Gale ordering described above corresponds to the ordering $I \leq_\sigma J$, namely $\sigma^{-1}I \leq_G \sigma^{-1}J$.

The condition that, for one of these subsets I , the corresponding minor of the matrix $P(T)$ has vanishing determinant $\Delta_I(P(T)) = 0$ means that there is a linear dependence between the vectors P_{i_k} , hence under the Statistical Semantics Hypothesis a semantic relation between the w_{i_k} . Thus, the matroid stratum $\mathcal{S}_{\mathcal{M}} \subset Gr(N, M)$ containing the point $p(T) \in \mathcal{S}_{\mathcal{M}}$ determined by the text T describes, for the given contexts c_i of the text T , all the choices of words w_{i_k} , $k = 1, \dots, N$ in the dictionary for which the semantic vectors P_{i_k} are independent. This can be seen as the maximal amount of semantic information that can be extracted from the text and its contexts.

Recall that the positive (or totally non-negative) Grassmannian $Gr_{\geq 0}(N, M)$ is the subset $Gr_{\geq 0}(N, M) \subset Gr(N, M)$ of matrices A such that for all $\Delta_I(A) \geq 0$, for I as above. The intersections of the matroid strata with the positive Grassmannians $\mathcal{S}_{\mathcal{M}}^{\geq 0} = \mathcal{S}_{\mathcal{M}} \cap Gr_{\geq 0}(N, M)$ are cells, the *positroid cells* of [Pos06].

In particular, the condition that the point $p(T)$ lies in the positroid cell $\mathcal{S}_{\mathcal{M}}^{\geq 0}$, that is, that all $\Delta_I(A) > 0$, for all $I \in \mathcal{M}$, is equivalent to the existence of continuous paths γ_I , for each $I \in \mathcal{M}$, where $\gamma_I(0) = P(T)$ and $\gamma_I(1)$ is a matrix where the I -minor is the identity, and for all $t \in [0, 1]$ one has $\gamma_I(t) \in \mathcal{S}_{\mathcal{M}}^{\geq 0}$. This condition

can be regarded as expressing the fact that the choice of words w_{i_1}, \dots, w_{i_N} for the contexts c_1, \dots, c_N of the text T contains a maximal amount of semantic information. Indeed, the case where the corresponding minor would be the identity, would correspond to a case where the word w_{i_k} is entirely specified semantically by the context c_k and by none of the other c_j with $j \neq k$.

2.6. Information Retrieval case. We now focus on the other case mentioned above, the “information retrieval” setting, where we have $N \geq M$, that is, where the list of words is, for example, the list of words in a query, and one wants to locate texts, or contexts within a text, that are semantically most relevant for that query. In this case, we can assume that the number of words searched is no greater than the number of contexts.

The setting is similar to what we described before, except that we now consider the case where the matrix $P(T)$ determines a point in the Grassmannian $Gr(M, N)$. The minors $I = \{i_1, \dots, i_M\}$ correspond to choices of contexts c_{i_k} in the text T in response to a query given by the words w_k . As before the condition $\Delta_I(P(T)) > 0$ corresponds to those assignments of a context to each word of the query that best matches it semantically.

2.7. Literary texts and their statistical processing. D. Yu. Manin in his article [Man12] suggests that literary texts (prose/poetry) require qualitatively different methods of statistical processing in order to make explicit what puts them apart from texts produced in ordinary speech.

Here we only mention a different kind of contexts used there ([Man12], p. 286).

Namely, a context in his sense is a fragment of text with a blank, a hole where different words might occur, like “a-*–b”. This would allow one to extract statistical data allowing one to say that “words x and y often occur in the same contexts”. Presumably, this fact would then reflect semantic relationships between x and y.

In the limiting case where x can occur in all the same contexts as y, and with the same frequencies, that would mean that x and y are exact synonyms. Or, if x can share contexts with u and v, but u and v do not share contexts, then they probably represent two very distinct meanings of x.

In this paper, we do not try to study semantic spaces and paths in them relevant to this approach. We only mention that it might be a very interesting project.

3. Projective Spaces and Flag Varieties

We describe here two variants of the construction above, aimed at encoding more explicitly the fact that a linguistic text has an ordered linear structure that is crucial to its semantic interpretation. We propose two modifications of the geometry described above that better encode this fact. One is based on regarding a text subdivided into contexts, as a collection of points determining a path in a projective space, rather than as a single point in a Grassmannian. The second is in terms of points in a flag variety.

3.1. Texts as Paths in Projective Spaces. Here we again consider the case where we have some fixed large vocabulary D of lexemes of size $M = \#D$, which contains at least all the words in the given text T . We also subdivide the text into contexts c_k , as before, but we do not necessarily assume that the total number N of contexts is smaller than M . Indeed, in this setting we could be dealing with a large corpus of texts and a large number of contexts. We again consider the semantic vectors $P_k(T) = (p_{ik})_{i \in D}$ that collect the probabilities (frequencies) of occurrence of words $w_i \in D$ in the context c_k of T . We regard each P_k as determining a point p_k in the projective space $\mathbf{P}^{M-1} \simeq Gr(1, M)$. Thus, a text T here corresponds to an ordered N -tuple of points in \mathbf{P}^{M-1} , where N is the number of contexts. We can think of this collection of points as an oriented path by drawing geodesic arcs between consecutive points. We denote by $\Gamma(T)$ the resulting path associated to a text T .

Given different texts T and T' , the comparison at the level of semantic vectors can be performed, in this setting, by computing the distance between the corresponding paths in the same ambient \mathbf{P}^{M-1} . This can be computed as the Fréchet distance between the two polygonal curves. The latter is defined as the infimum over reparameterizations by $[0, 1]$ of the maximum over $t \in [0, 1]$ of the distance between corresponding points

$$\delta(\Gamma(T), \Gamma(T')) = \inf_{\gamma, \gamma'} \max_{t \in [0, 1]} d_{FS}(\gamma(t), \gamma'(t)),$$

where $\gamma : [0, 1] \rightarrow \Gamma(T)$ and $\gamma' : [0, 1] \rightarrow \Gamma(T')$ are parameterizations of the two curves by $[0, 1]$, and $d_{FS}(x, y)$ is the Fubini-Study metric on \mathbf{P}^{M-1} . The Fréchet distance for polygonal curves is computable in polynomial time ([AlGo95]).

3.2. Texts as points in flag varieties. Another way to keep track of the linear ordering of contexts in a given text is by building larger subspaces, as more and

more contexts in the given texts are encountered in a linear reading of the text. Thus, if $P_k(T) = (p_{ik})_{i \in D}$ are the semantic vectors as above, one considers the vector spaces $V_k = \text{span}\{P_j : j = 1, \dots, k\}$. The spaces $V_1 \subset V_2 \subset \dots \subset V_N$ form a flag in \mathbf{R}^M . We denote by $F(d_1, \dots, d_\ell)$ the flag varieties of flags $W_1 \subset \dots \subset W_\ell$ with $\dim(W_k/W_{k-1}) = d_k$. We associate to a text T the point of the corresponding flag variety $F(1, \dots, 1, M - N)$ determined by the flag $V_1 \subset V_2 \subset \dots \subset V_N$ with $V_k = \text{span}\{P_j : j = 1, \dots, k\}$.

The natural Fubini–Study metric on projective spaces has an analog for Grassmannians and flag varieties. It is obtained from the curvature form of the first Chern class of the determinant line bundle of a hermitian vector bundle ([Dem88]), or else by considering these varieties as quotients of $SU(n)$ by subgroups, with the metric induced from the bivariant metric of $SU(n)$ ([Gri74]). Thus, one can compare texts viewed as points in Grassmannians or in flag varieties, by measuring their distance with respect to this metric.

4. From Lexemes to Semantic Dictionaries

We now consider a setting where, instead of a “lexemes dictionary” D of words, one passes to a “semantic dictionary” S where lexemes are grouped together according to some semantic description. This can happen in two different ways, based on supervised or unsupervised learning.

(a) *Supervised Learning*. In this case, also referred to as “sense tagging” (see [MaSch99]), lexemes are grouped together into semantic categories by assigning appropriate tagging. In this setting, the type of question we look at is to what extent the information contained in the semantic vectors computed for the initial lexical vocabulary still retains the correct information when passing to a quotient that corresponds to the identification by semantic categories.

(b) *Unsupervised Learning*. In this case, sense tags are not assigned, so that one cannot identify directly the corresponding semantic categories, but one can still obtain a “sense discrimination” by grouping together words into unlabelled groups using the information contained in the semantic vectors. In this setting, we will show that the resulting grouping can be studied in terms of *persistent topology* ([Ca09]).

4.1. Supervised Learning. We consider the case where we associate to texts T points $p(T)$ in a Grassmannian (either $Gr(N, M)$ or $Gr(M, N)$ depending on

relative size of vocabulary and contexts). We consider the case $N < M$. The other possibility can be treated similarly.

We want to consider also the case where we deal not with a single text T but with a corpus consisting of several texts. In this case, we need to assume that the vocabulary D , with $M = \#D$, is large enough to include all words that occur in all the texts of the corpus. Moreover, if we choose an ordering of the dictionary, as discussed previously, by order of apparition in a text, we can extend the order to the whole corpus, by choosing an order in which the different texts in the corpus are looked at. For the model with points in Grassmannians, or in flag varieties, we consider the case where the number N of contexts is fixed across all texts in the corpus. In the more general case where the number $N = N(T)$ varies across texts, we will be working with the model in which texts determine a sequence of points and an oriented polygonal path in a fixed projective space. In both cases, the question will be the behavior of the locus (in the Grassmannian, flag variety, or projective space) determined by the semantic vectors of all texts in the corpus, under a projection map that corresponds to passing from the lexical to the semantic dictionary.

4.2. Points in Grassmannians and Flag Varieties. At the level of the matrix $P(T)$ and the corresponding point $p(T)$ in the Grassmannian $Gr(N, M)$, one can view the operation of passing from the lexemes in D to the semantic categories in S as the effect of a projection $\pi_{M, M'} : Gr(N, M) \rightarrow Gr(N, M')$, where $M' \leq M$ is the size of the set of semantic categories considered, $M' = \#S$.

We regard a corpus $\mathcal{C} = \{T\}$ of texts T as a discrete sampling of a subvariety of the Grassmannian $Gr(N, M)$, under the hypothesis that the number of contexts is fixed and the size of the dictionary D is also fixed for all $T \in \mathcal{C}$. We denote by $\Pi_{\mathcal{C}} = \{p(T)\}_{T \in \mathcal{C}}$ the finite set of points on $Gr(N, M)$ corresponding to the texts in the corpus. Given the finite set $\Pi_{\mathcal{C}}$, we consider possible algebraic subvarieties $X_{\mathcal{C}} \subset Gr(N, M)$ that interpolate the points $p(T) \in \Pi_{\mathcal{C}}$, namely algebraic subvarieties $X_{\mathcal{C}}$ of $Gr(N, M)$ with $\Pi_{\mathcal{C}} \subset X_{\mathcal{C}}$.

We recall some results about projectability of subvarieties of Grassmannians, see [ArRa05]. A subvariety $X \subset Gr(N, M)$ is k -projectable, for some $0 \leq k \leq N - 1$, under $\pi_{M, M'} : Gr(N, M) \rightarrow Gr(N, M')$ if any two N -planes in the image of X only meet along linear spaces of dimension less than k . The case $k = N$ corresponds to X being isomorphically projectable to $Gr(N, M')$. Note that k -projectability also implies that no two N -planes in X can intersect in dimension greater than or equal

to k .

If the variety $X_{\mathcal{C}}$ associated to a corpus \mathcal{C} of texts is k -projectable to $Gr(N, M')$, this means that the N -planes given by the images $\pi_{M, M'}(p(T))$ and $\pi_{M, M'}(p(T'))$ of any two points $p(T), p(T')$, with $T, T' \in \mathcal{C}$, will intersect in at most a $(k - 1)$ -dimensional space.

The size of the intersection between the N -planes of T and T' is a measure of dependence between the respective semantic vectors, hence of the semantic relatedness of the two texts. If in the variety $X_{\mathcal{C}}$ every two N -planes intersect in dimension less than k , but the variety is not k -projectable under $\pi_{M, M'} : Gr(N, M) \rightarrow Gr(N, M')$, this means that there is loss of semantic information in the matching of words (and their semantic categories) to contexts in the texts of the corpus.

There are strong algebro-geometric restrictions on k -projectable varieties. For example, it is shown in [ArRa05] that the Veronese embedding of \mathbf{P}^n is the only variety in $Gr(d - 1, dn + d - 1)$ that can be projected to $Gr(d - 1, n + 2d - 3)$ so that any two $(d - 1)$ -planes meet in at most one point.

We have only discussed here the case where we associate texts to points in Grassmannians. The case of points in flag varieties is similar, with similar questions about k -projectable subvarieties.

4.3. Paths in Projective Spaces. We then consider the case where the size $N = N(T)$ of contexts in a text is varying with $T \in \mathcal{C}$. In this case, instead of working with texts defining points in a Grassmannian, it is more convenient to adopt the viewpoint where texts determine polygonal paths in a projective space \mathbf{P}^{M-1} with $M = \#D$ the size of the dictionary. In this case, we are looking at a similar question about k -projectable subvarieties in projective spaces.

More precisely, we consider again algebraic subvarieties $X_{\mathcal{C}}$ of \mathbf{P}^{M-1} that contain all the paths $\Gamma(T)$ for $T \in \mathcal{C}$. As a weaker condition, we can just assume that the variety $X_{\mathcal{C}}$ contains the set of points $\Pi_{\mathcal{C}} = \{p_k(T) : T \in \mathcal{C}, k = 1, \dots, N(T)\}$. If $X_{\mathcal{C}}$ is also geodesically complete, then it contains also the paths $\Gamma(T)$.

We then consider a projection $\pi_{M, M'} : \mathbf{P}^{M-1} \rightarrow \mathbf{P}^{M'-1}$ that corresponds to performing some identification of the vocabulary by grouping lexemes according to a choice of semantic categories, with $M' = \#S$.

We are then looking at the problem of whether it is possible to project isomorphically a subvariety $X_{\mathcal{C}}$ of \mathbf{P}^{M-1} that contains the points $\Pi_{\mathcal{C}}$ (and possibly the

collection of paths $\Gamma(T)$) to the quotient $\mathbf{P}^{M'-1}$. Again, there are strong restrictions on the existence of such isomorphically projectable subvarieties. For example, it is shown in [Ar01] that the only n -dimensional variety that can be isomorphically projected from $Gr(1, 2n + 1)$ to $Gr(1, n)$ is the Veronese variety, that is, the embedding of \mathbf{P}^n in $Gr(1, 2n + 1)$ via $O_{\mathbf{P}^n}(1)^{\oplus d}$.

When the variety X_C is not isomorphically projectable from \mathbf{P}^{M-1} to $\mathbf{P}^{M'-1}$, there is some loss of information in the semantic vectors, when the identification of words according to semantic tags is performed. In such cases, which will be typical in view of the very restrictive condition of isomorphic projectability, one can describe the effect of the identification on semantic vectors by analyzing the change of topology in the polygonal path $\Gamma_C = \cup_{T \in \mathcal{C}} \Gamma(T)$. We describe ways of approaching computationally such topology changes.

4.4. Persistent Topology. It was understood in recent years that clusters of data points can exhibit interesting topological structure that can be useful in analyzing large data set, see [Ca09] for a general introduction and overview of the field of persistent topology. Applications of persistent topology to Linguistics were recently discussed in [PorGhGuCLDMar15].

Given a set Π of points in a metric space, one considers a family of simplicial complexes, parameterized by a real number $\epsilon > 0$, the so called Vietoris–Rips complexes $R(\Pi, \epsilon)$. Here the n -th term $R_n(\Pi, \epsilon)$ is the vector space spanned by all the unordered $(n + 1)$ -tuples of points in Π where all pairs have distance at most ϵ . There are inclusion maps $R(\Pi, \epsilon_1) \hookrightarrow R(\Pi, \epsilon_2)$ when $\epsilon_1 < \epsilon_2$. These induce maps in homology $H_n(R(\Pi, \epsilon_1)) \rightarrow H_n(R(\Pi, \epsilon_2))$. In analyzing the dependence on ϵ of the ranks of these homology groups one discards as “noise” those generators that arise and disappear within a small range of values of ϵ , while one regards those generators that persist for sufficiently long intervals of values of ϵ , the “persistent generators”, as signaling the presence of actual structure in the data.

Persistent topology of the set Π_C in the Grassmannian. Persistent topology can also be used to enrich the semantic comparison of different texts, when we assign to each text in a corpus a point in a Grassmannian or in a flag variety, as discussed above. The simplest level of comparison would be to cluster together the points corresponding to the various texts by separating them into groups according to the relative distances in the ambient metric. The resulting groups are dependent upon the scale of the neighborhoods of points, and the number of different groups of semantic similarity correspond to the rank of the zeroth order persistent homology

of the Vietoris–Rips complex. Thus, more refined information about how texts cluster together by semantic similarity is obtained by additionally considering also the first and higher dimensional persistent homology.

We consider, as above, a projection $\pi_{M,M'} : Gr(N, M) \rightarrow Gr(N, M')$ and the image $\pi_{M,M'}(\Pi_{\mathcal{C}})$. In the case where the set of points $\Pi_{\mathcal{C}}$ does not fit on an interpolating variety that is isomorphically projectable, we can analyze the change in the semantic proximity of texts by analyzing the differences between the persistent topology of $\Pi_{\mathcal{C}}$ and of $\pi_{M,M'}(\Pi_{\mathcal{C}})$. This can be seen by computing the number of persistent generators, in various degrees, of the homology of the respective Vietoris–Rips complexes. The case of points in flag varieties can be treated analogously to the case of points in Grassmannians.

Persistent topology and paths in \mathbf{P}^{M-1} . In a similar manner, one can use persistent topology to analyze syntactic proximity of texts in the point of view where we assign to each text in a corpus a path in projective space \mathbf{P}^{M-1} . In this case, we again associate to a corpus \mathcal{C} a simplicial complex, where the zero-cells are all the points $p_k(T) \in \mathbf{P}^{M-1}$, for all texts $T \in \mathcal{C}$, and all the one-cells are the geodesic arcs connecting consecutive pairs of points $p_k(T)$ and $p_{k+1}(T)$. The higher dimensional skeleta are then constructed as in the Vietoris–Rips complex, by adding an n -dimensional simplex whenever an $n + 1$ -tuple of points $\{p_{k_0}(T_0), \dots, p_{k_n}(T_n)\}$ where the geodesic distances between all pairs of these points are less than a fixed scale ϵ . This may require introducing additional one-cells in the complex.

As in the case of points in Grassmannians and flag varieties, when we consider a projection $\pi_{M,M'} : \mathbf{P}^{M-1} \rightarrow \mathbf{P}^{M'-1}$, we can study the effect of the projection on the persistent topology of the set of paths $\Gamma_{\mathcal{C}} = \cup_{T \in \mathcal{C}} \Gamma(T)$ and its image $\pi_{M,M'}(\Gamma_{\mathcal{C}})$, by associating complexes as indicated above to $\Gamma_{\mathcal{C}}$ and $\pi_{M,M'}(\Gamma_{\mathcal{C}})$ and comparing generators of the respective persistent homologies.

4.5. Unsupervised Learning. In the case of unsupervised learning, a grouping corresponding to “sense discrimination” is obtained solely on the basis of the semantic vectors and the position of the corresponding points in the ambient variety, without any external tagging of words by semantic categories. In the setting of unsupervised learning, the grouping together of subsets of the M lexical dimensions into putative semantic categories is itself performed solely on the basis of the semantic vectors. A simple way to search for semantic relatedness in an unsupervised context is to identify frequent co-occurrences within the same contexts (see Section 2.4 of [TuPa10]). Many co-occurrences arise for purely syntactic reasons,

but those tend to be between words that belong to different parts of speech, while co-occurrences that carry semantic significance are more often found between words in the same part of speech, see [BuHi06], [ChiaBRP90], [SchPe93], [TuPa10].

4.6. Syntactic dependence of semantic vectors. Clearly, the vectors $P_k(T) = (p_{ik})_{i \in D}$, associated to the contexts c_k in a text, depend on both syntactic and semantic information and there is a priori no obvious way to distinguish between the dependence on syntax and on semantics. However, a possible way to make these semantic vectors more syntax independent would be to consider a training corpus of different language translations of the same texts, with marked matching paragraphs and matching word dictionaries, and average the semantic vectors $P_k(T, L)$ over the set of languages L . This can be done either by simply averaging the vectors, or else by considering the corresponding points $p_k(T, L)$, for all languages L , in the fixed ambient \mathbf{P}^{M-1} , and replace them by the barycenter $\bar{p}_k(T)$ computed with respect to the Fubini-Study metric on \mathbf{P}^{M-1} . This has the effect of reducing the purely syntactic contribution, especially if the set of languages chosen contains languages with sufficiently different set of syntactic parameters. Of course, it is not possible to entirely decouple semantics from syntax, as the syntactic-semantic interface is very rich (see for example [Ha13], [Va05]), but this averaging method can at least partially reduce the influence of those effects that are due to syntax alone.

5. Geodesically convex neighborhoods and semantic spaces

In the setting above, we have associated to texts in a corpus a collection of points (or of paths) in an ambient geometric space (a Grassmannian, or a flag variety, or a projective space). We have also seen that, when we group together words in the lexicon by semantic categories, geometrically we look at how the set of points and paths behaves under a projection map of the ambient variety. In this section we use the same general geometric picture, and we consider coverings by convex open sets. These local neighborhoods correspond to grouping together texts by semantic similarity. The convexity property corresponds to the possibility of interpolation and will be compared in Section 7 with the approach of Gärdenfors on conceptual spaces as “meeting of minds”, cf. [Gä00], [Gä14], [WaGä13].

5.1. Geodesic convexity and good coverings. Recall that a subset $U \subset X$ in a Riemannian manifold X is said to be *geodesically convex* if for arbitrary points $p \neq p' \in U$ there is a distance minimizing geodesic arc connecting them that is

entirely contained in U . In particular, a geodesically convex U is topologically a contractible set. Moreover, a non-empty finite intersection of geodesically convex open sets U_i is also a geodesically convex open set. If X is compact, we can assume the number of open sets in such a covering to be finite. Their size (measured as the diameter) in such a covering is bounded. We then say that $\mathcal{U}_\epsilon = \{U_i(\epsilon)\}_{i=1}^n$ is a *good ϵ -covering* of the compact Riemannian manifold X , if the U_i are geodesically convex with $\epsilon = \max_i \{\text{diam}(U_i(\epsilon))\}$.

In particular, we consider such coverings for the Grassmannians $Gr(N, M')$, flag varieties $F(1, \dots, 1, M' - N)$, and projective spaces $\mathbf{P}^{M'-1}$, with the respective metrics discussed above, and where M' is the size of the semantic vocabulary, after semantic identifications have been performed on the initial lexical vocabulary of size $M \geq M'$, as discussed in the previous section. We view points $p_k(T)$ that lie within the same convex neighborhood $U_i(\epsilon)$ of a good ϵ -open covering by geodesically convex sets as being semantically related. In particular, we are interested in considering good ϵ -open coverings that are generated by starting with a collection $U_k(\epsilon, T)$ of geodesic balls of radius $\epsilon/2$ centered at the points $p_k(T)$ associated to a text T in a corpus. Consider the case where $p_k(T) \in \mathbf{P}^{M'-1}$. The cases of points in Grassmannians and flag varieties are analogous. We construct an ϵ -open covering of the ambient variety by starting with the collection $\{U_k(\epsilon, T) : k = 1, \dots, N(T); T \in \mathcal{C}\}$ and we complete it to an ϵ -open covering of $\mathbf{P}^{M'-1}$ by adding enough additional sets $U_i(\epsilon)$ covering the complement of $Y_{\mathcal{C}, \epsilon} := \cup_{k, T} U_k(\epsilon, T)$. We let \mathcal{U}_ϵ denote the resulting covering of $\mathbf{P}^{M'-1}$ and we write $\mathcal{U}_\epsilon(\mathcal{C}) \subset \mathcal{U}_\epsilon$ for the covering of $Y_{\mathcal{C}, \epsilon} \subset \mathbf{P}^{M'-1}$ by the $\{U_k(\epsilon, T) : k = 1, \dots, N(T); T \in \mathcal{C}\}$.

As it is customary in topology, we can associate to a given good ϵ -covering $\mathcal{U}_\epsilon(\mathcal{C})$ the simplicial complex given by its Čech complex $\mathcal{N}_\star(\mathcal{C}, \epsilon) := \mathcal{N}_\star(\mathcal{U}_\epsilon(\mathcal{C}))$, with geometric realization $\mathcal{N}(\mathcal{C}, \epsilon) := |\mathcal{N}_\star(\mathcal{C}, \epsilon)|$. Note that, while the geometric realization of the Čech complex of the full covering \mathcal{U}_ϵ of $\mathbf{P}^{M'-1}$ is just homotopy equivalent to $\mathbf{P}^{M'-1}$ (see [Se68] and also [Du1] for a generalization), the geometric realizations $\mathcal{N}(\mathcal{C}, \epsilon)$ of the subcomplexes $\mathcal{U}_\epsilon(\mathcal{C})$ of \mathcal{U}_ϵ in general depend on the corpus \mathcal{C} and will in general not be homotopy equivalent to the ambient space.

One can then study, for a given corpus of texts \mathcal{C} , how the homotopy type, and invariants such as homology, of the simplicial space $\mathcal{N}(\mathcal{C}, \epsilon)$ vary with the scale ϵ . According to the usual approach of persistent topology, those features that change rapidly with the scale are attributed to random fluctuation, while persistent features can be identified with actual structures.

5.2. Geodesically convex neighborhoods, Čech complexes, and neural codes. As in the previous subsection, we consider simplicial complexes $\mathcal{N}_\star(\mathcal{C}, \epsilon)$ obtained as the Čech complex of the collection $\mathcal{U}_\epsilon(\mathcal{C})$ of the geodesically convex balls $U_k(\epsilon, T)$ of diameter ϵ around the points $p_k(T) \in \mathbf{P}^{M'-1}$, for $k = 1, \dots, N(T)$, the number of contexts in the text T and for T varying in a given corpus \mathcal{C} . Their geometric realizations are denoted, as above, by $\mathcal{N}(\mathcal{C}, \epsilon) = |\mathcal{N}_\star(\mathcal{C}, \epsilon)|$.

Following the approach of [CuItVCYo13], we associate a code $C = C(\mathcal{C}, \epsilon)$ to the collection $\mathcal{U}_\epsilon(\mathcal{C})$ of geodesically convex balls. This is a code $C \subset \{0, 1\}^m$, where $m = \sum_{T \in \mathcal{C}} N(T)$. Here we assume chosen an ordering of the texts $T \in \mathcal{C}$, with $n = \#\mathcal{C}$, so that we identify the set of contexts

$$\{c_1(T_1), \dots, c_{N(T_1)}(T_1), \dots, c_1(T_n), \dots, c_{N(T_n)}(T_n)\}$$

of the entire corpus \mathcal{C} with the set $\{1, \dots, m\}$. The code words $w \in C$ are those elements $w \in \{0, 1\}^m$ such that

$$\left(\bigcap_{i \in \text{supp}(w)} U_{k_i}(\epsilon, T_i) \right) \setminus \left(\bigcup_{j \notin \text{supp}(w)} U_{k_j}(\epsilon, T_j) \right) \neq \emptyset,$$

where $\text{supp}(w) = \{i \in \{1, \dots, m\} : w_i = 1\}$.

According to the “nerve theorem” ([Ha02], Corollary 4G.3), as discussed in [CuItVCYo13] and [Ma15], the homotopy type of the space $Y_{\mathcal{C}, \epsilon} = \cup_{k, T} U_k(\epsilon, T)$ is equal to the homotopy type of the nerve $\mathcal{N}(\mathcal{C}, \epsilon)$ of the complex $\mathcal{N}_\star(\mathcal{C}, \epsilon)$. In particular, the persistent homology of $Y_{\mathcal{C}, \epsilon}$ is the same as the persistent homology of $\mathcal{N}(\mathcal{C}, \epsilon)$.

This is the setting used in [CuItVCYo13] to reconstruct information about the topology of the stimulus space from knowledge of the associated neural code. Neural codes and the problem of how they encode the structure of the stimulus space have been studied extensively in neuroscience, especially in relation to vision (see [CuItVCYo13] and references therein). The study of neural codes in the linguistic setting is presently less extensive: neural codes for syntax, based on data of neurosurgical procedures, have been studied (see [BikSza09]). A detailed criticism of a possible linguistic approach to neurosemantics is given for instance in [Eli05], while a proposal for semantic representation of linguistic data via shared neural codes (for auditory, visual or somatosensory inputs) is analyzed in [Poe06].

We argue for a proposal of the simplicial complexes $\mathcal{N}_*(\mathcal{C}, \epsilon)$ and their persistent homotopy type as possible computational models of neural codes for neurosemantics, at least up to homotopy. Namely, instead of the usual approach to measuring semantic relatedness of texts on the basis of angular distances of semantic vectors, one can consider topological notions of relatedness and proximity, in terms of deformability and homotopy equivalence of the complexes $\mathcal{N}_*(\mathcal{C}, \epsilon)$.

6. Spectral decompositions and Riccati flows

6.1. Singular Value Decomposition. Typically, the word–document semantic matrices discussed above are very sparse, with often only a small percentage of entries being non–zero. It is known that this creates problems in measuring semantic similarity with the usual cosine method (see [TuPa10]), as the method easily assigns zero to non co–occurring words even though they are semantically related.

In order to circumvent this problem, one can perform a dimensional reduction based on a singular value decomposition (SVD). This represents the semantic matrix P as a product $U\Sigma V^T$, where U and V are, respectively, and $M \times M$ and an $N \times N$ unitary matrix and Σ is an $N \times M$ matrix with the singular values on the diagonal, of rank r equal to the rank of the original semantic matrix.

6.2. Latent Semantics. The technique known as “latent semantics” (see Section 4.3 of [TuPa10]) then considers truncations of the matrix $U\Sigma V^T$ to a rank $k < r$ approximation $U_k \Sigma_k V_k^T$ obtained by considering only the k largest singular values. This has the effect of creating a low-dimensional linear mapping between words and contexts, which reduces noise and improves the estimates of semantic similarity, or “discover latent meaning” in the terminology used in vector space semantics.

Thus, according to this procedure, the process of analyzing semantic relatedness based on the given word–context semantic matrix, involves a singular value decomposition and a truncation according to the largest singular values. We will see in the rest of this section that these operations also have a very natural geometric interpretation in terms of the geometry of projective spaces and Grassmannians.

6.3. Term co-occurrence matrix. In order to obtain the singular value decomposition and restrict to the largest k singular values, one considers the symmetric matrix $A = P^T P$, the term co-occurrence matrix, and its spectral decomposition. The truncations discussed above can then be obtained by applying power

methods to separate out the span of the eigenvectors of the largest k eigenvalues of $A = P^T P$ from the complementary space. For further discussions of “semantic spectrum” and “eigenword” decomposition see for instance [DhFU15], [WiDa11].

6.4. Perron–Frobenius and Riccati equation. If there is only one top eigenvalue one can apply the usual Perron–Frobenius theory. Let $Sp(A) = \{\lambda_1, \dots, \lambda_N\}$ with $|\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_N|$. In the case we considered in the previous sections where $N \leq M$ and the rank is N , the matrix A determines an action on \mathbf{P}^{N-1} , and the sequence of points $x_m = A^m x_0$, for an assigned initial point $x_0 \in \mathbf{P}^{N-1}$, converges to the point in \mathbf{P}^{N-1} corresponding to the line spanned by the Perron–Frobenius eigenvector of A . Moreover, as discussed in [AmMa86], [MaAm92], in a local chart corresponding to vectors with first component equal to one, we have

$$A : x_m = \begin{pmatrix} 1 \\ y_m \end{pmatrix} \mapsto x_{m+1} = Ax_m = \begin{pmatrix} 1 \\ y_{m+1} \end{pmatrix},$$

with

$$y_{m+1} = \frac{A_3 + A_4 y_m}{A_1 + A_2 y_m}$$

where

$$A = \begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix},$$

where A_4 is an $(N-1) \times (N-1)$ -matrix and A_1 a number. The recursion relation of the sequence y_m is then given by

$$y_{m+1} - y_m = (A_3 + A_4 y_m - y_m A_1 - y_m A_2 y_m)(A_1 + A_2 y_m)^{-1}.$$

The above can be viewed as a discretization of the matrix Riccati equation

$$\frac{d}{dt}y(t) = A_3 + A_4 y(t) - y(t)A_1 - y(t)A_2 y(t),$$

in particular, both equations have the same stationary solutions given by solutions to

$$A_3 + A_4 y - yA_1 - yA_2 y = 0.$$

Thus, in order to find the limit $x = \lim_m x_m$, or equivalently the stationary solution $y_{m+1} = y_m$ of the difference equation above, one can consider the Riccati flow to the

same fixed point. For this reformulation of the Perron–Frobenius theory in terms of a matrix Riccati equation in a projective space, see [AmMa86], [MaAm92].

6.5. Latent Semantics and flows on Grassmannians. In a similar way, it is shown in [AmMa86], [MaAm92] that the selection of the span of the eigenvectors of the k largest eigenvalues of the matrix $A = P^r P$ can be performed dynamically in terms of a Riccati flow on the Grassmannian $G(k, N)$. More precisely, for a given k -dimensional vector space $V \in G(k, N)$ and a matrix $A \in GL_N$, we have $AV \in G(k, N)$ given by $AV = \{Av : v \in V\}$. Thus, given an initial point $V_0 \in G(k, N)$ one can consider the power sequence $V_{m+1} = AV_m$. If $\text{Spec}(A) = \{\lambda_1, \dots, \lambda_N\}$ with

$$|\lambda_1| \geq |\lambda_2| \geq |\lambda_k| > |\lambda_{k+1}| \geq \dots |\lambda_N|,$$

and U is the span of the eigenvectors corresponding to λ_i with $i = 1, \dots, k$, then the sequence of points V_m in $G(k, N)$ converge to the point corresponding to the space U , for every choice of initial V_0 with $V_0 \cap W = \{0\}$, where W is the span of the eigenvectors with eigenvalues λ_i with $i = k + 1, \dots, N$.

For a choice of complementary subspaces $U \in G(k, N)$ and $W \in G(N - k, N)$, and a morphism $L \in \text{Hom}(U, W)$, consider the element $U_L \in G(k, N)$ given by the subspace

$$U_L = \left\{ \begin{pmatrix} u \\ Lu \end{pmatrix} \mid u \in U \right\} \subset U \oplus W.$$

If the matrix A in the decomposition $U \oplus W$ has the form

$$A = \begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix}$$

then

$$AU_L = U_{(A_3 + A_4 L)(A_1 + A_2 L)^{-1}}$$

in this local chart on the Grassmannian $G(k, N)$. Thus, one obtains a corresponding sequence

$$L_{m+1} = (A_3 + A_4 L_m)(A_1 + A_2 L_m)^{-1}$$

which can be written as a difference equation

$$L_{m+1} - L_m = (A_3 + A_4 L_m - L_m A_1 - L_m A_2 L_m)(A_1 + A_2 L_m)^{-1}.$$

As before, the stationary solutions can be equivalently obtained as the stationary points of the matrix Riccati flow

$$\frac{d}{dt}L(t) = A_3 + A_4L(t) - L(t)A_1 - L(t)A_2L(t).$$

This shows that the latent semantics method based on singular value decomposition and truncation to the top k singular values for P can be reformulated in terms of a geometric flow on a Grassmannian.

7. Relation to Gärdenfors’ “meeting of minds”

7.1. Where the minds meet. In [Gä14], Gärdenfors developed an approach to semantic spaces based on the metaphor “meeting of minds” (see [WaGä13]) and on models of “conceptual spaces” developed in [Gä00]. The main idea is that meaning is emergent in communication (see Section 5.1 of [Gä14]). Typically, coming to a common understanding of meaning in communication is seen as a fixed point problem taking place in a convex space which describes some configuration domain, such as colors, some kind of actions, etc. Communication is modeled in terms of a partitioning of this domain determined by the transmitter and a sample set of points in the domain obtained by the received, and the common understanding is achieved by the construction of a Voronoi partition common to both sets of points, see Section 5.4.1 of [Gä14].

In our setting, the geometry of semantic spaces is not dictated by conceptual spaces determined by preassigned external semantic categories as in [Gä14], but rather the geometry of an ambient space (a Grassmannian, or a set of paths in a projective space) built out of corpora of texts and the frequencies of occurrences of lexemes in contexts of these texts. However, we can still develop an approach to communication as a fixed point problem leading to a common semantic interpretation between different users, which resembles, in a different geometric setting, the “meeting of minds” approach of Gärdenfors.

Consider a set A of different users. All users have access to the same dictionary D of lexemes, while each user $\alpha \in A$ has access to a certain corpus of texts \mathcal{C}_α , and derives semantic information from the analysis of occurrences of the words of D in the contexts of the texts $T \in \mathcal{C}_\alpha$. Thus, each user $\alpha \in A$ obtains a matrix $P_\alpha(T)$ of semantic vectors for each text $T \in \mathcal{C}_\alpha$. Assuming each user has analyzed the entire corpus \mathcal{C}_α , and used the information available in all texts $T \in \mathcal{C}_\alpha$ to obtain semantic

information, we obtain, for each lexeme $w_k \in D$ and for each user $\alpha \in A$, a semantic vector $P_{\alpha,k} = (p_{\alpha,ki})$, where the index i ranges over all the contexts $c_i(T)$ of all the texts T , listed in a given order in the corpus \mathcal{C}_α . We can view all these semantic vectors $P_{\alpha,k}$ inside a larger vector space that corresponds to the union $\mathcal{C} = \cup_\alpha \mathcal{C}_\alpha$, where we add zero entries to the vector $P_{\alpha,k}$ whenever a certain text T in some corpus \mathcal{C}_β is not also contained in \mathcal{C}_α . In this way, for a given lexeme $w_k \in D$, the different users arrive at somewhat different semantic interpretations, depending on the different texts they had access to. This difference is measured by the different position of the vectors $P_{\alpha,k}$ in this ambient space. In a similar way, if we consider the entire dictionary, or just some subset of lexemes, we obtain for each user a different semantic matrix P_α , computed as above over all texts T in $\mathcal{C} = \cup_\alpha \mathcal{C}_\alpha$. As before, we regard these matrices as points p_α in a Grassmannian $Gr(M, N)$, where M is the number of lexemes considered and N is the overall number of contexts in all the texts in the entire union \mathcal{C} of corpora. Here we typically are in the situation where we are seeking a common semantic understanding of a small number of lexemes using a large number of context and corpora, hence $M < N$.

Given this finite collection $\{p_\alpha\}_{\alpha \in A}$ of points in a Grassmannian $Gr(M, N)$, which represents the different positions in semantic space the different users arrived at by analyzing the occurrence of the same list of lexemes in the corpora available to them, we need a simple geometric procedure that arrives to a common position in semantic space and that can be implemented interactively as a sequence of approximations. A simple such procedure consists of taking the geodesic barycenter of the set $\{p_\alpha\}_{\alpha \in A}$. In fact, more generally one can consider a weighted distribution of the points p_α , where each p_α is assigned a weight $\lambda_\alpha \geq 0$ with $\sum_\alpha \lambda_\alpha = 1$. The additional information contained in the weights λ_α can be some a priori knowledge of the higher reliability or relevance of some corpora \mathcal{C}_α with respect to others, which would make the semantic matrix P_α obtained by some user more reliable than that obtained by some other user. Given the set $\{p_\alpha\}$ in $Gr(M, N)$ and the respective weights λ_α , the barycenter p_B is determined by the condition

$$\sum_\alpha \lambda_\alpha \delta^2(p_\alpha, p_B) = \min_{p \in Gr(M, N)} \left\{ \sum_\alpha \lambda_\alpha \delta^2(p_\alpha, p) \right\},$$

where $\delta(x, y)$ is the geodesic distance. Assuming that all the points p_α lie sufficiently close to each other (as would be the case if there is enough overlap between the corpora available to different users) so that they are contained in a single geodesically

convex neighborhood $U \subset Gr(M, N)$, the potential function

$$V(p) = \sum_{\alpha} \lambda_{\alpha} \delta^2(p_{\alpha}, p)$$

is a strictly convex function on the neighborhood U and has therefore a unique minimum. The barycenter is then the point p_B where $V(p)$ achieves its minimum.

It can be also described as the unique fixed point of the map $p \mapsto p - h \nabla V(p)$, where $\nabla V = g(dV, \cdot)$, g being the Riemannian metric tensor, and h is a finite increment in a discretized gradient descent. Recursively, p_B is then approximated by $p_{k+1} = p_k - h \nabla V(p_k)$.

In a similar way, one can consider simplicial Vietoris–Rips complexes $\mathcal{N}_*(\mathcal{C}_{\alpha}, \epsilon)$ obtained by different users based on different corpora \mathcal{C}_{α} . After again considering them inside a larger common projective space, one can construct a new complex which is their common barycentric subdivision. The homotopy type and persistent homology of the resulting complex can then be treated as a model of the “meeting of minds” in our setting.

8. Semantic vectors, Zipf’s law, and Kolmogorov complexity

8.1. Zipf’s law. As observed in [Lowe01], constructions of semantic spaces based on semantic vectors should take into consideration the fact that the distribution of linguistic data is skewed towards high count data, according to the empirical Zipf’s law.

Given a corpus of texts \mathcal{C} and a word w (in the sense of a lexeme from a dictionary of words), let $F_{\mathcal{C}}(w)$ denote the number of tokens of the given word that appear in the corpus, and $F_{\mathcal{C}}(w)/N$ the relative frequencies, where $N = \#\mathcal{C}$ is the size of the corpus. Let $\{w_k\}$ be an enumeration of the dictionary words by decreasing frequencies. Then Zipf’s law states that $\log(F_{\mathcal{C}}(w_k)) = \kappa(N) - B \log(k)$, for a constant $\kappa(N)$ depending on the corpus size and with the power law B satisfying $B \sim 1$. It was shown in [Ma13] that if one postulates that, in Zipf’s original explanation as “minimization of effort”, the word “effort” means Kolmogorov’s complexity, then Zipf’s law with exponent 1 becomes a consequence of properties of the related universal Levin probability distribution.

In the construction of semantic spaces, when one counts co–occurrences of words in given contexts with certain given vocabulary lexemes, one encounters a situation

where low frequency words may be more significant for semantic association, but produce very sparse semantic matrices, while high frequency words provide more reliable statistics, but are less significant in determining semantic association, as they tend to appear in almost every context. Semantic vectors based on low frequency words will have high variance, and Zipf's law predicts that the amount of additional data required in order to reduce the variability is expressed by a power law relation.

8.2. Latent semantic analysis. In *latent semantic analysis* this phenomenon is accounted for by introducing weights assigned to the vocabulary entries, so that the estimated probability (frequency) $p_c(w, \ell)$ of co-occurrence of a given word w with a given lexeme ℓ in a context c is weighted by $S(\ell)^{-1} \log(1 + p_c(w, \ell))$, where the denominator is given by the entropy $S(\ell) = -\sum_{c \in C(T)} p_c(\ell) \log(p_c(\ell))$, with $p_c(\ell)$ the probability of occurrence of ℓ in the context c . In this way, if ℓ is equally distributed in all contexts, as one expects for the most frequent words, the entropy is maximal and the weighted co-occurrence is less significant, while if ℓ is likely to occur only in a smaller number of contexts the co-occurrence is weighted more, as more semantically significant. Other methods that can be used for taking into account the effects due to Zipf's law and the different semantic significance of words with different frequencies are surveyed in [Lowe01].

This type of considerations based on Zipf's law apply to any suitable construction of semantic spaces, including the geometric construction we discussed in the previous sections. In particular, one can similarly consider introducing appropriate weights in the construction of the semantic matrix $P(T)$ of a text, that we discussed before, so that, in addition to counting occurrences in contexts $c \in C(T)$, one also keeps into account how uniform or non-uniform the distribution over contexts is, measured in terms of the Shannon entropy of the resulting probability distribution.

References

- [AlGo95] H. Alt, M. Godau, *Computing the Fréchet distance between two polygonal curves*, Int. J. Comput. Geom. Appl. Vol.5 (1995) 75–91.
- [AmMa86] G. Ammar, C. Martin, *The geometry of matrix eigenvalue methods*, Acta Appl. Math. Vol.5 (1986) N.3, 239–278.
- [Ar01] E. Arrondo, *Projections of Grassmannians of lines and characterization of Veronese varieties*, J. Alg. Geom., Vol.1 (2001) 165–192.

[ArRa05] E. Arrondo, R. Paoletti, *Characterization of Veronese varieties via projections in Grassmannians*, in “Projective Varieties with Unexpected Properties: A Volume in Memory of Giuseppe Veronese” (C. Ciliberto, A.V. Geramita, B. Harbourne, R.M. Miró-Roig, K. Ranestad Eds.) Walter de Gruyter, 2005.

[BikSza09] D. Bickerton, E. Szathmáry, *Biological Foundations and Origin of Syntax*, MIT Press, 2009.

[BuHi06] A. Budanitsky, G. Hirst, *Evaluating WordNet-based measures of semantic distance*, Computational Linguistics, Vol.32 (2006) N.1, 13–47.

[Ca09] G. Carlsson. *Topology and data*. Bull. AMS, vol. 46 (2009), no. 2, 255–308.

[ChiaBRP90] C. Chiarello, C. Burgess, L. Richards, A. Pollock, *Semantic and associative priming in the cerebral hemispheres: Some words do, some words don’t ... sometimes, some places*, Brain and Language, Vol.38 (1990) 7–104.

[CuIt08] C. Curto, V. Itskov. *Cell Groups Reveal Structure of Stimulus Space*. PLoS Computational Biology, vol. 4, issue 10, October 2008, 13 pp. (available online).

[CuItVCYo13] C. Curto, V. Itskov, A. Veliz-Cuba, N. Youngs. *The neural ring: An algebraic tool for analysing the intrinsic structure of neural codes*. Bull. Math. Biology, 75(9), pp. 1571–1611, 2013.

[Dem88] J.P. Demailly, *Vanishing theorems for tensor powers of a positive vector bundle*, in “Geometry and analysis on manifolds” (Katata/Kyoto, 1987), pp.86–105, Lecture Notes in Math., Vol.1339, Springer, 1988.

[DhFU15] P. Dhillon, D.P. Foster, L.H. Ungar, *Eigenwords: Spectral Word Embeddings*, Journal of Machine Learning Research 16 (2015)

[DuI] D. Dugger, D.C. Isaksen, *Hypercovers in topology*, preprint <http://www.math.uiuc.edu/K-theory/0528/>

[Eli05] C. Eliasmith, *Neurosemantics and categories*, in “Handbook of Categorization in Cognitive Science”, pp.1035–1054, Elsevier, 2005.

[Gä00] P. Gärdenfors. *Conceptual spaces: The geometry of thought*. Cambridge, Mass. MIT Press, 2000.

[Gä14] P. Gärdenfors. *Geometry of Meaning: Semantics Based on Conceptual Spaces*. Cambridge, Mass. MIT Press, 2014, 343+xii pp.

- [Gri74] P. Griffiths, *On Cartan's method of Lie groups and moving frames as applied to uniqueness and existence questions in differential geometry*, Duke Math. J., Vol.41 (1974) 775–814.
- [Gui68] P. Guiraud. *The semic matrices of meaning*. Social Science Information, 7(2), 1968, pp. 131–139.
- [Ha02] A. Hatcher. *Algebraic Topology*. CUP, Cambridge, 2002.
- [Ha13] M. Hackl, *The syntax-semantics interface*, Lingua, Vol. 130 (2013) 66–87.
- [InLe04] P. Indefrey, W. J. M. Levelt. *The spatial and temporal signatures of word production components*. Cognition 92, 2004, pp. 101–144.
- [JeLe94] J. D. Lescheniak, W. J. M. Levelt. *Word frequency effects in speech production: retrieval of syntactic information and of phonological form*. Journ. of Experimental Psychology: Learning, Memory and Cognition, 20, 1994, pp. 824–843.
- [Li] L. Lica. *The Distinction between WHICH and THAT. With Diagrams*. <http://home.earthlink.net/~llica/wichthat.htm>
- [Lowe01] W. Lowe, *Towards a theory of semantic space*, in “Proceedings of the 23rd Conference of the Cognitive Science Society” (2001), pp. 576–581.
- [MaAm92] C. Martin, G. Ammar, *The geometry of the matrix Riccati equation and associated eigenvalue methods*, in “The Riccati equation”, pp.113–126, Comm. Control Engrg. Ser., Springer, 1991.
- [Ma13] Yu. I. Manin. *Zipf's law and L. Levin's probability distributions*. Functional Analysis and its Applications, vol. 48, no. 2, 2014. DOI 10.107/s10688-014-0052-1. Preprint arXiv:1301.0427
- [Ma15] Yu. I. Manin. *Neural codes and homotopy types: mathematical models of place field recognition*. Moscow Math. Journal, vol. 15, Oct.–Dec. 2015, pp. 1–8 . arXiv:1501.00897
- [Man12] D. Yu. Manin. *The right word in the left place: Measuring lexical foregrounding in poetry and prose*. www.researchgate.net
- [MaSch99] C.D. Manning, H. Schuetze, *Foundations of statistical natural language processing*, MIT Press, 1999.
- [Me16] I. Mel'čuk. *Language: from Meaning to Text..* Ed. by D. Beck. Moscow & Boston, 2016.
- [Poe06] D. Poeppel, *Language: Specifying the Site of Modality-Independent Meaning*, Current Biology, Vol.16 (2006) N.21, R930–R932.

- [PorGhGuCLDMar15] A. Port, I. Gheorghita, D. Guth, J.M Clark, C. Liang, S. Dasu, M. Marcolli, *Persistent Topology of Syntax*, arXiv:1507.05134
- [Pos06] A. Postnikov. *Total positivity, Grassmannians and networks*, preprint arXiv:math/0609764 [math.CO].
- [SchPe93] H. Schütze, J. Pedersen, (1993). *A vector model for syntagmatic and paradigmatic relatedness*, in “Making Sense of Words”, pp. 104–113, Oxford, 1993
- [Se68] G. Segal, *Classifying spaces and spectral sequences*, Inst. Hautes Etudes Sci. Publ. Math. Vol.34 (1968) 105–112.
- [TuPa10] P.D. Turney, P. Pantel, *From frequency to meaning: vector space models of semantics*, Journal of Artificial Intelligence Research, Vol.37 (2010) 141–188.
- [Va05] R.D. van Valin, Jr. *Exploring the Syntax-Semantics Interface*, Cambridge University Press, 2005.
- [WaGä13] M. Warglien, P. Gärdenfors, *Semantics, conceptual spaces and the meeting of minds*, Synthese, Vol.190 (2013) N.12, 2165–2193.
- [WiDa11] P. Wittek, S. Darányi, *Spectral Composition of Semantic Spaces*, in “Quantum Interaction”, Lecture Notes in Computer Science, Vol.7052 (2011), pp. 60–70.
- [Yo14] N. E. Youngs. *The neural ring: using algebraic geometry to analyse neural rings*. arXiv:1409.2544 [q-bio.NC], 108 pp.