

WORD CO-OCCURRENCE AND SIMILARITY IN MEANING

Some Methodological Issues

Franco Lancia

E-mail: franco.lancia@tlab.it

in Salvatore S., Valsiner J. (eds.), *Mind as Infinite Dimensionality*,
Roma, Edizioni Carlo Amore (forthcoming, 2008)

ABSTRACT

Starting from the observation that word co-occurrence analysis needs to be anchored to the theory of meaning, various issues are discussed with a view to understand what happens when the words become numbers and the software outputs (i.e. tables and charts) become texts to be interpreted. In particular, with reference to the representation of the word co-occurrences into vector-space model, linguistic and semiotic theories are presented as tools for discussing assertions as “two (or more) words that tend to occur in similar linguistic contexts (i.e. to have similar co-occurrence patterns) tend to resemble each other in meaning”. Two main references are used: to structural linguistics (particularly Z.S. Harris, L. Hjelmslev and A. Greimas) and to semiotics (particularly C.S. Peirce and U. Eco), considering meaning either within the structural relationships between expression and content forms or as result of abductive inference. Finally, looking at software outputs as multi-semiotic objects, i.e. like a sort of texts to be interpreted, the discussion addresses hermeneutic questions.

1. Introduction

Within social sciences, word co-occurrence analysis is widely used in various forms of research concerning the domains of content analysis, text mining, construction of thesauri and ontologies, etc.. In general, its aim is to find similarities in meaning between word pairs and/or similarities in meaning among/within word patterns, also in order to discover latent structures of mental and social representations.

My work experience in this field, both as researcher in cultural psychology and as software architect, leads me to think that many methodological issues, whether concerning the definition of object studied (i.e. word-co-occurrence), the choice of specific techniques for data analysis or the interpretation of results, are very difficult to be resolved without taking into account a theory of meaning which works as

a theory of method as well.

In particular, assuming that in order to achieve their aims researchers use statistical tools, from my point of view the reference to a theory of meaning is a prerequisite in accounting for two main processes, one upstream and the other downstream of algorithm applications. That is:

- a) the way of arranging raw data (i.e. words and texts) into data matrices;
- b) the way of interpreting the outputs (i.e. tables and graphs).

In other words, this theory is a tool for understanding what happens when words become numbers and tables and/or the graphs become *texts* to be interpreted.

To start with, let us consider the first question, i.e. the involvement of the theory of meaning in relation to the representation of raw data into data matrices. I will refer mainly to vector-space model, which allows us to represent each context unit as a vector including all the words co-occurring within it and, at the same time, allows us to represent each word as a vector including all the context units in which it occurs. Then, using appropriate algorithms, it is possible to examine the words within “semantic spaces”.

As the following analysis process depends on how each matrix has been built, each researcher, according to his specific goals, each time has to decide:

- a) which objects (i.e. context units) to put into rows;
- b) which features (i.e. words) to put into columns;
- c) which values (i.e. numbers) to put into cells.

For example, the context units (a) can be “word windows” (or “context windows”) of fixed length, sentences, paragraphs or entire documents. The word list (b) can include the content words only (e.g. nouns, verbs, adjectives) or all the words with a threshold value of occurrence; the same words can be lemmatized or not, the multiword expressions (e.g. “United States”) can be detected or not, and so on. Finally different values (c) can be used either to indicate if the *j*-word is present (“1”) or absent (“0”) within the *i*-context or to indicate how many times (e.g. “5”) is in it.

Now let’s suppose that we have built a data matrix in which each row-vector represents a different context unit (c_1, c_2, \dots, c_n) corresponding to a sentence or paragraph, and each column-vector represents a different word (w_1, w_2, \dots, w_n). By coding the presence (“1”) and the absence (“0”) of each word within each context unit

we can obtain a representation as in Fig. 1.1.¹

Note that, by means of a simple transformation, the same data can be represented within a square matrix (Fig. 1.2)² in which the words are row and column headings while each cell contains the number of context units in which the word w_i co-occurs with the word w_j ; but, to illustrate the argument, we can refer to the representation in Fig. 1.1.

| | w_1 | w_2 | w_3 | ... | w_m |
|-------|-------|-------|-------|------|-------|
| c_1 | 0 | 1 | 1 | ... | 0 |
| c_2 | 1 | 1 | 0 | ... | 1 |
| c_3 | 1 | 0 | 1 | ... | 0 |
| | | | | | |
| c_n | 0 | 1 | 0 | ... | 1 |

Fig. 1.1: Rectangular co-occurrence matrix

| | w_1 | w_2 | w_3 | | w_m |
|-------|-------|-------|-------|------|-------|
| w_1 | | 5 | 13 | | 8 |
| w_2 | 5 | | 2 | | 11 |
| w_3 | 13 | 2 | | | 6 |
| | | | | | |
| w_m | 8 | 11 | 6 | | |

Fig. 1.2: Square co-occurrence matrix

As often in this kind of matrix the word-columns are hundreds (or thousands), for its analysis multidimensional methods which perform a dimensional reduction are required.

The logic of this process is shown in the following pictures concerning the analysis of a matrix “A” (see Fig. 1.3) consisting of 20 rows (i.e. the context units labeled with numbers) and 10 columns (i.e. the content words labeled with letters). By using different techniques, the same matrix can be transformed and represented in different ways. Just to quote the most popular:

- by using Correspondence Analysis we can obtain a table reordered by rows and columns (see Fig. 1.4)³ and two graphs in which, in a bi-dimensional space, the row points (see Fig. 1.5) and the column points (see Fig. 1.6) are plotted;
- by using some distance measures we can obtain a similarity or dissimilarity matrix (see Fig. 1.7 and 1.8) and plot it by using a Multidimensional Scaling (see Fig. 1.9).

¹ We can also use a representation that weights each word on its IDF (inverse document frequency) and in which we scale each row to have Euclidian norm equal to 1.

² A similar way of representing data has been used by Osgood (1959), although in his “contingency analysis” each column-vector did not represent a word but a “content category”.

³ In this case, rows and columns have been reordered by using the coordinates of the first factor obtained by Correspondence Analysis. An analogous result can be obtained by using some clustering techniques.

| | A | B | C | D | E | F | G | H | I | J |
|----|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 2 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 |
| 3 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |
| 4 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| 5 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 |
| 6 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 7 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 8 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| 9 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 10 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| 11 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 12 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 13 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 14 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| 15 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 16 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 17 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |
| 18 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| 19 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| 20 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |

Fig. 1.3: Matrix "A"

| | F | D | I | E | B | J | H | G | C | A |
|----|---|---|---|---|---|---|---|---|---|---|
| 4 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| 8 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| 19 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| 10 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| 14 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| 18 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| 3 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| 17 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| 20 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| 7 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 12 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 15 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 2 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 5 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 16 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Fig. 1.4: Matrix "A" reordered

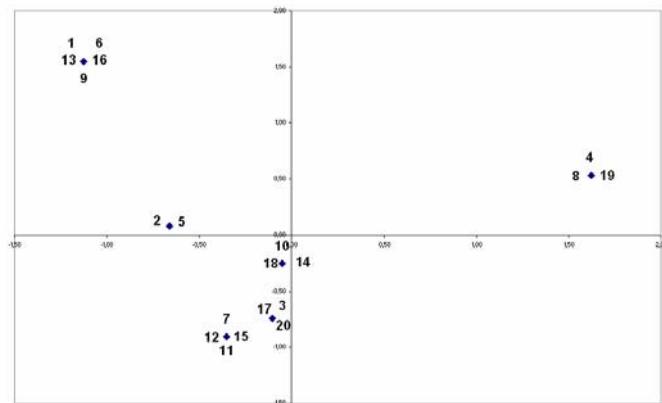


Fig. 1.5: Graphical display of the row profiles of Matrix "A".

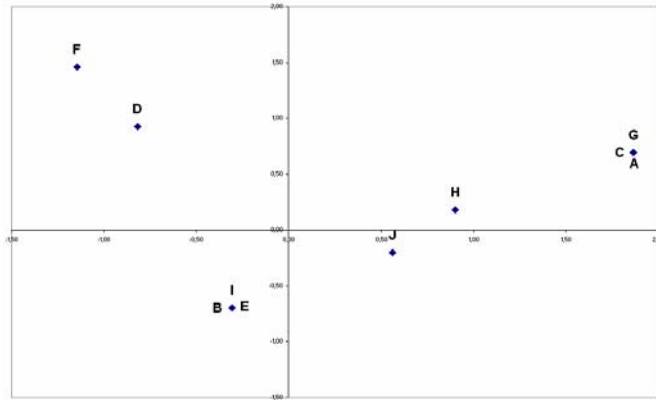


Fig. 1.6: Graphical display of the column profiles of Matrix "A".

| | A | B | C | D | E | F | G | H | I | J |
|---|------|------|------|------|------|------|------|------|------|------|
| A | | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 | 1,00 | 0,71 | 0,00 | 0,58 |
| B | 0,00 | | 0,00 | 0,46 | 1,00 | 0,22 | 0,00 | 0,35 | 1,00 | 0,58 |
| C | 1,00 | 0,00 | | 0,00 | 0,00 | 0,00 | 1,00 | 0,71 | 0,00 | 0,58 |
| D | 0,00 | 0,46 | 0,00 | | 0,46 | 0,84 | 0,00 | 0,39 | 0,46 | 0,32 |
| E | 0,00 | 1,00 | 0,00 | 0,46 | | 0,22 | 0,00 | 0,35 | 1,00 | 0,58 |
| F | 0,00 | 0,22 | 0,00 | 0,84 | 0,22 | | 0,00 | 0,00 | 0,22 | 0,00 |
| G | 1,00 | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 | | 0,71 | 0,00 | 0,58 |
| H | 0,71 | 0,35 | 0,71 | 0,39 | 0,35 | 0,00 | 0,71 | | 0,35 | 0,82 |
| I | 0,00 | 1,00 | 0,00 | 0,46 | 1,00 | 0,22 | 0,00 | 0,35 | | 0,58 |
| J | 0,58 | 0,58 | 0,58 | 0,32 | 0,58 | 0,00 | 0,58 | 0,82 | 0,58 | |

Fig. 1.7: Similarity Matrix

| | A | B | C | D | E | F | G | H | I | J |
|---|------|------|------|------|------|------|------|------|------|------|
| A | | 3,87 | 0,00 | 3,61 | 3,87 | 3,16 | 0,00 | 1,73 | 3,87 | 2,45 |
| B | 3,87 | | 3,87 | 3,46 | 0,00 | 3,87 | 3,87 | 3,46 | 0,00 | 3,00 |
| C | 0,00 | 3,87 | | 3,61 | 3,87 | 3,16 | 0,00 | 1,73 | 3,87 | 2,45 |
| D | 3,61 | 3,46 | 3,61 | | 3,46 | 1,73 | 3,61 | 3,16 | 3,46 | 3,61 |
| E | 3,87 | 0,00 | 3,87 | 3,46 | | 3,87 | 3,87 | 3,46 | 0,00 | 3,00 |
| F | 3,16 | 3,87 | 3,16 | 1,73 | 3,87 | | 3,16 | 3,61 | 3,87 | 4,00 |
| G | 0,00 | 3,87 | 0,00 | 3,61 | 3,87 | 3,16 | | 1,73 | 3,87 | 2,45 |
| H | 1,73 | 3,46 | 1,73 | 3,16 | 3,46 | 3,61 | 1,73 | | 3,46 | 1,73 |
| I | 3,87 | 0,00 | 3,87 | 3,46 | 0,00 | 3,87 | 3,87 | 3,46 | | 3,00 |
| J | 2,45 | 3,00 | 2,45 | 3,61 | 3,00 | 4,00 | 2,45 | 1,73 | 3,00 | |

Fig. 1.8: Dissimilarity Matrix

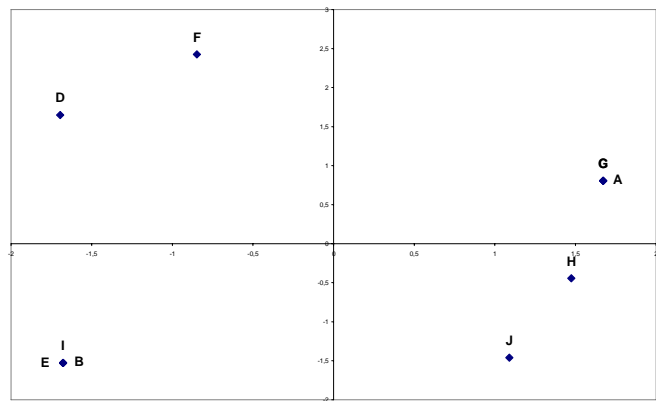


Fig. 1.9: MDS map

The reason why these kinds of outputs are so popular depends on the fact that multidimensional analysis allows us to represent the entire structure of data matrices and allows us to discover *new* information patterns by highlighting similarities and differences between objects (i.e. rows) or features (i.e. columns). Nevertheless all the pictures illustrated above are often used as a scientific evidence of banal assertions like the following:

two (or more) words that tend to occur in similar linguistic contexts (i.e. to have similar co-occurrence patterns), tend to be positioned closer together in semantic space.

We can notice that the evidence involved in this (i.e. the closeness in semantic space) is only produced (i.e. explicable) by the application of certain statistical algorithms and that, scientifically speaking, the “semantic” character of the space considered is irrelevant. In fact, the matrix “A” (see Fig 1.3) could contain “co-occurrences” of animal species within ecosystems, of molecules within drugs, of traits within images, and so on. In other words, this kind of assertion concerns the general problem of pattern recognition and its semantic connotation is just an idiosyncratic accident. As we can expect that all the elements which tend to co-occur in similar *x*-contexts tend to be positioned closer together in a *x*-space, if there were no instrumental mediation (i.e. the use of statistical tools), the same assertion could be considered a sort of tautology.

From my point of view, even though within a research project the choice of statistical tools has a strategic relevance, the most important decision involving theory of meaning concerns the transformation of texts into data matrices. In fact, not by chance, this decision is often used to mark the ideological and artificial watershed between “qualitative” and “quantitative” methods.

As the way how texts are arranged into a data matrix (Fig. 1.3) requires a specific kind of data transformation (e.g. words and context units represented by zero-one vectors), is this kind of reduction, characterized by a specific coding system which allows us to grasp only some specific similarities and differences, which need to be discussed first.

For example the “18” row-vector of matrix “A” (Fig. 1.3) could represent the following utterance:

“Did you read this book? It’s the latest work by Harris”.

More precisely, arranged in alphabetical order, using a sort of lemmatization (e.g. “latest” becomes “late”) and including the content words only, its representa-

tion could be as follows:

| "appreciate" | "book" | "good" | "Harris" | "late" | "moment" | "private" | "read" | "you" | "work" |
|--------------|---------------|--------|-----------------|---------------|----------|-----------|---------------|--------------|---------------|
| 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |

Fig. 1.10: Representation of an utterance

The question is: using a similar representation, in which each word has a sole attribute (i.e. it is, together with other words, an "element" of the same set), which kind of meaning can we extract and analyse? Or rather: what kind of meaning can we infer?

Also given that, in order to extract meaningful patterns, co-occurrence analysis requires comparison (similarities and differences) between a lot of context units, the question still stands.

Further question: within the "A" matrix there are two row-vectors ("10" and "14") which are identical to the "18"⁴. For example one of these ("10") could be the following:

"You can read the latest book of Harris to improve your work."

Even though from an intuitive point of view the two utterances are just concerning "similar" contents, as the matrix "A" is organized by a few number of columns (i.e. features) and by a specific coding system (i.e. the presence/absence of each word within each context unit) their row-vectors ("18" and "10") are identical.

The fact is that, in the "normal" life and in the "real" world, we can distinguish a lot of objects and features only because we use a wide and sophisticated system of categories (i.e. n-dimensional representations). On the contrary, the way how this kind of data matrices are arranged follows the logic of "possible words".

To stress this point let's suppose that, in order to build our representations of the world we could use just three categories (A, B, C). If so, in this kind of possible world, about each object present in it, we could just say that it has or doesn't have the "A", "B", "C" features.

For example, if our three features were "white" ("A"), "square" ("B") and "human" ("C"), when seeing a blackbird flying we can just say that it is something neither human nor square nor white. The same for describing a cat which purrs, a fire which blazes and an apple which falls from the tree. To be precise, in the possible world which we are referring to, all these three "objects" will be represented as identical. The same, as we know, happens within the logic of prejudice when two or more people are considered as identical just because they share one or two features (e.g. to be Hebrew or Muslim)⁵.

⁴ In the "real" text analysis, as the word-columns are hundred this case is very rare.

⁵ According to the principle of generalization proposed by Matte-Blanco (1975), the logic of unconscious and of the emotions treats the elements of each class as identical.

2. Word co-occurrence analysis as a de-construction and construction process

Going back to our issue, the use of techniques for data reduction doesn't imply think about the world within reduced dimensions; on the contrary, the phenomena which we are interested in exploring can be represented within low-dimensional spaces just in order to find new interpretations of them, that is in order to propose new dimensions for their representations. In other words, within the social sciences, before constructing a new representation of objects studied, we need to use analytical tools and perform some de-construction processes.

Generally, in my view, each analysis process:

- starts with the de-construction of some phenomenon, that is to say its reduction, as it would appear to some of its relevant features;
- requires the construction and/or use of some equivalence classes (i.e. categories) which allow "classing a variety of stimuli as forms of the same things" (Bruner, Goodnow & Austin, 1956, p. 2);
- and, by constructing and/or using categories, it proceeds by means of inferences which bring into play various theories.

In particular, the de-construction process carried out by a co-occurrence analysis leaves out three features of word/sentence meaning:

- a) the *reference* to the extra-linguistic context (or situation), that is the *indexicality* beloved of the ethnomethodologists;
- b) the *sequential order* of the words within the linguistic contexts, that is text *cohesiveness* and the anaphoric processes;
- c) the semantic effects of *speech acts* that is all the relationships between the utterances and their *enunciation* processes.

For example, considering the vector represented in Fig. 1.10 compared with the the utterance which it correspond to:

- a) we cannot know the referents of "you" (Who is?) and of "this book" (Which is?);
- b) we lose the anaphoric link of "the latest work" with "this book";
- c) we lose the effect of the question form "Did you read...?" (e.g. it could be an ironic question).

In some ways, after the de-construction process, the remaining utterance meaning seems to be a sort of *propositional content*, i.e. a quasi-sentence without speaker and without textual/situational context.

However, in thinking about co-occurrence analysis, as about other analysis processes, the de-construction effects (“What features do we leave out?”) are not the main concern; rather, it is a question of verifying how, by using categories (theories and/or models), we can achieve a useful and valid *construction* of knowledge.

As an exemplary case of the de-construction/construction process, I quote Propp’s work *Morphology of the Folktale*. At the beginning of second chapter (“The Method and the Material”) he says:

We are undertaking a comparison of the themes of these tales. For the sake of comparison we shall separate the component parts of fairy tales by special methods; and then, we shall make a comparison of tales according to their components. The results will be a morphology (i.e. a description of the tale according to its component parts and the relationship of these components to each other and to the whole).

What methods can achieve an accurate description of the tale? Let us compare the following events:

1. A tsar gives an eagle to a hero. The eagle carries the hero away to another kingdom.
2. An old man gives Sucevko a horse. The horse carries Sucevko away to another kingdom.
3. A sorcerer gives Ivan a little boat. The boat takes Ivan to another kingdom.
4. A princess gives Ivan a ring. Young men appearing from out of the ring carry Ivan away into another kingdom, and so forth.

Both constants and variables are present in the preceding instances. The names of the dramatis personae change (as well as the attributes of each), but neither their actions nor their functions change. From this we can draw the inference that a tale often attributes identical actions to various characters. This makes possible the study of the *tale according to the functions of its dramatis personae*. (Propp, 1928, pp. 19-20).

To achieve his aim, i.e. the *construction* of a morphology, Propp *de-constructs* every tale by segmenting it into sequences; then, by comparing all sequences with each other, he infers (i.e. *constructs*) equivalence classes, and he names each of them as a *function* (that “is understood as an act of the character, defined from the point of view of its significance for the course of the action”; Propp, 1928, p. 21), and so on. At the end, after having distinguished a set of functions, Propp’s method allows us to construct a table in which each row represents a tale and each column represents a function. In other words, the construction of the morphology passes through a sort of co-occurrence analysis.

Now, going back to the example in Fig. 1.10 and using a category proposed by the semiologist Hjelmslev (1943), we can state that the vector-space representation involves a de-construction of the raw/original *expression-form* (E1: “Did you read ...”) but, at the same time, we can state that this representation proposes a new *expression-form* (E2: see the vector in Fig. 1.10). More precisely, E2 proposes a representation which loses the “linearity” of the text characterizing E1 (e.g. “work” is after “Harris” and before “latest”).

How E1 and E2 are interrelated and how each of them interrelates with the *content-forms* will be discussed below. Now we can consider the following assertion recurring in scientific papers:

two (or more) words that tend to occur in similar linguistic contexts (i.e. to have similar co-occurrence patterns) tend to resemble each other in meaning.

In order to comment on the implications of this, let us first agree about what we intend by “resemblance of meaning” (or similarity of meaning). In particular, do we mean cases of synonymy or the fact that two (or more) words belong to the same “semantic field”? In the first case (i.e. synonymy) we use a thesaurus-oriented approach, in other words within the same linguistic context and with no significant change of meaning, we can replace one word with another (e.g. replace “Islamic” with “Muslim”); in the second case, because the structure of semantic fields relies on social/cultural events, we must carefully evaluate co-occurrences of words like “Islamic” and “terrorist” in specific contexts, that is to say we must move within the boundaries of text mining and/or of content analysis practices.

With reference to the latter, taking “co-occurrence” as a synonym of “contingency”⁶, let us recall some of Osgood’s claims:

contingency analysis depends only upon the presence of symbols, not how they are linked... The contingency method provides evidence for non-chance structure; interpretation of this structure is still the job of the skilled analyst...

Finally, we may note that there was considerable discussion in the conference over whether or not contingency among categories implied similarity of meaning. Certainly, if the meaning of a concept is identified either in terms of what is referred or in terms of a location in a n-dimensional space, then the fact of association is not indicative of semantic similarity. References to COMMUNISM may frequently lead to references to CAPITALISM, but this does not necessarily imply that these concepts are either similar in reference or in psychological meaning. On the other hand, there are certain relationships between association and meaning that should be indicated. First, association between concepts (e.g. GOD and DEVIL, CAPITALISM and COMMUNISM, SOLDIER and SAILOR), even though they may be common opposites semantically, will often reflect the fact that they share certain attributes, are of the same “class”, or (linguistically) are alternatives within the same structural frames.” (Osgood, 1959, 76-77).

In fact, many co-occurrence analyses show “opposites semantically” closer together in semantic space; but, in order to determine whether this evidence is a sort of counter-example of an expected similarity in meaning, let us first decide which theory of meaning to refer to. For instance, if we refer to Greimas’s model (1966), the pairs “god/devil” and “capitalism/communism” can be considered two cases of

⁶ Osgood himself said “contingencies or co-occurrences of categories in the same units” (1959, p. 61).

*isotopy*⁷: the first religious and the second political. Moreover, according to Greimas's semantics, each of them can be considered as *actant* within the same semiotic square (see Fig. 2.1).

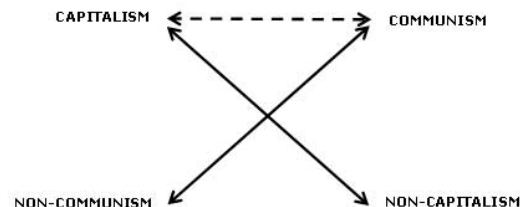


Fig. 2.1: A semiotic square

We don't know whether, by using the expression "structural frames", Osgood meant to refer to structural linguistics (or structural semiotics), but his intuition was appropriate. The fact is that, being persuaded that his method had a "defensible psychological rationale", he was interested in a genetic explanation of the object studied

On such grounds it seems reasonable to assume that *greater*-than-chance contingencies of items in messages would be indicative of associations in the thinking of the source (Osgood, 1959, p. 55)

But between structural and genetic explanations there is a significant difference. While the first makes inferences within the same domain (i.e. that of *sign*⁸), the second aims to connect two different domains: that of sign and that of *source*.

Before going any further, because – inevitably – I have already made reference to various theories, I would prefer to be explicit.

Assuming that meaning has no existence on its own and can be described only in relation to something else, on the one hand (1) I shall consider meaning within language as *sign structure* and, on the other hand, (2) I shall consider it as the result of *inference processes*. Then I think I will have a minimum of equipment for explaining what happens (see above) when words become numbers and when tables and/or graphs become "texts" to be interpreted.

In the first case (1), I assume a structural point of view as defined in the following claim by Hjelmslev:

⁷ See below section 6.

⁸ According to Hjelmslev (1943), expression and content can be considered as belonging to the same domain: that of the *sign*.

By structural linguistics we mean a set of investigations based on the assumption that it is scientifically legitimate to describe language as being primarily an autonomous entity of internal dependences, or, in a word, a structure (1959, p. 21)⁹.

This point of view, like Ricoeur (1969), postulates the view of language as a closed system in which every sign only refers to other signs. And, according to Saussure (1916), it leads us to consider the sign as a two-sided entity: both *signifiant* (*signifier*) and *signifié* (*signified*) or, in Hjelmslev's (1943) terms, *expression* and *content*.

Within this framework, “meaning” is synonymous with “content” and as such it is only one of two functives that contract the “sign function”:

The sign function is in itself a solidarity. Expression and content are solidary – they necessarily presuppose each other. An expression is expression only by virtue of being an expression of a content, and a content is content only by virtue of being a content of an expression. (Hjelmslev, 1943, p. 48-49).

I shall comment on other aspects of the structuralist standpoint later. Now I shall just refer to the hypothesis, initially proposed by Saussure (1916) and subsequently by several authors (e.g. Jakobson, 1963; Barthes, 1964), whereby the relationships between linguistic elements (i.e. signs as expressions and as contents) can be analysed as syntagmatic and/or as paradigmatic relationships. The former regulate the combination of linguistic elements within contexts (one “near to” the other), the latter determine the possibility of replacing a linguistic element with one that has something in common with it (one “in place of” the other).

Moving on to the second perspective, i.e. considering the sign as the result of an inference, according to Eco (1984), if the relationships between signs were only of a structural type, the same relationships could all be governed by *codes* and all the relationships between expression and content pairs could be represented in the form of a *dictionary* (and/or an *ontology*, I would add); but the construction of meaning seems to require a network of *encyclopedic* knowledge which can't be described in its entirety. In particular, the Italian semiologist argues that the “stands for” relationship – referring to definition of the sign as “something which stands for something else” – is neither a simple equivalence nor a correlation between entities, but a kind of implication; in other words, the content (i.e. the meaning), rather than being the result of a substitution (“in place of”), is the result of an interpretation process. And the interpretation process requires encyclopedia-like semantics which take into account the contextual selections.

⁹ “On comprend par linguistique structurale un ensemble de recherches reposant sur une hypothèse selon laquelle il est scientifiquement légitime de décrire le langage comme étant essentiellement une entité autonome de dépendances internes, ou, en un mot, une structure”

In this type of semantics each link between sign and contexts of meaning which we refer to is a result of an abductive inference. In fact, following Peirce, Eco proposes a conception of the sign and of the inference as a triadic function.

From among Peirce's definitions of the sign we can just remember the following:

I define a *Sign* as anything which on the one hand is so determined by an *Object* and on the other hand so determines an idea in a person's mind¹⁰, that this latter determination, which I term the *Interpretant* of the sign, is thereby mediately determined by that Object. A sign, therefore, has a triadic relation to its Object and to its Interpretant. (Peirce, C.P. 8.343).

Using a diagram, the same definition can be represented as follows

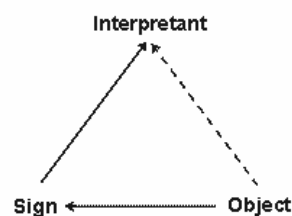


Fig. 2.2: Peirce's model of the sign

As the triadic structure of abduction, also called "making an hypothesis", Peirce defines it as "the inference of a *case* from a *rule* and a *result*" (C.P. 2.623). To illustrate his concept, he uses the following much-quoted example:

Suppose I enter a room and there find a number of bags, containing different kinds of beans. On the table there is a handful of white beans; and, after some searching, I find one of the bags contains white beans only. I at once infer as a probability, or as a fair guess, that this handful was taken out of that bag. This sort of inference is called making a hypothesis. (Peirce, C.P. 2.623).

According to the author, the logical structure of this sort of abductive inference is the following:

¹⁰ In some ways, the reference to "person's mind" is problematic; in fact, in a confidential context, Peirce added an interesting gloss: "I define a Sign as anything which is so determined by something else, called its Object, and so determines an effect upon a person, which effect I call its Interpretant, that the latter is thereby mediately determined by the former. My insertion of "upon a person" is a sop to Cerberus, because I despair of making my own broader conception understood." (Welby, 1977, pp. 80-81).

Rule: All the beans from this bag are white.

Result: These beans are white.

Case: These beans are from this bag.

As a process: starting from a *Result* (the minor premiss) and referring to a *Rule* (the major premiss), the abduction infers a *Case* (the conclusion) which, looking at Fig. 2.2, is a sort of Interpretant.

More generally, we make abductive inferences whenever – by identifying some character(s) – we recognize a *token* as the occurrence (or instance) of a *type*, i.e. whenever we recognize something as element of a class: a bird as belonging to a species, a car as belonging to a brand, a word as belonging to a root or to a semantic class, and so on.

Now, coming back to the question of meaning: if we adopt the view of structural semantics, it is a *content* (i.e. one of two sides of the *Sign*), or alternatively, if we adopt a Peircean view, it is an *Interpretant*. In both cases it is a term of a function, the former dyadic and the latter triadic. How we can refer to one and/or the other theory as a useful tool depends on the issues we intend to account for. It is only a methodological, not an ideological, question; in fact, purporting “to have a foot in both camps” (or “to play both sides of the street”) is a typical ideological assumption.

Briefly, it seems to me that the structuralist framework is more useful for explaining (see above) what happens when words become numbers, whereas the Eco-Peirce standpoint is more useful for explaining how we interpret “texts” as software outputs (table and/or graphs).

Moreover, because within concrete hermeneutical processes (i.e. when we analyse texts as specialists in some scientific domain) we carry out *interpretative trajectories* by interconnecting features of meaning, theoretical constructs and extralinguistic factors, both Eco’s encyclopedia-like model and the *descriptive semantics* (and/or the *interpretative semantics*) proposed by Rastier (Rastier 1987; Rastier, Cavazza & Abeillé 1994) seem to me useful tools which - as a sort of metalanguage - can help us (a) to define our steps within the hermeneutic circle (i.e. the interpretative process of moving between part and whole) and (b) to keep sight of the structural constraints of language.

Now, starting from the structural point of view, I shall just consider some implications of the claim henceforth referred to as our *c.b.e.* (i.e. as the claim to be explained)

two (or more) words that tend to occur in similar linguistic contexts (i.e. to have similar co-occurrence patterns) tend to resemble each other in meaning.

In particular, since we can't take into account how this claim can be interpreted differently within different traditions of structuralist research, I shall consider two implications only: firstly, one originating from Leonard Bloomfield and developed by the work of Zellig S. Harris; and secondly, one originating from Ferdinand de Saussure and developed (among others) by Louis Hjelmslev and Algirdas J. Greimas.

3. Harris's distributional hypothesis

Premise: the fact that, when reporting co-occurrence analyses, some scholars quote Harris's work depends on variants of our *c.b.e.* For example, the following:

two (or more) words that tend to have similar *distributional* patterns (i.e. to have similar co-occurrence patterns) tend to resemble each other in meaning

The fact is that, if we assume a quasi-synonymy between "distributional" and "co-occurrence" patterns, the concept of co-occurrence undergoes a significant shift. Inasmuch as there is no point in nurturing misunderstandings, it is worth remembering that according to Harris distribution concerns the "freedom of occurrence of portions of an utterance relatively to each other" (Harris, 1951, p. 5)

Here are some of his definitions:

The ENVIRONMENT or position of an element consists of the neighborhood, within an utterance.... 'Neighborhood' refers to the position of elements before, after, and simultaneous with the element in question...

The DISTRIBUTION of an element is the total of all environments in which it occurs, i.e. the sum of all the (different) positions (or occurrences) of an element relative to the occurrence of other elements.

Two utterances or features will be said to be linguistically, descriptively, or distributionally equivalent if they are identical as to their linguistic elements and the distributional relations among these elements (Harris, 1951, pp. 15-16)

From these definitions it is possible to study the "combinations (mostly sequences) of elements, and to state their regularities and the relations among the elements" (ibid., p. 17).

To put it briefly, in Harris's writings the notion of distribution is indivisible from that of "combination" (or "sequence"). In fact "distributional or combinatorial analysis" (1952, p. 109) is the method of his *descriptive linguistics*, which always aims to account for *how* departures from randomness (or equiprobability), through constraints on the freedom of occurrence of elements in respect to each other, produce patterns of non-randomness, that is information.

In his mind, simply because the parts of a language do not occur arbitrarily relative to each other, descriptive linguistics can produce a phonology (comparing phoneme distribution), a morphology (comparing morpheme distribution), and a syntax (comparing sentence distribution). In all three cases, Harris's "procedure" requires two major steps: "the setting up of elements, and the statement of the distribution of these elements relative to each other" (1951, p. 6). The second (i.e. the statement of distribution), by means of substitution tests, produces "equivalence classes" (or substitution sets).

As Harris claims:

We group A and B into substitution set whenever A and B have the same (or partially same) environments X (Harris, 1981, p.17)

Let's follow one example of his procedure:

Suppose our text contains the following four sentences: *The trees turn here about the middle of autumn*; *The trees turn here about the end of October*; *The first frost comes after the middle of autumn*; *We start heating after the end of October*. Then we may say that *the middle of autumn* and *the end of October* are equivalent because they occur in the same environment (*The trees turn here about -*), and that this equivalence is carried over into the latter two sentences. On that basis, we may say further that *The first frost comes* and *We start heating* occur in equivalent environments... After discovering which sequences occur in equivalent environments, we can group all of them together into one equivalence class... In our example, *The trees turn here in* (T₁) and *The first frost comes after* (T₂) are all members of one equivalence class T, while *the middle of autumn* (E₁) and *after the end of October* (E₂) are members of another equivalence class E... Our text fragment can be structurally represented by a double array, the horizontal axis indicating the material that occurs within a single sentence or subsentence, and the vertical axis (here broken into two parts) indicating the successive sentences:

| | | | |
|----------------|----------------|----------------|----------------|
| T ₁ | E ₁ | T ₃ | E ₂ |
| T ₁ | E ₂ | T ₃ | E ₃ |
| T ₂ | E ₁ | T ₂ | E ₄ |

In this double array, the various symbols in one horizontal row represent the various sections of a single sentence or subsentence of the text, in the order in which they occur in the sentence...The vertical columns indicate the various members of an equivalence class, in the order of successive sentences in which they occur."(Harris, 1952, pp. 113-116)

An important feature of this procedure is that it requires the transformation of raw data into some array (or matrix) in which the sequential order of the members must be respected. So, "A consecutive (or seriate) discourse of one or more persons is thus the fullest environmental unit for the distributional investigation" (Harris, 1954, p. 15). Therefore, when analysing these types of contextual co-occurrence, any statistical (i.e. algorithmic) approach must use probability computations, such as hidden Markov models.

On the other hand, if we try to represent Harris's four quoted sentences as zero-one vectors¹¹, we can obtain a result like the following:

| after | autumn | come | end | first | frost | heat | middle | October | start | tree | turn |
|-------|--------|------|-----|-------|-------|------|--------|---------|-------|------|------|
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 |
| 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 |

Fig. 3.1: Harris's examples in a zero-one representation

As we can see, if we zero the distributional properties arising from the “sequential” order, the task of building equivalence classes and finding similarities of meaning becomes arduous. Nevertheless, the representation of word co-occurrences as in Fig. 3.1, even if it treats texts as “bags-of-words”, works well in research processes that use measures of similarity (e.g. the cosine coefficient), clustering algorithms and/or some types of MDS or SVD¹². Obviously, this applies when using a lot of sentences or documents and – above all – when the aim is to find (or construct) semantic fields or classes and not to establish the “exact” meaning of the words; also, bear in mind that the interpretation of every semantic space “is still the job of the skilled analyst” (see the above quotation from Osgood).

Now let us consider Harris's hypotheses concerning the theory of meaning.

He said that “difference of meaning correlates with difference of distribution” (Harris, 1954, p.14); but, as regards a scientific analysis of meaning and meaning as an explanation of linguistic phenomena, he remained somewhat sceptical.

As Leonard Bloomfield pointed out, it frequently happens that when we do rest with the explanation that something is due to the meaning, we discover that it has a formal regularity or “explanation”. It may still be “due to meaning” in one sense, but in accords with a distributional regularity (Harris, 1954, p. 13)

In his opinion “descriptive linguistics has not dealt with the meanings of morphemesit can only been able to state the occurrence of one linguistic element in respect to the occurrence of others” (Harris, 1952, p. 108).

According to Nevin (2002, p. xxi), in much of the literature of linguistics, syntax and semantics are distinct rubrics, but - for Harris - they are two faces of the same socially maintained phenomenon, linguistic information. To be precise, in analysing the “formal features” of language, he was interested not in building a

¹¹ In Fig. 3.1, the first row represents the sentence *The trees turn here about the middle of autumn*, the second represents the sentence *The trees turn here about the end of October* and so on.

¹² MDS: Multidimensional Scaling. SVD: Singular Value Decomposition (mostly used in factorial techniques).

theory of *meaning* but in showing how the restrictions on combinations of linguistic elements have the cumulative effect of creating *information* (Ryckman, 2002, p. 30). In the end, not by chance, Harris proposed a theory of language as a self-contained, self-organizing, and evolving system (Harris, 1991).

As he himself claimed:

This is so because of the unique status of language as system which contains its own metalanguage. Any description we make of a language can only be stated in a language... We cannot describe a language without how our description can in turn be described. And the only way to avoid an infinite regress is to know a self-organizing description, which necessarily holds also for the language we are describing even if we do not use this fact in our description (Harris, 2002, p.10).

From a semiological standpoint, we could say that by taking the linguistic empiricism of Bloomfield to extremes (“The only useful generalizations about language are inductive generalizations”, Bloomfield, 1933, p. 20) and not recognizing the relative autonomy of the “content forms”, that is by recognizing only the structure of “expression forms”, Harris was unable to conceive the possibility of an autonomous metalanguage.

In fact, according to Hjelmslev, metalanguage is conceivable only by distinguishing the expression and the content planes, because the metasemiotics are “semiotics whose *content plane* is a semiotic” (1943, p. 114). In other words, according to Barthes (1964), a metalanguage is a system in which the content plane is a signification system (see Fig. 3.2).

| | | | |
|---------------|-------------------|-------------------|----------------|
| Meta-language | EXPRESSION | CONTENT | |
| | Language | EXPRESSION | CONTENT |

Fig. 3.2: Language and metalanguage

The implications of this model for our *c.b.e* will become clear in the following paragraphs.

4. According to Hjelmslev: the distinctiveness of *form contents*

In structural linguistics both Harris and Hjelmslev excel because of the scientific rigour of their methods but, while the former proceeds by careful description and inductive inference, the latter prefers to refine his conceptual equipment through a deductive approach.

In particular, by specifying the relationships between two concept pairs (sub-

stance/form and expression/content), Hjelmslev made possible the structural analysis of content forms (or *figurae*). Of course, Harris too admitted the existence of content forms and – in my opinion – Hjelmslev would subscribe to the following claim by his colleague:

If one wishes to speak of language as existing in some sense on two planes – of form and of meaning – we can at least say that the structures of two are not identical, though they will be found similar in various respects (Harris, 1954, p.9).

This very point, concerning the asymmetries between expression and content planes, is a crucial problem which Hjelmslev was trying to clarify. In order to summarise his hypotheses, I will use the following diagram (Fig. 4.1)

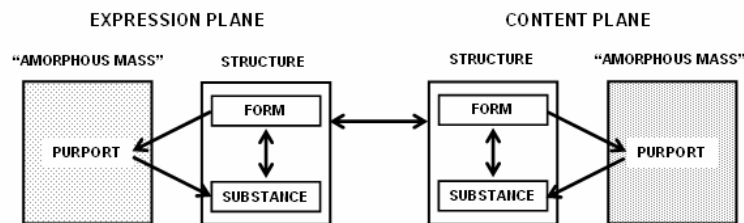


Fig. 4.1: Overview of Hjelmslev's model

The Danish semiologist used the same model to study both expression and content plane structures, by distinguishing in each one a form and a substance. As to the respective *purports*, they are also "amorphous" because - like the Kantian "thing in itself" – they are unanalysable (obviously, from the standpoint of structural linguistics).

According to Barthes (1964), this model allows us to distinguish in every sign (e.g. a simple word like "woman")

- 1) a substance of expression: for instance the phonic, articulatory, non-functional substance which is the field of phonetics, not phonology;
- 2) a form of expression, made of the paradigmatic and syntactic rules (let us note that the same form can have two different substances, phonic and graphic);
- 3) a substance of content: this includes, for instance, the emotional, ideological, or simply notional aspects of the signified, its 'positive' meaning;
- 4) a form of content: it is the formal organization of the signified among themselves through the absence or presence of a semantic mark. (1964, II.1.3).

As regards content structures, Hjelmslev proposed the following much-quoted example

| | |
|--------------|---------------|
| <i>green</i> | <i>gwyRDD</i> |
| <i>blue</i> | <i>glas</i> |
| <i>gray</i> | |
| <i>brown</i> | <i>llwyd</i> |

Fig. 4.2: A Hjelmslev example

In this case, the *purport* is the “amorphous continuum” of the colour spectrum “on which each language arbitrarily sets its boundaries” (1943, p. 53); the two different *forms* (the two columns in Fig. 4.2) are the grids displaying the continuum within languages/cultures: English (left) and Welsh (right); the different *substances* are the colours (as contents/concepts and not as expressions/nouns).

In Welsh, ‘green’ is *gwyRDD* or *glas*, ‘blue’ is *glas*, ‘gray’ is *glas* or *llwyd*, ‘brown’ is *llwyd*. That is to say, the part of the spectrum that is covered by our word *green* is intersected in Welsh by a line that assigns a part of it to the same area as our word *blue* while the English boundary between *green* and *blue* is not found in Welsh. Moreover, Welsh lacks the English boundary between *gray* and *brown*. On the other hand, the area that is covered by English *gray* is intersected in Welsh so that half of it is referred to the same area as our *blue* and half to the same area as our *brown*. (Hjelmslev, 1943, 53).

These words suggest a hypothesis similar to the so-called “linguistic relativism” proposed by Sapir (1949) and Worf (1956), which states that differences between cultures are widely determined by differences between their respective languages; or to that of Lotman (1990), which considers languages as clusters of semiotic spaces and their boundaries interacting within the semiosphere.

But, in order to keep sight of our questions, let us verify how the Hjelmslev hypothesis concerning the relationships between expression and content planes can help us to understand something about the logic of word co-occurrence analysis.

In the first instance we discover that the asymmetry between the two planes contradicts the belief that single words are signs; in fact, the sign being a two-sided entity (expression/content), many words can be subdivided into two or more morphemes (i.e. minimal units of meaning). For example, the word “subdivided” is made of three morphemes, each with its meaning: “sub”+ “divid”+ “ed”. The first (“sub”) is a grammatical element (a prefix) which, combined with a word, produces – at the same time - derived forms (expression plane) and meaning shifts (content plane). The second (“divid”) is the stem of the verb and means a kind of action. The third “ed” is an inflexional form indicating time.

This asymmetry leads to the discovery that, in terms of expression, every mor-

pheme can be subdivided into a number of phonemes; whereas, in terms of content, it can be associated with a number of meanings. In particular, if – according to Eco (1984) – we consider any dictionary as an encyclopedia, each word can be associated with a potentially infinite set of meanings.

Since I agree with Eco’s hypothesis, I should like to clarify a point. It is evident that potentially infinite sets cannot be represented in a co-occurrence array; but, in our case (and according to Eco), the problem of being faced with potentially infinite links concerns *interpretation* as a function which connects expression and content-planes. As to the vectors of our hypothetical array, the headings of its rows and columns are in any case a finite set of expressions.

But if, from a scientific standpoint, the concept of word is blurred, how we can proceed?

Well, an exploration of the literature on content forms produces some interesting and useful discoveries. In particular, the notion of *lexie* as proposed by the linguist Pottier (1974) allows us, in a way, to reverse the process whereby the analysis of content units leads to the breaking down of single words; on the contrary, since each lexie is a sign (or signifying unit) referring to a unitary content, not only single words but also sequences of two or more words can be considered as referring to a unitary meaning¹³. To be exact, Pottier distinguishes several kinds of lexies: the *simple* (e.g. boy, run), the *compound* (e.g. biotech, self-oriented) and the *multiplex* (e.g. United States, Chamber of Commerce).

Of course, the problem remains that each lexie, as a sign, can refer to a set of semantic traits (content plane), i.e. it can refer to a *semie* (Pottier 1974, p. 79) understood as a set of *semes*¹⁴. For example, “boy”, considered as a dictionary entry, refers to the set “human”+“male”+“young” But from a linguistic standpoint, as Martinet (1960) has suggested, it comes to distinguish the first and the second articulation, i.e. as in the grid below:

| | EXPRESSION | CONTENT |
|---------------------|-------------|------------------------|
| First articulation | /boy/ | “boy” |
| Second articulation | /b/+/o/+/y/ | “human”+“male”+“young” |

Fig. 4.3: First and second articulation

¹³ According to Pottier, the lexie, originating from an associative habit, resides in our lexical memory (1974, pp. 265-266) whereas the word is «une unité construite, intermédiaire entre le morphème, unité de construction, et la lexie, unité mémorisée de fonctionnement» (1992, p. 38). To put it differently, both the word and the lexie are composed of one or more morphemes, but the latter only (i.e. the lexie) is properly a sign.

¹⁴ From now on I will use the notion of “seme”, proposed by Greimas (1966), to refer to “minimal” units of content plane, i.e. I will use “seme” as a synonym of semantic trait.

While the smallest units of meaning on the expression plane are the morpheme (first articulation), the smallest units of meaning on the content plane are the semes (second articulation)¹⁵. Now, coming back to the problem of asymmetries between the two planes, in addition to noting that *the set of phonemes is finite and the set of semes can be potentially infinite*, we can observe that, from the standpoint of meaning, the order (or sequence) of phonemes within each morpheme is a constraint; for example, changing the sequence /b/+o/+y/ in /o/+b/+y/ we obtain a word (“oby”) devoid of meaning. On the contrary, if we change the order of semes (e.g. from “human”+“male”+“young” to “young”+“male”+“human”) the meaning is unaltered.

This fact, as we shall see later, has significant consequences for our *c.b.e.*, as defined by the following statement:

two (or more) words that tend to occur in similar linguistic contexts (i.e. to have similar co-occurrence patterns) tend to resemble each other in meaning

Firstly, as in the previous arguments, we could replace *word* with *lexie*; not least because we would then have a single label for a set of cases (e.g. single words, multi-words, word phrases, etc.). But also, if we continue to use *word* as a synonym of *lexie*, we must reckon with the fact that every word (i.e. every column of our hypothetical co-occurrence matrix) refers to a set of semes (i.e. semantic traits). This implies that, as regards its meaning (or content), *every word – as such – is made up of seme co-occurrence*. And – a not inconsiderable aspect – *this kind of co-occurrence does not imply the sequential order of its elements*.

In fact, it is no accident that, if we follow a dictionary-like approach, the componential analysis allows us to construct some sort of co-occurrence matrix (see Fig. 4.4).

| | to sit | from firm material | for one person | on leg(s) | with backrest | with arm rests |
|---------------------|--------|--------------------|----------------|-----------|---------------|----------------|
| siège (seat) | + | 0 | 0 | 0 | 0 | 0 |
| chaise (chair) | + | + | + | + | + | - |
| fauteuil (armchair) | + | + | + | + | + | + |
| tabouret (stool) | + | + | + | + | - | - |
| canapé (settee) | + | + | - | + | + | 0 |
| pouf (pouf) | + | - | + | - | - | - |

Fig. 4.4: Seats according to Pottier (1965)¹⁶

¹⁵ On the expression plane, the units of second articulation (i.e. phonemes) are meaningless.

¹⁶ Key: “+” = presence; “-” = absence; “0” = neutral.

5. Greimas's structural semantics: a way to account for contextual meaning

In his work *Sémantique structurale* (1966), Greimas builds on Hjelmslev's hypotheses concerning content forms and suggests new analytical tools. Here I will consider only a few: those I believe most useful in accounting for our *c.b.e.*

Firstly, I would cite Greimas's classification of the semes (i.e. semantic traits) as *nuclear* and *contextual*, starting with the "formula" (1966, p. 53)

$$\text{sememe } Sm = Ns + Cs$$

where *sememe* (*Sm*) indicates a content unit of the first articulation (see Fig. 4.3 above), while *Ns* and *Cs* indicate *nuclear semes* (or semic nuclei) and *contextual semes* respectively, that is the minimal content units of the second articulation.

The former (*Ns*) are specific and invariable, that is they characterize every *sememe* independently of the contexts in which they occur. The latter (*Nc*) are shared among two or more *sememes* which *co-occur* within the same contexts.

Because the *contextual semes* allow us to group the sememes into classes, Greimas proposes calling them *classemes*. Moreover, for indicating the "meaning effects" due to contextual semes (or classemes), Greimas suggests using the term *isotopy* (etym: *iso* = same; *topos* = place).

The way Greimas illustrates his findings is particularly interesting. In order to seek out the semic nucleus of the sememe "*tête*" ("head") he uses a dictionary. By contrast, when seeking out the contextual semes (or classemes) he starts from the sentence "*le chien aboie*" ("the dog barks"). In this last case, he argues that there are two "contextual classes" of subjects which can combine with "barks":

firstly, the animal class:
the dog
the fox
the jackal etc.
and secondly, the human class:
the man
Diogenes
this ambitious etc. (1966, p. 59)

In a way, the co-occurrence of "barks" with the two contextual classes leads us directly to consider the relationship between co-occurrence and resemblance of meaning. In this case, the resemblance is between each of two contextual classes. But we are avoiding a significant aspect of the problem: Greimas evokes the subjects (animal and human) by resorting to his reminiscences (or, in Eco's words, to his encyclopedic knowledge); in a different way, "real" co-occurrences occur in utterances, spoken or written by someone.

But *isotopy* is precisely the “meaning effect” due to:

the interactivity - within the syntagmatic chain – of *classemes* which shape the homogeneity of speech-utterance (Greimas & Courtés, 1979, 197)¹⁷.

To put it another way, the superabundance of contextual *semes* (or *classemes*), by shaping the relationships between words (or *sememes*) in sentences and texts, allows us to perceive cohesiveness of meaning and thematic similarities, that is to say it makes possible a type of pattern recognition.

In this case, examples of single sentences can be trivial; however, let us consider the following:

- *That student is reading the software manual.*
- *Waitress! Can you bring us the menu?*
- *The film was greeted with applause by the press.*

Now, simulating a typical procedure of co-occurrence analysis, let us rewrite them, only using the content words, eliminating the auxiliary verbs and – in each sentence – respecting alphabetical order.

In this way we obtain the following result:

- *manual, reading, software, student*
- *bring, menu, waitress*
- *applause, film, greeted, press*

In some ways, within each context unit, the “meaning effect” of three different isotopies remains. Obviously, many counter-examples can be found; but, in this case, they are pushing against an open door. In fact, this is not a question of collecting a lot of “correct” examples and, by means of *inductive* procedures, imagining a general law (or rule); rather, it comes down to finding a category of analysis (an equivalence class) by means of *abductive* inferences.

Moreover, the question of co-occurrence analysis doesn’t only affect the relationships of words within context units but, overall, it involves the relationships (similarities and differences) between numerous context units, that is to say between numerous sentences and/or texts.

Coming back to the vector-space model (see section 1), because every context unit is represented as a zero-one vector¹⁸ we can use measures of similarity (e.g. the cosine coefficient) to compare each of them with each other. In this way, also

¹⁷ «itéractivité, le long d’une chaîne syntagmatique, de *classèmes* qui assurent au discours-énoncé son homogénéité».

¹⁸ The fact that in many cases the cell values are occurrences (i.e. frequencies) doesn’t change the logic of our argument. Like the representation of time (see watches), the representation of context units can be digital (zero-one) or analogue (occurrence values).

using clustering algorithms, we can still address our *c.b.e.*; but, at this point, we can add some explanatory items.

For example, using Greimas's categories, we can rewrite it as follows:

two (or more) words that tend to occur in similar linguistic contexts (i.e. to have similar co-occurrence patterns) tend to resemble each other in meaning

BECAUSE

two (or more) words that tend to occur in similar linguistic contexts share some contextual semes, that is to say they – probably – refer to the same isotopy.

This concerning the reason why words co-occurring in similar contexts tend to have similar meaning.

Similarly, because we can scroll through our co-occurrence array (see Fig. 1.3) by rows or by columns, we can also say:

two (or more) context units in which similar word patterns tend to occur (i.e. to have similar co-occurrence patterns) tend to resemble each other in meaning

BECAUSE

two (or more) context units in which similar word patterns tend to occur share some isotopies.

This concerning the reason why documents that contain similar word patterns tend to have similar topics.

In both cases, observation of the co-occurrence patterns concerns the expression plane, whereas the explanation involves the relationships between expression and content planes. Therefore the meaning of “because” is not to be understood scientifically as the *Covering Law Model*¹⁹(Hempel, 1961), but within a semiotic framework. That is to say, in our case, “because” does not refer to an inductive but to an abductive inference.

In fact, as Rastier argues (1987, pp. 11-12), the recognition of an isotopy is not the simple observation of a *given* but the result of an interpretative process. According to Rastier, within the “*présomption d’isotopie*” (*isotopy assumption*) an overall vision precedes the attribution of meaning to the single elements, i.e. it is a process “où la classe détermine l’élément, et où le global détermine le local” (in which class determines the element and the global determines the local).

In Peircean terms, as the result of an abductive inference, the isotopy is an Interpretant. In fact, according to Eco (1984, p. 193), by means of contextual selections the isotopy leads “to the semantic result of a coherent interpretation”. Or, in

¹⁹ According to this model, by assuming the validity of one (or more) general law, events can be explained and/or anticipated by referring to the occurrence of specified conditions. In this case, the *explanans* (that which allows us to say “because...”) includes a law and a set of specific conditions.

other words, the recognition of an isotopy follows the selection of some topic (i.e. semantic field) and, from a semantic standpoint, is a verifiable property.

6. Outputs as texts to be interpreted: from semiotics to hermeneutics

In the previous sections of this paper we referred to various semantic theories and examined the implications of arranging raw data (i.e. words and texts) in co-occurrence data matrices. Now let us consider, from a semiotic standpoint, the interpretation of the outputs, i.e. the results of data processing²⁰.

In some ways, when we examine these outputs – whether tables or charts – we are faced with a kind of *multi-semiotic* object and a kind of text to be interpreted. To understand the implications of this hypothesis, we must firstly consider the semiotic character of the tables. In fact, the shape of charts results from information already present in output tables, even though their visual representation offers new patterns to our interpretative skills.

Because the specificity of output tables consists in assembling information as numerical values, we must first clarify the semiotic character of the numbers.

Referring to Hjelmslev's semiotics, we remember that every number, "*per se*", is not a sign but an expression form without content. This amounts to saying that it becomes a sign, i.e. it assumes *value*, only if we refer to a system of content forms. For example, only because we refer to the decimal system do we interpret the string "28" as "twenty-eight" (i.e. two tens plus eight units); that is to say we make a sort of *hyper-coded abduction*, in which – as Eco suggested (1983) – the Peircean "rule" is given automatically.

To illustrate the inferential character of this process, let me give an elementary example. When we encounter the string "101", according to linguistic convention we recognize the sequence "one-zero-one"; but, in order to establish what it means, we need a frame of reference. To be exact, if "101" is to be interpreted as a numerical value, its meaning varies depending on the notation system; in fact, in a binary system it means "five", whereas in a decimal system it means "one hundred and one"

But this is not enough. While still referring to the decimal system, the same number can assume other meanings. For example, in the coding system used by many hotels, "101" means "the first room on the tenth floor". In other words, when numbers become signs, referring to Hjelmslev's model, we can use them in either a

²⁰ As every analyst knows, the way of arranging data matrices has more than one consequence in the output structure.

denotative or a connotative system (see Fig. 6.1).

| | | | |
|-------------------|----------------|----------------|-------------|
| EXPRESSION | | CONTENT | Connotation |
| EXPRESSION | CONTENT | Denotation | |

Fig. 6.1: Denotation and connotation

Coming back to output tables, by taking them together we can observe that not only do they assemble words and numbers, but – at the same time – they determine different meanings for the same numbers. For example, using a Correspondence Analysis different output tables can report the number “0.5” (“zero point five”), but it can be a coordinate, an absolute or relative contribution, a percentage, and so forth. In other words, in different tables the same number can have different connotations.

It remains to establish why output tables can be considered as texts; indeed the fact that they can be interpreted is a necessary but not a sufficient condition. We must remember that, from a semiotic standpoint, the definition of text neither requires a necessary reference to the linguistic character of signs included in it (e.g. words, images, sounds etc.), nor does it require that the signs included in it have a sequential order. But, in any case, a text must have boundaries and its signs must be interpreted by reference to some code which allows us to establish correlations between expression and content forms. Otherwise, as Eco (1992) argues, if anything can be considered as text and if every text can say anything, we have no criteria for limiting the interpretation; that is, rather than pursuing hermeneutic aims, we are moving within a sort of Hermetism.

Having said that, my hypothesis is that the possibility of using *sorting criteria* is the most important property of output tables; in fact, only because we can sort the data according to some statistical norm can we uncover a *loaded meaning order*. To be precise, by choosing both a reference measure and a value column (i.e. a dimension), whenever we obtain a *syntagmatic chain* (see above: section n. 2) in which the closeness of the word-labels (each “near to” the other) makes sense.

Let me give an example involving the use of a *T-LAB*²¹ tool to analyse a corpus concerning tobacco problems. In this case the raw data matrix, structured as in Fig. 1.1., contains 8,665 rows (i.e. elementary contexts, roughly sentences) and 493 columns (i.e. words). A statistical algorithm using Correspondence Analysis produces – among other things - output tables with *Test-Values*²². This measure,

²¹ T-LAB package (www.tlab.it) offers many tools co-occurrence analyses, comparative analysis and thematic analysis.

²² A detailed description of this measure is found in Lebart, Morineau & Piron (1995: 123-125).

which correlates with absolute contributions²³, has two important properties: a threshold value to reject the null hypothesis ($p = 0.05$) and a sign (-/+)²⁴. This means that by sorting the values in increasing or decreasing order, according to whether the values are considered to be on the "negative" pole (-) or "positive" pole (+), it is possible to obtain a syntagmatic chain for each dimension (i.e. factorial axis). For example, Table 7.1 summarizes the characteristics of a factorial axis such as "cigarette consumption" versus "smoking controls", on the left and right extremity respectively of the same dimension.

| Pole (-) | Test Value | Pole (+) | Test Value |
|---------------------|------------|------------------------|------------|
| <i>tar</i> | -27.847 | <i>subsection</i> | 35.261 |
| <i>taste</i> | -24.401 | <i>amend</i> | 30.937 |
| <i>flavor</i> | -23.805 | <i>Section</i> | 30.447 |
| <i>brand</i> | -22.673 | <i>clause</i> | 27.790 |
| <i>ultra-light</i> | -21.981 | <i>Federal_Cigaret</i> | 27.006 |
| <i>smoker</i> | -19.886 | <i>establish</i> | 24.808 |
| <i>product</i> | -19.866 | <i>strike</i> | 24.391 |
| <i>male</i> | -17.783 | <i>federal</i> | 22.403 |
| <i>full</i> | -17.576 | <i>Secretary</i> | 19.774 |
| <i>Marlboro</i> | -17.564 | <i>health</i> | 18.952 |
| <i>satisfaction</i> | -16.716 | <i>appropriate</i> | 16.842 |
| <i>female</i> | -15.865 | <i>air_quality</i> | 16.629 |
| <i>cigarette</i> | -15.754 | <i>flight</i> | 16.056 |
| <i>purchase</i> | -15.682 | <i>committee</i> | 15.652 |
| <i>recall</i> | -14.551 | <i>cabin</i> | 15.584 |
| <i>score</i> | -14.530 | <i>aviation</i> | 15.116 |
| <i>nicotine</i> | -14.420 | <i>agency</i> | 14.361 |
| <i>choice</i> | -13.284 | <i>refer</i> | 13.832 |
| <i>level</i> | -11.858 | <i>congress</i> | 13.819 |
| | | | |

Table 7.1: Test Values obtained by Correspondence Analysis

Referring to previous sections of this paper (2 and 5), we can recognize expressions of both "cigarette consumption" and "smoking controls" as, at the same time, results of *abductive inference* and of *isotopy assumption*. In more detail: neither the

²³ The absolute contributions are measures which allow us to quantify the part played by each point (i.e. either row or column) in accounting for the inertia of each factorial axis (see Greenacre, 1984, pp. 67-70).

²⁴ The values with a statistical significance ($p = 0.05$) are either minor/equal to "-1.96" or major/equal to "1.96".

sub-table on the left nor the one on the right contains words like “*consumption*” and “*controls*”; but

- a) each of the sub-tables allows us to construct *some clusters*²⁵ (i.e. equivalence classes of semantic traits). For example, on the one hand, “*tar*”, “*taste*”, “*flavor*” and “*nicotine*” share characteristics referring to “*cigarette consumption*”; and, on the other hand, “*subsection*”, “*amend*”, “*section*”, “*clause*” and “*Federal Cigarette Labeling and Advertising Act*” share characteristics referring to “*control*” organisms.
- b) both the sub-tables, as a whole, propose a syntagmatic chain which – as a *text* – defines the boundaries of a linguistic *co-text* and refers to an extra-linguistic *context*²⁶. Therefore, because every text is produced within a specific social practice, we can use a *hermeneutic* approach; that is to say we can make inferences in an encyclopedia-like space.

In other words, taking paradigmatic relationships into account, i.e. the “associative relations” (Saussure, 1916) between semantic traits, we start from syntagmatic structures and always need to refer to their constraints, by bearing in mind how the different forms are determined by the logic of statistical tools. Then, according to the different steps of the analysis, we refer to bi-polar organization of single factorial axes (see above), to the shape of semantic spaces, to the significant characteristics of clusters, and so on.

Going on to the multi-semiotic character of the charts, in order to understand how they enable our interpretative skills, we can firstly consider the shape in Fig. 7.2. (see below): thirty-eight points in a bi-dimensional space. Also, if we don't have information about what the points represent and about what measures are used, we can activate our inferential processes. For example, if we have a statistical background, we can start by seeing the group of points on the left as a cluster. This implies that we are establishing a correlation between iconic expression forms and statistical content forms. According to Hjelmslev's model, in order to analyse the relationships between “forms” and “substances” in the content plane, in this case we need to refer to a shape like the one in Fig. 7.3. In it we recognize the forty words of Table 7.1.

²⁵ I have borrowed this concept from Rastier (1994)

²⁶ The distinction between co-text and (pragmatic) context derives from Bar-Hillel (1954)

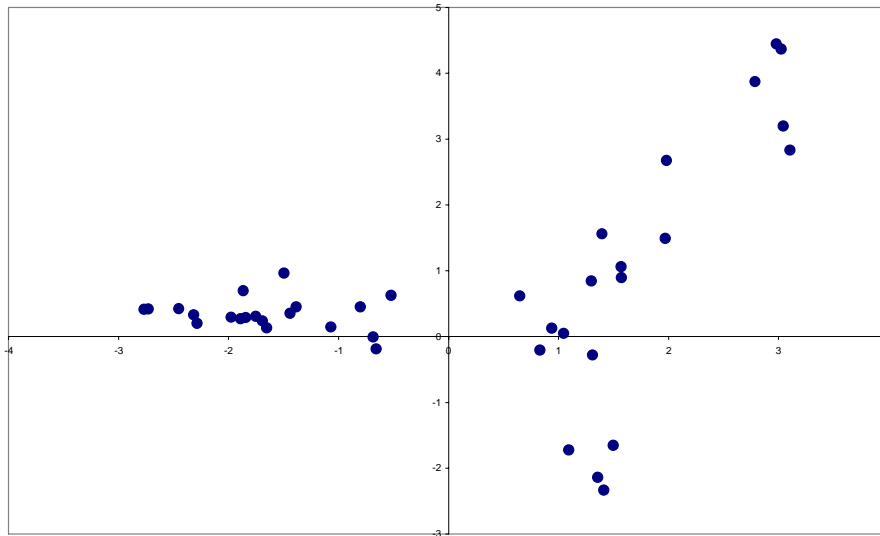


Fig. 7.2: A bi-dimensional space

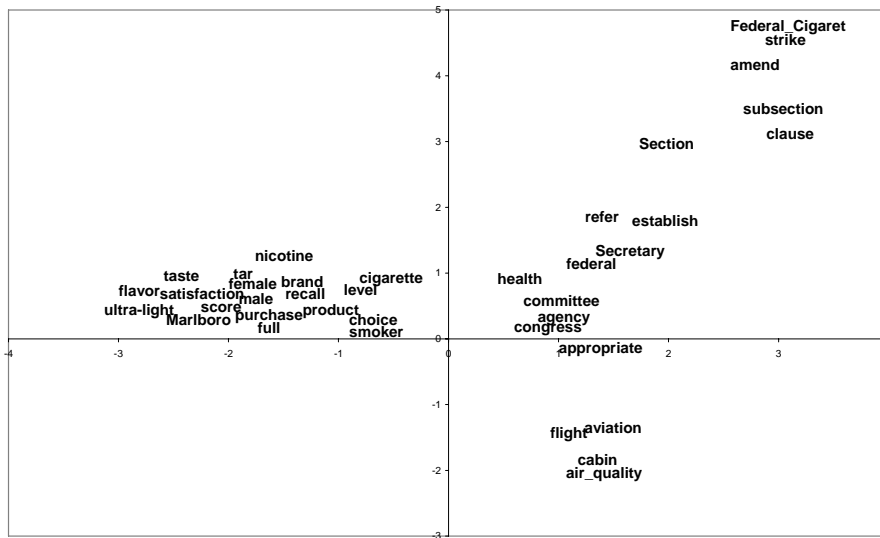


Fig. 7.3: A bi-dimensional space as multi-semiotic object

As multi-semiotic objects, Figs. 7.3 and 7.4 “talk” two different languages; in fact they are “coded” by different statistical models: Correspondence Analysis in the first and Multidimensional Scaling in the second. Both can be interpreted by looking at the dimensions (“x” and “y” axes) and/or at clusters (two or more words close together in the *semantic space*); but we cannot ignore that, in order to “synthesise” the information, the first “extracts” n-1 dimensions (where “n” is the number of words/columns), whereas the second extracts only a few dimensions.

Obviously, the possibilities of exploring word relationships like those in Fig. 7.4 also depend on the software we use. For example, using T-LAB we can click on each word (e.g. “flavor”) and obtain further charts like the one in Fig. 7.5, where the word selected is placed in the center and the others are distributed around it, each at a distance inverse to its degree of association.

In this case the chart refers to one “ordered” display which contains cosine coefficients³⁰, i.e. a measure of the “one to one” relationships between the selected word and all the others. But the evidence, e.g. the closeness of “flavor” with “taste” and “satisfaction” is not merely a linguistic and/or statistical issue. To clarify this, I suggest we compare Figs. 7.5 and 7.6 (see below).

Clearly, in order to understand the association between “Mary Magdalene” and “Grail” we need to know something about the *Da Vinci code* (Dan Brown); on the other hand, the association between “flavor” and “satisfaction” *seems* to be a fore-gone conclusion. But, in both cases, the links are a cultural artefact.

In all probability, in the next few years, the software for this kind of analysis will resemble present-day video-games, but in any case the “rules” of the game, that is the rules of interpretation, will continue to refer to linguistic and statistical models. However, this is not enough: in order to construct meaning – i.e. “new knowledge” – we cannot limit ourselves to *iconography*. In fact, because software outputs (e.g. charts) are icons, we need some *iconology*³¹; we need, for example, to refer to interpretative models of human/social sciences. Thus, we need to integrate statistical and hermeneutical tasks.

³⁰ In particular, in this case the cosine measures are as follows: full (0.343); taste (0.330); satisfaction (0.316); smoker (0.166); tar (0.144); product (0.141); positive (0.117); purchase (0.111); brand (0.104); choice (0.100); rate (0.094); score (0.077); recall (0.067); cigarette (0.057); difference (0.052); nicotine (0.035); male (0.034); female (0.026); information (0.005).

³¹ The distinction between iconography (referring to the description and classification of icons) and iconology (referring to the interpretation of icon symbolism) was suggested by E. Panofsky (1955).

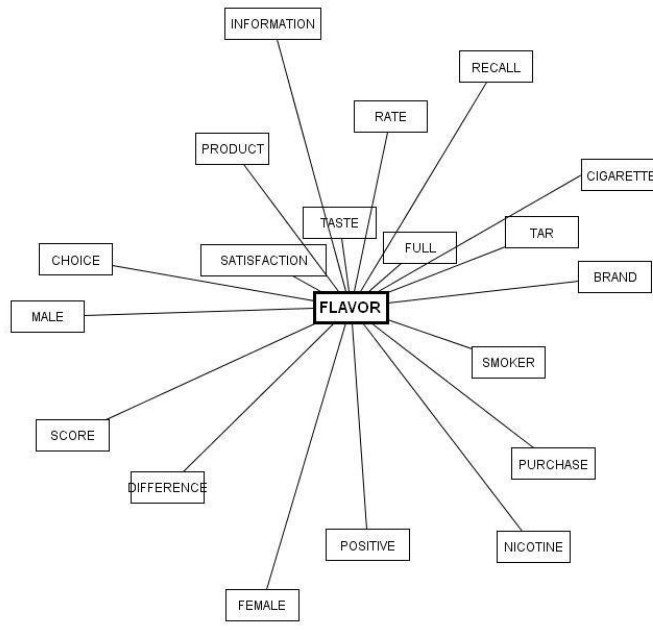


Fig. 7.5: Word Associations (Flavor)

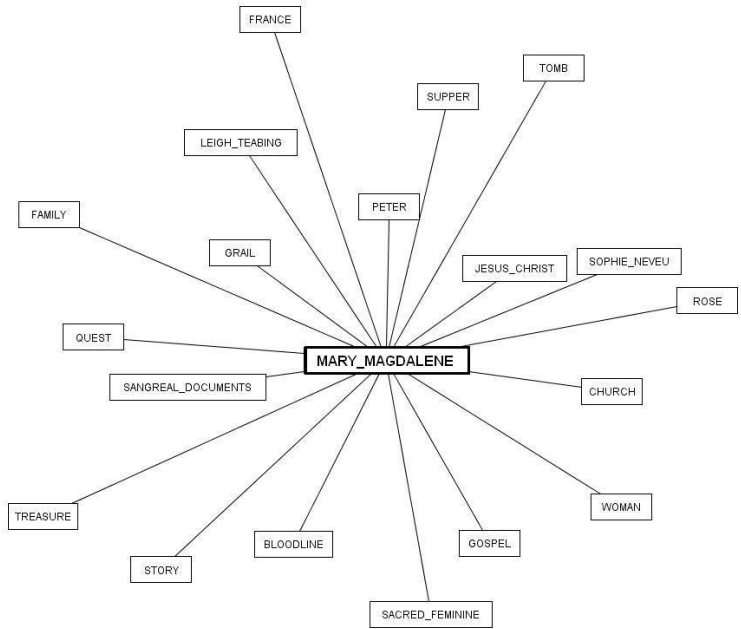


Fig. 7.6: Word Associations (Mary Magdalene)

7. Provisional conclusions

Referring to scientific literature, at the beginning of our brief excursion we focused on the following assertion:

- a) two (or more) words that tend to occur in similar linguistic contexts (i.e. to have similar co-occurrence patterns), tend to be positioned closer together in semantic space;
- b) two (or more) words that tend to occur in similar linguistic contexts (i.e. to have similar co-occurrence patterns) tend to resemble each other in meaning.

At this point, using semiotic tools, we can better understand how – referring to co-occurrence analysis – the *resemblance of meaning* appears as a sort of closeness within semantic spaces which are, at the same time, *given* and *constructed*, being, at the same time, *in* and *out of* our mind. In fact, as Saussure said, the associative relationships unite terms “*in absentia*” (1916, p. 123), whereas the relationships “*in praesentia*” - those we see or hear in different contexts - are only a basis for our inferential activity. In other words, the relationships “*in praesentia*” – i.e. concerning the expression forms – are a basis for identifying specific content forms as *seme clusters*, *isotopies*, *themes*, *cultural models* and so on.

According to Rastier and Cavazza, interpretative semantics offers both a rational representation of semantic content and a description of the interpretative process as a whole (1994, p. 232). Similarly, Pottier (1992, pp. 16-17) argues that semantics allows us to describe the pathways, both of the “*énonciateur*” (who talks) and of the “*interprétant*” (who interprets), as symmetrical processes: the former, moving from the “*monde référentiel*” (referential world), through “*conceptualisation*” produces “*discours*”; the latter, moving from “*discours*” and through its “*compréhension*” (understanding), can produce actions in the world.

I myself (Lancia, 2004) have used a model (see Fig. 8.1) for representing the cycle of text production and of text interpretation. In this model “*world*” stands for anything that can be represented, including texts (i.e. objects, events, social relations, emotions, etc.), whereas “*representations*” and “*representations of representations*” – both resulting from psycho-socio-cultural processes – indicate two ways of building texts, i.e. two ways of relating expression and content forms.

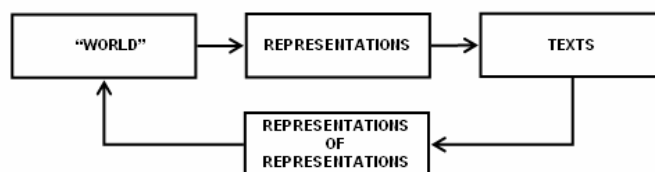


Fig. 8.1: Production and interpretation of texts.

According to this model, all the phases of text analysis, including those which are computer-assisted, propose *representations of representations*. But, in the end, this is only a way of describing something which needs to be explained. In particular, in my own view, we need to explain each time why, in specific contexts, people choose specific words and assemble them in a meaningful order. On this matter, semiotics only offers generalisations, a sort of metalanguage to talk about. Other scientific disciplines have the task of finding specific answers.

Here, that is within a methodological framework, we can just consider how the “representations of representations” constructed by the semiotic models and the geometric logic of multidimensional analysis are interrelated.

According to Greimas’s semiotics (Greimas & Courtés, 1979), the meaning construction requires grasping similarities (“and ... and” relations) and differences (“or ... or” relations) at the same time. In particular, according to its models, similarity/difference relations are organized from elementary structures of signification which can be represented by semiotic squares.

For example, a semiotic square concerning tobacco consumption could be as follows (Fig. 8.2).

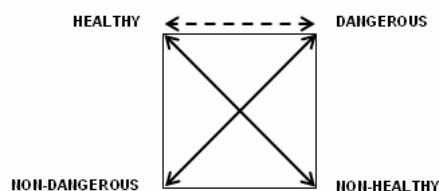


Fig. 8.2: Semiotic Square

Even though this kind of bi-dimensional representation can have psychological and cultural connotations, its structure comes from Aristotle’s logic. More precisely, the relation on the top side (“healthy” vs “dangerous”) concerns contrary opposition, the relations on the diagonals concerns contradictory opposition and the relations on the left-right sides concern complementary terms. In the same square, contrary/contradictory oppositions refer to differences (“or ... or”), whereas complementary relations refer to similarities (“and ... “and”).

In semiotics’ terms each “dimension” of this square is a “semantic axis”.

The same expression (i.e. “semantic axis”) is often used in order to interpret the bi-dimensional space obtained by means of statistical techniques (see above Figg. 7.3 and 7.4). I myself interpreted the first dimension of factorial space depicted in Fig. 7.3 as a “semantic axis” in which “*cigarette consumption*” and “*smoking controls*” are opposed.

According to Benzecri, one of the mathematicians that has contributed most to defining the Correspondence Analysis model:

Understanding a factorial axis means finding what is similar, firstly all that is on the right of the origin (barycentre), and secondly all that is on the left of it, and then expressing concisely and exactly the opposition between the two extremes (BENZÉCRI J.P & F., 1984, p. 302).³²

With this assertion, while describing an interpretation method, the author communicates a specific conception of factors as organizers of oppositions (i.e. contrasting relationships) between sets of objects which, within the “semantic space”, are represented by closer points (“all that” is on the right and “all that” is on the left of the origin).

In the same way, this assertion can be considered an illustration of abduction inference.

Even though the semiotic square can be used within an abduction process, the construction of its axes follows just logic rules and its structure is in same way invariant. On the contrary, both the construction and the interpretation of factorial axes, as they depend on how the data are arranged within matrices or charts, allow us to discover meaning structures (i.e. similarities and differences) strictly linked to the contexts which the analysis comes from. Not by chance, we refer to Correspondence Analysis like an “exploratory” method.

Some years ago, Richard Rorty (1992) claimed that texts are no different from other objects (“objects as rocks and trees and quarks”) and that Eco’s description of the universe of semiosis and of human culture – like a labyrinth structured according to a network of interpretants – “seems to be a good description of the universe *tout court*” (ibid., p. 99). According to Rorty’s pragmatism, we can assume that the origin and structure of texts can be studied like the origin and structure of living creatures; but – to use a metaphor – a text has no genome. In other words, each text is a *token* without *type* and, in accounting for its meaning, we cannot refer to general laws. We can only make explicit the rules that we follow for constructing the representation of the object studied.

We can construct our interpretative trajectories each time within specific encyclopedic frames and, if we use statistical tools, we must not forget that, in constructing representations of the texts analysed, at the same time they fix the limits of our interpretation.

To put it in semiotics’ terms:

(a) when analysing texts we “observe” the expression forms only;

³² “... interpréter une axe, c’est trouver ce qu’il y a d’analogie d’une part entre tout ce qui est écrit à droite de l’origine, d’autre part entre tout ce qui s’écarte à gauche ; et exprimer avec concision et exactitude, l’opposition entre les deux extrêmes.”

(b) the way how they are arranged, also within “semantic spaces”, depends on the logic of content structures, including those of analytical tools;

(c) in consequence we can just explain how and why, in specific contexts, we make inferences concerning specific relationships between expression and content structures.

References

- BARTHES, R. (1964). *Eléments de sémiologie*, Paris: Seuil. (English translation by A. Lavers & C. Smith. New York: Hill and Wang, 1968)
- BAR-HILLEL, Y. (1954). Indexical Expressions. *Mind*, LXIII, pp. 359-379.
- BENZECRI J.P & F. (1984), *Pratique de l'analyse des données. Analyse des correspondances & Classification*, Dunod, Paris.
- BLOOMFIELD, L. (1933). *Language*. New York: Holt, Rinehart & Winston
- BRUNER, J.S., GOODNOW, J.J., AUSTIN, G.A.(1956). *A Study of Thinking*. New Brunswick & London: Transaction Publishers.
- ECO, U. (1981). *Significato*. In Enciclopedia, XII, pp. 831-875. Torino: Einaudi.
- (1983). *Horns, Hooves, Insteps* in ECO, U. and SEBEOK, T.A. (Ed.) *The Sign of Three: Dupin, Holmes, Peirce*, pp. 199-219. Bloomington: Indiana University Press.
 - (1984). *Semiotica e filosofia del linguaggio*. Torino: Einaudi. (English edition, *Semiotics and the Philosophy of Language*.. Bloomington: Indiana University Press, 1984).
 - (1990). *I limiti dell'interpretazione*. Milano: Bompiani (English edition, *The Limits of the Interpretation*.. Bloomington: Indiana University Press, 1994).
 - (1992) Ed. *Interpretation and Overinterpretation*. New York: Cambridge University Press.
- GREIMAS, A.J. (1966). *Sémantique structurale*. Paris: Larousse.
- GREIMAS A.J., COURTES J. (1979), *Sémiotique. Dictionnaire raisonné de la théorie du langage*, Paris. Haschette.
- GREENACRE, M.J. (1984). *Theory and Applications of Correspondence Analysis*. New York: Academic Press
- HARRIS, Z.S. (1951). *Methods in Structural Linguistics*. Chicago: Univ. of Chicago Press.
- (1952). Discourse analysis. *Language* 28, N 1, pp. 1-30. (Repr. in Harris 1981: 107-142).
 - (1954). Distributional structure. *Word* 10, N 2-3, pp. 146-162. (Repr. in Harris 1981: 3-22).
 - (1981). *Papers on syntax*. Ed. by H. Hiz. Dordrecht/Holland: D. Reidel.
 - (1991). *A theory of Language and Information: a mathematical approach*. Oxford & New York: Clarendon Press.
 - (2002). *The background of transformational and metalanguage analysis* In Nevin E.B. (Ed.), 2002, pp. 1-15.
- HEMPEL C.G. (1952), *Fundamentals of Concept Formation in Empirical Science*, Chicago: The University of Chicago Press.
- HJELMSLEV, L. (1943). *Omkring sprogteoriens grundlaeggesle*. Kobenhavn: Munksgraad. (English translation by F.J. Whitfield, *Prolegomena to a Theory of Language*. Madison: The University of Wisconsin Press, 1961).
- (1959). *Essais linguistiques, Travaux du Cercle Linguistique de Copenhague* (vol. XII). Copenhagen: Nordisk Sprong-og Kulturforlag.
- JAKOBSON, R. (1963). *Essais de linguistique générale*. Paris : Editions de Minuit.
- LANCIA, F. (2004). *Strumenti per l'analisi dei testi. Introduzione all'uso di T-LAB*. Milano: FrancoAngeli.

- LEBART, L., MORINEAU, A., PIRON, M. (1995). *Statistique exploratoire multidimensionnelle*. Paris: Dunod
- LOTMAN, M.Y. (1990). *Universe of Mind: A Semiotic Theory of Culture*. Shukman A. tr. Bloomington and Indianapolis: Indiana University Press.
- MATTE-BLANCO, I. (1975). *The Unconscious as Infinite Sets. An Essay in Bi-Logic*. London: G. Duckworth & Company Ltd.
- MARTINET, A. (1960). *Eléments de linguistique générale*. Paris: Colin.
- NEVIN, E.B. (ED). (2002). *The legacy of Zellig Harris. Language and information into 21st century. Vol. I. Philosophy of science, syntax and semantics*. Amsterdam/Philadelphia: J. Benjamins Publishing Company.
- OSGOOD, C. R. (1959). The representation model and relevant research methods. In Ithiel de Sola Pool (Ed.), *Trends in content analysis* (pp. 33-88). Urbana: University of Illinois Press.
- PANOFSKY, E. (1955). *Meaning in the Visual Arts. Papers in and on Art History*. New York: Doubleday.
- PEIRCE, C.S. (1931-58). *Collected Papers*. Cambridge: Harvard University Press.
- POTTIER, B. (1965). *La définition sémantique dans le dictionnaires*, in « Travaux de linguistique et de littérature », III, n. 3, pp. 33-40.
- (1974). *Linguistique générale, théorie et description*. Paris: Klincksieck..
 - (1992). *Sémantique générale*. Paris: P.U.F..
- PROPP, V.JA. (1928). *Morfologija skazski*. Lelingrad: Academia (English translation by L. Scott, *Morphology of the Folktale*. Austin: University of Texas Press, 1968).
- RASTIER, F. (1987). *Sémantique interprétative*. Paris: P.U.F.
- (1995), “La sémantique des thèmes ou le voyage sentimental”, in F. Rastier (Ed.), *L’analyse thématique des données textuelles*. Paris : Didier, 1995, pp. 223-249.
- RASTIER, F., CAVAZZA, M., ABEILLE, A. (1994). *Sémantique pour l’analyse: de la linguistique à l’informatique*. Paris : Masson (English translation by R. Lawrence Marks, *Semantics for Descriptions. From Linguistics to Computer Science*, Stanford: CSLI Publications, 2002).
- RYCKMAN, T.A. (2002). *Method and theory in Harris’s Grammar of Information*. In Nevin E.B. (Ed.), 2002, pp. 19-37.
- RICOEUR, P. (1969). *Le conflict des interprétation*. Paris: Seuil
- RORTY , R. (1992). “The pragmatist’s progress”, in ECO U. (Ed.). *Interpretation and Over-interpretation*. New York: Cambridge University Press.
- SAPIR, E. (1949). *Culture, Language and Personality*. The Regents of the University of California
- SAUSSURE (de), F.(1916). *Cours de Linguistique générale*, Lusanne-Paris: Payot. (English translation by W. Baskin, *Course in General Linguistics*. New York: McGraw-Hill, 1966)
- WELBY, V. (1977). *Semiotic and Significs: The Correspondence Between Charles S. Peirce and Victoria Lady Welby*. Ed. by Charles S. Hardwick & J. Cook. Bloomington: Indiana University Press.
- WHORF, B. (1956). *Language, Thought and Reality*. Cambridge: Mass. MIT Press