Ludovic Lebart
John Wisdom
Jean-François Steiner
Marie Piron

The Semiometric Challenge:

Words, Lifestyles and Values

//

The Semiometric Challenge:

Words, Lifestyles and Values

© *L2C* April 2014

ISBN 978-2-953777222

Table of Contents

Preamble : How semiometry began 9

Introduction : What is Semiometry ? 12

Chapter 1
Words, respondents… a structure ! 15

1.1 One questionnaire: a list of words 17
1.2 A representative survey 21
1.3 Principal components analysis 24
1.4 The stable axes: "semiometric axes" 26
1.4.1 Axis 1 : A methodological axis of participation 27
1.4.2 Axis 2 : « Duty / Pleasure » 28
1.4.3 Axis 3 : « Attachment / Detachment » 29
1.4.4 Semiometric plane of axes 2 and 3 30
1.4.5 Axis 4 : « Sublimation / Materialism » 32
1.4.6 Axis 5 : « Idealization / Pragmatism» 33
1.4.7 Axis 6 : « Humility / Sovereignty» 34
1.4.8 The following axes 35
1.5 The semiometric planes 37
1.5.1 The multiple facets of "Pleasure" 37
1.5.2 The multiple facets « Duty » 43
1.5.3 The semantic dimensions of words themselves 44
1.6 Conclusion 45

Chapter 2
Stability of the Semiometric structure 47

2.1 Empirical stability per sub-populations 48
2.1.1 Diachronic stability 48
2.1.2 Stability per country 53
2.1.3 Stability according to sex categories 65

2.1.4 Stability according to age 68
2.2 Confidence intervals for eigenvalues 69
2.3 Stability through re-sampling 71
2.3.1 Partial Bootstrap 72
2.3.2 Total Bootstrap 73
2.4 Structural stability 80
2.5 Stability with respect to coding 82
2.6 Analysis with classical factor analysis 84
2.7 Conclusion 86

Chapter 3
Scorings, willingness to answer, attitudes 87

3.1 Scoring, participation, and size factor 88
3.1.1 The scoring effect 88
3.1.2 The participation effect 88
3.1.3 The size factors 89
3.2 Creating and positioning control variables 89
3.3 Scoring in four European countries 92
3.4 Typology derived from control variables 94
3.5 Relationships between the scores and the characteristics of individuals 97
3.6 The *rating planes* in Spain, UK, and Germany 102
3.6.1 Scoring effects in Spain 102
3.6.2 Scoring effects in Britain 103
3.6.3 Scoring effects in Germany 104
3.7 Correction of scores by double centring 105
3.8 Evolution of scoring (France, from 1990 to 2002) 110
3.9 Conclusion 114

Chapter 4
The spontaneous choice of words 115

4.1 Setting up the experiment 116
4.2 Exploring the responses 121
4.3 Principal axes of spontaneous responses 125
4.4 Characteristics of respondents 127
4.5 Semiometry and open questions together 132
4.6 Conclusion 138

Chapter 5
Semantic structures and Semiometry 141

5.1 Semantic neighbourhood of the semiometric questionnaire 142
5.1.1 An attempt to describe a purely semantic network 142
5.2 A semantic network of French verbs 148
5.3 "Internal semantic fields" 153
5.3.1 Hierarchical representation of semantic proximity 155
5.3.2 Planar representation of semantic proximity 156
5.3.3 Semantic nuances provided by Semiometry 159
5.4 The two facets of meaning: denotation and connotation 161
5.4.1 Mental mechanisms implemented by the respondents 162
5.4.2 Semantic space: connotations and correlations 163
5.4.3 A structured and semantically continuous space 165
5.4.4 Meaning of the dimensions of the semiometric space 169

Chapter 6
Towards a few interpretations 171

6.1 A first approach 172
6.2 A psychoanalytical interpretation 176

Chapter 7
Some semiometric applications 183

7.1 The words characteristic of males and females 184
7.2 The characteristic words of two age groups 186
7.3 Television addicts 187
7.4 The characteristic words of two groups of young people 188
7.5 The characteristic words of two age groups over 60 190
7.6 Persons who attributed a "7" to the word *God* 191

7.7 Changing values of the French between 1996 and 1998 193
7.8 Conclusion *(chapter 7)* 197

Conclusion 198

Appendix 1
Elements of Multivariate Descriptive Statistical Analysis 204

A1.1 A reminder: principles of exploratory multivariate methods 205
A1.2 Factorial methods: technical aspects 208
A1.3 Principal Components Analysis 210
A1.4 Correspondence Analysis 215
A1.5 Logarithmic analysis 217
A1.6 Factor analysis (common and specific factors) 217
A1.7 Methods of hierarchical clustering 218
A1.8 The self-organizing (Kohonen) maps 220
A1.9 Validation Tools 234

Appendix 2
Multilingual lists of words 231

References 236

Preamble

How Semiometry began...

Semiometry stemmed from an experience that almost every man has undergone in his lifetime, at least once during his very early childhood, consciously and lucidly, or in such a fleeting way that the mind has not kept any memorable trace of it.

I mean that moment when, suddenly, for some mysterious reason, we cease to communicate with the people and objects that surround us, when they stop talking to us, when they do not tell us anything anymore, when we find ourselves alone amidst a mineral world filled with dead objects, and in which, with our ice-bound minds, we can no longer find our bearings. A few moments or a few days earlier – the phenomenon was triggered off in a more or less brutal way and its duration is extremely variable – the world was full of life, sounds, colours, tastes, smells, sensations, objects and beings that we are attracted to, and others which repelled us. Whether we loved them or hated them, we were able to name them, and their names meant something to us. Names or nouns that were endowed with a suggestive power, which, solely by uttering them, could give rise to a procession of sensations akin to their signatures: a *friend* was something warm and precious, rare too, solid, very solid, yet at the same time very fragile. A *house* was also something warm, but warm in a different way; while the warmth from the *friend* radiated from a point in the middle of the chest, that of the *house* seemed to come down from some unknown place, outside of us, and we were slowly infused with happiness. The heat from the friend was also more intense, more violent than that of the house. It was a red heat while the heat of the house was instead rather bluish. The word *war* also meant something. It conjured up another world of sensations that seemed to belong to an area that was poles apart from the area that belonged to the word *house*, and that seemed to have a common territory with the word *friend*. The word *tree* also had a meaning and so did the

words *river, path, school, summer, forest, sadness* and even the word *boredom*, not to mention *despair*.

In short, thanks to the signs that allow us to describe it, the world made sense, it was intelligible for us. We knew how to find our way in it. It was our natural environment, so natural that the thought did not even occur to us that we could be excluded from it, other than by death. We were so accustomed to seeing it again every morning when we woke up, always identical to itself – with the exception of a few objects from time to time, objects which had migrated from one area to another – that we had not thought of this obvious truth – so strongly obvious, however, once one has realized it – that meaning is not consubstantial with the world around us, that meaning is only one property that we attribute to it, only another reality that we project onto it. It was necessary to experience the *loss of sense* to make us understand this very commonplace fact that meaning is a phenomenon both separate and independent from the objects through which it is embodied, that it is a reality, at least psychological, if not physical, and therefore we must be able to study it independently of the objects for which it is used as an attribute and through whose mediation it is commonly perceived.

But this first experience contained another one, which, in a way, was the next logical step: not only was there sense, but also senses. Thus to the word *castle*, could not only be attributed (or it could be lacking) a sense, but once it possessed this property, it necessarily follows that it also had ONE or MORE meanings.

And this second experience, in turn, implicitly contained a third one: the various meanings revealed through the words were more or less close to one another. Thus the sense of the word *castle* was intuitively closer to that of *home* than it was to the word *mathematics*, but even closer to the word *palace*. The meaning of the word *house* seemed equidistant from the words *castle* and *palace*. The sense of the word *mathematics* was so far from the meaning of the latter three that it was hard to tell if it was closer to one of them in the same way that three very close points seem equally distant from a fourth, if the latter is at a great distance; though, if we look closer, the meaning of *castle* seemed slightly less distant from that of *mathematics* than the sense of the word *house* was. As for the meanings of other words, they were just intuitions, but so tenuous, so subtle, so elusive that they vanished as soon as we tried to reason with them, and yet reappeared almost instantaneously, like an indelible stain, the moment we turned the harsh light of our attention away from them. These intuitions, which made

us attribute distance of meaning between words, therefore had a foundation: they were the expression of a reality, which although seemingly inaccessible with an objective observational tool, was nonetheless stable, and therefore theoretically measurable. One could therefore imagine a space that contained all the words of a language, arranged one with respect to the others, depending on the distance separating their meaning. It would be a sort of geographical dictionary in which words are not arranged in alphabetical order – which, however practical as this is, is nonetheless a relatively arbitrary classification – but arranged, one with respect to the others, according to their meaning: an organic dictionary in which each word is defined (mathematically) from its position in the set containing all the others. With reference to Ferdinand de Saussure, who imagined a science of signs, and who proposed to call it "semiology", I baptized my endeavour Semiometry[1].

[1] The idea that there is a "space of meaning" that can be modeled goes back to Charles E. Osgood. He, indeed, in the '50s, invented the first device designed to rebuild this space. This device differs radically from that used by Semiometry, and I knew nothing of the work of Osgood when I coined it, but it was appropriate to name the man who had introduced such a rich concept as that of the "space of meaning", cf. Osgood (1965), Osgood et al (1957).

Introduction:

What is Semiometry?

If we had to answer the question: “What is Semiometry?” It is easy to answer naively: “It is a long list of words to which a score is assigned by thousands of people (according to the more or less pleasant or unpleasant character of the words)”.

Let us try all the same some more academic definitions:

- Semiometry is the descriptive technique of certain types of semantic links between words;

- Semiometry is a more fundamental research tool allowing us to get nearer (cautiously and circumspectly) to concepts such as the “collective subconscious” or the individual one, also allowing one to highlight and validate structural features (the finesse of which sometimes leaves one speechless ...);

- Semiometry is a tool to describe systems of values and lifestyles, for the purpose of psycho-sociological studies, and applications in marketing.

Semiometry is actually a bit of all those things, and that is what this book aims to show.

The first “naïve” definition suggests an experimental protocol, both developed and rigorous, to which the first chapter of this book is devoted. This experimental phase, which consists of many repeated surveys over time and space (especially in many Western countries), constitutes the basis of the whole semiometric methodology, a basis which will be validated by powerful statistical processing, albeit blind, *i.e.* independent of any interpretive framework. Note that, at the presentation level in this first chapter as in those that will follow, the technical developments will be reported in the appendix to allow a more straightforward reading of the

contents, results, and general methodology, smoothly and without any digressions.

The theme of the stability of structures (stability over space and time) is the subject of the second chapter, and is central to the problem of Semiometry. This stability is established from repeated independent investigations, but also from the more technical proof of validity (re-sampling techniques belonging to computational statistics). It is this stability that allows one to assert the reality of certain structural features observed; and which gives semiometric maps, widely used in applications, the value of *tracking grids*.

The structural stability for a given questionnaire (including the choice of a list of words) being established in different countries and at different times, there still remains the problem of the intrinsic character of this structure, that is to say, its independence *vis-à-vis* the questionnaire itself, and not only *vis-à-vis* the population to whom this questionnaire has been submitted.

The next three chapters evoke our methodological research on this theme.

To understand better the nature of the statistical data observed, various experiments were conducted: the study of certain semantic networks attached to the words chosen, and then the more general study of semantic networks (chapter 3); an “open” questionnaire allowing the interviewee to choose spontaneously, without any preliminary list, the words he or she finds pleasant or unpleasant (Chapter 4). These experiments are also an opportunity to address other avenues of research generated by the semiometric toolbox, and the multidimensional structure it reveals. The fifth chapter is devoted to the technical problem, often evaded, of the link between the attitude of the respondent in relation to the questionnaire or survey and the actual content of the questionnaire or the purpose of the investigation.

Slightly in the background compared to the previous techniques and empirical considerations, a tentative interpretation of the stable structures of Semiometry is proposed in the sixth chapter. Whatever the reader's level of attachment is to the ideas presented, he or she is unlikely to be insensitive to the amazing convergence mentioned.

Finally, in the seventh chapter, devoted to applications of semiometric tools, we shall present different types of applications: the synchronic comparison of two populations (male/female, young/old, believer/non believer, etc.) and the diachronic comparison of a sample of the French population at two different times.

Chapter 8 is actually a conclusion, which can only be temporary as part of an endeavour that is so perilously multi-disciplinary; furthermore it is also an invitation to further research in an area that we propose to call "structural statistics".

At the end come the annexes, austere warehouses of materials, results, technical reminders, in which the reader, motivated by the previous chapters, can find answers to many questions raised during his/her trip through this multidimensional space.

CHAPTER 1

Words, respondents… a structure!

The socio-economic studies using opinion surveys generally characterize the individuals interviewed (consumers, customers, voters, patients, investors, readers ...) by what is commonly called descriptive variables, or basic socio-demographic traits (sex, age, educational level, occupation, marital status, income, to mention the most common). These characteristics have undoubtedly a capacity to predict behaviour. In the field of marketing studies, however, it has appeared necessary, over the previous decades, to describe more accurately the respondents from questionnaires using more subjective or more psycho-sociological variables. It is usually under the name of "lifestyle" studies or studies of "socio-cultural trends" that these attempts at description have been conducted[1].

If one had to find among existing methodologies the approach closest to the one we mean here by "semiometry", we probably ought to refer to lifestyle studies, despite significant differences in motivation, parentage, and even principles[2].

These studies bring into play questionnaires that are usually quite bulky, mainly encompassing matters of opinion.

---

[1] We can find an overview and critical synthesis of these studies in the book by Pierre Valette-Florence: *Lifestyles. Critical Review and Prospects*. Nathan, Paris, 1994 (in French).

[2] See Steiner and Auliard (1992).

Take the example of two types of survey related to lifestyle studies whose questionnaires are or were transparent, that is to say, available in full to users: The survey system "Agoramétrie"[1] and the one about "Living conditions and aspirations of the French[2]".

For the first type mentioned, let us take, as an example, the wording of one of the permanent questions of the questionnaire: "*Do you: strongly disagree, more or less disagree, perhaps agree, more or less agree, fully agree, with the following sentence: '*We must restore the death penalty'*."*

For the second batch of surveys cited, and in the same vein, next to issues also involving a level of agreement, are questions like the following[3]: "*Among these opinions, what is the one closest to yours: marriage is: 1. '*A union that is not dissolvable*', 2. '*A union that can be dissolved in very serious cases'*, 3. '*A union that can be dissolved by simple agreement of both parties*', 4. '*Do not know'*?"*.

The semiometric questionnaire may be regarded as an extremely refined form of the lifestyle one, because, as we shall see, one does not give one's opinion on proposals presented in the form of more or less terse sentences, but one assigns scores to words based on the pleasant or unpleasant sensations they are evocative of.

We shall present in this chapter the questionnaire, then how the basic semiometric tool works: the survey (how and by whom we fill in the questionnaires), and the statistical processing with the emergence of the *basic semiometric structure*.

---

[1] This survey system was founded and directed by Jean-Pierre Pagès and Georges Morlat (see Fabre et al., 1981).

[2] The system of survey "Living conditions and aspirations of the French" (CREDOC), developed in 1978 with the collaboration of Jacques Antoine, cf. Lebart and Houzel (1980), Lebart (1986) (in French).

[3] Cf. Nicole Tabard: *Besoin et Aspirations des familles et des jeunes, Coll. Etudes CAF, 16, CNAF, Paris, 1974.*

1.1 One questionnaire: a list of words

Semiometry is based on a list of 210 words (see Table 1.1) supposed to represent, directly or indirectly, the main values of Western society. Those words were selected on the basis of four criteria: non-consensuality, semantic univocity, semantic stability, and evocative power. In addition, in order to ensure some representativeness while avoiding, at least partly, too rough a projection of the subjectivity of the author, this list has been established, originally, from some three hundred roots[1]. The original list, which included 306 words, was subsequently worked on again, based on the results, first of successive statistical analyses, so as to highlight areas of too high and too low a density; and, secondly, on creative groups who were asked to imagine the missing words.

This initial list was reduced, for economic and ergonomic reasons, first to 286 words, then to the current 210 word list. If by doing so, this has inevitably led to a loss of information, it did not affect the stability of the structure revealed by the complete word list.

These words can be either nouns, adjectives, or verbs. Some nouns are accompanied by the indefinite article. The choice that was made concerning the grammatical category of the word and the article accompanying it is intuitive and therefore more or less arbitrary. This choice corresponded to the desire to find the signifier that, first, best fits the underlying concept that it was intended to measure. Secondly, we chose a word the stimulus effect of which would be the strongest, but also the sound of which would be the most euphonious.

[1] Note that, for instance, 95% of the words of the first part of the Old Testament [Pentateuch] stem from these 300 roots. The question of the definition of a corpus of words that could claim a certain "representativeness" of concepts with which man perceives and describes the world and its relationship with him seemed to Jean-Francois Steiner as simple to ask and as difficult to solve when he learnt about, fortuitously, this particularity of the biblical text. The privileged position that the text had occupied in Western culture and its attempt to account for the major spiritual, emotional, sensitive experiences, etc. between the moment of one's birth and that of one's death, led him to think that from the meaning of these 300 roots could be established a list of words that would serve as a matrix for the devised questionnaire.

While in some cases the choice was necessary for obvious reasons: nudity and not *a nudity*, theatre and not *a theatre*, an island and not *the island*, a wall and not *the wall*; in others, it was less immediate, and we chose what seems to us the most suggestive form: *to seduce* and not *seduction*, *to criticize* and not *criticism*, *desire*, not *a desire* or *to desire*, *a perfume* and not *perfume*, a *wedding* and not *marriage*, *humble* and not *humility*, *intimate* and not *intimacy*, *the desert* and not *a desert*...

In many lifestyle surveys, the wording of the questions generally concerns the value or opinion of the interviewee with respect to some proposed item. This has the effect of placing him/her in a position to judge from the outside, thus leading him/her more or less to rationalise the response. Sensing that the information collected would be all the better in quality if it were provided from a more subconscious level, it seemed more appropriate to try to measure the affect triggered off by the word. The latter would play (hopefully) the role of a kind of stimulus that would bypass consciousness and thus thwart the inevitable distortions produced by the multifarious forms of censorship that consciousness places between the subject and his/her subconscious mind. Thus respondents were asked to mark the words in terms of the sensation, pleasant or unpleasant, that were conjured up when read[1].

[1] In practice, the semiometric questionnaire is self-administered. This procedure certainly has known disadvantages, particularly as regards the rate of return, but it appeared to constitute the protocol most suited to a questionnaire asking from the interviewee coolness, concentration and discretion. This mode of execution also avoids the bias that could be created by the voice of an investigator interpreting the words according to his/her own sensations.

Table 1.1 The Semiometric words list

Persistence	To_criticize	Free	Morals	Respect
To buy	Danger	To recover	Death	To dream
To admire	A challenge	War	A wall	Rebellion
To love	The desert	To inherit	Music	Wealth
Ambition	Desire	A hero	A mystery	Rigid
Soul	Disorder	Honest	To swim	To laugh
Friendship	Detachment	Honour	A birth	Robust
Anguish	God	Humble	A nest	To break
An animal	Different	Humour	Noble	Red
A tree	Discipline	An island	A knot	Craftiness
Money	Softness	Huge	Black	Sacred
An armour	Doubt	Immobile	Nudity	A sacrifice
Art	Dynamic	Stranger	To obey	Wild
Clever	Water	Industry	The ocean	Science
Attachment	School	Infinite	Gold	Secret
To attack	To economize	To forbid	A storm	To seduce
Audacity	To write	To interrogate	Original	Sensuous
An adventurer	Effective	Intimate	Peace	To tend
An entrepreneur	An effort	An inventor	Forgiveness	A soldier
A jewel	Elegance	Irony	Perfume	A peak
Blue	The elite	A game	Patience	Flexibility
Bohemian	An emotion	Justice	Homeland	Masterly
A present	Childhood	A maze	The skin	Sublime
Countryside	To teach	Lightness	Perfection	Tenderness
A caress	Together	A book	Poetry	Theatre
A ceremony	To climb	Logic	Politeness	Tradition
A certainty	Eternal	The law	Precious	To betray
Change	A foreigner	The moon	Accuracy	Work
Charitable	An escape	Magic	A priest	Practical
Carnal	The family	A house	To produce	Green
Hunt	A fault	To master	Property	Victory
A researcher	To fertilize	Wedding	To protect	Emptiness
To command	Feminine	A mask	Prudence	To age
Trade	Firmness	Material	Power	Virile
Solid	Fire	Maternal	To punish	Speed
Confidence	Loyalty	Mistrust	Purity	Headstrong
Comfort	A flower	Metallic	A question	Voluptuous
To conquer	A river	Slimness	Refined	
To console	Faith	Fashion	Reason	
To_build	A border	Moderation	A reward	
Courage	A rifle	Modesty	Meditation	
A creator	Liveliness	Mellow	To ponder	
A cry	Glory	A mountain	A rule	

Figure 1.1: Excerpt from the "word questionnaire"

This exercise is to complete a questionnaire about your reactions to everyday language. It's not difficult, and you can answer as if it's a game.

On the following pages you will find a list of words. We would like you to mark them on a scale of -3 to +3 according to how they make you feel, between very negative (-3), or very positive (+3). The scale allows you to express the extent to which the words make you feel negative or positive.

	Very negative	Fairly negative	Slightly negative	No feeling	Slightly positive	Fairly positive	Very positive
Word	-3	-2	-1	0	+1	+2	+3

This is not an exam. There are no correct answers. **The best answers are the most spontaneous ones.** Some words may raise mixed feelings, making you hesitate. In this case, it's very important not to stop and think; put the first mark that comes to mind and skip to the following word.

Remember: It is vital that you give an answer to every word (do not skip any) and also that you ONLY mark one box per word

	Very negative						Very positive
Absolute	-3	-2	-1	0	+1	+2	+3
Persistence	-3	-2	-1	0	+1	+2	+3
To buy	-3	-2	-1	0	+1	+2	+3
To admire	-3	-2	-1	0	+1	+2	+3
To love	-3	-2	-1	0	+1	+2	+3
Ambition	-3	-2	-1	0	+1	+2	+3
The soul	-3	-2	-1	0	+1	+2	+3
Friendship	-3	-2	-1	0	+1	+2	+3
Anguish	-3	-2	-1	0	+1	+2	+3
An animal	-3	-2	-1	0	+1	+2	+3
Attachment	-3	-2	-1	0	+1	+2	+3
To attack	-3	-2	-1	0	+1	+2	+3
Audacity	-3	-2	-1	0	+1	+2	+3

Also motivated by the same desire to obtain answers that are as "natural" as possible, the interviewees were recommended in a short introductory text, to answer only once – and as spontaneously as possible - by allowing

themselves to be guided by their instinctive reactions. Furthermore they were advised, were they to hesitate (certain words necessarily cause mixed feelings) to give "any score" and to move on to the next word. Finally, to neutralize order effects, twenty questionnaire versions corresponding to as many random permutations of the words have been used.

The scale retained consists of seven selected response items ranging from -3 to +3, the sign (-), which evokes frustration, being quite naturally regarded as an unpleasant sensation and (+), which connotes the contrary, rewarding experiences, a pleasant sensation[9]. But should a central modality (zero) which in fact corresponds to a non-answer be proposed? It was tempting not to include this modality in the scale so as to compel answers. However, it appeared that the involving nature of the questionnaire made it a duty to offer the interviewees the possibility not to give their evaluation of certain words.

This "word questionnaire" is subsequently complemented with other questionnaires comprising socio-demographic information and answers to questions of attitude or behaviour. These sources of information are correlated in order to dispose simultaneously for each individual of two types of information: first basic characteristics and variables of interest and also the data provided by the semiometric questionnaire.

1.2 A representative survey

The sample survey remains one of the fundamental tools for obtaining knowledge about economic and social reality. National or international statistical institutes use them extensively. Political polls and opinion polls, somewhat used badly or incorrectly over the last three decades of the twentieth century, may have tarnished the image of the tool in the eyes of the general public, but they have in no way impaired either a survey's usefulness or the demand for it or the interests professionals have for it.

The theory of probability indicates how and with which margin of error the observation of a small sample of a population can inform us about the entire population ... but, above all, this rigorous mathematical foundation should

[9] In terms of the statistical processing that will follow, this scale is strictly equivalent to a scale of 1 to 7, selected for coding responses to eliminate the minus sign (-) which weighs down data files, and we will use normally score 1 (rather than score -3), score 2 (note rather than -2) up to score 7 (corresponding to score 3). The initial median score 0, becomes score 4.

not mislead us in any way. The survey is an art in the same way as arts and crafts or works of art are: a lot of knowledge, know-how, expertise, and ethics are necessary to give this tool its scientific status. The physicist knows he can not measure without making any measurement error and, furthermore, he/she does not know the error he/she commits (otherwise he/she would know the true value of the measured thing); but, he/she strives to determine the magnitude of this error. And at the cost of this approach, in fact relatively modest, very gradually he/she builds up scientific knowledge.

It is also the approach of the statistician who carries out a survey... but, in this case, the complexity of the tool requires a real "investigation into the investigation" in order to carry out an assessment.

The design of the questionnaire, the sampling design and sample selection, the phase of data collection (face to face, by telephone, mail, the Internet), the codification of information, consistency checks thereof, and any procedures or corrective weighting are all steps that require real-time control. Such control is preferably achieved by an independent body, independent of both the institute that conducted the survey and the survey sponsors[10].

A powerful, useful, complex, fragile tool, this is how a representative survey can be described and qualified. This is why its results cannot be delivered without a minimum amount of methodological scrutiny and comment. The reader should not be surprised if the results presented to him/her only appear as the tip of a large methodological iceberg.

> To know what the words evoke for respondents (be they English, Italian, German, Spanish, French or American, and from many other countries), numerous surveys have been carried out. These surveys were conducted at different times for the same country, or sometimes simultaneously in different countries.

These repeated surveys in space and time are certainly expensive, but they give valuable indications on the stability of the structures and associations observed. They will be completed in the next chapter, by more technical stability studies.

---

[10] See for example the role of JICs (Joint Industry Committees) with respect to investigations concerning audience measurement media. In France the CESP (*Centre d'Etude des Supports de Publicité*: Centre for the Study of Media Advertising) has been assuming this role since 1957.

Since structure has been brought up (and this concept will in fact be a *leitmotif* throughout this book), let us specify here that representative surveys are more robust tools for highlighting structures (or correlations) than they are for measuring levels (average values, percentages). A measure of correlation between two variables, for example, will be much less sensitive to the perfect representation of a sample than the measure of the average of these variables in the sample would be. This property is difficult to prove in general, however it can give rise to experimental probes tailored to each specific situation, based on simulations of sample disturbance.

Table 1.2: Surveys from 1990 to 2002

Country	Sample	Year
France	2,764	1990
Great Britain	1,849	1995
France	2,764	1996
Spain	2,983	1996
France	2,764	1998
Italy	2,606	1998
Germany	3,065	1999
Greece	1,062	2000
France	2,763	1999
Canada	1,865	2000
France	2,764	2001
France	2,763	2002

Sources: Taylor-Nelson-Sofres group

The observation instruments required must then be able to structure the corpus of words. To do this, they will rely on exploratory techniques that provide synoptic representations of huge data sets without referring to strong assumptions of any statistical nature or a particular model. They can exhibit then, as a series of planar maps, associations between words and between individuals or groups of individuals.

In this chapter, we shall refer primarily to the procedures of multivariate statistical analysis by considering principal components analysis (see

Annexes A1.2 and A1.3). Other techniques (classification and Kohonen maps) will occur in subsequent chapters.

The results, which will be presented, are based, as far as France is concerned, on six independent samples for years 1990, 1996, 1998, 1999, 2001 and 2002 involving 16,582 individuals surveyed, representative of the French population aged 18 years and over, and who have responded to the 210 word questionnaire[11]. Other results concerning notably Germany, Spain, and the United Kingdom will be presented in Chapters 2 and 5.

1.3 Principal Components Analysis

This is a technical description of multidimensional arrays of measurements closely related to the technique known as factor analysis, which has been used by some psychometricians since the early twentieth century. It will be a privileged instrument of description of the corpus of scores to the 210 words of the questionnaire by the 16,582 persons interviewed between 1990 and 2002.

The principle of this type of analysis is both simple and intuitive[12]. We have to represent all survey respondents, on the one hand, and all the words in the questionnaire, on the other, as two "clouds of points", the former being defined with respect to the latter.

The proximity between words is defined by the scores that respondents have awarded (two words scored similarly will be close, and all the closer as the similarity is high); while respondents' proximity will be defined by all 210 of the scores they have ascribed (two respondents will be close if they have ascribed a score to all the words in a similar way: thus we can speak of people with similar response profiles).

[11] These data are then translated into the form of a rectangular array composed of 16,582 lines (respondents) and 210 columns (words).

[12] Although based on a theory of algebraic geometry which is more elaborately outlined in Annex A1.1.

These results are presented graphically in a reduced space the dimensions of which are called *principal* components[13].

Rules for reading graphs

The coordinate of a variable (word) on an axis is the correlation coefficient of the variable with the new artificial variable (sometimes called a factor: it is the set of coordinates of respondents on the axis) which the axis represents. The axes are qualified by all the words that are well correlated with them. Respondents occupying extreme locations on the positive side of an axis will have given a high score to the words that are highly associated with that axis (on the average). On the contrary, such respondents will have under-rated the anti-correlated words (negatively correlated) located in the other side of the axis. On examining a factorial plane (a plane generated by two axes) we can visualize the correlations between variables and identify groups of individuals giving the same scores to the same words highlighted by the plane.

On the planar representation, the proximity of two words is interpreted in terms of correlations of scores: two words are close if they are rated similarly by the respondents. On the contrary, they will be far apart if opposite marks are assigned to them. Two respondents will be close if they have given similar scores to each of the words.

In fact, respondents (anonymous) only interest us here in terms of their identification (variables called socio-demographic characteristics such as age, sex, occupation, etc..) or in terms of their responses to a question of opinion, attitude or behaviour, asked in the same survey. We thus interpret the cloud of respondents from these variables or any other issue raised by the survey. This phase aims at positioning these variables as *supplementary* elements, variables not belonging to the purely semiometric part of the questionnaire: these variables are not involved in the composition and definition of the axes, they just illustrate it *a posteriori*.

These allow us to obtain the best approximate display of distances between respondents on the one hand, and between the words, on the other[14]. As the

[13] These are new variables (which can be described as latent variables or hidden variables, or artificial variables) which are actually linear combinations of initial variables, i.e. linear combinations of scores given to words.

[14] These new synthesis variables are quantitative (or numerical), independent of each other (more precisely, uncorrelated), and of decreasing importance (the importance being measured by the variance or dispersion). The first components of higher variances reflect the strongest structures. They are then ranked from the most dominant to the most marginal.

respondents are anonymous, they will be described by their basic characteristics (supplementary variables). Thus, a display of proximity between words and socio-professional categories, sex or age, for example, can shed light on the interpreting of the average scores given by each of these categories.

The box above provides the first rules of interpretation of principal components analysis: for each axis the principal words that are correlated positively and negatively with it will be given, and the populations the axis differentiates[15].

1.4 The stable axes: "semiometric axes"

Only the first six main axes (or "factors") that seem interpretable will be presented. Chapter 2 will show that these six axes have satisfactory stability properties (stability as measured by experimental criteria - stability over time and space - and also stability according to more technical statistical criteria). The following axes are less stable in samples of normal size (from 1000 to 2000 people), but this does not mean they only represent noise[16]. In fact, we limit ourselves to the most invariant and robust part of the structure observed.

A tentative interpretation will be proposed[17], and we shall then endeavour to qualify each axis. At each pole of an axis, a concept will emerge. Each axis is then qualified by two opposing concepts.

> We shall return, in Chapter 6, in greater detail to the problems raised by interpretation and we shall then propose a personal reading grid to describe the meaning of the axes, and, beyond that, a hypothesis on the nature of the "object" which semiometry observes.

---

[15] These populations are identified by the technique known as "projection of supplementary categorical variables"; the socio-demographic characteristics of respondents on the axis appear as the mean point of the individuals concerned (see Appendix A1.9.3).

[16] The stability of the more remote axes depends on the size of the sample. Thus only the first four factors are explicit in a sample of 500 subjects; the first six axes are from a sample of 1,000 respondents, but the other two axes (7 and 8) only started to become clear with 5,000.

[17] If the calculations appear to be "objective", the interpretation is of the order of freer comments, suggesting several possible names for the axes. But bear in mind that these are only conventional labels that summarize (and therefore distort) the psychological and semantic reality of axes, which depends on all the words that are correlated.

1.4.1 Axis 1: A methodological axis of participation

The first axis, which will be named "*the axis of participation*" has a different status from others. It is an axis related to the methodology of the survey, which reflects the attitude of the respondents to the questionnaire, whereas the other axes concern the very content of the questionnaire, i.e. the semiometric structure itself[18].

On one side of the axis, we find the following words[19]:

(+) *Courage, Efficient, Dynamic, Reward, To protect, Honest, Robust, Confidence, Elegance, Respect, Purity, Accuracy, Softness, To admire, Politeness, To built, Charitable, Comfort, Liveliness, Honour, To produce, Reason, To recover, Precious , Maternal.*

On the other side:

(-) *To betray, Disorder, War, Anguish, Fault, Danger, To break, Emptiness, Death, Rebellion, Storm, Black, Doubt, To attack, Irony, Maze, To criticize, Rifle, Cry, To age, Detachment, Wall, Fire, Wild , Desert.*

The categories of individuals particularly affected by these poles are respectively:

(+) Women, seniors, individuals with a low level of education, retirees, people living in small towns or rural areas.

(-) Young, men, in jobs, individuals with higher education, people living in the largest cities.

This axis contrasts words that extol the values that are consensual enough and express a certain social conformity (*Courage*[20]*, Efficient, Dynamic, Precision, Reward* ...) with words evocative of values which rather tend to transgress what is socially valued (*To betray, War, Disorder*, *Anguish* ...)

---

[18] The first axis accounts for 12% of the total variance. Recall that, unlike a technical prejudice fairly common, this percentage can in no way be interpreted as a percentage of information. The total variance is weighted by considerable statistical noise, and only serves as a measurement reference for comparing the analysis (see Annex A1.9 on tools for validation).

[19] The orientation of the axes is arbitrary. The fact that the sign (+) corresponds to values commonly considered "positive" is purely fortuitous. The categories are ordered from the most significant (statistically) to the least significant.

[20] By convention, in the main text, the words of the semiometric questionnaire are written in italics beginning with a capital letter.

The concept "conformism" corresponds to a population composed of rather old people, females and people with a low level of education, opposed to a population represented by males, young people, professionals and graduates of higher education.

However, the fact that, generally speaking, the first axis of principal components analysis performed on standardized variables often reflects a "size effect", should prompt us to greater caution in that ever perilous exercise of interpretation[21]. Indeed, the words: *Effective*, *Dynamic*, etc. are among the words that receive the highest scores while: *To Betray*, *War*, etc. have scores among the lowest. We will return in greater detailed and more technically to the interpretation of this axis in Chapter 5, entirely dedicated to methodological problems related to scoring.

1.4.2 Axis 2: "Duty / Pleasure"

On one side of the axis, are the following words:

(-) *Discipline, To obey, Homeland, Morals, Soldier, To economize, Industry, Priest, Rule, Law, Firmness, Faith, Work, Sacrifice, Honour, God, Elite, To forbid, Rigid, To punish, Border, Hunt, Tradition, Meditation, Metallic*.
On the other side:

(+) *Sensual, To dream, Adventurer, Original, Island, Nudity, Wild, Voluptuous, Lightness, Ocean, Bohemian, Desire, Carnal, Emotion, To seduce, Escape, Moon, Humour, Mellow, Storm, Revolt, Magic, Mystery, Music, Skin*.

The categories of individuals particularly concerned by these two poles of the axis are respectively:

(-) Individuals over 55 years of age, retired people, farmers.

(+) Individuals under 30 years of age, graduates between 30 and 55, executives or employees.

---

[21] The size effect is usually due to the fact that in many collections of measures, the correlations are, overwhelmingly, positive: in the case of measurements on people or animals, the size is the main source of dispersion: briefly, all measurements are simultaneously larger in big individuals and smaller in the smaller ones. In most applications, the axis following the factor of size, the *form factor* is more structural: it provides information on the network, more subtle and hidden, of correlations that remain after eliminating the size effect.

The following words: *Discipline, Morals, Homeland, To obey, Soldier, Priest* ... evoke the values of order and tradition that relate, in perhaps a ludicrous way, to the "marriage of the aspersorium and the sword" (i.e. the church and the army) referred to at certain periods of the history of France, while the words: *Sensual, To dream, Lightness, Wild, Adventurer, Voluptuous* ..., seem to correspond to another element, that of motion and pleasure, which reached its heyday during the events of May 1968. By convention, we shall call this axis: "Duty / Pleasure"[22].

This axis is strongly structured by age, and "Duty" is a value that characterizes quite clearly the elderly; and "Pleasure" is a more salient value among younger people. But this axis continues to exist even if we analyze separately the subpopulation of individuals aged between 35 and 55. This is certainly an axis related to age, but, as we shall see, it is far from being characterized by age alone. The interpretation of this axis is richer and more complex[23].

1.4.3 Axis 3: "Attachment / Detachment"

On one side of the axis, are the following words:

(+) *Jewel, Present, Comfort, Reward, Gold, House, Money, Wedding, Wealth, Liveliness, Family, Perfume, Loyalty, To inherit, Flower, Free, Fashion, Softness, Tenderness, Elegance, To recover, Peace, Maternal, Caress , To laugh*.

On the other side:

(-) *Danger, Death, To break, Storm, Anguish, Empty, To punish, Maze, Cry, To criticize, Doubt, Rebellion, To attack, Desert, Detachment, Fault, Fire, Irony, To age, Wall, Effort, Mistrust, Question, To interrogate, Stranger*.

The categories of individuals especially concerned by the poles of this axis are respectively:

(+) Women, individuals without their high school certificate, workers or employees.

(-) Males, graduates, executives, craftsmen-traders or intermediate occupations.

---

[22] This axis represents 6% of the total variance. This percentage is highly significant statistically, as are the percentages relative to the axes whose description follows.

[23] A more complete interpretation is proposed in Chapter 6.

Words like: *Jewel, Present, Comfort, Wedding, Family, House, Liveliness* ... seem to evoke a happy life together. These words are over-scored by women compared to men, while on the other hand, the words: *Danger, Death, To Punish, Emptiness, To break, To attack* ... which are rather over-scored by men, appear as words of rupture and dereliction. Conventionally, we shall call this axis: "Attachment / Detachment".

This axis[24] is marked by the opposition: Male / Female. But this is by no means an axis that is limited to this opposition. In fact, this axis is found again if we conduct separate analyses of male respondents and female respondents[25].

1.4.4 The semiometric plane of axes 2 and 3.

This *semiometric plane* (Figure 1.2) will be called the *principal semiometric plane* (it is, in fact, generated by the two most important axes, if one omits the first axis, which is, as we have seen, what we call a "methodological axis of participation"). This plane is widely used in the iconography and communication of semiometric analyses[26]. It describes the axes simultaneously: "Duty / Pleasure" (horizontal axis) and "Attachment / Detachment" (vertical axis). This representation of the plane can somehow supplement the interpretation of the previous axis by extending it with a further dimension. We can thus talk about the four quadrants of this plane, "Pleasure / Attachment" and "Pleasure / Detachment" and "Duty / Attachment" and "Duty / Detachment". The interpretation is richer and more nuanced. It is explained in greater detail in the fifth section of this chapter devoted to the analysis of planes spanned by some pairs of axes.

---

[24] This axis represents 4.2% of the total variance

[25] This point is developed in section 2.1.3 of the second chapter on the stability of the observed structures.

[26] Chapter 7, which presents a number of semiometric applications will often use a representation of words in this plane generated by axes 2 and 3.

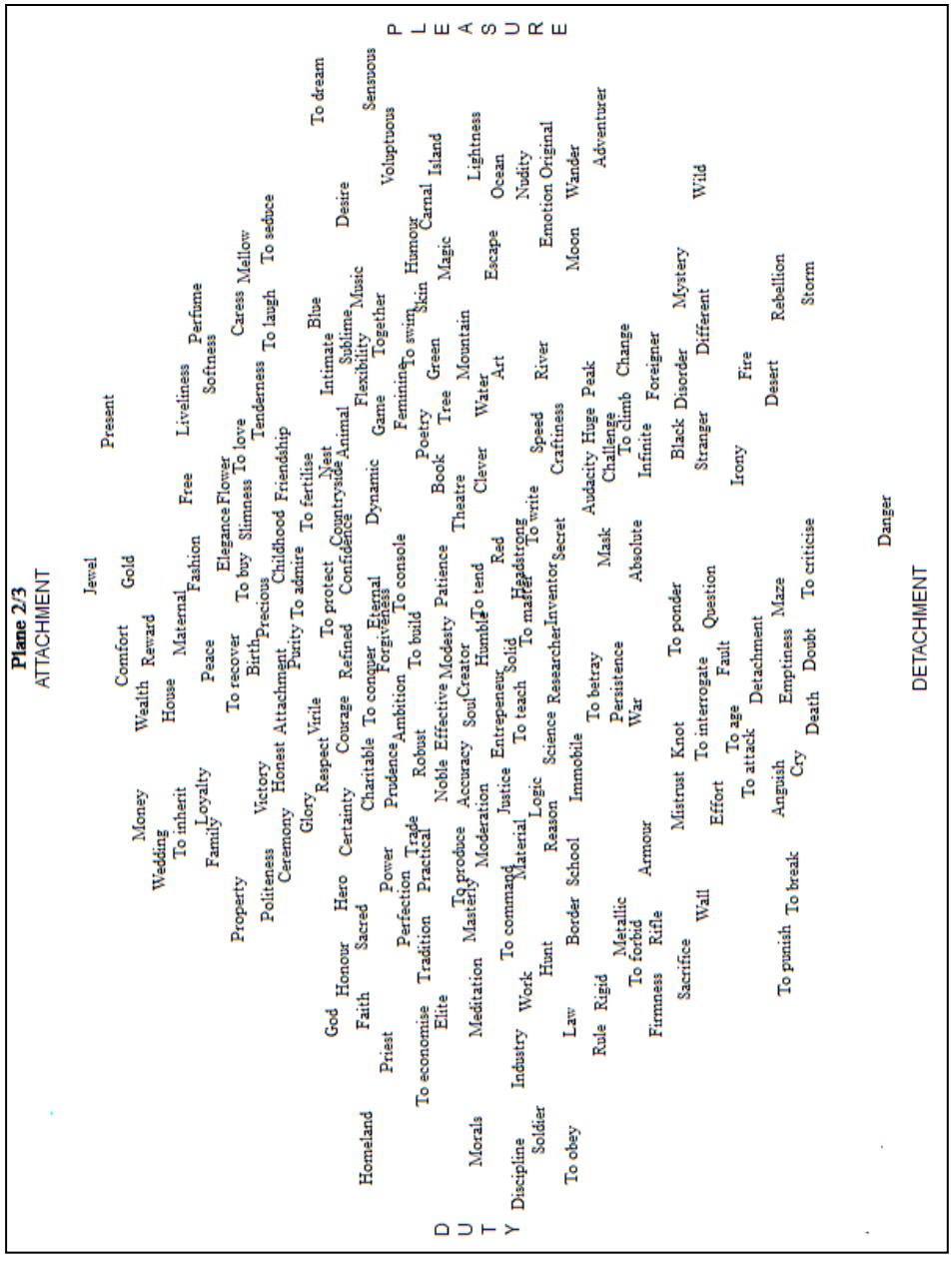


Figure 1.2: Semiometric Map (2.3) - The dimensions of “Pleasure” and “Duty” with "Attachment - Detachment". Two facets of “Pleasure”, “Pleasure-Attachment” and “Pleasure-Detachment”. Two facets of “Duty” (*comments in section 1.5*)

1.4.5 Axis 4: "Sublimation / Materialism"

On one side of the axis, we find the following words:

(+) *Book, Art, Poetry, Theatre, Tree, School, To teach, Forgiveness, To write, Meditation, Charitable, Stranger, To think, Nest, Peace, Flower, Soul, Moderation, Patience, Researcher, Modesty, To console, River*.

And on the other side:

(-) *Wealth, Power, Gold, Money, Glory, To conquer, Speed, Ambition, To command, Craftiness, Rifle, To inherit, To seduce, Property, To attack, Victory, Jewel, Fashion, Desire, Precious, Sensual, Armour, Rigid, Elite , Free*.

The categories of individuals especially concerned are respectively:

(+) Individuals over 55 years of age, graduates, women, executives.

(-) Individuals under 30 years of age, without a high school certificate, workers, employees, artisans, shopkeepers, farmers, men.

Words like: *Wealth, Gold, Silver, Glory, To conquer, Speed, To command* ... evoke attachment to worldly things, faith in materialistic values, in short, the worship of Mammon, while the opposite group of words at the other end of the axis: *Book, Art, Theatre, Poetry, Soul, School, Stranger, To think*, ... bring to our minds a different system of values, more spiritual, more sophisticated, more sublimated, those belonging to a secularized kingdom of heaven. Conventionally, we shall call this axis: "Sublimation/Materialism"[27].

This axis compares and contrasts two very different profiles of respondents: young men without the high school certificate and older women with degrees. Like the previous axes, it is not limited to an opposition between these categories, because we can find a similar axis if we conduct an analysis inside these very categories.

[27] We call this axis " Sublimation/ Materialism" for reasons which are explained in Chapter 6 (*Towards a few interpretations*).

1.4.6 Axis 5: "Idealization / Pragmatism"

On one side of the axis, are the following words:

(+) *God, Faith, Soul, Priest, Sacred, Meditation, Eternal, Ceremony, Masterly, Infinite, Noble, Jewel, Poetry, Absolute, Wedding, To love, Hero, Moon, Magic, Lightness, Fashion, Mask, Glory, Purity, Mystery*.

On the other side:

(-) *Effective, Accuracy, Logic, Solid, To master, Clever, Practical, Robust, To produce, To ponder, Headstrong, To build, Dynamic, Material, Effort, Audacity, Honest, Entrepreneur, Inventor, Confidence, Firmness, Reason, Nudity, Science, Industry*.

The categories of individuals especially concerned by the two poles of this axis are respectively:

(+) Women, young people.

(-) Males, artisans, shopkeepers, intermediate professions, managers, workers, people between 40 and 50 years of age, people living in the largest cities.

The words: *God, Faith, Soul, Sacred, Eternal, Infinite*, etc. express a desire for transcendence which reason can not account for, a need to look beyond too rigorous a logic, more dreamlike or irrational forms of expression; while the words: *Effective, Precision, Dynamic, To build, Practical, Headstrong*… connote a need for rationality, pragmatism and the desire to live in an entirely logical world, dedicated completely to action, which is reminiscent of the Enlightenment ideal. Conventionally, this axis will be called "Idealization / Pragmatism"[28].

This fifth axis[29] compares men and people still working with women and young people.

---

[28] We call this axis "Idealization / Pragmatism" for reasons which are explained in Chapter 6 (Towards a few interpretations).

[29] This axis corresponds to 2% of the total variance, a figure that remains however highly statistically significant especially given the number of variables and the number of respondents.

1.4.7 Axis 6: Humility / Sovereignty

On one side of the axis, we find the following words:

(+) *Birth, Mistrust, Fault, Doubt, Family, Rigid, Tenderness, Wedding, To forbid, Caress, Mask, To obey, Childhood, Prudence, Maternal, Cry, Anguish, Immobile, Knot, Loyalty, To punish, Emptiness, Politeness, Ceremony*.

On the other side:

(-) *Creator, Noble, Art, God, Audacity, Elite, Faith, To master, Robust, Entrepeneur, Efficient, Concrete, Sacred, Voluptuous, Inventor, Soul, To command, Meditation, Clever, Absolute, Priest, Sovereign, Sublime, Theatre*.

The categories of individuals particularly concerned by these two poles are respectively:

(+) Workers, employees, persons under 40 years of age, people without the high school certificate, women, people living in rural areas.

(-) Executives, seniors, retirees, people with a high education level, people living in major metropolitan areas.

The interpretation of this axis[30] is less straightforward. Indeed, if we find at one extreme end words like: *Noble, Creator, Art, God, Faith, Elite, Audacity* ... that evoke a kind of elitism, an aspiration towards a certain kind of sovereignty or "distinction" in the terminology of Pierre Bourdieu; the words on the other one are divided into two semantic sub-sets: *Birth, Family, Tenderness, Wedding, Childhood , Maternal*, on one hand, and *Mistrust, Fault, Doubt, Rigid, To obey, To forbid*, on the other hand. A synthesis seems at first glance to be difficult to achieve. However, the fact that this axis contrasts individuals, rather old ones with a high level of culture, with individuals whose social status is more modest, leads us to think that this axis reflects the confrontation between two systems of values that really and truly exist. Conventionally, without going any further into interpreting it, this axis will be called: "Humility / Sovereignty".

[30] The sixth axis corresponds to 1.8% of the total variance.

1.4.8 The following axes

The following axes are more difficult to interpret. When statistical analyses concern samples composed of less than a thousand individuals, axes 7 and 8 do not systematically appear. They only acquire a good level of stability starting with five thousand individuals[31]. We attempt here only to describe these axes without venturing to interpret them and without trying to give them a name[32].

- **Axis 7**

On one side of the axis, we find the following words:

(-) *Audacity, Faith, Secret, Sublime, Ambition, To master, Purity, Charnel, Desire, To seduce, To console, Respect, Challenge, Craftiness, Persistence, Sacred, To love, Eternal, Sensual, Headstrong, Forgiveness, Soul, Intimate*.

(+) *River, Tree, Mountain, Theatre, Book, Ocean, Countryside, Flower, To swim, Water, Metallic, House, Animal, Peak, Art, Science, Industry, Island, Moon, To inherit, Trade, Red Jewel, Poetry*.

The categories of individuals especially concerned are respectively:

(-) Individuals under 30 years of age and especially the age group 15-19 years of age belonging to a family of executives, high educational level.

(+) Individuals over 55 years of age, retired, and spouses of retirees, low educational attainment.

If we make a distinction between the two opposing groups on axis 7: on one hand, we have the elderly having no occupation any longer and no

---

[31] This analysis, performed here on a sample of over 11,000 individuals totalling all four of the first French surveys, were also conducted on a sample of 1,000 and 5,000 individuals. The sample size improves the quality of the axes of high rank. Axes 7 and 8 represent respectively 1.3% and 1.2% of the total variance (percentages highly statistically significant for 210 variables and 5 000 individuals).

[32] We must distinguish between: a) Highlighting the stability of an axis that is a "statistical fact" established in an objective, repetitive, automatic way from the collection of data, b) the description of this axis, which contains an element of subjectivity as does any natural language description involving the choice of words and concepts; c) naming an axis, which implies a simplification and a drastic choice, and therefore a more important element of interpretation; d) the proper interpretation of the axis, which is not always possible, which uses categories, theories or models, explicit or implicit. In this chapter, the first six axes have been described and named. An interpretation will be proposed later in Chapter 6. The following axes are simply described, with interpretative comments that have no other ambition than to make the description a little less terse.

education, and on the other: young people from intellectually advanced backgrounds, students either in high school or in higher education, and therefore not yet having joined the workforce, it seems more difficult to identify concepts, and underlying them, the values that they are opposed to.

This is more of a divide between different mentalities: the younger and more educated are fond of abstract concepts: *Audacity, Faith, Secret, Sublime, Ambition, To master, Purity*, etc. They contrast with the older, uneducated people, having already been faced with the realities of life, focusing on concrete objects (especially those referring to nature): *River, Tree, Mountain, Book, Ocean, Flower, House, Moon, Jewel*, etc.

- **Axis 8**

On one side of the axis we find the following words:

(+) *Theatre, Book, To write, Refined, Poetry, Art, Fashion, Perfume, Jewel, Elegance, To criticize, School, Noble, To teach, Slimness, To interrogate, Doubt, Lightness, Justice, Irony, Question, Precious, To break, To ponder*.

On the other side:

(-) *Countryside, Mountain, Summit, Rifle, Island, Water, Adventurer, To climb, Animal, Ocean, River, Faith, Armour, To swim, God, Tree, Family, Priest, To hunt, Sacred, Speed, Savage, Storm, Soldier*.

The categories of individuals especially concerned are respectively:

(+) Older women, with a high level of education, executives and their wives.

(-) Men under thirty years of age, workers and farmers, low education. It's a little like the dichotomy between Beauty and the Beast or between refined people and unrefined ones, with on one hand elegance and female refinement: *Theatre, Book, To write, Refined, Poetry, Art, Fashion, Perfume* ... and on the other hand,*: Countryside, Mountain, Peak, Rifle, Adventurer, To climb, Animal*[33].

As with the previous one, this axis appears only when working on samples containing more than five thousand individuals.

[33] Opposition, however, can lead, in certain circumstances, to a troubling attraction. We have in mind the couple formed by Lady Chatterley and her gamekeeper. This ambivalence leads us to think that this axis is not merely sociological (we mean here: between social classes) but, to some extent, psychological.

1.5 The semiometric planes

We have just presented the first six stable dimensions, and briefly described the two following ones. Considering the axes in pairs (We are dealing here with principal planes and not principal axes) sheds new light on semiometric structures. This brings added meaning and hones the description of the axes and interpretation thereof. The example of the plane of axes 2 and 3 has already been mentioned (Figure 1.2). We will comment in more detail on this plane [plane (2.3)], then consider some other planes.

1.5.1 The multiple facets of "Pleasure"

By considering simultaneously the second axis "Duty / Pleasure" and the third axis, "Attachment / Detachment", one obtains two opposing components of "Pleasure", which we shall call "Pleasure-Attachment" and "Pleasure-Detachment".

"Pleasure-Attachment" (top right corner of Figure 1.2[34]) is characterized by the words: *Softness, Caress, Perfume, Liveliness, Present, To laugh, Tenderness, Mellow, To seduce…*, whereas "Pleasure-Detachment" (bottom right of Figure 1.2) is epitomized², with the words: *Adventurer, Wild, Storm, Revolt, Mystery, Light, Bohemian, Original, Emotion…*

Comments / Suggestions for interpretation

If "Pleasure-Attachment" evokes the happiness of being alive through a fresh exhilarating sense of excitement, a world of happiness where every moment seems an eternity of sweet pleasure, "Pleasure-Detachment" has a more pungent taste, it sounds harsher, rougher, darker, more violent, more tragic. There are indeed two ways of enjoying life, but the latter is that of the warrior ("*guerrier*") and the former the... "*repos du guerrier*" (hero's welcome).

We can now see how subtle a contribution may be made to the study of semantic words by exploiting the results of a survey carried out among the general public!

[34] This figure is presented above (see sections 1.4.3 and 1.4.4).

This contribution is probably quite unexpected, even inappropriate, for linguists and semioticians. It is all the more remarkable as it is absolutely not an opinion poll on the subject matter here, 'pleasure' or 'semantic proximity' (the protocol would probably be very heavy) but a simple scoring of words one by one[35].

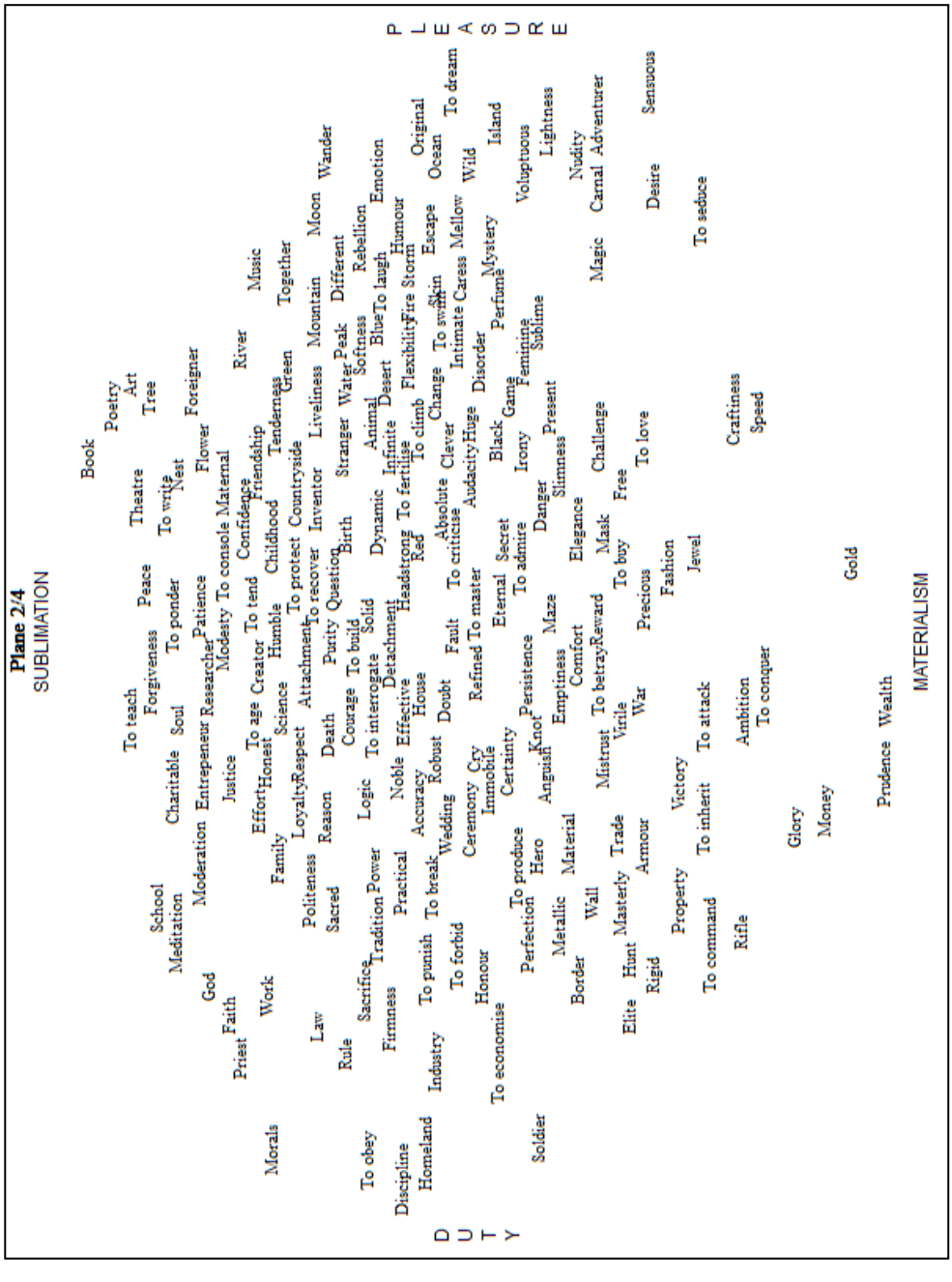
Let us continue our extension of "Pleasure" now by taking into account simultaneously axis 2 "Duty / Pleasure" and axis 4 "Sublimation / Materialism". Figure 1.3 shows the plane [2.4] thus obtained.

We can then distinguish, in Figure 1.3, "Pleasure-Sublimation" (above and right) which is characterized by the words: *Art, River, Stranger Music, Bohemian, Tree, Poetry…*, and "Pleasure / Materialism" (bottom right), with the words: *Sensual, Desire, To seduce, Carnal, Magic, Adventurer, Nudity, Voluptuous, Lightness…*

Comments / Suggestions for interpretation

"Pleasure-Sublimation" is subtle: the senses have been soothed over by the contemplation of the beauties of art and nature, while "Pleasure-Materialism" is looking for rougher, more immediate benefits, closer to that other form of nature characterized by our natural instincts. Note that the diagonal brings out a contrast between art and sexuality or, more precisely, between a domesticated, elaborate, and under control sort of sexuality and a spontaneous and natural one.

[35] The 11,055 respondents may not moreover suspect that by giving these marks that stable statistical facts will be automatically extracted from the correlation matrix marks (symmetric matrix of 210 rows and 210 columns containing 21,945 correlation coefficients).


Plane 2/4
SUBLIMATION
MATERIALISM
DUTY
PLEASURE
Book
Poetry
Art
Tree
Theatre
To write
Nest
Foreigner
Flower
Maternal
Confidence
Friendship
Tenderness
Green
River
Music
Together
Countryside
Inventor
Liveliness
Mountain
Moon
Wander
Different
Rebellion
Emotion
Humour
Original
Ocean
To dream
Wild
Island
Lightness
Nudity
Adventurer
Carnal
Magic
Sensuous
Desire
To seduce
Craftiness
Speed
To love
Challenge
Present
Slimness
Elegance
Mask
Free
To buy
Precious
Fashion
Jewel
Gold
Wealth
Prudence
Glory
Money
To conquer
Ambition
To attack
War
Virile
Mistrust
To betray
Reward
Maze
Emptiness
Comfort
Anguish
Knot
Persistence
Danger
To admire
Irony
Game
Black
Disorder
Perfume
Mystery
Storm
Fire
Flexibility
Blue
To laugh
Softness
Peak
Water
Stranger
Animal
Desert
Infinite
To fertilise
To climb
Change
To swim
Skin
Intimate
Caress
Mellow
Escape
Feminine
Sublime
Voluptuous

By considering now the plane spanned by axis 2 "Duty / Pleasure" and axis 5 "Idealization / Pragmatism", we obtain "Pleasure / Idealization" (top right corner of Figure 1.4) characterized by the words: *Island , To dream, Lightness, Ocean, Bohemian, Moon, Magic, Mystery, Sublime, Perfume, Desert, Infinite…, and* "Pleasure / Pragmatism" (bottom right corner of Figure 1.4), characterized by the words: *Nudity, Audacity, Clever , Sensual, Revolt, Humour…*

Comments / Suggestions for interpretation

"Pleasure-Idealization" is the pleasure associated with dreaming and the contemplation of phenomena whose territory is situated in the marches of fleeting reality, in this no man's land where objects seem to have not yet taken on a clear and definitive shape. "Pleasure-Pragmatism", on the other hand, does not listen to stories but loves to scrutinize objects in a raw manner and without any illusion.

Finally, still considering axis 2 "Duty / Pleasure" together with axis 6 "Humility / Sovereignty", we obtain "Pleasure / Humility"(top right corner of Figure 1.5 representing the plane [2,6]) characterized by the words: *To laugh, Desire, Caress, Softness, Liveliness, Tenderness, Craftiness*, and "Pleasure / Sovereignty" (bottom right corner of Figure 1.5), characterized by the words: *Original, Voluptuous , Sublime, Fire, Peak, Art…*

Comments / Suggestions for interpretation

"Pleasure-Humility" is reminiscent of "Pleasure-Attachment", but on a more innocent, more childlike level, which seems like a reminiscence of the state of paradise before the 'Original Sin'. "Pleasure-Sovereignty" would rather indulge, in search of a sublime and heady feeling.

Although these versions of "Pleasure" fade away gradually as the order of the axes increases, making the interpretation and comments less and less confident, we may be surprised by the finesse and consistency of this "semantic decomposition"[36]

[36] Our comments and suggestions are illustrative, and in a way, optional. Their presence should not overburden the scientific status of Figures 1.2 to 1.5, which are derived from the data table automatically and are reproducible.

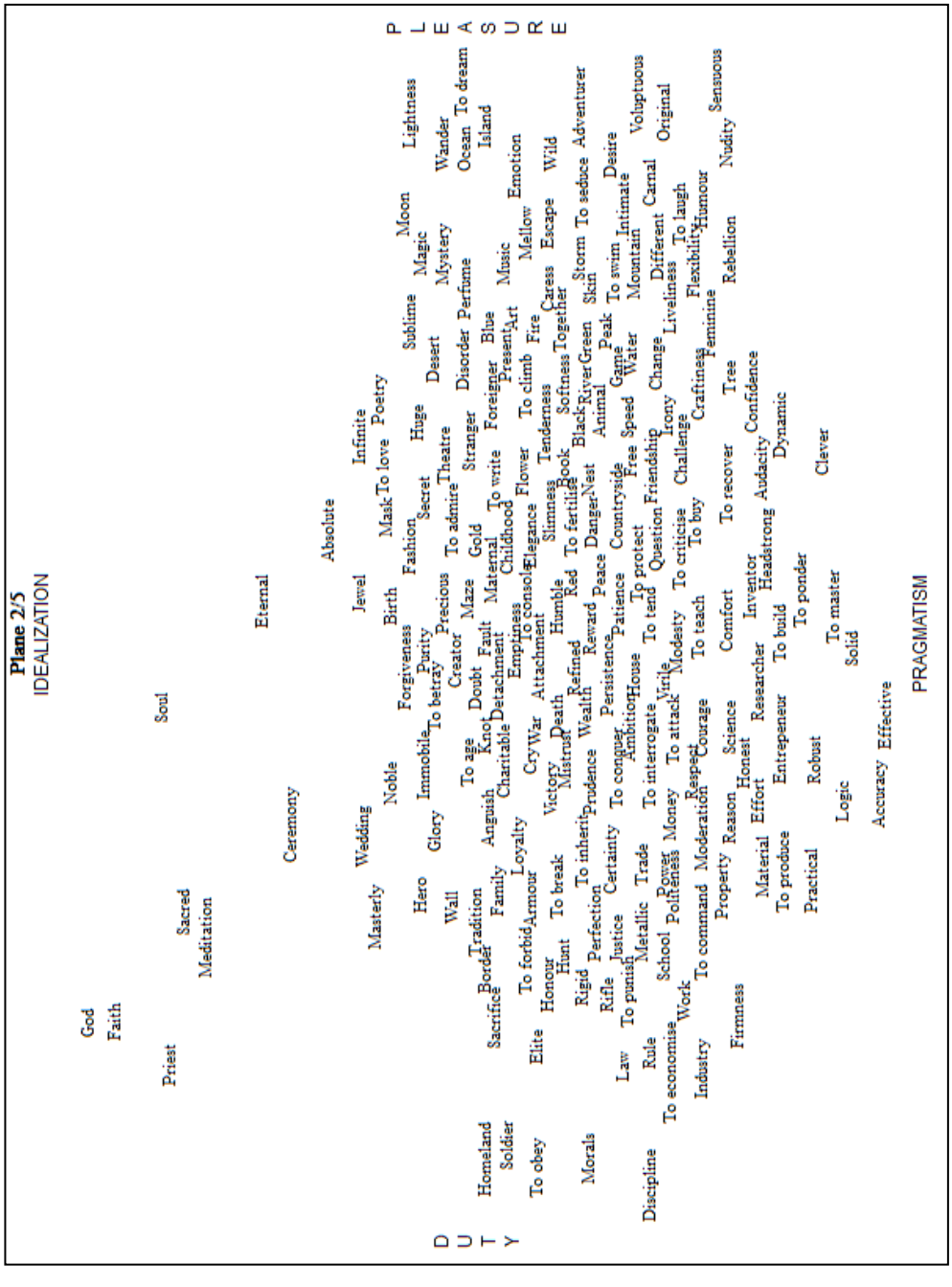


Figure 1.4: Semiometric Map (2.5) - Facets of Duty and Pleasure: Intersection with "Reason / Idealization"

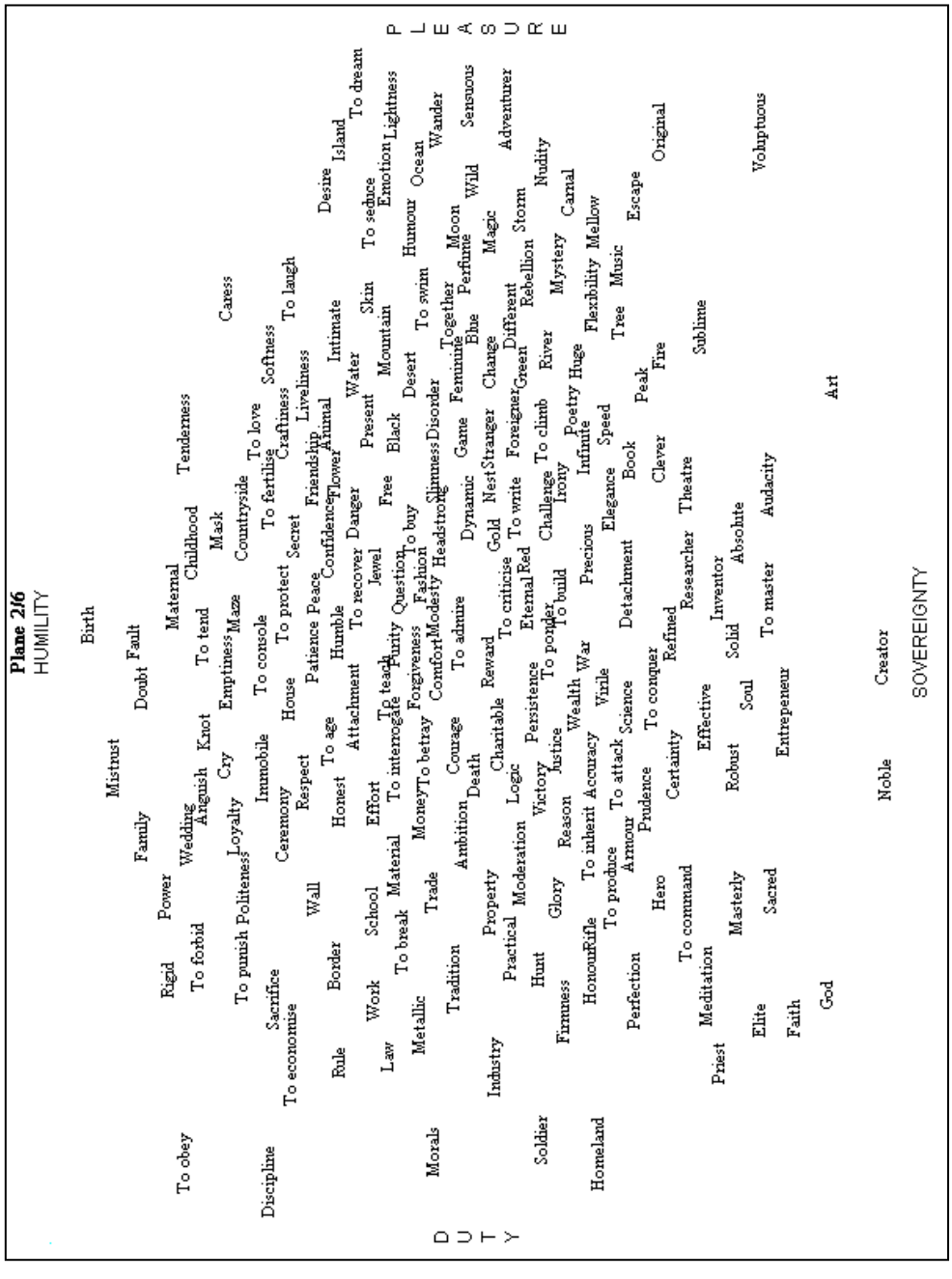


Figure 1.5: Semiometric Map (2.6) - Facets of Duty and Pleasure: Intersection with "Humility / Sovereignty"

1.5.2 The multiple facets of « Duty »

The same examination of factorial components (Figures 1.2 to 1.5), focusing this time on the left side of the horizontal axis (axis 2 "Duty / Pleasure") provides in the same manner the different facets of "Duty". This time we invite the reader to write his own comments after consulting the above figures.

By considering simultaneously axis 2, "Duty / Pleasure" and axis 3, "Attachment / Detachment", we obtain "Duty/ Attachment" (top left of Figure 1.2), characterized by the words: *Loyalty, Ceremony, Property, Politeness, Wedding, Family…*, and "Duty / Detachment" (bottom left of Figure 1.2), with the words: *Sacrifice, To forbid, Firmness, Law, Rule, To obey, To punish, Rigid, Frontier, Metallic , Wall, Death…*

By considering simultaneously axis 2 and axis 4 "Sublimation / Materialism", we get "Duty / Sublimation" (top left of Figure 1.3), characterized by the words: *Modesty, Meditation, Charitable, Justice, Moderation, Respect, God, Soul…*, and "Duty / Materialism" (bottom left of Figure 1.3), with the words: *Property, Glory, Money, Power, Elite, Sovereign, Trade...*

By considering simultaneously axis 2 and axis 5 "Idealization / Pragmatism", we get "Duty / Idealization" (top left of Figure 1.4), characterized by the words: *Eternal, Immobile, Soul, Sacred, Ceremony, Sovereign, Tradition…*, and "Duty/ Reason" (bottom left of Figure 1.4), characterized by the words: *Property, Prudence, Material, Reason, Honest, To produce, Utility...*

By crossing axis 2 with axis 6, "Humility / Sovereignty", we obtain "Duty / Humility" (top left of Figure 1.5), characterized by the words: *Distrust, Rigid, Loyalty, Prudence, Sacrifice, Wedding, Honest, To obey…*, and "Duty / Sovereignty" (bottom left of Figure 1.5), with the words: *Sacred, God, Sovereign, Elite, To command, Hero, Perfection, Glory…*

The mere mention of aspects or dimensions of "Duty" confirms the observation made about the dimensions of "Pleasure": bearing out the consistency of our findings and the finesse of the results. We may add that this procedure of systematic extension of a concept according to different axes can reveal the profound multidimensional topography of words, irreducible to one axis or plane.

1.5.3 The semantic dimensions of words themselves

After extending the concepts that the axes account for: "Duty", "Pleasure", "Attachment", "Detachment", "Sublimation", "Materialism", "Idealization", "Pragmatism", "Humility" and "Sovereignty", we shall focus on the words taken from the list.

The same word can, of course, correlate with several axes, each one of these "epiphanies" revealing its semantic dimensions.

The word *Wedding*

Thus the word *Wedding* in "Duty"[37], connoted by: *Loyalty, Honest, Respect, Courtesy*... is not the same as that of "Attachment", connoted by: *Comfort, Maternal, Birth, Trust, Friendship, Tenderness and Softness…*; the former referring to marriage viewed as an institution, the latter as part of a love relationship between two people (see Figure 1.2).

The word *Revolt*

Similarly, the *Revolt* in "Pleasure", connoted by: *Adventurer, Original, Bohemian, Escape…*, is not the same as that of "Detachment" connoted by: *To break, To criticize, Doubt, Danger, Death*... In one case, it evokes the manifestation of a desire to escape from the limits imposed by social constraints to access a larger life; in the other, a propensity to break these constraints even if it means destroying oneself (Figure 1.2).

The word *Border*

Similarly, the *Border* in "Detachment", connoted by the words: *Rigid, Firmness, To break, Wall, Mask…* is not the same as the *Border* in "Duty", connoted by: *Marriage, Property, Rule, Morals*... The first is the border that isolates the subject by rejecting others; the second is the border which distinguishes, organizes and guarantees order, both socially and mentally (Figure 1.2).

The word *Theatre*

Similarly, the word *Theatre* pertaining to "Idealization" (axis 5), connoted by: *Ceremony, Jewel, Mask, Magic…*, is not the same as the

[37] We denote by the expression *Wedding* of "Duty" the component (or coordinate) of the word Wedding on the axis of "Duty". Similarly, *Revolt* of "Pleasure" means: the component of *Revolt* on the axis of "Pleasure".

Theatre pertaining to "Sublimation" (axis 4) connoted by: *Art, Books, Music, Poetry, To write*... One is the spectacle which we witness; the other the work that is created.

The word *God*

As for the word *God*, it correlates with four semiometric axes, thereby uncovering for us some of its semantic components: the *God* of "Duty" (on axis 2) that the words: *Sacrifice, Rule, Morals, Discipline*… connote; and the *God* of "Sublimation" (on axis 4) that the words: *Forgiveness, To console, Charitable* connote; the *God* of the "Idealization" (on axis 5) connoted by: *Ceremony, Theatre, Magic, Mystery*; and the *God* in "Sovereignty" (on axis 6) connoted by: *Absolute* and *Creator*. The first refers to the *God* of the Law of the Old Testament, the second, the God of love in the New Testament, the third, the religious spectacle, and finally the fourth, the ineffable *God* of transcendence.

Conclusion

Let us make a first assessment of the results achieved, and also an inventory of issues that arise at this point, an inventory that will announce in a motivated way this time, the contents of the chapters to come.

Our approach is highly empirical and exploratory: a few simple ideas, an expensive campaign of measurements, then statistical processing, both classical and versatile enough to extract most of the structural features that may crop up.

We must of course ensure the reality of what has been observed, and then understand the nature and scope of new information highlighted.

There has been much discussion about words, and little about individuals or respondents. We did indeed mention which group of individuals characterized some axes (the elderly more on the "Duty" side; the youth more on the "Pleasure" side, for example). But people interviewed are anonymous, and could only intervene at this stage of our work through the categories to which they belonged.

Now it is one of the properties of the statistical method used to position the individuals on the same axes ... the mean point (or: centre of gravity) of a category of individuals is the one that we used to position the group in question. We will have, in general, far more information about these individuals outside of the **characteristic** variables: for example, the make

of their vehicles, the titles of the magazines they usually read, commuting time and means of transportation to and from work, and as much information that can be associated to the semiometric coordinates of individuals on the principal axes[38].

The analytical results have revealed structural features in the form of the privileged associations of words. The existence of these structural features, i.e., their stability, invariance, their possible generalization (or statistical inference) to all the populations concerned will be studied in detail from three angles: per duration, per country, and from statistical validation procedures called *re-sampling*. The work of validating the observed structures will be the subject of Chapter 2.

The reader intrigued throughout this chapter by the special treatment given to the first axis will find the required explanation in Chapter 3, which is a sort of "technical zoom" on the specifics of this axis, reinforced by international data and the impact that concerns in fact the methodology for all surveys.

With the results obtained by submitting a list of words to thousands of people having being tested, strengthened and better understood, it remains to study their dependence with respect to this list. The operational character of the list is clear, but what about its possible universality? Chapter 4, based on a new collection of empirical material, broaches this ambitious question.

The richness and subtlety of the observed structures in semantic terms throughout the previous sections raise the question of relations between semantics and semiometry. The process is difficult because semantics is not a discipline flagged out, rather a research topic torn apart by its own inter-disciplinary character. Chapter 5 will bring in some materials based on new observations.

Finally, Chapter 6 will go deeper into the interpretations outlined above, well aware of the risks posed by such work, while Chapter 7 will allow you to roll up your sleeves, by showing you a small sample of applications of the methodology.

[38] This opens the way for semiometric marketing applications justifying the approximation made earlier at the beginning of the chapter with investigations on lifestyles.

CHAPTER 2

Stability of the Semiometric structure

We shall study in this chapter the stability of the structure highlighted in Chapter 1, and try to answer the following question: is this structure a statistical accident or an artifact devised by the methods used? Bear in mind that we defined this structure as a set of six clearly identifiable main axes, to which we gave conventional names, although everyone can recognize the words associated with the axes, regardless of the name chosen.

Initially, stability will be tested empirically on different samples, collected at different times for the same country (France), bringing our attention to *stability over a period of time*. Next, samples from different countries will be analyzed. In this case, we will be dealing with geographical *stability*. This stability concerns in fact mainly the Western world. It has been established from seven independent sample surveys in seven countries. Interestingly enough, the eighth survey, which relates to a country outside the Western world (China / Hong-Kong) provides us with quite a different structure, emphasizing the importance of cultural background. Finally, limiting ourselves to the set of Western countries, the semiometric axes will be recalculated in specific sub-populations within a country (France), studying men and women and different age groups separately. We can speak of *internal stability* in this case. Some exceptions and peculiarities are nevertheless observed. They will be limited, but interesting to interpret, as small deviations from a model or standard can be significant. We can thus notice some subtle differences between Northern and Southern Europe. The axes are stable, but their

order of appearance (i.e. their statistical importance in terms of variance) may vary. We shall also see later on, in Chapter 3, through the study of the only diachronic series available, the slow drift of the French population within this structure over the last twelve years.

In a second step, this stability will be tested in a more technical manner, using the tools of statistical validation and simulation. The samples will be disturbed by several variants of a *re-sampling method* known as bootstrapping[39] to test the semiometric structure. The structure, let's make it clear from the outset, will be largely insensitive to these disturbances. Obviously, as regard Western countries, this chapter will engage in an exercise, which is perhaps not very exciting for a reader in search of new adventures ... as it in fact demonstrates that we always find the same thing, diachronically, and geographically (within the Western World) or even within certain sub-categories of individuals. The reader who trusts us concerning the surprising stability of the semiometric structure can move on to the next chapter, and simply refer back to this one, if in doubt, during his future exploration within the richness and complexity of the universe of words.

2.1 Empirical Stability per subpopulation

2.1.1 Diachronic Stability

The semiometric questionnaire was handed out in France over a period of twelve years (1990, 1996, 1998, 1999, 2001 and 2002) for the purpose of marketing applications[40] and to monitor certain aspects of the changing values of the French. The comparison of these successive waves of surveys obtained from independent samples revealed a high degree of structural stability (see Table 2.1).

But even if semiometric structures do not intrinsically change, there are slight differences in the positioning of certain categories. Changes of low amplitude can be identified according to sex, age and especially profession. Thus, farmers have expressed more and more over the last decade the values near the pole "Detachment" (one pole of axis 3:

[39] The principle of "bootstrap" is outlined in Appendix A1.9.5.

[40] Application examples are discussed in Chapter 7.

"Attachment / Detachment") and they seem less characterized by the cluster "Duty" of axis 2 "Duty / Pleasure". Small-scale traders (craftsmen, shopkeepers) and workers were approaching the pole "Pleasure" of that axis in 1998, a pole characterized at that time mainly by managers and middle management. These category modifications in semiometric space, small but statistically significant, require, a multidisciplinary approach involving statisticians, sociologists, and economists in order to be interpreted. Such work is beyond the scope of the more specifically methodological presentation given in this book. We shall see in Chapter 3 that the way we attribute a score explains another significant development that cannot be detected by simply comparing the analyses performed independently on each sample. For an analysis of the changing values of the French, we refer you to paragraph 7.7 of Chapter 7 on the topic of diachronic comparisons between samples from the same population, surveyed at different times.

How to read table 2.1 : In 2002, for example, the word *Courage* has the highest correlation with axis 1. It is followed by *Dynamic*, *Softness*, *Effective*, etc. The word *Disorder* is the most negatively correlated, followed by *To Betray*, *War*, etc.

The same rule applies for tables 2.2 to 2.6.

Table 2.1: Words describing "axis 1" from 1990 to 2002 "Methodological axis of participation"

Axis 1: 2002	Axis 1: 1999	Axis 1: 1996	Axis 1: 1990
Courage	*Courage*	*Courage*	*To protect*
Dynamic	*Effective*	*To protect*	*Purity*
Softness	*Dynamic*	*Effective*	*Courage*
Effective	*Accuracy*	*Dynamic*	*Honest*
Liveliness	*Reward*	*Robust*	*Reward*
Respect	*Elegance*	*Honest*	*Effective*
Reward	*Robust*	*Politeness*	*To admire*
Honest	*Charitable*	*Purety*	*Dynamic*
Confidence	*Confidence*	*Elegance*	*Respect*
Central zone			
To break	*Death*	*Doubt*	*Death*
To attack	*Danger*	*To break*	*Emptiness*
Death	*To break*	*War*	*Fault*
Danger	*Fault*	*Emptiness*	*Danger*
Emptiness	*Emptiness*	*Fault*	*War*
Anguish	*Anguish*	*Danger*	*Anguish*
War	*Disorder*	*Anguish*	*To break*
To betray	*War*	*Disorder*	*Disorder*
Disorder	*To betray*	*To betray*	*To betray*

Table 2.2: Words describing "axis 2" "Duty / Pleasure"

Axis 2: 2002	Axis 2: 1999	Axis 2: 1996	Axis 2: 1990
Sensual	*Sensual*	*Sensual*	*Sensual*
To dream	*To dream*	*To dream*	*Adventurer*
Voluptuous	*Lightness*	*Adventurer*	*To dream*
Island	*Wild*	*Nudity*	*Wild*
Lightness	*Adventurer*	*Original*	*Original*
Mellow	*Island*	*Voluptuous*	*Nudity*
Original	*Original*	*Desire*	*Bohemian*
Ocean	*Voluptuous*	*Bohemian*	*Lightness*
Wild	*Nudity*	*Island*	*Voluptuous*
Central zone			
Rigid	*Rule*	*To unish*	*Law*
Priest	*Law*	*Rule*	*Honour*
Elite	*Industry*	*Industry*	*To forbid*
Soldier	*Priest*	*To economize*	*To economize*
Industry	*Soldier*	*Morals*	*Soldier*
Homeland	*To obey*	*Soldier*	*Homeland*
To obey	*Homeland*	*Homeland*	*Discipline*
Morals	*Morals*	*Discipline*	*Morals*
Discipline	*Discipline*	*To obey*	*To obey*

Table 2.3: Words describing "axis 3" from 1990 to 2002 "Attachment / Detachment"

Axis 3: 2002	Axis 3: 1999	Axis 3: 1996	Axis 3: 1990
Danger	*Danger*	*Danger*	*Danger*
Storm	*Death*	*Storm*	*Doubt*
Rebellion	*To punish*	*Death*	*Anguish*
To criticize	*Emptiness*	*Desert*	*To break*
Death	*To break*	*Maze*	*Cry*
To atttack	*To attack*	*Rebellion*	*Emptiness*
Cry	*Anguish*	*To break*	*Maze*
Fire	*Maze*	*To criticize*	*Death*
Detachment	*To criticize*	*Cry*	*Storm*
Central zone			
Wealth	*Gold*	*Free*	*Wealth*
Wedding	*Liveliness*	*To inherit*	*Reward*
Reward	*House*	*Reward*	*Caress*
Gift	*Reward*	*Gold*	*Liveliness*
House	*Family*	*Money*	*House*
Money	*Wedding*	*Wealth*	*Gold*
Comfort	*Comfort*	*Present*	*Comfort*
Family	*Present*	*Jewel*	*Present*
Jewel	*Jewel*	*Comfort*	*Jewel*

Table 2.4:Words describing "axis 4" from 1990 to 2002 "Sublimation / Materialism"

Axis4: 2002	Axis 4: 1999	Axis 4: 1996	Axis 4: 1990
Wealth	*Wealth*	*Power*	*Wealth*
Power	*Power*	*Money*	*Power*
Gold	*Gold*	*Wealth*	*Money*
Glory	*Money*	*Gold*	*Gold*
Speed	*Glory*	*Glory*	*Glory*
Money	*To conquer*	*Craftiness*	*To conquer*
Ambition	*Speed*	*Speed*	*Speed*
To conquer	*To command*	*Ambition*	*Property*
Central zone			
To teach	*To write*	*Art*	*Forgiveness*
To ponder	*Tree*	*Forgiveness*	*Soul*
School	*Foreigner*	*Charitable*	*Flower*
Tree	*School*	*Poetry*	*Charitable*
Theater	*Poetry*	*Meditation*	*Art*
Poetry	*Theatre*	*Nest*	*Tree*
Art	*Art*	*Tree*	*Poetry*
Book	*Book*	*Book*	*Book*

Table 2.5: Words describing "axis 5" from 1990 to 2002
"Idealization / Pragmatism"

Axis 5: 2002	Axis 5: 1999	Axis 5: 1996	Axis 5: 1990
God	*God*	*God*	*God*
Faith	*Faith*	*Faith*	*Faith*
Priest	*Soul*	*Soul*	*Priest*
Sacred	*Meditation*	*Priest*	*Soul*
Soul	*Sacred*	*Sacred*	*Eternal*
Meditation	*Priest*	*Eternal*	*Sacred*
Creator	*Creator*	*Ceremony*	*Meditation*
Eternal	*Eternal*	*Meditation*	*Jewel*
Infinite	*Noble*	*Sovereign*	*Ceremony*
Central zone			
To ponder	*To produce*	*Practical*	*Headstrong*
Material	*To master*	*Clever*	*Matérial*
To master	*Practical*	*To produce*	*Robust*
Headstrong	*Solid*	*To ponder*	*To master*
Practical	*Clever*	*To master*	*Effective*
Solid	*To build*	*Effective*	*Accuracy*
Logical	*Dynamic*	*Accuracy*	*To produce*
Accuracy	*Accuracy*	*Solid*	*Solid*
Effective	*Effective*	*Logical*	*Logical*

Table 2.6: Words describing "axis 6" from 1990 to 2002
"Humility / Sovereignty"

Axis 6: 2002	Axis 6: 1999	Axis 6: 1996	Axis 6: 1990
Creator	*Entrepreneur*	*Noble*	*Noble*
Art	*Creator*	*Creator*	*Elite*
Audacity	*To master*	*Art*	*To command*
Robust	*Art*	*Elite*	*Audacity*
Inventor	*Robust*	*God*	*God*
Effective	*Solid*	*Audacity*	*Creator*
Entrepreneur	*Inventor*	*Sovereign*	*Art*
Accuracy	*Audacity*	*Faith*	*Absolute*
Central zone			
Rigid	*To obey*	*To obey*	*Tenderness*
Fault	*Wedding*	*Doubt*	*Anguish*
Emptiness	*Family*	*Wedding*	*To obey*
Wedding	*Tenderness*	*To forbid*	*Rigid*
Mistrust	*Mistrust*	*Family*	*Fault*
Anguish	*Doubt*	*Fault*	*Doubt*
Family	*Fault*	*Mistrust*	*Mistrust*
Birth	*Birth*	*Birth*	*Birth*

Although we can observe significant category changes in the semiometric planes, the positions of words in these planes remain stable, as evidenced by Tables 2.1 to 2.6, which describe systematically the changes in axes 1 to 6 for the four years: 2002, 1999, 1996, 1990. In this series of tables, each axis for each year is described by the most extreme 20 words, viz. by the 10 words that are most positively correlated, and the 10 words that are most negatively correlated.

2.1.2 Stability per country

The semiometric questionnaire has been submitted to samples from several European countries[41], from Canada and the United States, and from China (Hong-Kong) for applications such as marketing. One of the unknowns in this major venture was the effect translating the questionnaire had on the results.

Table A2.1 presented in Annex A2 lists the correspondences between words in five languages. It shows both the extent and limits of the difficulties encountered. Many words do not pose major problems and others involve making choices, which are sometimes difficult[42].

The results are presented for eight countries: France, Britain, Germany, Italy, Spain, and Canada (Quebec), United States of America, China (Hong-Kong). A structural stability has been verified, exemplified by figures 2.1 to 2.6, all of them relating to Western countries. Figure 2.7 and 2.8 concern an exception: the case of China (Hong-Kong). In this latter case, the structure is markedly different, underlining the probable influence of cultural differences between Western and Asian countries.

Stability and nuances within the Western world

However, within the Western world, there are also some peculiarities which, without questioning the overall stability, bring out some cultural particularities.

---

[41] France, Great Britain, Germany, Italy, Spain.

[42] Thus, *God, Water, A tree, Metallic, A nest*, do not pose major problems as a first approximation, and for the countries concerned. However, *Money, To ponder, A priest*, are words whose translation is not immediate (the French word: *Argent* could be translated either by *Silver* or *Money* in English) and require a rational choice, even an arbitrary convention. But as we will see below, the results will also be stable through *perturbation* of the word list, and the translation may be regarded as an operation involving a disturbance within tolerable limits.

The first of these particularities is the exchange of axes 2 and 3, "Duty / Pleasure" and "Attachment / Detachment" between the countries of Southern Europe and Northern Europe: for the Spanish, French and Italians, the "Duty / Pleasure" axis appears before the "Attachment / Detachment" one. Conversely, for the Germans, British, Canadians and citizens of the United States, it is the dimension of "Attachment / Detachment" which appears in second position, and therefore comes before the "Duty / Pleasure" axis, thus now occupying the third position[43]. The axes themselves are not involved, but their hierarchy is[44].

Chapter 3 will show that this inversion of axes may be attributable, at least in part, to the "scoring effects" that characterize the first and third axes, and that reflect cultural differences in attitude toward the survey and/or questionnaire as well as behavioral differences with respect to a scale of scores.

This peculiarity is also found in the analyses that can be carried out within certain age groups. This will be discussed when these analyses are presented (Section 2.1.4).

The second peculiarity concerns axis 4, conventionally called "Sublimation / Materialism". If, on the side of the pole "Materialism", we find, for the five countries, a large number of common words such as: *Gold, Money, Jewel* (except Canada), *Wealth, To conquer, To win, To seduce*, etc. – all words which very clearly express the quest for enjoyment through the conquest and possession of material goods –, words which are opposite lie on the other end of the axis, are quite different from country to country.

We can divide these countries into three groups: France, where the concept of "Sublimation" is above all - but not exclusively - expressed by all the words of the questionnaire referring to the arts: *Book, Art, Theater* and *Poetry…*, Canada and Great Britain where the concept is mainly rendered by words related to nature: *Tree, Mountain, River, Ocean, Moon* and *Green*, whereas with Spain and Italy words appear like: *Death, Sacrifice, Humble, Fault* (except Spain), *Doubt…*, which are completely

[43] This result is confirmed by the axes of Norway, Finland and the Czech Republic, which are not reported here.

[44] We can therefore say that to describe all the correlations between words, the opposition "Duty / Pleasure" is more important (in term of explained variance) than the opposition "Attachment / Detachment" in southern Europe than in northern Europe.

absent in the other countries, with Germany and the United States providing a synthesis of these three groups[45].

An exception: the Hong-Kong semiometric survey

It did not come exactly as a surprise, but nevertheless, the difference of structure between our unique Asian survey and all the other surveys was beyond our expectations.

The translation (from the English version of the questionnaire) was a challenge, as usual. Some words are polysemic in one language, and not in the other, emphatic in one language, and derogatory in the other. However, the translation was carried out with the utmost care, involving several specialists and various iterations, under permanent control using back translation.

Figure 2.7 shows the space spanned by the two first principal axes. Unlike all the other surveyed countries, in which the first component is a dominant axis of "level of notation" (see next chapter for a thorough study of this methodological issue), we are dealing here with two highly dominant components [technically speaking dominant, in terms of eigenvalues].

Figure 2.7 shows the parabolic form of the cloud of words. The diagonal line marked out by the labels "INDIFFERENCE" and "INTEREST" opposes words with low scores to words with high scores.

However the plane (3, 4) represented in figure 2.8 vaguely resembles subspaces observed in other countries, but a deeper scrutiny discourages us from giving these axes one of the previous labels (Duty, Pleasure, Attachment, …).

Evidently, specialists of oriental culture and values could go further into the interpretation of the Hong-Kong survey. It seems probable that the questionnaire is not adapted to countries outside the Western World. But, to what extent? Why?

---

[45] We may, concerning such results, assume that cultural particularities can be the cause of the differences observed, between the French opposing *art* with *money*, the English, the *gold of the City* to the *English countryside* and the inhabitants of Southern Europe, Germany and the United States opposing *gold* with *death*. But, aware of our limitations in such a subtle field, we leave that task to our readers and specialists.

The relatively modest size of the Hong-Kong sample (795, to be compare with, e.g., 9094 for the US sample) suggests we wait for more data, in particular from other regions of China, and hopefully from more Asian countries[46].

Briefly, our conclusion about this section devoted to geographic stability is that the structures described by the first axes are amazingly similar in all the surveyed Western countries. The only Eastern country surveyed provides results notably distinct.

In any case, the interpretation of the observed patterns in Honk-Kong remains an open and stimulating piece of research.

[46] Note that we do not question the quality of the survey within the Hong-Kong area. Some socio-demographic variables (such as sex, age of the respondent) are consistent with the answers to the semiometric questionnaire [we can observe statistically significant locations of these variables in the plane (3,4) shown in Figure 2.8].

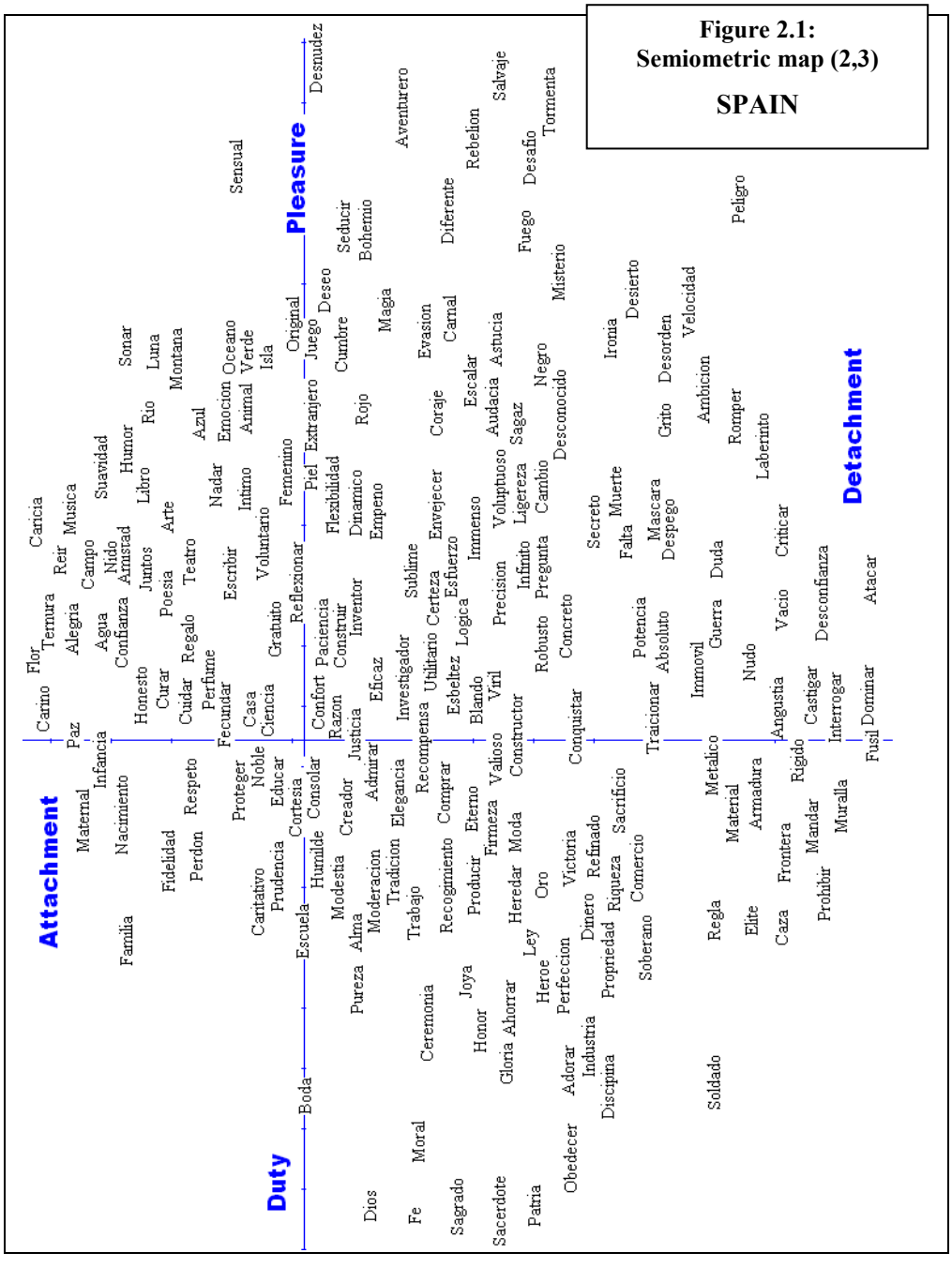


Figure 2.1:
Semiometric map (2,3)

SPAIN

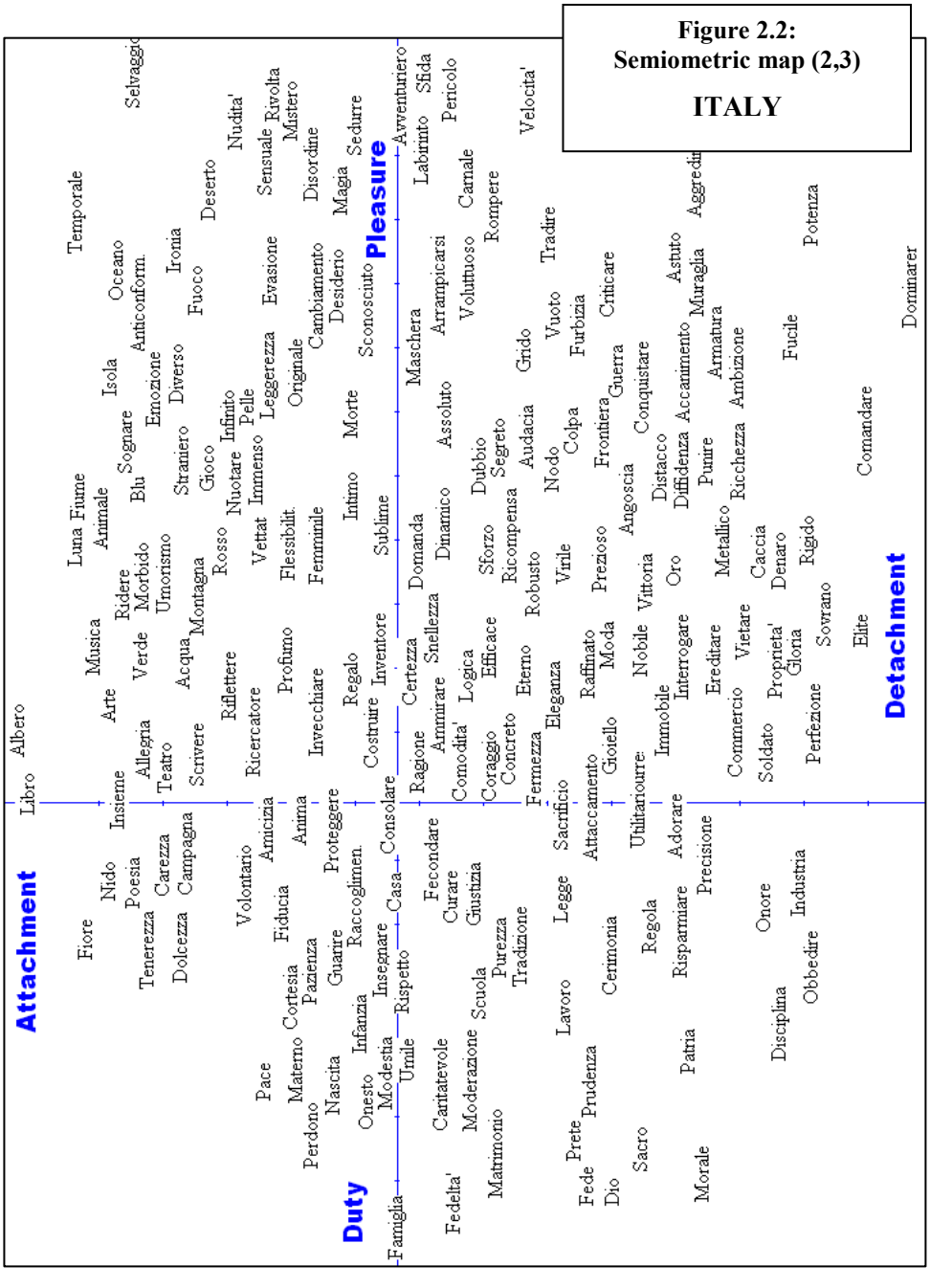
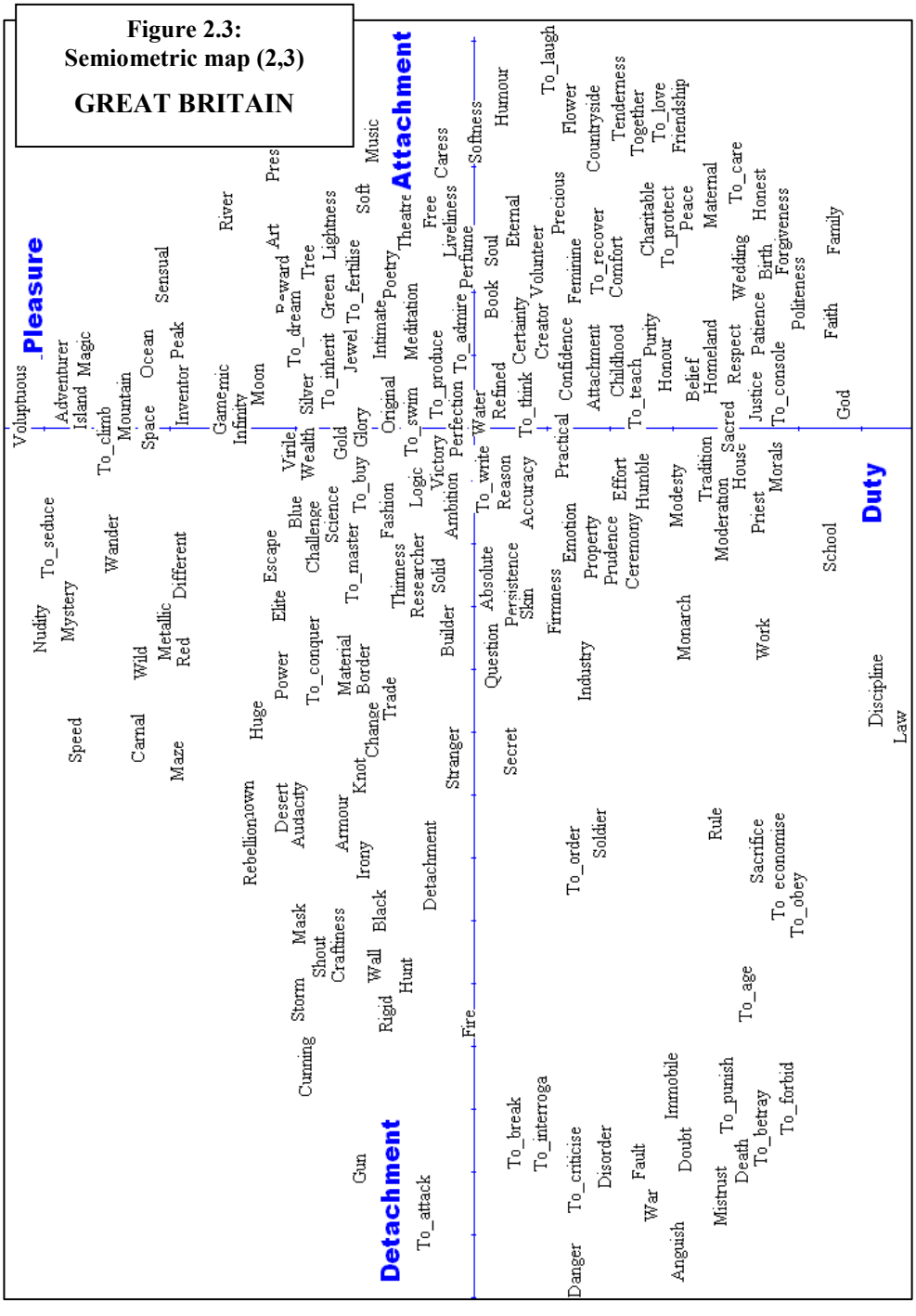


Figure 2.2:
Semiometric map (2,3)

ITALY

Figure 2.3:
Semiometric map (2,3)

GREAT BRITAIN



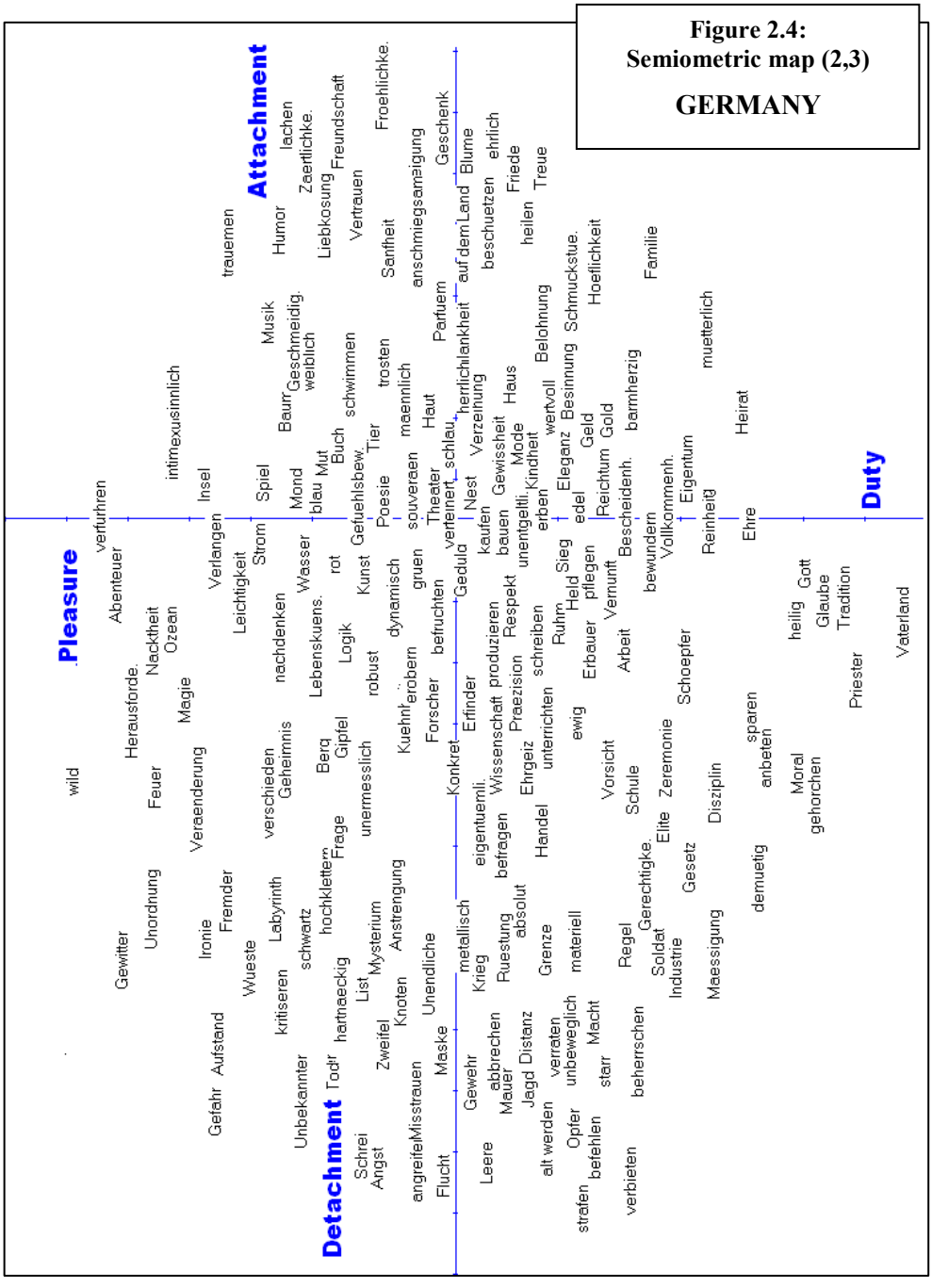


Figure 2.4:
Semiometric map (2,3)

GERMANY

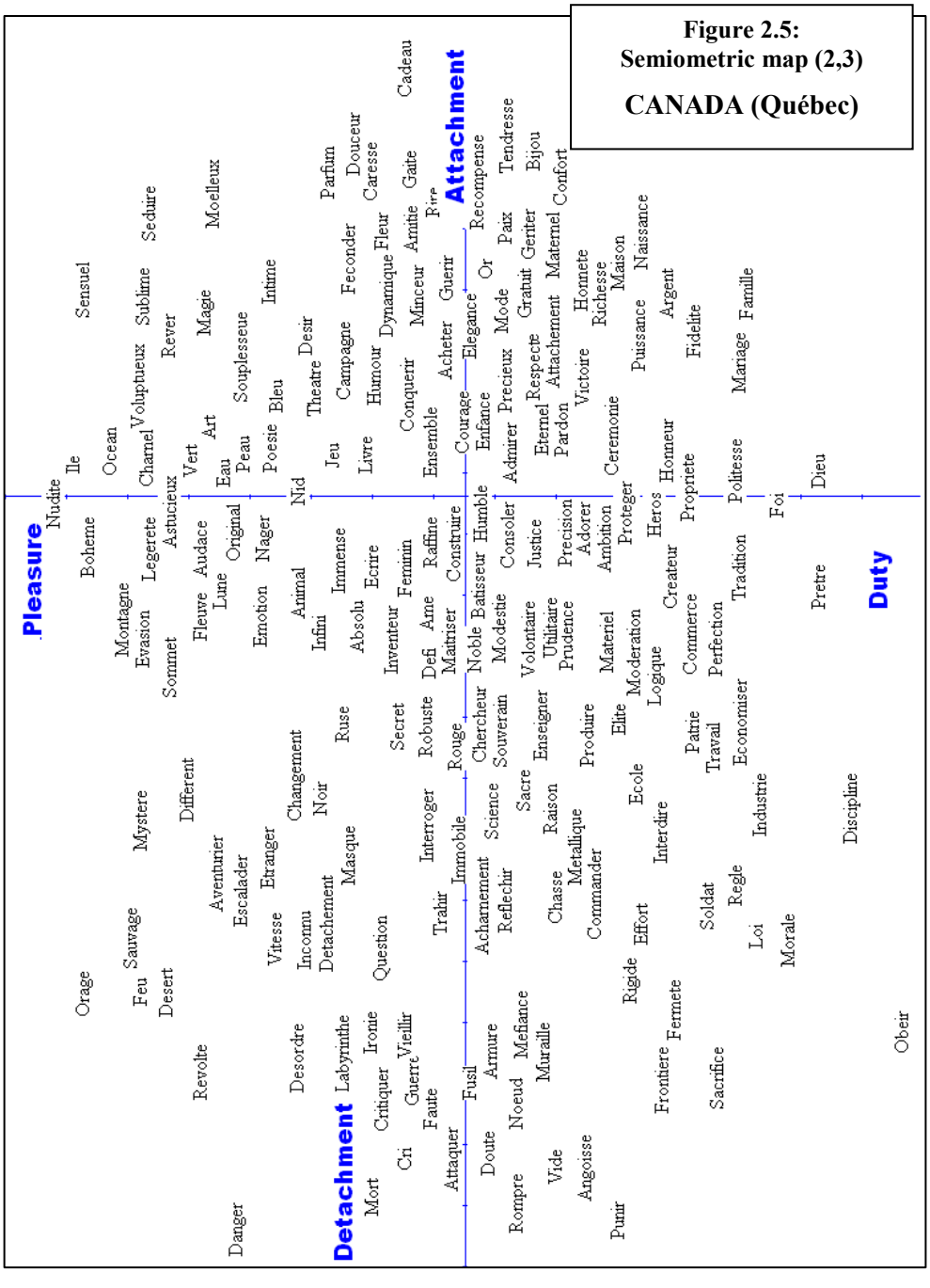


Figure 2.5:
Semiometric map (2,3)

CANADA (Québec)

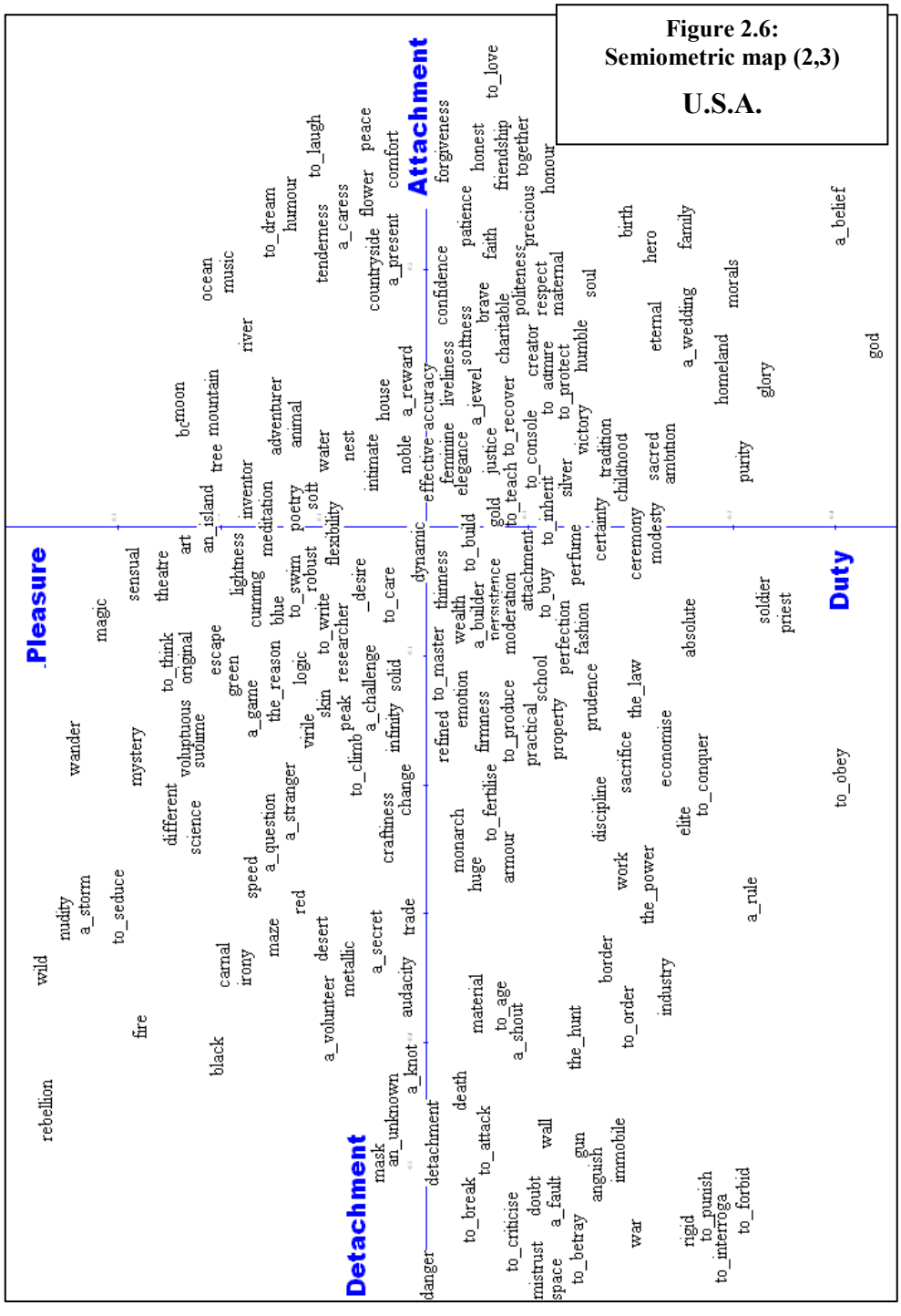


Figure 2.6:
Semiometric map (2,3)

U.S.A.

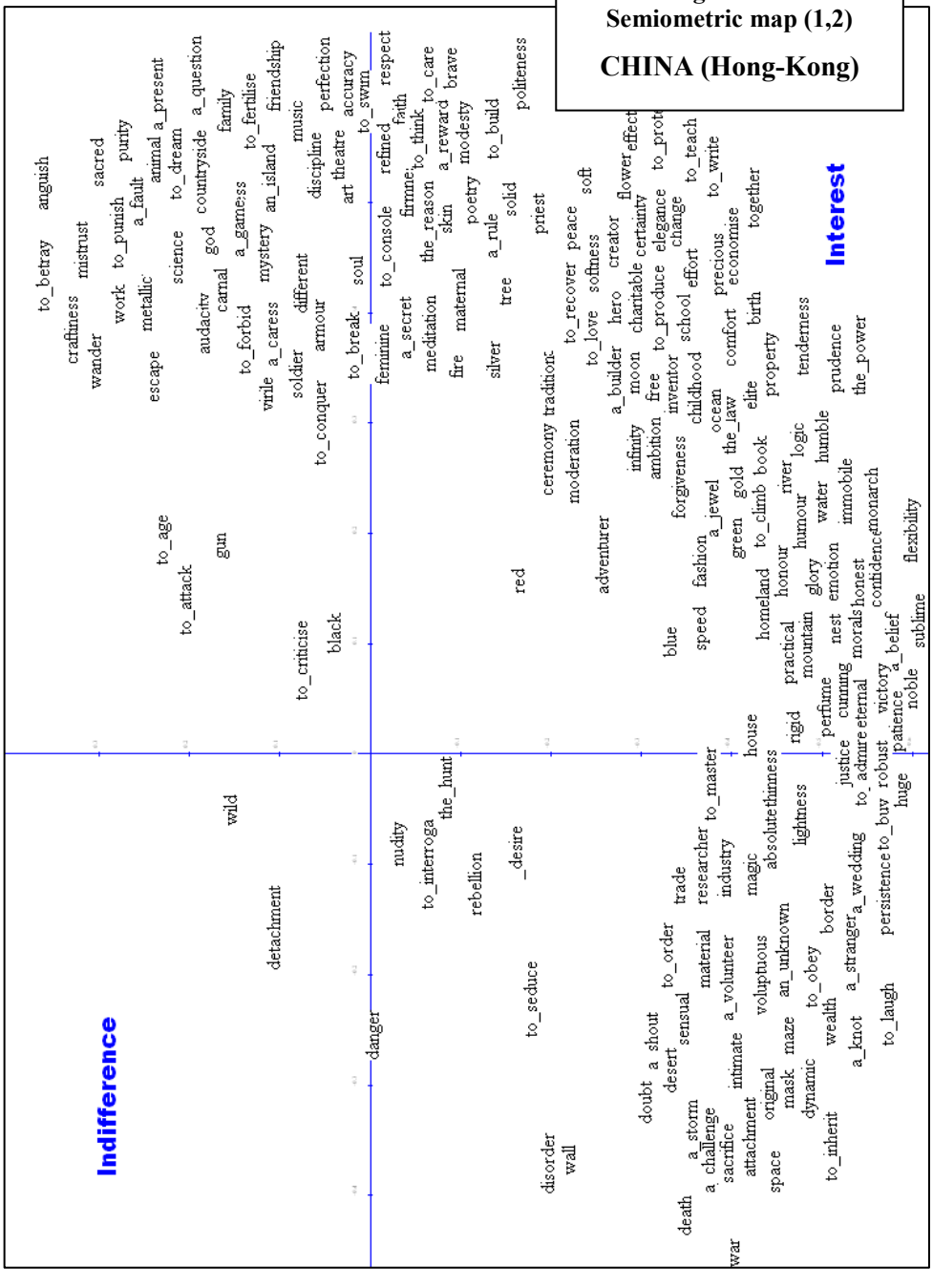


Figure 2.7:
Semiometric map (1,2)

CHINA (Hong-Kong)

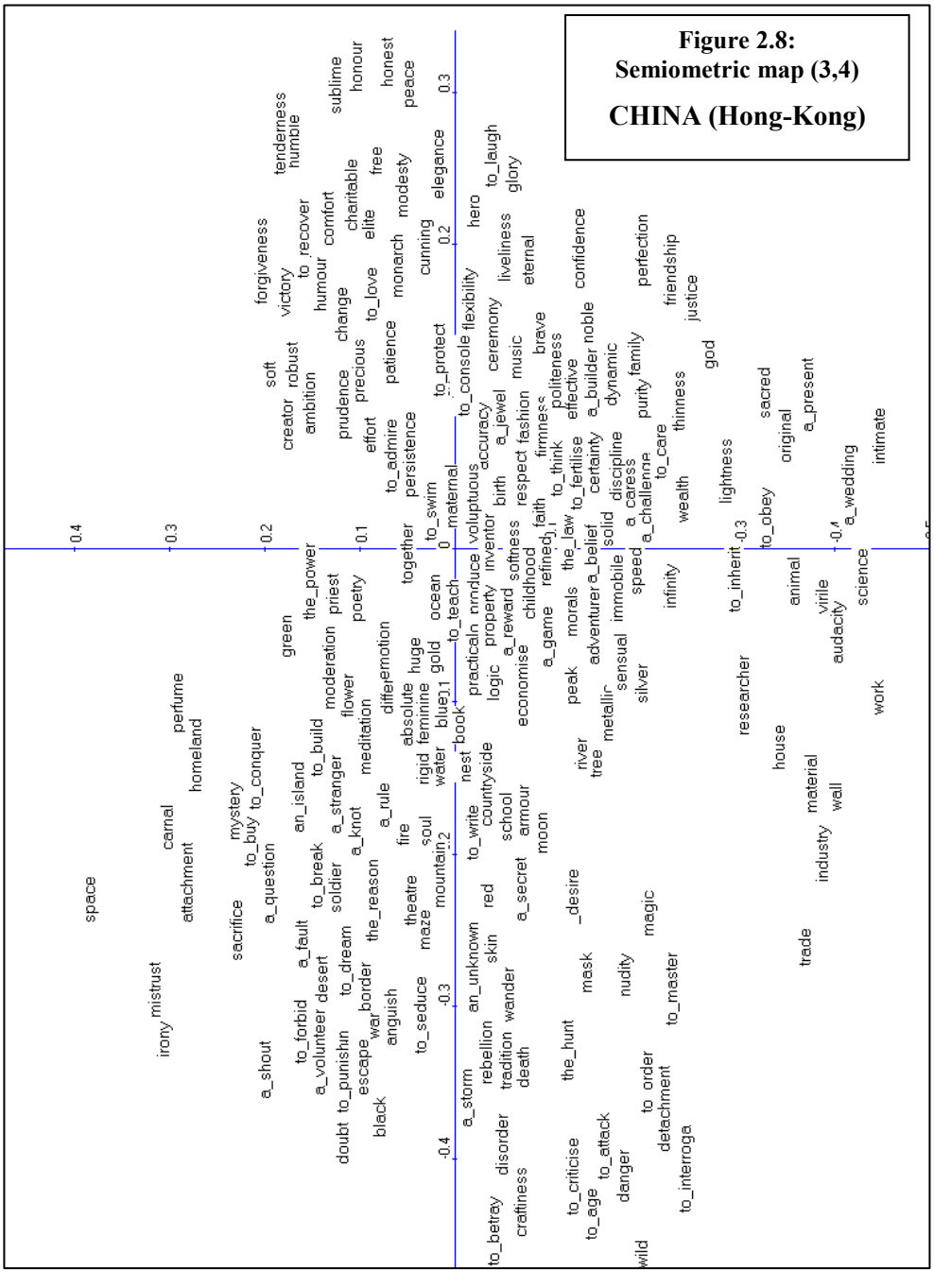


Figure 2.8:
Semiometric map (3,4)

CHINA (Hong-Kong)

2.1.3 Stability according to sex categories

We saw in the previous chapter that axis 3 (conventionally entitled: "Attachment / Detachment") and axis 5 (conventionally entitled: "Idealization / Pragmatism"), obtained from an analysis of the overall population, contrasted the male and female populations. If the variable "sex" is really the hidden explanatory variable responsible for the major oppositions observed, the analyses performed only on a sample of women or only on a sample of men must no longer show this type of axis.

However, the overall semiometric structures remain more or less preserved during analyses involving only men or women. We find both stability and permanence, which are rather unexpected. The structure is the same for men and women, but some of the features of this structure can significantly separate men from women[47]. The shades of meaning picked out mainly stem from the rankings of words in the explanation of the axes involved. These rankings may vary slightly without disrupting their initial interpretation. Table 2.7 presents, for each axis, the first ten words that are most strongly correlated to each one[48].

The first three axes of the semiometric structure (axes 2-3-4) are stable. Values revealed by these axes are similar for men and women, despite significant differences observed in the common analysis, particularly for the third axis. Thus the concept of "Attachment" on axis 3 is a value which is, however, more a feminine value although it is found in men. On the other hand, "Detachment" is more a masculine value but is also found among women.

---

[47] This phenomenon of the occurrence of the same structure at different scales irresistibly evokes structures called *fractals*. This phenomenon, like the analogy that has been attached to it, was already noticed and reported by Jean-Pierre Pagès, as part of his investigations in Agoramétrie. The analogy with the *fractal* world is all the more significant as it is prolonged, subdivision by subdivision of the sample, up to the individual, which would bear itself a similar structure. We shall return to this theme several times, until it is time for us to conclude and discuss open problems and future research.

[48] The first axis whose status is particular is not considered here. However, it also proves to be stable. A whole chapter is devoted to this first axis (chapter 5).

Table 2.7: Opposition, from one axis to another, from the analyses conducted separately on males and females.

Axis 2: "Duty / Pleasure"		Axis 3: "Attachment / Detachment"		Axis 4: "Sublimation / Materialism"		Axis 5 : "Idealization / Pragmatism"		Axis 6 : « Humility / Sovereignty"	
Males	**Females**	**Males**	**Females**	**Males**	**Females**	**Males**	**Females**	**Males**	**Females**
Wild	*Sensual*	*To punish*	*Danger*	*Power*	*Power*	*God*	*God*	*Birth*	*Wedding*
Adventurer	*To dream*	*Danger*	*To break*	*Wealth*	*Wealth*	*Faith*	*Faith*	*Mistru.*	*Birth*
Sensual	*Voluptuo.*	*Death*	*Death*	*Gold*	*Gold*	*Soul*	*Soul*	*Fault*	*Family*
To dream	*Nudity*	*To break*	*Cry*	*Money*	*Money*	*Priest*	*Sacred*	*Doubt*	*Ceremony*
Storm	*Adventur.*	*Anguish*	*Storm*	*Glory*	*Glory*	*Sacred*	*Priest*	*Rigid*	*To love*
Nudity	*Original*	*Maze*	*Desert*	*To conquer*	*To conquer*	*Meditati.*	*Meditation.*	*Caress*	*Doubt*
Original	*Lightness*	*Doubt*	*To criticize*	*Speed*	*Craftiness*	*Ceremony*	*Noble*	*Mask*	*Childhood*
Island	*Bohemian*	*Cry*	*Rebellion*	*Rifle*	*Ambition*	*Eternal*	*Creator*	*Tende.*	*Eternal*
God	*Rule*	*Weath*	*To inherit*	*Tree*	*Charitable*	*To ponder*	*Craftiness*	*Art*	*Entrepren.*
Faith	*Priest*	*Money*	*House*	*Forgiven.*	*Peace*	*To produce*	*Liveliness*	*Priest*	*To produce*
Priest	*To econom.*	*Free*	*Reward*	*To write*	*Poetry*	*Robust*	*Headstrong*	*Meditat*	*Clever*
To econom..	*Industry*	*Gold*	*Money*	*To teach*	*Theater*	*Clever*	*Honest*	*Soul*	*Accuracy*
Soldier	*Soldier*	*Liveliness*	*Wealth*	*Theater*	*Nest*	*Logical*	*To punish*	*Noble*	*Audacity*
Morals	*Homeland*	*Jewel*	*Present*	*School*	*Meditation*	*Solid*	*Politeness*	*Sacred*	*To master*
To obey	*Morals*	*Reward*	*Gold*	*Art*	*Art*	*To master*	*Mistrust*	*Faith*	*Robust*
Discipline	*To obey*	*Comfort*	*Comfort*	*Poetry*	*Book*	*Accuracy*	*Effort*	*God*	*Solid*
Homeland	*Discipline*	*Present*	*Jewel*	*Book*	*Tree*	*Effective*	*Power*	*Creator*	*Effective*

Table 2.8: Opposition, from one axis to another (analyses conducted separately in two age groups).

Axis 2:"Duty / Pleasure"		Axis 3: "Attachment / Detachment"		Axis 4: "Sublimation / Materiallism"		Axis 5: "Idealization / Pragmatism"		Axis 6: "Humility / Sovereignty"	
- 45s	+ 45s	- 45s	+ 45s	- 45s	+ 45s	- 45s	+ 45s	- 45s	+ 45s
Storm	*Rebellion*	*Flower*	*Present*	*Wealth*	*Wealth*	*God*	*God*	*Fault*	*Birth*
Wild	*Storm*	*Liveliness*	*Perfume*	*Gold*	*Power*	*Faith*	*Faith*	*Birth*	*Doubt*
Foreigner	*Wild*	*Softness*	*Mellow*	*Speed*	*Gold*	*Soul*	*Soul*	*Rigid*	*Mask*
Moon	*Bohemian*	*Tenderness*	*Voluptuous*	*Power*	*Money*	*Priest*	*Priest*	*Doubt*	*Caress*
Rebellion	*Desert*	*Peace*	*Jewel*	*To seduce*	*To conquer*	*Sacred*	*Sacred*	*Mistrust*	*Tenderness*
Fire	*Nudity*	*Present*	*To laugh*	*Money*	*Rifle*	*Meditation.*	*Meditation.*	*Caress*	*Ceremony*
Bohemian	*Original*	*Maternal*	*To dream*	*Sensual*	*Glory*	*Ceremony*	*Eternal*	*Knot*	*Fault*
Mystery	*Fire*	*Caress*	*Liveliness*	*Desire*	*To command*	*Eternal*	*Absolute*	*To forbid*	*Wedding*
To inherit	*Soldier*	*Emptiness*	*Danger*	*To age*	*Friendship*	*To ponder*	*Dynamic*	*Art*	*Clever*
To obey	*Wedding*	*Armour*	*To forbid*	*To teach*	*To teach*	*Robust*	*Practical*	*Noble*	*Noble*
Wedding	*Honour*	*To break*	*Cry*	*God*	*Book*	*Clever*	*Clever*	*Soul*	*Robust*
Homeland	*Wealth*	*Wall*	*Discipline*	*To obey*	*Peace*	*Practical*	*To produce*	*Sacred*	*Elite*
Property	*To econom.*	*Anguish*	*To break*	*Priest*	*To console*	*Logic*	*Material*	*Priest*	*Audacity*
Glory	*To inherit*	*Rifle*	*Anguish*	*Morals*	*Charitable*	*To master*	*Headstrong*	*Meditation.*	*Creator*
Wealth	*Homeland*	*Danger*	*Sacrifice*	*Faith*	*School*	*Accuracy*	*Effective*	*Creator*	*Effective*
Discipline	*Property*	*To punish*	*To obey*	*Meditation.*	*Confidence*	*Solid*	*Logic*	*Faith*	*Solid*
Money	*Money*	*To attack*	*To punish*	*School*	*Forgiveness*	*Effective*	*Accuracy*	*God*	*To master*

We have noticed for the pole "Attachment" in women, the presence of the words: *To buy, To inherit, Loyalty* and the absence of *Liveliness, Caress, To laugh, Softness* in the ten word list. So there are nuances in the characterization of the axis, shades of meaning that fade out or change if we use more words to describe this end of the axis.

2.1.4 Stability according to age

We have just seen that axes 3 "Attachment / Detachment" and 5, "Idealization / Pragmatism", seemed to be linked, in the overall analysis, to the sex of the respondents. Yet, they persisted when we eliminated the effect of sex categories, viz. by working separately on two sub-populations of females and males.

Now axis 2 "Duty / Pleasure" and to a lesser extent, axis 4 "Sublimation / Materialism" are themselves linked to the variable "age". What happens if one divides the population into two age categories: the "under 45s" and "over 45s" and if the investigations are carried out within each of these two sub-populations?

Again, there is a permanent structure except for a simple exchange of axes 2 and 3 for those over 45 years old, keeping in the background "Pleasure / Duty" with respect to "Attachment / Detachment" (see Table 2.8 above)[49].

However, the concept of "Pleasure" in "the under 45s" is similar, in terms of words that characterize it, to the notion of "Idealization": *Moon, Mystery, Art, Stranger, River, Ocean...* Similarly, the concept of "Duty" for the same sub-population, refers more to the concept of discipline and to some sort of material success, close to the pole "Materialism" of axis 4.

The "Attachment" of axis 3 is less materialistic in the "under 45s" (absent in the first 15 words are the following terms: *Gold, Reward, Money*) and more related to the words of "Pleasure" in the "over 45s" (*Voluptuous, Sensual, To seduce*). The "Detachment" is more aggressive and warlike for the younger group (*To attack, Rifle, Armor, War*) and more sacrificial for the older one (*To obey, Sacrifice, To forbid, Effort*).

[49] Not surprisingly, given the exchange of axes reported, the reader will see in Table 2.8 axis 2 of the "under 45s" compared to axis 3 of the "over 45s" (two columns) and vice versa (two columns).

The concept of "Sublimation" of axis 4 is less well defined for the under 45 whose values are partly reflected in the "Pleasure" of axis 2.

Note that for these stability studies according to age, like those related to sex in the previous section, the sub-samples are no more representative. This is obvious for the criteria in question (age or sex), but also for other criteria (occupation, education level, region ...), criteria which were not necessarily balanced within each age group or sex. As the sub-samples are no longer representative, the overall permanence of structures is even more surprising, and the small differences between observed semiometric structures seem ultimately modest.

2.2 Confidence intervals for eigenvalues

In this section, which is more technical, and in the following sections, stability and consistency of the semiometric structure will be tested using the tools of statistical validation and simulation. One is naturally led to ask a number of questions about the quality of representation: Do we really observe something? Do the data have a structure? Or, rather, are mere sampling fluctuations sufficient to contribute to explain the eigenvalue percentages obtained (percentages which, we must bear in mind, measure the importance of the axes)?

This question will first be treated within a classical statistical framework: there is indeed a procedure used to assess the confidence one can place on the first eigenvalues, i.e. the first axes. The magnitude of the confidence interval that will be calculated gives an indication of the stability of the eigenvalue with respect to the fluctuations due to sampling. The Anderson confidence interval (see appendix A1.9.4) for each eigenvalue somewhat describes their uncertainty. Table 2.9 and Figure 2.9, in the case of the first two eigenvalues, have confidence intervals for the first six eigenvalues in the case of three populations of different sizes.

The calculations focus on three semiometric samples of 2000, 10,000 and 15,000 individuals. The results are highly satisfactory in the sense that for a given sample size, most of the confidence intervals of successive eigenvalues do not overlap, indicating that they are significantly distinct. We can read, for example, in the first two rows of Table 2.9, that the upper bound of the interval of the second eigenvalue (11.60) is smaller than the lower bound of the interval of the first (24.00).

Table 2.9: Confidence intervals for the first six eigenvalues
(For three different sample sizes)

		Lower bounds	Eigenvalues	Upper bounds
Sample 2,000	Vp1	24.00	25.35	26.77
	Vp2	10.40	10.98	11.60
	Vp3	8.24	8.70	9.19
	Vp4	6.80	7.18	7.58
	Vp5	3.80	4.01	4.23
	Vp6	3.59	3.79	4.00
Sample 10,000	Vp1	25.49	26.19	26.91
	Vp2	10.07	10.35	10.63
	Vp3	8.58	8.82	9.06
	Vp4	6.82	7.01	7.20
	Vp5	4.04	4.15	4.26
	Vp6	3.58	3.68	4.78
Sample 15,000	Vp1	26.40	26.99	27.59
	Vp2	10.13	10.36	10.59
	Vp3	8.64	8.84	9.03
	Vp4	6.78	6.93	7.09
	Vp5	4.07	4.16	4.26
	Vp6	3.61	3.69	3.77

Figure 2.9 illustrates the (expected) decrease of the confidence intervals when the sample size increases. All this proves that the principal axes occupy significant directions, and that they are really and truly individualized. These methods in fact apply to observations distributed according to a normal or Gaussian distribution, but their robustness allows them to be used outside this framework. We shall see in the next paragraph how the methods of validation by re-sampling apply to the study of structure stability without any prior assumption on data distribution.

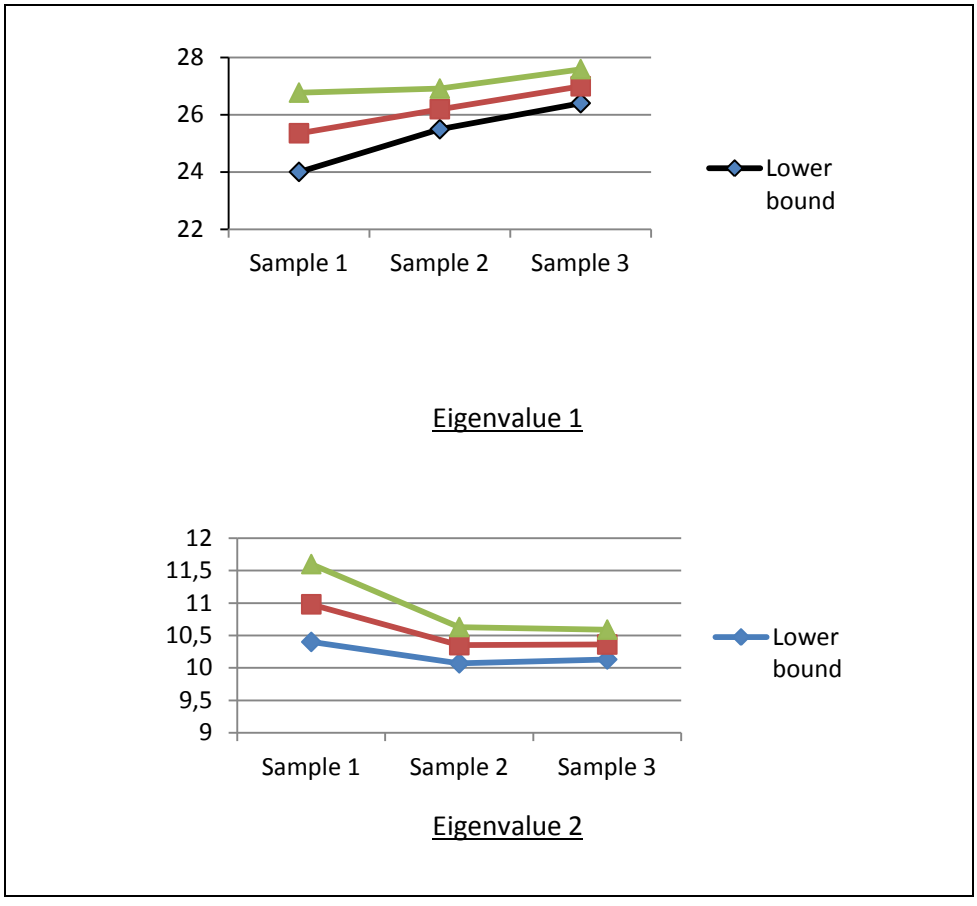


Figure 2.9: Changes in confidence intervals depending on the size of the sample.

2.3 Stability through re-sampling

The permanence of the structures observed on independent samples should satisfy us and convince us that when a set of words has been chosen, a series of axes or dimensions appear fairly systematically, not always in exactly the same order, but with very similar characterizations per word.

We know very little, however, about the stability of the positions of words with respect to one another. Their ranking on the axis has actually shown that they should be close to one population or another. But one can legitimately ask the question of how accurate the position of a word is on

the planes generated by two principal axes, taking into account the sampling fluctuations of our collections of answers.

Statisticians could not answer this question satisfactorily until recently. The power of new computing resources, however, has helped systematize the methods known as *re-sampling*. Among these methods, a statistical method of validation, known as the "bootstrap" method[50] can both verify the stability of the axes (which will only confirm the results of previous sections of this chapter), and assess the confidence that can be given to the position of each point on the principal planes.

The classic "bootstrap" method used here will disturb the original sample by simulating a large number of samples of similar size[51]. In this context, the simulations are called *replicates*. Two variants of this method will be used, the partial "bootstrap", and the total "bootstrap". They give similar results, confirming the stability of structures observed up to the sixth axis.

2.3.1 Partial "Bootstrap"

Figure 2.10 show, for a sample of 11055 respondents, some confidence ellipses obtained by the validation method according to principles of the partial "bootstrap" [52] relative to plane (2, 3). Plane (2, 3) is the first semiometric plane[53], the most often used in visualizations. We are dealing here with a sample, "France", comprising the four independent samples of 1990, 1996, 1999 and 2002, totaling 11,055 individuals, representative of the French population aged 18 and over. The ellipses are very small and therefore the fluctuations in the position of the points are very limited, despite the severity of the principle of *replication*.

---

[50] This metaphor coined by Efron (1979) refers to the expression "to pull oneself up by one's own bootstraps", which literally means "stand up by pulling one's own boot straps" or "move without help" (see Appendix A1.9.5).

[51] These samples are obtained by *n* random draws with replacement of the individuals in the initial sample of n individuals.

[52] The partial "bootstrap" aims to project the columns of the tables replicated (words) as additional elements on the axis of reference analysis, i.e. the analysis of the initial sample undisturbed. With 30 replications, we get 30 positions (for calculating density ellipses) for each word (see Appendix A1.9.5).

[53] Axes 2 and 3 are the two most important axes, since axis 1 is not part of the semiometric structure itself.

Figure 2.11 is the analogue of figure 2.10 for a sub-sample of 1000 respondents. Evidently, the confidence zones are larger, but the interpretation of the axes is by no means altered by this substantial reduction of the sample size.

Figure 2.12 concerns now the plane (5, 6), the last semiometric plane, for the 11,055 sample. We can observe that the ellipses relative to the plane are slightly larger than that of figure 2.10, while leaving the position of points acceptable accuracy.

By drawing from the sample with replacement, we can show that about 30% of individuals are excluded from each replication; other individuals, on the other hand, appear twice, three times, six times in exceptional cases.

2.3.2 Total "Bootstrap"[54]

Figures 2.13 and 2.14 show the same principal planes as before, now with larger ellipses[55]. Examining Figure 2.13 shows that areas of confidence overlap very little. Yet, this is a severe test (total bootstrap) and the semiometric plane (5, 6), was *a priori* less stable than the main plane (2, 3). The effects of sampling fluctuations are therefore low, confirming the stability of these axes and the robustness of the structures obtained.

[54] The total "bootstrap" aims to redo the complete principal components analysis on each sample replicated. As with the partial one, with 30 replications, we obtain 30 positions (for calculating density ellipses) for each word (see Appendix A1.9.5).

[55] In fact, the total "bootstrap" gives a pessimistic view of the precision of points, as principal components analysis is re-implemented for each replication, and therefore the projection planes are computed on perturbed data: the data are not only disturbed, but the visualization tool itself is disturbed (see Appendix A1.9.5).

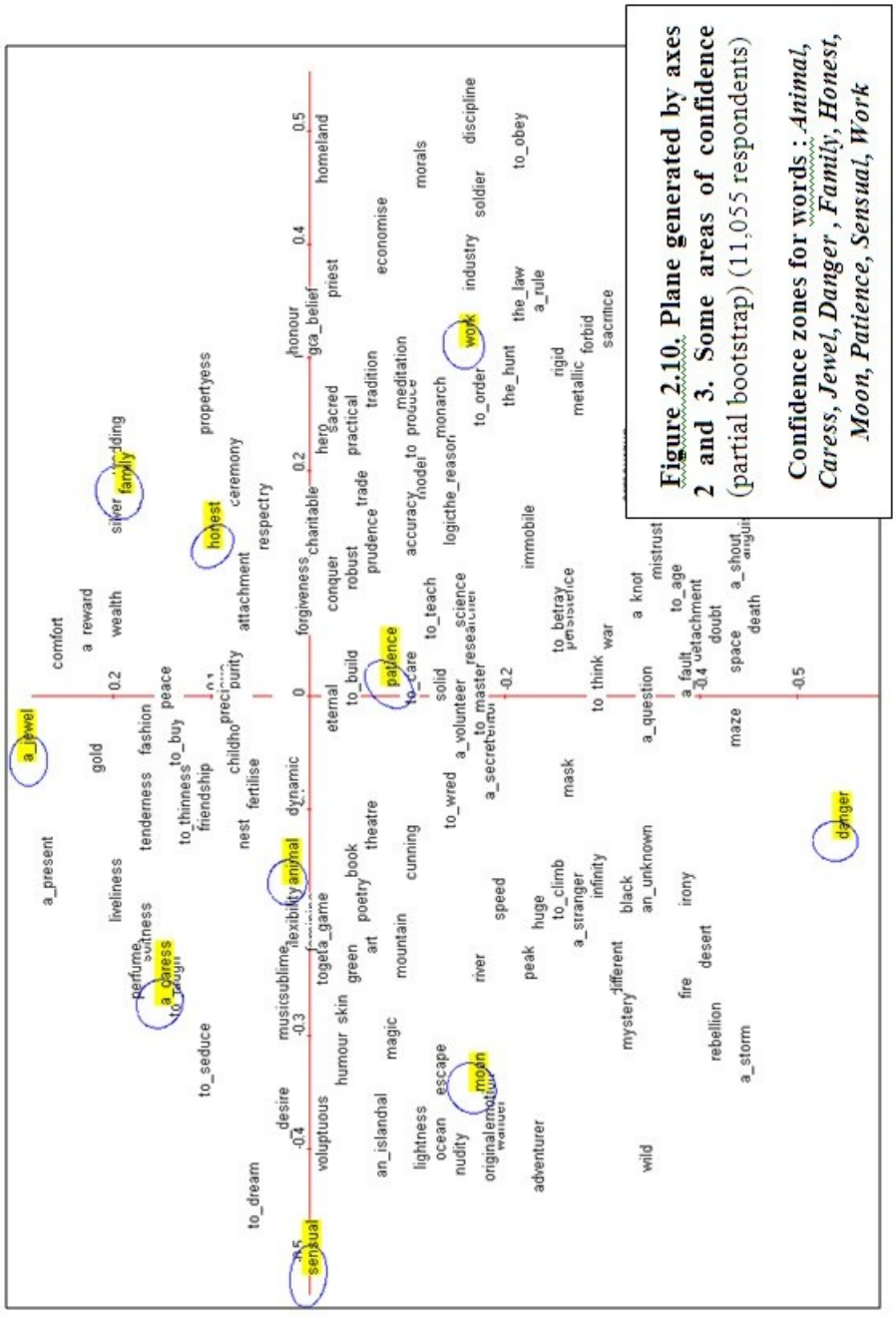


Figure 2.10. Plane generated by axes 2 and 3. Some areas of confidence (partial bootstrap) (11,055 respondents)

Confidence zones for words : ***Animal, Caress, Jewel, Danger , Family, Honest, Moon, Patience, Sensual, Work***

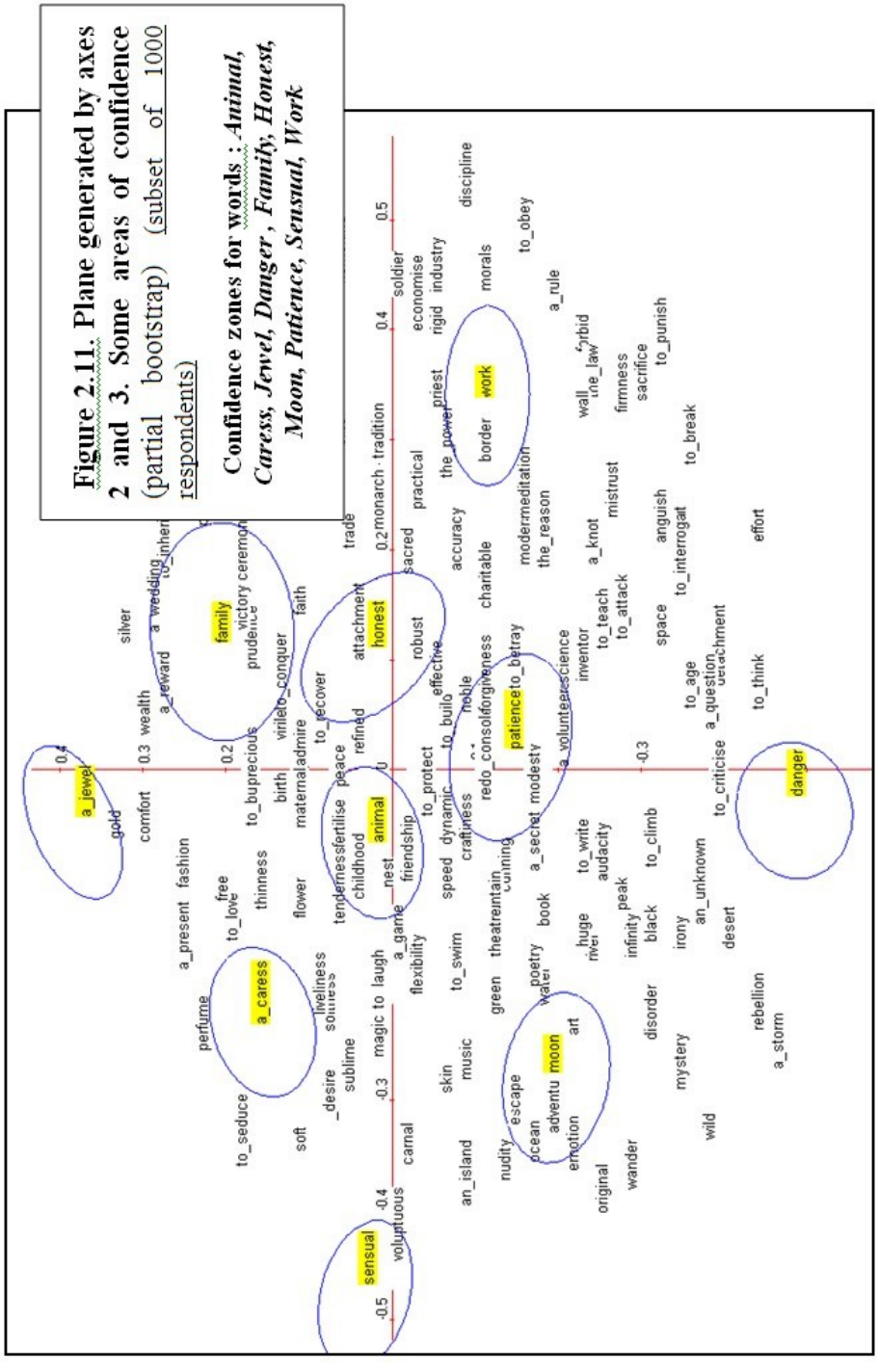
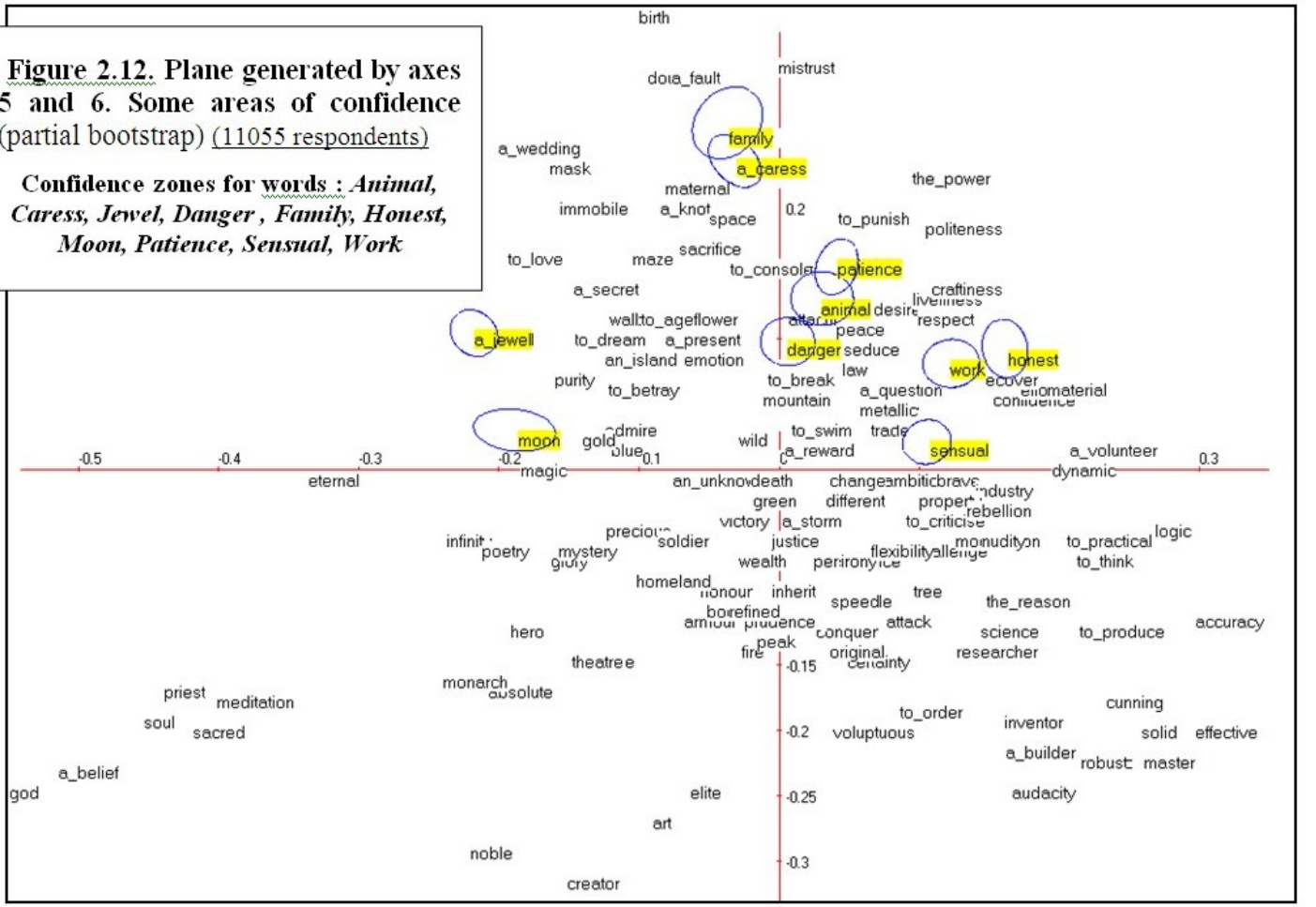


Figure 2.11. Plane generated by axes 2 and 3. Some areas of confidence (partial bootstrap) (subset of 1000 respondents)

Confidence zones for words : *Animal, Caress, Jewel, Danger , Family, Honest, Moon, Patience, Sensual, Work*

Figure 2.12. Plane generated by axes 5 and 6. Some areas of confidence (partial bootstrap) (11055 respondents)

Confidence zones for words : *Animal, Caress, Jewel, Danger , Family, Honest, Moon, Patience, Sensual, Work*



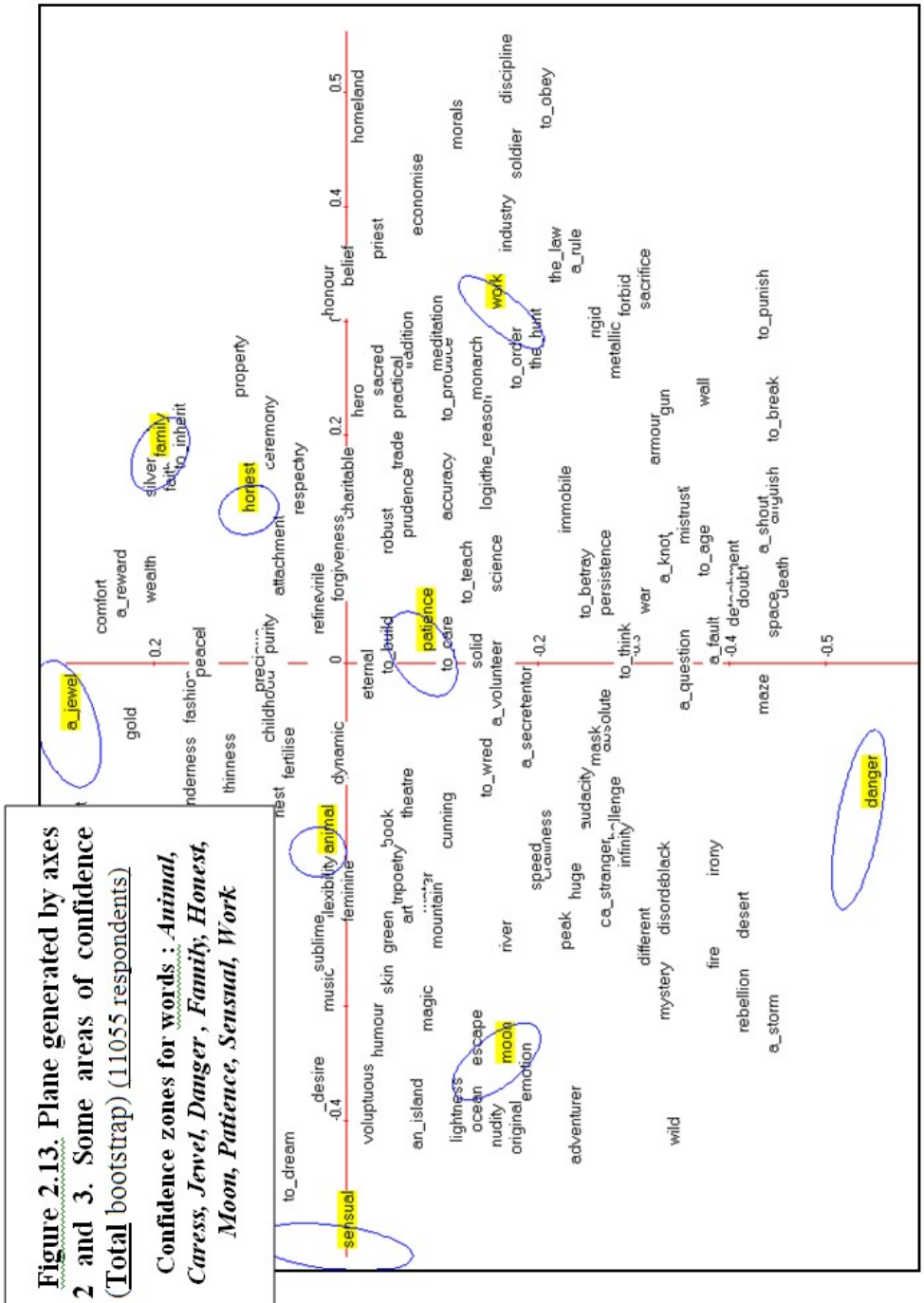


Figure 2.13. Plane generated by axes 2 and 3. Some areas of confidence (Total bootstrap) (11055 respondents)

Confidence zones for words : *Animal, Caress, Jewel, Danger , Family, Honest, Moon, Patience, Sensual, Work*

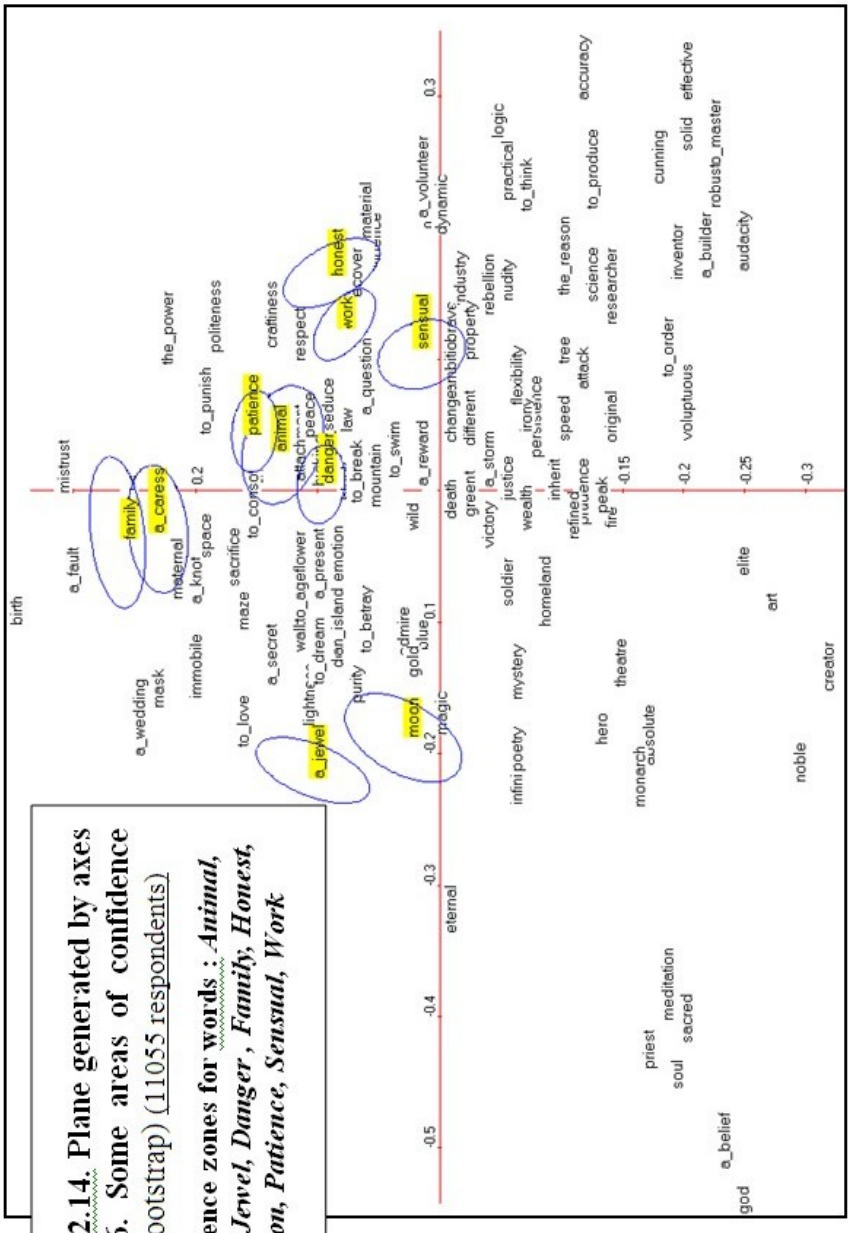
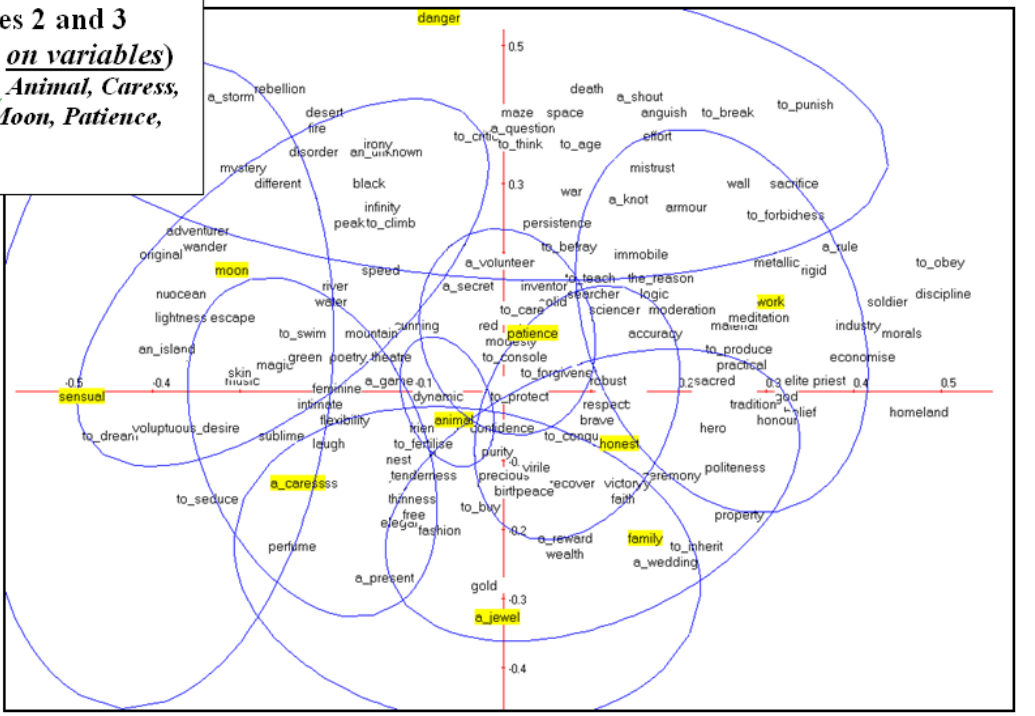


Figure 2.14. Plane generated by axes 5 and 6. Some areas of confidence (Total bootstrap) (11055 respondents)

Confidence zones for words : *Animal, Caress, Jewel, Danger , Family, Honest, Moon, Patience, Sensual, Work*

Figure 2.15 : Plane of axes 2 and 3
Confidence areas (*Bootstrap on variables*)
Confidence zones for words : *Animal, Caress, Jewel, Danger, Family, Honest, Moon, Patience, Sensual, Work*



In this rather daring experiment, we try to assess the stability of the axes vis-à-vis the composition of the set of words.

The 210 words are drawn at random with replacement for each replicate. A principal components analysis is performed on each replicate allowing us to draw confidence ellipses for the words.

We see for instance that some opposition between words along axes 2 and 3 remain significant (see section 2.4)

2.4 Structural stability

In the previous section of this chapter structural stability *vis-à-vis* the fluctuations of sampling was tested on all individuals using the "bootstrap" method. We are going to change register and ask a radically different question and one which is far more difficult: to what extent do the semiometric structures depend on the choice of words?

This question is actually more difficult because there is no universe of words similar to the universe of individuals who may be interviewed in a survey. This universe consists, for example, of all individuals aged 18 or older in a given country. Statisticians are used to treating the "individual" dimension or "observation" with which we can associate the concept of drawing from an urn and the corresponding mathematical models. There are few situations where one can talk about drawing or about a universe of variables[56].

We will nevertheless seek to test the structural stability *vis-à-vis* all the variables. We can then make the assumption that the variables (210 words) are selected by a random draw without putting them back into the urn containing all the "semiometricable"[57] words belonging to the French language (see section 1.1 of Chapter 1).

A draw from this universe every time implies the drafting of a new questionnaire and new field work, the cost of which would be prohibitive.

As in the case of samples of individuals, the "bootstrap" method provides a practical and economical solution to this problem: we will construct the replications of the set of words by drawing (with replacement) from the 210 words already selected.

While the classical "bootstrap" method, carried out on individuals, tests the stability of patterns of variables, i.e. words (see Section 2.2), the "bootstrap", carried out on the variables, is used to test the patterns that are

[56] There are circumstances where the problem of sampling arises for both variables and individuals. This old problem posed by Hotelling (1933) [author of a seminal paper about principal components analysis] has been particularly studied by Escoufier (1970).

[57] This, remember, concerns concepts - not grammatical words or tool words - part of the basic vocabulary, non consensual, semantically unambiguous - not polysemous - axiologically or emotionally charged - not the type of neutral words as *A table* nor technical words as *Camera*.

observable at the individual level. Now the individuals are anonymous (several thousand). The only interesting structures are those that are induced from their basic characteristics (sex, age, occupation, education level, activity, habitat types ...). The procedure described in Annex A1.2.4 allows one however to position the words corresponding to each sample replicated, including words that are missing from some replications[58].

The set of variables (210 words here) is much smaller than that of individuals (several thousand). On average, with each drawing with replacement, we can show that about two thirds of the initial variables are present, which is indeed a very severe gap in the original set, and thus a disruption that can be judged excessive in the data table.

Figure 2.15 shows the results of this procedure in terms of axes [2,3]: the confidence ellipses of the words are, of course, extremely dilated compared to those of the previous graphs, but the results are not disastrous. Along the second axis (horizontal) the area of elliptical confidence of the word *Sensual* remains opposed to that of the word *Work*. Similarly, along the third axis (vertical), the area of confidence of the word *Jewel* remains opposed to that of the word *Danger*. Interpretations of these axes are resistant to the drastic disturbances imposed on the collection of words.

The following axes show a stability that is not as clear, but one can think that a more extensive list of words (e.g. a list of 306 words, which was the case in early versions of Semiometry) would allow us to go further in applying this test of validity, which, moreover, is also extremely severe.

We must bear in mind and retain from this section that this procedure (novel) on "bootstrapping" variables provides a presumption - not proof - of structural stability *vis-à-vis* the choice of words. Chapter 4 dealing with "open" Semiometry will also make contributions to this problem of stability *vis-à-vis* the composition of the list, an interesting and important problem. However, it is not a crucial issue if Semiometry is used only for *comparing* and *contrasting*, which is the case in the context of current applications.

---

[58] In short: a "bootstrap" replication is sampling with replacement within 210 initial words. Some words are missing (they have zero weight), others will appear once (weight 1), twice (weight 2), etc. By replacing the zero weights by infinitesimal weights, which requires some additional technical improvements, the missing words will be automatically positioned as extra words (see Appendix A1.9.3).

2.5 Stability with respect to coding

Semiometric structure (existence and permanence of the six main axes characterized by the same words) appears stable geographically, diachronically, and *vis-à-vis* potential sampling fluctuations.

One could also argue, quite legitimately, that the basic coding is a rather arbitrary convention: for the respondent, it is primarily the order of the scores that matters, not the materialization thereof in the form of a scale of 1 to 7, with scores in arithmetic progression.

Would the underlined semiometric structure be the same if the encoding proposed was different (respecting however the order of the scores)?

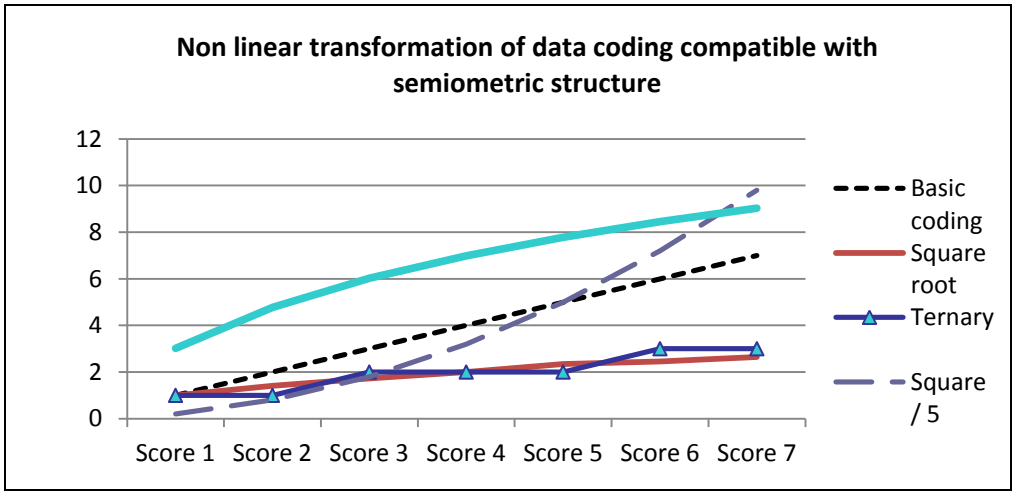


Figure 2.16: Various changes in the rating scale.

Note that any changes in the initial coding x into y such as: y = ax + b leaves the correlations invariant, and therefore leaves the entire semiometric structure identical. Thus, results are not modified by replacing (1, 2, 3, 4, 5, 6, 7) with (2, 4, 6, 8, 10, 12, 14) or with (11, 12, 13 , 14, 15, 16, 17).

To further distort the initial coding, we shall carry out *non-linear*[59] transformations (see Figure 2.16). We shall see that the structure is invariant for a large family of such transformations.

The result is as follows: for all these transformations, the first six semiometric axes are unchanged. This means that the axes appear in the same order and are characterized by the same words.

Table 2.10: Eigenvalues of the various tables recoded

Number	Base	Square	Log	Ternary	Sq. Root
1	26.6	28.7	23.5	22.1	25.1
2	10.2	10.3	9.6	8.2	9.9
3	8.7	8.4	8.7	7.1	8.8
4	7.0	7?0	6.4	5.9	6.8
5	4.0	4.0	4.0	3.6	4.1
6	3.7	3.7	3.3	3.2	3.5

We found that the first eigenvalue is a maximum for the square of the scores (codes strongly dilated from 1 to 49), and is a minimum for the ternary coding, squeezing the 7 values into three values.

The specific directions (or principal axes) are stable, i.e., characterized by the same groups of words. The eigenvalues, which describe the variances along these lines, are more sensitive to the coding without changing neither the order nor the meaning of the axes in any way.

It is interesting to note that the ternary[60] coding does not alter the structure: this indicates that the scale in 7 positions is probably quite redundant. But this does not necessarily prove (it should be checked, but this would mean returning to the field) that a scale in three positions, given directly to respondents, would restore the original structure.

---

[59] Nonlinear transformations following the initial arithmetic coding have been achieved (see Figure 2.16): a) $y = \sqrt{x}$; b) $y = x^2$; c) $y = \log_{10}(1+x)$; d) "ternary" coding defined below. In Figure 2.13, the square was divided by 5 and the logarithm multiplied by 10 to bring the different encodings on a scale comparable to the initial coding (or coding base on a dotted line). With a sample of 5,527 individuals (France), four new arrays of scores (5,527 x 210), were calculated by applying the transformations to the scores above.

[60] The ternary coding consists in agglomerating the seven values into three values: 1 for (1,2), 2 for (3, 4, 5) and 3 for (6, 7).

In contrast, a binary coding (for example: y = 1 for x = 1,2,3,4, y = 2 for x = 5,6,7) which is too destructive of information, significantly changes the structure for the worse. The first axis is approximately preserved (eigenvalue 19.3, lower than all the first eigenvalues of Table 2.4.1) but there is a rotation within the plane (1, 3). Surprisingly, the axes 4, 5, 6, are nevertheless preserved. With such coding, some words like: *War, Death, To betray…*, only get 1 from all individuals, others, such as *Peace* or *Tenderness* getting 2. Such words, whose variances are zero, no longer intervene in the correlations and are excluded from the structure.

2.6 Analysis through classical factor analysis

The method of analysis in common and specific factors, which dates back to the early twentieth century (see Annex A1.6), is related to principal components analysis, but generally occurs in a very different context: it is no longer a mere instrument of observation, but a model. This model states here that each of the 210 scores (thus corresponding to each word) depends on a small number of common factors and one specific factor (see the exact model in Appendix A1.6).

The common factors would correspond, for example, to the first six dimensions[61], and the specific factors in some way would summarize the 204 following axes, which are not taken into account.

This analytical model, which also requires some additional constraints, has little chance of being validated by a rigorous statistical test, even with thirty or fifty odd common factors; not least because the initial scores are discrete and not continuous variables. Yet, the adjustment of such a model (using the procedure outlined in Annex A1.6) with, for example, 12 common factors, generates, for the first six factors (in fact the coefficients of the words on these factors, called *saturations*), the first six axes of the semiometric structure.

The application of this method is contained in the chapter devoted to the stability of results because almost all of the previous conclusions on the interpretation of the axes also apply to common and specific factor analysis. The results have therefore resisted significant disruption of basic data,

[61] Principal components analysis in this case is a special case of factor analysis with 210 common factors (we keep all axes) and no specific factor. The specific factor is present to explain through a model the discarded axes.

changes in the coding, and also significant modification of the method of data reduction.One of the interesting byproducts of this latter type of analysis is the calculation, for each word, of its specific variance (specificity: a variance unexplained by common factors) whose complement at 1 is the common variance (sometimes called *community*).

Table 2.11: Words with highest and lowest common variances

Strongest community			**Weakest community**		
Words	Specificity	Common	Words	Specificity	Common
Faith	0.332	0.668	*Justice*	0.823	0.177
God	0.362	0.638	*Hunt*	0.825	0.175
Priest	0.447	0.553	*Border*	0.825	0.175
Meditation	0.480	0.520	*Detachment*	0.828	0.172
Sacred	0.491	0.509	*To swim*	0.829	0.171
Power	0.525	0.475	*To forbid*	0.832	0.168
Discipline	0.538	0.462	*Metallic*	0.836	0.164
Effective	0.538	0.462	*Change*	0.839	0.161
Sensual	0.559	0.441	*Game*	0.849	0.151
Wealth	0.562	0.438	*Stranger*	0.853	0.147
Money	0.568	0.432	*Knot*	0.854	0.146
Glory	0.571	0.429	*Escape*	0.856	0.144
Gold	0.585	0.415	*Black*	0.866	0.134
Poetry	0.587	0.413	*Persistence*	0.870	0.130
To seduce	0.588	0.412	*War*	0.891	0.109
Tree	0.592	0.408	*To betray*	0.893	0.107
Soul	0.593	0.407	*Immobile*	0.914	0.086
Elegance	0.595	0.405	*Red*	0.925	0.075

Table 2.11 lists the 18 words whose scores are best explained by common factors (words having the strongest *communities*) and the 18 words whose scores are less well explained by common factors (words having the highest specific variances)[62].

We find in the first column the words that are well explained by the set of the first axes of Semiometry, and in the second one, words that are poorly explained by these first axes.

[62] A rather similar result could be found by summing, for the first six axes, the squares of the coordinates of the words on the axes, which are also the correlations of words with these axes (Table A2.3 in Annex 2). These sums represent very roughly the *communities*, while their complements at 1 show the specificities. This is only an approximation because the real *communities* of factor analysis in common and specific factors are obtained through an iterative process (Appendix A1.6).

2.7 Conclusion

The word list and the survey protocol (the questionnaire and the wording of the basic questions) having been defined, the structure extracted from the data table appears stable in all the surveyed Western countries. This structure consists of six main ordered axes. It does not change from one sample to another, from one year to another, does not undergo changes (or undergoes minor changes) from one country to another - despite the risks taken in translating the questionnaire, if we except the case of China (Hong-Kong survey).

The semiometric structure is found and repeated within sub-populations (sex, age group), suggesting a fractal phenomenon (at this point, this is only an analogy).

This stability is also verified through standard statistical tests of re-sampling ("bootstrap") which confirm the previous more empirical experiments. A non-standard test ("bootstrap on variables") was a first attempt to assess the degree of dependence of the structure obtained with respect to the composition of the list of words.

The digital encoding chosen for the answers may be distorted using several mathematical functions without altering the structure, which frees us from the sensitivity of correlations with the coding, more or less conventional at the start.

Finally, the classic factor analysis, or analysis in common and specific factors, reproduces the same first dimensions (as does the logarithmic analysis discussed further in Chapter 3). We are no longer dealing with just stability, but are presuming that the semiometric structure is robust.

Chapters 4 and 5 will relate several experiments providing further answers to the question, much more complex and ambitious, of any possible intrinsic character of the structure.

CHAPTER 3

Scorings, willingness to answer, attitudes

In this chapter on methodology, we return to the first axis called *the involvement and scoring axis* (Chapter 1), which was excluded from the semiometric structure in previous chapters. Indeed, it is our responsibility to report back to the reader on the results and arguments that led to giving this axis a special status. The study of both scoring styles and the degree of involvement in the survey naturally leads to the poorly understood and often eluded problem of the relationship between the attitude toward a survey and, on the other hand, the content and quality of responses to that survey. The first axis of semiometric analysis, we have seen, is always stable and dominant[63].

This chapter will confirm the hypothesis that this first axis is one that can be described as *methodological*; in short, an axis which does not concern the content of the questionnaire and which is much less bound up than the following axes with the socio-demographic characteristics of the respondents.

The specialists in the field of multidimensional processing of survey data often encounter such axes, which are most of the time dominant in terms of explained variance. Sometimes called a *scoring axis* or an *axis of attitudes toward the survey*, and in a context quite different (especially in biometrics), a *size factor*, the first axis of semiometric analysis is all three of these at the same time.

The wealth and scope of the database available has allowed for a thorough study of this phenomenon, and even, in so doing, has allowed us to bring

[63] It always accounts for about 12% of the variance, while the second axis rarely exceeds 5%.

some contributions to the difficult and crucial problem of respondents' participation in a survey.

3.1 Scoring, participation, and size factor

3.1.1 The scoring effect

Certain scoring effects are known and identified during psycho-sociological and psychometric tests, opinion surveys and/or marketing ones. The first effect usually described is that of the scoring *level*. There may exist – or there exist – good and bad scorers. The second effect relates to the scoring *scale*. Some people use the whole scoring scale proposed (here: all the scores go from 1 to 7)[64], whereas others use only the central part of the interval (e.g. scores of 2 to 6, or only scores 3, 4, 5). The combination of these two effects produces idiosyncratic scoring styles leading to specialization in specific areas of the scale (e.g. scores 1, 6, 7).

3.1.2 The participation effects

This is a fundamental concept and a very disturbing one for statisticians, tantamount to a continuation of the "refusal to be interviewed" attitude, which is probably the biggest threat *vis-à-vis* the quality and sustainability of statistical information systems in general, and surveys in particular. Are we sure that the fact that someone accepts to fill in a questionnaire (such acceptance is in itself the source of bias, but about which no one knows very much) implies, on the part of all the respondents, the same adherence to the rules of the game and to the spirit of the survey?[65] This type of acceptance will not be independent of the *scoring* effects: a questionnaire full of non-responses or responses that systematically concern the average of a scale is ultimately closer to a refusal than to an acceptance. It could be even a refusal that does not admit itself as such, almost a non-response to the refusal.

---

[64] Remember that scores: 1, 2, 3, 4, 5, 6, 7, correspond, in the questionnaire, to scores: -3, -2, -1, 0, +1, +2, +3. The two encodings are equivalent for the correlation calculations and for all statistical processing performed.

[65] In fact, we are certain for the contrary, but we rarely have available, in the databases collected, items to prove it or analyze it. From this point of view, the different semiometric files available and the *general* nature of the questionnaire provide favourable circumstances for this methodological study.

3.1.3 The size *factor*

This is quite a different phenomenon, because it is linked to the very drafting of the questionnaire itself. However, in the case of semiometry, it is inextricably linked to the *scoring* effects.

Early studies on factor analysis in the early twentieth century have dealt with scores in tests of intelligence or academic tests. Consider this latter example: the first axis will often be the "general aptitude" factor (called "general intelligence or the g factor" by Spearman (1904), but this was and still is a disputable issue). In these kinds of applications, with few exceptions, all the scores are positively correlated with one another. The explanation being that there are, roughly, good students and bad students. The former had above average scores in virtually all subjects, the latter lower scores.

In a much simpler way, if we measure the length and diameter of eggs from hens, we will find that these two measures are correlated, because the main source of variability is the size of the egg: the first principal axis will distinguish between large eggs and small eggs, hence the name "size factor". The second principal axis, sometimes called the "shape factor" will instead show that with eggs of the same size, the eggs are more or less spherical. We will see that such a phenomenon is superimposed on the *scoring* effects in the case of Semiometry.

3.2 Creating and positioning control variables

Each respondent is initially characterized by 210 scores in the semiometric questionnaire and by their responses to socio-demographic questions (sex, age, etc.). From the 210 scores, we have created new technical variables.

Each individual will then be also characterized by:

- A new variable: the sum of the scores calculated from all 210 words (or equivalently, their average scoring, which will be denoted *Mean_score*)
- Seven new variables which are: the number (called: Score_1) of responses "1", the number (called: Score_2) of responses "2", (…), the number (called: Score_7) of responses "7" (the sum of these 7 variables is equal to 210, the total number of scores).

➢ Two new variables derived from the previous scores which are grouped median scores called: Score_345 = Score _3 + Score _4 + Score _5, and the high scores: Score_67 = Score_6 + Score_7.

Table 3.1: Basic statistics of control variables

	Mean	(%)	Minimum	Maximum
Score_1	11.43	(5.4%)	0	89
Score_2	09.89	(4.7%)	0	69
Score_3	15.51	(7.4%)	0	78
Score_4	40.05	(19.1%)	0	146
Score_5	41.67	(19.8%)	0	124
Score_6	46.83	(22.3%)	0	139
Score_7	44.62	(21.2%)	0	148
(Total)	(210.00)	(100%)		
Score_345	97.23	(46.3%)	0	188
Score_67	91.45	(43.5%)	0	156
Mean_score	4.95		2.39	6.21

The basic statistics of these new control variables were calculated on a representative sample of 5,527 respondents in France (see Table 3.1 for some statistics about these variables).

We can read, for example, on the first row of Table 3.1, that the number of scores "1" used by a respondent is on average 11.43 (it represents 5.4% of the 210 scores). Column Minimum shows that there is at least one individual who has never used a "1", since the minimum equals "0". Column Maximum shows that there is at least one respondent who used 89 times a "1", and no one else has used this score "1" more often.

The first three scores are less employed than scores 4 and 5, and those scores 4 and 5 are less used than score 6 (the most frequent) and score 7.

The last row of Table 3.1 is different in nature, since we are no longer counting scores per individual, but the averages of all the scores given by a respondent. We can see that the average score awarded by an individual is 4.95 (greater than 4, the arithmetic mean of the seven numbers 1, 2, 3, 4, 5, 6, 7) but this can vary from 2.39, for some questionnaires, up to 6.2[66].

[66] Do not confuse the average scoring for an individual (calculated on all the 210 scores for words given by this individual) with the average scorings for each of the 210 words (computed over all individuals of a sample), given in appendix A2.2. We learn for example in this appendix that *War*

Table 3.2 shows the correlations of these control variables with the first six semiometric axes, as calculated in chapters 1 and 2.

Table 3.2:Correlations of control variables with the semiometric axes

	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5	Axis 6
Score_1	-.08	-.19	.48	-.13	-.16	.06
Score_2	.31	-.07	.31	-.10	-.06	.20
Score_3	.46	-.06	.12	-.03	.07	.14
Score_4	.55	.17	-.01	.04	-.01	-.22
Score_5	.41	.11	-.20	-.00	.07	.02
Score_6	-.23	.11	-.15	.02	.02	.13
Score_7	-.77	-.20	-.01	.02	-.01	-.02
Score_345	.79	.18	-.09	.02	.06	-.11
Score_67	-.91	-.12	-.11	.03	.00	.06
Mean_score	-.92	-.05	-.36	.10	.07	-.03

At first glance, we can notice that the high correlations of all these variables (which are not directly related to the content of the questionnaire), mainly concern the first axis (and to a lesser extent, the third axis with which the variable, Score_1, has a correlation of 0.48).

The average control variable (remember that this is the average of 210 scores attributed by each individual) is strongly correlated (negatively, -0.92) to Axis 1, as well as the frequency of scores 6 and 7 (-0.91).

However, these scores, 6 and 7, are not opposed on this axis to 1 or 2, as might have been expected, but to the aggregate of scores 3, 4, and 5 (their correlation of 0.79 with the first axis is the strongest positive correlation).

Thus the first axis, closely linked to the frequencies of the highest scores (6 and 7), and contrasting them with the frequencies of the median scores (3, 4, 5), is both a *scoring* factor and a *size* factor.

The fact that individuals choose to attribute a score by only using the central part of the scale (scores 3, 4, 5) while others do not hesitate to use extreme scores (1 or 7) is itself a *scoring* factor. The fact that Axis 1 (and to a lesser extent Axis 3) is highly correlated (in absolute value) with the average score makes up the *size* factor. But the extreme scores are not involved in a balanced way, because the high scores are much more frequent than the low ones (scores 6 and 7 represent 43.5% of the scoring, although scores 1 and 2 represent only 10.1% of these scores). The

is the word most poorly scored on average (1.26) followed by *To betray* (1.42), while the word *Peace* is the highest scored on average (6.72), followed by *Tenderness* (6.67) .

correlations of the control variables with Axis 1 (and Axis 3) show that the *scoring* factor and the *size* factor are linked. We shall return in detail to these complex relationships in sections 3.5 and 3.6.

3.3 Scoring in four European countries

The control variables that overall characterize how scores are attributed were calculated for four countries (France, Spain, Great Britain, Germany). This paragraph shows how attributing a score is highly variable. The impact of these scoring styles on semiometric structures is however limited.

Table 3.3 gives an overview generalizing the first column of Table 3.1.

Table 3.3: Four scoring styles in four European countries.

Country: Score	Spain	France	Great Britain	Germany
Score_1	14.71	11.43	11.11	10.07
Score_2	9.67	9.89	11.87	11.21
Score_3	14.89	15.51	1993	22.49
Score_4	33.31	40.05	44.36	43.01
Score_5	39.72	41.67	46.46	50.04
Score_6	44.17	46.83	38.27	40.15
Score_7	53.32	44.62	37.70	32.40
Score_345	87.88	97.23	111.09	116.19
Score_67	97.61	01.45	75.88	72.43
Sample size	2 984	5 527	924	3.066
Mean_score	5.0	4.95	4.76	4.73

One reads on this table, for example, that 14.71 is the average number of score 1 assigned by a Spanish respondent, and that 32.4 is the average number of score 7 assigned by a German respondent.

The order of the columns of this table is not neutral: we have voluntarily placed the countries in order of decreasing average scores (see also Figure 3.1, which schematically shows the last line of Table 3.1). Spain is the country which gives on average the highest scores to words (5.0), while Germany gives the lowest ones (4.73).

We notice an opposition between the Southern and Northern European countries, an opposition often borne out in international surveys[67].

[67] These differences between average scores are statistically highly significant for all pairs of countries except for the couple United Kingdom - Germany (the size of the British sample does not allow us to conclude accurately). For France, for example, the standard deviation of average scores

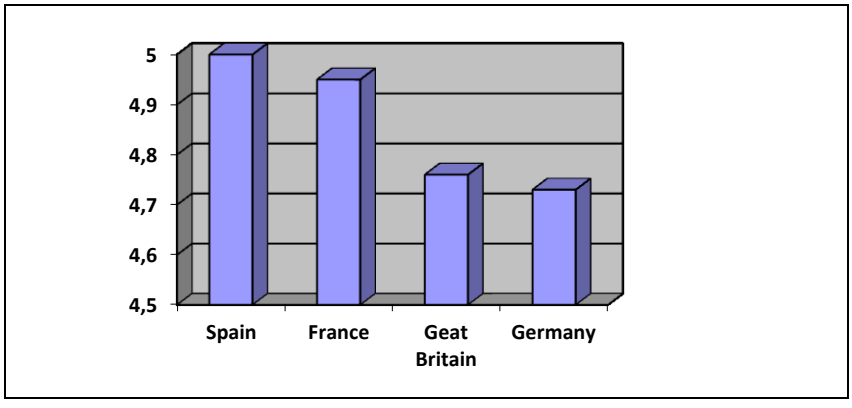


Figure 3.1: Average scores given to words in the four countries

- Uncommitted or reserved "scorer"?

Figure 3.2 illustrates the variety in the range of score frequencies. It is clear that, for example, the right bar of the histogram for each country (corresponding to score 7), decreases steadily from Spain to Germany, while average scores also decrease.

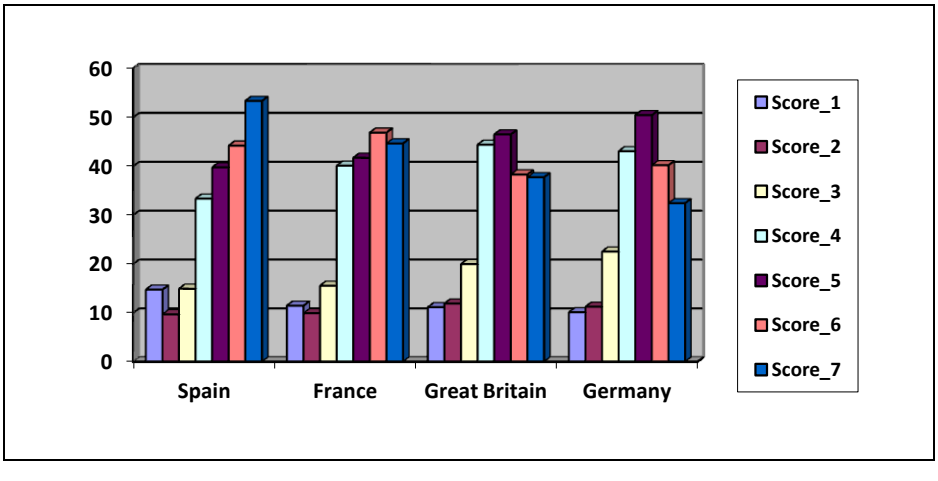


Figure 3.2: Summary of frequencies of the seven marks per country.

(this is, remember, the averages calculated for each individual on his/her 210 marks) is 0.38, the average of these mean scores being 4.95. In a sample of 5,527 respondents, the confidence interval ($p = 0.05$) of this mean is [4.94, 4.96].

Figure 3.3, by combining scores 3, 4 and 5, and scores 6 and 7, better highlights the cross-national variations of the *scoring* effect. Indeed, the countries of Northern Europe use intermediary scores much more readily; however, we cannot really decide if this is due to subtlety in the way of scoring, or caution, or mistrust or even lack of interest. In any case, we are dealing with major cultural differences, which erode the relevance of international comparisons in levels or averages within Europe itself, a subcontinent which nevertheless forms a relatively homogeneous group on a worldwide scale (age structure, level of education, standard of living, etc.).

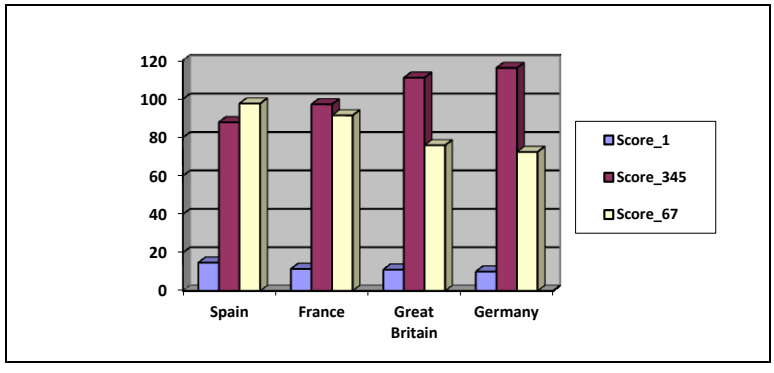


Figure 3.3: Extreme scores (1 and 6/7) and median scores (3/4/5)

But, perhaps surprisingly, these scoring effects cannot shake off the strong structural stability observed, dependent on the correlations between scores and much less dependent on the level of scores.

3.4 Typology derived from the sole control variables

To pursue further study of this methodological investigation, a "control analysis" will be performed on the French sample (the largest one). This will lead to a new ***principal components analysis*** of individuals. The individuals will no longer be characterized by 210 scores, but by the 7 control variables, *i.e* by the numbers of scores 1, 2, (...) 7 that they have given, without worrying about which words they attributed these scores to.

One of the first by-products of this analysis is the correlation matrix of the scores (Table 3.4). These correlations, computed on 5,527 individuals, are

all highly significant statistically (with two exceptions: the values 0.01 and -0.02)[68].

The strongest positive correlations concern, on the one hand, the number of scores "5" and the number of scores "3" (0.46: the people who use a lot of "5s" use a lot of "3s" as scores, these "scorers" are moderate). Moreover, they concern the number of "1s" and the number of "7s" (0.45: this link is all the more remarkable as score "1" is (from Table 3.1) four times less common than "7"; such scores are given by *extreme scorers*.)

The strongest negative correlation (-0.65) in absolute value concerns the following scores: "7" and "5". This reflects a negative association between a strong score and an average one; in other words, an opposition between *extreme "scorers"* and *moderate "scorers*".

Table 3.4: Correlation between the frequencies of each of the seven scores among respondents

	Score_1	Score_2	Score_3	Score_4	Score_5	Score_6	Score_7
Score_1	1.00						
Score_2	0.07	1.00					
Score_3	-0.22	0.31	1.00				
Score_4	-0.15	-0.21	-0.24	1.00			
Score_5	-0.49	-0.02	0.46	-0.11	1.00		
Score_6	-0.37	0.17	0.01	-0.41	0.10	1.00	
Score_7	0.45	-0.26	-0.45	-0.34	-0.65	-0.31	1.00

The first principal axis of this "control analysis", derived from the previous correlation matrix, is described in Table 3.5, which contains the most characteristic variables of the axis: the active variables (variables: Score_1 (...) Score_7 defined above) and the additional variables or illustrative variables [69] which make up composite scores (Score_345, Score_67) and the individual average of the scores. Moreover, this table contains the words recorded by the same 5,527 individuals.

The result is quite spectacular if one bears in mind the fact that individuals were characterized by a single numerical assessment of the scoring given

[68] For 5,527 individuals, any correlation greater (in absolute value) than 0.03, can be considered significant at a 0.05 threshold.

[69] See Appendix A1.9.3 for some hints about supplementary variables: in the case of principal components analysis, the coordinate of a supplementary variable on an axis is its correlation coefficient with the axis. It is calculated using all coordinates of the individuals on the axis.

(the seven control variables), regardless of the words that these scores were attributed to.

In terms of active variables, this first control axis compares the numbers of extreme scores ("7" and "1") with the number of average scores ("3", "4", "5") well. In terms of supplementary variables, the average score is located on the side of extreme scores, and the aggregate scores ("3", "4", "5") on the side of average scores. Most importantly, and this is the most striking result, we find nearly the same words as those characterizing the first semiometric axis.

Table 3.5: Analysis of the frequencies of seven marks: Description of the first axis of the "control analysis"

Warning: The words here are "supplementary" variables, projected afterwards on the axes

The most negatively correlated variables to the first axis		*The most positively correlated variables to the first axis*	
Score_5	-.80	*Score_7*	.85
Score_3	-.66	*Score_1*	.68
Score_345	-.64	*Score_67*	.56
Score_6	-.44	*Mean_score*	.42
Score_2	-.29	***Softness***	.39
To betray	-.17	***Honest***	.36
To break	-.16	***Courage***	.35
Fault	-.15	***Confidence***	.35
Disorder	-.14	***Dynamic***	.34
To punish	-.13	***Comfort***	.34
To attack	-.13	***Liveliness***	.34
Anguish	-.13	***Politeness***	.34
To criticize	-.13	***Respect***	.34
Danger	-.12	***To protect***	.34
Wall	-.11	***Reward***	.33
Doubt	-.10	***Tenderness***	.31
Emptiness	-.09	***Friendship***	.31

If one refers to the description of this first axis (chapter 1), we find in common, on the positive side, the first eleven words in the right column of Table 3.5. On the negative side, there are seven words in common with the left column of the same table.

In this sense, the first semiometric axis deserves being called the *methodological axis*. It is not necessary to have at one's disposal the 210 scores (that is to say, those extracted from Table 5527 x 210) to build it.

Quite independently from the content of the survey, the seven control variables (Table 5,527 x 7) are sufficient.

It now remains to understand why these particular words characterize the axis. "Scorers" we called "moderate" (respondents who use only the central part of the scale of scores) are characterized by words whose average score is low (in the left column of Table 3.5. See also the Table of mean scores of words in Appendix A2.2), and "anti-characterized" by words whose average score is high (in the right-hand column of Table 3.5). There is one reason which is simply mechanical: the variable "words" are centred relative to the average scoring for each word. As *moderate "scorers"* use the bottom of the scale of scores relatively little, they attribute scores which, once centred, will mechanically appear as high scores[70]. But if one centres the scores *per individual*, in other words, if for each respondent, we subtract the average scoring for each of the 210 marks (so that all respondents have the same average score) we always obtain the same characterization of the first axis of the control analysis from the words found in Table 3.5. So the mechanical effect mentioned is not sufficient to explain why those words characterize the two styles of scoring[71]. We will return below to this procedure of centring using the scores per individual (Section 3.7).

3.5 Relationship between the scores given to words, the way they are attributed, and the characteristics of individuals

The previous sections have shown that the average score for each individual (denoted *Mean_score*) was strongly linked to the first semiometric axis, and to a lesser extent, the third axis. We will now examine in more detail these

[70] Let's give another example of this phenomenon already mentioned in the previous chapter: *To betray* has an average rating of 1.42 for all individuals. Those who scored "2" or "3" to *To betray* thus lie above the overall average, and therefore appear to have relatively scored that word well. The word *Softness* itself has an average scoring of 6.37. So a score of 5 or 6, i.e. a good scoring for a moderate scorer, will be a bad score compared to the average.

[71] We cannot exclude the hypothesis that the style of scoring is linked to certain traits of respondents, with an effect on the choice of words superimposed with what we have designated as the *mechanical effect*. The semiometric questionnaire itself does not allow us to separate these effects, but the additional open-ended questionnaire (Section 5 of Chapter 4) has led to some results in favour of this hypothesis.

connections: first from the new categorical variable the four categories of which are called Mean1, Mean2, Mean3, Mean4 (a division into 4 classes of values of the average *Mean score*). After consulting the histogram of average scores, the five class limits selected for this distribution into four classes are: 1, 4.5, 4.95, 5.4, 7.

Table 3.6 gives the test values[72] of each of the four categories projected as supplementary elements on the first five semiometric axes. We see that the first axis is characterized by exceptional test values.

Table 3.6: Importance of individual mean scores

Average score per individual					
Semiometric axis in 4 classes	Test values				
	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5
Mean1 – low_mean	41.9	2.5	10.8	-3.0	-2.0
Mean2 – mean_2	28.8	1.3	14.8	-4.1	-2.7
Mean3 – mean_3	-29.3	-1.7	-5.0	1.4	.3
Mean4 – high_mean	-39.9	-1.9	-24.9	6.9	5.5

The third axis is also important, although its magnitude is not comparable. It is not independent of the other axes, but the connection is much weaker.

Figure 3.4, meanwhile, shows the plane of the first two semiometric axes, the behaviour of the four categories, denoted “Mean score low”, “Mean score = 2”, “Mean score = 3”, “Mean score high”, in relation to a series of categorical variables schematized by small circles to highlight their concentration along the vertical axis (cross-tabulation of age and sex, cross-tabulation of age and education, cross-tabulation of sex and education).

[72] Remember that the *test value* is a simple conversion of the coordinate on an axis into a quantity that, under the assumption of independence (random distribution of individuals concerned about the mean of the axis) behaves like a standard normal variable. In other words, if the category is not related to the axis (i.e. if it is composed of individuals randomly distributed on the axis), the test-value must be (approximately) between -2 and +2. This parameter makes it possible to assess quickly the statistical significance of the location of a category on an axis (*see appendix* A1.9.1).

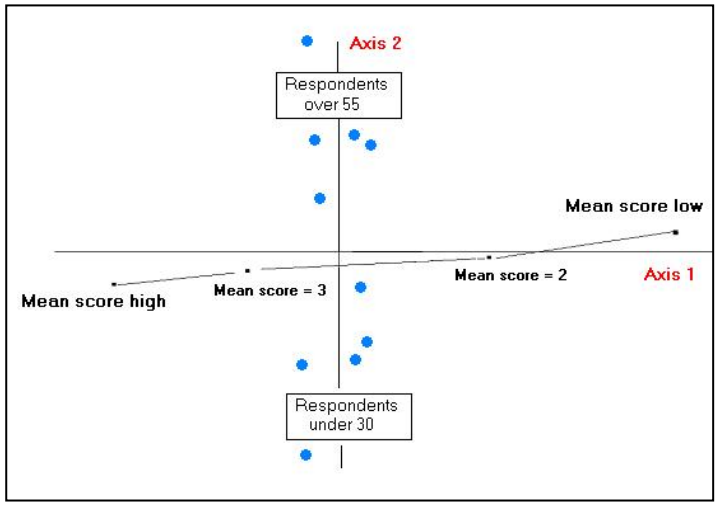


Figure 3.4: Position of the categories of the average scores in the plane (1, 2)

The small circular symbols describe the location of various supplementary categories involving age, level of education, sex)

It is clear that the categories of the average score characterize the first axis much more than any other socio-demographic variable, and that they characterize the second axis very little. In fact, we find the result mentioned in the description of Axis 2 (Chapter 1) according to which the second axis contrasts, for both males and females separately, the young respondents with older people.

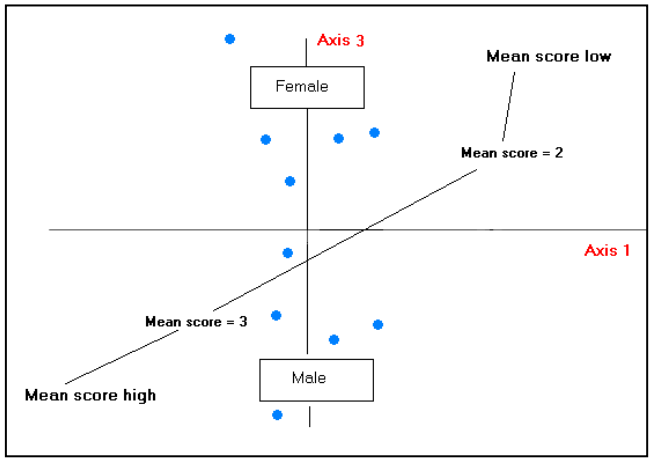


Figure 3.5: Position of the categories of the sum of the scores on the plane (1.3)

(The small circular symbols describe the location of several categories involving age, level of education and sex)

Figure 3.5 shows that contrary to the second semiometric axis, the third axis is related to the average score.

Although three times smaller in terms of explained variance than the first axis (4.2% against 12%), Axis 3 is the only axis, with the first one, to be clearly linked to that average score. Axis 3, (vertical in Figure 3.5) which compares, as we saw in Chapter 1, males (especially educated males) to females (particularly poorly educated females), is still much more linked to socio-demographic characteristics than Axis 1. It was shown however in Chapter 2 that this axis remained with sub-samples consisting only of men or women.

It is clear that the *size factor* is not Axis 1 set apart, but a direction, almost diagonal (a little closer to Axis 1, in fact, because the scale on the vertical axis is expanded in this figure) on the plane (1, 3). This diagonal direction is represented diagrammatically by the trajectory of categories ranging from medium low to medium high.

Now let us project onto the same plane (1/3)[73] the words (an extract, for greater readability) and the control variables (the frequencies of each of the seven scores for each individual). Figure 3.6, thus obtained, provides an overview of the wealth and complexity of the connections between *scores* and *how to attribute scores*, in other words between the *results of a survey* and the *attitude toward this survey*[74].

The position of words on the plane (1/3), highly off-centre with respect to the origin, is typical of the existence of a *size factor*. Respondents whose midpoint is the origin of the axes thus have very different attitudes with respect to all the scores: those at the top right give low scores to all the words while those in the bottom left give high scores.

But the trajectory of the frequencies of scores shows that the average score conceals very different components. With a simple *size factor*, this trajectory should have been nearly a straight line, going from low scores in the upper right to the high scores at the bottom left. This is not the case because, on the *size* effect, an *attitude* effect is superimposed or a

---

[73] This plane is derived, remember, from a classic semiometric analysis taken from a table with 5,527 rows and 210 columns,

[74] Figure 3.6 is only an outline. The dots representing scores 1 and 7 are outside the frame in the direction indicated by the arrows (the precise coordinates of score points are given by the columns *Axis 1* and *Axis 3* of Table 3.4).

participation effect: some respondents use all the possible scores; others stick to the average ones, and others prefer to use the extreme scores.

1) Respondents located in the upper left quadrant (Quadrant I) are "real good scorers" who attribute a "7" to a word which is very pleasant for them (Cadeau *[Present],* Récompense *[Reward],* Guérir *[To recover],* Tendresse *[Tenderness],* Respect *[Respect]*) and a "1" otherwise for unpleasant words (Danger *[Danger],* Vide *[Emptiness],* Faute *[Fault],* Punir *[To punish]*). The semiometric words bring those of "Attachment" and the acceptance of the whole scoring scale together.

This shows that there is no independence between the questionnaire content and the attitude toward the questionnaire.

2) Respondents located in the lower left quadrant (Quadrant III) are "good systematic scorers"; their scores contain fewer scores "1s" and "2s": they do not clearly reject the words.

3) Respondents located in the upper right quadrant (no words present) (Quadrant II) are "bad systematic scorers". They use the bottom of the scoring scale, but also reject the words (*Danger, Emptiness, Fault, To punish*). The *size factor* in the traditional sense contrasts Quadrant II and Quadrant III.

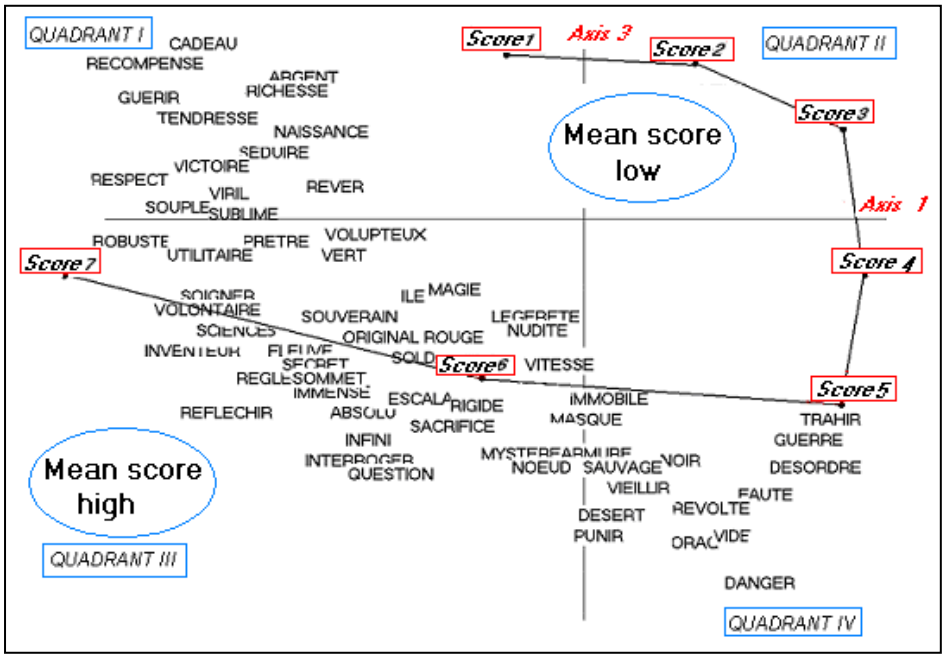


Figure 3.6: Location of words (*in French*), **frequencies of scores in the plane spanned by axes 1 and 3**

4) Finally, respondents located in the lower right quadrant (quadrant IV) use mainly the central part of the grading scale. They use relatively few scores 1s and 7s. The semiometric words can now bring the words belonging to "Detachment" and the refusal of any rating scale nearer.

This plane (1/3) may be called the *scoring plane*. Indeed, we find along the two dimensions of the plane, the significant outlying positions of the control variables (score frequency, average scores).

In Chapter 2, we noticed that the order of axes for axes 2 and 3 were inverted; the axis conventionally called "Attachment / Detachment" occupying the third position in France, Italy and Spain; and the second position in the countries of Northern Europe (Germany, Great Britain, Finland, Norway) and the USA.

This axis is the only one concerned by the scoring style (with the first axis that we have excluded from the semiometric structure itself).

We can assume that its change in rank is specifically related to cultural "differences in scoring" identified and demonstrated in this chapter.

3.6 The *rating planes* in Spain, UK, and Germany

We shall check that the structural features observed in the case of France are still valid for the three additional countries that we selected[75].

3.6.1 Scoring effects in Spain

Figure 3.7 is analogous to Figure 3.6 for the Spanish semiometric collection based on 2,984 individuals. Only one word out of 4 (randomly selected) appears on the chart.

As before, the score frequencies are included with the status of any additional elements; moreover, the active variables of the PCA are the 210 words.

[75] Similar results relating to all the countries for which semiometric data are available are not published here.

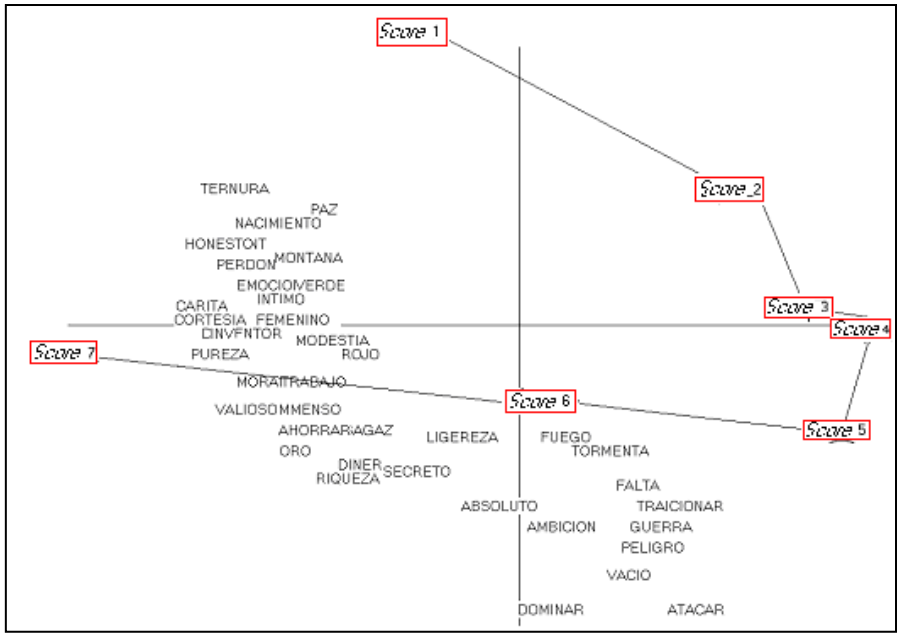


Figure 3.7: Diagram of the plane (1, 3) (Spain)

As for France, the plane generated by axes 1 and 3 brings together the *scoring effect*, with the same interpretations[76].

The words belonging to "Attachment" (Ternura: *Tenderness,* Nacimento: *Birth,* Paz: *Peace*…) on the side of the higher frequency of extreme scores; the words belonging to "Detachment" (Atacar: *To attack,* Dominar: *To master,* Vacio: *Emptiness,* Guerra: *War,* Traicionar: *To betray*…) (nearer the high frequency of average grades {scores 4s and 5s}).

3.6.2 Rating effects in Britain

The diagram in Figure 3.8 is on a much smaller sample (924 individuals) and must therefore be interpreted more carefully.

[76] See the multilingual table A2.1 in the appendix of words for a translation.

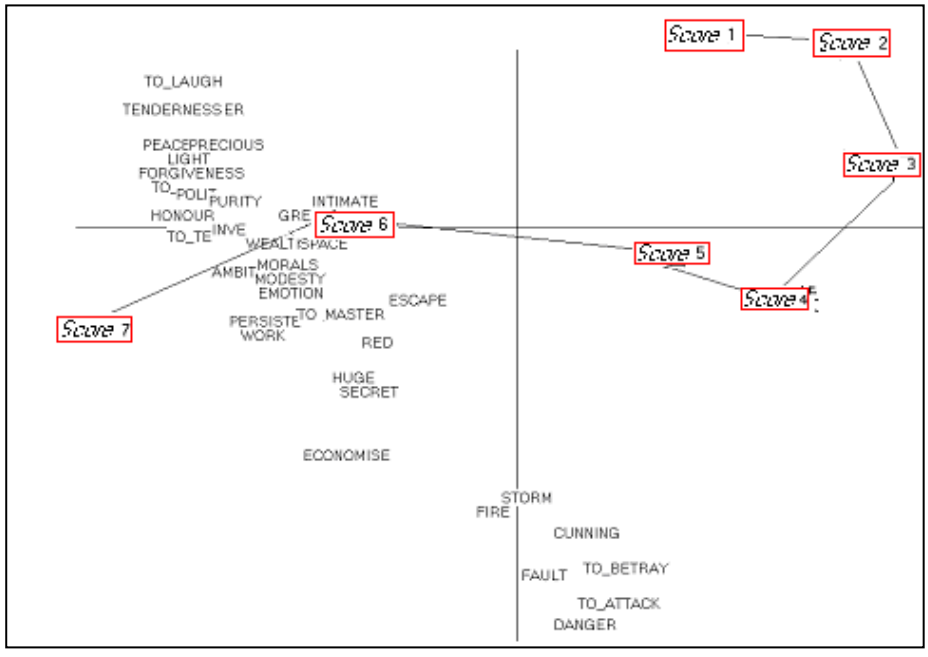


Figure 3.8: Schematic diagram of the plane (1, 2) (Great Britain)

Due to inverted axes reported several times (notably in Section 2.1.2 of Chapter 2) we now no longer show the plane generated by axes 1 and 3, as was the case for France and Spain, but the plane generated by axes 1 and 2.

The positions of the word-points in this plane are then quite similar to those observed in the cases of France and Spain for the planes defined by axes 1 and 3. The points representing the frequency of scores (positioned, remember, as additional elements) have a slightly different configuration. Scores ("7", "6", "5", "4", "3") range along the first axis, while for France and Spain, this was true only for marks ("7", "6", "5").

3.6.3 Rating effects in Germany

Figure 3.9 shows how the words and the frequencies of scores are structured in the plane (1/2) of Germany. Here the survey was conducted on a sample of 3,066 respondents.

The bend in the curve of score frequency is around "4", when it was "5" for France and Spain. Nevertheless, the general organization remains the same in the four countries: a concentration of *scoring effects* in a two-

dimensional sub-space, a non-linear behaviour of the scores whose path curves so as to contrast the extreme scores with the average scores.

We find the presence of the words of "Attachment" in people using the whole scoring scale; the words belonging to "Detachment" on the side of the intermediate scores[77].

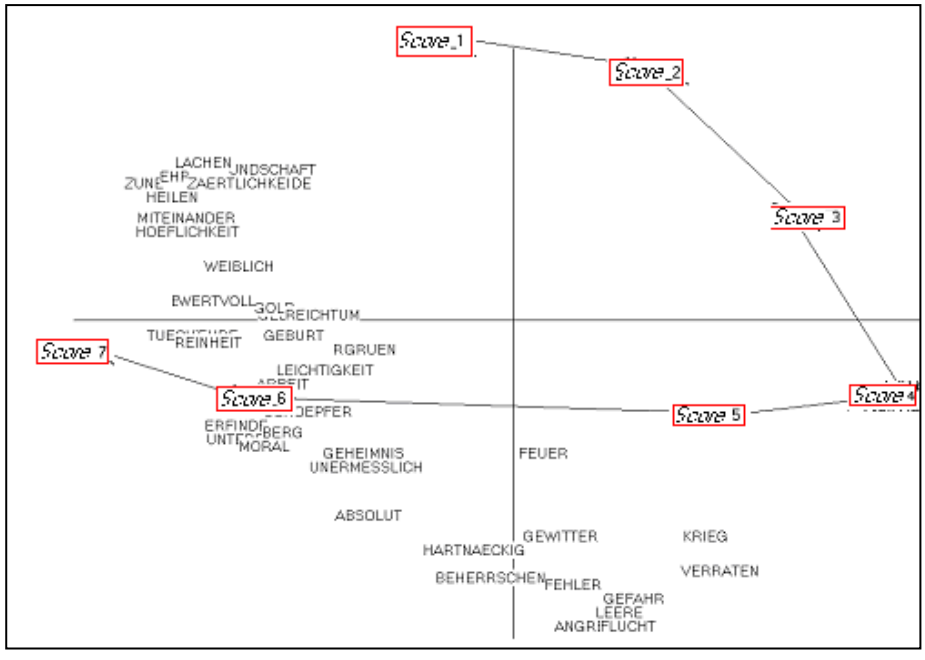


Figure 3.9: Diagram of the plane (1, 2) (Germany)

3.7 Correction of scores by double centring

In some applications, the *size factor* is corrected by centring for each respondent. To eliminate the differences between good scorers and bad ones, we proceed in this way so as to carry out a centring per individual in addition to the usual centring i.e. the centring of variables[78]: one subtracts

[77] See the multilingual table A2.1 in the appendix of words for a translation.

[78] Similar corrections are made in marketing surveys, but also in some scoring of exam or competitive exam papers, when many copies are distributed at random between scorers, and we wish to eliminate a potential *corrective* effect on scores, bringing the scores of each scorer to the same average common to all the scorers.

from each of the 210 scores given by a respondent the average of those 210 scores. The corrected scores for all the individuals then have a zero mean[79].

We saw above (last line of Table 3.1) that the mean scores of individuals ranged from 2.39 (for the *lowest scorers* of the sample) to 6.21 (for the *highest scorers*). After the transformation of centring "per individual", all the individuals have the same average score.

This transformation will profoundly affect the first principal axis (in terms of variances) because this axis was mainly related to the average score of the respondents (Table 3.2 has indeed shown that the absolute value of the correlation between the first axis and the average score is 0.92).

Table 3.7: Parameters of both standard and *centred per individual* analyses
(French sample)

	Eigenvalue (*Standard analysis*)	Eigenvalue (*Centring per individual*)		Percentage (*Standard analysis*)	Percentage (*Centring per individual*)
Axe 1	26.64	13.33		12.69	6.35
Axe 2	10.24	11.21		4.88	5.34
Axe 3	8.79	8.05		4.19	3.83
Axe 4	7.05	5.10		3.36	2.29

We can actually see in table 3.7 that the new first axis is twice as small in terms of variance. It is no longer a strong dominant axis.

In short, this transformation will change the space spanned by the first axes in the following way: the new first axis is a fusion of former axes 1 and 3. It remains close to the former first axis, while being strongly correlated to the old third axis (Attachment / Detachment). It also appears that the other axes are roughly preserved, but shifted (the new third axis being the former fourth axis, etc.), although the order has been somewhat inverted.

[79] This is perhaps not intuitive, but centring on columns (to ensure that all words have the same average score: zero, per individual) after rows centring (to ensure that all respondents have the same average score: zero) does not remove the effect of rows centring. In other words, we get a table immediately doubly centred.

Figure 3.10 shows the new plane (1/2) where the major changes are to be found. The frequencies of use of the scores are projected onto this plane; and are then boxed in, in two different ways: the words belonging to "Attachment" (top left of the graph) and "Detachment" (bottom right of the graph).

The new Axis 1, after centring per individual (the horizontal axis of Figure 3.10), is correlated[80] with the former Axis 1 (0.92) and the former Axis 3 (0.78).

The new Axis 2 is strongly correlated with the former Axis 2 (0.88) (axis "Duty / Pleasure") with now a slight association with the former Axis 3 (0.51), a detectable association in Figure 3.10 from the position of the words in boxes.

The polygonal trajectory of the scoring frequencies shows that centring per individual obviously does not solve the problem of the contrast between extreme scorers (on the "Attachment" side of the plane) and average scorers (on the "Detachment" side).

[80] These are correlations calculated from the coordinates of 210 words.

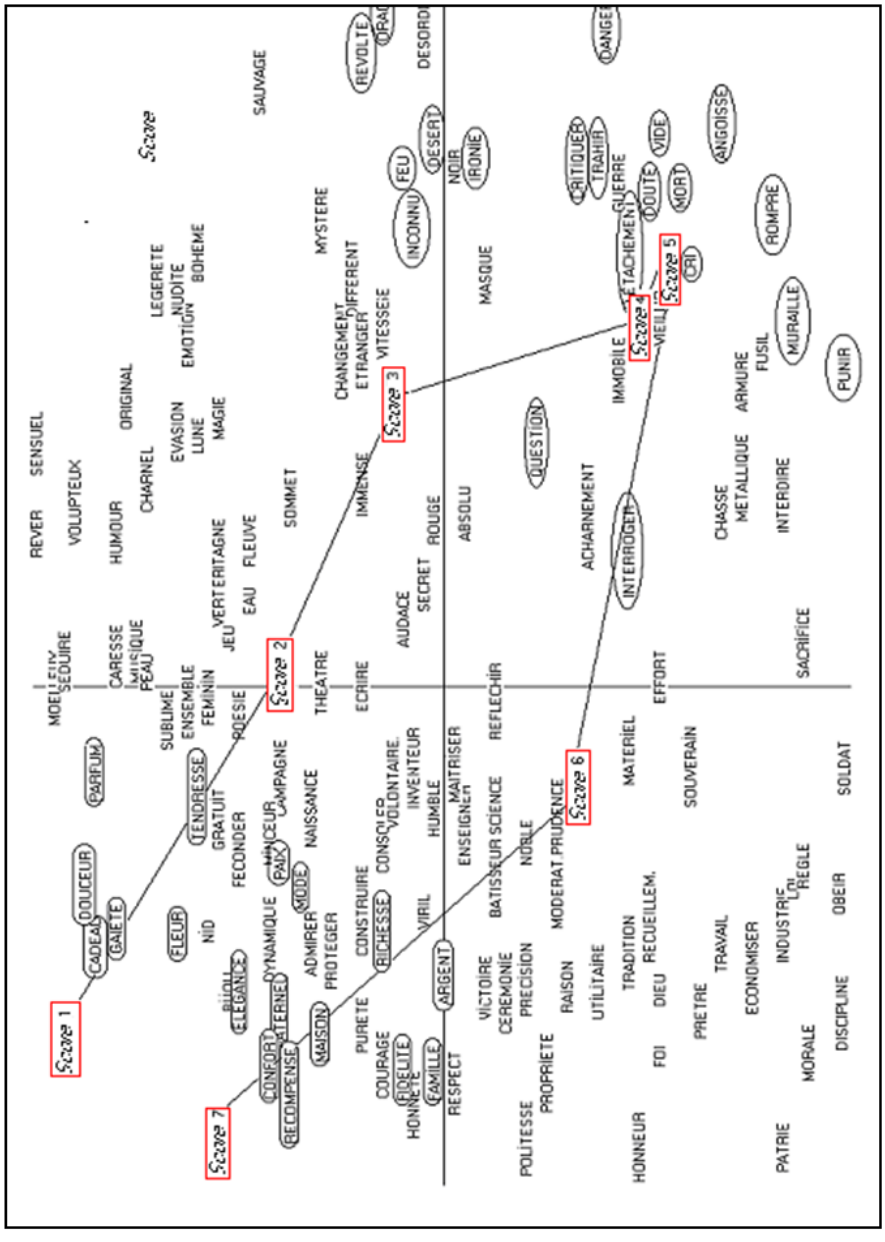


Figure 3.10 : Map (1.2) - double-centring *(French sample of 5,727 respondents)*

Should we have centred** a priori **the scores given by each individual?

We chose not to perform any transformation beforehand on the scores (except for the standardization of variables inherent to the technique of principal components analysis on correlation matrices) for several reasons, which we can now explain in the light of the experiments presented in this section.

• In all the surveys and in all the countries studied, the level of rating (average score per individual) and the scoring structure (the frequency of each of the seven scores) form a complex pattern that is not reducible to a preliminary non-linear transformation.

• Centring per individual eliminates the *size* factor (mean individual scores) but not the *scoring* effect (the scoring specialization in the central part of the scale).

• The analysis of untransformed data shows the complexity of the relationship between the attitude toward the questionnaire and the content of the survey.

• The option of analysing the raw data, which consists in considering both the attitude toward the survey and the scoring style as an integral part of the phenomenon under study, without prior mechanical simplification, seems heuristically richer, more scientifically ambitious. This option belongs to an exploratory perspective consisting in "not reducing *a priori* the scope of the observable data", thus not jeopardizing future research.

An interesting alternative: logarithmic analysis

Among the various attempts at recoding and processing data, one technique has produced interesting results: logarithmic analysis (see Appendix A1.5).

This is a transformation that leads directly to the various semiometric axes such as they have been studied in previous chapters, but the first axis has been taken out. As in the case of double centring, Axis 3 appears first, followed by Axes (2, 4, 5, 6). But these axes are, so to speak, very pure: it is exactly the axis "Attachment / Detachment" which appears first, then the axis "Duty / Pleasure", etc.. (The correlation coefficients with classic semiometric axes are above 0.95).

Nevertheless, the rules of interpretation of the results are less well known and less wide spread than those of the principal components analysis, and it did not seem to us appropriate to use this method to present the semiometric axes. For the moment it could be a possible research avenue[81].

3.8 Evolution of scoring in France from 1990 to 2002

We have seen how scoring varies from country to country. We shall now see that the way to score also varies over time. The study will focus on the French sample (France being the only country for which longitudinal data series are available).

The result is quite spectacular, but as we have done so far, we will endeavour to separate the statistical facts that we highlight, supported by appropriate validation procedures, and their interpretation, about which we are only making proposals. Those proposals are open to contributions and criticisms from other interdisciplinary teams.

The analytical work focuses on six surveys of comparable numbers of respondents for the following periods: 1990, 1996, 1998, 1999, 2001, 2002, totalling 15,684 individuals.

Semiometric structure (the first six axes, and the order of appearance of these axes, viz. their relative importance in terms of variance) has been stable over the past six years, but this stability masks a clear evolution in the way of scoring: the people interviewed use fewer and fewer extreme scores and therefore more and more median ones.

Figure 3.11 illustrates this evolution, especially between the 90-96 period, contrasting it with the past four years. Another feature of this decline in participation (or commitment) of respondents toward the survey is the growing importance of the first axis, which we termed the *methodological axis*, an axis opposing the extreme responses to the median ones.

Figure 3.12 shows the evolution of the first eigenvalue (the variance on the first axis) during this period. However, the frequencies of the scores and the importance of the first axis only reflect a small portion of the changes observed from 1990 to 2002.

[81] Logarithmic transformation, frequently used in economics, has classically the effect of making the distributions more symmetrical and the relationships linear. Intuitively, it is not too surprising to see it operate successfully in a context which we have described in previous chapters as non-linear and non-symmetrical.

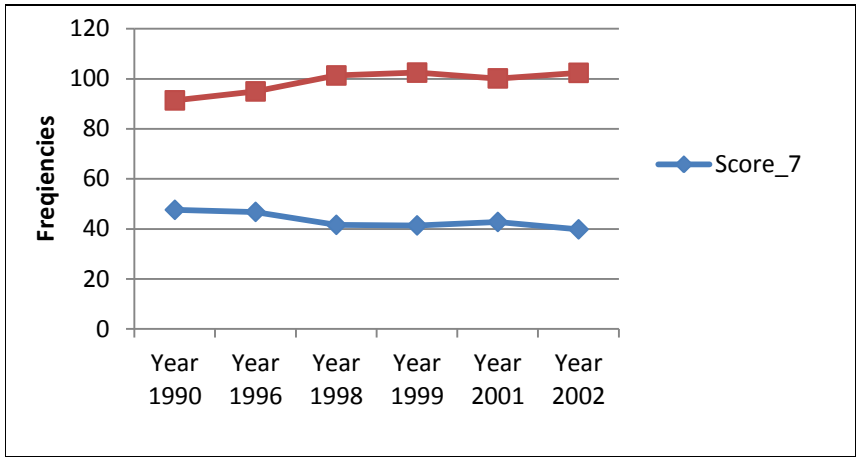


Figure 3.11: Evolution of the frequencies of median scores and 7s.
The propensity to use moderate scores is increasing.

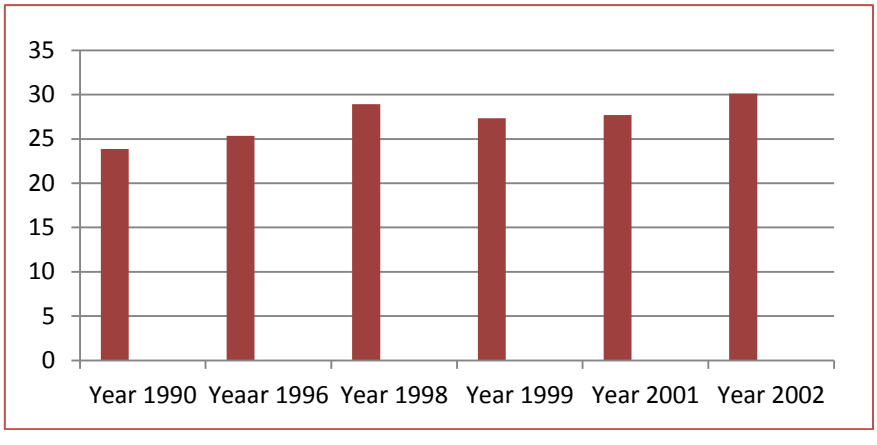


Figure 3.12: Evolution of the largest eigenvalue: the first axis is becoming increasingly important over time.

In fact, the year-points occupy very typical positions[82] in the plane of Axes 1 and 3, a plane which has been studied in detail in previous sections of this chapter.

These positions were plotted on Figure 3.13 which also includes, in dotted lines, the "parabola of the score frequencies" shown previously in

[82] These positions are projections of the six categories of the dummy variable "year" on the semiometric axis [1,3] according to the technique of supplementary variables (Appendix A1.9.3).

Figures 3.6 to 3.10, all assigned to those planes (1.3). Figure 3.13 also contains a shaded ellipse representing approximately the area occupied by the words. In the upper part of this ellipse are the words: *Elite, Creator, Faith, Sacred, Effective, Robust, Moderation, Reward, Present, Jewel…* These words are characteristic of the early years of the period[83].

You can find among the above words, those which are characteristic of the good "scorers", identified by the first axis, and those words belonging to "Attachment" on the third axis.

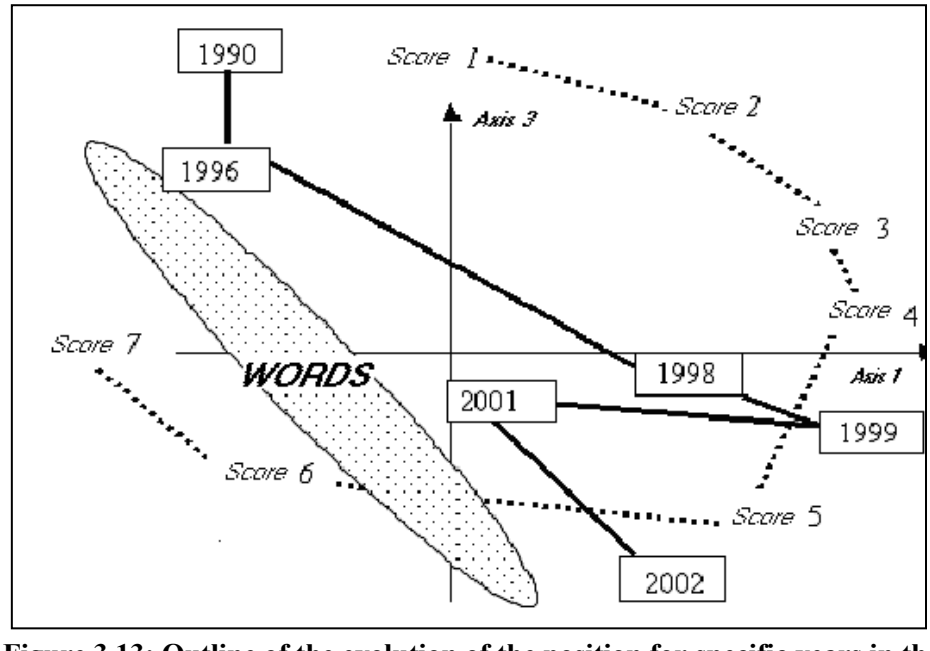


Figure 3.13: Outline of the evolution of the position for specific years in the plane of the axes (1.3). The gray ellipse schematically represents the area occupied by the words.

In the lower part of the ellipse are the words: *Fault, Rule, To punish, Mask, Desert, Danger, Anguish, Disorder, To betray*. These words characterize the last years of the period considered.

[83] The proximity of the upper part of the ellipse with points representing the early years in Figure 3.13 suggests an association between the corresponding words and those early years. One then checks, regardless of the principal components analysis, that the average scores of those words for 1990 and 1996 are significantly higher than their average scores over the entire period (test-values > 5).

These are, for some, the "non-participation" words, one of the poles of the first axis; and for others, the words relating to "Detachment", one of the poles of the third axis.

Despite the somewhat unusual position of 2001, the trajectory of years is strongly oriented toward the pole "Detachment" and simultaneously - and perhaps correspondingly - a trend consisting in scoring in a more central and less pronounced way.

In this sense, France is moving closer to the two countries studied in Northern Europe (Great Britain and Germany), whose predilection, we have seen, goes for the median scores (see Table 3.3 and Figure 3.3). The frequency of scores (3, 4, 5) varies in France from 91 to 102, without however reaching the 116 of Germany, or even the 111 of Britain. The average number of scores "7s" attributed decreases from 47 to 39 between 1990 and 2002, but remains higher than 37 for Great Britain and 32 for Germany. On the other hand, the order of the axes does not vary over time, and we do not observe the interchange of Axes 2 and 3, which characterizes both the countries of Northern Europe and the USA for which data are available.

It is quite difficult to distinguish what would be a change in the way of scoring (in the sense of a growing disengagement manifested by fewer and fewer differentiated scores) and a more general socio-cultural evolution toward disengagement and the values of "Detachment". This latter development may anyway affect the attitudes toward surveys in general and those related to semiometric surveys in particular.

What is interesting about the conceptual and technical framework within which we are working is to present the statistical facts in all their complexity, before carrying out any modelling, and thus making the basic material readable and open to various interpretations. Remember that we are dealing with the opposition on the first axis reflecting the contrast between the extreme scores and average ones. It appears that the periods are significantly more (i.e. significantly, in a statistical sense) separated by the third axis[84]. There is a strong presumption in favour of a structural shift towards the values of "Detachment" that would not be a mere artefact based on the way to score, but which could itself be a component of a broader trend.

[84] The test values of positions for 1990, 1996, 2002 are respectively 9.1, 7.7 and -10.6 on the third axis (the vertical axis of Figure 3.13) whereas they are worth -5.1, -5.1, and 2.5 on the first axis (the horizontal axis of Figure 3.13) ..

3.9 Conclusions

As was stated from Chapter 2 on, the semiometric structures (word configurations) are stable. Statistical material and the experiments of this chapter have shown that this stability persists despite scoring styles that are quite different from one country to another.

By showing that the first axis of analysis could be obtained simply from the frequency of use of each of the seven scores (regardless of the words that are assigned these scores), the "methodological" character of this axis, which we have up until now sidelined, keeping it away from the semiometric structure, has been established: it mainly concerns the form and not the content of questionnaires and surveys.

The *average scores* are much more affected by the *scoring styles* than the *correlations between scores*. The order of the axes is sensitive to encoding the data, whereas the axes themselves are much more robust.

The interchanged axes, observed between the countries of Southern Europe (France and Spain in this chapter) on the one hand, and the countries of Northern Europe (Germany and UK in this chapter, and USA, not dealt with in this chapter) on the other, may be due to the differences in scoring styles, or deeper structural differences[85].

A key lesson from this chapter is that the attitude toward a questionnaire is part of the information to be analyzed and interpreted; that this attitude leads to stable patterns, which can hardly be ruled out without mutilating the basic information.

In most sample surveys, the questionnaires do not contain evidence allowing us to characterize the attitudes related to the actual conditions of gathering information and to the adherence of the respondent to the issues relating to the questionnaire.It is somewhat to the merit of the semiometric questionnaire to allow this characterization. With this tool, we can study not only what the answers say, but also, to some extent, what is meant by the word "answer".

---

[85] The statistical fact is described by the following sentence: *the axis "Duty / Pleasure" is less important in terms of variance explained, in northern Europe than in Southern Europe*. One possible interpretation perspective could be, for example, the following: this axis contrasts the younger people to older people. One can imagine that, for historical reasons, cultural (and even climate, ethnic ones, etc.) the opposition between the generations is less pronounced in northern Europe than in the South. Evidently, several other explanations are possible.

CHAPTER 4

The spontaneous choice of words

This chapter attempts to answer the following question: can we achieve a kind of *open Semiometry* or *spontaneous Semiometry* by directly asking respondents to mention the words that seem pleasant or unpleasant spontaneously?

The experiment consisted in including the two following open questions in a sample survey of the general population:

"No one can explain why, perhaps because of what they evoke, we find some words pleasant, others unpleasant.

1. As far as you are concerned personally, what are the words that you find most pleasant? (Cite as many words as possible).

2. What are the words that you find most unpleasant? (Cite as many words as possible)."

The principle behind this type of questioning is not new. A famous precedent was given by the writer Albert Camus, who had been asked what his ten favourite words were[86].

The adjectives *pleasant* and *unpleasant* have been used deliberately here in the question wording to remain as close as possible to the spirit and wording of the semiometric questionnaire (but a similar survey on

[86] Albert Camus, *Notebooks*. His answer was: *the world, pain, earth, mother, men, desert, honour, misery, summer, sea.*

preferences would certainly also be just as interesting). If the exercise is not new in principle, its realization in the case of a sample survey with a sample supposed to represent the general population was, to our knowledge, a new one.

4.1 Setting up the experiment

We have at our disposal the questionnaires of 1191 respondents. The corresponding responses have produced an "artificial text" 41,547 words long (in fact, word *occurrences*, or: *tokens*) from 7,170 distinct words (or: *types*) cited. Many words have been cited only once, and therefore play no role in calculating the distances between respondents[87].

Among the 7,170 distinct words, there are 1,466 or 20.4%, which are cited four times or more. If we restrict ourselves to the artificial text composed of these 1,466 words appearing at least four times, we get a text of 33,950 occurrences, which makes up 82% of the initial corpus (as is the case for most texts, the skewed distribution of the frequency of use of words can allow us to reconstruct a significant part of the text with a limited number of words)[88].

Table 4.1 shows, for example, the answers of the first three survey respondents (words preceded by a *sharp* symbol "#" are cited as unpleasant words)[89]. We find, indeed, among the words quoted, words already encountered in the semiometric questionnaire, but there are obviously many "consensual" words which were voluntarily eliminated when constructing the list of 210 words.

---

[87] The survey was conducted in France by TNS-Sofres in 1995. We are dealing here with translations of the words selected by the respondents.

[88] For usual texts, the distribution of frequencies of words is often referred to as "Zipf law" (in fact, it is a Pareto distribution). It is interesting to observe a similar asymmetric distribution in the present case of artificial sequences of words: We are not dealing with real texts.

[89] Some words, like *fire, money, cigarettes, rain, order*, may be included (for different individuals) with or without the sharp, viz. be cited as unpleasant by some, and as pleasant by others. It was therefore necessary to provide special coding to one of the word families. We chose to attach a particular symbol (#) to the words considered unpleasant.

Table 4.1: Examples of free responses

Respondent 1:

sweet, cute, wonderful, charm, fullness, child, philosophy, beach, garden, animals, finesse, hot, sun, silk, nature, journey, joy, kindness, love, freedom, friend, caress, peaceful, natural, serenity, feeling, love, music, holiday, happiness.

#hate, #evil, #villain, #thirsty, #vomit, #casual, #nerve, #snake, #skeleton, #scam, #fat, #bulging, # reason, #cherub, #vomit, #solitude, #hemoglobin, #hypocrisy, #plump, #arrogant, #enemy, #sorrow, #sink, #incrustation, #sequester, #vicious, #dirt.

Respondent 2:

spring, sunshine, joy, health, happiness, child, holiday, travel, money, home, ceremony, flower, garden, restaurant, promenade, friend, beauty, company, tinker, knit, leisure, magazine, celebrates, discovery, visit, sea, mountain, destiny, reading.

#cold, #winter, #Snow, #Ice, #sect, #pollution, #disease, #drug, #solitude, #idleness, #ugly, #war, #dead, #suffering, #poverty, #misery, #waiting, #corruption.

Respondent 3:

flower, joy, kindness, love, tender, pretty, gay, friendly, smiling, caress, forest, mountain, nature, beauty, chocolate, cherry, bread, wine, woman, bed, sleep, bread, wife, child, life, travel, family, friend.

#kill, #evil, #hypocritical, #drama, #rape, #selfish.

Table 4.2 lists the words appearing over 63 times[90]. The four words most frequently mentioned are not in the semiometric questionnaire: three words mentioned as pleasant (*love, sun, vacation*), and a word cited as unpleasant (*#disease*).

Then the following words appear: *#war, #death, friendship, children, family, flower*, all six (*Childhood*[91] for *children*) are included in the semiometric questionnaire. The following three words: *happiness, travel*, and *sea* are absent from it (we do find *Water, Ocean, River* in the semiometric questionnaire, but not *sea*). There are obviously a lot of redundancy or implication in the words quoted spontaneously such as

[90] This threshold, which is still very high, was here chosen for reasons of space. The table is available in its entirety from the authors.

[91] Remember that words from the semiometric questionnaire are written, by convention, in italics with uppercase.

(*health, #disease, #cancer, #AIDS*) or (*friendship, friend*), (*#dead, #deaths*), *(#theft, #thief*), (*#crime, #murder*), (*love, to love*)[92].

We also noticed that nouns make up the overwhelming majority of words quoted, verbs and adjectives appear in this list initially in ambiguous forms; *laugh* (153) and *smile* (145), are probably cited most often as nouns and not verbs, given their context.

We shall see that the parts of speech often appear as sequences in a single answer; in the same way as *friend(ly)* (122), is more often a noun than an adjective. The first three adjectives cited would be *#black (106), blue (94), handsome (69)* (they are also nouns in French), and the first verbs cited are *eat (83), sleep (64), love (63)*.

We have noticed that there are also more current themes related to the news: (*#pedophilia, #taxes (95), #drugs (133), #attack (59), #sect (77)*), words which had been excluded *a priori* from the semiometric questionnaire.

One of the findings that emerges when we consider the entire corpus of the original responses, not only the most frequent words, is the total absence of spontaneous quotes of some of the words belonging to the closed semiometric questionnaire.

For this sample of 1191 respondents, 173 words (out of the 210 of the semiometric questionnaire) are spontaneously cited as *pleasant* or *unpleasant*. If we retain only the first 600 respondents, only half the words of the semiometric questionnaire were spontaneously mentioned by interviewees (some of them may, however, appear over 200 times - this is the case for: *flower, war, death, friendship*)[93].

---

[92] The favorite words of Albert Camus appear in the following order of frequency: *sea (286), #pain (104), #misery (99), summer (91), mother (22), [#mother (1)], honour (16), land (14), desert (11) [#Desert (6)], man (not men) (6) [man (2)], world (1) [#world (1)*] . This rapprochement with the words of Camus is admittedly anecdotal, but it nevertheless has the merit of bringing out the true nature of our questionnaire: it is not about issues of concern, or concepts, or keywords referring to fundamental problems, but merely sensations, either pleasant or unpleasant.

[93] Of course, if we increased the sample size, we would collect more and more distinct words, and we would eventually get the list of words in the questionnaire. However, as is the case with any corpus of text, the number of distinct words of the collection of responses would increase much more slowly than the total number of their occurrences.

Table 4.2: Words most frequently mentioned spontaneously

Words	*frequencies*	*Words*	*frequencies*	*Words*	*frequencies*
love	731	#rain	135	forest	88
sun	628	flowers	135	#noise	87
holiday	498	#drug	133	#prison	86
#disease	475	#rape	131	animals	85
#war	439	spring	129	to_eat	83
#death	421	kindness	129	#fear	83
friendship	379	garden	124	book	82
child	366	life	123	#policy	82
family	358	friend	122	#jealousy	82
flower	328	walk	121	reading	80
happiness	305	#calamity	119	water	80
travel	292	heat	116	hello	80
sea	286	birth	115	tolerance	78
music	277	#AIDS	113	#grief	78
#unemployment	265	#solitude	107	#sect	77
#accident	263	#black	106	heaven	76
joy	258	quiet	104	#pedophilia	75
nature	209	#pain	104	#death	75
#violence	191	children	102	mom	72
health	189	pleasure	101	light	71
affection	184	#poverty	100	beach	71
beauty	180	chocolate	100	#theft	70
#rage	173	marriage	99	leisure	69
money	171	#misery	99	beautiful	69
#lie	167	#pollution	96	nice	68
freedom	167	rest	96	perfume	68
#racism	165	countryside	96	tree	67
hate#	165	#tax	95	cinema	67
peace	156	blue	94	restaurant	67
mountain	156	sport	93	cat	67
#cancer	156	#hospital	93	honesty	65
laughing	153	#divorce	91	thank_you	64
smooth	151	#selfishness	91	#hunger	64
#cold	148	summer	91	#dirt	64
house	147	#hypocrisy	90	#villain	64
smile	145	#injustice	90	to_sleep	64
day	143	#famine	88	sharing	63
baby	141	#outstanding	88	#thief	63
work	137	#intolerance	88	to_love	63

Table 4.3:
The 37 words of the semiometric questionnaire not mentioned spontaneously

Absolute	*To conquer*	*Humble*	*Moderation*	*To ponder*
Persistence	*Challenge*	*To interrogate*	*Wall*	*Rule*
Armor	*Detachment*	*Inventor*	*Knot*	*Rigid*
Clever	*Elite*	*Maze*	*Gold*	*Robust*
Attachment	*To teach*	*Logic*	*To produce*	*Sacred*
Entrepreneur	*To climb*	*Magic*	*Question*	*Practical*
Ceremony	*To fertilize*	*Mask*	*Meditation*	*Virile*
Solid	*Firmness*			

This experiment consisting in spontaneously citing words in a representative survey is probably quite original, and the "ability to be spontaneously mentioned" of certain words has not been studied, to our knowledge, in a systematic way. We are dealing, at least partly, with very frequent words in everyday language; yet, there is no written corpus conducive to estimating these frequencies. The criterion for extreme frequency was not taken into account when we drew up the semiometric list.

There are in the answers a frequency concentration of consensual words, and also a highly dispersed distribution of relatively rare or idiosyncratic words (e.g. *pigwidgeon, honeysuckle, abdomen, Popocatepetl, clafoutis, candelabra*), stemming probably from the playful nature of the filling in exercise of such a questionnaire. The probability of finding relatively neutral words (like words in the semiometric questionnaire such as: *To teach, To interrogate, Question, Practical*), is very low. The number of words that appear only once (i.e. hapaxes) in these open or free responses is considerable (4,266 out of 7,170 distinct words)[94].

The open nature of the questions favours the quotation of synonyms, semantic neighbours (not necessarily belonging to our semiometric list) or inflections of a word (singular/plural, present/infinitive, male/female, etc.).

We find some dispersion in the frequency of synonyms and inflections. The most commonly used synonym consequently screens out the others.

---

[94] On the problems of lexical distributions, cf. Muller (1977, 1979), there is also a statistical modeling of rare events applied to the lexicometric field, cf. Efron and Thisted (1976), Baayen (2000).

The frequency of the former is excessively increased to the detriment of the latter[95].

Sections 4.2 and 4.3 below are assigned to exploring, for the first time, the corpus of responses; followed by Section 4.5, which will use the fact that a small proportion of respondents had answered the semiometric questionnaire during the previous year. This is done in an attempt to reconcile the two types of questionnaires.

4.2 Exploring the responses

The following analysis allows us to understand the nature of responses to these open questions better. As in the previous sections, tables, cross-tabulating words and responses, are described using the technique of *correspondence analysis*, briefly presented in Appendix A1.4. The analysis is performed initially on the words quoted as pleasant. The first of the tables analyzed corresponds to the frequency threshold minimum 4, viz. the words quoted more than four times. It contains 1191 rows (respondents) and 592 columns (words).

Table 4.4 shows the typical words of the first principal axis (or *factorial axis*) of this table with 1191 rows and 592 columns. What is noteworthy is that the words in the right column (positive coordinates) are much farther away from the origin (almost twenty times for the first words) than those in the left one (negative coordinates), concentrated near the origin of the axes. On the other hand, the words on the right column are semantically fairly homogeneous, whereas those in the left column could have inspired the French poet Jacques Prévert, who wrote under the title ”Inventory” a celebrated poem merely consisting of a motley collection of words.

---

[95] The word *Armour* in the semiometric questionnaire is never spontaneously mentioned in this sample, but the following graphic forms of the same root are cited: *#armament (1), #arms (12), #armed (1)1, #army (11)*. Also never mentioned spontaneously are the words (also absent from the semiometric questionnaire) *breastplate, shield, sword, sabre, sword*, which, like *armour*, are more reminiscent of history or fiction than a danger or a current debate. In contrast, we find: *rifle (14), #bomb (34), #bombs (2),#cannon (6), #bombing (3), #submachine gun (1), #machine gun (1), #missile (1).*

Table 4.4: Extreme Words on the first axis
(Analysis of open responses for pleasant words)

Left side (very close to origin)		Right side (far away)	
drawing	-.48	polite	9.23
friends	-.46	smiling	8.60
storm	-.44	dynamic	7.81
horizon	-.44	courageous	7.03
concert	-.43	helpful	6.58
comet	-.41	adorable	5.35
milk	-.40	nice	5.23
radio	-.40	clean	4.78
plant	-.39	sincere	4.58
fire-wood	-.39	just	4.54
daughter	-.39	goodnight	4.41
train	-.39	sociable	4.34

The explanation for this phenomenon is simple, (and announces the findings of the analyses of the subsequent chapter): the great mass of word-points is concentrated around the origin, and a small sub-group (about twenty out of 592 words) being very distant and made up of correlated words, is the only group responsible for one axis.

We notice that, on the right of the axis, we have almost only adjectives, in contrast to the mass of nouns on the left, often referring to objects, substances, material concepts.

The following axes highlight small groups of words, without any axis having a bipolar interpretation. Clearly, again, there are locally privileged groups, but there exists no overall structure of these groupings.

To demonstrate the hypothesis according to which the structural complexity observed is produced by infrequent words, analyses were reiterated while increasing the threshold of minimum frequency. Retaining only the 158 words occurring more than 25 times in the corpus of responses to the first open question (*pleasant* words), the overall structure is significantly improved: the first two axes are defined by very few words, but these have the merit to appear at least 25 times, and therefore the results are more significant from a statistical perspective.

Given the shortcomings of factorial methods to describe such structures[96], we have used *Kohonen maps (a.k.a.: Self Organizing Maps)*, which strive to combine a classification procedure and a factorial plane (see Annex A1.8).

Figure 4.1 shows such a *"Self-Organizing Kohonen Map"*. So we have a graphic description of associations between words appearing more than 25 times in all the responses. Two words belonging to the same box are often cited together by the same person. This is still true, but to a lesser extent, if they are located in neighbouring boxes. If the boxes are far apart, the corresponding words are, however rarely (or not at all) mentioned simultaneously.

The first finding that can be made looking at the graph refers to a weakness in the spontaneous questionnaire: the role of associations of ideas when one chooses words. We can observe such associations just by reading the raw responses, but Figure 4.1 confirms that this is not an accidental phenomenon[97].

Groups such as (*summer, vacation*), (*dad, mom*), (*cake, chocolate*), (*sorry, thank you, hello*) can indeed be an indication that the questionnaire is being filled in somewhat automatically.

Here are three examples of actual responses for which this type of association is quite clear:

- *"nice, friendly, helpful, thank you, kindly, excuse me, pardon, hello, obedient"*
- *"hello, good evening, thank you, sorry, excuse me"*
- *"hello, thank you, please, pardon"*

[96] There is in this case a lack of clarity due to the overload of graphics (many overlapping points). The analyses are still interesting.

[97] This phenomenon is even more visible with a lower threshold, thus with less frequent words, but the corresponding Kohonen maps are consequently too bulky for a publication in the present format.

equality justice fraternity candor	respect intelligenc e honesty goodness	politeness kindness understandin g	harmony		pretty beautiful		excuse_m e thank_you hello
solidarity sincerity	generosity faithfulnes s loyalty	tenderness	serenity softness calm		good		
Tolerance humour courage	sharing peace liberty confidence	nature hope omfort	to live quietnes s pleasure	island snooze to eat woman to sleep			happy to love
youth gaiety convivialit y	joy beauty friendship	life to laugh relaxation love	rest heat	dream bed	star bird water color	green caress blue	soft cuddle
well-being	success passion good luck happiness	party Present	smile spare time child song		sand light book flower sky	perfum e kiss	rose dad mom baby
	cleanness	birth marriage play money birthday	summer vacation	travel sport sun meal encounter spring beach music	sea countrysid e	cat friend	chocolate
kind amiable		work health gift	weekend leisure family	car walk house Champagn e animals	dance	snow dog	theater painting
pleasant		travels grandchildre n	flowers	TV restaurant	byke rambling mountain reading cinema	rain birds garden cookin g boat	wine greenery fruit forest singing tree

Figure 4.1: Kohonen Map (*Self Organizing Map*) representing the associations between words in the open-ended (or free) responses

Also shown below is a long answer that illustrates how associations[98] were formed gradually during the response:

[98] Mechanisms of word associations have been studied in other contexts, by psychologists, including: Ferrand and Alario (1998).

• "*campaign, landscape, friendship, happiness, intelligence, measurement, sensitivity, towards, old woman, companion, baby, child, mother, father, grandmother, grandfather, aunt, uncle, cousin, grass, tree, moon, star, ocean, river, stream, dawn, twilight, innocence"*.

It is clear that the closed questionnaire has the advantage of avoiding the excesses and unintentional overweighting of certain themes. This tendency to associate the words consecutively also concerns the grammatical parts of speech. We have seen that adjectives are opposed to nouns on the first axis (Table 4.4). A list that begins with an adjective (or a noun, or a verb) will tend to continue with adjectives (or nouns, or verbs).

4.3 Principal axes of spontaneous responses

If we keep only the words appearing over 10 times (164 distinct words, corresponding to 5474 occurrences, 444 respondents), we begin to make out the stable dimensions that are approaching certain semiometric axes.

The principal plane – plane (1, 2) – of the correspondence analysis performed with this threshold of frequency (the plane (1, 2) is not shown here) opposes almost all of the words to two small clusters: "*thank you, excuse me, hello*", and "*pleasant, gentle, kind*" (a triangular structure already mentioned earlier in this chapter).

This phenomenon is clearly visible on the Kohonen map (Figure 4.1) and moreover concerns the same table (individuals - words). Indeed, we find, well isolated in the upper right corner of Figure 4.1, the three previous forms of politeness, surrounded by empty boxes, and the other words mentioned (*pleasant, gentle, kind*) in the lower left corner of the Figure[99].

Figure 4.2 shows the arrangement of words on the plane (3, 4), a plane generated by the third and fourth axes. We notice on that plane a lot of words evocative of leisure and pleasure on the left side of the vertical axis[100]. On the right side, we can read words such as "*politeness, solidarity, honesty, courage, truth* "... Here is an opposition "Duty / Pleasure" which is

[99] This allows us to highlight that the Kohonen maps, being a non linear method, have a great power of compression: we can read information on several axes simultaneously.

[100] Using the technique of projection of supplementary elements, we could see that the word "*#work*", i.e. the rejection of work, is, in fact, located in this area. The display is limited to *pleasant words*.

reminiscent of axis 2 of the semiometric structure... obtained here without referring to a fixed list of words. We will also see that axis 4 (the vertical axis of Figure 4.2) is not unrelated to axis 3 of the semiometric structure.

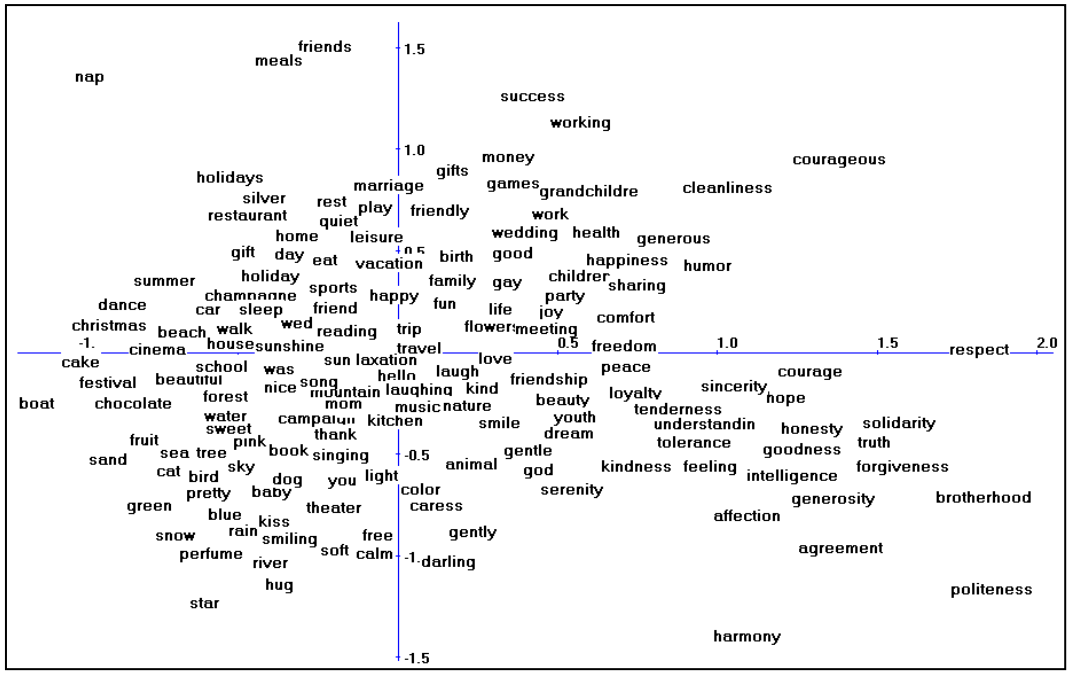


Figure 4.2 Spontaneous answers: Principal plane (3, 4) (3: horizontal, 4 :vertical). Proximities between spontaneously mentioned words.

To conclude on this first attempt at exploring the responses to open questions, we must bear in mind the extremely noisy aspect of the data collected in this way, compared to a collection of scores for a list of words identical for all respondents. The distance between individuals will depend indeed on the small number of words they may have in common, and individuals who have no words in common will have undifferentiated distances between one another. The existence of this considerable noise does not prevent us from finding structural features, but the current size of the sample limits the scope of work on individual data.

There are, nonetheless, some similarities that suggest that a structure related to the semiometric structure can occur spontaneously and independently of any closed questionnaire.

4.4 Spontaneous selection and characteristics of respondents

One technique used frequently in the statistical processing of open-ended questions is to analyze not the answers themselves, but the tables of aggregated responses, which are less sensitive to individual fluctuations. The answers are indeed very "noisy", but hopefully, by using this technique, we may find patterns that are undetectable directly from non-aggregated responses. *A priori* groupings of respondents were made from some of their available features. This provides an opportunity to see that sex and age, considered separately, or even better, simultaneously, are not independent of the words cited as pleasant or unpleasant.

Table 4.5: Words characteristic of extreme ages

Spontaneous characteristic words	Test values	Probability
Age: Under 30		
1 pleasure	4.00	.000
2 eat	3.78	.000
3 sleep	3.48	.000
4 hug	2.63	.004
5 baby	2.37	.009
6 rest	2.29	.011
7 chocolate	2.22	.013
Age: Over 55 years		
1 politeness	3.38	.000
2 courage	3.00	.001
3 fraternity	2.77	.003
4 travel	2.74	.003
5 thank you	2.42	.008
6 read	2.38	.009
7 affection	2.37	.009
8 cleanliness	2.37	.009
9 forgiveness	2.22	.013

The first basic variable, used for clustering responses, is age, for which two extreme categories are retained: those aged under 30 and those aged over 55 (Table 4.5). Only the pleasant words appearing at least 16 times are

taken into account. Words characteristic[101] of young people (*pleasure, eat, sleep*...) and older people (*politeness, courage, brotherhood*,...) are reminiscent of the two ends of the horizontal axis of Figure 4.2, an axis that is rather similar to semiometric axis 2.

The second basic variable is the sex of the respondents. Words most characteristic of men and women are shown in Table 4.6. Test values, which are lower, show that sex is less discriminating than age, the latter bringing great heterogeneity even within each group, that of men and that of women.

Four new categories of respondents are now obtained by combining sex with two age classes: men under 30, men over 55, women under 30, women over 55 (Table 4.7).

Tables 4.6 and 4.7 have few words in common: the cross-tabulated category variable "age x sex" is a more relevant variable than sex alone. Among the words common to both tables, we note: *courage* for men*, mom* and *chocolate* for women.

Table 4.6: Words characteristics of males and females

Spontaneous characteristic words	Test values	Probability
Males		
1 walk	2.82	.002
2 joy	2.41	.008
3 courage	2.30	011
4 sport	2.25	.012
Females		
1 mom	3.32	.000
2 chocolate	2.53	.006
3 nice	2.20	.014
4 book	2.09	.018

[101] These words are unusually frequent words in the category relative to their average frequency in the whole sample. The difference between the internal frequency of the category and the overall frequency is converted into *test-value*, i.e. in the standard normal variable under the assumption of independence of frequencies (see Appendix A1.9.1). The last column of Table 4.5 gives equivalent information in terms of probability (the test-value is converted into the more classical *p-value*).

However, the words characteristic of younger people and older people, in fact, described in Table 4.5 either men or women, but not both at the same time. Thus, among the words, characteristic of the younger, the words, *fun, sleep, eat*, characterized mainly men, whereas the words, *hug, baby, chocolate*, characterized women. It was therefore important to cross-tabulate both sex and age variables.

The sample size does not allow the taking into account of more elaborate cross-tabulations.

Table 4.7: Words characteristic of four sex-age categories

Spontaneous characteristic words	Test values	Probability
Males under 30		
1 sleep	3.24	.001
2 pleasure	3.03	.001
3 eat	2.84	.002
4 leisure	2.29	.011
Males over 55		
1 courage	3.59	.000
2 brotherhood	2.80	.003
3 property	2.68	.004
4 health	2.24	.012
Females under 30		
1 chocolate	3.06	.001
2 baby	3.02	.001
3 animals	2.46	.007
4 mom	2.26	.012
5 hug	2.13	.016
6 summer	2.12	.017
Females over 55		
1 thank you	3.00	.001
2 affection	2.83	.002
3 politeness	2.53	.006
4 hello	2.14	.016

Figure 4.3 shows, in the form of proximity graphs, a summary of the links between the six categories (two similar categories have common

lexical profiles) and between words (two similar words have similar socio-demographic profiles)[102].

The representation obtained confirms the complementarity, one might almost say the additive effects, of these two basic variables.

Indeed, these two variables spread out along orthogonal directions, age, horizontally, opposing the older categories on the left, the younger ones on the right; sex, in a vertical direction, opposing men, below, and women at the top.

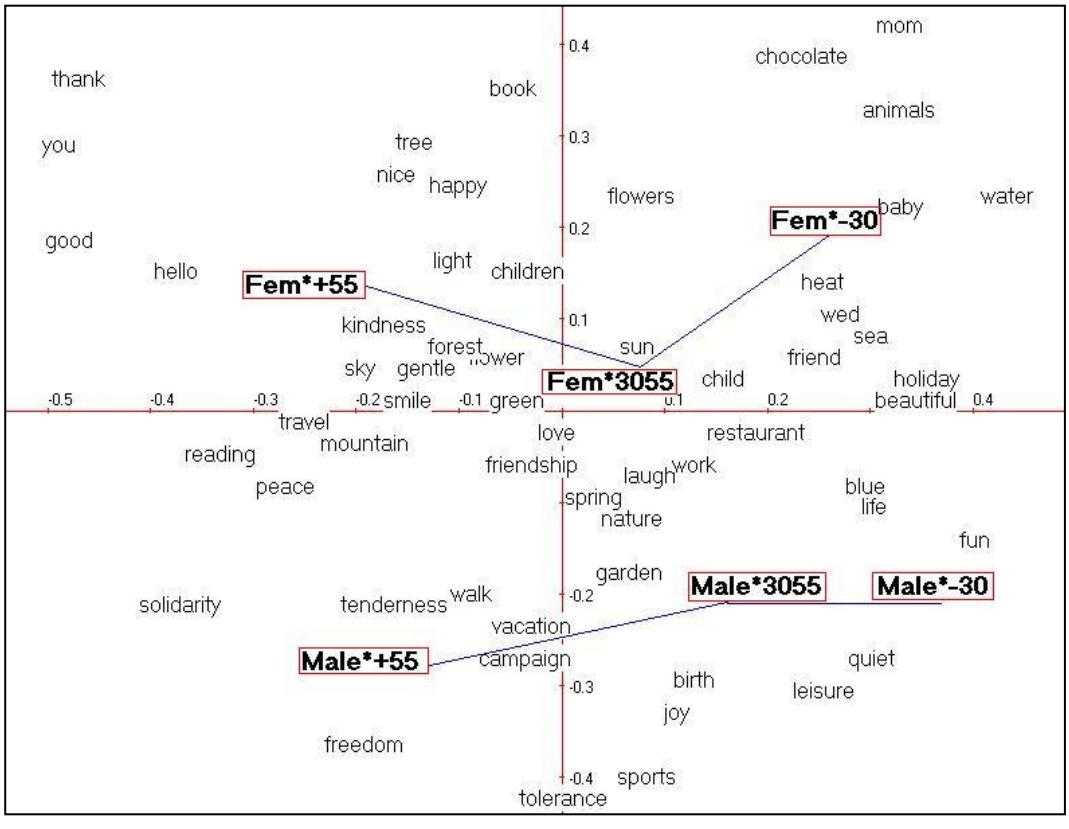


Figure 4.3: Spontaneous quotation of words per sex and age
(Plane of the first two axes).

[102] This is the first principal plane of a correspondence analysis of the contingency table (called lexical table) whose rows are the words appearing at least 16 times, and whose columns are the six categories (sex-age), including, this time the intermediate age group. The description tool is the same as in Figure 4.2, but it now applies to aggregate data, rather than individual data.

What is noteworthy, and what may have escaped the notice of readers unfamiliar with the principle of these methods of description is that nothing in the six columns of the input array being analyzed indicates that they originate from the cross-tabulation of two variables. The only information that determines the position of the column-points on the chart is the lexical profile of these columns; consequently all the words spontaneously mentioned as pleasant in the categories correspond to these columns.

Finally, this figure is a readable synopsis of the results in Tables 4.5, 4.6 and 4.7. It describes, in the context of a nuanced continuum, the oppositions between the male and the female for a given age, together with the extent of internal lexical variations for each sex categories, depending on age[103].

We still must furnish the elements of validation provided by the test values in the tables. These are given by the *bootstrap confidence ellipses* (already used in Chapter 2, in a different context)[104].

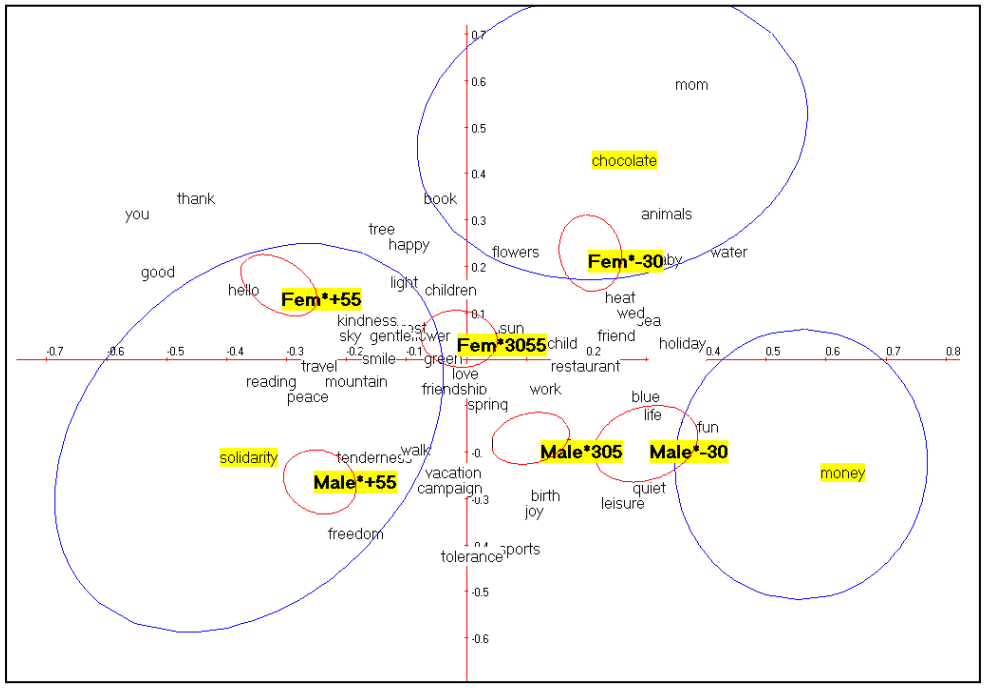


Figure 4.4: Confidence Ellipses for categories and words

[103] Remember that semiometric axes 2 and 3 respectively oppose, in the same way, men and women, the younger and older people.

[104] On the technical aspects of the method: See Appendix A1.9.5.

Figure 4.4 shows the confidence ellipses of the category-points *age x sex* (small ellipses in the central part of the graph). It is clear that the pattern observed is stable, despite the moderate size of the sample. As is always the case for this type of lexical table, the confidence ellipses for the words are much larger (the three ellipses selected as examples concern the three words: *Money, Chocolate, Solidarity*).

Nevertheless, the large size of these ellipses does not fundamentally alter the interpretation of the proximity observed. In the bottom right of the graph, the word *money* is still characteristic of young men, regardless of its location in its confidence ellipse. In the same way, on the left hand side, the word: *solidarity* remains characteristic of the elderly. Above, the words: *chocolate* is characteristic of women in this exercise of spontaneous pleasant words.

To conclude on Section 4.4, devoted to the proximities between words mentioned spontaneously and some basic characteristics of respondents, we notice the richness and consistency of the material collected while allowing respondents full freedom to answer. The responses might be noisy, colourful (one might say so for some of them), their analysis shows patterns that are indicative of the potential of this type of data collection.

4.5 Bringing semiometry and open questions together

We have been able to collect semiometric scores for 335 people among the 1,191 respondents answering the two open questions requiring them to name spontaneously pleasant or unpleasant words.

We have defined in Chapter 1 a semiometric structure[105]. In this section, we have been able to check that such a structure is still valid for the relatively small group of 335 respondents, at least regarding the first four axes[106]. This would not unduly surprise us if this group had been taken at random from the sample panel. Yet these are people who accepted to respond to the open questionnaire, and we might have feared that they could represent a biased sample. Apparently this is not the case.

---

[105] The structure has been defined by a set of easy interpretable principal axes resulting from the Principal Components Analyses (PCA) of tables of scores.

[106] Thus, the PCA of the (335 x 210) table of scores provides us with 4 principal axes likely to be interpreted in a similar fashion as in Chapter 1.

This allows us to answer the following question: what are the words spontaneously mentioned that characterize the first semiometric axis.

The words quoted spontaneously are to be considered as *supplementary variables* (as well as sex or age) and will be projected onto the principal axes (see Appendix A1.9.3).

We saw that, for the whole corpus of responses to the open-ended questions, about half of the words in the semiometric questionnaire appear at least once. But a statistical study of the closeness of the two questionnaires requires us to retain only the words quoted with some frequency. The closeness will be sometimes surprising because few words mentioned spontaneously belong to the semiometric questionnaire.

While the position of semiometric words on the axes is characterized by their correlation coefficients with these axes, the position of words mentioned spontaneously, projected as additional categories on the same axes, is characterized by the test-values that take into account the population concerned (the number of people who cited the word) by converting the coordinate on the axis into a standardized normal variable[107].

4.5.1 Links between spontaneous words and the first axis

In the semiometric analyses of the previous chapters, we have considered the first axis as a methodological axis of survey participation, deemed to be outside the semiometric structures. This axis has been given special treatment (chapter 3). The position of spontaneous words on this first axis will be interesting from a methodological point of view. This axis contrasts, as we have seen, the respondents fully utilizing the scale proposed for the scoring with respondents who use only the central part of the scale.

The following words, which are the most characteristic of individuals who use the full scale of scores for the semiometric analysis of 335 individuals are: *Courage, Politeness, Hero, Honour, To protect, Robust, Tradition, Dynamic, Refined, Elegance, Honest* ..., with correlation coefficients with the axis ranging from -0.49 to -0.59.

The words characteristic of those who use only the central part of the scale are: *To betray, Anguish, Rebellion, Danger, Disorder, Death*..., with

[107] Remember that the value thereof will be approximately between -2 and +2 if the respondents who cited the word could be considered as randomly drawn from the sample (See Appendix A1.9.1).

correlation coefficients with the axis, much lower, ranging from 0.17 to 0.29.

The words quoted spontaneously characterizing individuals who use the scale of scores fully, [followed by their test-values in parentheses] are: *confidence* (-2.9), *love* (-2.8), *hello* (-2.7) *thank you* (-2.4) *courtesy* (-2.1), *honest* (-2.1). We also find the notions of politeness and honesty. Note also, for the same individuals, the words considered to be the most unpleasant: #*incest* (-2.2), #*assassin* (-2.1), #*pedophile* (-2.1).

We obviously could not see appearing on this axis the words of the semiometric questionnaire: *Hero, To protect, Robust, Tradition*, *Refined and Elegance*, which are never mentioned spontaneously as agreeable words.

We saw that, from the semiometric questionnaire, there were no words strongly characterizing the respondents occupying the positive part of the axis, i.e. respondents utilizing the central part of the scale. The assumption was made that these were indeed respondents who were reluctant to confide, or "were not playing the game of the questionnaire fully". These respondents raise problems in all procedures of sample surveys. Like abstainers, to whose still observable fringe they belong, they are the weak point or the penumbra of this instrument of observation[108].

We cannot expect, either, to see mentioned as pleasant words: *To betray, Anguish, Revolt, Fault, Danger, Disorder, Death*... In a very telling way, the most characteristic word of these respondents is #*constraint* (indeed regarded here as a particularly unpleasant word, with a test-value of 3.7, which is the highest of all the words on the axis). Then come the words: *life* (2.7), *rain* (2.5), o*cean* (2.5), *intelligence* (2.3), *women* (2.2), and unpleasant words (except for #*constraint* already cited) #*sadness* (2.4), #*spider* (2.3), #*racist* (2.3), #*work* (2.2).

These are also younger people, better educated, and they are most often men. We now learn that these are above all people who reject constraints (the translation of a statistical fact: they are characterized very significantly by spontaneously citing #*constraint* as an unpleasant word). The rejection of the word *work* (and probably the constraint that it represents) goes in the same direction. The words: *life, rain, ocean, intelligence,* also occupy

[108] Sometimes we designate, not without discomfort, these differences in attitudes as the *scoring effect*. We can not exclude such an effect, but naming it does not solve all the methodological problems posed by the attitude of the respondent with respect to the questionnaire and the survey (see Chapter 3).

significant positions (more than 2.3 standard deviations from what would be their positions assuming a random distribution of responses) but the interpretation is trickier. The word *rain* is interesting because it appears more as a disagreeable word (37 citations in the sub-sample used here) than as a pleasant word (only 5 citations).

But the position of those who chose *rain* as a pleasant word is sufficiently and characteristically salient on the first axis so that the centroid of these respondents is statistically significant. It is regrettable that the sample size of this experiment does not allow us to characterize these respondents better: people who appear to be less conventional, more individualistic, but nevertheless willing to answer!

4.5.2 Links between spontaneous words and the axes: 2, 3, 4.

Table 4.8 gives the most characteristic spontaneously mentioned words in axes 2 to 4. Remember that these are the axes stemming from the principal components analysis of the 210 semiometric words, those spontaneously mentioned here being projected afterwards.

As axes 5 and 6 are not well reconstructed on our small sample of 335 respondents, they will not be discussed here. Words in bold are test values greater than or equal to 2.3, and thus occupy positions that should be interpreted. The locations of words that correspond to test values less than 2.3 will not be interpreted, although they provide a pool of words that might be candidates for being much more statistically significant if the sample were larger.

*- **Axis 2** (Duty / Pleasure)*

The first column of this table refers to axis 2, conventionally designated as the axis of "Duty / Pleasure" which, remember, opposes the words: *Discipline, To obey, Homeland, Morals, Soldier, To economize, Industry, Priest, Rule*... to the words: *Sensual, To dream, Adventurer, Original, Island, Nudity, Wild* …

Table 4.8: Projections on the semiometric axes of words spontaneously cited

Axis 2	Vtest	*Axis 3*	Vtest	*Axis 4*	Vtest
education	**-2.7**	**girl**	**-3.4**	**grandfather**	**-3.7**
happy	**-2.4**	**hiking**	**-2.9**	**#hate**	**-3.1**
generous	**-2.3**	**river**	**-2.8**	**grandmother**	**-3**
union	**-2.3**	**#nazism**	**-2.7**	**peace**	**-3**
clean	-2.2	**#contempt**	**-2.4**	**#alcoholism**	**-2.8**
helpful	-2.2	**wind**	**-2.4**	**fraternity**	**-2.8**
visit	-2.2	**#fatigue**	**-2.3**	**labour**	**-2.8**
thank you	-2.2	**#hate**	**-2.3**	**grandchildren**	**-2.7**
good-mood	-2	eating	-2.2	**#war**	**-2.6**
wisdom	-2	speed	-2.2	**compliance**	**-2.6**
feeling	-2	affection	-2.1	**meeting**	**-2.5**
work	-2	drink	-2.1	**tree**	**-2.4**
aircraft	-1.9	idleness	-2.1	**#drug**	**-2.3**
care	-1.9	bread	-2.1	**truth**	**-2.3**
----------	-----	----------	-----	----------	-----
#weapon	2.2	Champagne	1.8	sleep	1.5
#racism	2.2	holiday	1.8	dining	1.5
kiss	2.2	meeting	1.8	ocean	1.5
delight	2.2	TV	1.8	fun	1.5
green	2.2	vacation	1.8	sports	1.5
heat	**2.3**	kindness	1.9	weekend	1.5
ocean	**2.3**	beautiful	1.9	money	1.6
#pollution	**2.4**	#nasty	2	beautiful	1.6
color	**2.4**	nice	2	horse	1.6
peck	**2.4**	husband	2	restaurant	1.6
desire	**2.4**	cute	2	evening	1.6
husband	**2.4**	#divorce	2.1	#anxiety	1.7
star	**2.4**	pretty	2.1	cute	1.7
laugh	**2.5**	success	2.1	food	1.7
wind	**2.5**	#storm	2.2	#nightmare	1.8
#duty	**2.6**	pleasant	2.2	#sorrow	1.8
baby	**2.7**	gem	2.2	#pay	1.8
island	**2.7**	summer	2.2	wedding	1.8
#intolerance	**2.8**	**#fire**	**2.3**	girl	1.9
#violence	**2.8**	**birthday**	**2.3**	happy	1.9
mom	**2.8**	**Present**	**2.3**	vacation	1.9
#pedophilia	**2.9**	**darling**	**2.4**	#obligation	2
campaign	**2.9**	**family**	**2.4**	friendliness	2
sleep	**2.9**	**marriage**	**2.4**	**fortunate**	**2.3**
book	**3.1**	**birth**	**2.8**	**meadow**	**2.5**
music	**3.1**	**amiable**	**3**	**#bill**	**3**
pleasure	**3.9**	**kindness**	**3.1**	**desire**	**3.1**
#money	**4.9**	**#rain**	**3.2**	**speed**	**5.3**

The words describing the side "Duty" (mandatory in the semiometric questionnaire in its closed form) are not mentioned spontaneously as

pleasant words, so we cannot expect to find them in the responses to our open question.

First of all, we find *education (-2.7), happy (-2.4), generous (-2.3), union (-2.3)*, which are, arguably, the pleasant components of duty (rather subtly, happy is opposed to pleasure on this axis). However, the first column contains many more significant words (with test values greater than or equal to 2.2) in its lower part that corresponds to the half-axis "Pleasure".

Among the most significant words, there is *#money* (hence quoted as unpleasant, with a test value of 4.9) and the word *pleasure* (3.9), which is not a word in the semiometric questionnaire, but has been chosen to interpret the axis from almost the beginning of the method and therefore well before the experiment with open questions. Then we find the words *music, books, sleep, country*, etc. Altogether more than 28 words spontaneously mentioned have test-values greater than or equal to 2.2.

In short, the words mentioned spontaneously shed light on the interpretation of the first semiometric axis, and confirm that of the second.

- Axis 3 (Attachment / Detachment)

The second column of Table 4.8 shows the place of words along the third semiometric axis conventionally designated as the axis of "Attachment / Detachment". Here again, the words relating to "Detachment" in the semiometric questionnaire (*Danger, Death, To break, Storm, Anguish, Emptiness, To punish...*) are unlikely to be spontaneously mentioned as belonging to the most pleasant words.

We find, for individuals on the axis of "Detachment" (mostly men, often young), the following words spontaneously mentioned: *girl (-3.4), hiking (-2.9), river (-2.8), wind* (-2.4). This is a less inhuman and less caricatural "Detachment" than the one suggested by semiometry, especially if one adds to the list *eat, speed, affection, drink, idleness, bread*, for the pleasant words, *#nazism, #fatigue, #hate,* for the words quoted as unpleasant[109].

In the "Attachment" area (bottom of the second column), however, there are words which are found in the semiometric questionnaire such as *birth (2.8), marriage (2.4) (Wedding* in the semiometric questionnaire*), gift(s)*

[109] Remember again that the over-scored words in Semiometry are not necessarily well scored words; these are words that score above the average (which can be very low, and unknown to the respondent, since the mean score is calculated retrospectively on all responses).

(2.3), Jewel (2.2). The most characteristic words are: *#rain* (3.2) considered this time as an unpleasant word, and *kindness* (3.1).

The open questionnaire indeed confirms and enriches the semiometric interpretations for this axis.

***- Axis 4** (Sublimation / Materialism)*

This axis, which is characterized by the opposition "Sublimation / Materialism" has, in its negative part ("Sublimation"), some of the words mentioned spontaneously such as *grandfather, grandmother, peace, fraternity, work, grandchildren, compliance, meeting tree, truth*.

We find the words *tree* and *peace* that are common with those of Semiometry. The word *book*, spontaneously mentioned, we have seen, clearly characterizes the positive part of Axis 2 ("Pleasure"), next to the words *music* and *enjoyment*. When the word *book* is mentioned spontaneously as a pleasant word, it is indeed a book for fun, not a school book or one for education (the two words, school, education, accompanying *book* on semiometric axis 4). On the other hand, giving a scoring to the word *book* in the semiometric list implies, perhaps, taking into account more semantic features of the word *book*.

*Grandfathe*r and *grandmother* (as well as *grandchildren*) are missing from the semiometric questionnaire, but are associated with *peace* and *serenity* that characterize this axis. On the positive side of the axis ("Materialism"), the spontaneously mentioned word that characterizes the most the half axis is: *speed* (with the outstanding test value of 5.3). It is also a semiometric milestone for this dimension, as indeed is the second most characteristic word: *desire* (3.1)[110].

Despite the different nature of the questions and the small size of this sample, there is an undeniable consistency between the responses to the two types of questions, closed and open.

4.6 Conclusion

These open question experiments are very instructive. Let us just mention the three main results obtained at the end of this collection and the statistical treatment of original data.

[110] This analysis of axis 4 will be considered with caution given the sample size.

The first result is that one can obtain from a modest-sized sample, a *spontaneous semiometry* from the open questions of the type proposed in this section, *i.e.* without any constraints. We can indeed obtain local associations, schematized by the Kohonen map in Figure 4.1, and factorial planes with a fairly strong relationship with the planes of Semiometry (Figure 4.2), but not all of the latent dimensions are stable and interpretable, as the closed questionnaire reveals.

This partially negative result helps to understand better what role a "closed" list of words identical for each respondent[111] can play.

Mentioning words spontaneously induces a limitless dispersion of vocabulary, reducing the significance of the distances between individuals in a correlative manner. Moreover, many words rich in meaning and value are *a priori* neither pleasant nor unpleasant, and therefore are unlikely to appear in the spontaneous responses. Note also that the consensual words (*love, holidays*, etc.) weigh down the collection without contributing to any decisive information.

The second and third results below concern more specifically the study of respondents taking part in the open questionnaire who also had responded earlier to the semiometric questionnaire.

The second result is also more a methodological confirmation here than a discovery. It deals with the statistical role of scores as opposed to a mere mention of presence or absence. The scoring can refer to a mean score for each word, an average score unbeknown to respondents. We can give a word a bad score, and yet attribute a score above an average score, unknown at the time of the interview. This treatment, elementary as it may seem, allows one to work exclusively on individual differences, and to obtain bipolar axes with acceptable stability properties[112].

The third result concerns the value of spontaneity as a complement and illustration for the interpretation of the semiometric axes. Unable alone to

---

[111] This does not, moreover, exclude varying the word order within the list to eliminate a possible order effect.

[112] Thus, for axis 3, *Death* (average score: 1.76), *Anguish* (average score: 1.84) characterize individuals on the half-axis "Detachment", while *Family* (average score: 6.48) and *Tenderness* (average score: 6.67) are words that characterize the opposite semi-axis, "Attachment". We calculated the average scores on all individuals from scores from 1 to 7. Clearly in these circumstances a person who attributed the rating 2 to *Death* and *Anguish* (and thus above the average scores of these words) and 6 to *Family* and *Tenderness* (i.e. below average scores) may find himself on the "Detachment" side while having a legitimate sense of attributing the wrong score to the first two words and having attributed a good rating to the last two.

generate the bipolar axes, the corpus of open-ended questions can enrich the interpretation of the axes derived from the semiometric questionnaire. For, if spontaneous citing can not supply all the pairs of respondents belonging to the same operational distance (as, for example, all pairs of individuals, who have cited no common word, are at comparable distances apart), it can characterize almost without any limit these individuals[113].

It suffices, in fact, that some individuals have cited a word for its position to be statistically tested on the semiometric axes. This has enabled us, for example, to enhance with nuanced terms the somewhat austere semiometric structures of "Duty" and "Detachment". And also to understand better the nature of the first axis which is a subject already dealt with in the previous chapter.

[113] In more technical terms, the open question provides inefficient active variables, but useful and enlightening additional or illustrative variables (See Appendix A1.9.3).

CHAPTER 5

Semantic structures and Semiometry

The stability of the semiometric structure revealed in the previous chapter raises a series of questions relating both to the nature and the origin of the observed structural features.

One of the fundamental issues concerns the role and importance of purely semantic relationships within the semiometric structure. Note that the semantic relationships are certainly not easy to define or measure, and in the discipline known as *Natural Language Processing*[114], issues affecting semantics remain among the toughest.

In this chapter, we will try to answer the following question: *Are the semantic similarities between words responsible for most of the structure observed or not*? In other words, in a nutshell: *Is the structure that we observe a linguistic structure rather than a psychological or psycho-sociological one*?

Two words with similar meanings are scored similarly and therefore correlated. Is the observed pattern none other than the network composed of those semantic links? The stability of the structure would result from this, since language is relatively stable over time. This stability is also virtually the same for different age groups, sex, etc. Such a semantic network must also overcome - more or less - the translation of the questionnaire, hence the relative stability from one country to another.

To provide some answers to these questions, this chapter describes three pieces of research corresponding to as many sections.

The first section with the heading “Semantic Neighbourhood of the Semiometric questionnaire”, is assigned to the following experiment: the

[114] A discipline from which are derived mainstream products such as spelling checkers, translation software, document search engines.

210 words of Semiometry are described by their *synonyms*, or at least their *semantic neighbours*, the latter being provided by a thesaurus. In other words, there will be no more questionnaires to be filled in by a representative sample… the language itself responds through some of the tools that were forged to describe and understand it. Of course, thesauri are not perfect instruments, and they differ significantly. The experiment will also allow us to see the extent of the difficulties that lie ahead.

The second section entitled "The Semantic network of French verbs", describes a much broader experiment, beyond he selection of 210 particular words. It takes all the verbs of the French language[115] to explore the semantic structure of these verbs with respect to semantically neighbouring terms. This experiment confirms the existence of local semantic nearness or proximity, but there exists no dominant stable principal axis.

The third section, the title of which is "The internal semantic fields" shows how the answers to the semiometric questionnaire and correlations between words derived from these responses actually describe, and quite finely too, links of a semantic nature. This result was not obvious *a priori* because, and let's bear this in mind, the questions raised only relate to the *pleasant* or *unpleasant* aspect of the words taken from the list.

5.1 Semantic Neighbourhood of the Semiometric Questionnaire

To understand whether the semantic similarities between words are responsible for most of the semiometric structure, we will confront the semiometric list of words to a thesaurus.

5.1.1 An attempt to describe a purely semantic network

In this first experiment, each semiometric word is described by its "synonyms", collated from an external source[116]. Table 5.1 gives a short excerpt taken from the collection of words thus obtained from the first

[115] In fact, the 829 most common verbs in the classic Bescherelle grammar book, including many very rarely used verbs, some even almost unknown.

[116] In the French version of this book, the thesaurus was the "Dictionnaire de synonymes et contraires" by Henri Bertrand du Chazaud (Robert, 1994). In this English version, synonyms and semantic neighbours have been supplied by the thesaurus of the "Institute for Cognitive Science of the CNRS" (National Centre for Scientific Research).

words: *Absolute* to *Anguish*. As is apparent, the number of semantic neighbours is very variable.

The number of neighbours can be very large and can reach 185 words for the word *To break* (a word which does not appear in the excerpt from Table 5.1); or this number can be very small (even almost nonexistent, for example, the adjective *Metallic* also outside Table 5.1 has only one neighbour in this dictionary).

From this new collection of texts a matrix is built containing in lines the 210 basic words, and in columns all the words found in the collection, viz. all the neighbours of the 210 semantic words. The table, at the intersection of row "i" and column "j" contains the value "1" if the word "j" is among the semantic neighbours of the word "i", and contains the value "0" otherwise. If some words have common semantic neighbours, a calculation of appropriate distance will designate them as being close from each other.

Table 5.1: Examples of semantic neighbours for the first nine words[117]

Absolute
actual, arbitrary, authoritarian, autocratic, categorical, certain, complete, conclusive, dead, decided, decisive, definite, definitive, despotic, dictatorial, downright, entire, flat, implicit, indubitable, infrangible, inviolable, out_and_out, outright, peremptory, perfect, plenary, positive, pure, rank, real, right_down, sheer, stark, supreme, sure, thorough, total, true, tyrannical, unadulterated, unalloyed, unbounded, unconditional, unconditioned, unequivocal, unlimited, unmitigated, unmixed, unqualified, unquestioning, unrestricted, utter
Persistence
assiduity, constancy, continuity, determination, diligence, doggedness, drive, durability, endurance, grit, indefatigableness, insistence, perseverance, perseveration, perseverence, persistency, pertinacity, pluck, resolution, tenaciousness, tenacity
To_buy
acquire, acquirement, acquisition, bargain, bribe, corrupt, deal, get, obtain, obtainment, pick_up, procure, procurement, purchase, steal, take
To_admire
adore, applaud, appreciate, esteem, look_up_to, praise, respect, revere, venerate
To_love
admire, adore, affection, ardour, attachment, baby, be_crazy_about, be_gone_on, be_keen_on, be_mad_about, be_nuts_about, be_partial_to, be_wild_about, beau, being_in_love, beloved, cherish, crush, darling, dear, dearest, desire, devotion, dig,

[117] Owing to the international aspect of the semiometric sample surveys, the order of the words had to be the same for each language. This order corresponds to the alphabetical order of words in the original French questionnaire. See Appendix 2 for five versions of the questionnaire.

dote_on, endearment, enjoy, fancy_man, fancy_woman, flame, fondness, heart, honey, infatuation, like, like_to, liking, love_to, loved_one, lover, mistress, paramour, passion, pet, prize, regard, relish, suitor, swain, sweetheart, sweetie, tenderness, treasure

Ambition

aim, aspiration, craving, desire, dream, eagerness, goal, intent, longing, objective, purpose, pursuit, target, yearning

Soul

being, creature, essence, heart, intellect, life, mind, psyche, quintessence, spirit

Friendship

affection, amity, association, attachment, closeness, companionship, familiarity, fellowship, fondness, friendliness, friendship, harmony, intimacy, society

Anguish

agony, distress, dolour, grief, heartache, hurt, misery, pain, severe suffering, severe sufferings, sorrow, suffering, torment, torture, woe

Using a method commonly applied to statistical text analysis (correspondence analysis of lexical tables, see appendix A1.4), one can construct a table of distances between words, and plot these distances as proximities between points.

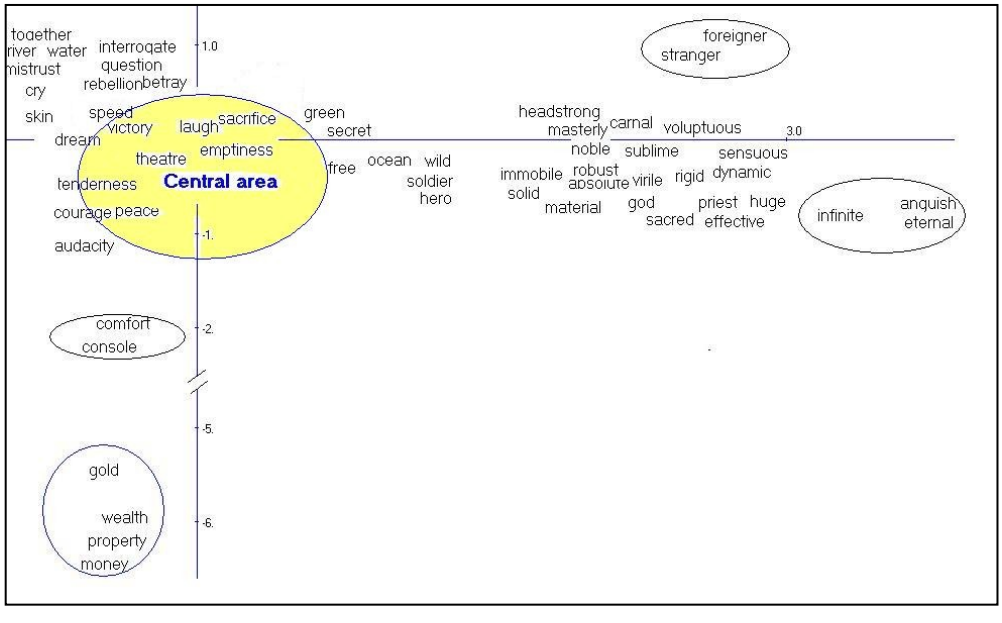


Figure 5.1 Association between semiometric words according to their synonyms

Sketch of the principal plane (axes 1 and 2) from a correspondence analysis of the lexical table cross-tabulating the 210 words (rows) and their synonyms (columns)

Figure 5.1 shows that we no longer observe regular, balanced clouds of points like those of the semiometric graphs presented in the previous sections, but clumps of words that are opposed to all the others.

In fact, the semantic distances derived from a mere list of synonyms are rather coarse. In the semiometric questionnaire, it is the most common meaning of a word which is proposed to the respondent, whereas a thesaurus strives to include all the possible meanings, and sometimes mentions semantic neighbours that are almost homonyms. For instance, we find in our thesaurus *stone* as a synonym of *jewel*, or likewise *cobalt* and *low* as synonyms of *blue,* but also *urine* as a synonym of *water*, *manure* as a synonym of *fertile*, etc.

But the most striking difference between semiometric axes (as described in chapters 1 and 2) and the dimensions derived from the analysis of thesauri concerns the extremities of the axes. We observe meaningful bipolar oppositions in the case of semiometry, and nothing of the kind when dealing with synonyms.

This appears to be due to the non-transitivity of semantic similarities, and the complexity of the notion of semantic distance. Non-transitivity can be expressed as follows: if word A has a similar meaning to word B, and if word B has a meaning similar to word C, then A does not necessarily have a similar meaning to word C. Iterating this process, one can even quite quickly discover an antonym of A[118]. Thus, two words far from each other in figure 5.1 have not necessarily opposite meanings. The confrontation of semantic links taken from a dictionary with those we can derive from semiometric scoring helps us to understand the nature of the statistical structure that we observe.

Note that there are, among the 210 words of Semiometry, words that have no semantic neighbours in common with the others (in fact with more than two others) such as the words: *Moon, Tree , Rifle, Island.* Understandably, if we take the example of *Moon*, its semantic neighbours in the used thesaurus (*daydream, lunar_month, lunation, moon_around, moon_on, moonlight, moonshine, satellite, synodic_month*), are too specific, and unlikely to be among the neighbours, even distant neighbours,

[118] Thus, *abandon* is a semantic neighbour of *give*, *give* a neighbour of *share*, which is itself a neighbour, of *join* which can be considered as an antonym of *abandon*. There are generally longer chains, simply leading to semantic indifference. As the sequence: *literature - poetry - music - vibration*. The first and last term of the chain do not have - or have little - relationship, without being opposites in any way.

of the other words. The same goes for *Rifle,* the semantic neighbours of which being: (*despoil, firearm, foray, go, gun, loot, pillage, plunder, ransack, reave, rob, scattergun, shotgun, strip*). In the case of semiometric surveys, the correlations observed between these words and others are not semantic entities; they belong to a much more general perceptual context.

Are synonymous words over-weighted in semiometry ?

There are, however, within the semiometric list of words, words that have incontestable semantic links between one another. Those words will be simultaneously close on semiometric maps and in semantic graphs[119].

Thus the words *Carnal, Sensual, Voluptuous* contained in the semiometric questionnaire are semantic neighbours of one another, and have several neighbours together in the semantic thesaurus that we used (e.g. *animal, bestial, erotic, lascivious, libidinous, lustful*). However, this is an exception in the questionnaire. There are no other groups of three words as close semantically as they are.

One could then argue that this partial redundancy (partial because those words still have clearly distinct meanings) gives undue weight to the underlying concepts.

We may then ask the question: Is this group of three words responsible for some of the structural features observed and described in the previous chapter?

This is not the case, as is easily verified: the basic axes the stability of which we have described in chapters 1 and 2 are not changed if we remove from the score file, two out of the three words in question. Even more striking, the first six axes are preserved if we delete from the list all three words *Carnal, Sensual* and *Voluptuous*.

The second axis, the one the most correlated to the three words deleted, and thus the most likely to be affected by this deletion, is surprisingly stable, as evidenced in Table 5.2, which shows the most extreme words on the axis. In both cases, these lists are the opposite of *Homeland, Money, Property, Soldier, Rigid, To economize, To obey, Discipline*.

[119] The tool chosen to give a pleasant visual assessment of semantic graphs is again the Kohonen self-organizing map. See section 3 of this chapter, and Appendix A1.8 for details on this *neural* method and justification of its use in this context.

Table 5.2: List of the most positively correlated words with the second axis

With 210 words	*With 207 words*
Art	*Art*
Bohemian	*Bohemian*
Book	*Book*
Original	*Original*
Sensual	*Poetry*
Poetry	*Storm*
Rebellion	*Theater*
Voluptuous	*Music*
Storm	*Rebellion*
Wild	*Ocean*
Ocean	*Wild*

Thus, semantic proximity, particularly important in the case of these three words, seems to play a secondary role in the establishment of a stable underlying structure.

Another example of semantic proximity not conveyed by Semiometric proximity: the word *Animal*, which we saw was among the synonyms of *Carnal*[120], *Sensual, Voluptuous*. *Animal* is not a neighbour of these words in the semiometric planes. It moves away especially along the second axis "Duty / Pleasure". It must be said here that there is ambiguity between the noun *animal* and the adjective *animal*, which can indisputably have different synonyms. The semiometric questionnaire does indeed specify "*An animal*".

To conclude this subsection, we note that:

– the semantic structure of the semiometric list such as it is described by a *thesaurus* (instead of a set of *responses to a questionnaire*) does not give rise to stable axes (or stable principal directions).

– locally, there are semantic associations on semiometric maps, but there are notable exceptions.

– the large stable oppositions observed in the semiometric field have a psycho-sociological or even a marked socio-demographic character (with links to certain axes of sex, age, occupational groups), and do not belong to the register of synonyms-antonyms.

[120] The word *Carnal* does not appear in Table 5.2 because it only occupies the thirteenth position on the second axis.

5.2 A semantic network of French verbs

The second experiment, unlike the previous, is independent of the corpus of 210 words. It deals with all the common French verbs (the 829 most frequent verbs in the "Bescherelle"[121] classical grammar book, which we will now consider).

The question is much broader and more ambitious: does a structure emerge, one similar to that observed during the semiometric analysis when we were describing the semantic distances between all the verbs of the French language?

Each word will be further described by all of its synonyms, which will help determine distances between verbs and thus we will be able to observe the structure induced by these distances.

Why did we choose the verbs? We did so because they constitute a much smaller and homogeneous set than the set of nouns or adjectives, and therefore it is possible to be (almost) exhaustive.

The results confirm those observed in the preceding paragraph: the semantic structure is essentially local, and “the large semantic distances” do not give rise to stable axes, or interpretable oppositions.

The corpus of verbs and their synonyms (similar to the text in table 5.2 above, but on this occasion related to 829 verbs) comprises 17,446 occurrences of words (which are obviously all verbs).

The same text contains 3,839 distinct verbs. This number is much larger than the initial 829 verbs because verbs, less frequently used, may appear as synonyms of the verbs given by the Bescherelle grammar. Table 5.3 provides some examples: synonyms selected from those of the verb ***to cut.***

Two verbs will be all the closer if they have many synonyms in common. Obviously, many pairs of verbs will have no synonym in common, and the distances between these pairs will be equivalent.

[121] Bescherelle, *Conjugation, Dictionary of 12,000 verbs*, Hatier, Paris, 1990.

Table 5.3: Synonyms[122] of the verb: *to cut*

abbreviate, abridge, axe, baseball swing, bisect, bite, break, bring down, carve, carve out, chisel, chop, chop off, contract, crop, cross, curtail, cut away, cut back, cut back on, cut down, cut down on, cut off, cut out, cut-back, cutback, decrease, dilute, diminish, diminishing, dip, divide, down, engrave, etch, excise, fell, foreshorten, garb, gash, gather, gouge, groove, hack, hew, hit, hurt, incise, intersect, knife, lacerate, lessen, lop, lop off, lower, mangle, mow, nick, notch, pare, pare down, part, penetrate, pierce, portion, proportion, prune, puncture, rationalize, reap, reduce, restrict, rip, roll back, scale down, score, scratch, sever, shape, share, shave, sheer, shorten, slash, slew, slice, slit, slue, snip, snip off, split, stab, sting, style, swerve, swing, tailor, thin, thin out, trend, trim, trim back, trim down, truncate, veer, weaken, whittle, wound

As is usual in statistical textual analysis, it is necessary to introduce a threshold for the minimum frequency of words.

Table 5.4 gives the most frequent verbs in the corpus, now limited to the verbs appearing at least 20 times[123].

First of all, we will only retain here those verbs appearing at least four times in the corpus of all the verbs and their synonyms. This leaves 1619 distinct words, and reduced to a lesser extent, the size of the overall corpus, which now has 13,660 occurrences. The minimum frequency of 4 still leaves in the corpus some verbs that are relatively uncommon.

To get a quick overview of the semantic graph of the verbs of the French language, we are finally led to perform a correspondence analysis (see Annex A1.4) of a binary array with 1619 rows and 829 columns[124].

Having chosen 4 as the frequency threshold, we can see that the group of words (*vibrate, agitate, shake, shiver, tremble, shudder...*) is opposed to the group (*scream, moan, yell out, summon, confirm, promise, notify, corroborate, attest, witness, narrate, relate, proclaim, recite...*) on the first axis, and is opposed to the group (*demoralize, exhaust, dull, tired,*

[122] It does not make sense to translate a list of synonyms from one language to another. The same verb could be repeated several times. This list is then obtained from an English thesaurus.

[123] The complete table is available from the authors.

[124] The word located at the intersection of row *i* and column *j* of this table is *1* if word i is one of the synonyms of word *j*, and is *0* otherwise.

massacre, darken, deface, despair, grieve, blacken, sorry, disgrace, corrupt, contaminate...) on the second axis.

Table 5.4: Excerpt from the list of verbs

(listed here by decreasing frequency (50 to 20) of the number of synonyms)

Verbs	*frequencies*	*Verbs*	*frequencies*	*Verbs*	*frequencies*
remove	50	distinguish	22	delete	22
fix	39	disturb	22	directing	22
take	36	rule	22	settle	21
examine	33	understand	22	guarantee	21
contact	33	drive	22	determine	21
stop	33	encourage	22	draw	21
result	32	provide	22	approve	21
assemble	32	link	22	mark	21
push	31	rule	22	raise	21
support	31	yield	22	declare	20
stand for	30	stir	25	attack	20
rise	30	destroy	25	return	20
show	28	release	25	train	20
protect	28	tie	24	develop	20
maintain	28	kill	24	reverse	20
admit	28	indicate	24	unite	20
available	28	enter	24	imagine	20
cut	28	observe	24	enjoy	20
cover	28	remember	24	estimate	20
exhibit	27	subtract	24	hold	20
arrange	27	crush	24	combine	20
receive	26	excite	24	grant	20
meet	26	collect	23	educate	20
win	26	establish	23		
support	26	associate	22		

By taking a new threshold with a minimum frequency of five (verbs with less than five occurrences in the corpus are removed), we have 1265 distinct verbs left in a corpus composed of 12,244 occurrences.

With this new threshold of five, we have the group of words (*wish, dream, hope, aspire, pray, stand, beg, implore, ask...*) opposed to two new groups on the first two axes: the group - less homogeneous - (*eat, despair, blacken, cool discourage, tire, depress, dismantle, deface, shave...*) on the first axis.

The same word group is opposed to the group (*agglutinate, correlate, aggregate, combine, conjoin, weld, juxtapose, connect, moister, attach, marry, braze*...) on the second axis.

We could still increment the frequency threshold to show that we continue to obtain triangular structures (one group opposed to two in the first factorial plane engendered by the first two axes), and, moreover, these are unstable structures, since the groups characteristic of the principal axes change when the threshold of minimum frequency changes.

In fact, the geometric analysis of the multidimensional cloud of points of the verbs shows that this cloud is nearly spherical[125]. This quasi-sphere includes, at the periphery, "lumps" which are clusters of semantically related verbs. These "lumps" create the principal axes according to their size, which depends also on the minimum frequency threshold chosen initially. Such a shape is therefore very far from the elliptical structure of the cloud of words in Semiometry[126], with its stable and significantly different principal axes.

For thresholds of about 12 to 18, i.e. for verbs appearing twelve to eighteen times as synonymous with other verbs of the French language[127], we clearly witness the aforementioned groups; however, the major oppositions observed on the principal axes are disappointing and unexpected.

Thus, for several of these thresholds we can see appearing at the end of one of the axes, the group:

[*conjoin, attach, marry, couple, combine, unite, connect, bring together, collect, assemble, associate, match, join, mix, combine, link, bind, blend*].

But this group of verbs, relatively homogeneous and quite well related to the pole "Attachment" of the third axis of Semiometry, is often opposed to the group:

[*claim, demand, desire, request, aspire, seek, explore, inspect, probe, scan, search, call, think, consider, appreciate*].

---

[125] This "sphericity" is especially true because of the similarity of the first *eigenvalues* whose percentages of variance are worth, for the first five: 0.77, 0.73, 0.69, 0.68, 0.66 (in the case of the analysis with threshold 4).

[126] Remember that the statistical significance of the Semiometric axes was tested using (1) *Anderson confidence intervals*, (2) repetitions of independent samples in time and space, and (3) "bootstrap" methods, cf. Chapter 2.

[127] With thresholds so high, there is no question of seeing rare verbs.

It can never be found opposed to the group, frequent yet more antonymous than the previous one:

[*Identify, separate, distinguish, discern, abandon, remove, return, reject, remove, extend*].

In fact, we rejoin some conclusions of the previous section. The notion of synonymy such as it is treated in dictionaries or thesauri, takes us away from the spirit of the semiometric questionnaire, which asked whether the words presented evoked pleasant or unpleasant sensations. Both the very *context* of the questionnaire (a deliberately diverse list of words) and the *time* available for its completion involve taking the suggested words in their most common acceptations.

On the contrary, a thesaurus strives to be comprehensive and not to omit the secondary meanings, metaphorical, familiar or slang.

Besides the statistical analysis, let us illustrate these difficulties around the concept of semantic distance with a reflection of Gaston Bachelard, evoked by Roland Barthes in his *Mythologies*, in which, somehow, "wine is the opposite of water .... " The distance here is more axiological than semantics, and water may be the opposite of drought for a farmer, of ice for an Eskimo, of solid for a physicist, of earth for a sailor, or of wine for the average French whose values and myths are humorously decoded by Barthes in the late fifties[128].

Similarly, a simple proverb as "the perfect is the enemy of the good"[129] owes its strength to the somewhat paradoxical challenge of some evident semantic proximities[130].

About the experience on all the French verbs and its great ambition: to discover the semantic network and to relate it to the stable structure observed in semiometry, we must also see, as could be foreseen, that the verbs expressing values are very infrequent in all 829 verbs chosen. By

---

[128] « Bachelard has already given a 'substantial psychoanalysis' of this liquid (wine) at the end of his essay on the "Dreams of the will" [Rêveries de la volonté], showing that wine is somewhat 'the juice of sun and earth', that the basic context is not the 'wet', but 'the dry', and as such, the mythical substance which is the most opposite to wine is water. » (Roland Barthes, *Mythologies*, Seuil, 1957).

[129] "the better is the enemy of the good" would be a more accurate translation of the original version attributed to Voltaire ["Le mieux est l'ennemi du bien"]

[130] Many style clauses exploit this topological complexity of semantic space, which, for statisticians, more prosaically, is a space that can not easily be provided with a Euclidean distance.

nature, the verb is related to the action, movement, time, change. In our corpus of synonyms, technical or descriptive verbs abound and somehow overshadow more abstract verbs linked to emotions or relationships *(love, judge, know, care for, admire, ban …*).

Finally, we conclude with a remark that applies to both the two previous experiments: whether dealing with nouns, adjectives or verbs, the semantic network is not homogeneous[131].

Some words have many synonyms, others have little or not at all. A word without synonym is no less important than another.

In a later phase of research, the previous experiments could be extended, beyond synonymy, to corpora based on analogy or on dictionary definitions and collections of quotations.

5.3 "Internal semantic fields"

We go back in this section to classical semiometric results exemplified in chapters 1 and 2. We shall call *internal semantic field* for a specified word all the words (taken within the list of 210 words chosen *a priori*, hence the adjective *internal*) that are significantly correlated with this given word. The distance used is all the smaller as the correlation coefficient between the two words is high[132]. This denotation ("Internal semantic fields") is justified by the *a posteriori* interpretation of the proximity observed. It was not clear whether scores based only on pleasure (pleasant, agreeable) or displeasure (unpleasant, disagreeable) generated semantic proximity. This section shows the consistency and fineness of the proximity observed. Some words have a rich and dense internal semantic field, defined by high correlations with many other words in the list. This is the case for the words *Effective* and *Courage*. Whereas words such as *Mountain*, *Voluptuous*, and *Mystery*, close to a small number of words, have a narrower scope.

A single word can obviously, within certain limits, be correlated with words that are not correlated among one another. Thus the semantic fields of *Courage* and *Effective* are not the same despite their strong correlation

---

[131] In more technical terms, but more precise, the semantic graph is far from being a regular graph, its vertices (the verbs) have widely varying degrees, some vertices being even isolated.

[132] The square of the distance between two words *a* and *b* whose correlation coefficient is *r(a, b)* is written $d^2 (a, b) = (1 - r (a, b))$.

together with the membership of certain words such as *Robust* and *Dynamic* in both their fields.

Semantic field of ***Effective***

Accuracy (.51)[133]*, Concrete (.40), Robust (.40), Courage (.39) Cunning (.38), To master (.38), Dynamic (.37), Reason (.37), Confidence (.35), To produce (.35), Practical (.35), Certainty (.34), Honest (.34), Headstrong (.34), To build (.33), Logic (.32), Reward (.32), To ponder (.32), Respect (.32) Entrepreneur (.31), Firmness (.31), Honour (.31), Softness (.30).*

Semantic field of ***Courage***

Honest. (.41), Dynamic (.40), Respect (.40), Practical (.39), Confidence (.38), Honour (.38), Elegance (.36), Politeness (.36), Friendship (. 35), To recover (.35), Robust (.35), Charitable (.34), Accuracy (.33), Headstrong (.33), Softness (.32), To protect (.32), Purity (.32), Reason (.32), Liveliness (.31), Reward (.31), Certainty (.30).

Semantic field of ***Voluptuous***

Sensual (.47), Carnal (.45), Desire (.34), Mellow, (.34) Nudity (.32), To seduce (.31), Flexibility (.31), Sublime (.30), Feminine (.25), To dream (.24), Cunning (.23), Audacity (.23), Softness (.23), Intimate (.23), Skin (.23), Caress (.22), Lightness (.22), Original (.22), Art (.20) Elegance (.20).

Semantic field of ***Mountain***

Peak (.47), To climb (.34), Countryside (.33), River (.32), Tree (.31), Ocean (.31), Island (.26), Moon (.26), Flower (.24), Music (.23), Water (.22), Animal (.21), Blue (.21), To build (.21), Nest (.21), Softness (.20).

Semantic field of ***Mystery***

Storm (.26), Magic (.25), Adventurer (.24), Emotion (.24), Original (.24), Secret (.24), Wild (.23), Infinite (.22), Fire (.21), Ocean (.20).

Semantic field of the word ***God***

Faith (.74), Priest (.65), Meditation (.60), Sacred (.55), Soul (.48), Creator (.33), Homeland (.32), Forgiveness (.29), Charitable (.27), Morals (.27), Hero (.26), Honour (.26), Noble (.26), Ceremony (.25), Eternal (.25),

[133] The quantities in parentheses are correlation coefficients with the word in bold. Note that all correlation coefficients cited are statistically highly significant given the sample size (11,055).

Tradition (.25), Wedding (.23), Sovereign (.22), Family (.21), Purity (.21), Discipline (.20), Elite (.20).

This first selection of some internal semantic fields may be generalized to all the words in the questionnaire. A possible generalization is obtained through an automatic clustering of the words, using the previously defined distances between all pairs of words. However, in this case, the most popular algorithms do not allow a word to belong to several clusters.

5.3.1 Hierarchical representation of semantic proximity

Words that are most alike when it comes to giving them a score, and consequently connoting the way they are felt, will now be grouped automatically. The method of principal components analysis presented in chapter one makes such an overall representation of correlations, and therefore proximities.

But the two classification methods applied here, *hierarchical classification*[134], and Kohonen *self-organizing maps*[135], describe perhaps with more finesse local proximity, which is, as shown in the previous section of this chapter, the only one to be interpreted in semantic terms.

To do this, we shall first of all carry out a *hierarchical clustering* of the 210 words provided with the distances defined above. The principle of this process (the algorithm) is to group the words in pairs by gradual agglomeration, thus providing a hierarchy of *partitions*.

We retain here the partitions nested in 12, 24 and 36 classes, represented by Table 5.5 which maps the *dendrogram* of the hierarchy. The results speak for themselves. The resulting classes group together words that are not true synonyms[136], but belong to the same *semantic halo*.

Consider, for example, the partition in 12 classes (the coarsest of the three partitions) whose classes are identified in the first column of Table 5.5 by the symbols 1/12, 2/12, (...) 12/12. Grouped in this partition are the words relating to the concept "Sublimation" (class 1/12: *Absolute, Huge, Infinite, To admire, To love, Secret, Sublime…).*

---

[134] The principle of the methods of hierarchical classification is briefly presented in Appendix A1.7.

[135] The principle of self-organizing maps is briefly presented in Appendix A1.8.

[136] The list of 210 semiometric words excluded *a priori* the presence of synonyms that would have introduced unnecessary redundancy.

Other classes correspond to topics like the art of writing (class 3/12: *To write, Book, Poetry, Theatre*), nature (class 4/12; *To climb, Peak, Island, Animal, Green*…), emotion (Class 5/12: *Emotion*, *To dream, Humour, Red, Wander, Rebellion*).

Observe that we find, in the 28/36 class, the semantic field of the word *Effective*. The word *Courage* does not belong here, and joins the words: *Respect, Politeness, Honest*.

5.3.2 Planar representation of semantic proximity

The representation of the classes nested in Table 5.5 does not allow us to appreciate the relative distances between clusters of words obtained by hierarchical classification.

A Kohonen map (Figure 5.2) is a way of visualizing data that groups the words in classes while keeping as much as possible the initial topology of the word-space. It thus reflects the proximity between classes that are similar (see Annex A1.8).

Thus, the words *Poetry* and *Theatre* are close to the following groups: (*Nest, Book, River, To write, Art)*; (*Green, Sublime, To swim, Music, Mountain, Water, Blue);* (*Peak, Secret, Moon, Huge*) and (*Justice, Meditation, Creator, Soul*). (See Table 5.1).

Table5.5: Hierarchical clustering in 12, 24 and 36 nested clusters

1/12	1/24	1/36	Absolute, Huge, Infinite
		2/36	To admire, To love, Eternal, Precious, Secret, Sublime
2/12	2/24	3/36	Softness, Feminine, Intimacy, Mellow, Skin, Flexibility
		4/36	Caress, Desire, To seduce
	3/24	5/36	Carnal, Nudity, Sensual, Voluptuous
3/12	4/24	6/36	Art, To write, Book, Poetry, Theater
4/12	5/24	7/36	To climb, Island, Moon, Mountain, Ocean, Peak
	6/24	8/36	Animal, Tree, Blue, Countryside, Water, Flower, River, To swim, Nest, Green
5/12	7/24	9/36	Emotion, Escape, Lightness, Magic, Original, To dream
		10/36	Humour, Game, Music, Red
		11/36	Wander, Change, Different, Foreigner, Stranger
	8/24	12/36	Persistence, Adventurer, Challenge, Craftiness, Speed
	9/24	13/36	Desert, Fire, Mystery, Black, Storm, Wild
		14/36	To criticize, Disorder, Irony, Rebellion
6/12	10/24	15/36	War, To betray
		16/36	Anguish, Cry, Danger, Detachment, Doubt, Fault, Maze, Mistrust, Death, To break, Emptiness, To age
	11/24	17/36	Armour, To attack, Hunt, Rifle
	12/24	18/36	Immobile, Mask, Wall, Knot
		19/36	Border, To forbid, Metallic, To punish, Rigid, Sacrifice
7/12	13/24	20/36	Jewel, Elegance, Slimness, Mode, Perfume
		21/36	To buy, Present, Comfort, Free, House, Reward
	14/24	22/36	Ambition, To conquer, Glory, Power, Victory
		23/36	Money, To inherit, Gold, Property, Wealth
8/12	15/24	24/36	Noble, Perfection, Refined
		25/36	To command, Elite, Hero, Honor, Homeland, Soldier, Masterly, Tradition, Virile
	16/24	26/36	Trade, Industry, Material, To produce, Practical
	17/24	27/36	Discipline, To economize, Firmness, Law, Morals, To obey, Rule
9/12	18/24	28/36	Clever, Audacity, Certainty, Solid, Dynamic, Effective, Logic, To master, Accuracy, Reason, Robust, Headstrong
	19/24	29/36	Effort, To interrogate, Question, To ponder
	20/24	30/36	Entrepreneur, Researcher, To build, School, To teach, Inventor, Science, Work
10/12	21/24	31/36	Soul, Creator, God, Faith, Priest, Meditation, Sacred
11/12	22/24	32/36	Childhood, Together, To fertilize, Maternal
		33/36	Ceremony, Family, Wedding, Birth
12/12	23/24	34/36	Courage, Loyalty, Honest, Politeness, Respect
		35/36	Friendship, Confidence, Liveliness, To recover, Peace, To laugh, Tenderness
	24/24	36/36	Attachment, Charitable, To console, Humble, Justice, Moderation, Modesty, Forgiveness, Patience, To protect, Prudence, Purity, To tend

Wealth Gold To inherit Glory Money	Reward Precious Mode Slimness Free Elegance Comfort Present To love To buy	House To fertilize Softness Family Childhood Confidence Countryside Attachment Friendship	Tenderness To protect Peace Birth Maternal To recover Liveliness Flower Loyalty	To tend Respect Prudence Politeness Patience Humble Honest To console	Work Rule Moderation Law Effort School
Virile Victory Power Property To conquer Trade	Certainty Ambition	Refined Wedding Ceremony To admire	Purity Forgiveness Charitable	Modesty To teach Courage	Practical Reason To produce Accuracy Logic Firmness
Soldier Material Industry To economize Discipline To command	Tradition Masterly Perfection Homeland Honor Hero Elite	Sacred Priest Noble Faith Eternal God	Meditation Creator Soul	Justice Effective To build Solid Researcher Entrepreneur	Headstrong Science Robust To ponder To master Inventor
Sacrifice Rigid To obey Metallic To forbid	To interrogate Border Hunt	Absolute	Theater Poetry	Nest Book River To write Art	Dynamic Audacity Clever Tree
To break To punish Knot Wall Mistrust Immobile Rifle	Question Mask Detachment Persistence Cry Armour	Red Infinite Stranger Foreigner To climb	Peak Secret Moon Huge Animal	Green Sublime To swim Music Mountain Water Blue	Flexibility Skin Game Intimate Humour Feminine Together
To age Emptiness To betray Death Maze War Fault Doubt Danger To attack Anguish	Craftiness Rebellion Storm Black Irony Disorder Desert critize To	Speed Wild Mystery Fire Different Challenge Change	Original Nudity Magic Lightness Escape Emotion Bohemian Adventurer	Voluptuous Sensual Ocean Island Carnal	To seduce To laugh To dream Perfume Mellow Desire Caress

Figure 5.1: Proximity between words described by a Kohonen map
(or: Self-Organizing Map)

The class "*Meditation, Creator, Soul*" is itself close to "*Modesty, To teach, Courage*" and "*Purity, Forgiveness, Charitable*", "*Refined, Wedding, Ceremony, To admire*", "*Sacred, Priest, Faith*"

The classes are probably less homogeneous and sometimes less consistent than *hierarchical partitions*; nevertheless, representation by *self-*

organizing Kohonen maps lays greater emphasis on consistency between classes.

Kohonen maps (or Self Organizing Maps) have, however, the disadvantage of being more unstable than hierarchical clustering and are better adapted to sets of unstructured data. They are better suited to open-ended questionnaires than to the analysis of already highly structured corpuses of words.

5.3.3 Semantic nuances provided by the Semiometric approach

On the one hand the local texture of the geometrical shape of the semiometric cloud of words is compatible and consistent with the most usual semantic links, on the other the global structure contains new pieces of information likely to enrich a semantic analysis of the language.

The fact that a questionnaire based on hedonic aspects (pleasure) can lead to semantic subtlety almost came as a surprise, and remains difficult to explain.

Let us take an example of the enrichment allowed by the semiometric analysis. There are pairs of words, such as *Irony* and *Humour*, which have many common semantic neighbours (*mind, caustic, malice, wit, witty, funny, mischievous, malicious, sharp, satirical, mocking, projection, satire*) and are therefore close semantically, words which Semiometry separates significantly. Thus, on the third axis "Attachment / Detachment", *Irony* is on the side of "Detachment" an "*Humour*" on the side of "Attachment" (table 5.6).

Humour and irony are two distinct ways to practice mockery. Humour is soft (*Softness, Tenderness*), the irony is abrasive (*Savage, Disorder*). The humour is used to create links (*Friendship, Present, Trust, Caress*), irony is used to break links (*Critical, Revolt, To attack, Detachment, Change, Craftiness*). Humour is a game (*To laugh, Game, To seduce*), irony is a drama (*Danger, Mystery, Adventurer, Challenge, Black*).

Let us recall that a correlation matrix relative to 210 variables (words) contains 21,945 coefficients (21,945 = 210 x 209 /2). We have at our disposal a stock of ammunition (so to speak) to repeat the previous analysis (limited here to the pair: *Irony* - *Humour*) for every pairs of words! The "semantic halo" mentioned above is in fact multidimensional: to the somewhat binary notion of synonymy is substituted a complex "semantic

vicinity" involving the atmosphere, the psychological, social or moral context, the usefulness, the appearance, and, eventually, the emotions.

Table 5.6: 25 most strongly correlated words with *Humour* and *Irony*

Humour	***Irony***
To laugh (47.16)	*Fire (28.77)*
Liveliness (39.20)	*To criticize (27.61)*
Original (29.85)	*Rebellion (26.74)*
Sensual (29.49)	*Original (24.09)*
Dynamic (28.80)	*Storm (23.80)*
To dream (28.42)	*Humour (20.36)*
Clever (28.39)	*Audacity (19.69)*
Friendship (27.75)	*Danger (19.46)*
Desire (27.15)	*Different (19.12)*
Softness (27.11)	*Wild (18.53)*
To seduce (25.72)	*Mystery (18.4)*
Music (25.57)	*Voluptuous (17.96)*
Tenderness (25.52)	*Disorder (17,89)*
Game (25.24)	*To ponder (17.34)*
Present (23.59)	*Nudity (17.08)*
Flexibility (23.08)	*To attack (16.25)*
Confidence (23.01)	*Emotion (15.95)*
Feminine (22.91)	*Detachment (15.87)*
Ocean (22.65)	*Change (15.68)*
Emotion (22.50)	*Craftiness (15.50)*
Tree (22.46)	*Adventurer (15.03)*
Caress (22.10)	*Black (15.01)*
Headstrong (22.04)	*Challenge (14.98)*
Mellow (22.02)	*Sensual (14.78)*

Conclusion of section 5.3

The semiometric structure is only very partially reduced to a semantic structure. Locally, around a point representing a word, we can find semantic neighbours, but the major oppositions responsible for the stable axes described in the previous chapters are not observed spontaneously from simply analyzing semantic proximities.

However, the *internal semantic fields* provided by the semiometric analyses provide us with elements of interpretation that go far beyond the usual tools of semantic investigations.

5.4 The two facets of meaning: denotation and connotation.

This section reflects on the results of the three preceding sections to analyze et propose a tentative interpretation of the semiometric space[137].

According to linguistic theory, a word has two types of meanings: the denotative meaning, neutral and objective, given by the dictionaries, and the connotative meaning, emotional and subjective, which results from the various evocations of the word.

Example:

The word Sea denotes *a vast expanse of salt water* whereas it connotes *vastness, freedom, adventure, depth, storm, shipwreck*, etc..

Because it measures the emotional charge contained in the words, Semiometry is based on the connotative meaning of words rather than their denotative meaning. This explains why the semantic space reconstructed from a thesaurus, which is based essentially on the denotative meaning of words, has little similarity with that shown by Semiometry.

Besides the distinction between *positive connotation* (producing pleasant sensations) and *negative connotatio*n (entailing an unpleasant effect on the respondent), we must then consider also *collective connotations* (closely related to the cultural environment of the respondent) and *individual connotation* (depending on the personal experience of the respondent).

Going back to the example of the word Sea: 1) Whereas freedom is generally a positive notion, shipwreck is a negative one. 2) If deepness is a widespread connotation of the word Sea, the connotations of this word is not the same for a fisherman who must spend the year at sea for a living whatever the weather, and for a sailor who is temporarily yachting whenever the weather is fine. Similarly, White connotes “purity” in Western civilization, while it connotes “grief” for several Asian countries. To sum up, a connotation can be positive or negative, collective or individual.

---

[137] This section, like the forthcoming interpretative Chapter 6, has been written mainly by Jean-François Steiner.

5.4.1 Mental mechanisms implemented by the respondents as they complete the semiometric questionnaire.

A posteriori interrogations of respondents[138] show that, when asked whether a word is pleasant or unpleasant, they identify at first, subconsciously, and almost instantly, all acquired notions and all situations (be it facts or fantasies) that the word evokes. Each situation is then assigned an emotional score, positive, negative or neutral depending on whether these evocations are pleasant, unpleasant or indifferent. In a second phase, they sum up all these feelings in an "average sense", and finally in a third phase, they express the emotional charge of the latter as a score from -3 to 3. This score is -3 if the "average feeling" is totally negative, 3 if it is totally positive, or else it could be an intermediate score depending on the intensity of sensation. This score will be 0 if the evocation of the connotations of the word does not cause any sensation, or if it generates contradictory feelings between which they can not decide, or even if the evocation, because of its traumatic nature, triggers a phenomenon of censorship.

Example: the word *Death*.

If everyone agrees on the denotation of the word (its neutral meaning given by dictionaries of the language, for example, according to the Oxford Dictionary *The action or act of dying or being killed* [139]), in general, in our

---

[138] The question asked was: Why did you put such a score to this word? Some respondents were able to justify it immediately, explaining what the word reminded them or what particular particular image was associated to it. Example: the verb *To S*wim (score = 2) : “Swimming is good for health”. "I love swimming in the sea when the weather is nice and is quiet but I hate pools, they smell and there are people that will fall on your head or splash you. But I never go to the pool”. Others could not be more specific, but in most cases, it was enough to raise one or two other issues like" Why did you put 2 and not 3? ". For extreme scores, especially negative responses, the answers were easier. Example, the word War (for a score = -3): "Because it kills innocent people and powerful people may get some benefit. Moreover, it is terrible to risk its own life and to be bound to kill" It might be objected that this is a phenomenon of post hoc rationalization, but rationalization itself is often a psychological defense mechanism based on explanations and never on evocations (similar to the *free associations* of psycho-analysis).

[139] This lack of ambiguity about the denotative meaning of words is a very important point. Indeed, if a word has several denotative meanings, it necessarily refers to emotional experiences of different nature. The respondent is then confused, and the response is blurred. Distinct respondents may also allude to different directions, entailing non-

society today, the word is associated with the notion of loss, sadness, grief, etc.. , But for men who suffer, for whom life on earth is a *valley of tears*, it is often associated with the concept of peace and rest, while for some believers or those who worship entities such as fatherland, revolution or any other cause to which they attach more value than to their lives, it may be associated with the notion of glory (cf. Among many other examples, the song of the *Girondins* at the time of the French Revolution: "Dying for our own country is the most beautiful fate, most envied "). So the former attributes the score -3 to the word Death and the latter a positive score, up to 3. But for the experienced woman/man who has thought, suffered and hoped, Death may evoke *loss*, *peace* and *glory*. To account for this plurivalence she or he will assign a score close to 0.

5.4.2 Construction of semantic space: connotations and correlations.

We see that what we have called the semantic field of a word forms a set sometimes quite disparate in terms of denotative meanings, but semantically coherent in terms of connotative meanings.

Examples:

War: Gun, Death, Attack, Betrayal, Revolt, Danger ...

Peace: Mountain, Politeness, Attachment, Victory, Book,...

Escape: Wild, Bohemian Revolt, Storm, Fire, Adventurer ...

Ambition: Certainty, Power, To conquer, Glory, Wealth, Property ...

Science: Researcher, Inventor, Industry, School, Logic, Builder ...

Change: Unknown, Foreign, Different, Original, Adventure, Challenge ...

It is obvious that for a given individual, these words could evoke other ones, either because they are not part of the *Semiometric* corpus, or because that individual has a sensitivity different from that of the sample of

---

comparable scores. This is why avoiding the presence of polysemous words has been a major concern in constructing the questionnaire.

respondents. But it is equally clear that these over-rated words can legitimately be seen as connotations of pivotal words.

We note, moreover, that if two words, for a sufficient number of respondents, share several connotations of similar sign (all positive or all negative), they correlate positively. Similary, we note that if two words have different connotations and provoke opposite feelings they correlate negatively, whereas if two words have little or no common connotations, they do not correlate at all.

Examples:

• Words *War* and *Gun*, which correlate to .24, have 13 common connotations*: Armor, Attack, The hunt, Danger, Fire, Mistrust, Metallic, Death, Storm, To break, Soldier, Craftiness, Speed.*

• The words *Discipline* and *Firmness*, which correlate to .43, have 25 common connotations: *To obey, Order, Morals, To punish, Rule, Law, Work, Industry, To produce, Homeland, Reason, Sacrifice, Elite, School, Border, Soldier, To save, Honor, Rigid, Effort, To interrogate, Accuracy.*

• The words *God* and *Sacred*, which correlate to .63, have 24 common connotations: *Faith, Priest, Meditation, Patrie, Ceremony, Sovereign, Hero, Tradition, Marriage, Honor, Morals, Soldier, Elite, Glory, Family, Discipline, To obey, Purity, Soul, Creator, Forgiveness, Charitable, Noble, Eternal.*

• The words *Construct* and *Builder*, which correlate to .44, have 17 common connotations: *To produce, Homeland, Practical, Material, Rule, School, Morals, Trade, Property, Honest, Creator, To teach, Robust, Inventor, Effective, Solid, Accuracy.*

• The words *War* and *Peace*, which correlate to -0.29, have no common connotation and are connoted by words causing opposing sensations. Connotations of the word *War*: *Gun* (avg. 2.26), *Death* (1.77), *To attack* (2.61), *To betray* (1.42), *Revolt* (3.27), *Danger* (2.58). Connotations of the word *Peace*: *Tenderness* (6.67), *Politeness* (6.24), *Attachment* (6.00), *To protect* (5.95), *Liveliness* (6.46), *Justice* (5.37), *Humour* (6.27), *Faith* (6.48), *Friendship* (6.59)

• The words *Science* and *Gun*, which correlate to 0.00, have only two common connotations, *To produce* and *Accuracy*.

5.4.3 A structured and semantically continuous space.

Local semantic consistency

We have shown (Section 5.3) what we have called the "internal semantic field" of some words, that is to say all the words that are significantly correlated with them. We saw that they had a high semantics affinity, at least in terms of connotative meaning. Experience shows that, whatever the chosen word, words that correlate with it are endowed with the same affinity. The location of a word in the semiometric space being determined by its correlation with the 209 remaining words, we can conclude that around each point of this space exists a local semantic consistency. The question one may ask is whether this creates a continuity of the local semantic consistency throughout the whole space, briefly: is it possible to walk within this space without ever being stopped by a break in continuity of the semantic point of view of connotative meaning. Intuition leads us to think so, but it deserves to be evidenced.

Semantic continuity: semantic chains

The experiment consists in browsing the periphery of planes by jumping from word to word and checking whether a semantic gap is encountered[140].

We will scan clockwise the words located at the periphery of the principal planes with the constraint that the correlation between two words is not less than .12.

Route path (in the French semantic space)

Plane 2 / 3.

Starting from the word Jewel (the top of the vertical axis 3) are: **Jewel**> Gold> Present> Perfume> Caress> To seduce> Desire> Voluptuous> Lightness> Island> Ocean> Nudity> Wander>Adventurer> Wild> Revolt > Storm> Danger> Doubt> Death> Anguish> To break> To punish> Sacrifice> Firmness> Rule> To obey> Discipline> Morals> Faith> God> Honour> Politeness> Family> To inherit> Silver> Rich> Comfort> **Jewel**.

Plane 3 / 4.

Starting from the word Meditation (top of the vertical axis 3) are: **Meditation**> Moderation> Forgiveness> Art> Poetry> Theatre> Music> Peace> Confidence>

[140] The proximities between words located in the periphery of a principal plane correspond to the highest correlations.

Friendship> Flower> Maternal> To recover> Tenderness> Softness> Caress > Perfume> Elegance> Jewel> Present> Reward> Gold> Silver> Wealth> Property> Desire> To seduce> Glory> Victory> To conquer> Ambition> Order> Rifle> To attack> Mistrust> Wall> Space> Danger> Cry> To break> Doubt > Desert> Storm> Rebellion> Death> Detachment> Sacrifice> Effort> Researcher> To teach> Patience> **Meditation**.

Plane 5 / 6.

This case is somewhat unusual because, as we have seen (paragraph 1.4.7), the cluster "Humility", which opposes the pole "Sovereignty", consists of two semantic entities belonging to radically different registers, a first register "regressive" characterized by the words: Countryside, Animal, Birth, Tenderness, Maternal, Humble ... and a second, "dysphoric" (desperate), consisting of the words: Rigid, Mistrust, Fault, Knot, Doubt, Anxiety ... We must therefore follow, at least part of the journey, a double route [141].

We start now from the word *Family* (the top of the vertical axis 6).

Route 1: **Family**> To obey> To forbid> Fault> Anxiety> Mask> Secret> Wall> Desert> Wander> Magic> Perfume> Precious> Elegance> Glory> Homeland> Power> Victory> To conquer> To command> Certainty> Audacity> Builder > Strong> To produce> Dynamic> Volunteer> Practical> Material> Discipline> To obey> **Family**.

Route 2: **Family**> Birth> Childhood> Maternal> Wedding> To love> Jewel> Lightness> Perfume> Precious> Elegance> Glory> (...)> **Family**.

It is equally possible to find such "semantic routes" in the 11 planes defined by all pairs of the five semiometric axes (Duty / Pleasure, Attachment / Detachment, Sublimation / Materialism, Idealization / Pragmatism and Humility / Sovereignty).

Route path in the semantic space derived from Chinese data (Hong-Kong)

Plane 2 / 3 starting from the sublime (the top of the vertical axis 3) are: **Sublime**> Victory> Justice> Strong> To admire> Lightness> Persistence> Wealth> To obey> To inherit> Wall> Sacrifice> Death> Disorder > Wild> To attack> To

[141] The cluster "Sovereignty" consists of a perfectly coherent set of words (Noble, Creator, Art, Elite, God, Audacity, Sovereign, Faith, Soul, Strong ...), we can explain this singularity by the existence of a triangular structure in which the node "sovereignty" would be opposed to two peaks located in the same plane, a node "regressive" and a node "dysphoric." If this singularity is troubling from a semantic point of view, it does not lack any sense from a psychological point of view, the regressive and dysphoric having in common a “lack of being”.

betray> To punish> Question> Mystery> Island> Flower> Refined> Loyalty> Respect> Politeness> Friendship> Modesty> Charitable> Free> Tender> Humble> Honest> **Sublime**.

Two links came as a surprise, Persistence and Lightness, Persistence and Wealth. Yet the links, not obvious to a Western mind, may well exist since the first pair of words correlates at .20 and the second at .33. These figures are statistically significant despite the relatively modest size of the sample (795). We don't know yet whether we are dealing here with a problem of translation (based on denotations!) or with a question of cultural connotations.

Connotations and correlations

If these examples constitute an evidence of the strong link between the connotations of a given list of words and the correlations between these words (correlation derived from the notes of pleasure or displeasure attributed to them by a given population) it remains still necessary to elucidate the process that translate connotations into correlations.

In fact, the explanation is closely related to the very definition of linear correlation. It is obvious that if we introduce the same word twice in a questionnaire, we get (thanks to the noise produced by the repetition of a same word, which is never neutral) exactly the same answers and the correlation between these two words will be close to 1. Conversely, if two words lead the entire sample interviewed to exactly opposite answers (-3 instead of 3, -2 instead of 2 and -1 instead of 1), these two words have a correlation of -1. Similarly, if there is no identity and no opposition between the responses, the correlation will be around zero. And it will be more or less positive or negative, depending on whether the answers are more or less the same or somewhat opposed.

It does not seem illegitimate to conclude that the semiometric space, reconstructed from the whole set of correlations between the 210 words, is a space of connotations and not, as we have once considered the hypothesis, a space of denotation. That is to say that the links between words are determined, not by their objective meaning as given in the usual dictionaries or thesauri, but by their subjective sense built by each person from his/her culture, his/her education and his/her personal experience.

Note that this space of connotations rebuilt through the analysis of semiometric data, is an "average" space of connotations within the sample of respondents (or within a specific country if the sample is representative, as is the case here) that is, the collective synthesis of all individual

connotations, which is, perhaps, what Jung called "collective unconscious". This explains the results obtained (semiometric planes and internal semantic fields) make sense for any individual belonging to the same culture as the interviewed sample.

The first conclusion is then that the semiometric space is a space that is both cultural and emotional, finding its origin in education and experience. It necessarily leads to another conclusion: this space is also both individual and collective, as the observed semantic structure obtained from the interrogation of a group of individuals, makes sense for everyone in the same culture[142].

Perhaps we can conceive that space as an emotional memory that contains the sum of our learning, in which would be stored all of our memories with, attached, a specific satisfaction index. These memories are then arranged according to their affinities (connotations). Each time we find ourselves facing a new situation, we refer to these memories to decide whether that situation is favorable or unfavorable, whether we should accept or flee. The semiometric space is then a particular representation of a connotation space.

Are the dimensions of the semiometric space an artifact induced by the choice of words or the revelation of a latent structure?

Two hypotheses are considered regarding the origin of the five main dimensions identified by the semiometric approach.

They can be induced by the construction of the questionnaire, by involuntarily overweighting, for example, the words related to "Pleasure", "Duty", "Attachment" , "Detachment ", etc.. or they can pre-exist within the minds of the respondents.

If these dimensions are an artifact induced by the questionnaire, we must find them regardless of the sample interviewed. Now, if indeed they appear in all tests performed on samples of populations belonging to the Western world, they disappear when dealing with data from a national sample to another world (cf. the analysis in from the Chinese data (section 2.1.5)). Indeed, the experiment attempted on a sample of 795 Chinese from Hong

[142] If, as we have seen (Chapter 2), we find essentially the same space in all the countries of Judeo-Christian culture, as shown in paragraph 2.1.2. It is not the same when we submit the semiometric questionnaire to a sample of Asians (Chinese from Hong Kong) (see section 2.1.5).

Kong shows that the "Chinese semantic space", it is as semantically coherent as the Western space, but also that it is organized around other dimensions, albeit difficult to be interpreted by a Western mind.

Then, we can make the hypothesis that the questionnaire does not create the structure, it reveals it, similar to X-rays showing the hidden features of the skeleton. If there were no words, for example, related to "Pleasure" and "Duty", the axis 2, most likely, would not emerge. But if the axis did not exist in what might be called the "mental space" of the respondents, the presence of the words would not make this axis appear. In the Chinese study mentioned in section 2.1.2 of Chapter 2, the dimensions "Duty / Pleasure", "Attachment / Detachment", etc.. do not seem to structure the Chinese psyche, despite the fact that the words to describe these dimensions are present in the questionnaire. For an axis to appear, two conditions should hold: 1 / it must pre-exist in the "mental space" of the respondents, and 2 / the questionnaire must contain the words that characterizes that axis.

5.4.4 Origin and meaning of the dimensions of the semiometric space.

The "Leibnizian" question remains: "Why is there something rather than nothing?". That is to say, in this case, why the semiometric space has an ellipsoidal shape and not spherical, why is it stretched in different directions, almost always the same.

Indeed, nothing *a priori* foreshadowed that the connotations space was structured along specific directions. These directions or axes describe what we consider as major oppositions that divide all the societal choices of a population as well as the souls of its individual members (see section 1.4 and chapter 6).

If, as we had a moment conceived the hope at the beginning of this research, we had found the same axes in all studies conducted with samples covering all populations (regardless of their culture), we could have deduce that these dimensions are induced by the "shape" of the human psyche. One might have inferred that these directions are a property of the human psyche as well as spoken language. But the Chinese case proves that it is not. We must therefore assume that, unlike the semantic coherence of space, which is the product of an innate faculty of man, these dimensions are of cultural origin. These dimensions are the “matrix” that formats,

within a given culture, the human psyche. This matrix organizes the sets of drives and determines human perceptions, behaviors and attitudes.

Space of connotations and space of meaning.

This space of meanings that we had originally envisioned in our research seems to well and truly exist in the minds of men living in society[143]. It could be the result of a combination of a natural disposition of man to create associations between his perceptions, from the sensations associated with them. A kind of cultural coding organizes then these perceptions according to patterns characterizing what might be called the "form" of different cultures[144].

It is thanks to the existence of this connotations space, acting as a sort of Cartesian reference, that we are able to understand the world around us. We are thus able to position ourselves in relation to the multitude of objects that make up the world; to position any new experience, to recognize our enemies and our friends; to evolve with a sense of security whose absence would make the stay on earth unbearable. In short! the existence of this connotations space gives us the feeling that the world we live in makes sense and that our life is not this "tale , told by an idiot, full of sound and fury, signifying nothing ", as claimed by Macbeth[145].

Although sometimes the world can give that feeling... Who has never felt disoriented, distraught by the sudden occurrence of an unpredictable and incomprehensible event, precisely, for lack of linking it to past experiences, assign it a place in this connotations network through which we "read" the world?

---

[143] If the idea of a cultural coding in animals seems inconceivable, the faculty of the most advanced of them to create such associations seems, however, more than likely. This faculty conditions their ability to learn. For Pavlov's dog, the sound of the bell "connotes" (positively) food, just like a stick "connotes" (negatively) a prohibited behavior, in the same way that the word Sea connotes (positively and negatively) the vastness, freedom, adventure, depth, storm, shipwreck, etc. for most persons.

[144] See next chapter, dedicated to a tentative analysis of the “form” of the Judeo-Christian culture.

[145] William Shakespeare, *Macbeth*, act 5, scène 5.

CHAPTER 6

Towards a few interpretations

In previous chapters, we have presented the semiometric structures revealed by six stable dimensions. These structures were validated using empirical tests and resampling procedures. This was followed by the study of the attitudes *vis-à-vis* the questionnaires and surveys, emphasizing the inevitable interaction between the *scoring* effects, the attitudes and the content of the results. Moreover, the current collection of data has been compared with what would be a collection compiled from open-ended questions. New experiments and statistical analysis strived to identify the relationship between the observed structure and some semantic networks. The contribution of the semiometric approach to semantic research has been stressed.

We have somewhat developed an "objective" support, a *set of statistical facts*, while remaining parsimonious with substantial comments. Any interpretative comments so far have had only a pedagogical role: covering, with a little flesh, a technical and methodological skeleton. It is the multidisciplinary character of this field of study, and therefore the multiplicity of possible interpretations, which led us to preserve as much as possible the factual basis, reproducible, from which several "subjective" interpretations can now be formulated. This chapter, written especially by Jean-François Steiner, resolutely changes the tone and runs the risk of proposing and arguing in favour of a few possible interpretations of the observed structures located at several levels: historical, sociological, literary, psychological and even, in some respects, theological.

6.1 A first approach

When we look into the semiometric structure a little carefully, such as it appears on the maps of the different factorial planes, this produces in us this strange feeling – almost Proustian – made out of a mixture of feelings of strangeness and familiarity that we experience in front of an object, or a being, albeit well known, but one we discover in a situation where we do not expect to meet him/her, as, for example, when you are sitting in a restaurant at a table near a celebrity. We fully recognize his/her features but we dare not believe it is him/her because, for us, the person belonged up until then to another world with which ours did not communicate.

Thus, the second axis "Duty / Pleasure" for example, has on one side the words: *Sensual, To dream, Adventurer, Original, Nudity, Bohemian, To escape, Rebellion…* and on the other: *Discipline, To obey, Morals, Soldier, Rule, Law, Work, God, Family, Tradition*, etc.. This axis evokes a dichotomy that seems to go through our own lives as well as political life, being at the heart of both our deepest conflicts and societal challenges.

On one hand, there is the famous "Law and Order" (*Law, Rule, Discipline*) of the Anglo-Saxon countries but also the somewhat outdated Vichy trilogy "Work, Family, Homeland", the covenant of the sword, the Army: *Soldier, Rifle*, and the aspersorium, symbol of the Roman Catholic Church: *God, Priest*. On the other hand, one finds all the themes of May 68, the dream of a world all dedicated to enjoyment, without coercion and without any law, where it is forbidden to forbid, and where dreams should be taken for reality (*Sensual, To dream, Nudity, Bohemian, To escape, Rebellion*).

But this dichotomy – and that's where this structure highlighted by Semiometry continues to be troubling – is not only a contemporary political phenomenon. Two centuries ago, Goethe had already described it, but placing it at the heart of his being: "Two souls, alas! – he had Faust say – "live in my chest and repel one another. One, thirsty for pleasure, is attached to the world with the full force of my flesh; the other, driven by an unknown force, seeks to draw me to the lofty mansions of my ancestors" [146].

---

[146] "Zwei Seelen wohnen, ach ! in meiner Brust,
die eine will sich von der andern trennen :
Die eine hält, ein derber Liebeslust,
sich an die Welt mit klammernden Organen :
Die andre hebt ewaltsam sich vom Dunst

The third axis "Attachment / Detachment"[147], with on one side the words: *Jewel, Present, Wedding, Liveliness, Family, Perfume, Flower, Free*, etc. and, on the other: *Danger, Death, To break, Anguish, Emptiness, To punish, Cry, Doubt, Rebellion, Question, Stranger*, etc. evokes God's ultimate injunction, through the mouth of Moses, to the Hebrews, just before they entered the land of Canaan: "See, I have set before thee this day life and good, and death and evil"[148], an injunction that the famous slogan of the beatniks, "Make love not war" may be considered, in some respects, to be a faraway avatar. This opposition between the state of being abandoned and the fullness of being is at the heart of the biblical question and much of Jewish mysticism with the two fundamental antagonistic concepts: *Shechina* (the divine presence) and *Galuth* (the exile of the soul): "From the depths of anguish I cried out to you and you answered me, O Lord, deliver my soul…"[149] says the psalmist. The whole drama of the descent into Egypt - *Mitzraim* in Hebrew, a word that literally means the two anxieties - and the coming up to Canaan, the country "flowing with milk and honey" is based on this same dichotomy between the two states of the soul, sometimes frozen with anguish and sometimes overflowing with happiness.

The fourth axis "Sublimation / Materialism", along with its opposition between, on one side, the words: *Book, Art, Poetry, To teach, To write, To ponder, Researcher, Soul, God, Faith*, etc.; and on the other: *Wealth, Gold, Money, Glory, To conquer, Ambition, To seduce, Desire, Sensual*, etc., describes an opposition that is found in many internal conflicts or societal ones. It is ultimately the Christian conflict that opposes the soul and body, wealth and salvation, the world here on earth, all dedicated to the worship of Mammon and the other world, full of sublime bliss. This same conflict is found today, in a secular context, in the opposition between art and money, research and trade, the NGOs and the multinational corporations, etc. Everyone will find, from Bossuet to Tolstoy, a thousand examples in Western literature of works whose dramatic device is organized around this

---

Zu den Gefilden hoher Ahnen."

[147] Recall the phenomenon of inversed axes "Duty / Pleasure" and "Attachment / Detachment" between the "Latin" countries: France, Italy and Spain and northern countries: Germany, Great Britain, Norway, Finland, Hungary and the Czech Republic. The axis "Duty / Pleasure" has an eigenvalue greater than that of the axis "Attachment / Detachment" in the Latin countries, and vice versa.

[148] Deuteronomy 30.15.

[149] Psalm 120.

conflict. Let us quote this response to Armande, from Henriette in Molière's *Learned Ladies*[150]:

> "Thus shall we both, in our contrary ways, imitate our mother: you on the side of the soul and of noble desires, I, through the coarser pleasures of the senses; you, by the productions of light and spirit, I, sister, in more material ways".

The fifth axis "Idealization / Pragmatism", with the words: *God, Faith, Soul, Sacred, Eternal, Ceremony, Infinite, Noble, Jewel, Poetry, Absolute, To love, Hero, King, Queen, Castle*[151]*, Fashion, Mystery*…, on one side and: *Effective, Accuracy, Logic, Concrete, Clever, Practical, To produce, Headstrong, Inventor, Reason, Science…*, on the other, evokes the opposition between our need to dream and to overreach, and our need to believe in the realities of another order, magical or transcendent; and our antagonistic need to understand and act in a rational world governed by simple laws and fully accessible to our intelligence. It can be interpreted as the axis of Blaise Pascal's conflict between the heart, which is the head and source of faith, and reason for which these mysteries remain forever impenetrable. Or the axis opposing Don Quixote and Sancho Panza: the former, the knight errant who fled the sad reality of his melancholy and his poverty to enter a world of fantasy, where windmills are perceived as giants, where a poor Spanish inn becomes a mighty castle, and where a wretched farm girl appears as a "peerless princess". Don Quixote, who escaped from distressing trivial everydayness into the sublime beauty of a world reconstructed by dreams, is compared to the brave Sancho Panza, who is desperately trying to impose his common sense as an earthy mud-besmirched peasant on the dreamlike delirium of his master.

However, this axis can also be interpreted as that of the conflict between, from the Counter Reformation on, the proponents of grace and those who claimed, such as the Molinists, that salvation was obtained by deeds[152]; or

---

[150] "Ainsi, dans nos desseins l'une à l'autre contraire,
Nous saurons, toutes deux, imiter notre mère :
Vous, du côté de l'âme et des nobles désirs,
Moi, du côté des sens et des grossiers plaisirs ;
Vous, aux productions d'esprit et de lumière,
Moi, dans celles, ma sœur, qui sont de la matière." *The Learned Ladies*, Act I Scene 1.

[151] In an earlier version of the questionnaire comprising 286 words, among which the words *King, Queen* and *Castle*, the latter correlated with the half axis "Idealization".

[152] See the case of the Unigenitus Constitution which poisoned the reign of Louis XIV, the Regency and a large part of the reign of Louis XV, a sort of Dreyfus affair before the letter, dividing French society for decades. The vehemence with which Saint Simon mentions it in his Memoirs sounds like a distant echo of the earthquake that it represented for the French society: "This same month of March (1711) saw the beginnings of the case that produced the Unigenitus Constitution, so

in the second half of the 18th century, the rationalist movement of Enlightenment whose proponents argued that man should have no other master than reason, whereas the religious movement postulated, at the origin of Creation, the existence of an unsurpassable mystery. The latter was responsible for maintaining the sumptuous pump of the Church, the disturbing and melancholic beauty of baroque music and the exuberance of baroque architecture.

The sixth axis "Humility / Sovereignty" contains the words: *Birth, Distrust, Doubt, Family, Rigid, To forbid, To obey, Childhood, Maternal, Immobile, Loyalty, Emptiness, Politeness*, etc. to which must be added *People, Crowd* and *Equality* belonging to the 286 word questionnaire on one side, and: *Noble, Creator, Art, God, Audacity, Elite, To master, Entrepreneur, To command*, etc. on the other. This axis refers to the opposition between the so-called "little people" or "humble people" as Jules Romains[153] called them and the "Masters of the Universe". The former are people who are subjected to their fates and whose horizon is shrivelled up by sociological factors: by a lack of education and, perhaps, a lack of love. The latter, to borrow a phrase which is a little outdated but very telling, are the "great of this world" or "the Superb" from the Latin adjective *superbus* meaning overbearing or arrogant, according to Jules Romains. They, on the contrary, have been brought up to think that they belong to a higher class, which was ordained to dominate others and control their destinies.

In another register, this axis is also a sociological one, because it opposes executives, senior citizens, people living in cities with over 200,000 inhabitants, Paris and the Mediterranean regions, on one hand, with the workers, employees, individuals less than forty years old, people without any school leaving certificate, and rural residents on the other. This axis is even an ideological[154] one, which makes its handling somewhat tricky.

---

fatal to the Church and State, so shameful for Rome, so fatal to religion, so advantageous to the Jesuits, the Sulpicians, the Ultramontanes, the ignorant, the people of nothingness, and especially to any kind of rogue and villain, whose consequences, directed as mush as possible on the model of the revocation of the Edict of Nantes, have engendered disorder, ignorance, deception, confusion everywhere, with such violence that it still remains; and under the oppression of which the whole kingdom trembles and groans, which, after more than thirty years of the more frantic persecution ever experienced, in all kinds and in all professions, a weight that extends to everything, and which dwells forever".

[153] *Men of Goodwill*, books 5 and 6.

[154] In the 286 word questionnaire, the words *People, Crowd* and *Equality* correlate with the half axis "Humility" and are thus opposed to the words *Noble, Elite, King, Queen and Castle*.

As we shall see below, in an attempt to interpret the semiometric structure through a Freudian interpretative grid, this axis can be also understood as the projection, if not of an internal conflict, at least of an intra-psychic tension to which man is subject because, during his development, he needs to build his ego through striving to ensnare his instinctual world.

But before we embark on this perilous adventure, note by the way that, perhaps contrary to what Goethe asserted, more than two souls, alas! inhabit our chests. Indeed, Nietzsche had already sensed this when he replied: "If a German says: Two souls, alas! inhabit my chest, he would be far from telling the whole truth"[155].

6.2 A Psychoanalytical Interpretation

From 1920 on, Freud developed a theory known as the "second topic", in which he describes the psychic apparatus as consisting of three great *agencies*: "the *id*, instinctual pole of the personality, the *ego*, an *agency* which consists in representing the interests of the whole person (...), the *superego*, finally, a *facet* which judges and criticizes, made by internalizing parental prohibitions and requirements"[156].

It seems to us that his description of the *id*: "... it (the *id*) includes everything that has been constitutionally determined, therefore, above all, the impulses emanating from the somatic organization", "the *id* tends to satisfy the innate needs of the individual", "The *id* obeys the inexorable *pleasure principle*", corresponds fairly well to the pole "Pleasure" of axis 2, of which a large portion of the words (*Sensual To dream, Nudity , Wild, Voluptuous, Desire, Carnal, Caress, Emotion, Mellow, Storm, Skin, Revolt,* etc.) are characteristic. That conception evokes precisely the free expression of impulses, the satisfaction of innate needs and pleasure-seeking, or, more exactly, the submission to the *pleasure principal*.

In contrast, the description of the *superego*: "a sort of precipitate of the long period of childhood (...) through which extends parental influence and, (...) transmitted by the parents, the influence of the family traditions, racial and national, as well as the requirements of the immediate social environment which they stand for", of which "the primary task is always to

---

[155] *Beyond Good and Evil*, F. Nietzsche.

[156] *Vocabulary of psychoanalysis*, J. Laplanche and J.-B. Pontalis.

curb the rewards", which dictate "severe restrictions", and which may even appear in a "hard and cruel[157]" manner; this description corresponds very closely to the pole "Duty" of that very same axis with the words: *Family, Homeland, Law, Tradition, To economize, To forbid, Sacrifice, To punish, Rigid*, etc.

Reflecting on the nature of instincts, their orientation and on how they express themselves in human behaviour, Freud suggested, in this "second topic", that they fell into two main categories: "that leads us to distinguish two kinds of drives: those that seek to lead life to death and the others, sexual drives, which tend indefinitely to renew life and which accomplish this"[158]. In 1938, on the eve of his death, he returned to this dichotomy and was more specific about it: "After hesitating and even procrastinating for a long time, we decided to admit the existence of only two basic instincts: Eros and the destructive impulse (...) The aim of Eros is to build ever larger units, thus its aim is to conserve: this is the link. The purpose of the other drive, on the contrary, is to break the relationship, and so to destroy things".

Our intention is obviously not to comment on the relevance of this meta-psychology. It has been the subject of much controversy, which, incidentally, has now been abandoned, for lack of fighters, who were merely reviving the old theological debate inaugurated by Saint Augustine, between the dualists, who claimed that the world was created by two forces – those of good and those of evil –, and the monists who posed the existence of God as the origin and source of everything. What seems to us much more interesting is to show how this dichotomy gives a fairly accurate account of the opposition expressed on semiometric axis 3 "Attachment / Detachment". And indeed, we can consider that virtually every word of the pole "Attachment": *Jewel, Present, Reward, House, Wedding, Liveliness, Family, Loyalty, Tenderness, Peace, Caress,* but also: *Gold*, *Silver, Wealth, Fashion , Perfume*, in a less immediate sense, albeit obvious, relate to objects intended to create links between people or between people and things; while the pole "Detachment" (*Danger, Death, To break, Storm, Anguish, Emptiness, To punish, To criticize, Doubt, Revolt, To attack, Fault, Irony, To age, Wall, Suspicion, Stranger,* etc.) evokes the destruction of relationships, through violence or absence, with oneself or with others.

---

[157] *Outline of Psychoanalysis*, S. Freud.

[158] *Beyond the Pleasure Principle*, S. Freud.

If the two previous axes can be interpreted quite easily from Freud's "second topic", the next two ("Sublimation / Materialism" and "Idealization / Pragmatism") can be less easily reduced to a literal psychoanalytical interpretation. However, the pole "Sublimation" of axis 4, with words connoting religious spirituality (*Soul, God, To pray, Meditation*), those describing the intellectual activities (*Science, Researcher, Inventor, To think, Question*) and those referring to different expressions of art *(Poetry, Book, Art, Music, Theatre*) is not without evoking the psychological phenomenon that Freud called *sublimation*: "Sublimation is a process postulated by Freud to account for human activities seemingly unrelated with sexuality, but which find their resilience in the force of the sexual drive. Freud described, as sublimation, primarily *artistic activity and intellectual inquiry*"[159], "The sexual drive makes available to intellectual work extraordinarily large amounts of strength, and because of this peculiarity, especially accentuated here, this drive is able to shift its objective without losing most of its intensity. We call this ability to exchange the original sexual aim for another one, which is no longer sexual, but which is psychically related, the capacity for sublimation"[160], "The historians of civilization seem willing to admit that with this diversion of impulsive sexual forces away from instinctual sexual purposes and guidance toward new goals - a process that deserves the name of sublimation - powerful components are acquired, taking part *in all the intellectual constructs*[161]". Certainly, at any time, Freud does not mention religious spirituality as a manifestation of the mechanism of *sublimation*, but it is reasonable to think that we should only see here the effect of the virulence of his atheism. Indeed, the links between religion and art, as well as science in its most pure manifestations, are too well known not to allow us to extirpate this exclusion.

If we accept this interpretation of the half axis "Sublimation" as corresponding to the phenomenon of *sublimation*, we can understand its antagonistic half axis in terms of human activities in which sexual urges are expressed in a more immediate and less elaborate way, a hypothesis, that the meaning of the words, *Wealth, Gold, Money, Glory, To conquer, Ambition, Trade, To command, To attack, To seduce, Desire, Sensual, Carnal, Voluptuous*, etc.. is far from contradicting.

---

[159] *Vocabulary of Psychoanalysis*, op. cit. Emphasis added.

[160] *Gesammelte Werke*, VII, 150 S. Freud.

[161] *Three Essays on the Theory of Sexuality*, S. Freud.

In the light of this interpretation, axis 4 would be formed by the opposition between people who express their sexuality in a relatively basic way; and those who are more sensitive to the cultural constraints that they have undergone during their childhood. With the multiplicity of taboos that any powerful fruitful civilization continues to decree against sexuality, the latter group has managed to channel their sexuality into other objectives, representing – at least up until recently – more socially elevated ideals.

As one penetrates deeper into the dimensions of the semiometric structure, interpreting the axes becomes more difficult, which should not surprise us, since the amount of information provided by each new axis continues to decrease. This difficulty is increased once more by the fact that if the pole "Materialism" of axis 4, and the pole "Pragmatism" of axis 5 are defined by radically different sets of words (*Gold, Money, Wealth ...*, for the first and *Logic, Accuracy, Effective…*, for that second one), the pole "Sublimation" of axis 4, and "Idealization" of axis 5, exhibit disturbing similarities: the same words of religious spirituality: *God, Priest, Faith, Soul...* and the same words of artistic expression: *Poetry, Theatre, Music, Art*. Yet, despite these similarities, there are two important differences between the poles "Sublimation" and "Pragmatism", which show they should not be confused. Besides the presence of words such as *King, Queen* and *Castle* next to "Sublimation", the big difference between the poles "Sublimation" and "Idealization" lies in the fact that words connoting intellectual activities: *Inventor, Scientist, To think, Science , Reason*, which correlated positively with the half axis "Pragmatism", negatively correlated with the half axis "Idealization" while some words of "Materialism": *Jewell, Gold, Fashion, Hero, Ceremony*, which negatively correlated with the half axis "Sublimation" now correlate positively with the half axis "Idealization".

This kind of exchange, which occurs when one moves from axis 4 to axis 5, if it has been for a long time a stumbling block for our interpretive attempts, perhaps provides a key for understanding axis 5 in terms of Freudian theory.

Indeed, Freud noted the existence of a psychic phenomenon, which has a number of similarities with *sublimation* to the extent of sometimes being mistaken for it, but which must nevertheless be regarded as radically independent in its mechanism: *idealization*. Our hypothesis is that if the pole "Sublimation" of axis 4 corresponds, at least partly, to the psychic phenomenon that Freud called "sublimation", the pole of axis 5 designated

by us as "Idealization", which has much in common with the former, corresponds effectively to the phenomenon of what he called "idealization".

But let us listen to Freud: "The formation of the *ego*'s ideal, he writes[162], is often confused with the sublimation of impulses, to the detriment of a clear understanding. A person who has exchanged his narcissism against the veneration of a high *ego* ideal has not necessarily been successful for all that in sublimating his libidinal impulses. The *ego* ideal requires, it is true, this sublimation but it can not get it through force: sublimation remains a particular process; the ideal may well encourage it (sublimation) to begin the process; but it remains completely independent of such an incitement (...) The formation of the ideal increases (...) the demands of the *ego*, and it is the former that acts most strongly in favour of repression; sublimation is the outcome that will satisfy these requirements without causing repression".

After having highlighted the links between the two phenomena, Freud defines them respectively and relatively, from the standpoint of his theory: "Sublimation is a process which concerns the object libido and consists in the fact that the drive is directed toward another purpose, away from sexual satisfaction: the emphasis is laid on the detour away from sexuality. Idealization[163] is a process concerning the object and by which this object is magnified and exalted psychically without its nature being changed. (...) Thus, provided that sublimation is a process concerning the drive and idealization a process concerning the object, the two concepts must be kept separate from each other"[164].

Axis 5 is likely to contrast, "idealization" with "rationalization". The former is: "a psychic process by which the qualities and value of the object are brought to perfection" (*God, Queen, King, Faith, Priest , Poetry, Ceremony, Jewel, Theatre, Sacred, Soul, Castle*) whereas the latter: "a

---

[162] *On narcissism: a beginning*, S. Freud.

[163] Freud operates a shift between the "ideal of the ego" and the *idealization* of the object, without explaining more beforehand, as if, for him, the two mechanisms were identical. The missing term of the ellipsis is, probably, the phenomenon of projective identification by which the subject merges with the object, i.e. the subject, having not completely cast off his narcissism, projects the ideal image he has of himself on the object, thus bestowing it with narcissistic libido. Freud confirms, moreover, this hypothesis and he also wrote: "In many forms of amorous choice, it becomes very obvious that the object is used to replace an ideal of one's own *ego*, not reached. We love the object because of perfections to which we aspired for our own *ego* and now we would like to procure it through this detour to satisfy our narcissism" (*Group Psychology and the Analysis of the Ego*).

[164] *On narcissism: a beginning*, op. cit.

process whereby the subject seeks to provide a coherent explanation from a logical point of view (...) with an attitude, an action, an idea, a sense, whose true motives are not perceived [165]" (*Effective, Precise, Logical, Concrete, to Master, Clever, Practical, To think, Reason, Science*, etc.). Note that "idealization" like "rationalization" is often used in defensive strategies against anxiety. The fact that they belong to mental strategies of a paradoxical nature could perhaps explain, at least partly, the often implacable character of conflicts that are structured around these two attitudes.

The interpretation of the sixth axis from the point of view of Freudian theory is much less obvious. As we have remarked above, we could see the trace of an intra-psychic conflict between the *ego* in its reiterated effort to pull away from the gravity of the instinctual world[166] (*Noble, Elite, Creator, Strong, Audacity, To command*), and the temptation to yield to the call for an impossible return to the idyllic state of childhood and nature (*Family, Birth, Childhood, Water, Animal, Countryside*).

This could be the axis of the opposition between the *ego* and the *id* – provided those *agencies* indeed exist! ; or, more exactly, the axis of opposition between people who give more value to their *ego* as their sovereign authority, those, perhaps, whom Marxists call "petty bourgeois individualists"[167] and, on the other hand, those who prefer to give up their individuality and blend into a whole which transcends and absorbs them[168] .

But ours is a very loose interpretation that we are unable to substantiate with any theoretical texts, as we have tried to do, with varying degrees of success, for the previous axes.

Our aim, moreover, was not intended to give a comprehensive and definitive interpretation of the dimensions of the semiometric sphere, but only to suggest some possible interpretations in order to make the reader share the sense of wonder that we experience sometimes in front of how,

---

[165] *Vocabulary of Psychoanalysis*, op. cit.

[166] See the famous "Wo es war, soll ich werden" (Where there was the *id*, the *ego* must appear) of Freud, that sounds like a distant echo of the first commandment of God to man after having created him from dust: "Be fruitful and multiply and fill the earth and enslave it", which some commentators interpret as an injunction ordering man to dominate his instincts.

[167] Widespread translation (involving the pejorative epithet *petty*) of: "petits bourgeois individualistes".

[168] When we project the consumers of different types of products on the axes, the consumers of luxury goods, whose social function of enhancing the status of the *ego* is well known (prestigious Champagne brands, branded watches, etc.) always appear on axis 6 on the side of "Sovereignty".

from a simple score assessing the amount of pleasure (or displeasure) that is attached to the connotations of some words, the latter are organized in a *semantic space* of unexpected depth.

CHAPTER 7

Some semiometric applications

Designed originally as a tool to reconstruct a hypothetical symbolic space that would structure, at least partially, our subconscious perception of the world, Semiometry has proved to be, in quite a fortuitous manner, a tool that could be used to some advantage in psycho-sociological and marketing applications. Thus, since 1986, alongside a research program and theoretical reflection on the nature of the structure highlighted by Semiometry, a partnership has developed between Jean-Francois Steiner and the International Communication Taylor Nelson Sofres Group (TNS) to develop this approach through opinion polls and marketing.

This book is not the place to present a review of hundreds of studies performed both in the field of marketing and in sociology, psycholinguistics or semantics. Our intention is only to illustrate some of these applications with some examples. The cases presented are far from being exhaustive and covering all applications. We have chosen them more for their educational value (the value of their descriptive and explanatory power) than for their actual operational interest[169]. Commercial applications, which remain the property of the organizations and companies that sponsored the studies, involve trademarks or brand names for which a degree of confidentiality is generally desired. They will not be presented here. The reader will however have no trouble imagining that instead of the categories used in this chapter (sex, age…) we can involve consumers of such a product, the owners of specific makes of cars, the voters of a particular candidate or party.

[169] There are other applications that rely on the ability that Semiometry provides to measure distances between individuals and groups of individuals in a stable space in time and space. They are cited here for reference. These applications are in the field of qualitative media planning and human resources. Other applications, finally, use the predictive power of semiometric rating to improve the quality of merging files or to predict missing data (see, for example, Auliard and Steiner, 1992).

The method involves comparing the average score attributed by a first population (called the "target population") to the 210 words in the semiometric questionnaire, with the average score assigned to these words by a second population (called the "reference population"[170]), then applying a statistical test to deviations in averages in order to define words significantly over-scored and under-scored by the latter population. These words are then presented on the most pertinent plane (generated by a pair of principal axes). This type of display allows for an easy reading of the results, but more specifically allows one to assess the semantic consistency of these results. We shall see as we peruse the examples that the combination of characteristic words in specific areas of the semiometric space provides the reader with a confirmation of the significance of the results by taking into account the correlations between words thanks to suggestive visualizations.

An alternative method (a more classical one) to compare sub-populations in a space whose points are the respondents consists in projecting the centres of gravity (or "average points") of the individuals belonging to these sub-populations onto the principal planes (see Appendix A1.9.3). This procedure is actually complementary to that just described, and in practice, must be implemented simultaneously. It does, however, lead to results that are less suggestive. It is therefore not presented here.

7.1 The words characteristic of males and females

Figure 7.1 is drawn from a diagram of the semiometric plane [2,3] which will serve as the backbone for most of our examples. The words on a black background are those that the men over-score compared to women (and women, therefore, under-score); the words framed on a white background, are those underscored by men, and therefore, words that women over-score[171]. As can be seen, the result is quite a caricature.

---

[170] This second population is generally the complement of the first. However, when we suspect that structural effects could mask the result that one seeks to highlight, we filter the two populations with a disruptive variable. e.g. when you want to study a product consumed primarily by individuals of one demographic category, we filter the two populations by this category. Thus, for a women's magazine, if the sample on which we work is representative of the national population, we oppose its readers to all women and not to the rest of the population. The words that characterize this readership would essentially be words over-scored by women.

[171] For the sake of clarity, only the first twenty words are presented here. In fact there are 46 words over-scored by men and 73 for women at a 1.96 threshold. Similarly, for the same reason, in

For men: *Violence, Danger, Detachment* and *Sexuality*; for women: *Softness, Elegance, Poetry, Art,* and *Fashion*. Note that the most characteristic word for men is the word *Rifle* (test-value: 27.9) and the most characteristic one for women is *Jewel* (test-value: 25.5), both of which are usually considered symbolically sexual for their respective sexes. Remember that, to qualify this interpretation, the same principal plane (spanned by axes 2 and 3) is also obtained through a principal components analyses conducted separately for men and for women, a phenomenon about which we have already spoken in terms of *fractal structure*. Some men, of course, over-score the word *Jewel* and so do some women for the word *Rifle*. When both sexes are analyzed simultaneously, the two distributions of men and women on the third axis "Attachment / Detachment" largely overlap, but are very slightly offset[172].

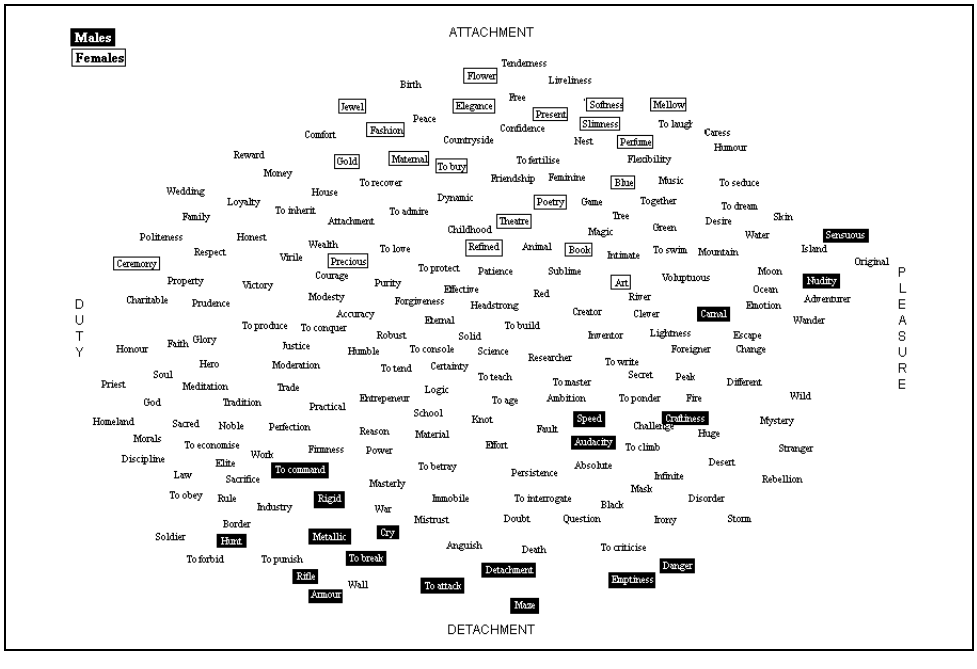


Figure 7.1 Words characteristic of men and women in the main semiometric plane [axes 2 and 3].

(The locations of the words are slightly altered to avoid overlapping of identifiers)

---

the following examples are included only the first twenty words. Note that the tags representing the words have been moved (as slightly as possible) to avoid overlaps between identifiers. .

[172] On this third axis, the centres of gravity (mean score) of men and women are opposed, with a 20.0 value-test (for the full sample of 15,686 individuals).

7.2 The characteristic words of two age groups

The two age groups selected for this example are those (under 45, 45 or older). Figure 7.2, whose interpretation rules are the same as those in Figure 7.1, shows a sharp contrast between words characteristic of the two age groups. The words on a black background are over-scored by the less than 45 year-olds.

Axis 2, "Duty / Pleasure", is closely linked to age, here again the result is not surprising. The question raised by such an outcome is which of these words are among those that are the product of a conflict of age and those which are the product of a generation conflict. Indeed, if *Homeland* and S*oldier* appear to reflect a changing society, and *Moderation* and *Meditation*, a maturing of the individual, what can we say about P*oliteness, Morals* or *Work*? We could hope that rejection of the latter by the younger segment of society is only temporary and due to that passion to make the best out of life, characteristic of youth.

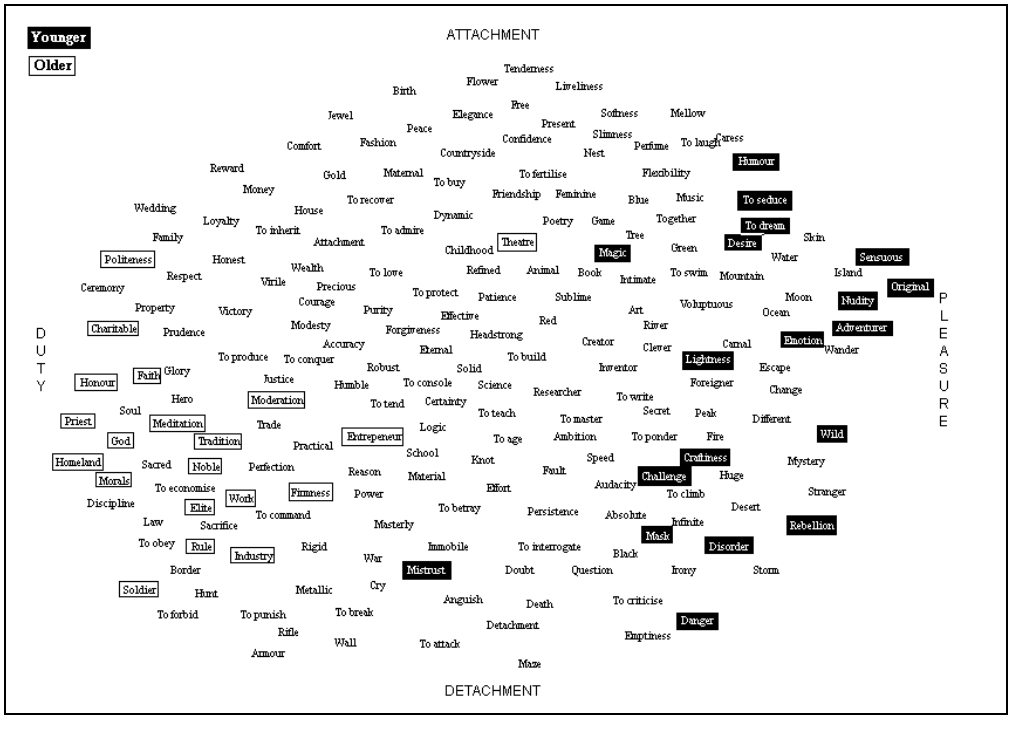


Figure 7.2 Words characteristic of two age groups in the main semiometric plane [2,3]

The presence of the word *Theatre* over-scored by the older group could be quite surprising. If the test-value which measures its significance, had not been very high (7.16), one might think that this is a random statistic result[173]. We shall see in an example presented later (see Section 7.5) that, far from being accidental, it reveals a phenomenon of age - life cycle or generation? - not devoid of interest.

7.3 Television addicts

This result comes from a study for a French television station that covered program audience measurement. The group, which we conventionally define here as "TV addicts", does not consist, as is customary, of the biggest consumers of TV, i.e. individuals who spend more than five hours per day in front of the screen, but the upper quartile of individuals watching the greatest number of programs.

The most discriminant axis being axis 4 ("Sublimation / Materialism"), the results are presented this time in the plane (2,4).

Figure 7.3 shows that the "big TV buffs" only over-score one word, *Hunt*, and that at a threshold of 2.53, which is not markedly significant, given the above remarks on multiple comparisons.

For ten words under-scored, nine (*Book, Poetry, Researcher, To ponder, Science, To write, Patience, To protect* and *Tree*), correlate significantly with the half-axis "Sublimation".

The portrait, which this result draws of the "TV buffs" seems rather bleak. Their only common feature is to reject what, in some ways, contributes to the greatness of man: his ability to control drives and channel them to goals of greater social value.

---

[173] Test-**Values** (see Appendix A1.9.1) should in theory be outside the range [-1.96, +1.96] to be statistically significant at the usual 5% (or: 0.05) threshold in the case of a single test. But we are dealing here with a phenomenon known as *multiple comparisons*: we carry out as many tests as there are variables. This is a difficulty which has long been recognized by statisticians (see Appendix A1.9.2). Several techniques are proposed, leading to recommending a more conservative threshold, the calculation of which depends on the type of test used. In our case, the situation is even more complex, because words are far from being independent and there is no readily available statistical tool to take into account this interdependence.

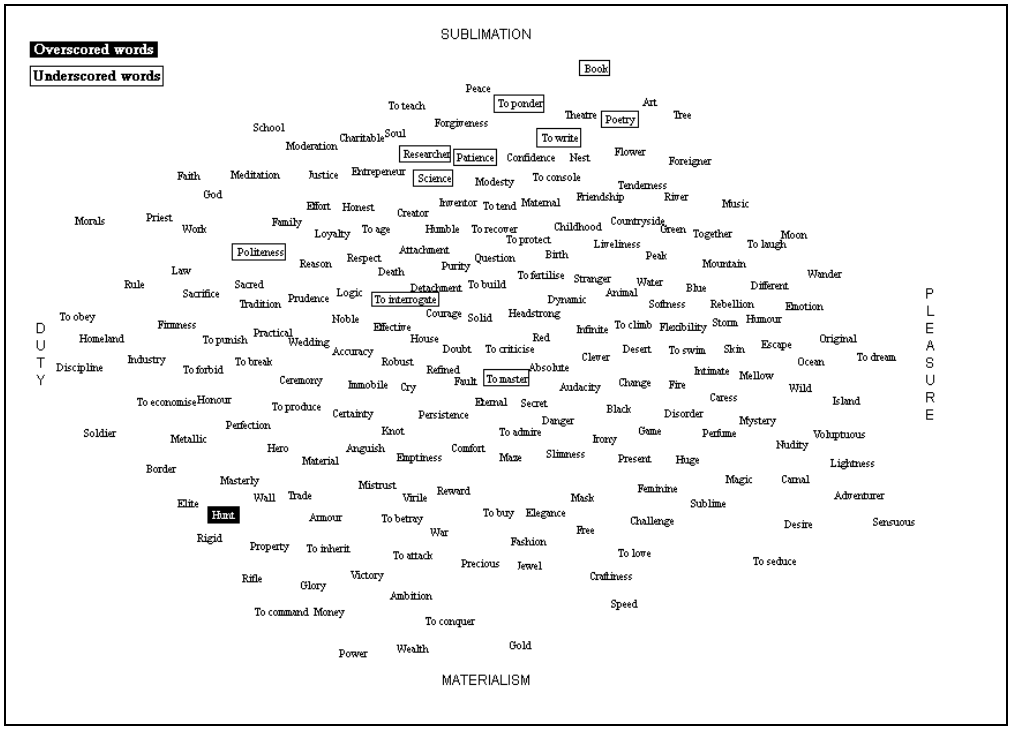


Figure 7.3 Words characteristic of "TV addicts" in the semiometric plane [2,4]

(Reminder: The locations of the words are slightly altered to avoid overlapping of identifiers)

7.4 The characteristic words of two groups of young people (15-19 year-olds and 20-24 year-olds)[174]

In this application, we calculate the words that these two specific populations over-score at the same time (relative to their respective reference populations). Those words are shown on the most relevant factorial plane.

[174] For readability, when dealing with two populations, only the overrated words are plotted. Among them, we distinguish three categories: the words overrated by both categories and the two sets of words characterizing exclusively each category.

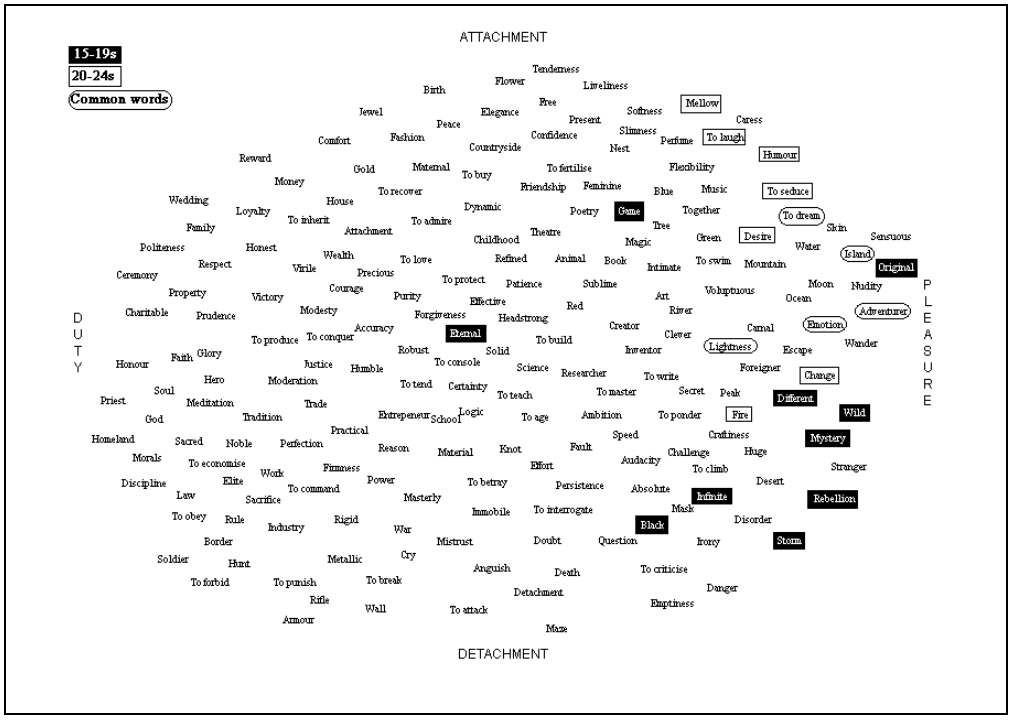


Figure 7.4 Words characteristic of the 15-19 year-olds and the 20-24 year-olds in the semiometric plane [2,3]

With the words: *Wild, Mystery, Revolt, Storm, Black*, the 15-19 year-olds clearly illustrate the opening sentence of Paul Nizan's "Aden Arabie": "I was twenty, I would not let anyone say that this is the prime of life". But, after the fateful passage of 20, young people seem to start taking more pleasure in life, with words like: *Mellow, To laugh, Humour*. The reason is, perhaps, because they find ways to do this: *Desire, To seduce*.

Anyone who remembers the stormy years of his/her adolescence, will agree easily with these results. But semiometry can show in a way that is both very strong and obvious this slow shift of bitter detachment, marked by the expression of the instinct of destruction of the teenager who is still seeking his/her identity and place in society, towards a more harmonious relationship with himself/herself and with others.

This result illustrates the amazing power of statistics to reveal phenomena whose presence, because it is too diffuse, sometimes seems to defy observation.

7.5 The characteristic words of two age groups (above 60)

The two age groups (60-64 years of age, 65 years and over), as expected, both invest heavily in the values of "Duty". Among the first twenty words most over-scored, fifteen for both groups are words pertaining to the domain of "Duty". Among these fifteen words, nine are common: *Morals, Homeland, Industry, Honour, Firmness, Tradition, Elite, Moderation* and *To produce*. Between the six words over-scored by one group but not the other, there are very few shades of meaning – which fade, moreover, if more words are taken into account.

The group of 60-64 year-olds are distinguished by more "relational" values: *Politeness, Respect, Honesty*; those of 65 years of age and over, break away from these contingencies to focus their attention on the hereafter: *Meditation, Priest, Sacred*. But, there are observable differences on the most sensitive axes 4 ("Sublimation / Materialism") and 5 ("Idealization / Pragmatism").

If members of those two age groups invest equally in the values of "Sublimation" and "Idealization" rather than "Materialism" and "Pragmatism", people aged 65 and over do it in a clearer way: six words pertaining to "Sublimation": *Book, Theatre, Art, Charitable, Meditation* and *Priest*, against four for the 60-64 year-olds: *Justice, Maternal, Honest, God,* and five words belonging to "Idealization": *Theatre*[175]*, Noble, Meditation, Priest, Sacred* against one word, *God* - although the word God comes in twenty-sixth position for the 65 year-olds and over, with a test value of 3.93.

[175] A single word, of course, can be correlated with several axes, each of its "epiphanies" revealing one of its semantic dimensions. It is clear that the word *Theatre* in its dimension "Idealization (associated with the words: *Ceremony, Queen, King, Jewel, Castle, Perfume, Mask, Magic*, etc) is not the same word *Theatre* in its dimension "Sublimation" (associated with the words: *Researcher, To teach, Inventor, Question, Art, Book, Foreigner, Peace, Forgiveness, Modesty*). Just as the word *Wedding* in its dimension "Duty" (associated with the words *Loyalty, Honest, Respect, Politeness*, etc.), is not the same as *Wedding* in its dimension "Attachment" (associated with *Comfort, Maternal, Birth, Confidence, Friendship, Tenderness and Softness*), the former referring to the marriage rather seen as an institution, the latter as part of a loving relationship between two people. There is likewise the word *Border* in its dimension "Detachment" (associated with the words *Rigid*, *To break, Wall, Mask, Anguish*). This is the border that isolates the subject by rejecting the other, somewhat a "bad border", whereas the *border* in its dimension "Duty" (associated with *Wedding, Property, Rule, Morals*), is the frontier that organizes, creates and protects the social order as much as a mental order: we may call it the "good border".

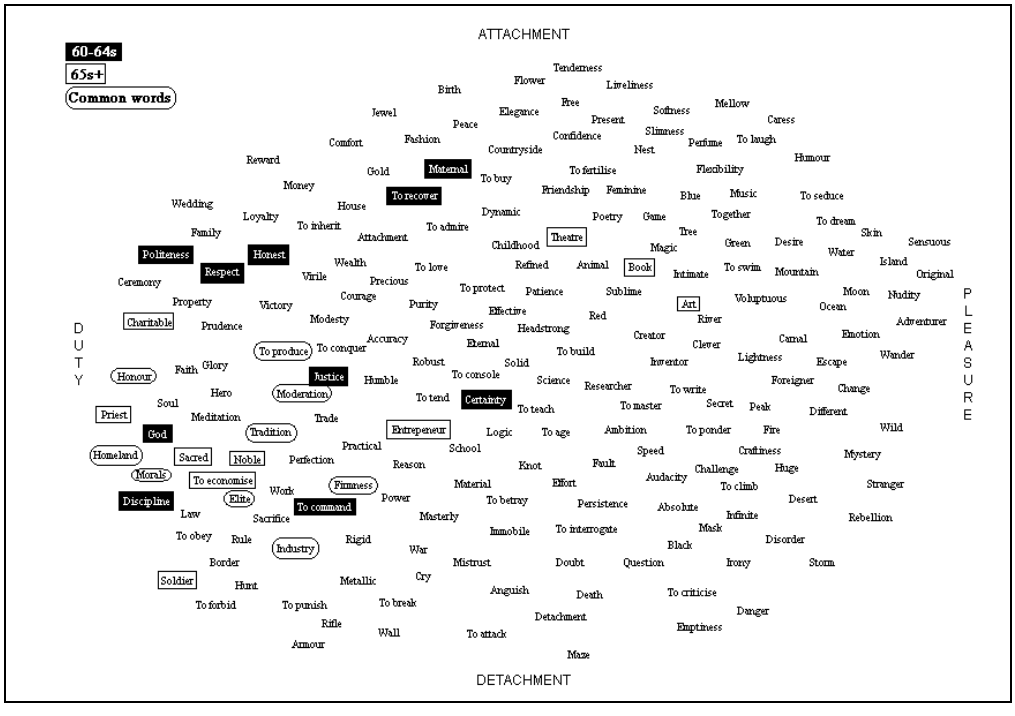


Figure 7.5. Words characteristic of the 60-64 year-olds and the 65 years and over in the semiometric plane [2,3]

(Reminder: The locations of the words are slightly altered to avoid overlapping of identifiers)

This shift along axes 2, 4 and 5 may be an effect of age: after 65, impulses start to be less strongly expressive and it is therefore easier to control them and subsequently, to sublimate them. This shift is presumably due to retirement from active working life which leaves little room for imagination and reflection and little scope too for a more contemplative life.

7.6 Men and women who have attributed a "7" to the word *God*

To understand the results shown in Figure 7.6 fully, we must bear in mind how we proceeded: to eliminate the effect of sex categories whose importance and significance would have masked the subtle differences between male and female scoring. We have compared men and women who

have assigned a "7" to the word *God*[176] separately with those who have given this term a lower score. Both results were then projected onto the same map.

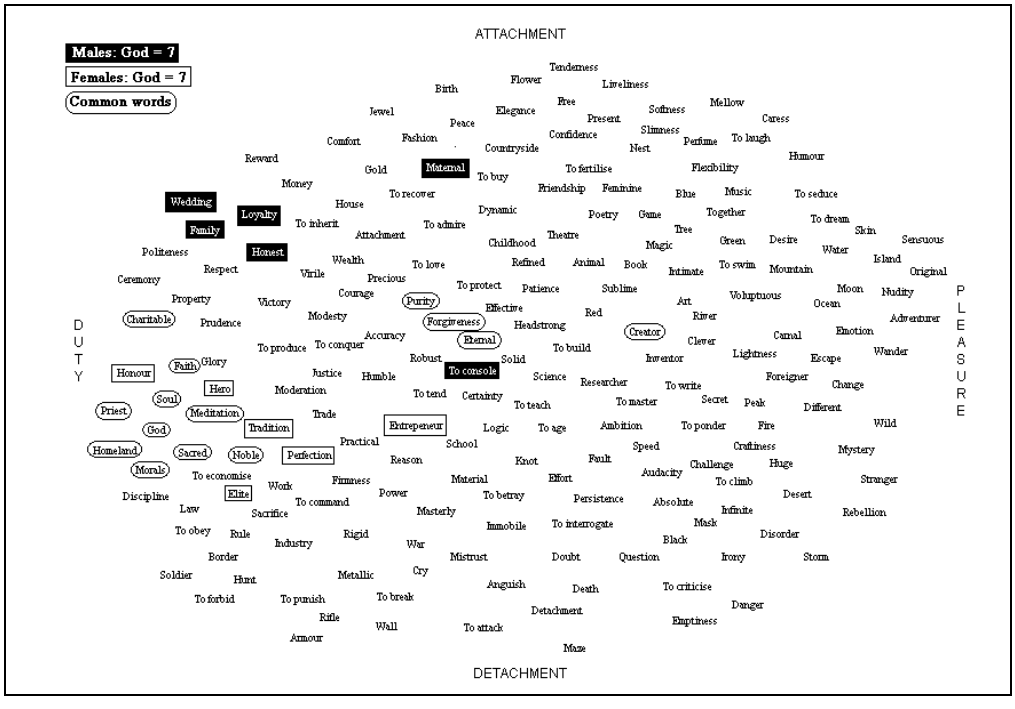


Figure 7.6. : Words characteristics of men and women who have given a "7" to the word God in the semiometric plane [2, 3]

The words on a black background are those characteristic of men responding that the word *God*, when evoked, conjures up very pleasant sensations compared to those who find this word less pleasant or very unpleasant. The words framed on a white background are the words characteristic of women for whom the word *God* causes very pleasant sensations compared to other women. The words on a rounded rectangle background are the words both have in common.

The result as it appears on the map is so paradoxical that, initially, we were inclined to think it is due to a coding error or manipulation mistake.

Indeed, the words over-scored by men are ones referring to the values, traditionally deemed feminine: *Wedding, Loyalty, Family, Maternal, To*

[176] Remember that the terms proposed to the interviewees range from - 3 (very unpleasant sensation) to +3 (very pleasant feeling) but, for reasons of convenience for capturing and processing, the scores are then recoded from 1 to 7.

console, while the words evoked by women seem rather to belong to the register of "male" values: *Honour, Hero, Tradition, Entrepreneur, Perfection, Elite*.

However, once these results have been checked, this is not so, and we must therefore ask ourselves what these strange results are due to.

We first notice that the words that characterize men are more about the relationship with the mother than with women in general. They primarily express a demand for protection against neglect, *Loyalty,* and against grief, *To console* for which the child asks from his mother in his distress.

Symmetrically, the words that characterize women express the same demand for reinsurance, but which are addressed to the father, "this hero whose grin is so sweet"[177], this builder, the custodian of a tradition on which the fragile ego of the child is being built. Of course, we are dealing more with ideal images of the father and mother, products of childhood fantasies, than with any objective reality. But do we not say that the concept of God is the product of an ideal image of the father (or mother)?

So, what this result shows is that, by virtue of a surprising *chasse-croisé*, the God of men is statistically rather maternal: this would be the God of mercy; while the God of women is rather paternal: this would be the God who said to the Apostle, "Peter you are a rock, and upon this rock I will build my Church, and the powers of evil will not prevail against it"[178].

But we are neither theologians nor psychologists, only statisticians attempting to measure complex phenomena with the greatest precision and accuracy possible. The interpretations presented only intervene in order to illustrate. They are just meant to draw the attention of relevant specialists, and suggest further research.

7.7 Changing values of the French between 1996 and 1998

The last example presented relates to comparisons between two samples from the same population surveyed at two different times. As noted previously, in the synchronic analyses, the reference population is (almost) always the complement of the study population. In the diachronic analyses,

---

[177] « *Mon père, ce héro au sourire si doux…* » from the poem : « *Après la bataille* », La Légende des Siècles, Victor Hugo.

[178] Matthew 16, 18.

it consists of a second sample having the same structure but questioned beforehand. We shall therefore take care, when observing the results, not to confuse the two approaches.

The first characterizes, at some point, the values of a given population relative to another one. If this other population is the complement of the study population compared to the total population, which is usually the case, we can consider that the words it over-rates represent what we may designate as its *relative value system*[179].

The second approach characterizes a given population relative to a similar population observed at an earlier time. In this case, words that it over-scores do not indicate its value system but changes during the period between the two observations. This type of analysis helps to highlight certain aspects of the evolution over time of the values of a population.

This application was made from two samples tested for their basic characteristics of 2,764 individuals representative of the French population. These samples are taken from the mail panel of TN Sofres called *Métascope*. This panel, like any panel, may present unidentifiable bias due to the fact that, by definition, it does not take into account people who do not accept to take part in it. But the permanence of the rules governing its management makes it fertile ground for comparisons diachronically. It seems unlikely that the differences observed between the two samples can only be artefacts, induced by structural changes between those two samples. Many signs, however, lead us to believe that those differences reflect a change in sensibility (if not in mentality) of the French during this period.

The analyses conducted in Section 3.8 of Chapter 3 have also revealed the same trend during the whole period from 1990 to 2002, on a much broader set of six successive surveys. In Figure 7.7, the words on a black background are those over-scored by the French in 1998 compared to 1996.

This shift along the axis “Attachment / Detachment” characterizes, as we have seen in section 3.8, the entire period from 1990 to 2002. We noted then that the attitude toward the questionnaire, as evidenced by the way

[179] We make the assumption that the set of words over-scored by one sub-population relative to another of which it is its complement can be considered to represent its *relative value system*, For example, the words over-scored by the readers of such and such a national daily paper in relation to all readers of the whole press represent the value system of the former from the perspective of the whole. It is clear that if one opposes the readers of this newspaper to all the French population, the result will be significantly different.

people attributed a score, varied during the same period in the sense of indifference and with an increased distance *vis-à-vis* the investigation.

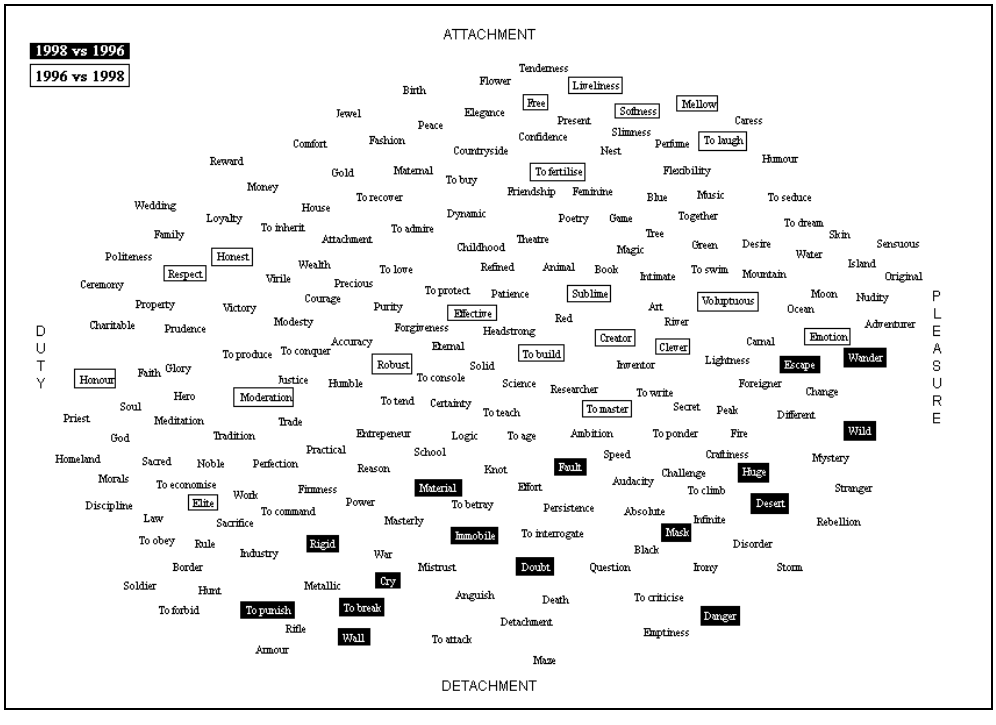


Figure 7.7: Words characteristic of 1996 and 1998 in the semiometric plane [2, 3]

(Reminder: The locations of the words are slightly altered to avoid overlapping of identifiers)

One could interpret this development as a crisis of pessimism or, in the vein of the tentative interpretation of Chapter 6, from a Freudian standpoint, the emergence of a destructive impulse in French society.

This interpretation, moreover, was consonant with diagnoses coming from extremely diverse backgrounds, such as Alain Ehrenberg (1998) or Tony Anatrella (1995). From the latter, we will retain the following description: “We are in the suicidal ambiance of a society that favours rupture and breaking up. We must break apart and smash to feel we exist (divorce and suicide to break free, abortion and euthanasia through distress or compassion: curious motivations linked to the death drive). We must freak out to escape from life better by consuming anaesthetics, stimulants or hallucinogens; we must remain unsophisticated, morbid (...) we must

break the social bond, eradicate desires and emotions, ones we do not know how to deal with, and destroy all institutional reality"[180].

The propos may seem exaggerated, but our results, produced by a system of measures which is both blind and rigorous, seem to be leaning in the same direction. "We must break and smash", "break the social bond", "remain unsophisticated, morbid", but is that not what is expressed by the words *Wall (*test-value: *11.1), Desert (10.1), Mask (8.7), Rigid (8.6), To punish (8.3), To break (8), Doubt (7.9), Danger (7.6), Fault (6.9), Immobile (6.6), Wild (6.1), Cry (5.6)* .... "We need (...) to eradicate the desires and emotions that we do not know how to deal with", but is that not what is expressed by the relative rejection of the words: *Sublime (-9.3) Voluptuous (-8.8), Emotion (-8.6), Mellow (-7.7), Softness (-7.2), To laugh (-7.1)*, ... "and destroy any institutional reality": *Creator (-11.6), Elite (-10.6), Robust (-9.2), To master (-7.8), To build (-7), Effective (-11), Respect (-8.9), Honest (-7.8), Honour (-7.7)?*

It seems as if there is a family resemblance between the previous clinical discourse and the litany of words, selected according to blind statistical criteria. Both approaches appear to describe the same social phenomenon, the same social reality in which we are immersed and which we can hardly grasp in its entirety spontaneously, not being able to step outside it, to distinguish the facts from the fantasies that we project onto it... in the hope - often sadly disappointing... - of neutralizing their anxiety-producing effects.

But this is, of course, only an interpretation, and we leave each reader to formulate her/his own idea.

However, the study of evolution, presented in Section 3.8 of Chapter 3, *which takes into account six years of surveys distributed between 1990 and 2002, covering more than 15,000 interviews, confirms the shift toward the* "Detachment" of the entire population[181]. Remember that this

[180] In his book "*Non à la société depressive [Down with the depressed society]*" (Champs, Flammarion, Paris, 1995), Tony Anatrella, a priest, a psychoanalyst and professor of social psychiatry, attempts an explanation for the emergence of the destructive impulse in modern societies, from the lack of an ideal. We can also read with great interest the work of Alain Ehrenberg, Senior Research Fellow at the CNRS, "*La Fatigue d'être soi*" (Ed. Odile Jacob, Paris, 1998) which shows fairly convincingly how we have replaced "the old bourgeois guilt (and) the struggle to free ourselves from the laws of our fathers (Oedipus), the fear of not being up to scratch, the emptiness and helplessness that result from this fear (Narcissus)" , i.e. how we have substituted for the neurosis, "disease of the law", depression, the "disease of inadequacy".

[181] Note that this shift affects both men and women, although these two categories of respondents occupy significantly different positions along the "Attachment / Detachment" axis in the beginning and at the end of the period.

"Detachment" cannot be separated from what might be called a "detachment *vis-à-vis* the investigations and the questionnaire themselves", a more methodological disengagement, manifested by a reduced use of extreme scores.

7.8 Conclusion *(chapter 7)*

Characterizing groups of individuals (viz. in practice consumers of a brand or product, customers of a large service company, readers of a newspaper or magazine ...) by words, and placing these words in a space whose dimensions are meaningful, is the method of exemplifying the semiometric tool selected in this chapter. In the purely methodological context of this current work, we have worked with more conventional groups of individuals, essentially defined by basic socio-demographic variables, showing the openings and possibilities of the tool in the fields of sociological or socio-economic applications.

As in previous chapters, commentaries and interpretations are mostly invitations to comment and interpret. They are examples ... about examples. How could it be otherwise, given the richness and complexity of this new landscape for statistical semantics?

The interpretations of more commercial or industrial work, which are part of a strategy and a well-defined decision-making framework, will be generally more targeted. In an era of ergonomics, human-machine interfaces, sensory analysis, media and multimedia, and general awareness of the quality of life, the positioning that the tools of Semiometry allow provide decision-makers with materials for understanding the most enigmatic links in the production-consumption chain.

However, we will have reached our goal if, from this small selection of examples, the reader can get an idea of the richness and uniqueness of the semiometric approach, and can now have a better glimpse of the opportunities offered in terms of description and analysis.

Conclusion

Never has a survey questionnaire been so simple in the design of its questions. Never has a questionnaire been so neutral, so basic one might say, in terms of reactions to the stimuli offered: either pleasant or unpleasant. So we have a simple list of words which, however arbitrary it may seem, has revealed structural features in the form of stable dimensions.

We had to carry out tens of thousands of questionnaires to validate these structural features in time and space, and to establish a benchmark for the socio-cultural positioning of individuals and groups of individuals. Despite the perilous ordeal of translating the questions into another language, the same traits appear in several different countries, thus confirming that the results do not depend on the exact composition of the list, and that the resulting structure is neither local nor national. ...

Because of its transparency and distance vis-à-vis current affairs, this marker is suitable for intergenerational, interregional, and international comparisons. But we can also position attitudes, opinions, products, media outlets etc. in this benchmark, by conducting *ad hoc* surveys.

This amazing tool has itself been the subject of lot of experiments the details of which have been given in the preceding chapters, and will continue to be so.

The only stumbling block left would be the choice of the word list. Several of the items presented in this work show indeed that this choice does not take away the obstacle, and that the operational nature of this list heralds in other results and other developments.

A pragmatic way to "experience the difficulty", if not to remove it, is to say that we do not observe a universal structure, but the trace of what could be a universal structure. A "total questionnaire", as fictional and unrealistic as the "total Ph.D" of Ionesco, would allow us to observe that structure. The list of words chosen only contains a trace, a kind of projection of the list, conditioned both by the criteria of cost and feasibility. Whereas it is almost impossible to prove that this "projection" is the best *vis-à-vis* these criteria, we have shown that it is neither fragile nor uninteresting.

It is not surprising if we learn from this experience - and perceive avenues of future work and research - in very different registers, within

mathematical statistics, sociology, linguistics, and finally, survey methodology.

- How does one go from pleasure to meaning?

The first question is prompted by the sheer simplicity of the way we question. How can questions based only on pleasant or unpleasant sensations lead to such a fine local semantic structure... or to such a stable pattern of values? Emotion and satisfaction may play an important role in the acquisition of language as they do in all forms of learning, and even, surely, in the gradual building of a language. At the word level, few other bipolar qualifications could apply exhaustively. The operative nature of the responses was 'pleasant' proof of the appropriateness of the scale we chose.

Further research, however, remains to be done in order to improve the analysis of the mechanisms that have actually led from the pleasant to the relevant. In the case of the open-ended questionnaire, used in Chapter 4, in which one no longer scores a fixed list of words, it is possible that the concept of "importance" is more appropriate than that of "agreeable" to make a more stringent selection and a more stringent filtering, and limit in doing so, the character, perhaps too playful, in the free context of this exercise. This playfulness contributes to enhancing, as we have seen, the extreme dispersion of the vocabulary in the individual responses.

Work on open-ended questions should therefore stimulate further research, including quite cumbersome experimental phases.

- Lessons on surveys designed to measure changes ... and also on developments in the surveys

It is generally accepted that measures of differences between answers to subjective questions, differences over time or between classes, are much more justified and prudent than absolute measurement. The results obtained about the simultaneous drift of both the instrument of observation and the observed reality are a (small) stone in the garden of those who think it is always permissible to observe the evolution of an issue whose formulation is fixed in time. The extreme homogeneity and the relative timelessness of "issues" of the semiometric questionnaire allow us to study these phenomena of attitudes and the problems of participation and scorings rather finely. Indeed, the evolution over time does illustrate the lack of

neutrality and therefore the involvement of the observational instrument in both the outcomes and quality of these results[182].

- The semiometric model and "double convergence"

Among the vital contributions of statistical techniques known as exploratory multivariate data analysis, we can mention on one hand the renewed interest brought to individuals in surveys, or more exactly brought to the "individual dimension"[183]; and, on the other hand, the consideration of variables that can be very numerous. Correspondingly, the classical probabilistic models, more unrealistic in multivariate cases than in the case of one or two variables, are used less, often leaving room for empirical validation methods based on computationally intensive calculations such as bootstrap techniques, which have been widely used in this book. Individuals have gained in importance at the same time as the variables have become more trivial, i.e. the rows (in general: individuals) and the columns (in general: the variables) of the classical table of data have played a more and more symmetrical role[184].

Hence the potential for a *double convergence* with the rows (individuals) and / or columns (variables), a convergence, which generalizes somehow the law of large numbers and other statistical limit laws, based on the infinite multiplication of individuals alone ... These problems are just mentioned here, but belong to an important research area.

This new review of data tables has been, and will be, because these innovations are far from having penetrated all disciplines, very rewarding, as they are for many applications in the human sciences. Multiplying the variables changes the nature of the problems.

The market research practitioners are well aware that a few very rich interviews are sometimes worth more than hundreds of poor ones... which clinicians have known for a long time now. Ultimately, let's take a chance

---

[182] Note that the very gradual degradation of participation in surveys (concerning both the *a priori* rejection and the quality of completion) is a phenomenon widely observed among the pollsters and the control bodies of the profession. The semiometric questionnaire, and also, as we have seen, the open-ended questionnaire in Chapter 4, can help us to study aspects of this phenomenon, in particular, the commitment and level of participation of the respondents.

[183] Individuals as statistical data (or observations) served mainly to compute averages or correlations, and were only rarely represented on charts with identifiers.

[184] The operation of compression at the basis of factorial methods, singular value decomposition, give a symmetrical role to rows and columns (see Annex A1.2).

with the following joke: in the *Essays of Montaigne* (sample size: 1) we learn much more about man in general and men of the sixteenth century in particular, than we would have obtained from a satisfaction survey with several thousands of his contemporaries. Montaigne is also well aware when he wrote: "Each man bears the entire form of the human condition", an idea which Sartre expresses more succinctly: "Every man is the whole of mankind"[185].

And what about Semiometry in all this? It makes up the emblematic example of this paradigm, the *leading case* of the double convergence. Indeed, there is a parent population of rows (the UK population, for example) and a parent population of columns (the vocabulary of the English language, for the same example). Few questionnaires are in such a favourable theoretical situation. But just taking random words in a dictionary would lead, a hundredfold, to the difficulties encountered in connection with the open questions in Chapter 4[186]. In parallel with the theoretical work that would be interesting to conduct within this complex statistical model, future experiments should also address the problems of selecting and sampling words.

- Overall structure, fractal or individual?

If the measurements are made on individuals, the correlations clearly reflect the relationships measured at the individual level. Some attribute a high score to both the words *Discipline* and *Homeland*, others a lower score, but there are few individuals who attribute a high score to *Discipline* and a low one to *Homeland*. The correlation stems precisely from this internal consistency for each individual; a consistency that the composition of the questionnaire allows us to check, and that some minor redundancies existing in the questionnaire can be confirmed. If there is a correlation structure, statistically significant, within a sample of individuals taken at random, we must find the same structure in another random sample, or in a sub-sample taken also at random.

It is not about this classic phenomenon that one can speak of *fractality*, but about the more paradoxical phenomenon which follows: categories

---

[185] Sartre, J.-P. (1980) *Words*, Gallimard, Paris.

[186] Open a dictionary and point blindly and randomly to a series of words "*, fold, hectare, monazite, poppy, prelude, resignation, spindle* ...." The limits of feasibility are immediately noticeable: we must be restricted to words known *a priori* by any sample of respondents (a small fraction of the dictionary entries) and then introduce a series of constraints that ultimately lead to a similar protocol (but not necessarily identical) to that which prevailed during the first tests of Semiometry (see Chapter 1).

seem to disagree on some axes significantly. By segmenting according to these categories, we take a deliberately non-random subsample. And one is wonderstruck, this time, to find the same structure within these sub-categories. In the case of the opposition between men and women, the surprise stems perhaps from a premature identification of certain archetypes. The highly significant association between the words belonging to "Attachment" and individuals belonging to the female group suggests that the difference between sex categories is solely responsible for the formation of the axis "Attachment / Detachment". The existence, nevertheless, of the same antithesis within each category therefore shows that the opposition between males and females is in no way responsible for the structure[187].

In fact, what is called "fractality" could be just the "quasi-individual" validity of a structure, which could be both semantic and psycho-cultural. The local proximity observed on semiometric maps are influenced by semantic links that are universal because they are specific to the vocabulary of a language common to all the respondents. The main divergences observed on the principal axes come from latent factors that could depend on some psychic organization, inseparable from a general cultural context (references, values), also common to all the respondents, with occasional shades of meaning according to different socio-demographic categories. The reader will have understood that the use of the conditional tense means in fact that the issues are for deeper future research.

- Toward structural statistics

Highlighting hidden structures is not a routine activity of statistics, although this approach historically emerged during the early twentieth century - in the context of specific models – with the works of psychologists using factor analysis.

We have been interested in this book in the concept of *statistical facts*, considering the tools used (principal components analysis, classification, Kohonen maps) as mere instruments for observing the *multidimensional* as microscopes or X-ray equipment are instruments for observing *small* or

[187] It is stable even in the usual socio-demographic groups... but the order of axes may, however, be modified, thus the second axis, strongly linked to the age of the respondents, goes to third place if we investigate a narrow age range like the 40-45 year-old age group.

opaque things[188]. Using a large variety of validation procedures (specific methods of re-sampling, recoding, various analytical transformations, extensive empirical procedures), we have shown that *valid structures* could be extracted without simply contemplating *suggestive patterns*.

In fact, we are only beginning to be able to implement tools that can give a scientific character to data analysis designed decades ago[189] ... and it is this research approach to find *statistical facts* or *validated structural features* which was termed *structural statistics*. The work presented in previous chapters could pave the way to the beginning of a particular branch of the discipline.

[188] The theories and mechanisms that govern the implementation of a microscope have little or no relation to the morphological properties of animalcules, cells or other objects observed. They do not imply any modelling of the objects observed.

[189] While many statisticians have used these observational instruments with the discernment and critical distance necessary, some uses were unfortunate, in particular because of the lack of a conceptual framework able to accommodate new visualizations. A rather striking analogue situation is the release of the first microscopes as described by Francois Jacob in *the Logic of Life*: "When Leeuwenhoeck contemplated for the first time a drop of water under a microscope, he found an unknown world, with forms swarming ... But the theory and ideas had little to do with all this world ... this discovery could only feed conversation ... "

APPENDIX 1

Some elements of Multivariate Descriptive Statistical Analysis

The exploratory statistical analytical methods used in previous chapters aim at describing large data sets, at extracting these structures and at validating them. They belong to exploratory multivariate statistics, data analysis, or data mining. Those three expressions are roughly equivalent in the uses of this book. We have used the expression structural statistics to highlight the emphasis put on the validation phase of structures. These methods generalize classical descriptive statistics, using fairly intuitive mathematical tools, but ones that are more complex than the means, variances and empirical correlation coefficients of elementary descriptive statistics.

We find in this appendix the principles of the techniques used or mentioned in previous chapters: principal components analysis is the basic principal axes technique for the applications in Semiometry. Some developments of the work noted [SEM 2006][190] will be included and will be supplemented by more recent work on validation methods, and in particular, on the techniques known as bootstrapping, or Kohonen maps, or on techniques of analysis less frequently used such as logarithmic analysis. Note that all the computations involved in this book have been carried out with the free software DtmVic (Acronym for *Data and Text Mining, Visualization, Inference, Classification*"). DtmVic can be freely downloaded from the website: www.dtmvic.com. One of the provided example concerns semiometric data.

[190] *Statistique Exploratoire Multidimensionnelle*, L. Lebart, M. Piron, A. Morineau, Dunod, 2006 (in French). This book is an updated version oft he book: *Multivariate Descriptive Statistical Analysis* (L. Lebart, A. Morineau, K. Warwick), J. Wiley, New York, 1984.

A1.1 A reminder of the principles of exploratory multivariate methods

The exploratory multivariate methods cover a large number of techniques that aim to describe and synthesize the information contained in large data arrays.

A1.1.1 Geometric representations and scattering of points

Initially, the data are in the form of large rectangular arrays, denoted **X**. The rows (or lines) (i = 1, ..., n) in the array represent the n individuals, the subjects or respondents surveyed, for example, and the columns (j = 1, ... m) represent m variables that can be measures, or scores measured on those individuals.

To understand the principle of exploratory multivariate statistical methods, it is useful to represent geometrically the set of n individuals (n rows) and the set of m variables (m columns) as two clouds of points, each set being described by the other one.

Table A1.1 : Example of Array X of scores (1 to 7) attributed to: m = 7 words, for n = 12 respondents

words respondents	tree	gift	danger	morality	storm	politeness	sensual
R01	7	4	2	2	3	1	6
R02	6	3	1	2	4	1	7
R03	4	5	3	4	3	4	3
R04	5	5	1	7	2	7	1
R05	4	5	2	7	1	6	2
R06	5	7	1	5	2	6	5
R07	4	2	1	3	5	3	6
R08	4	1	5	4	5	4	7
R09	6	6	2	4	7	5	5
R10	6	6	3	5	3	6	6
R11	7	7	6	7	7	6	7
R12	2	2	1	2	1	3	2

We then define, for the two clouds, the distances between row points and between column points that reflect the statistical associations between individuals (rows) and between variables (columns).

In the case of Semiometry, a word (variable) is a point whose coordinates are the scores given by the n individuals (respondents): the

cloud of m words is embedded within an n-dimensional space. Similarly, an individual is a point whose coordinates are the scores for the m words, the cloud of n individuals is embedded within an m-dimensional space.

Figures A1.1 and A1.2 illustrate, from Table A1.1 containing the scores for 7 words given by 12 respondents, the representation of both those intrinsically linked clouds of points.

The cloud of word points is built within the space of respondents, here from only two individuals, R04 and R08, since two dimensions make it possible to have a graph on a plane (see Figure A1.1).

words respondents	tree	gift	danger	morality	storm	politeness	sensual
R04	5	5	1	7	2	7	1
R08	4	1	5	4	5	4	7

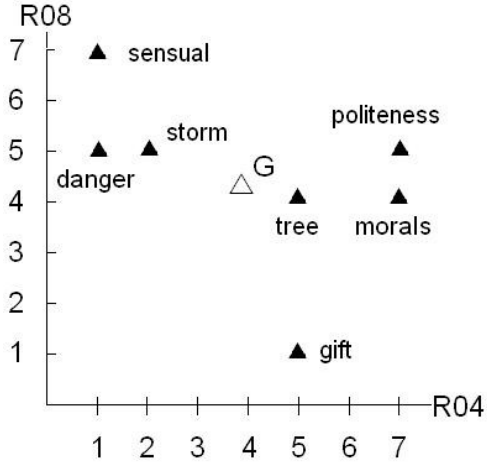


Figure A1.1: Representation of the cloud of words in the space of two respondents "R04" and "R08"

Similarly, the cloud of 12 respondents is built in the space of variables, here from two words, **Morals** and **Sensual**, viz. within a two dimensional space (see Figure A1.2). (Please, check from the corresponding columns of Table A1.1)

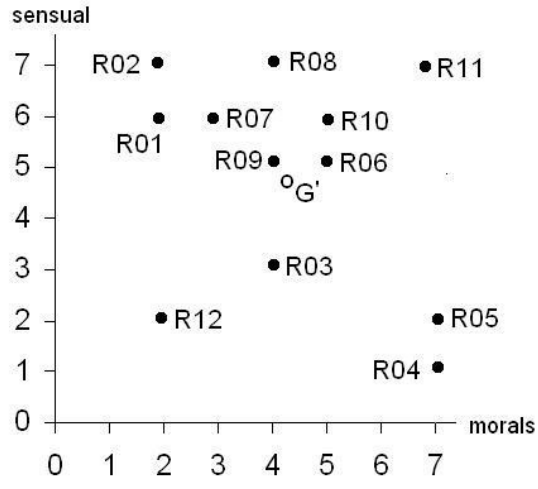


Figure A1.2: Representation of clouds of respondents in the space of words "Sensual" and "Morals"

For each cloud, the mean point called centre of gravity is shown. This is G for the centre of gravity of the scores assigned by respondents (see Figure A1.1) and G' for that of respondents who rated the two words Sensual and Morality.

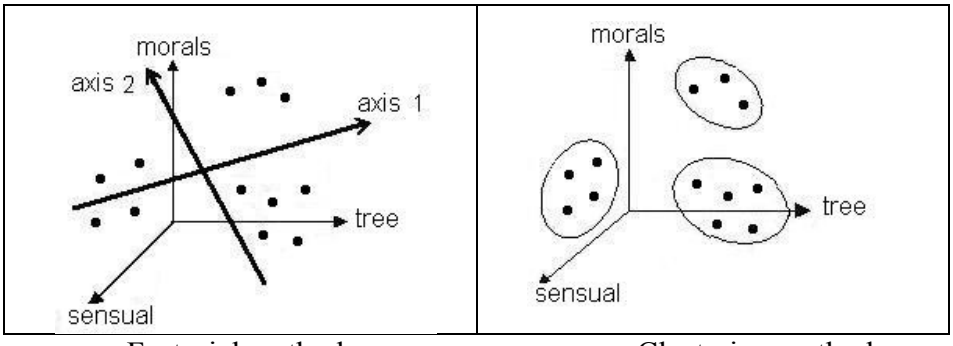
A1.1.2 Principle and methods of analysis

While it is still possible to calculate distances between the rows and distances between the columns of an array **X**, it is not possible to visualize them immediately (the geometric representations associated usually involving spaces of more than two or three dimensions): it is then necessary to carry out transformations and approximations to obtain a flat (planar) representation.

Tables of distances associated with these geometric representations (simple in principle but complex because of the large number of dimensions of the areas concerned) can be described by the two main families of methods: the factorial methods (or principal axes methods) and clustering.

The first one consists in finding the main directions according to which the points deviate most from the average point.

The second is to search for groups or clusters of individuals that are as homogeneous as possible (Figure A1.3).



Factorial method (search for principal axes) | Clustering method (search for homogeneous groups)

Figure A1.3: Two major families of methods

These methods often involve, in the same way, individuals (rows) and variables (columns). The comparison of spaces of individuals and spaces of variables enriches the interpretations.

A1.2 Factorial methods: technical aspects

Factorial methods[191] can simultaneously handle large amounts of data and their system of correlations. Through a technique producing a kind of data compression, they can bring out the internal structure of the data, especially as planar graphical displays.

- Search for factorial (or: principal) subspaces

The goal is to find sub-spaces of smaller dimensions (between three and ten, for example) that best fit the cloud of individual-points and the cloud of variable-points, so that the proximities measured in these sub-spaces reflect as much as possible the actual proximities. This gives a representation space, the factor space, defined by the principal "axes of inertia". It is possible to represent the points of the cloud in this system of axes (see Figure A1.4). These axes achieve the best fit of all the points according to

[191] They include in the French statistical literature of the last thirty years all the techniques of representation using the "principal axes": principal components analysis, single and multiple correspondence analysis, factor analysis or common and specific factor analysis.

the classic least square criterion which involves minimizing the sum of the squared differences between points and axes.

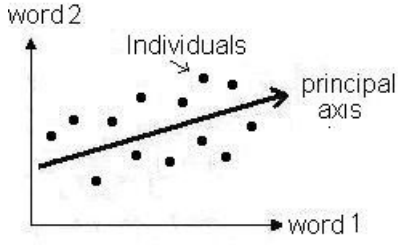


Figure A1.4: Fit of the cloud of individual points in the word space

The first of these axes corresponds to the line of maximum elongation of the cloud of points, the second axis maximizes the same criterion while being orthogonal to the first axis, and so on for the following axes, which are all mutually perpendicular. This orthogonality implies the absence of correlation between pairs of axes.

The fitting procedure is exactly the same for both clouds. We can then find simple relationships linking the axes calculated in both spaces, one for the individuals and one for the variables (*transition relationships*).

The vector of the coordinates of points on each axis, called a factor, is a linear combination of initial variables. We denote by φ_α and ψ_α the factors corresponding to the axis α in the space noted $\mathbb{R}^m$ (a space whose n points have as coordinates the m words) and in the space noted $\mathbb{R}^n$ (a space whose n points have for coordinates the n individuals).

X is the array of data having undergone preliminary transformations (reduced centered variables, for example), **X'** is its transpose.

$\mathbf{u}_1$ is the unit vector that characterizes the first axis. $\lambda\mathbf{u}_1$ is then the eigenvector of the matrix **X'X** corresponding to the largest eigenvalue [SEM 2006].

More generally, the subspace with q dimensions, which fit best (in the least squares sense) to the cloud, is generated by the first q eigenvectors of the matrix **X'X** corresponding to the q largest eigenvalues.

The two clouds of points, that of the words and that of the respondents, are intrinsically linked and, in fact show two facets of the same structures:

in one case, the factors describe the correlations between words; in the other, the associations between respondents.

Each of the factorial planes of visualization used throughout this book corresponds to a pair of factors. The most used plane of Semiometry is the (φ_2, φ_3) plane.

The elements (words or individuals) involved in the computation of the axes are the active elements. We also introduce, in the analysis, some supplementary elements (or illustrative elements) that do not participate in the formation of the axes but are projected subsequently onto the factorial planes. These elements can be of utmost importance in the interpretation of the planes (see Section A1.2.4).

- *Basic techniques and derived methods*

The nature of the information, its coding in the data array, the specific application domain will introduce variations in the factorial methods. Those used here are in fact derived from two basic techniques, principal components analysis and correspondence analysis.

The principal components analysis applies to an array of numerical measurements and will be used, within the framework of Semiometry, to process an array of scores. Examples of textual data analysis, presented in Chapter 4, are based on correspondence analysis applied to lexical tables (contingency tables, cross-tabulating words and texts).

A1.3 Principal Components Analysis: technical aspects

Principal Components Analysis (Hotelling, 1933) applies to variables with numerical values (measurements, rates, words etc.) represented as a rectangular array of measures **R** whose general term is $\mathbf{r}_{ij}$, whose columns are variables, and whose rows represent the individuals about whom these variables are measured. In Semiometry, the variables are the words, the rows the respondents, and the numerical values in the entries of the table, the scores.

A1.3.1 Geometric Interpretations

Geometric representations between the rows and the columns of the data table allow one to visualize the proximities, respectively between individuals and between variables (see Figures A1.1 and A1.2 above).

In $\mathbb{R}^m$, two individual points are very similar, if, overall, their m coordinates are very close. Both respondents concerned are therefore characterized by almost equal values for each variable. The distance used is the usual Euclidean distance.

In $\mathbb{R}^n$, if the values taken on by two particular variables are very close for all respondents, these variables will be represented by two very close points in this space. This may mean that these variables measure the same thing or they are bound by a particular relationship.

But the units of measurement of the variables can be very different, and therefore make it necessary to transform the data table beforehand.

A1.3.2 Problems of scale of measurement and data transformation

We want the distance between two individuals to be independent of the units of the variables for each variable to play a similar role. For this, we assign to each variable j the same dispersion by dividing each of its values by its standard deviation s_j with

$$s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(r_{ij} - \overline{r}_j)^2$$

.

Furthermore, we are interested in how individuals differ from the mean. We then place the midpoint at the centre of gravity of the cloud of individuals. The coordinates of the midpoint are the average values of the variables noted

$$\overline{r}_j = \frac{1}{n}\sum_{i=1}^{n} r_{ij}$$

Taking this point as the origin amounts to subtracting for each variable its mean: $\overline{r}_j$.

In this way, we correct the scales by transforming the data array **R** into a new table X as follows:

$$x_{ij} = \frac{r_{ij} - \overline{r}_j}{s_j\sqrt{n}}$$

The variables thus reduced and centered all have a variance $s^2(x_j)$, equal to 1, and an average $\overline{x}_j$ null, and become comparable. Other preliminary transformations are possible (see Section 2.5 of Chapter 2).

A1.3.3 Analysis of the cloud of n respondents

Data transformation leads to a translation from the original to the centre of gravity of this cloud and changing (in the case of the analysis we call normalized) the scales on the different axes.

To achieve the analysis of the cloud of respondent points in $\mathbb{R}^m$, the matrix **X'X** to be diagonalized in this space, is the correlation matrix whose general term is :

$$c_{jj'} = \sum_{i=1}^{n} x_{ij} x_{ij'} = \frac{1}{n} \sum_{i}^{n} \frac{(r_{ij} - \overline{r}_j)(r_{ij'} - \overline{r}_{j'})}{s_j s_{j'}}$$

$c_{jj'}$ is the correlation coefficient between variables j and j '.

The coordinates of n individual points on the factorial axis are the n components of the vector **Xu**.

Figure A1.5 below illustrates the representation of the cloud of respondents for a table of 12 respondents having rated 7 words (already presented in section A1.1) in the principal plane (2, 3)[192]. Respondents R01 and R02 have given in the same way, very contrasted scores and have given high marks to Tree and Sensual and low grades to Morals and Politeness, and are therefore near in the plane and differentiate with respondents R05 and R04 who have expressed themselves reversely on these words. Respondent R08 is distinguished by having highly rated Danger without giving a high score to the other words, while R11 highly rated all the words.

A1.3.4 Analysis of the cloud of variables (words)

The factorial coordinates $\varphi_{\alpha j}$ of variable points on the α axis are the components of

[192] The plane (2, 3) has been considered the primary plane of Semiometry given the specific nature of the first axis (axis responsible of a “size effect”, cf. Chapter 3)

$$\mathbf{u}_\alpha \sqrt{\lambda_\alpha}$$

and we have:

$$\varphi_{\alpha j} = cor(j, \boldsymbol{\psi}_\alpha)$$

The coordinate $\varphi_{\alpha j}$ of a variable-point j on the axis α is none other than the correlation coefficient of this variable with the factor ψ_α (a linear combination of initial variables) considered as an artificial variable whose coordinates are made up of the n projections of the individuals on this axis.

The factorial axes are mutually orthogonal, we thus obtain a series of artificial variables, uncorrelated with each other, called principal components[193], which summarize the correlations of all initial variables.

Table of scores (1-7) given to 7 words by 12 respondents (Reminder)

	tree	gift	danger	morals	storm	politeness	sensual
R01	7	4	2	2	3	1	6
R02	6	3	1	2	4	1	7
R03	4	5	3	4	3	4	3
R04	5	5	1	7	2	7	1
R05	4	5	2	7	1	6	2
R06	5	7	1	5	2	6	5
R07	4	2	1	3	5	3	6
R08	4	1	5	4	5	4	7
R09	6	6	2	4	7	5	5
R10	6	6	3	5	3	6	6
R11	7	7	6	7	7	6	7
R12	2	2	1	2	1	3	2

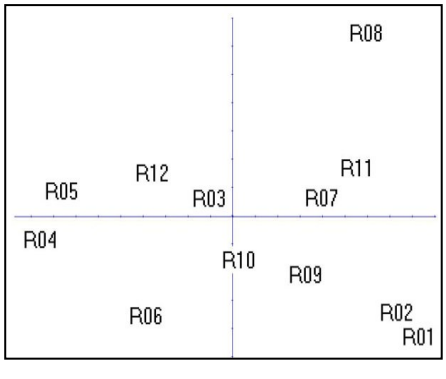
In Figure A1.5-b, as on the corresponding correlation matrix, Politeness and Morals are highly correlated and to a lesser extent, Storm and Sensual. We can observe the behavior of respondents: R01 and R02 are going in the direction of good raters of both words Tree and Sensual and poor raters of Morals and Politeness in contrast to the respondents R04 and R05.

[193] Principal components analysis translates only linear relationship between variables. A low correlation coefficient between two variables means that they are linearly independent, whereas there may be a nonlinear relationship.

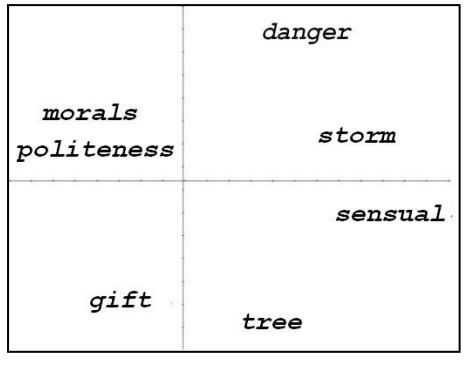
Variables strongly correlated with an axis will contribute to the definition of that axis[194]. This correlation can be read directly on the chart as it is the coordinate of point j on the axis.

Correlation matrix

	tree	gift	danger	morals	storm	polite	sensual
tree	1.00						
gift	.55	1.00					
danger	.29	.14	1.00				
morals	.16	.62	.36	1.00			
storm	.51	.09	.54	-.01	1.00		
polite	.00	.63	.23	.91	-.05	1.00	
sensual	.56	-.08	.45	-.30	.68	-.37	1.00



A1.5-b: Representation of respondents on the plane (2,3)



A1.5-b: Representation of words on the plane (2,3)

Figure A1.5 : Principal Components Analysis in the table of scores of 7 words by 12 respondents

We have to especially focus on variables with the highest coordinates and we can then interpret the principal components based on the clustering of some of these variables and on the opposition with others.

[194] The example is obviously not representative enough for the plane to be interpreted. It is just intended to bring the data table and the results nearer.

Note that while all the variable points lie on a sphere of radius 1, centered at the origin of the axes[195], the planes of fit cut the sphere according to large circles (of radius 1), the circles of correlations, within which are positioned variable-points. In this book, the circles are not plotted on the factorial planes representing the words for a better readability of labels (the boxing of factorial planes would have indeed led to a significant reduction of scale).

A1.4 Correspondence Analysis

Correspondence analysis[196] applies primarily to a contingency table **K** (cross-tabulation), with n rows and p columns, which describes the distribution of the population according to two qualitative (or categorical) variables with n and p categories. The rows and columns thus play similar roles. In Chapter 4, the analysis is applied to a table cross-tabulating 1,191 respondents spontaneously answering open questions with the 592 words cited at least 4 times as being agreeable. A second analysis focuses on 158 words that appear over 25 times.

- Notation

Let $k = \sum_{ij} k_{ij}$ be the sum of all the elements k_{ij} of the contingency table K. In the case of the open questions table, k represents the number of times that the 592 words were mentioned spontaneously.

We note $f_{ij} = k_{ij} / k$, the relative frequencies, with $\sum_i \sum_j f_{ij} = 1$.

---

[195] The analysis of the cloud of variable points in $\mathbb{R}^n$ is not performed with respect to the centre of gravity of the cloud (unlike that of individual points) but with respect to the origin. The distance of a variable j to the origin O is expressed by:

$$d^2(O, j) = \sum_{i=1}^{n} x_{ij}^2 = 1$$

[196] Introduced and studied systematically as a flexible technique of exploratory analysis of multidimensional data by Benzécri J.-P. (1973), correspondence analysis has other precursors, in particular, L. Guttmann, C. Burt, C. Hayashi (1956), and has given rise to scattered and mutually independent works.

We note : $f_{i.} = \sum_j f_{ij}$, $f_{.j} = \sum_i f_{ij}$, the marginal relative frequencies.

The contingence table K is transformed into both an array of row profiles $f_{ij} / f_{i.}$ and an array of column profiles, $f_{ij} / f_{.j}$.

Point i of $\mathbb{R}^m$ has the coordinates: $f_{ij} / f_{i.}$ for every j ≤ m.

Likewise, point j of $\mathbb{R}^n$ has the coordinates : $f_{ij} / f_{.j}$ for every i ≤ n.

We can observe a significant difference between correspondence analysis and principal components analysis: the transformations on the table in both spaces are identical (because the row set and the column set play similar roles).

- *Chi-square distance (χ^2) and distributional equivalence*

The distances between two row points i and i', on the one hand, and between two column points j and j', on the other, are given by the following equations:

$$d^2(i,i') = \sum_{j=1}^{m} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2 \quad d^2(j,j') = \sum_{i=1}^{n} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2$$

The distance of the χ^2 offers the advantage of verifying the *distributional equivalence principle*. This principle ensures the robustness of the results of the correspondence analysis vis-à-vis the arbitrary division into categories of the nominal variables. It is expressed as follows: If two rows (resp. columns) of the contingency table have the same profile (i.e.: are proportional) then their aggregation does not affect the distance between the columns (resp. rows). This aggregation results in a new row point (resp. column point) with the same profile and to which is assigned the sum of the frequencies of the two row points (resp. column points). This property is important because it guarantees a certain invariance of the results vis-à-vis the nomenclature chosen for the construction of the categories of a qualitative variable.

A1.5 Logarithmic analysis

Logarithmic analysis, proposed by Kazmierczak (1985), similar but not identical to the *Spectral Maps* of Lewi (1976), realizes the property of distributional equivalence of correspondence analysis on tables that are not necessarily contingency tables. Kazmierczak uses and generalizes the principle of Yule, which states that one does not change the distance between two rows or the distance between two columns of a table by replacing the rows and columns of this table by any other rows and columns that are proportional (this is actually a generalization of the principle of distributional equivalence).

Logarithmic analysis involves taking the log of data (after possible addition of a constant in case of possible negative or null data). Then, after centering both rows and columns, the data table is submitted to a non-normalized principal components analysis, which coincides here with a singular value decomposition [SEM 2006].

Note that if **R** is a table of data (n, m) and if **A** and **B** are two diagonal matrices respectively of dimensions (n, n) and (p, p) with positive diagonal elements, the **ARB** matrix gives rise to the same logarithmic analysis as the matrix **R**. This invariance property has had the effect of suppressing the first semiometric axis (size effect) without altering the following axes (Section 3.7 of Chapter 3).

A1.6 Factor analysis *(into common and specific factors)*

Factor analysis (or analysis into common and specific factors) is probably the oldest model of latent variables[197]. These models were mainly developed by psychologists and psychometricians. The developments to which they give rise are complex and diverse. We can consult on this point the classic books of Harman (1967) and Mulaik (1972)[198].

We should also mention the work of Anderson and Rubin (1956) and Lawley and Maxwell (1963), who placed factor analysis in a classic inferential framework.

---

[197] The original principles of the method were given by Spearman (1904) (univariate analysis) and Garnett (1919), Thurstone (1947) (multivariate analysis).

[198] In econometrics, we usually distinguish between functional models, or fixed-effect models (such as multiple regression and the linear model as a whole), and structural models or random effect models (models of latent variables).

- *The model of factor analysis*

This model proposes to reconstruct, from a small number q of factors, the correlations between m observed variables. We assume the existence of an *a priori* model:

$$\underset{(m,l)}{\mathbf{x}_i} = \underset{(m,q)}{\mathbf{\Gamma}} \underset{(q,l)}{\mathbf{f}_i} + \underset{(p,l)}{\mathbf{e}_i}$$

In this equation, $\mathbf{X}_i$ represents the i-th vector observed from m variables ; Γ is a table (m, q) of unknown coefficients (with q < m) ; $\mathbf{f}_i$ is the i-th value of the random vector and the unobservable common q factors ; and $\mathbf{e}_i$ the i-th value of the unobservable vector de residuals, the latter represent the combined effect of specific factors and a random disturbance.

We denote **G** the table (n,p) whose i-th row represents the observation i. Likewise **F** denotes the unobservable table (n,q) whose i-th row is $\mathbf{f}_i^{'}$ and E the unobservable table (n,p) whose i-th row is $\mathbf{e}_i^{'}$. The model linking all the observations to the hypothetical factors is written:

$$\underset{(n,m)}{\mathbf{X}} = \underset{(n,q)}{\mathbf{F}} \underset{(q,m)}{\mathbf{\Gamma}'} + \underset{(n,m)}{\mathbf{E}}$$

In this equation, only **X** is observable. As such, the model is indeterminate.

The identification of this model and the estimation of parameters raise complex problems. A series of additional *a priori* assumptions allows that identification. The application within this book is in Section 2.6 of Chapter 2.

A1.7 Methods of hierarchical clustering

Automatic clustering techniques[199] are designed to produce clusters of objects or individuals described by a number of variables or characteristics. The circumstances of use are substantially the same as in descriptive factorial analysis methods presented in previous sections. In chapter 3, the

[199] Clustering is a branch of data analysis and a fundamental step in many scientific disciplines. It has given rise to numerous and diverse publications including: Sokal and Sneath (1963) and Benzécri (1973).

clustering is performed on all 210 words from the coordinates of these words on the principal axes.

There are several families of clustering algorithms: hierarchical algorithms that provide a hierarchy of partitions of objects; and algorithms that lead directly to partitions such as the methods of aggregation around mobile centers (Aka: K-means algorithm). The principles, common to various techniques of ascending hierarchical clustering, are simple. One must create, at each step of the algorithm, a partition obtained by aggregating the closest elements pairwise.

- *The basic algorithm of hierarchical clustering*

The basic algorithm of ascending (bottom up) hierarchical clustering produces a hierarchy starting from the partition in which each element to be classified constitutes a class, leading to the partition consisting of one single class containing all the elements.

For n elements to be classified, it is composed of n steps. At the first stage, there are therefore n elements to classify. We construct the distance matrix between the n elements and we look for the two closest, which are aggregated into a new element.

We construct a new distance matrix resulting from the aggregation, by calculating the distances between the new element and the remaining elements. We are now in the same conditions as in step 1, but with only (n-1) elements to be classified.

We look again for the two closest new elements, which we aggregate. The process is reiterated until there is only one element containing all the objects and which is the final partition.

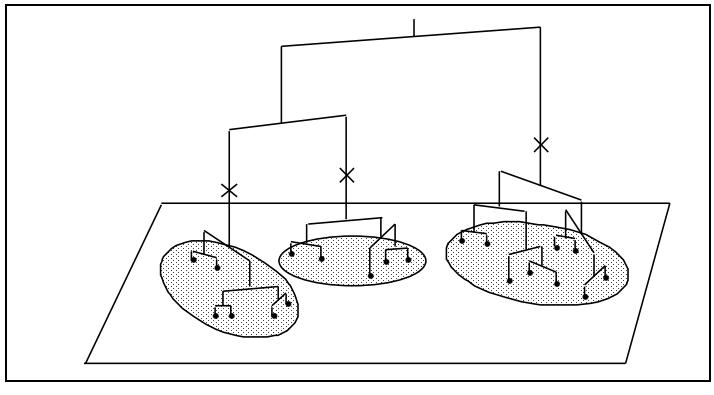


Figure A1.6: Dendrogram or hierarchical tree

The algorithm does not provide a partition in q classes of a set of n objects; but a hierarchy of partitions, which appears under the form of trees also called dendrograms, and containing n - 1 partitions (see Figure A1.6). The value of these trees is that they can give an idea of how many classes actually exist in the population. Each cut of a dendrogram provides a partition.

In Table 3.1 of Chapter 3, we chose three cuts of the dendrogram in 36, 24, and 12 classes. The coarsest of the cut into 12 classes is shown in the first column, in 24 classes in the second column, and the finest cut in 36 classes, in the third column. It is followed by the list of words that make up the 36 classes.

A1.8 The self-organizing maps ((*Kohonen maps*)

The goal of self-organizing maps is to classify a set of observations in order to maintain the initial topology of the space in which they are described. Like the neural networks to which they are associated, these maps obtain good performances for pattern recognition[200]. They were used in Chapters 3 (Section 3.1) and 4 (Section 4.2).

- *The principle*

Kohonen maps seek to represent in a space with two (sometimes three) dimensions the rows or columns of a table in accordance with the concept of neighborhood in the space of the elements to be classified. Like PCA, it is useful to imagine at the start all the data (words) as a cloud of points in a high-dimensional space (that of the individuals or respondents).

The principle is to consider a map as a rectangular grid (sometimes hexagonal). This grid, once unfolded, fits the shape of the cloud of points as best as possible. The grid nodes are the *neurons* of the map. Each point in the original cloud is projected onto the node that is the closest. In fact, each point, first described in a multidimensional space, is represented in the end by two coordinates giving the position of the node on the map: the space is reduced. The set of points assigned to a single neuron are close in the original space.

[200] Introduced by Teuvo Kohonen in 1981, they can be considered as neural methods (cf. Kohonen, 1989). They give rise to several applications such as text analysis, medical diagnosis, industrial process controls, and robotics.

We define *a priori* the notion of neighborhood between classes. These neighborhoods can be chosen in various ways but generally are assumed to be directly contiguous in the rectangular grid (representing 8 neighbors for a neuron).

Neighboring observations in the space of variables of dimension q belong, after clustering, to the same class or to neighboring classes.

- *The algorithm*

The learning algorithm for classifying m points is iterative. The initialization consists in associating with each class k (node) a temporary centre C_k (with q components) chosen randomly in the q-dimensional space containing the m words to classify.

At each step, we choose a word i at random to be compared to all the provisional centers and we assign the word to the closest center $\mathbf{C}_{k_0}$ (closest in the sense of a given distance *a priori*). Then we bring the center $\mathbf{C}_{k_0}$ nearer to the word i, together with the centers neighboring $\mathbf{C}_{k_0}$, which is expressed in step t by:

$$\mathbf{C}_k(t+1)= \mathbf{C}_k(t)+\varepsilon(i(t+1)- \mathbf{C}_k(t))$$

where i (t+1) is the word presented at stage t+1, ε is a parameter of adaptation, positive and less than 1. This expression only concerns the centre $\mathbf{C}_{k_0}$ and its neighbors.

This algorithm is similar to the k-means algorithm, but in this latter case, there is no notion of neighborhood between classes; and we only modify at each step the position of the centre $\mathbf{C}_{k_0}$.

Like the k-means algorithm, this algorithm is suitable for applications in which data is abundant and in which there is no need to store it in the central core of the computer.

A1.9 Validation Tools

Throughout this book, we have used the concepts of test value and supplementary (or: additional) category.

Test Values (Section A1.9.1) are tools of elementary statistical inference, but very versatile and useful, especially if the user is aware of problems of multiple comparisons that often appear (section A1.9.2) .

The technique of supplementary categories (Section A1.9.3) is a fundamental validation tool for principal axes methods. It allows for an *external* validation of the results, which is both a test of consistency and an enrichment of interpretation.

The two other validation tools used in this work are the confidence intervals of Anderson and the bootstrap re-sampling procedure.

Anderson confidence intervals (Section A1.9.4) are used in Section 2.2 of Chapter 2 to validate eigenvalues.

The bootstrap re-sampling procedure (Section A1.9.5) is used in Sections 2.3 and 2.4 of Chapter 2 to highlight the stability of semiometric structures, and in Section 4.4 of Chapter 4 to validate a textual analysis.

A1.9.1 What is a test value?

A test value is a criterion to assess quickly whether a category of a nominal variable (i.e. a category of respondents) has a significant position on an axis. For this, we test the hypothesis that a group of individuals, corresponding to a given category of an external categorical variable (such as the category *Female*, for example) may be considered to be drawn randomly, without replacement, from the sample under consideration.

In the case of a true random selection, the mean point (or center of gravity) of the sub-cloud representing the group (i.e. the category) departs somewhat from the center of gravity of the overall cloud corresponding to the entire sample. We then convert the coordinate of this category on the axis into a *test value* which is, under this assumption of independence, the realization of a standard normal variable. In other words, assuming that a variable has a random distribution on the axis, the corresponding test value has 95% chances of being included in the interval [-1.96, +1.96].

We then consider as occupying a significant position the variables whose test values are greater than 1.96 in absolute value, which corresponds

approximately to the usual threshold of 5% probability. Often the test values are well above this threshold. They are then used to sort the categories from the most significant to the least. The test value systematizes the notion of *t-value* often used in the literature.

Suppose a category j concerns n_j individuals. If these n_j individuals are drawn at random (this is called the null hypothesis H_0) among the n individuals analyzed (supposed to be drawn without replacement), the average of n_j coordinates drawn at random from the finite set of n values $\psi_{\alpha i}$ (coordinates of the respondent i on axis α) is a random variable $X_{\alpha j}$ such that:

$$X_{\alpha j} = \frac{1}{n_j} \sum_{i \in I(j)} \psi_{\alpha i}$$

with the expectation $E(X_{\alpha j}) = 0$ and for variance[201]:

$$Var_{H_0}(X_{\alpha j}) = \frac{n - n_j}{n - 1} \frac{\lambda_\alpha}{n_j}$$

In the formula giving $X_{\alpha j}$, I(j) is the subset of respondents characterized by the category j. The coordinate φ_{aj} of the variable j is proportional to the random variable $X_{\alpha j}$ and is written:

$$\varphi_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} X_{\alpha j}$$

And we therefore have $E(\varphi_{\alpha j}) = 0$ and:

$$Var_{H_0}(\varphi_{\alpha j}) = \frac{n - n_j}{n - 1} \frac{1}{n_j}$$

The quantity:

$$t_{\alpha j} = \sqrt{n_j \frac{n-1}{n-n_j}} \varphi_{\alpha j}$$

Is a measure, in terms of standard deviations, of the distance between the variable j, and the origin, on the factorial axis. We call this quantity a "test value". According to the central limit theorem, its distribution tends towards a standardized normal distribution.

---

[201] . This is the classical formula giving the variance of the mean in a drawing without replacement of n_j objects from n, depending on the total variance λ_α , which is also, in the case of factorial coordinates, the eigenvalue corresponding to axis α.

It should be noted that the test values are meaningful only for the supplementary categories (see next section), or active variables with low absolute contributions, viz. behaving in fact as supplementary variables[202].

Test values allow us to identify the significant variables quickly for the interpretation of an axis or a factorial plane.

A1.9.2 The problem of multiple comparisons

The simultaneous calculation of several test values or several probability thresholds comes up against the pitfall of multiple comparisons, well known to statisticians, cf. O'Neill and Wetherill (1971), Saville (1990), Westfall and Young (1993), Westfall et al. (1999), Hsu (1996).

Suppose we projected 100 additional categories (see next section A1.9.3) that are truly randomly drawn. The test values attached to these variables are then all the realizations of standard random reduced independent variables.

Under these conditions, on average, from 100 test values calculated, five are outside the interval [-1.96, +1.96] and will be, in appearance only, significant. The 5% threshold is meaningful only for a single test, and not for multiple tests[203].

We solve this problem in practice by choosing a more stringent threshold. The threshold the most stringent and pessimistic that we can imagine is the "Bonferroni threshold" (the initial threshold is divided by the number of tests: in the case of 210 tests: $0.05 / 210 = 2.4\ 10^{-4}$). The corresponding unilateral value test is 3.49. This value provides us with a prudent safeguard against excessiveness[204].

As has been reported in the body of the text (see, for example, notes in Section 7.2 of Chapter 7), the interdependence of words does not allow us to apply the results for multiple comparisons blindly. What can we conclude indeed when several words with similar meanings have simultaneously test values of about 1.96? These are not significant one by

---

[202] Coordinates on an axis of individuals corresponding to an active category cannot be considered to be drawn at random, since that category has helped build the axis.

[203] Test values can, above all, sort the supplementary categories in order of decreasing interest, which is invaluable for the interpretation of factors.

[204] See, for example, Hochberg (1988), Perneger (1998).

one by applying the Bonferroni threshold, but they confirm and validate one another when considered simultaneously.

A pragmatic solution (in the multidimensional case): the bootstrap.

The bootstrap validation technique, which will be discussed later in this Appendix, makes a valuable contribution to the difficult problem of multiple comparisons. The bootstrap involves replications of samples that take into account all the variables simultaneously, and consequently the interdependence of variables.

This is an overall test, rather than separate tests for each variable. An illustration is given in Figure 4.4 of Chapter 4, which shows simultaneous confidence areas for words, some of which appear to be significantly distinct. In this case, the tests are not performed in isolation or in series, but simultaneously.

A1.9.3 Usefulness of supplementary elements

Principal axes analyses produce representation subspaces for individuals and/or for variables. They rely on elements (individuals or variables) known as **active**.

It is possible to introduce other supplementary points (or supplementary elements) that did not intervene in the composition and definition of the axes. We may want to know their positions in the factorial space[205]. We then project these points after the construction of the factorial axes in this new referential framework. This projection is very simple using the so-called transition formulae, whether this be in principal components analysis or correspondence analysis.

This is the case when one wants to place the words, numerical variables, in the space of control variables (see Section 3.4). We calculate, retrospectively, their coordinates on the factorial axes.

---

[205] We can cite three reasons for making a point a supplementary element:

1) to enrich the interpretation of axes by the variables (thematically or in nature different from that of active elements) that did not participate in their construction;
2) to adopt the perspective of forecasting by projecting the additional variables onto the space of individuals. These will be "explained" by the active variables, and
3) to bring out the essential structure that could be obscured by the existence of active points of low mass which could distort the cloud.

This is also the case when one wishes to characterize the semiometric axes by socio-demographic factors (variables) of the population being surveyed (see Section 1.4).

These criteria actually define groups of individuals, and are regarded either as categories of categorical variables or as individuals, added as supplementary elements.

The centers of gravity of these groups are positioned in the space of the variables. The test value is used to appreciate the significance of the locations of these groups on the axis. This procedure could be used as an alternative method for comparing subpopulations in Chapter 7.

A1.9.4 Anderson Confidence Intervals

Anderson (1963) calculated the limiting distributions of eigenvalues λ_α from a principal components analysis without necessarily assuming that the corresponding theoretical values are distinct.

If the eigenvalues λ_α of the theoretical covariance matrix Σ are distinct, the eigenvalues $\hat{\lambda}_\alpha$ of the empirical covariance matrix S follow asymptotically the normal distribution with expectation λ_α and variance $2\lambda_\alpha^2/(n-1)$ where *n* is the size of the sample.

We deduce the confidence intervals approaching the threshold 95%:

$$\lambda_\alpha \in \left[\hat{\lambda}_\alpha\left(1-1.96\sqrt{2/(n-1)}\right)\ ;\ \hat{\lambda}_\alpha\left(1+1.96\sqrt{2/(n-1)}\right)\right]$$

The amplitude of the intervals gives an indication about the stability of the eigenvalue vis-à-vis the sampling fluctuations (assumed to be normal). Overlapping of the intervals of two consecutive eigenvalues therefore suggest the equality of these values. Thus the user can avoid interpreting an axis, unstable according to this criterion.

Anderson confidence intervals concern in fact both the eigenvalues of covariance matrices as well as the matrices of correlations. The simulations show that the confidence intervals obtained are generally conservative: the percentage of coverage of the true value is usually higher than the confidence level announced.

In all these cases, both the asymptotic nature of the results and the underlying assumption of normality[206] lead us to consider the results as merely indicative.

A1.9.5 Bootstrap techniques

With the results of principal axes techniques, some questions about the validity of the obtained patterns naturally arise: Are there any criteria to test the stability of a structure and validate it? What was the incidence of dealing only with a sample of individuals? And also a more difficult question arises: What are the consequences of the selection of variables?

We saw in Chapter 2 that in an attempt to partially answer these questions, we had recourse to empirical validation methods. They disrupt the original table by additions or withdrawals of array elements, individuals or variables (weight, coding, etc..). The assumption is that if the perturbations performed on the samples do not affect the observed patterns in the subspaces, the latter are assumed to be stable and the highlighted structure could be "significant."

Re-sampling methods propose to systematize this approach[207]. The nonparametric bootstrap is well adapted to the problem of the validity of the patterns observed on a principal plane: based on simulations, it allows for calculating areas of confidence for the locations of row points and column points.

- *Principle of the bootstrap*

The bootstrap technique, introduced by Efron (1979), consists in simulating s (s is generally greater than 30) samples of the same size n as the original sample. They are obtained by drawing randomly with replacement from the n individuals observed at the outset: the latter all have the same probability 1/n of being chosen. Some individuals appear more

---

[206] Muirhead (1982) has shown that the hypothesis of existence of the first four moments for the theoretical law of the sample was sufficient to validate these intervals.

[207] These methods are computationally intensive techniques, based on the technique of simulations of samples arising from one single sample. Made possible by the increase of computational power, these techniques are substituted in some cases for more traditional procedures that rely on restrictive assumptions. They are the only possible procedures when the analytical complexity of the problem does not allow for classical inference.

than once and have therefore a high weight (2, 3 ,...) while others are absent (zero weight).

This method is used to analyze the variability of simple statistical parameters by generating confidence intervals of these parameters. It can also be applied to many problems for which we cannot analytically estimate the variability of a parameter. This is the case for multi-dimensional methods when assumptions of multi-normality are rarely verified. Principal components analysis is an application domain that has given a great deal of work using the bootstrap re-sampling method.

Take the example of estimating the correlation coefficient r between two variables or between a variable and a factor (principal axes). The principle consists in calculating the correlation coefficient for each replicated sample (drawing with replacement of pairs of observations). The frequency distribution of the correlation coefficient is then obtained (histogram of the s values of r coefficients corresponding to s replications). Then one calculates from that histogram the required confidence intervals. The bounds of the confidence interval can be estimated directly by the quantiles of the simulated distribution.

One obtains an estimate of the accuracy of the value of r on the base sample without assuming a normal distribution of data.

To estimate the factorial coordinates from a principal components analysis, the principle is the same as for the correlation coefficient: we perform on each simulated sample a principal components analysis and then we draw a frequency distribution for each of the components. The bootstrap method gives in most cases a good picture of the statistical accuracy of the estimate on a sample. Theoretical research conducted by Efron, in particular, show that for many statistical parameters, the confidence interval corresponding to the simulated bootstrap distribution and that corresponding to the actual distribution are generally of the same amplitude.

- Implementation and computation of confidence zones

There are several procedures for assessing the stability of factorial coordinates through Bootstrap techniques. Gifi (1981) and Meulman (1982), Greenacre (1984) have conducted early work in the context of the analysis of single or multiple correspondences. In the case of principal components analysis, Diaconis and Efron (1983), Holmes (1989), Stauffer et al. (1985), Daudin et al. (1988) have dealt with the problem of choosing the appropriate number of axes.

To take into account the replicates, we must refer to a common factorial space. Several variants are possible.

We have based Chapter 2 on two techniques called the *total bootstrap* and the *partial bootstrap*.

The total bootstrap consists in carrying out as many principal components analysis as there are replications, with a series of transformations to find homologous axes during successive diagonalization of the s correlation matrices of replicate C_k (C_k corresponds to the k- th replication). These transformations are: changes of sign of the axes, rotations or permutations of the axes. This method was proposed by Milan and Whittaker (1995).

In the partial bootstrap method, proposed by Greenacre (1984) in the case of correspondence analysis, it is not necessary to calculate the eigenvalues and eigenvectors for all the replications: the principal axes, calculated on the undisturbed original data, play a special role (the initial correlation matrix **C** is indeed the expectation of the disturbed matrices $\mathbf{C}_k$).

The partial bootstrap method is based on the projection as supplementary points of the replicated points on the subspaces of reference provided by the principal axes of the initial correlation matrix **C** = **X'X**.

In the following transition formula, we replace matrix **X** with the replicated matrix $\mathbf{X}_k$ to obtain the corresponding replicate $\mathbf{u}_q(k)$.

$$\mathbf{u}_q = \frac{1}{\sqrt{\lambda_q}} \mathbf{X}'\mathbf{v}_q$$

where $\mathbf{u}_q$, $\mathbf{v}_q$ are respectively the q-th eigenvectors of **X'X** et **XX'** and λ_q the associated eigenvalue.

More precisely, the projection[208] of the k-th replication of the m variables (words) is given by the vector $\mathbf{u}_q$ (k) of $\mathbb{R}$ such that :

[208]. The projection of bootstrap replications in the context of principal components analysis uses the fact that the coordinate of a variable on a factorial axis is none other than its correlation coefficient with the variable coordinates of the individuals on the axis. We therefore calculate the replications of this coefficient, which amounts to re-weighting for each replication, the individuals with bootstrap weights that characterize a selection without replacement. We obtained as a byproduct, replications of the variance on the axis, which are obviously different from what would be replications of eigenvalues.

$$\mathbf{u}_q(k) = \frac{1}{\sqrt{\lambda_q}} \mathbf{X}'\mathbf{D}_k \mathbf{v}_q$$

and $\mathbf{D}_k$ denotes the diagonal matrix (n, n) of the bootstrap weights associated to the k-th replication.

In the case of partial bootstraps, the analyses of the matrices $\mathbf{C}_k$ are by no means necessary, since the eigenvectors are obtained from the principal components analysis of matrix **C**.

Bootstrap variability is thus observed better on the original permanent referential, which is also the only one that has not been disturbed. This technique, tested empirically, largely meets the users' concerns in the case of principal components analysis.

-Bootstrap on variables

Replications are classically obtained by drawing with replacement the n individuals. To test structural stability vis-à-vis the set of words, we propose to replicate this set using the total bootstrap method.

We thus implicitly assume that all the words of the questionnaire are a sample of m words randomly extracted from all of the "semiometrisables" words of the considered language.

This “sample of words” will undergo the same perturbations as the sample of individuals in the case of usual bootstrap

APPENDIX 2

Multilingual lists of words

FRENCH	ENGLISH	GERMAN	SPANISH	ITALIAN
l'absolu	absolute	absolut	el absoluto	l'assoluto
l'acharnement	persistence	hartnaeckig	el empeño	l'accanimento
acheter	to buy	kaufen	comprar	comprare
admirer	to admire	bewundern	admirar	ammirare
adorer	to love	anbeten	adorar	adorare
l'ambition	ambition	der Ehrgeiz	la ambición	l'ambizione
l'âme	soul	die Seele	el alma	l'anima
l'amitié	friendship	die Freundschaft	la amistad	l'amicizia
l'angoisse	anguish	die Angst	la angustia	l'angoscia
un animal	animal	ein Tier	un animal	un animale
un arbre	tree	ein Baum	un árbol	un albero
l'argent	silver	das Geld	el dinero	il denaro
une armure	armour	die Ruestung	una armadura	un'armatura
l'art	art	die Kunst	el arte	l'arte
astucieux	cunning	schlau	sagaz	astuto
l'attachement	attachment	die Zuneigung	el cariño	l'attaccamento
attaquer	to attack	angreifen	atacar	aggredire
l'audace	audacity	die Kuehnheit	la audacia	l'audacia
aventurier	adventurer	ein Abenteuer	un aventurero	avventuriero
un bâtisseur	a builder	ein Erbauer	un constructor	un costruttore
un bijou	a jewel	ein Schmuckstueck	una joya	un gioiello
bleu	blue	blau	azul	blu
bohème	wander	ein Lebenskuenstler	bohemio	anticonformista
un cadeau	a present	ein Geschenk	un regalo	un regalo
la campagne	countryside	auf dem Land	el campo	la campagna
une caresse	a caress	eine Liebkosung	una caricia	una carezza
une cérémonie	ceremony	eine Zeremonie	una ceremonia	una cerimonia
une certitude	certainty	eine Gewissheit	une certeza	una certezza
le changement	change	die Veraenderung	el cambio	il cambiamento
charitable	charitable	barmherzig	caritativo	caritatevole
charnel	carnal	das sexuelle	carnal	carnale
la chasse	the hunt	die Jagd	la caza	la caccia
un chercheur	researcher	ein Forscher	un investigador	un ricercatore
commander	to order	befehlen	mandar	comandare
le commerce	trade	der Handel	el comercio	il commercio
concret	solid	konkret	concreto	concreto
la confiance	confidence	das Vertrauen	la confianza	la fiducia
le confort	comfort	die Bequemlichkeit	el confort	la comodita'
conquérir	to conquer	erobern	conquistar	conquistare
consoler	to console	trosten	consolar	consolare

FRENCH	ENGLISH	GERMAN	SPANISH	ITALIAN
construire	to build	bauen	construir	costruire
le courage	brave	der Mut	el coraje	il coraggio
un créateur	creator	ein Schoepfer	un creador	un creatore
un cri	a shout	ein Schrei	un grito	un grido
critiquer	to criticise	kritiseren	criticar	criticare
le danger	danger	die Gefahr	el peligro	il pericolo
un défi	a challenge	eine Herausforderung	un desafio	una sfida
le desert	desert	die Wueste	el desierto	il deserto
le désir	desire	das Verlangen	el deseo	il desiderio
le désordre	disorder	die Unordnung	el desorden	il disordine
le détachement	detachment	die Gleichgueltigkeit	el despego	il distacco
Dieu	God	Gott	Díos	Dio
différent	different	verschieden	diferente	diverso
la discipline	discipline	die Disziplin	la discipina	la disciplina
la douceur	softness	die Sanfheit	la suavidad	la dolcezza
le doute	doubt	der Zweifel	la duda	il dubbio
dynamique	dynamic	dynamisch	dinámico	dinamico
l'eau	water	das Wasser	el agua	l'acqua
l'école	school	die Schule	la escuela	la scuola
économiser	economise	sparen	ahorrar	risparmiare
écrire	to write	schreiben	escribir	scrivere
efficace	effective	tuechtig	eficaz	efficace
un effort	effort	eine Anstrengung	un esfuerzo	uno sforzo
l'élégance	elegance	die Eleganz	la elegancia	l'eleganza
l'élite	elite	die Elite	la élite	l'elite
une émotion	emotion	eine Gefuehlsbewegung	una emoción	un'emozione
l'enfance	childhood	die Kindheit	la infancia	l'infanzia
enseigner	to teach	unterrichten	educar	insegnare
ensemble	together	miteinander	juntos	insieme
escalader	to climb	hochklettern	escalar	arrampicarsi
éternel	eternal	ewig	eterno	eterno
un étranger	a stranger	ein Fremder	un extranjero	uno straniero
une évasion	escape	eine Flucht	una evasión	un'evasione
la famille	family	die Familie	la familia	la famiglia
une faute	a fault	ein Fehler	una falta	una colpa
féminin	feminine	weiblich	femenino	femminile
la fermeté	firmness	die Standhaftigkeit	la firmeza	la fermezza
le feu	fire	das Feuer	el fuego	il fuoco
la fidélité	faith	die Treue	la fidelidad	la fedelta'
une fleur	flower	eine Blume	una flor	un fiore
un fleuve	river	ein Fluss	un río	un fiume
la foi	a belief	der Glaube	la fé	la fede
une frontière	border	eine Grenze	una frontera	una frontiera
un fusil	gun	ein Gewehr	un fusil	un fucile
la gaieté	liveliness	die Froehlichkeit	la alegría	l'allegria
la gloire	glory	der ruhm	la gloria	la gloria
gratuit	free	unentgeltlich	gratuito	gratuito
guérir	to recover	heilen	curar	guarire
la guerre	war	der Krieg	la guerra	la guerra
hériter	to inherit	erben	heredar	ereditare
un héros	hero	ein Held	un héroe	un eroe

FRENCH	ENGLISH	GERMAN	SPANISH	ITALIAN
honnêté	honest	ehrlich	honesto	onesto
l'honneur	honour	die Ehre	el honor	l'onore
humble	humble	demuetig	humilde	umile
l'humour	humour	der Humor	el humor	l'umorismo
une île	an island	eine Insel	una isla	un'isola
immense	huge	unermesslich	immenso	immenso
immobile	immobile	unbeweglich	immovil	immobile
un inconnu	an unknown	ein Unbekannter	un desconocido	uno sconosciuto
l'industrie	industry	die Industrie	la industria	l'industria
l'infini	infinity	die Endlosigkeit	el infinito	l'infinito
interdire	to forbid	verbieten	prohibir	vietare
interroger	to interrogate	befragen	interrogar	interrogare
intime	intimate	intim	intimo	intimo
un inventeur	inventor	ein Erfinder	un inventor	un inventore
l'ironie	irony	die Ironie	la ironía	l'ironia
un jeu	a game	ein Spiel	un juego	un gioco
la justice	justice	die Justiz	la justicia	la giustizia
un labyrinthe	maze	ein Labyrinth	un laberinto	un labirinto
la légèreté	lightness	die Leichtigkeit	la ligereza	la leggerezza
un livre	book	ein Buch	un libro	un libro
la logique	logic	die Logik	la lógica	la logica
la loi	the law	das Gesetz	la ley	la legge
la lune	moon	der Mond	la luna	la luna
la magie	magic	die Magie	la magia	la magia
une maison	house	ein Haus	una casa	una casa
maîtriser	to master	beherrschen	dominar	dominarer
un mariage	a wedding	eine Heirat	una boda	un matrimonio
un masque	mask	eine Maske	una máscara	una maschera
maternel	maternal	muetterlich	maternal	materno
la méfiance	mistrust	das Misstrauen	la desconfianza	la diffidenza
métallique	metallic	metallisch	metálico	metallico
la minceur	thinness	die Schlankeit	la esbeltez	la snellezza
la mode	fashion	die Mode	la moda	la moda
la modération	moderation	die Maessigung	la moderación	la moderazione
la modestie	modesty	die Bescheidenheit	la modestia	la modestia
moelleux	soft	anschmiegsam	blando	morbido
une montagne	mountain	ein Berg	una montaña	una montagna
la morale	morals	die Moral	la moral	la morale
la mort	death	der Tod	la muerte	la morte
une muraille	wall	eine Mauer	una muralla	una muraglia
la musique	music	die Musik	la música	la musica
un mystère	mystery	ein Mysterium	un misterio	un mistero
nager	to swim	schwimmen	nadar	nuotare
la naissance	birth	die Geburt	el nacimiento	la nascita
un nid	nest	ein Nest	un nido	un nido
noble	noble	edel	noble	nobile
un nœud	a knot	ein Knoten	un nudo	un nodo
noir	black	schwarz	negro	nero
la nudité	nudity	die Nacktheit	la desnudez	la nudita'
obéir	to obey	gehorchen	obedecer	obbedire
l'océan	ocean	der Ozean	el océano	l'oceano

FRENCH	ENGLISH	GERMAN	SPANISH	ITALIAN
l'or	gold	das Gold	el oro	l'oro
un orage	a storm	ein Gewitter	una tormenta	un temporale
original	original	eigentuemlich	original	originale
la paix	peace	der Friede	la paz	la pace
le pardon	forgiveness	die Verzeihung	el perdón	il perdono
un parfum	perfume	ein Parfuem	un perfume	un profumo
la patience	patience	die Geduld	la paciencia	la pazienza
la patrie	homeland	das Vaterland	la patria	la patria
la peau	skin	die Haut	la piel	la pelle
la perfection	perfection	die Vollkommenheit	la perfección	la perfezione
la poésie	poetry	die Poesie	la poesía	la poesia
la politesse	politeness	die Hoeflichkeit	la cortesía	la cortesia
précieux	precious	wertvoll	valioso	prezioso
la précision	accuracy	die Praezision	la precisión	la precisione
un prêtre	priest	ein Priester	un sacerdote	un prete
produire	to produce	produzieren	producir	produrre
la propriété	property	der Besitz	la propriedad	la proprieta'
protéger	to protect	beschuetzen	proteger	proteggere
la prudence	prudence	die Vorsicht	la prudencia	la prudenza
la puissance	the power	die Macht	la potencia	la potenza
punir	to punish	strafen	castigar	punire
une question	a question	eine Frage	una pregunta	una domanda
raffiné	refined	verfeinert	refinado	raffinato
la raison	the reason	die Vernunft	la razón	la ragione
une récompense	a reward	eine Belohnung	una recompensa	una ricompensa
le recueillement	meditation	die Besinnung	el recogimiento	il raccoglimento
réfléchir	to think	nachdenken	reflexionar	riflettere
une règle	a rule	eine Regel	una regla	una regola
le respect	respect	der Respekt	el respeto	il rispetto
rêver	to dream	trauemen	soñar	sognare
la révolte	rebellion	der Aufstand	la rebelión	la rivolta
la richesse	wealth	der Reichtum	la riqueza	la ricchezza
rigide	rigid	starr	rígido	rigido
rire	to laugh	lachen	reír	ridere
robuste	robust	robust	robusto	robusto
rompre	to break	abbrechen	romper	rompere
rouge	red	rot	rojo	rosso
la ruse	craftiness	die List	la astucia	la furbizia
sacré	sacred	heilig	sagrado	sacro
un sacrifice	sacrifice	ein Opfer	un sacrificio	un sacrificio
sauvage	wild	wild	salvaje	selvaggio
la science	science	die Wissenschaft	la ciencia	la scienza
un secret	a secret	ein Geheimnis	un secreto	un segreto
séduire	to seduce	verfurhren	seducir	sedurre
sensuel	sensual	sinnlich	sensual	sensuale
soigner	to care	pflegen	cuidar	curare
un soldat	soldier	ein Soldat	un soldado	un soldato
un sommet	peak	ein Gipfel	una cumbre	una vettat
la souplesse	flexibility	die Geschmeidigkeit	la flexibilidad	la flessibilita
souverain	monarch	souveraen	soberano	sovrano
sublime	sublime	herrlich	sublime	sublime

FRENCH	ENGLISH	GERMAN	SPANISH	ITALIAN
la tendresse	tenderness	die Zaertlichkeit	la ternura	la tenerezza
le théâtre	theatre	das Theater	el teatro	il teatro
la tradition	tradition	die Tradition	la tradición	la tradizione
trahir	to betray	verraten	traicionar	tradire
le travail	work	die Arbeit	el trabajo	il lavoro
utilitaire	practical	praktisch	utilitario	utilitario
vert	green	gruen	verde	verde
la victoire	victory	der Sieg	la victoria	la vittoria
le vide	space	die Leere	el vacío	il vuoto
vieillir	to age	alt werden	envejecer	invecchiare
viril	virile	maennlich	viril	virile
la vitesse	speed	die Geschwindigkeit	la velocidad	la velocita'
volontaire	a volunteer	eigenwillig	voluntario	volontario
Voluptueux	voluptuous	lustvoll	voluptuoso	voluttuoso

References

Anatrella T. (1995*), Non à la société dépressive*, Champs, Flammarion, Paris.

Anderson T. W., Rubin H. (1956), Statistical inference in factor analysis, *Proc. of the 3rd Berkeley Symp. on Math. Statist.*, 5, p 111-150.

Anderson T. W. (1963), Asymptotic theory for principal components analysis, *Ann. Math. Statist.,* 34, p 22-148.

Baayen R. H. (2001), *Word Frequency Distributions*, Kluwer Academic Publishers, Dordrecht.

Barthes R. (1957), *Mythologies*, Editions du Seuil, Paris.

Benzécri J.-P. (1982), *Histoire et préhistoire de l'analyse des données*, Dunod, Paris.

Benzécri J.-P. (1989), Essai d'analyse des notes attribuées par un ensemble de sujets aux mots d'une liste, *Cahiers de l'Analyse des Données*, Vol XIV, 1, p 73-98.

Benzécri J-P. (1973), *L'Analyse des Données,* Tome 1 : *La Taxinomie*, Tome 2 : *L'Analyse des Correspondances*, Dunod, Paris.

Benzécri J-P. (1992), *Correspondence Analysis Handbook*. Marcel Dekker New York.

Bréchon P. (2000), *Les valeurs des Français. Evolutions de 1980 à 2000*, Armand Colin, Paris.

Chateau F., Lebart L. (1996), Assessing sample variability in the visualization techniques related to principal components analysis: bootstrap and alternative simulation methods, in : *COMPSTAT96,* A. Prats (ed), Physica Verlag, Heidelberg, p 205-210.

Cottrell M., Rousset P. (1997), The Kohonen Algorithm: a powerful tool for analysing and representing multidimensional qualitative and quantitative data, in: *Biological and Artificial Computation* : From Neuroscience to Technology, J. Mira, R. Moreno-Diaz, J. Cabestany, (eds), Springer, p 861-871.

Daudin J.-J., Duby C., Trécourt P. (1988), Stability of principal components studied by the bootstrap method, *Statistics*, 19, p 241-258.

Diaconis P., Efron B. (1983), Computer intensive methods in statistics, *Scientific American*, 248, p 116-130.

Dupuis F., Lebart L. (2009), Visualization, validation and seriation. Application to a corpus of medieval texts. in *Historical Linguistics 2007*, John Benjamins Publishing Company, Philadelphia.

Efron B. (1979), Bootstraps methods : another look at the Jackknife, *Ann. Statist*., 7, p 1-26.

Efron B. (1982), The Jackknife, the Bootstrap and other resampling plans, *SIAM*, 1982, p 116-130.

Escoufier (1970), Echantillonnage dans une population de variables aléatoires réelles, *Publications de l'Institut de Statistique de l'Université de Paris*, Vol XIX, fasc. 4, p 1-47.

Fabre J., Morlat G., Pagès J-P, Stemmelen E. (1981), Les Structures de l'Opinion Publique, *Le Progrès Technique*, n° 22-24, ANRT, Paris.

Ferrand L., Alario F.-X. (1998), Normes d'associations verbales pour 366 noms d'objets concrets, *L'année psychologique*, 98, p 659-709.

Freud S. (1978), *Abrégé de psychanalyse,* PUF, Paris.

Freud S. (1985), *Trois Essais sur la théorie sexuelle*, Gallimard, Paris.

Garnett J.-C. (1919), General ability, cleverness and purpose, *British J. of Psych*., 9, p 345-366.

Gifi A. (1981), *Non Linear Multivariate Analysis*, Department of Data theory, University of Leiden, (updated version: 1990, same title, J. Wiley, Chichester).

Gower J.C. and Hand D.J. (1996), *Biplots*. Chapman and Hall, London.

Greenacre M. (1984), *Theory and Applications of Correspondence Analysis*, Academic Press, London.

Harman H.H. (1967), *Modern Factor Analysis,* Chicago University Press, Chicago.

Hayashi C. (1956), Theory and examples of quantification, (II), *Proc. of the Institute of Statistical Mathematics,* 4, (2), p 19-30.

Hochberg, Y. (1988), A sharper Bonferroni procedure for multiple tests of significance, *Biometrika*, 75, p 800-803

Holmes S. (1989), Using the bootstrap and the RV coefficient in the multivariate context, in : *Data Analysis, Learning Symbolic and Numeric Knowledge*, E. Diday (ed.), Nova Science, New York, p 119-132.

Hotelling H. (1933), Analysis of a complex of statistical variables into principal components, *J. Educ. Psy.* 24, p 417-441, p 498-520.

Hsu, J. C. (1996), *Multiple Comparisons: Theory and Methods*, Chapman & Hall, London.

Jacob F. (1970), *La logique du vivant*, Gallimard, Paris.

Kazmierczak J.-B. (1985), Analyse logarithmique : deux exemples d'application*, Revue de Statist. Appl.*, 33, (1), p 13-24.

Kleiweg P. (1996), *Een inleidende cursus met practica voor de studie* Alfa-Informatica, Master's thesis, Rijksuniversiteit Groningen, 1996.

Kohonen T. (1989), *Self-Organization and Associative Memory*, Springer-Verlag, Berlin.

Laplanche J., Pontalis J.-B. (1994), *Vocabulaire de la psychanalyse*, PUF, Paris.

Lawley D. N., Maxwell A. E. (1963), *Factor Analysis as a Statistical Method*, Methuen, London.

Lebart L. (1986), Qui pense quoi en France ? Evolution et structure des opinions en France de 1978 à 1984, *Consommation, Revue de Socio-Economie*, Dunod, p 3-22.

Lebart L. (2004), Validité des visualisations de données textuelles,. in: *Le poids des mots, Actes des JADT04: Septièmes Journées Internationales d'Analyse Statistique des Données Textuelles,* (G. Purnelle, C. Fairon, A. Dister, eds), Presse Universitaires de Louvain, p 708-715.

Lebart L. (2006), Assessing Self Organizing Maps via Contiguity Analysis. *Neural Networks*, 19, p 847-854.

Lebart L. (2007), Which bootstrap for principal axes methods ? In: *Selected Contributions in Data Analysis and Classification.* P. Brito et al. (eds), Springer, (2007), p 581-588.

Lebart L., Houzel Y. (1980), Le système d'enquête sur les conditions de vie et aspirations des Français, *Consommation, Revue de Socio-Economie*, Dunod, p 3-25.

Lebart L., Morineau A., Tabard N. (1977), *Techniques de la description statistique.* Dunod, Paris.

Lebart L., Morineau A., Warwick K. (1984), *Multivariate Descriptive Statistical Analysis*, Wiley, New York.

Lebart L., Piron M., Steiner J.-F. (2003), *La Sémiométrie*, Dunod, Paris. . [in French. Downloadable from www.dtmvic.com].

Lebart L., Piron M., Morineau M., (2006), *Statistique exploratoire multidimensionnelle*, Dunod, Paris.

Lebart L., Salem A. (1994), *Statistique textuelle*. Dunod, Paris.

Lebart L., Salem A., Berry L. (1998), *Exploring Textual Data*, Kluwer Academic Publishers, Dordrecht.

Lewi P. J. (1976), Spectral mapping, a technique for classifying biological activity profiles of chemical compounds. *Arzneim. Forsch*. in: *Drug Res.* 26, p 1295-1300.

Meulman J. (1982), *Homogeneity Analysis of Incomplete Data*, DSWO Press, Leiden.

Milan L., Whittaker J. (1995), Application of the parametric bootstrap to models that incorporate a singular value decomposition, *Appl. Statist*. 44, 1, p 31-49.

Muirhead R. J. (1982), *Aspects of Multivariate Statistical Theory*, J. Wiley, New York.

Mulaik S. A. (1972), *The Foundation of Factor Analysis*, McGraw Hill, New York.

Muller C. (1977), *Principes et méthodes de statistique lexicale*, Hachette, Paris.

Muller C. (1979), Peut-on estimer l'étendue d'un lexique ? In : *Langue Française et Linguistique Quantitative*, p 399-425, Slatkine, Genève.

O'Neill, R., and G. B. Wetherill. (1971), The present state of multiple comparison methods (with discussion), *Journal of the Royal Statistical Society*, Series B, 33, p 218-250.

Osgood C. E. (1965), Cross-Cultural Comparability in Attitude Measurement via Multilingual Semantic Differentials, in: Steiner I., Fishbein M. (Eds): *Current Studies in Social Psychology*, Holt, Rinehart and Winston, New York.

Osgood C. E., Suci G. J., Tannenbaum P. H. (1957), *The Measurement of Meaning*, University of Illinois, Urbana.

Perneger T.,V. (1998), What is wrong with Bonferroni adjustments, *British Medical Journal*, 136, p 1236-1238

Riffault H. (1994), *Les valeurs des Français*, PUF, Paris.

Saville, D. J. (1990), Multiple comparison procedures: The practical solution, *American Statistician*, 44, p 174-180.

Sokal R. R., Sneath P. H. A. (1963), *Principles of Numerical Taxonomy*, Freeman and co., San-Francisco.

Spearman C. (1904), General intelligence, objectively determined and measured, *Amer. Journal of Psychology*, 15, p 201-293.

Stauffer D. F., Garton E. O., Steinhorst R. K. (1985), A comparison of principal components from real and random data, *Ecology*, 66, p 1693-1698.

Steiner J.-F., Auliard, O. (1992), La sémiometrie: un outil de validation des réponses, In : *La Qualité de l'Information dans les Enquêtes* / Quality of Information in Sample Surveys, ASU, (L. Lebart ed.), Dunod, Paris, p 241-274.

Stoetzel J. (1983), *Les valeurs du temps présent : une enquête européenne*, PUF, Paris.

Tabard N. (1974), *Besoin et Aspirations des familles et des jeunes,* Coll. Etudes CAF, 16, CNAF, Paris.

Thiria S., Lechevallier Y., Gascuel O., Canu S. (1997), *Statistique et méthodes neuronales*, Dunod, Paris.

Thurstone L. L. (1947), *Multiple Factor Analysis*, The University of Chicago Press, Chicago.

Valette-Florence P. (1994*), Les styles de vie. Bilan critique et perspectives*, Nathan, Paris.

Westfall, P. H., Young S. S. (1993), *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, J. Wiley, New York.

Westfall P. H., Tobias R. D., Rom D., Wolfinger R. D., Hochberg Y. (1999), *Multiple Comparisons and Multiple Tests Using the SAS System*, SAS Institute.

Young G. A. (1994), Bootstrap: more than a stab in the dark, *Statistical Science*, 9, p 382-418.