Analyzing the language of everyday life: how Textual Statistics can support Time Use Surveys. The Italian Experience

Il linguaggio della vita quotidiana: la Statistica Testuale a supporto delle Indagini sull'Uso del Tempo. L'esperienza italiana(⋆)

Alessio Canzonetti[1], Michelangelo Misuraca[2], Maria Clelia Romano[3]

[1] Dip. di Studi Geoec., Ling., Stat., Stor. per l'A.R., Università "La Sapienza" di Roma
e-mail: alessio.canzonetti@uniroma1.it

[2] Dip. di Matematica e Statistica, Università "Federico II" di Napoli
e-mail: michelangelo.misuraca@unina.it

[3] Istituto Nazionale di Statistica (ISTAT)
e-mail: romano@istat.it

Riassunto: La Statistica ha tra i suoi compiti quello di assicurare ad una comunità un'informazione dettagliata e completa su tutti quei fenomeni ritenuti di pubblico interesse. Tradizionalmente nelle indagini statistiche svolte a tal scopo la conoscenza derivante dallo studio di dati non strutturati come quelli *testuali* è stata limitata dalla complessità e dall'onerosità del loro trattamento, e relegata funzionalmente al solo scopo di "aggiungere" informazione documentale all'oggetto d'analisi. Obiettivo di questo lavoro è mostrare come le tecniche proprie della Statistica Testuale siano oggi mature per offrire un supporto sostanziale allo studio di fenomeni complessi e contribuire alla creazione di *fonti statistiche*. Tale possibilità è descritta, in particolare, nel quadro dell'indagine sull'*Uso del Tempo* realizzata dall'ISTAT nel 2002-2003.

Keywords: textual statistics, time use survey, multiple factorial analysis, ontologies

1. Introduction

The use of textual dataset as an investigation tool always belongs to the qualitative research traditions, for instance in the case of in-depth interviews, but the textual information has been also used in sample surveys, as a mean for obtaining knowledge in the form of free answers to *open-ended question*, with different purposes.

First of all resorting to those questions has allowed, in the frame of survey design and particularly in the pre-test step for the collection of information dealing with the analyzed phenomena, when it was not available *a priori* knowledge or when it was not possible to rely on previous surveys regarding the same research field. Moreover their use has ever been very suitable in treating themes of particular sensitiveness, in which it is important to "not force" and/or influence the respondent subjects in the answering process.

The open-ended questions leave indeed the individuals involved in the survey to express themselves in the way they prefer, limiting the risk of non-sample errors due

(⋆) This work has been carried out by the authors in the frame of the ISTAT - University of Rome "La Sapienza" research protocol *Application of Text Mining techniques to textual information in the daily diaries*.

to the intermediation role played by interviewers, or to a bad definition of the answering modalities in structured (closed-ended) questions. On the other hand, their use has always been limited because of the unavailability of reliable encoding systems that allowed a good level of synthesis, therefore a profitable use with respect to the goals of the investigation.

The development of tools for analyzing, from a quantitative point of view, text written in natural language, and therefore the automatic treatment of the open-ended questions (e.g. Lebart, 1982), it has opened undoubtedly a new perspective in social researches.

The possibility of integrating the classical data obtained from structured questions with textual information, originally unstructured, has been more and more deepened by using multivariate descriptive techniques, in the furrow of the French *Analyse de Données* point of view (but actually also developed in the other several schools that have embraced J.P. Benzécri original idea). We can not only consider the use of textual data as *external information* or in general as *meta-information* in the analysis and the interpretation of the socio-economical phenomena, but nowadays it is proper to speak of a "Textual Statistics", or more specifically of a "Textual Data Analysis", as an indipendent and original research sphere (cfr. Benzécri, 1981; Lebart and Salem, 1988; Bolasco, 1992).

In this viewpoint any collection of text (newspaper years, technical handbooks or patents, literary works, political speeches, and so on) can be analyzed, in a statistical way, with the aim of studying the latent structures and relations "hidden" among the words but more ambitiously for discovering and extracting Knwoledge from the textual datasets.

This opportunity it is not sufficiently explored in the Official Statistics, where the main aim is to collect and represent in a usable and easy-readable way the information of public interest on collectives phenomena. The use of such data has still viewed as too problematic, because of the complexity and the expensiveness of the pre-treatment processing and often for the lack of suitable tools.

For instance, it is necessary to remark that in the open-ended question analysis the "importance" of the different textual forms is sometime rather ambiguous (Lebart, 2004).

The observed lexical frequencies are fairly artificial, since the same question is sometimes administered to hundreds or thousands of individuals, and the juxtaposition of the answers constitutes, by construction, a redundant text. We have in fact two different kind of statistical units: the interviewed subjects (the units commonly analyzed by the statisticians in sample surveys) and the used textual forms (the units typically considered in a textual data analysis framework). Some statistical tests will be classically meaningful for the forms but not for the individuals, so that inferring the results to the reference population is not so easily performed.

The effort of the Italian National Statistical Institute (ISTAT) in collecting and recording a huge amount of textual data, for the first time in an official survey, represents a challenge that would be unthinkable only few years ago. It shows that the Textual Statistics techniques can contribute in describing collective behaviours and attitudes in a substantive way, by helping the traditional detecting tools in focusing on the "obscure matter" of the social universe.

Aim of this paper is to deep the possible interactions between the Official Statistics problems in describing collective phenomena and the solutions offered by Textual Data Analysis in discovering, extracting and visualizing Knowledge from language. Particularly it will show the effectiveness of Text Mining strategies in Time Use studies for describing and profiling everyday life styles in the different social strata of a population.

2. Social and Individual Time: the Time Use Surveys

The notion of *time perception* commonly used is so abstract and general that different human societies, with a wide range of cultural and economical backgrounds, would not know how to use it in a same way. However the abstraction is not necessarily referred to the concept of time, but properly to the ways of organize the job and the daily life or to the prevailing knowledge forms of the culture in which the definitions are formulated.

The time, in a "social" meaning, gives an objective measurement of the amount of hours and minutes spent by an individual in the single activities as a member of a community, but at the same time is rich in shades and hints. As "individual" time it quickly flows or slows down, with respect to the subjective perceptions of the activity lengths, or to how much is to choose an activity rather than another.

The social sciences have initially started their interest in studying the time distribution among the human activities within a day, a week or a year, dealing with the most complex theme of the analysis of working class conditions. Nevertheless only a part of the literature can be considered in the circle of the *time budgets* studies. The first example of investigation were the researches of G.E. Bevans (1912) and P.A. Sorokin and C.O. Berger (1935). The last one can be considered as a model for the following development of the analysis on *Time Use* phenomenon.

The time budgets definition was born in analogy with the so called *family budgets*. Both have to be referred as a "funds destination": as the family budgets report the destination of the household income in different expense items, the time budgets report the allocation in 24 hours (or in another interval of meaningful social time) of the activities as feeding, working, taking care of children or relatives, reading, watching tv, and so on. In the f.b. the monetary aspects of people life are showed up, while in the t.b. we consider all that social, cultural and custom aspects not valuable in monetary terms.

The tool which through it is possible to record the complete sequence of individual activities, the places in which it habitually develops it and with whom, is the *diary*. This peculiar kind of structured questionnaire is more suitable than others for recording the frequent events as the daily activities, allow the subjective description of activities and so to gather a wide variety of perceptions on the common life expectations and rythms.

2.1. The 2002-2003 ISTAT Time Use Survey

The 2002-2003 Multi-purpose Time Use Survey (TUS) on households was carried out by the ISTAT Institute between April 2002 and March 2003 (excluding the 9 midweek "bank holidays"). A sample of 21,075 households has been interviewed, summing up about 55,000 individuals. The surveying tools consisted in an individual questionnaire, a household questionnaire, a weekly diary for 15 years old people and older (aimed to determine the number of paid working hours in a given week), and a daily diary for 3 years old people and older, to examine in details the time use in a specific day randomly assigned to each respondent.

The survey on Time Use has been projected dealing with the EUROSTAT "Hetus" project (*Harmonized European Time Use Surveys*). Differently from the EUROSTAT recommendations on the interviewed age threshold, it was established to include 3-9 years old children for the good results obtained in the previous eighties' experience for this population segment.

The most important surveying tool is the daily diary, a substantive information source with a high level of detail on the individual time organization. By scheduling the diary it is possible to know in which way each respondent shares the time in 24 hours (144 ten minutes slots) in terms of activities, transfers, visited places and person with whom the activities are carried out. This information is unique, because it is not possible to discover these aspects by means of traditional structures questions.

The *main* and *parallel* activities were described as free text (very often in the form of sentences) 10 minutes by 10 minutes, as well as the locations in which the activities took place and the means of transport used by the respondents to move from a place to another one. The respondents were asked to indicate with whom time was spent, for each interval of 10 minutes, ticking the respective box.

The great effort made by households in meeting the survey requirements comes through very clearly by reading the diaries, which in many cases are real life stories. It is necessary to remark that some diaries have been completed in a sketchy fashion, they are not so clear and detailed, and sometimes clearly show respondents unwillingness. Nevertheless in many diaries the respondents have carefully described how they spend their time and how the various activities carry on along the day with extreme detail.

For the first time it has been decided to record (and to reverse into electronic format) not only the activity codes but also the textual information scheduled in the diaries (Romano *et al.*, 2004). According to EUROSTAT guidelines and to what has been carried out in most European countries, the activities surveyed are classified according to a common analytical encoding scheme, revised in order to fit it to the ISTAT requirements and highlight peculiar national aspects. Nevertheless, indipendently from the encoding scheme adopted by the different countries (manual or computer-assisted), nobody else (but Italy and France) has fully recorded the textual descriptions contained in the diaries.

In order to prevent non-sample errors in the encoding process an assisted procedure based on *tri-grams* recognition has been carried out (Romano and Cappadozzi, 2004).

3. Preparing daily diaries for statistical analysis

A *corpus* can be analyzed using methodologies belonging to qualitative or quantitative research. In the first case the attention is focused on the different contexts underlying the text, aiming at classifying the most interesting with respect to some criteria. In the second case the main purpose is to decompose the text into basic units, perform some analysis for discovering the most frequent content bearing terms and graphically representing the latent association structures (Balbi and Misuraca, 2005).

The choice of textual elementary units is still an open problem in Textual Data Analysis and generally in the automatic text processing based on a statistical viewpoint. The easier solution that directly comes up is to select words as basic units. Because of the peculiar nature of the analyzed phenomenon, it is possible to consider several strategies according to different information needs of the users (Misuraca, 2004).

It is important to consider the necessary pre-processing that leads to a set of structured data starting from the textual information contained in a *corpus*. By means of a parsing procedure the text is “read” as a sequence of characters (*form*) separated by a blank or in general by a *delimiter* character, yelding as a main result the list of all the recognized terms (namely, the *vocabulary*) with their occurrences. The different steps (normalization,

segmentation and lexicalization, implicit lemmatization or stemming) are justified with respect to the research goals, depending on the fact that a *paradigmatic* or a *syntagmatic* point of view has been focused.

The paradigmatic approach points out the lexical aspect of the text, dwelling upon the vocabulary expressed both in terms of lexical or lemmatized units (*vertical researches*). The last ones in particular are obtained by considering the dictionary entry words and can be very useful in avoiding some ambiguos cases (e.g. omographs terms belonging to different headwords) and "simplifying" the language variability.

The syntagmatic approach points out the semantic/conceptual aspects of the text (*orizzontal researches on contexts*). Dividing a text in elementary units leads to a notable loss of information about meanings, sense, style, and all those phenomena that are generated by the combination of different signs. Following the proposal of Bolasco (1992) sense significant units called "textual forms" are selected taking into account meta-information on the role played by words in the text and the contexts in which they are used. In this way it is possible to consider the textual data in the frame of a knowledge discovery process, allowing the definition of homogeneous statements classes or semantic classes.

The textual data analyzed in this paper belong to the daily diaries entries, i.e. the free description of each ten minutes slot scheduled by the respondents, concerning main activities, parallel activities, places and means of transport used by respondents (Table 1). The text has been edited by previously trained recorders, encoding the activities with the numerical codes scheme, and then subjected to a double correction process, *deterministic* and *non-automatic*, but only for verifying the correspondence between activity descriptions and numerical codes.

Table 1: *Strings collected from the daily diaries, by type of information.*

	Main activity	Parallel activity	Place or Mean of transport
Strings (sentences)	1,457,246	414,979	1,221,372
- different strings	*240,252*	*58,345*	*31,791*
Average length of strings	17.9	18.6	9.3
Occurrences (words)	4,973,359	1,477,047	2,887,369
- different words	*29,478*	*12,960*	*9,292*
Average frequency	168.7	113.9	310.7

Source: *ISTAT, 2002-2003 Time Use Survey*

In this work only the diaries of 24-65 years old people have been selected, to deeply study a sub-sample of quite homogeneous units, avoiding the trivial evidences due to peculiar age *strata* (e.g children or elderly people).

Moreover, it has decided to take into account only the subjects declaring that the described day was in their opinion an "ordinary day", in opposition to the peculiar days (the respondent was travelling, had a particular day at work, and so on). A sample of 24,103 individuals has been considered, in order to profile and describe the various kinds of days according to classes of individuals with similar characteristics and behaviours. For each individual a set of variables on socio-demographic aspects have been available, as shown in Table 2.

Table 2: *Examples of individuals structured information available in the survey.*

Household Characteristics	Family type Family size Family tie
Individual Characteristics	Sex Age Study degree Civil status
Working characteristics	Working condition Working position Activity field
Day info	Day surveyed Day weight

Textual diaries entries has been processed with the purpose of improving data quality, even if in this case the treatment has been not so "in depth" as in other standard analyzing strategies because of the peculiar surveyd phenomenon(1). First of all an orthographic correction process has been performed for limiting the noise caused by most frequent terms misspelling. It has remarked that in diary processing this is a large problem both due to respondent bad handwriting and to recorders typing errors.

The text is *normalized* in order to reduce the possibility of data splitting, for instance by converting all the capital letters to the lower case or conforming the different transliteration of a unique term, mainly proper nouns. By using a "list-based" normalization and a *lexicalization* procedure it has been possible to discover and tag some polywords and multiwords of interest, mainly referred to places and means of transport.

On the other hand a semantic tagging has been performed on multiwords in order to obtain a conceptual categorization of several daily aspects.

4. Visualizing and profiling everyday life styles in TUS

In order to obtain a rappresentation of the individuals on the basis of the activities carried out within the day, firstly it has tried to detect which main activities point out. This type of investigation has been implemented by comparing the vocabulary of the diaries with a *frequency lexicon*. The last one is a list of forms extracted from a representative *corpus* with the number of ocurrences, in our case the Standard Italian Language (Bolasco and Morrone, 1998). The comparison between these two lists allows to obtain the "use degree" of each form in common (*underusage* and *overusage*), as a deviation between the form frequencies in the analyzed text and in the frequency lexicon (Bolasco, 1999).

Therefore forms with the higher deviation values represent the so called *peculiar language*, that in our case is the vocabulary subset that includes the topics, the actors and the typical actions in the *corpus*.

Aiming at highlighting the daily activities, the peculiar vocabulary has been extracted. In particular it has performed a lemmatization on the forms (e.g. *vado* is changed in

(1) All the pre-processing steps, the lexicometric analysis and the information retrieval process has been performed with *TALTAC 2.0*, a software containing a wide library of tools for lexical and textual analysis

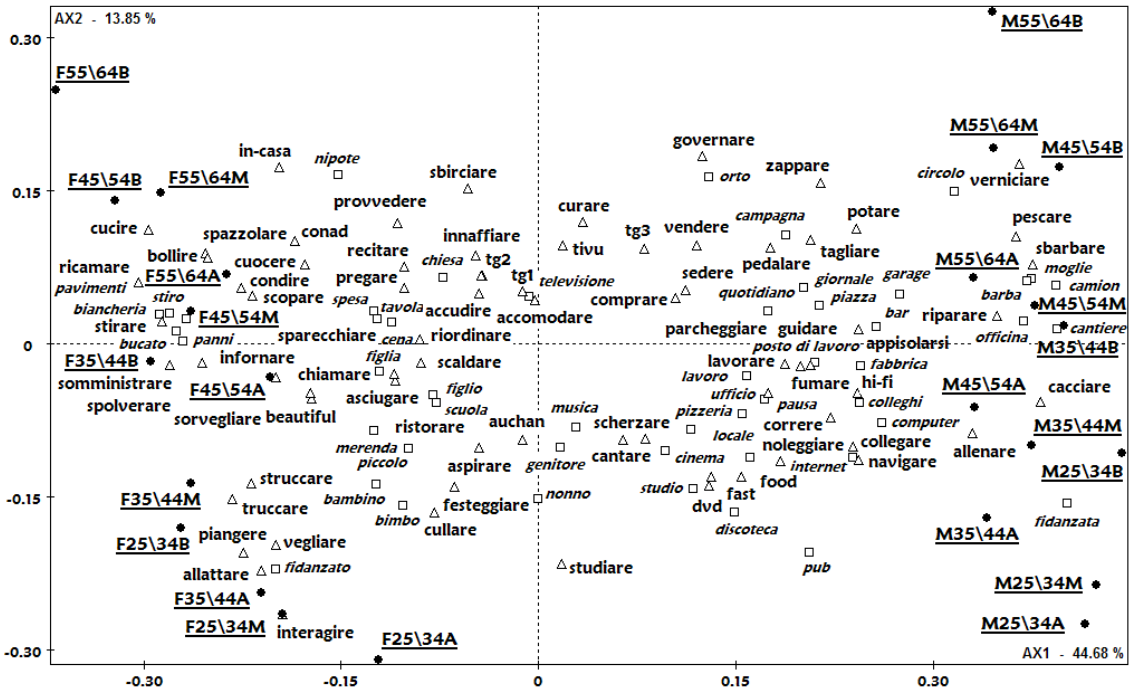
Table 3: *The 30 higher deviation forms.*

Form	Occurrence	Deviation	Form	Occurrence	Deviation
preparo	29,863	1422.4	ascolto	28,486	368.4
dormo	27,997	1333.5	pranzo	26,814	342.5
guardo	53,415	1078.7	guardato	18,062	337.8
casa	411,561	1070.9	svegliata	6,632	315.5
lavo	21,117	836.6	accendo	3,765	310.4
lavato	14,269	832.3	auto	29,616	300.2
mia	261,114	666.9	vado	44,492	295.1
vesto	6,200	626.4	piatti	12,652	281.4
lavata	7,084	584.3	macchina	35,420	278.0
cenato	6,612	545.4	dormendo	5,432	274.1
colazione	27,612	521.6	cucina	23,021	268.6
esco	17,514	510.3	radio	27,065	266.6
pranzato	6,970	497.8	cena	18,850	262.6
svegliato	8,258	481.5	doccia	6,966	256.5
parlo	40,500	416.7	ho	167,707	255.4

andare), by comparing the resultant vocabulary with a frequency lexicon of lemmatized forms. In this way a $\{forms \times text\}$ matrix has been obtained, cross-tabulating the lemmatized forms on the rows and the sub-occurences on the columns, i.e. the number of times each form has been used in each fragments group. In this case the respondents has been classified in 18 classes regarding sex, age and study degre.

By performing a Correspondence Analysis it is possible to explore the lexical profiles and extracting useful information, evaluating characteristics and differences. In particular nouns and adjectives have been considered as active points (squared labels), while verbs and adverbs has been projected as supplementary points (triangled labels), in order to underline mainly places and actors and undirectly the connected actions.

Figure 1: *First factorial map of peculiar vocabulary in TUS.*



The factorial map (Figure 1) shows on the first axis a clear sex-based "polarization". It belongs to women the objects and the actions properly referred to the house and family care (*biancheria, spesa, cucinare, stirare*, etc.). On the other hand it can be noticed for the men objects, places and actions tipically referred to job and leisure activities (*posto di lavoro, computer, allenare, pedalare*).

On the second axis it is possible to read, from the top to the bottom side, the different classes of age and the classes of study degree (A = high level, M = middle level, B = low level). It seems to be interesting that the higher study degree has a "rejuvenating" effect, because individuals with a high level are close to the youngest.

This kind of representation allows to summarize the textual information in a very readable way and supplies a suitable interpretation tool, on the basis of the varibles involved in the analysis, in the frame of a syntagmatic approach to text analysis.

However beside this kind of approach it is possible to go beyond the form literalness by detecting the concepts within the text, independently from the various ways in which they can be expressed. It needs to carried out full text queries, in a Text Mining framework, to find the presence or the absence of a specific activity and therefore create from the text a new variable, adding this type of information to the original data. Repeating this operation for all the interesting activities, the profiling of the individuals with respect to a conceptual rather than lexical logic has reached.

The choice of the activities to detect begins from the analysis of the peculiar verbs. The presence of forms within this subset of verbs. For instance the presence of verbs like *spazzare, riordinare, pulire, lavare,* has focused the attention on the "house care" concept. However it has been said that the words are often caracterized by a not negligible degree of semantic ambiguity: *pulire* can be referred to the house, but also to the own person or some other object. Therefore it is necessary to detect which are the several literal expression through which the "house care" can be expressed. The *concordance analysis* is the technique to use at this aim, by visualizing all the contexts, the phrases, in which a form appears. These lists of contexts allow to set the associations, the co-occurrences that can define a concept or a topic, e.g. with regard to the "house care" the verbs *pulire, lavare* with *pavimento, casa, soggiorno* and so on, or *sistemare*/*riordinare* with *cassetto, armadio, camera da letto*, etc.

These expressions tipically have a structure like "verb - object". In order to perform easier and faster queries, a semantic tagging procedure has been introduced. A list regarding the verbs and a list regarding the objects have been builded, then each form belonging to a list has been marked on the *corpus* with a semantic category, establishing an equivalence relationship between verbs and objects. Thus the query can be expressed in this way:

```
CATSEM(predicati di cura della casa) LAG5 CATSEM(oggetti di cura della casa)
```

The first and the second CATSEM contain the tags of the created semantic categories, while the LAG operator represents the maximum delay in terms of forms (five in the example), within which the occurrence of the second category form have to appear in the text. This kind of query detect and retrieve all the individuals that have declared to carry out an activity like *lavato i pavimenti, riordinato assieme a mia figlia i cassetti, pulendo un poco la casa*, and it marks them by mean of a variable counting people who carried out "house care" during the day.

Obviously a control process in the development of these queries has been carried out to verify the goodness and the reliability of the tagging. Every search produces in fact the list of the multiword expressions properly found in the text and their occurrences. The screening of this list allows to proceed to suitable corrections or queries refining. Moreover the possibility of turning to these lists allows a detailed analysis of the main concepts shading.

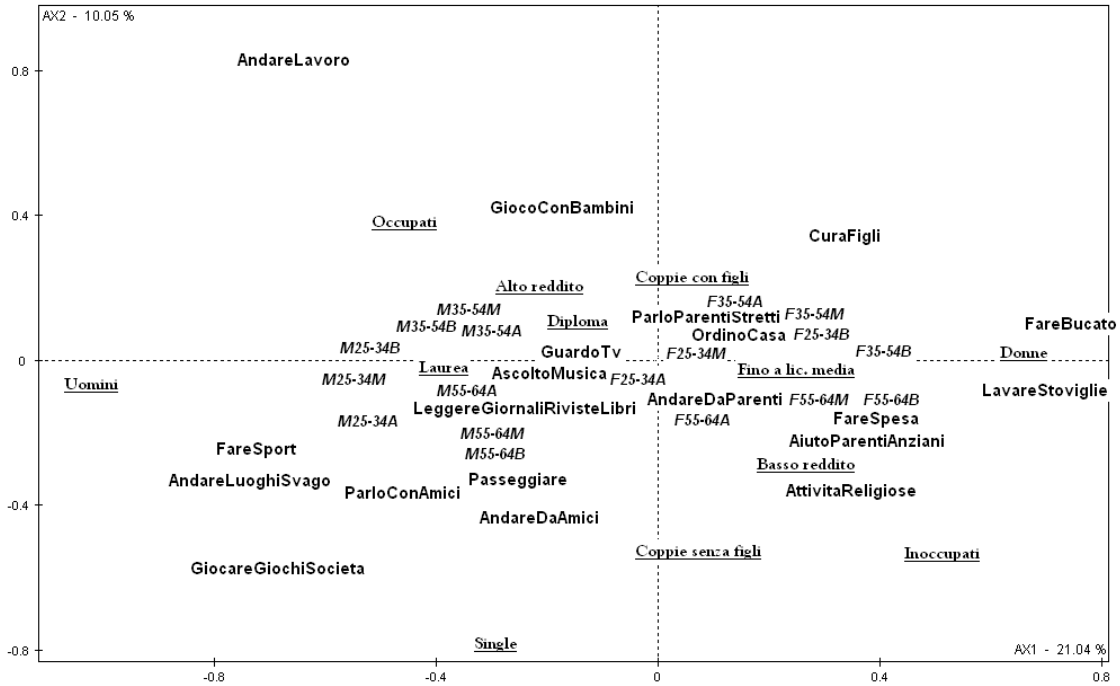
By using this procedure 20 variables have been created from the textual data regarding to several activities: *andare a lavoro*, *cura della casa*, *lavare le stoviglie*, *fare il bucato*, *fare la spesa*, *cura dei figli*, *aiutare parenti più anziani*, *giocare con i bambini*, *giocare giochi di società*, *leggere giornali/riviste/libri*, *guardare la TV*, *ascoltare musica*, *passeggiare*, *fare sport*, *attività religiose*, *andare in luoghi di svago*, *parlare con parenti*, *parlare con amici*, *andare da amici*, *andare da parenti*.

It has decided to not consider physiological activities (eating, sleeping, etc.), because of their role is not so discriminant in identifying significant differences in day profiles, and the activities characterizing specific classes of few individuals.

Now it is possible to "collapse" the fragments (daily diaries) matrix on the basis of whichever variable or combination of variables, obtaining therefore a table of frequencies having on the rows strata of individuals and on the columns the frequencies regarding the several activities for the cosidered nucleus. In our case we have obtained 3,073 different strata combining the modalities of the following variables: sex, age (recoded in 3 classes), study degree, family type, geographic area, family total income, working position.

It is possible to carry out a Correspondence Analysis on such matrix to point out not the similarity in terms lexical profiles, as showed before, but the proximity among the conceptual profiles, by gathering on the individuals showing likeness in the daily activities.

Figure 2: *First factorial map of daily activities.*



In order to make clear the results interpretation, some illustrative variable have been projected on the factorial plan (with undeline labels), beyond the points representing the 20 daily activities (Figure 2). In a similar fashion 18 strata of illustrative individuals built combining sex, age and study degree have been represented, i.e. the groups already considered in the analysis of the peculiar language previously shown (with italic labels).

On the first factorial axis the distinction between men and women is still pointed out. On the second axis it notices instead a distinction between the families with sons in opposition to singles and families without sons, and this is the key to read the position of the activities regarding the sons (*GiocoConBambini* and *CuraFigli* in the map). Actually the "mean" age, i.e. the one in which people start his family, is getting to be near the described area, while the young age (singles) and the elder one (widower or without sons living in the family) are characterized by free time activities. Observing the illustrative individuals, the study degree appears to be discriminant for women: within every class of age, it notices a lowering in study degreee proceeding from left to right side, that is the area of the domestic activities (house care).

The men seem to be involved in job activities (paid work) and leisures. A particular aspect concerns *giocare con i bambini* with respect to *cura dei figli*: the first is a male activity while the second one, that regards washing, dressing, attending the children, is typical of the mothers, in particular of low study degree ones. It appears therefore the figure of a father as sons' "friend" rather than tutor.

Another interesting aspect is that only women take care of the laundry and wash the plates and kitchenware, while the house care (label *ordino casa* in figure) assumes a much less clean position. Even if it remains a typical female activity, it is however possible to assert that the men help women in this task. Nevertheless, it should be taken into account of single men in the sample.

A further relevant element is the clear distinction, on the basis of the daily activities, between employed and unemployed (inclding also retireds and housewives). In terms of coded variable it is expressed in the opposition between middle-age men with upper-middle degree study and old women with low study degree.

In order to reach some average individuals and the respective average days, it has been carried out a cluster analysis starting from the Correspondence Analysis, obtaining 20 classes. In the following some interesting results are deepen.

In a first instance appear the young mothers (at least 75% in the class) with adequate economic resources (85%), employed (81%), with upper-middle study degree (81%), living mostly in the center and northeastern Italy. The prevalent activities carried out during the average day regard, however, not the job but the children and the elder relatives care, and the relationships with the other relatives in general terms.

The second average individual is represented, on the contrary, by the women (99%) unemployed (96%), with low study degree (87%), living in the southern Italy and on islands (50%), tendentially old or middle aged and without sons (at least 75%), or with sons who have left the family, with middle-lower incomes. The average day of these class is focused on the house care activities, like washing the kitchenware and the laundry, making the shopping and reordering the house.

Another average individual is represented by the men (97%), middle-age (92%), employed (86%), married and with sons (97%), living in the northern Italy (58%) with upper-middle study degree (64%), who spend their average day sharing time in work, children care, reading and watching on tv, and in smaller measure in sport activities too.

A further class is constituted by the old men (at least 68%), retired (81%), mostly living in the northern Italy and with low study degree (68%). The activities carried out reflect the stereotype of the retired people, addicted to reading or watching on tv, and going out to to walk, meet friends and play cards with them.

5. Comparing Official Statistics needs and Textual Statistics proposals

Textual data usually belong to a collection of text written in natural language, without any constraints but those dealing with the language in which they are expressed. Nevertheless stated the extreme complexity of the implicit grammatical and syntactic structure, differently from a table of numerical or categorial data in which classically the structure is given by the different roles played by rows and columns, textual data are treated as unstructured data.

This work has shown how the Textual Statistics techniques can be used for extracting knowledge from document dataset, and therefore transforming the unstructured data in structured ones, allowing a detailed analysis of complex phenomena.

The attempt of obtaining average days for categories of individuals has furnished some "seeds" that encourage future developments in such direction, even if several matters still remain opened. First of all in sample surveys as the Time Use one there is the necessity of comparing the reached results with those of the other European investigations. At the present this is achieved by using a shared system of numerical codes that allows an objective comparison of the different activities.

Is it possible to imagine in the future a similar level of comparison directly analyzing the language used for describing the everyday life stories? Comparing the interpretation deduced by the factorial representations of lexical and conceptual profiles with the results of the classical tables built from the codes, it notices that also starting from the text is possible to underline a clean distinction between the male and the female time, the time of whom is involved in working activities and of those taht instead spend more time among the domestic boundaries. Nevetheless, at this point the analysis is able to describe the activities but not to quantify how many time is spent in doing them. This is possible only by considering a different text structure, taking into account the so called "episodes", i.e. one or more intervals of 10 minutes in which the main activity, the parallel activity, the place and the mean of transport or the person with whom time spent by the respondent do not change. As a further development it would be very interesting to consider an external information on the activities lenght.

Far from a trivial and evident statement, this research sphere seems to reassure on the validity of the results obtained by a textual analysis on the respondent activities description. The *semic nuclea* derived from the semantic categorization of the textual content are close to the numerical encoding of the activities. In this frame Textual Statistics open to infinite ways of innovatively analyzing the survey data.

It is also necessary to remark that in this study the strategy chosen has transformed the information in a suitable form to be used in a classical survey framework. Once implemented some techniques of quantitative analysis on typically qualitative data as the textual one, it seems opportune to recall interpretative criteria derived from other research fields, allowing a wide perspective on the linguistic phenomena.

The treatment of natural language, particularly in a century more an more characterized

by a widespread cultural homologation, can not ignore the language evolutionary mechanisms and the communicative processes. In the frame of a multivariate descriptive analysis is necessary to widen the explanatory power of classical statistical analysis with Textual Data Analysis and Text Mining strategies, but at the same time it is important to co-opt in the interpretation phase the "best practices" of the qualitative analysis, to understand the modalities of language expressions used by the actors involved in the communicative processes.

References

Balbi S., Misuraca M. (2005) Pesi e Metriche nell'Analisi dei Dati Testuali, *Quaderni di Statistica*, Liguori, 7, 55–68.

Benzécri J. P. (Ed.) (1981) *Practique de l'Analyse de Données. Tome 3: Linguistique & Lexicologie*, Paris, Dunod.

Bevans G. E. (1913) *How Working Men spend their spare time*, Columbia University Press, New York.

Bolasco, S. (1992) Criteri di lemmatizzazione per l'individuazione di coordinate semantiche, in: *Ricerca qualitativa e computer*, Cipriani R., Bolasco S. (Eds.), (1995), Franco Angeli, Milano, 87–111.

Bolasco, S. (1999) *L'analisi multidimensionale dei dati*, Carocci, Roma.

Bolasco S. (2005) Statistica testuale e text mining: alcuni paradigmi applicativi, *Quaderni di Statistica*, Liguori, 7, 17–53.

Bolasco S., Morrone A. (1998) La construction d'un lexique fondamental de polyformes selon leur usage, in: *Actes des 4es Journées internationales d'Analyse statistique des Données Textuelles (JADT 1998)*, S. Mellet (Ed.), Univiversité Sophie Antipolis, Nice, 155–166.

Bolasco S., D'Avino E., Pavone P. (2005) *Analisi lessicale dei diari giornalieri con strumenti di statistica testuale e text mining*, Invited talk at the conference "I tempi della vita quotidiana", ISTAT, Roma, 20 december 2005.

Lebart L. (1982) L'Analyse Statistique des Réponses libres dans les Enquêtes socio-économiques, *Consommation/Revue de Socio-Economie*, Dunod, 1, 39–42.

Lebart L. (2004) Validité des visualisations de données textuelles, in: *Le poid des mots. JADT2004*, Purnelle G., Fairon C., Dister A. (Eds.), UCL Presses Universitaires de Louvain, 2, 708–715.

Lebart L., Salem A. (1988) *Analyse statistique de données textuelles*, Dunod, Paris.

Misuraca M. (2004) La visualizzazione dell'informazione testuale, *Doctoral Thesis*, Università di Napoli "Federico II".

Romano M. C., Cappadozzi T. (2004) Il processo di codifica dei dati testuali dell'indagine Multiscopo "Uso del tempo", in: *Le poid des mots. JADT2004*, Purnelle G., Fairon C., Dister A. (Eds.), UCL Presses Universitaires de Louvain, 2, 957–968.

Romano M. C., Camporese R., Vitaletti S. (2004) Time Use Survey in Italy, paper presented at the XXVI Internazional IATUR Conference *Time Use: What's New in Methodology and Application Fields*, ISTAT, Roma, 27-29 october 2005.

Sorokin P. A., Berger C. Q.(1939) *Time Budgets of Human Behaviour*, Harvard University Press, Cambridge (MA).