Text typology and statistics
Explorations in Italian press subgenres

Marina Santini

According to Biber's definition, text types are represented by groupings of texts which are similar in their linguistic form, while genre categories are assigned on the basis of use. It is important to stress that texts from a single genre might be classified into different text types.

In the present research, an experiment will be carried out to single out different text typologies among Italian press subgenres only on the basis of morphosyntactic features. The approach will be corpus-based and the aim merely exploratory.

A virtually random sample will be extracted from two Italian tagged corpora, the sample divided into 22 files, a file for each subgenre, and 41 morphosyntactic features will be counted per article within each subgenre. The raw frequencies will be normalised. The dataset built from the normalised frequencies (quantitative variables) will be submitted to Descriptive Statistics and Factor Analysis.

Descriptive statistics gives information about the distribution of the variables. It is a preliminary step in any exploration and gives a better understanding of the set of data.

Factor analysis studies the correlations among a large number of interrelated quantitative variables by grouping these variables into a small number of factors, which help to understand the structure of correlations or the underlying construct.

The aim of this research is to investigate to which extent statistical techniques can help in classifying texts in a steady and reliable way. *

1. Introduction

There is no general agreement on the criteria to use for the classification of texts. None of the systems proposed is comprehensive or generally accepted. Many criteria are available: internal, external, cultural, stylistic, etc.; many labels have been suggested: genre, register, typology, sublanguage, etc.; texts could be grouped according to their topic, the type of audience they address, their purpose, and so on. Different disciplines (linguistics, socio-linguistics, grammatical studies, corpus-analysis, literary criticism, rhetoric, etc.) often show clashing preferences on how to categorise texts, each field being keener on one aspect or another. All the categories suggested, however, have no neatly defined boundaries.

Rivista di Linguistica, 13.2 (2001), p. 339-374 *(ricevuto nel febbraio 2001)*

A remarkable attempt to find common and sharable criteria in defining text typology, mainly for electronic corpora needs, is being carried out by the EAGLES initiative. [1] Although still in a preliminary version, EAGLES guidelines on text typology contain useful indications on the different classifications and show how difficult it is to find unique criteria in this complex and extensively overlapping subject.

In order to show how text categorisation and labelling is deeply controversial among scholars, we now compare two different opinions.

1.1. Genre, Register and Text Typology

Biber's work has had a particularly large influence on British corpus linguistics. He has interests in many areas – from representativeness in corpus design to diachronic comparisons, from authors' styles to collocations, from anaphora handling to lexical bundles – but he uses a common approach in all of them, i.e. multivariate techniques. On the assumption that underlying co-occurrence patterns among linguistic features indicate sharing of communicative functions, he defines text types in linguistic terms, using factor analysis to detect hidden constructs.

In his well-known study on speech and writing, Biber (1988) used the term *genre* to refer to text categorisation carried out on the basis of external criteria. By genre he meant 'literary genre', i.e. general, cultural and widely accepted categories, such as novels, newspaper articles, public speeches, academic essays, etc.

In Biber (1995), he switched from the term genre to the term *register*, in order to emphasise the 'situation' in which a text is produced. He then used the word register to refer to all situationally-defined varieties, but he extended the covering of this term, including in it also named varieties within a culture, such as novels, letters, sermons, etc. Register distinctions are usually defined in non-linguistic terms, by differences in purpose, interactiveness, production, relations, etc. However,there are usually important linguistic differences among registers. Moreover, many texts are mixed and registers could be defined at any level of generality, for example an academic essay is a very general register, while the technical section of an essay on chemistry is a highly-specific register.

He used the term *text type*, instead, to refer to groupings of texts that are similar with respect to their linguistic form, regardless of genre or register classification. Text types are defined such that the

texts within each type are very similar from the linguistic point of view (lexical, morphological, syntactic, etc.). Only after having identified text types on linguistic grounds, can they be interpreted functionally in terms of purpose, production and other situational aspects.

Stubbs (1998), another authority, does not distinguish between text type and genre. He considers text types or genres as events which define the culture, i.e. they are both considered as conventional ways of expressing meanings. *Text types* are usually goal-directed and socially-recognised language activities, which form patterns and imply different ways of producing, distributing and consuming texts, while *genres*, indicating traditional categories in literary studies, like short stories, diary, biography, etc., refer to a distinction based not only on the aesthetic functions of language, but also on broader forms of cultural analysis, like science fiction, romance, and so on. The important point, however, is not knowing in some mechanical way which genre or typology a text belongs to, but knowing how the category can help to interpret it. The ability to identify different genres helps us to understand texts better. Misunderstandings of texts of different kinds can depend upon a lack of knowledge of the different conventions involved. The view that language varies systematically across text types or genres has implications for interpretation, that is texts are interpreted against a background of expectations, because their interpretation depends on both what they omit and what they express (Stubbs 1998: 12). This view of language variation has also the methodological implication that text study must be comparative. The most powerful interpretation emerges if comparisons of texts across corpora are combined with the analysis of the organisation of individual texts. However, as there is no convincing theory of how the frequency of linguistic features contributes to the meaning of individual texts, there is the need to combine the analysis of large-scale patterns across texts with the detailed linguistic study of them.

Stubbs criticises Biber because, even if Biber's approach provides a "powerful interpretative background of different genres" and "a powerful interpretative background for the analysis of individual texts", it provides "no analysis of the discourse structure of individual instances of genres" (Stubbs 1998: 34).

1.2. Corpora and Text Typology

Why is it useful to detect text typology and why is a more accurate classification of texts especially noteworthy? Corpus studies show that language in use is characterised by an astonishing

amount of regularities with endless variation. Detection of similarities and automated categorisation of textual entities play an important role in many areas, for example in the identification of hidden structures, retrieval of documents satisfying a query, resolution of morphosyntactic, syntactic, or semantic ambiguities, automatic abstracting, and so on.

1.3. Research on text typology on Italian

The majority of text typology studies on Italian is not based on machine-readable corpora. Investigations usually rely on a manual and qualitative inspection of selected features occurring in a number of texts belonging to different registers or genres. They are mainly based on observation of syntax and lexis rather than parts of speech.

For instance, Dardano (1994: 392) carried out much research on the language of the Italian contemporary press. Interesting findings were produced by accurate analyses made article by article, comparing and classifying genres and subgenres using qualitative manual micro-analysis. [2] Similarly, Sabatini (1990) plotted a comprehensive table including 30 features and 8 types of texts. In this table, each feature is specified with a dichotomic or binary attribute, + or –. Investigations on texts were done manually. The linguistic features included in the table are structurally complex.

One of the few examples of statistical investigations of a large machine-readable corpus was carried out by Pirrelli (1985). He applied Multidimensional Scaling (MDS) to the Italian corpus *La stampa periodica milanese della prima metà dell'Ottocento* ('Milanese periodicals in the first half of 19th century'). [3] The texts of this corpus were fully annotated, including macro-information (such as the kind of magazine and the (sub)genre), and micro-information (such as parts of speech and lemmas). (Sub)genres (for example, politics, theatre, entertainment, sciences, etc.) were defined on the basis of contents and style. The goal of the study was to investigate linguistic variation with respect to (sub)genres on the basis of parts of speech, which were handled as nominal data.

MDS is usually used to detect meaningful underlying dimensions, which could explain similarities or dissimilarities between the objects. MDS is a way to efficiently rearrange objects in order to obtain a configuration that best represents the distances between objects. Unlike other methods, it does not impose a preliminary hypothesis on the data.

In the present research, an attempt will be made to single out different text typologies among Italian press subgenres applying Biber's multidimensional approach [4] only on the basis of *morphosyntactic* features. The approach will be corpus-based and the aim merely exploratory. A virtually random sample will be extracted from two Italian tagged corpora, the sample divided into 22 files, a file for each subgenre, and 41 morphosyntactic features will be counted per article within each subgenre. The raw frequencies will be normalised. The dataset built from the normalised frequencies (quantitative variables) will be submitted to Descriptive Statistics and Factor Analysis.

1.4. A sample for Italian

The sample used in this research comes from two different corpora, LE-PAROLE and Elsnet. [5] A complete newspaper article has been taken as the minimum size of a text. The first step in creating the sample used in this research was to combine 250,000 tagged words from LE-PAROLE and 50,000 tagged words from ELSNET. From these 300,000 words, a further selection was made. A sample for multivariate analyses requires two main characteristics: a genre indicator and morphosyntactic annotation. Finally, the composition of the sample submitted to multivariate analyses was the following: 22 subgenres, 230 articles, 144,593 words. In general, a sample should be large enough so that correlations are reliable. As a rule of thumb, at least 5 cases should be included for each variable (Tabachnick & Fidell 1983: 603). As the number of observations increases, the reliability of correlations strengthens. The adequacy of sample size may be evaluated on the following scale: 50 cases = very poor; 100 cases = poor; 200 cases = fair; 300 cases = good, and so on (Comrey 1973: 200). The recommendation when the sample is not very large (as in this research) is a conservative interpretation of the results (Comrey 1973: 201).

1.5. Selection of Linguistic Features

Before attempting any comparison of texts, the linguistic features to be used must be selected. In theory, the widest possible range of potentially important linguistic features – especially those associated with particular communicative functions and therefore useful to single out different types of texts – should be included. In practice, it is often very hard to build a representative sample (corpus), including those cases (texts) and variables (frequencies of occurrences of lin-

guistic features) which could emphasise a specific methodology. In this study, the linguistic features selected are morphsyntactic, because this was the kind of annotation available in our corpora.

A total of 41 linguistic features were selected: adjectives, adverbs conjunctions, demonstrative determiners, indefinite determiners, possessive determiners, simple prepositions, other prepositions, interjections, cardinal numbers, ordinal numbers, demonstrative pronouns, indefinite pronouns, possessive pronouns, first person pronouns, second person pronouns, third person pronouns, relative pronouns (*che/cui*), other relative pronouns, determinative articles, indeterminative articles, common nouns, proper nouns, foreign nouns, infinitive, gerund, past participle, present participle, first person verbs, second person verbs, third person verbs, subjunctive, indicative: conditional, future, imperfect, present, simple past, imperative, predicative phrases, causative verbs, modal verbs.

2. *Methodology*

The datasets used in this research were handled and computed using Windows Excel 97. The rows, or cases, in the dataset represent the articles; the columns, or variables, are the linguistic feature frequencies of occurrence. The statistical package used in this study is *SPSS for Windows 9.0*. [6] The dataset (Excel files) were imported into SPSS Data Editor.

2.1. *Sample distribution*

SPSS uses the sample mean and standard deviation to construct the normal curve superimposed on the histogram. The histogram represents the distribution of the sample population. When bars on the left-hand-side are taller than the rest, as in our case (see the figure below), it means that the distribution is right-skewed. Skewness measures the symmetry of the sample distribution. In a positively skewed distribution most of the data is grouped below the mean, and a few data form a tail above the mean. In corpus analysis, much of the data is skewed (Fig. 1).

This was one of Chomsky's main criticisms of corpus data. [7] One of the answers to Chomsky's criticisms of the corpus-based approach appealing to the skewness argument is that skewness can be overcome by using lognormal distributions. In fact, it is possible to compute the base 10 logarithm for our variable (Fig. 2) in order to obtain

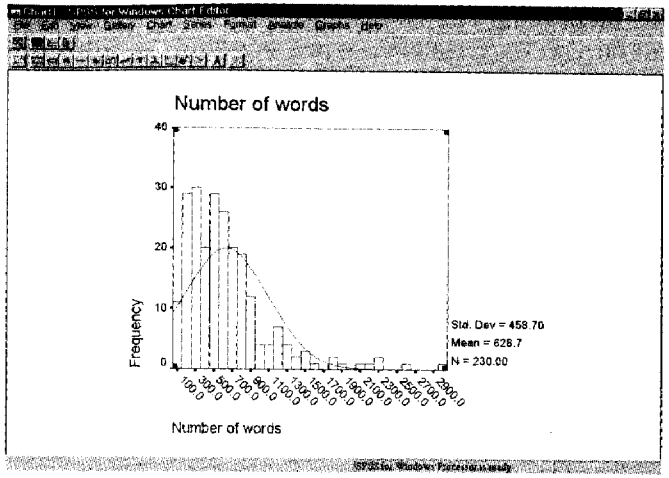

Figure 1. Distribution without normalisation.

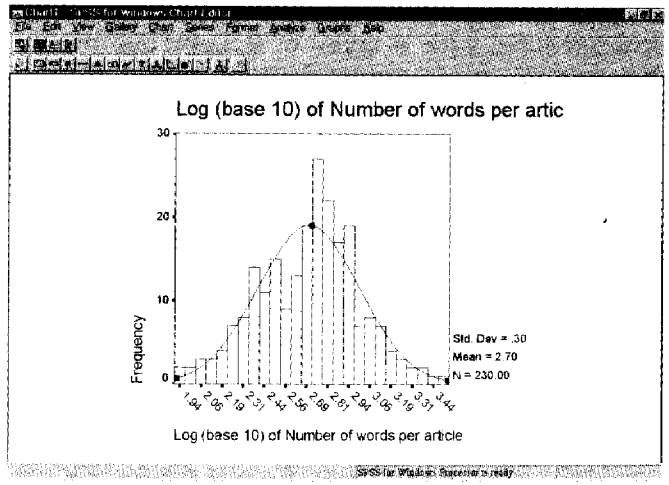

Figure 2. Log-normalised distribution of the new variable 'Number of words per article'.

a histogram displaying the sample values in log units. The transformation makes the distribution more symmetric than that for the untransformed data.

Besides log transformation, other transformations can be used, for example normalisation or standardisation (see below). Even if the linguistic feature frequencies of occurrence used in this research were normalised, normalisation of variables does not guarantee normality. Multivariate statistics, the kind of statistical techniques applied in this study, generally assume normality, but this assumption is usually difficult to meet. In theory, normalising the data increases the correlations, but in reality the advantage is not very outstanding; normality, however, is not needed as a standard requirement.

2.2. Descriptive statistics

Descriptive statistics are designed to give information about the distributions of variables. The SPSS Descriptive Statistics procedure was run on the dataset including all the linguistic feature frequencies of occurrence normalised to a value of 500 (Table 1).

Descriptive Statistics results do not enable characterisation of particular subgenres, but they provide an overview of the overall distribution of particular features in Italian press subgenres. Some features occur very frequently, for example 'Common Nouns' with a mean of 124 per 500 words; other features occur very infrequently, for example 'Interjections' with a mean of 0.2 per 500 words (Table 1). The variability in the frequency of features differs from one feature to another. For example, while 'Causative Verbs' seem to be more evenly distributed across the sample, with a maximum frequency of 4.5 and a minimum frequency of 0 per 500 words, other features show a wide gap, for instance 'Simple Prepositions' occur 173 times in some texts and not at all in other texts and the same is true for 'Indeterminative Articles', occurring 122 times in some articles and being absent from others. The most infrequent feature is 'Possessive Pronouns', while the most frequent is 'Common Nouns'.

Table 1. Descriptive statistics of the sample.

Linguistic features	Min	Max	Mean	Std. Dev.	Skewness	Kurtosis
Common Nouns	33.13	330.77	124.6620	23.6396	2.839	25.619
Simple Prepositions	.00	173.08	56.0409	12.9172	2.818	29.956
Other Prepositions	.00	113.64	40.1758	14.1579	.912	3.460
Determinative	2.73	80.77	37.5903	9.0325	.104	3.530
Adjectives	4.22	82.52	36.7778	10.4931	.195	1.794
Third Person Verbs	.00	61.51	34.0831	9.2950	-.300	1.553
Proper Nouns	.40	184.62	33.8031	27.0850	2.469	8.992
Indicative - Present	.00	58.50	26.9005	11.1358	.001	.269
Conjunctions	1.74	121.15	24.8648	10.5201	3.290	29.490
Adverbs	.00	67.96	23.6028	10.7810	.385	.722
Past Participle	.00	69.23	21.4950	10.5537	1.110	2.561
Cardinal Numbers	.00	111.93	13.9938	13.1549	2.778	13.942
Infinitive	.00	33.50	11.0095	6.8052	.561	.161
Indeterminative	.00	122.22	10.9545	10.9044	6.488	57.881
Relative Pronouns	.00	13.54	5.6575	2.9774	.049	-.391
Indefinite	.00	54.28	3.9039	4.4564	6.853	72.326
Indicative - Future	.00	18.52	3.0184	3.9920	1.749	2.663
Demonstrative	.00	54.28	2.7208	4.5571	8.172	83.545
Indicative	.00	28.07	2.7188	4.1508	2.682	9.658
Possessive	.00	17.08	2.6423	2.7710	1.602	3.821
Determinative	.00	13.08	2.5701	2.2304	1.328	3.326
First Person Verbs	.00	26.49	2.2461	4.0011	2.816	9.384
Indefinite Pronouns	.00	62.50	2.2375	4.5629	10.330	133.842
Modal Verbs	.00	8.48	1.8967	1.8826	.875	.110
Gerund	.00	22.99	1.8365	2.2638	4.339	33.919
Subjunctive	.00	12.33	1.7311	2.1040	1.876	4.633
Predicative Phrases	.00	9.67	1.6378	1.6886	1.365	2.856
Ordinal Numbers	.00	11.54	1.4426	1.6851	1.911	5.989
Conditional	.00	10.44	1.2412	1.7558	1.997	4.972
Indicative - Simple	.00	30.37	1.1354	3.7251	5.670	36.251
Foreign Nouns	.00	14.04	1.0058	2.4262	3.180	10.248
Other Relative	.00	51.91	1.0050	4.3235	10.204	110.052
Third Person	.00	11.42	1.0005	1.8574	2.991	10.466
First Person	.00	12.24	.9124	1.9535	3.102	10.923
Causative Verbs	.00	4.56	.5868	.8906	2.026	4.757
Second Person Verbs	.00	21.83	.4169	1.9755	8.355	79.803
Present Participle	.00	14.29	.4154	1.4568	7.630	67.752
Interjections	.00	21.77	.2905	1.8409	9.902	104.656
Imperative	.00	17.47	.2103	1.4544	9.730	103.216
Second Person	.00	5.52	.1507	.6355	5.540	35.043
Possessive Pronouns	.00	1.83	4.743E-02[8]	.2030	5.311	33.602

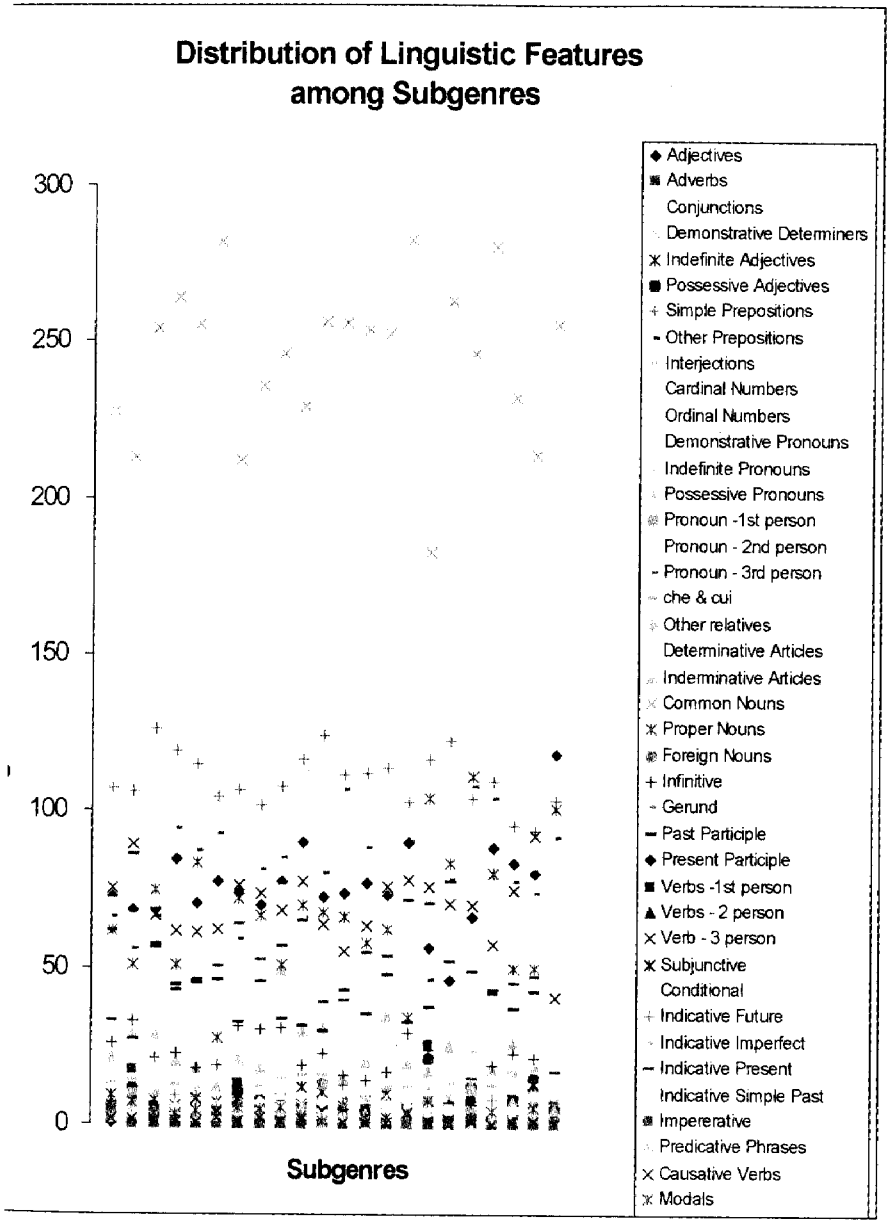

Figure 3. Scatter Plot of the Linguistic Features.

A graphical distribution of normalised linguistic features frequencies per subgenre is shown in Figure 3.

As noted above, frequencies of occurrence of individual features do not provide a complete description of textual variations or textual relations among subgenres. For these purposes, multivariate statistical techniques must be used.

Multivariate statistics includes that part of statistics concerned with multiple measurements made on one or several samples. The important thing in multivariate data analysis is that the multiple measurements are considered in combination, as in an interrelated system. Multivariate techniques are used to analyse complicated data, i.e. when there are many variables, all correlated to one another to a different degree. Most of the procedures of multivariate analyses are concerned with the problem of reducing the original number of variables in order to summarise them in fewer elements encompassing the most important information included in the original observations.

There are many kinds of multivariate techniques. In this research factor analysis was used.

2.3. Factor analysis

Factor analysis is a useful tool for generating hypotheses. The constructs, or factors, that emerge from a factor analysis are useful for understanding and describing relationships, but the correctness of any interpretation must be confirmed by evidence outside the factor analysis itself, because there is nothing in the factor analytic methods themselves that can demonstrate that one factor solution is stronger than another: the final choice among alternatives depends, in the end, on personal assessment.

Principal components and common factor analyses are often joined together under the heading 'Factor Analysis'. Although they are based on different mathematical methods, they can be used on the same data, and produce similar results. These procedures are often used in exploratory analyses to study the correlations among a large number of interrelated quantitative variables. Variables are grouped into a few factors [12] and, after grouping, the interpretation is simpler because the variables within each factor are more correlated with the variables in that factor than with variables in other factors. To get an empirical summary of a dataset PCA is the better choice (Tabachnick & Fidell 1983: 625).

Factor analysis includes the following major steps:
1. selection of the variables
3. computing the correlation matrix among the variables
4. extracting the unrotated factors
5. rotating the factors
6. determining how many factors to retain
7. interpreting the rotated factors

2.3.1. Variables and Normalisation

The variables included in a factor analysis must be quantitative. Algorithms have been used to count the linguistic features in the sample. The frequency counts of all linguistic features have been normalised to a text length of 500 words. Normalisation is crucial for any comparison of frequency counts across texts, because text length can vary widely from one article to another. A comparison of non-normalised counts will give an unreliable assessment of the frequency distribution in texts, because raw totals do not represent comparable frequencies of occurrence. By normalising the total counts to a text length of 500, i.e. computing how many occurrences of a given linguistic feature would occur if the text had been 500 words long, the frequencies of occurrence can be compared directly. The formula to normalise to a text length of 500 is the following:

raw_frequency / text_length * 500 = normalised_frequency

2.3.2. Correlation matrix

Factor analysis typically begins with the correlation matrix of the variables being studied. A correlation matrix is a square, symmetrical matrix. When the correlation matrix has substantial correlation coefficients in it, this indicates that the variables involved are related to each other, or overlap in what they measure. It is to be noted that the significance of a correlation coefficient of a particular magnitude will change depending on the size of the sample from which it is computed. With a large number of variables and many substantial correlations among the variables, it is difficult to observe all relationships. But factor analysis provides a way of handling these interrelationships by positing the existence of underlying factors that account for the values appearing in the matrix of correlated variables.

A matrix that is factorable should include several sizable correlations. The expected size depends, to some extent, on the number of cases in the sample, because larger samples tend to produce smaller

correlations, but if no correlation exceeds .30, factor analysis can be used only in its most exploratory and pragmatic sense, because there is probably nothing to factor analyse.

Sophisticated tests of factorability of the correlation matrix are available. One is *Bartlett's test of sphericity* which is a sensitive test of the hypothesis that the correlations in a correlation matrix are zero. The use of the test is recommended only if there are fewer than 5 cases per variable. Another test is *Kaiser-Meyer-Olkin measure of sampling adequacy* (KMO) tests whether the partial correlations among variables are small. It measures if the distribution of values is adequate for conducting factor analysis. If partial correlations are small, the value approaches 1. Values of .6 and above are required for good factor analysis.

2.3.3. Extraction of the unrotated factors

After the correlation matrix has been computed, the next step is to determine how many factors are needed to account for the pattern of values found in that matrix. This is done through a process called factor extraction.

The usual procedure is to extract factors from the correlation matrix until there is no appreciable variance left, that is, until the 'residual' correlations are all close to zero and their importance is negligible. After the 1st factor is extracted, the effect of this factor is removed from the correlation matrix to produce the matrix of first factor residual correlations. If a substantial number of values remains in the 1st factor residual correlations, however, it is necessary to extract a 2nd factor. If substantial values remain in the 2nd factor residual correlations, a 3rd factor must be extracted, and so on, until the residuals are too small to continue.

There are many methods to extract a factor, but they all end up with columns of numbers, one for each variables. These numbers are the *loadings* of the variables on that factor. The loadings represent the extent to which the variables are related to the hypothetical factors. They can be thought as correlations between the variables and the factors. A variable can also have a substantial negative loading on the factor, indicating that it is negatively, or complementarily, correlated with that factor. Loadings are then rotated (see next section 2.3.4).

Other important concepts are represented by commonality and eigenvalue. *Commonalities* represent the extent of overlap between the variables. That is, they are designed to show the proportion of variance that the factors contribute to explaining a particular variable. These values range from 0 to 1, with 0 indicating that common

factors explain none of the variance in a particular variable, and 1 indicating that all the variance in that variable is explained by common factors. However, for the default procedure at the initial extraction phase, each variable is assigned a commonality of 1.0.

The variance of the factors is commonly known as the *eigenvalue*. Eigenvalues are designed to show the proportion of variance accounted for by each factor (not by each variable as do commonalities). The first eigenvalue will always be the largest one and greater than 1.0 because, by default, it explains the greatest amount of total variance. It then lists the percent of the variance accounted for by this factor (the eigenvalue divided by the number of variables), and this is followed by a cumulative percent. For each successive factor, the eigenvalue printed will be smaller than the previous one, and the cumulative percent of variance explained will total 100% after the final factor has been calculated. If the variables in our dataset were independent of one another, there would be 41 components, each with a variance of 1.

One criterion for determining the number of useful factors for extraction is to exclude factors with variances less than 1, because they hardly correspond to a single independent variable. Often, for real data, there may be one or more eigenvalues close to 1, so one may need to request fewer factors than extracted by default. The place where there is a relatively large interval between values is usually taken as a cut point. It is necessary to examine the loadings for solutions with different numbers of factors to see which results provide the best interpretation of the data.

2.3.4. Rotation of the factors

The factors represented in an unrotated factor matrix are not easy to read. These unrotated factors are very complex because they relate to or overlap with many of the variables rather than with just a few; moreover they are not homogeneous and include many unrelated parts.

However, it is possible to 'rotate' the factor matrix into another form that is mathematically equivalent to the original unrotated matrix but which represents factors in a more interpretable fashion. In other words, rotation methods make the loadings for each factor either large or small, not in-between, and their interpretation becomes easier.

There are two general types of rotation: orthogonal and oblique. An *orthogonal rotation* generates factors uncorrelated with each other.

However, when it is supposed that underlying processes may be correlated, an oblique rotation is recommended.

In an *oblique rotation*, the factors are correlated and the loadings represent a measure of the unique relationship between the factor and the variables.

Anyway, different methods of rotation tend to give similar results if the pattern of correlations in the data is fairly clear, that is, a stable solution tends to appear regardless of the method of rotation used.

In this research, the oblique rotation Promax was applied to the dataset. In Promax rotation, an orthogonal rotated solution, usually Varimax, is rotated again to allow correlations among factors. While Varimax maintains orthognonal structure, requiring the assumption that the factors are uncorrelated, Promax permits oblique structure, that is, the moderate and low loadings are made lower than the orthogonal solution while the high loadings remain relatively high. When the factors represent underlying textual dimensions, it is assumed that the factors are correlated, therefore an oblique (Promax) rotation is recommended. [10]

2.3.5. Determining how many factors to retain

After the first few factors, there are typically several factors of lesser importance. There is no precise solution to determine the number of factors to be retained. There are mathematical criteria available, but decisions of this kind ultimately are based on the sample size employed, which has nothing to do with the nature of variables being studied.

Several signs are useful in trying to decide whether or not to stop extracting factors. As Comrey points out, the important thing about stopping the factor extraction is that it is better to extract too many factors rather than too few (Comrey 1973:101-102). [14] The recommended procedure is to extract enough factors to be relatively certain that no more factors of any importance remain. Between a larger or smaller number of factors, the more conservative procedure prevails: it is better to extract the larger number and then discard the unnecessary factors afterwards. Extracting too few factors would result in loss of information, because the constructs underlying the excluded factors would be overlooked, and in a distortion of the factorial structure, because multiple constructs would collapse into a single factor.

The importance of a factor (or set of factors) is evaluated by the proportion of variance or covariance associated with the factor after

the rotation. The proportion of variance attributable to individual factors differs before and after rotation because rotation tends to redistribute variance among factors.

2.3.6. Interpretation

The last step is to interpret the results using the knowledge about the variables and any other pertinent information at one's disposal. Variables that have high loadings in each of the rotated factors must be picked up, studied and some hypotheses must be formulated concerning what they share in common. On the basis of this analysis, each factor must be given an appropriate name that helps in identifying it. Interpretation and naming of factors depend on the meaning of the particular combination of observed variables that correlate highly within each factor. Interpretation of factors is facilitated by the output of the matrix of sorted loadings where variables are grouped by their correlation with factors. The number of variables should be several times as large as the number of factors. There should be at least 5 variables for each factor.

In an ideal world, the factors having the highest loadings should have excellent face validity and measure some underlying construct. In the real world, this rarely happens. The output of factor analysis requires considerable understanding of the data, and it is rare for the arithmetic of factor analysis alone to produce entirely clear results.

2.4. Factor Analysis on the Italian sample

The SPSS Factor command automatically computes the 'Descriptive Statistics' table, sorted by the variable list, and a correlation matrix. The correlation matrix contains Pearson correlations. The size of the correlation (positive or negative) indicates the extent to which two linguistic features vary together. A large negative correlation indicates that two features covary in a systematic, complementary fashion, i.e. the presence of the one is highly associated with the absence of the other. A large positive correlation indicates that the two features systematically occur together.

KMO was quite poor (.490), but Bartlett's test was more encouraging (.000). SPSS extracts automatically the number of factors the variables allow. In our case, 14 factors were extracted. The following figure shows the scree plot of the Experiment.

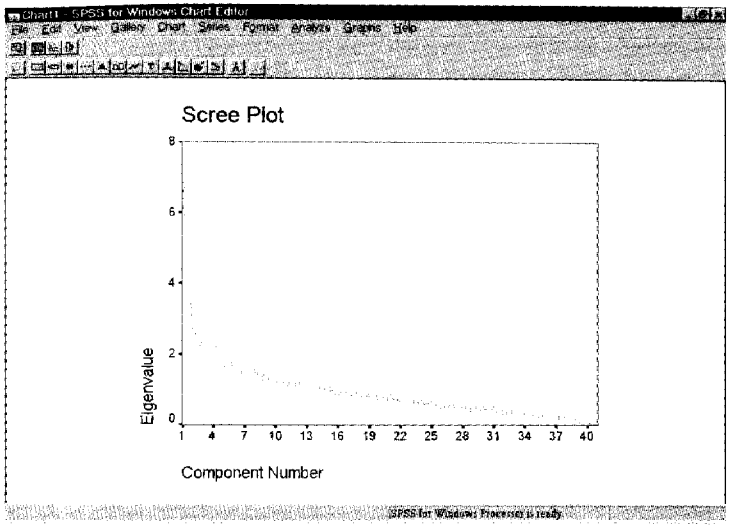

Figure 4. Scree Plot - 14 factors extracted automatically.

From the 14 factors, the 5 most relevant factors were retained. There are several techniques for determining the required magnitude of statistically significant loadings, i.e. the loadings not due to random patterns of variation. As a rule of thumb, only variables with loadings of .30 and above are interpreted. The greater the loading, the more the variable is a pure measure of factor. From Factor 6 onwards, loadings were very poor, that is why non-factorised features and their loadings on one or more factors have been ignored in this study. The complete factor loading matrix for the first 5 factors is given below; the barred figures represent weak duplicates that have not been included in the computation of factor scores.

Table 2. The factor loading matrix for the first 5 factors

	1	*2*	*3*	*4*	*5*
First Person Verbs	.992				
First Person Pronouns	.955				
Possessive Determiners	.695				
Third Person Pronouns	.497				
Indicative - Present	.483				
Possessive Pronouns	.371				
Causative Verbs		.714		~~.393~~	
Conditional	~~-.342~~	.683			
Infinitive		.652			
Modal Verbs		.632			
Subjunctive		.548			
Ordinal Numbers		-.341			
Imperative			.999		
Second Person Verbs			.997		
Second Person Pronouns	~~.421~~		.666		
Simple Prepositions				.905	
Common Nouns				.820	
Other Prepositions				.465	
Adjectives					.908
Past Participle					-.631
Predicative Phrases					.505
Conjunctions					.383
Adverbs					.370
Indicative - Future					
Third Person Verbs					
Cardinal Numbers					
Indeterminative Articles					
Relative Pronouns (che/cui)					
Demonstrative Determiners				-.379	
Indefinite Determiners					
Other Relative Pronouns					
Determinative Articles					
Gerund					-.336
Foreign Nouns		-.375			
Present Participle					
Proper Nouns					
Indicative - Simple Past					
Indicative - Imperfect					
Indefinite Pronouns					
Interjections					
Demonstrative Pronouns					

In general, 5 salient loadings are required for a meaningful interpretation of the construct underlying a factor. However, when there are only 2 or 3 loadings for one factor but showing outstanding values, that factor can be retained, even if very cautiously. That is why Factor 3 has been kept (see below).

Here is the summary of the factorial structure of the first 5 factors:

Factor 1

First Person Verbs	.992
First Person Pronouns	.955
Possessive Determiners	.695
Third Person Pronouns	.497
Indicative – Present	.483
Possessive Pronouns	.371

Factor 2

Causative Verbs	.714
Conditional	.683
Infinitive	.652
Modal Verbs	.632
Subjunctive	.548
Foreign Nouns	-.375
Ordinal Numbers	-.341

Factor 3

Imperative	.999
Second Person Verbs	.997
Second Person Pronouns	.666

Factor 4

Simple Prepositions	.905
Common Nouns .820	
Other Prepositions	.465
Demonstrative Determiners	-.379

Factor 5

Adjectives	.908
Predicative Phrases	.505
Conjunctions	.383
Adverbs	.370
Past Participle	-.631
Gerund	-.336

Some of the loadings are important (for example, .999 for imperative), but the number of variables per factor is not conspicuous, especially for Factors 3 and 4.

In the interpretation of a factor it is assumed that the linguistic features included in that factor co-occur because they share common functions. As factors have been rotated, linguistic features load mainly on one factor, and consequently each factor is characterised by the features most representative of the underlying construct they share.