

Text Typology and Statistics

Explorations in Italian press subgenres

Marina Santini

Abstract

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 give 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.

Author's note.

This paper is an excerpt taken from my master dissertation and adapted to the length of an article. It represents only one of the exploratory experiments I carried out for that work. The experiment described in this paper suffers from some limitations, for instance, the sample is extremely restricted (only 144,593 words), the domain (press subgenres) not very differentiated, the number of morphosyntactic features quite high, the outliers excluded from the final considerations and so on. However, regardless these limitations, this experiment opens some interesting prospects on the handling of Italian corpora.

I wish to thank the revisers of this article for their useful, appropriate and accurate suggestions.

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.

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 initiative¹. 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

¹ Expert Advisory Group on Language Engineering Standards (*EAGLES*) is a European Commission initiative, started in February 1993 within the Linguistic Research and Engineering program. For further details, visit <<http://www.ilc.pi.cnr.it>>.

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, 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 (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-analysis². Similarly, Sabatini (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 -.

² One of Dardano’s most interesting findings is that boundaries between different kinds of articles or subgenres are nowadays almost faded out: the language of the press appears to be largely mixed.

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 (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 approach⁴ 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

³ See Ciccone De Stefanis, S., Bonomi, I. & Masini, A. eds. (1983), *La stampa periodica milanese della prima metà dell'Ottocento. Testi e concordanze*, Orientamenti Linguistici 19, Vol. 1_Testi, Pisa.

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 Elsnets⁵. 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

⁴ See BIBER (1988), where the approach is described in details.

⁵ These two corpora were built within and belong to the *Istituto di Linguistica Computazionale* (ILC) in Pisa. For further details about the Institute, visit <<http://www.ilc.pi.cnr.it>>.

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 linguistic features) which could emphasise a specific methodology. In this study, the linguistic features selected are morphosyntactic, 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 represent 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.

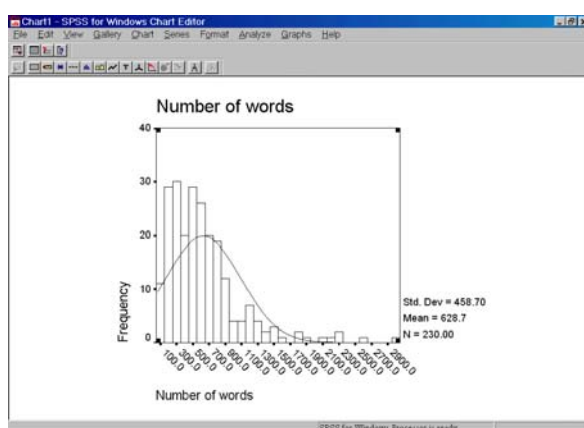


Figure 1. Distribution without normalisation.

This was one of Chomsky’s main criticisms of corpus data⁷. 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 (see Figure 2) in order to obtain a histogram displaying the sample values in log units. The transformation makes the distribution more symmetric than that for the untransformed data.

⁶ SPSS (*Statistical Package for the Social Sciences*) is a user-friendly program and makes statistical analysis accessible for inexperienced users. The proper use of the statistical procedures, however, requires understanding of many technical points.

⁷ A full description of Chomsky’s sceptical attitude towards corpus linguistics can be found in McENERY-WILSON (1996: 4-10).

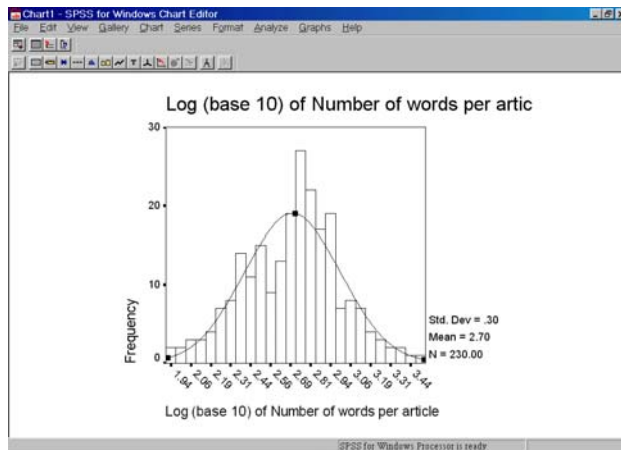


Figure 2. Log-normalised distribution of the new variable ‘Number of words per article’.

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 (see below).

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 (see the Descriptive Statistics Table below). 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'.

Descriptive Statistics of the Sample as a whole (descending order)

| Linguistic features | Min | Max | Mean | Std. | 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 |

Table 1. Descriptive statistics of the sample.

⁸ In the descriptive statistics table, the 'E' of 4.743E-02 'E' is a shorthand for 10^{\wedge} i.e. "times 10 to the power of", so $4.743E-02 \times (10^{\wedge}-2) = 4.743 \times (1/100) = 0.04743$.

A graphical distribution of normalised linguistic features frequencies per subgenre is shown in the following scatterplot:

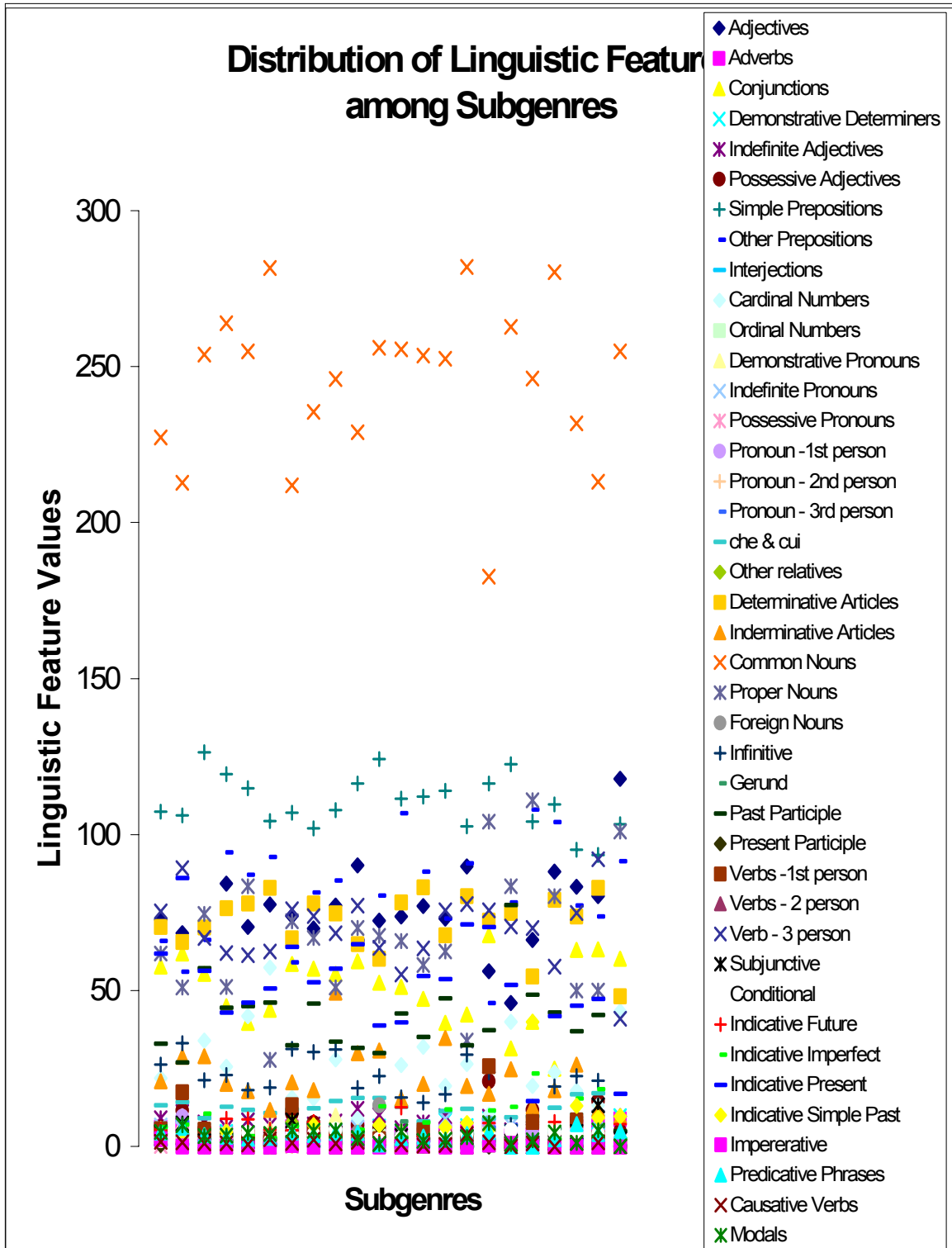


Figure 3. Scatter Plot of the Linguistic Features.

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 factors⁹ 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
2. computing the correlation matrix among the variables
3. extracting the unrotated factors
4. rotating the factors
5. determining how many factors to retain
6. 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:

⁹ PCA extracts ‘components’ and not ‘factors’, but here the term ‘factor’ will be used for both techniques.

$\text{raw_frequency} / \text{text_length} * 500 = \text{normalised_frequency}$

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).

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 2 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 orthogonal 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 recommended¹⁰.

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)¹¹. 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.

¹⁰ More generally, from a theoretical perspective, all aspects of language use appear to be interrelated to some extent, that's why there is no reason to hypothesize mathematically uncorrelated factors representing those aspects (BIBER 1988: 85).

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 2 linguistic features vary together. A large negative correlation indicates that 2 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

¹¹ Biber agrees with this approach, see BIBER (1988: 88).

positive correlation indicates that the 2 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.

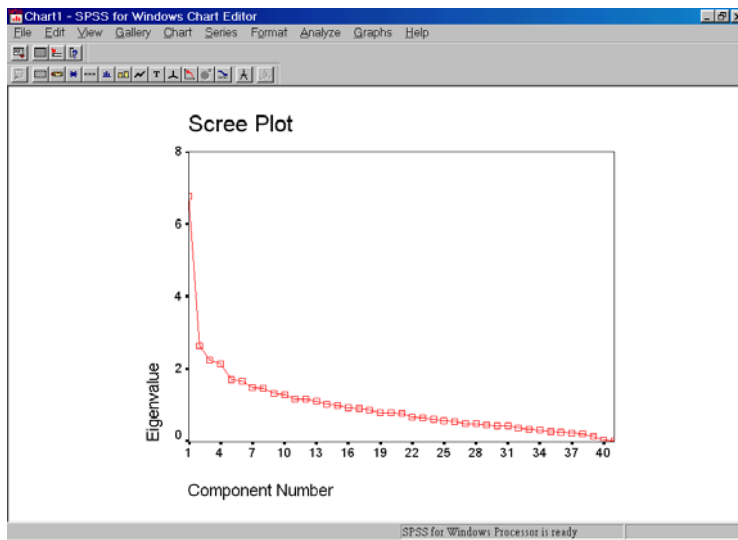


Figure 4. Scree Plot - 14 factors extracted automatically.

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.

| | 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 | -.424 | | .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 | | | | | |

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

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 | | Factor 2 | |
| First Person Verbs | .992 | Causative Verbs | .714 |
| First Person Pronouns | .955 | Conditional | .683 |
| Possessive Determiners | .695 | Infinitive | .652 |
| Third Person Pronouns | .497 | Modal Verbs | .632 |
| Indicative - Present | .483 | Subjunctive | .548 |
| Possessive Pronouns | .371 | ----- | |
| | | Foreign Nouns | -.375 |
| | | Ordinal Numbers | -.341 |
| Factor 3 | | Factor 4 | |
| Imperative | .999 | Simple Prepositions | .905 |
| Second Person Verbs | .997 | Common Nouns | .820 |
| Second Person Pronouns | .666 | Other Prepositions | .465 |
| | | ----- | |
| | | Demonstrative | -.379 |
| | | Determiners | |
| 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. An interpretation of these factors could be:

- a very strong verbal characterisation for **Factor 1**. Strong presence of 1st person pronouns, large use of present tense and 3rd person pronouns. It is to be noted that in most cases the use of personal pronouns in subject position is a marked feature in Italian. 1st person pronouns indicate an interpersonal focus and can be considered a marker of ego-involvement in a text. 3rd person pronouns mark reference to animate, typically human, referents apart from the speaker and addressee. They usually indicate that there is no direct participation and mark a reference to entities outside the immediate interaction. Present tense can refer to actions occurring in the immediate context or general and universal events, outside any temporal sequencing. In the last case, the present is used to give general definition. Possessive determiners and pronouns provide further information about nouns or referents. We call this factor *Speaker involvement* and it is characterised by speaker(s) describing objects or ideas.

- **Factor 2** has again a verbal character and denotes a more articulated syntax with causatives and modals, which are ‘supporting’ verbs bearing other verbal forms. Usually, infinitives need a matrix verb somewhere in the sentence. Modals and infinitives are marks of persuasion: either explicit (marking the speaker’s own persuasion or point of view) or argumentative discourse designed to persuade the addressee. In particular, modals indicate a direct involvement concerning the ability or the possibility of the occurrence of certain events. In particular, some uses of possibility modals represent a discursive strategy, since they can be used as "downtoners", with a general lowering effect as they mark politeness or deference (Biber 1988: 240-243). More sophisticated moods, like

conditional and subjunctive, which in most cases are not matrix verbs, confirm the presence of a more complex syntax. It seems that a more elevated and accurate use of the language is opposed to a large use of foreign words, which sometimes can show shabby laziness in finding the appropriate Italian synonym. We call this factor *Elaborated speech*.

- In **Factor 3** there is strong correlation of 2nd person verbs, especially imperatives, and pronouns. 2nd person pronouns typically express a high degree of involvement with the recipient of the act of communication. The imperative form is used to give orders, advice, to ask somebody to do something, give instructions, etc. The co-occurrence of 2nd person pronouns with the imperative forms implies that there is a direct reference to the addressee, i.e. the person who should carry out the action. We call this factor *Hearer involvement*.

- **Factor 4** includes nouns and prepositions. Nouns are the most important bearer of meaning and a high frequency of occurrence of nouns indicates high concentration of information. Prepositions are generally used to relate two or more elements and they are a common device used to integrate or consolidate information into a text. Demonstrative determiners represent the negative loadings and are usually associated to the immediate context, which can be text-internal or outside the text. We call this factor *Informational speech*.

- **Factor 5** seems richer, with a strong presence of adjectives and adverbs, and syntactically more complex, with conjunctions and predicative phrases. Adjectives, predicative forms and coordinations are all features that integrate and add information into a text. In particular, adjectives are used to elaborate nominal information, because they pack information in a synthetic and precise fashion, providing highly descriptive information. Adverbs can have a broad range of functions and can be used

for specification of manner, quantity, degree, and so on. Altogether they elaborate and expand the information presented in a text. High frequencies of occurrence of adjectives and adverbs usually indicate a high index of descriptive information. The negative loadings include past participles, a subordinate feature usually employed in more elaborated or sophisticated discourse, and gerunds, another subordinate feature which is a mark of a more complex syntax. We call this factor *Highly Integrated speech*.

The distribution of the subgenres along these factors/dimensions will be discussed below.

2.5 Factor Scores

Factor scores are particularly useful when the researcher wishes to perform further analyses involving the factors identified in a factor analysis. In the interpretation of a factor, an underlying aspect of hidden relations is sought to explain the co-occurrence among features grouped in the factor. When a set of features co-occurs frequently in texts, it means that they have a common function in those texts. When factors have been extracted using statistical techniques, the microanalysis of linguistic features becomes extremely important. Functional analysis of the features grouped in a factor allows identification of shared functions. It is important to note that, while factors or co-occurrence patterns are identified using a quantitative approach, interpretation of the relations (called *dimensions* by Biber) underlying a factor is a kind of plausible construction and requires confirmation.

One technique used to confirm a factor interpretation uses scores computed from the factors. In multidimensional research, factor scores represent *textual dimensions*. A factor score, or dimension score, can be computed for each text, so that similarities and differences among

subgenres, the *textual relations*, can be analysed with respect to these scores.

2.5.1 Technical Description

Once a factor has been isolated, it is necessary to relate this variable to other variables of interest. There are several methods to compute factor scores, but in this research they are computed according to the method used by Biber in his study on speech and writing variation.

After having singled out all those variables that have factor loadings on the factor above a selected cut-off value (in the present research a cut-off of .30 has been adopted), the scores of the variables having salient loadings (i.e. above the cut-off) on a factor were scaled to the same mean and standard deviation and then added up for each text.

Standardisation was performed on the normalised frequency values. All frequencies of occurrence were standardised to a mean of 0.0 and to a standard deviation of 1.0. The mean is the measure of the central frequency of a features; the standard deviation is a measure of the spread of frequency values of a features. When the frequency values are standardised, they are translated to a new scale. For example, 'Proper Nouns' have a mean value of 33.8 and a standard deviation of 27.0850, as shown in the Descriptive Statistics table above. Therefore, if a text has 33 proper nouns, it would have a standardised score of 0.0 for this feature because its frequency equals the mean. The standardised score of 0.0 indicates that this text is unmarked with respect to this feature. Instead, if a text has a frequency of 102, it would have a standardised score of 2.5. This score is computed using the following formula:

$$(102 - 33.8) / 27 = 2.5$$

The score of 102 is 2.5 standard deviations more than the mean 33.8, and the standard score of 2.5 shows that this text is quite marked with respect to proper nouns.

This procedure prevents those features that occur very frequently from having too much influence on the computed factor score. Frequencies standardised to a standard deviation of 1.0 retain the range of variation for each linguistic feature and at the same time standardise the absolute magnitude of those frequencies to a single scale.

A standardised score can be also negative, if the frequency of a feature in a text is markedly less than the mean frequency for the entire corpus. For example, the standardised score is -1.5 for Adjectives, reflecting the fact that there are fewer adjectives in this text than the mean number of adjectives in the corpus as a whole. The effect of this method of computation is to give each linguistic feature a weight in terms of the range of its variation rather than in terms of its absolute frequency of occurrence in texts. Standardised values, reflecting the magnitude of a frequency of occurrence with respect to the range of possible variation, is a more adequate representation for the purposes of the present study.

To illustrate the computation of factor scores in the present research, let us take Factor 1 of the Experiment as an example. The factor score representing Factor 1 is computed by adding together the standardised frequencies of occurrence of: First Person Verbs, First Person Pronouns, Possessive Determiners, Third Person Pronouns, Indicative–Present, Possessive Pronouns, i.e. the figures with positive loadings, for each text. No standardised frequencies of occurrence are subtracted in this case, because there are no negative loadings for this factor.

When some features have salient loadings on more than one of the factors, each feature is included in the computation of only one factor

score, the one on which it has the highest loading (in terms of absolute magnitude, i.e. ignoring + or -).

Subsequently, the mean of factor scores for each subgenre has been computed. There can be outlying cases among the factors, i. e. cases ‘unusual’ with respect to their scores on the factors. These are cases that have unusually large or small scores on the factors. Examination of these cases for consistency can be informative. If scatterplots between pairs of factors were plotted, these cases would appear along the borders.

2.5.2 Distribution of Subgenres along the Dimensions

Similarities and differences among subgenres can be considered with respect to each of the 5 factors/dimensions. Subgenres can be compared along each dimension using factor scores. The distribution along Factor 1, (see Figures in Appendix A), shows that ‘Interview’ has the highest verbal profile, followed by ‘Testimony’. Their connection with this factor is clear: verbs and pronouns are used in conversations or dialogue-like texts where interaction, immediate reference with the context and personal focus are stressed. ‘Interview’ is relatively strong also on Factor 2, followed by ‘Editorial’. This dimension is characterised by elaborated and more syntactically complex profile, with strong co-occurrence of causatives and modals, as well as conditional, subjunctive and infinitive. In Factor 3, ‘Analysis’ is very strong, followed by ‘Interview’. This dimension indicates hearer’s involvement with its strong presence of 2nd person pronouns and verbs. It is important to note how the factor score values are relative within each factor. For example, ‘Analysis’ is outstanding in Factor 3, but it shows its highest score in Factor 1.

Italian readers can compare the factor features directly, by having a look at the following articles shown in full in Appendix B (the rows below are extracted from the entire table):

Interview

| | 1 | 2 | 3 | 4 | 5 |
|---------------|------|-------|------|-------|-------|
| *%B318SICU85I | 7.63 | -1.08 | 0.33 | -5.71 | -2.29 |

Analysis

| | | | | | |
|---------------|------|-------|-------|-------|-------|
| *%A001SIPO85I | 5.99 | 24.09 | -0.67 | 16.46 | 15.89 |
|---------------|------|-------|-------|-------|-------|

Analysis

| | | | | | |
|---------------------------------|-------|-------|-------|-------|-------|
| *%CN00/04/88 INFORMATI COSI' | 39.27 | 10.39 | 96.17 | 40.56 | 48.64 |
|---------------------------------|-------|-------|-------|-------|-------|

Review

| | | | | | |
|--|-------|--------|-------|-------|-------|
| *%RE15-07-95 ALFRED HITCHCOCK CHE FILIBUSTIERE | 15.90 | -13.04 | -0.67 | 62.11 | 29.01 |
|--|-------|--------|-------|-------|-------|

Analysis

| | | | | | |
|----------------------------------|-------|-------|-------|------|------|
| *%SI00/04/88 UN MILIARDO? NO, | -1.40 | -1.69 | -0.67 | 1.65 | 4.35 |
|----------------------------------|-------|-------|-------|------|------|

2.6 Some remarks

It can be noticed that all the factors are affected by a relevant frequency communality of the linguistic categories they include. For instance, Factor 3 includes 3 infrequent linguistic features (Second Person Verbs, Imperative, Second Person Pronouns). This is due to the characteristics of the sample, and Factor 3 probably just reflect their co-occurrence in the texts where their instances are found. Furthermore, the feature "Imperative" shows a deep asymmetry in the occurrence of this features and the others and this fact raises the doubt that Factor 3 just catch a sampling bias. Moreover, the suspect that Factor 4 is the product of a general frequency factor has not been submitted to further control in this exploration, besides the influence of the outliers has not been measured. Even though normality is NOT a necessary assumption for PCA (NORUŠIS 1999: 320), skewed distributions and outliers can distort the results. The ideal solution against this state of things lies in modifying the sample, which should be more representative of all the linguistic variables considered. Probably, a better approach consists in performing the analyses on larger chunks of text (thus reducing the likelihood of dealing with points

equal to zero) and normalizing them to 1000. The skewness of the frequency distribution could be reduced by grouping the number of linguistic variables in couples or triplets in a single category. But this is the food for next research! The present paper just is tentative and exploratory, its goal is simply to highlight what is needed for a more representative investigation

3 Conclusions

The goal of this research was to explore the influence of morphosyntactic elements in the identification of different text types in Italian press subgenres. Observations of very specific aspects (morphosyntactic features), applied to a very restricted domain (Italian contemporary press subgenres), and using a relatively small sample was a challenging test for the statistical-multidimensional approach. The test, however, proved successful and the experiment carried out shows that, even though morphosyntactic features represent only a restricted perspective on a language, they can indeed give some hints about text typology and help to categorise texts.

Descriptive Statistics gave a first flavour of the characterisation of data, showing quantitative variations across the whole sample.

Factor analysis unveiled the underlying correlations among features and, consequently, among texts bearing the same features. Factor analysis is an important tool to obtain generalisation. Before its application to linguistics, only a limited number of features, selected upon linguists' intuition, were claimed to characterise texts. From our Experiment, 5 strong and outstanding morphological factors, or dimensions, came out (*Speaker involvement*, *Elaborated speech*, *Hearer involvement*,

Informational speech and Highly-Integrated speech), which can be traced in the texts shown in Appendix B.

Statistical techniques can prove very helpful in text classification and can yield steady and reliable results if applied knowingly. It is important to point out that text categorisation is a thoroughly interdisciplinary area. It is a new blend of many different disciplines: it requires a traditional linguistic background, which is important for the selection of linguistic features and for the interpretation of the results; a corpus-linguistic approach, because annotated corpora represent the foundations of the analyses; programming skills, because algorithms are needed for the countings of tag frequencies as well as for the identification of features not included in the annotated corpora; and, last but not least, statistical flare and expertise, because statistical techniques must be known down to their mathematical core in order to take full advantage of them.

One of the major concerns is represented by annotated corpora. If many kinds of different information were annotated in machine-readable corpora, more useful and richer results could be achieved in a shorter amount of time. In an ideal world, crowded with diversified, fully-annotated and standardised corpora, a quantitative approach could lead to more productive and helpful findings. The approach used in this study could be easily applied to a more extensive research on text typology aiming at text categorisation.

The identification of different typologies or categories of texts can make a difference to the way a text is interpreted, because the ability to identify different kinds of texts contributes to their understanding. In fact, the full understanding of a text entails the comprehension of the

conventions the text involves, and conventions are one of the key concepts in linguistics (STUBBS 1998: 12).

Text categorisation is a basic and, at the same time, sophisticated requirement in many cutting-edge areas, such as information retrieval, machine translation, artificial intelligence, and sublanguage studies¹².

¹² The state-of-the-art in text categorisation and text similarities can be found in LEBART-RAJMAN (2000).

Bibliographical References

- Biber, D. (1988), *Variations across speech and writing*, Cambridge University Press, Cambridge.
- Biber, D. (1995), *Dimensions of register variation*, Cambridge University Press, Cambridge.
- Comrey, A. L. (1973), *A First Course in Factor Analysis*, Academic Press, New York-London.
- Dardano, M. (1973), *Il linguaggio dei giornali italiani*, Edizioni Laterza, Bari.
- Dardano, M. (1987), "Il linguaggio dei giornali", in Jacobelli, J. (1987), *Dove va la lingua italiana*, Laterza, Bari, pp. 58-65.
- Dardano, M. (1994), "Profilo dell'italiano contemporaneo", in L. Serianni – P. Trifone eds. (1994), *Scritto e Parlato*, Vol. II, G. Einaudi Editore, Torino, pp. 343-430.
- Dardano, M., Giovanardi, C., Pelo, A. et alii (1992), "Testi misti", in Moretti, B., Petrini, D. & Bianconi, S. eds.(1992), *Linee di tendenza dell'italiano contemporaneo*. Atti del XXV Congresso Internazionale di Studi della Società di Linguistica Italiana, Lugano 19-21 settembre 1991, Bulzoni, Rome, pp. 323-351.
- Gorsuch, R. L. (1974), *Factor Analysis*, W. B. Saunders, Philadelphia – London-Toronto.
- Lebart, L. & Rajman, M. (2000), "Computing Similarity", in Dale, R., Moisl, H. & Somers, H. eds. (2000), *Handbook of Natural Language Processing*, M. Dekker, Inc., New York-Basel , pp. 477-505.
- McEnery, T. & Wilson, W. (1996), *Corpus Linguistics*, Edinburgh University Press, Edinburgh.
- Pirrelli, V. (1985), *La statistica multivariata nell'analisi linguistica del testo. Un esempio: lo Scaling Multidimensionale*, Unpublished dissertation, Università degli Studi di Pisa, supervisor: Antonio Zampolli.

Sabatini, F. (1990), "Analisi del linguaggio giuridico. Il testo normativo in una tipologia generale dei testi", in *Corso di studi superiori legislativi 1988-1989*, Cedam, Padova, pp. 675-724.

Stubbs, M. (1998), *Text and Corpus Analysis*, Blackwell Publishers, Oxford, (first published 1996).

Tabachnick, B. G. & Fidell, L. S. (1983), *Using Multivariate Statistics*, HarperCollins Publishers, Second Edition.

Appendix A

Column Charts – Distribution of Subgenres along each Factor/Dimension

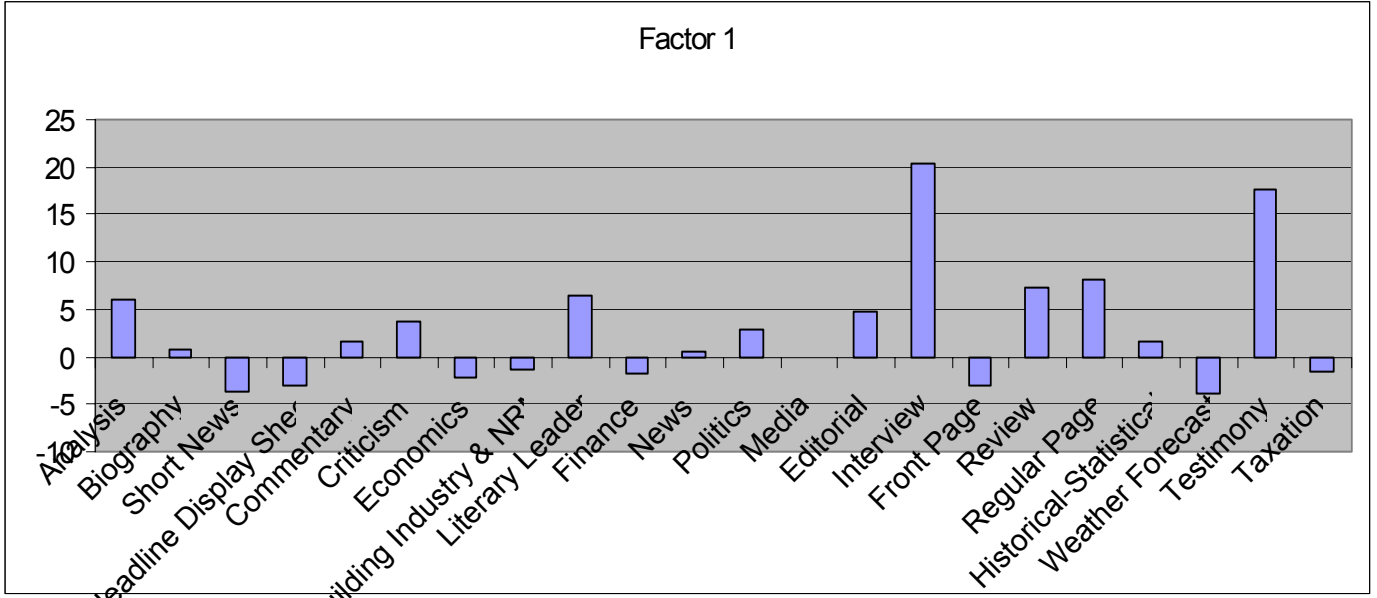


Figure 5. Factor 1 – Speaker involvement.

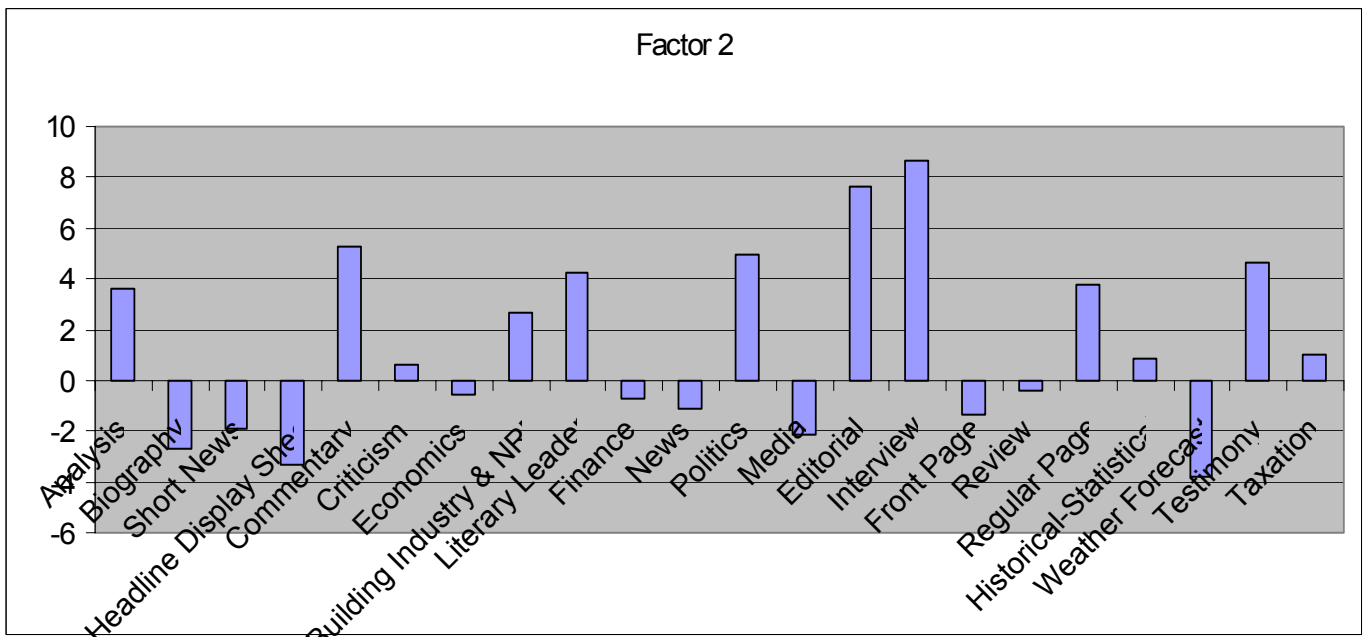


Figure 6. Factor 2 – Elaborated speech.

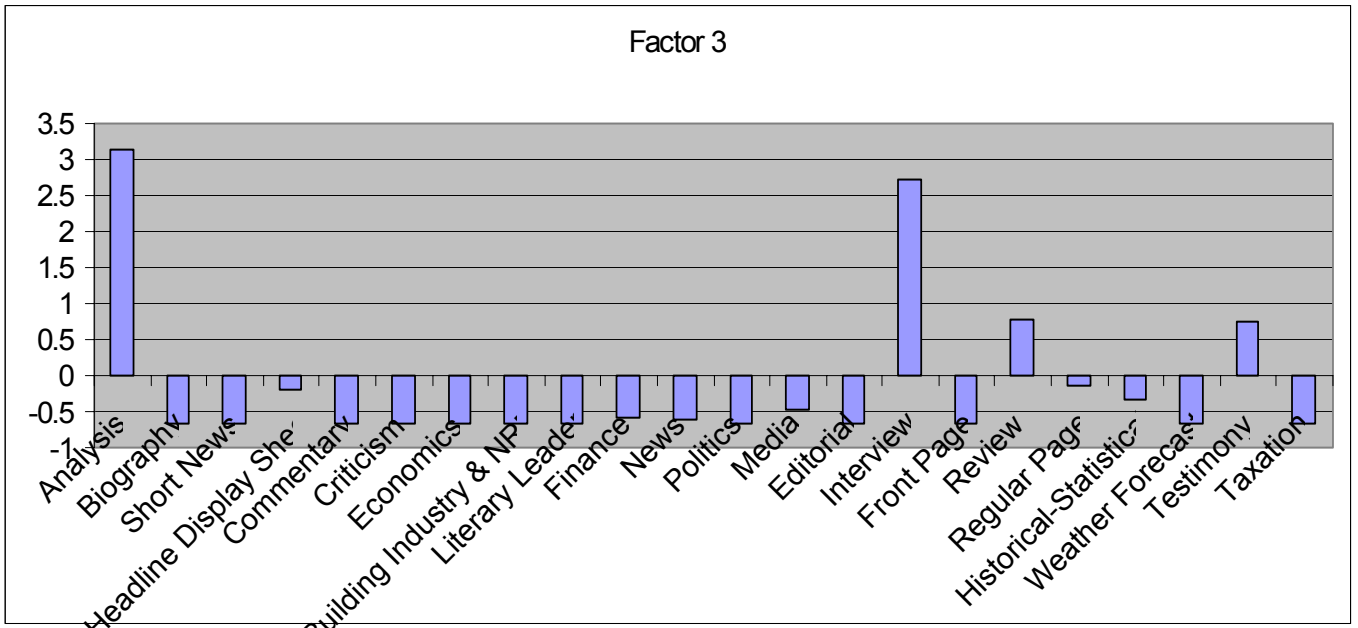


Figure 7. Factor 3 – Hearer involvement.

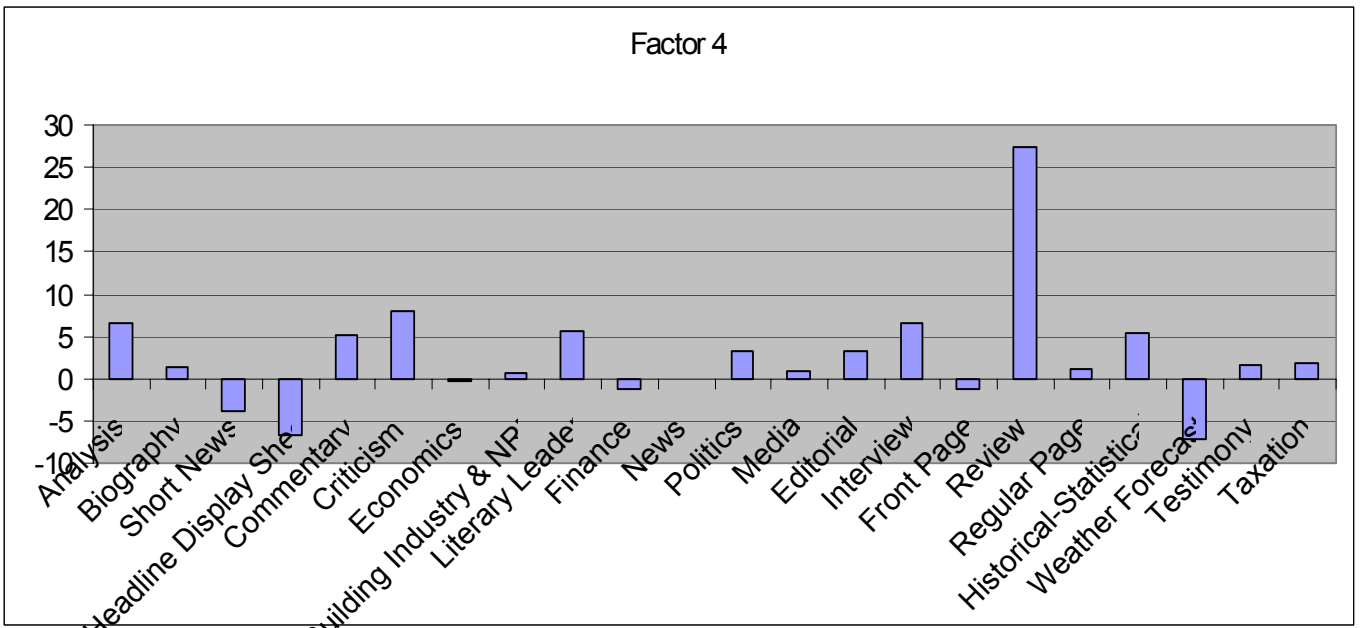


Figure 8. Factor 4 – Informational speech.

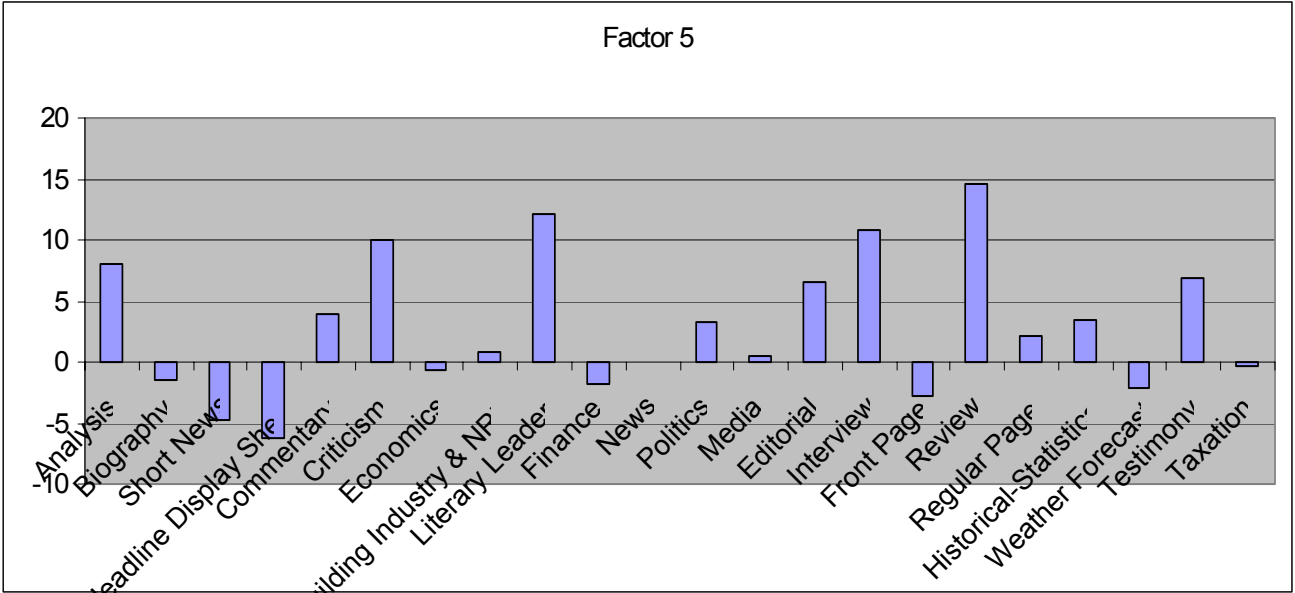


Figure 9. Factor 5 – Highly integrated speech.

Appendix B

The highlighted words show the features relevant to each dimension. Not all the occurrences of the features have been highlighted, in order to avoid a messy representation of the characterisation of the articles.

Dimension 1. Nome testo: %B318SICU85I

Noi olandesi **siamo** un popolo freddo. Perciò non **mi aspetto** tanta folla attorno al papa". Chi parla così è il neo cardinale Adriano Simonis, arcivescovo di Utrecht e Primate dei Paesi Bassi. Sta rispondendo in una saletta alle domande di alcuni giornalisti ."**Siamo** un popolo strano", dichiara Simonis, "un popolo non spontaneo. **Siamo** per natura contro il trionfalismo e contro tutte le manifestazioni troppo calorose. Il **nostro** non è un paese del sole, ma il paese delle brume. Anche per questo, per la visita del papa, non troverete manifesti per le **nostre** strade, non troverete imbandierate le **nostre** città." Qualcuno, però, si scaldato, facendo minacce al papa. ("La polizia **mi** dice che non si tratta di cose serie. È gente che ha qualche difficoltà ad accettare la visita del pontefice che viene da Roma, ed esprime così le proprie difficoltà". La Chiesa olandese non è ancora in pace. Perché? "Esiste una divisione per l'interpretazione dei documenti del Concilio Vaticano Secondo e anche per l'interpretazione della Sacra Scrittura. **Noi siamo** un popolo teologico. Che ama le discussioni dottrinali. Si sa che cosa dice il proverbio: 'Un olandese è un teologo, due olandesi fanno una Chiesa, tre olandesi fanno uno scisma'. Non voglio dire però che qui, da **noi**, ci sia ora uno scisma, ma una divisione seria c'è veramente. Si può notare, tuttavia, che, dopo il Sinodo olandese riunito a Roma nel 1980, la comunione tra i vescovi è aumentata, anche se abbiamo ognuno il proprio carattere e tutti insieme molte difficoltà da affrontare. **Spero** che la visita del papa migliori la situazione. **Io** ho fiducia che, dopo la sua presenza nel **nostro** paese e nella **nostra** Chiesa, **noi** tutti **potremo** intenderci meglio in ogni cosa, è per questo che, domani, aspetto con grande fiducia il papa Giovanni Paolo Secondo".

First Person Verbs

First Person Pronouns

Possessive Determiners

Possessive Pronouns

***Lots of "Indicative - Present" forms (not highlighted)

Dimension 2. Nome testo: %A001SIPO85I

L'ITALIA **deve** ora occuparsi dell'Europa, invece d'**andar** per sabbia in Medio Oriente. Presidente di turno della Cee, avrà la responsabilità di **stimolare** e **guidare** una Comunità che gira a vuoto e somiglia a Penelope, disfacendo di notte ciò che ha tessuto di giorno. Prendiamo l'esempio dell'Unione europea. Il primo rapporto del Comitato speciale dice chiaramente che la faremo solo se la Cee diventa una palazzina a due piani. Primo piano, le nazioni che vogliono l'unione, secondo, quelli che non la vogliono.

Se tutto va bene, avremo due Cee. Se ne deduce che l'Italia, mentre **dovrà gestire** l'allargamento della Cee alla Spagna e al Portogallo, sarà poi costretta a **restringere** l'Europa dei dodici, **facendola tornare** a sei, sempre che **voglia compiere** il passo avanti politico.

Ma la stessa cosa vale per l'integrazione economica, visto che Papandreu s'è fatto portavoce a Dublino dell'"Europa dei poveri", rivendicando pretese e aiuti in cambio di consenso. La prospettiva è di **avere** anche qui due Comunità: da una parte Francia, Germania, Olanda, Belgio, Lussemburgo; dall'altra parte i ("poveri", greci, spagnoli, portoghesi, gli irlandesi, che rafforzano il polo dissenziente formato da Gran Bretagna e Danimarca. L'Italia, che spesso fa la spola fra i due schieramenti, **potrà decidere** di **rafforzare** il primo. Ma saremo sempre dodici che diventano sei contro sei. CRAXI si promette dal semestre di presidenza molti successi, magari suggellati da un trionfale "vertice di Milano", da **tenere** a ridosso delle amministrative italiane, un occhio ai voti meneghini, un occhio alla gloria europeista. C'è da **sperare** che questo desiderio di successo personale **giovi** all'Europa, ma c'è pure da **dubitare** che il "pacchetto di Bruxelles" **contenga** il solito finale.

[...]

Causative verbs
Infinitives
Modal Verbs
Subjunctive

Dimension 3. Nome testo: %CN00/04/88 INFORMATI COSI'

[...]

Si dice spesso: "quando non **sei** soddisfatta di **te**, non abbatte**ti**, ma **superi** il problema gratificandoti in eccesso." Giusto. Se **vuoi** un make-up su misura per **te**, al Make-Up Studio (Via Madonnina 15, Milano, tel. 02/800403) **ti** insegnano in una sola seduta a rifarti il look. Se il problema sono le unghie, regalati un trattamento speciale. L'operazione semplice: l'unghia naturale viene ricoperta con una ceramica speciale, il materiale si solidifica in pochi istanti e viene limato per dare all'unghia la forma che preferisci. Tempo: un'ora. Costo: 120.000 lire. Dove: da Annette (Milano tel. 02/8399466; Bergamo, tel. 035/260612). Se vuoi "rimodellarci la figura", il sistema più rivoluzionario arriva dalla Svezia. Si chiama Thermo-Trim e procura ottimi risultati. Sulle zone da rimodellare (fianchi, cosce, sedere) vengono applicate placche che, riscaldate, mantengono il calore a 40x, favorendo l'aumento della circolazione sanguigna e del metabolismo dei grassi. Dodici sedute a intervalli di tre giorni possono dare perdita di volume e calo di peso.

PARTI CON UN CHEK-UP

E' l'inizio per una revisione di viso e corpo da cima a fondo. **Prenditi** una giornata (domenica?) o un intero week-end per iniziare il conto alla rovescia pre-estivo. Regola numero uno: **sfodera** tutta la tua capacità di attenzione e spirito d'osservazione. Regola numero due: **sii** imparziale. L'auto-test primaverile potrebbe anche rivelarsi impietoso e mettere in evidenza difetti e imperfezioni. Vietato deprimersi: una volta individuate le cause del problema, la soluzione a portata di mano. In fondo, è primavera. Riorganizzare, trasformare, rimettere in ordine, cambiare le solite abitudini, sono i verbi vincenti per affrontare la nuova stagione in perfetta forma fisica e mentale.

[...]

Imperative

Second Person Verbs

Second Person Pronouns

Dimension 4. Nome testo: %RE15-07-95 ALFRED HITCHCOCK CHE FILIBUSTIERE

New York - Nella famosa intervista che concesse a Francois Truffaut, Alfred Hitchcock sosteneva di aver girato il film Psycho per provare che la cinpresa poteva "farsi beffe della platea", di "suonare" anzi quella platea "come un organo", permettendo al regista di "provocare emozioni di massa". In un'altra occasione ebbe a dire che "la civiltà si è fatta così protettiva che non siamo più capaci di rabbrivire dalla paura". Unica cura: uno choc collettivo, che lui pensò di fornire appunto con Psycho nel 1960. In realtà, queste erano giustificazioni a posteriori. Janet Leigh, l'attrice che è passata alla storia del cinema soprattutto per la scena della sua carneficina nella doccia, rivela in Psycho: Behind the Scenes of the Classic Thriller (Harmony, pagg. 197, dollari 22) che la ragione principale per cui Hitchcock girò il film fu il denaro. Sul finire degli anni cinquanta libri e film del genere "horror" rastrellavano montagne di dollari in libreria e al box office. Hitchcock, molto semplicemente, si allineò alla voga. Janet Leigh, che ha già scritto un libro autobiografico, There Really Was a Hollywood, e che a ottobre esordirà nella narrativa con il romanzo House of Destiny, non è una vera scrittrice. Ma qui, con la collaborazione di Christopher Nickens, autore di quattro biografie di grandi star, rievoca con sufficiente candore e una conoscenza di prima mano le vicende "dietro le scene" di un film diventato un classico nonostante le premesse (un po' come successe a Casablanca).

[...]

Simple preposition
Common nouns
Other prepositions

Dimension 5. Nome testo: %SI00/04/88 UN MILIARDO? NO,

Costanzo Ciano, padre di Galeazzo, ministro delle Comunicazioni dal 1925 al 1935 e in seguito presidente della Camera, secondo la maldicenza **popolare era straordinariamente ricco**. Una lettera firmata giunta da Genova il 4 agosto alla presidenza del Consiglio segnala che Galeazzo Ciano, alla morte del padre avvenuta il 27 giugno 1939, aveva denunciato, per l'eredità lasciatagli da Costanzo, il valore di 100.000 lire. "Eppure", prosegue la lettera, "a tutti **è noto** che il Costanzo decedette nel giorno stesso in cui a Livorno cogli amici aveva celebrato il **raggiunto** miliardo." Tradotto in moneta **corrente**, 1 miliardo di lire del tempo equivarrebbe a circa 900 miliardi di oggi. Il patrimonio Ciano fu oggetto di indagine da parte della Commissione Casati. Per parare il colpo, Galeazzo scrisse una **patetica** lettera a Badoglio in cui elenca puntigliosamente quanto, a sua detta, gli aveva lasciato il padre: 1) Tre quarti della Società **tipografica** del giornale Il Telegrafo di Livorno. 2) Quattro edifici in Roma del valore totale, alla sua morte, di circa 5 milioni. 3) Titoli **industriali** così ripartiti: Romana elettricità, azioni 1400; Terni, azioni 300; Montecatini, azioni 2000; Valdarno, azioni 1000; Metallurgica, azioni 1000; Navigazione **generale**, azioni 300; Ilva, azioni 500; Anic, azioni 1000; Amiata, azioni 700; Imi, azioni 100; Consorzio credito opere pubbliche, azioni 24; Buoni del tesoro, lire 1.000.000; contante, lire 355.089,40@; conto **corrente postale**, lire 32.975. "**Sono sicuro**" concludeva Galeazzo, "che queste cifre, così lontane dalle **astronomiche** fantasie dei calunniatori **anonimi**, saranno dal **sereno** spirito di Vostra Eccellenza valutate non quale **disonorante** bottino di un profittatore, bensì come l'**equo** frutto di una vita **interamente operosa**" Lo stesso Galeazzo, nel memoriale di Verona, valuta l'eredità **paterna complessivamente** nella cifra, **veramente modesta**, di circa 8 milioni, a prescindere dal giornale che, come tutti i giornali, ha un valore molto **aleatorio**.

Adjectives

Predicative phrases

Adverbs

***Lots of "Conjunctions" (not highlighted).