A multidimensional approach to the classification of combining forms

Claudio Iacobini & Alessandro Giuliani

This paper aims to show the usefulness of multidimensional analysis techniques for linguistic classification, and to propose a solution to the much debated categorial status of combining forms providing a classification of such elements based on a detailed analysis of a representative corpus taken from the Italian language. Computational methods for statistical analysis of observed correlations allow for internally compact clusters that are well apart from each other, and also permit the establishment of descriptive variables in consideration of a given data set. This methodology satisfies the requirements of structural stability and flexible adaptability called for in linguistic prototype theory and at the same time it resolves both the problem of an optimal number of classes and predictivity of classification. This analysis gives both a structural description of the studied data set as a whole and a precise allocation of each analyzed item. The analysis, performed on a corpus of 563 Italian combining forms, is based on explicit criteria so that its linguistic interpretation is relatively straightforward. Solid empirical evidence is given to demonstrate that typical combining forms are bound lexemes (stems), consequentially their classification does not require any further linguistic category beyond those of affix and lexeme (the term "combing form" is a convenient descriptive tool for grasping together bound elements used to form morphological complex words not sharing all the characteristics of lexemes or affixes of a particular language). The results support the idea of a more widespread use of cluster analysis in linguistics, both from a methodological and empirical point of view.

1. Introduction

The establishment of category borders in the area of neoclassical compounding is a pertinent area of investigation, since numerous studies on the topic disagree as to the location of the phenomenon within word-formation processes, and especially as to the classification of the units that take part (the so-called combining forms, henceforth **CFs**). Neoclassical compounding share characteristics with native and foreign compounding. CFs are considered by some scholars as a particular instance of lexemes, by others as a particular instance of affixes, or a category distinct from the preceding two, and by others yet as an 'intermediate' category displaying both lexical and affixal features. [1] Moreover, CFs are mostly exogenous elements, predominantly borrowed from Latin and

Italian Journal of Linguistics 22.2 (2010), pp. 287-316 (Received February 2010)

Old Greek (hence the term neoclassical CFs) and employed in technical-scientific terms according to foreign word-formation patterns, which can be more or less integrated into the lexicon of a natural language and productively employed in word formation, but they also can result from recent processes of clipping and secretions (e.g. *eco-* 'ecology'; cf. Bauer 1998, Fradin 2000, Iacobini 2004a).

The unsatisfactory classification of both neoclassical compounding and of CFs mainly depends on the fact that qualitative descriptions must be based on the presence/absence of discrete categories, which by definition do not predict intermediate spaces. Discrete categories oblige us, therefore, to ignore the existence of items that correspond only partially to adopted categories (cf. ten Hacken 2000, on the need of establishing clear borders between lexemes and affixes, compounding and derivation). On the other hand, methods of inductive classification that adopt a corpus-based approach and scalar criteria (for example as in prototype theory) call for a selection of representative corpora and an adequate number of analytical criteria. Furthermore, these methods call for a complex network of interaction that is determined by the interaction of the data taken into consideration and the variables that are adopted to define and interpret them, with the opportune help of analytical methods.

In previous research Iacobini & Giuliani (2001) have demonstrated that the use of multidimensional analysis technique (henceforth **MDA**) allows us to obtain inductive classifications, in which the CFs are grouped in clusters characterized by the greatest internal cohesion and the greatest spacing between clusters. The classification obtained by automatic procedures has been compared with a linguistic classification of the same corpus (362 CFs). The comparison of these two methods highlights a robust agreement between the two classifications (contingency coefficient = 0.83, with a maximum possible of 1).

The objective of this paper is to reconfirm the usefulness of MDA in classifying linguistic elements, when it is necessary to consider a number of variables, and to provide a more reliable classification of CFs based on a higher number of items. With this objective we collected a 'new' set of 201 CFs and compared the results obtained by the 'new' classification with the 'old' set collected in Iacobini & Giuliani (2001), and with that conducted on the 'global set' (the 'old' and 'new' set put together for a total of 563 CFs). The net concordance of these results constitute further validation of the first classification and of the method of analysis. The grouping of clusters that were obtained for the 'global set', and the indications relative to the relevance of the descriptive criteria, constitute an example of classification based on explicit criteria, which we propose as a model for the classification of CFs (as well as for other linguistic items

and categories). Moreover, the demonstration of the consistency between the 'old' and the 'new' mathematical classification permits us to verify the predictivity of the classification in clusters and to provide a tool for automatical classification of further items, which in turn allows us to overcome drawbacks in the use of prototype analysis in linguistics.

This paper is set out as follows: Section 2 illustrates some of the characteristics and the limits of prototypical classification in linguistics; Section 3 briefly introduces MDA; Section 4 lists the criteria of our corpus selection and the linguistic variables used to for our analysis; Section 5 displays the statistical analysis of the 'new set', compares it with the analysis of the 'old set', and with the classification in clusters of the 'global set'; Section 6 presents the analysis from a linguistic point of view with a brief comment; Section 7 gives indications regarding the method of calculating the predictivity of our classification; Section 8 presents the conclusions; Appendix 1 lists all of the CFs analyzed (i.e. the 'global set') divided into clusters and ordered within each cluster from the most central to the most peripheral.

2. Characteristics and limits of inductive classification in linguistics

Inductive methods are used in all the branches of contemporary linguistics. The works of Lakoff (1987) and Taylor (1989) have had an important role in the diffusion of prototype theory (a theory developed by Eleanor Rosch and her school in the 1970s in studies of categorization in cognitive psychology). According to this approach categories are determined by the interaction between the data and the relevant attributes used to define them.

Notions of prototipicality and scalarity appear in studies devoted to highly varied phenomena, and are largely employed in linguistics at phonological, morphological, lexical, semantic, and syntactic levels. For example, consider their use in typology for the identification of fundamental categories such as 'parts of speech'. [2]

Prototype-based classifications are aimed at satisfying the requirements of both structural stability and of flexible adaptability; the objective is to classify the objects that do not totally correspond to the category determining features, as well as those that do. This goal may be reached, in that categories have an internal structure, and they are defined by a group of features that are not necessarily shared by all of the objects, or not shared to the same degree by each object that is a member of the category. Moreover, there is a scalar continuum, not only internal to each category, but also between the different categories. There are more

central members, those that respect the combination of the typical attributes of a category, and more peripheral members within each category. Failure of an object to exhibit some of these attributes does not of itself preclude its membership.

The main strength of prototype theory is that it permits us to classify items that do not fit into the 'classical categories', without extending the boundaries of the categories. This is also its main weakness. The possibility of explaining a number of almost infinite deviations from the norm (and simultaneously explain those that do not fulfill the criteria that permit a distinction in the number of categories utilized) renders prototype theory very weak from the point of view of predictivity.

For example, the classification that collocates compound-words along an axis that is defined by only two qualitative attributes (native vs. foreign compound), as illustrated in (1), though it at first glance seems quite convincing, does not have any explicative potential. The lack of this potential is because the collocation of neoclassical compounds in intermediate positions is based on impressionistic evaluations that are made with non-quantifiable values. This type of classification tells us nothing about the characteristics of words like *hydro-slide* or *geology*, nor about the degree of similarity that there may be between them, or with the prototypical examples of an English native compound (*house-boat*), or a foreign compound (*Zeitgeist*), nor does it foresee limits to the existence of other possible intermediate categories.

(1) *house-boat* *hydro-slide* *geology* *Zeitgeist*
NATIVE COMPOUND ‹———› FOREIGN COMPOUND
NEOCLASSICAL COMPOUNDS

3. Multidimensional analysis for prototype classification in linguistics

Prototype-based classifications in linguistics are usually represented in a bidimensional space which depicts a graphic simplification of a more complex reality. Such bidimensional spaces, at odds with the effectively measurable spaces arising from MDA, are nothing more than hypothetical constructs, in which the items are subjectively chosen by the scholar.

On the other hand, MDA produces new variables, called principal components, which automatically arise through the computation of the observed correlations established over a set of linguistic features in the consideration of a given data set. The metric definition of clusters (classes) in such spaces, where the CFs are allocated on the basis of a mathematical criterion (the clusters correspond to the generation of

classes endowed by a minimum within-cluster variance and a maximal between-clusters variance, i.e. clusters are the most internally compact and the most distantly spaced from each other), allows to solve the above mentioned fuzzy membership problem. [3]

Moreover, MDA allows us to evaluate the degree of importance (prototypical character degree, correspondent to the proximity of the item to its cluster centre) of the objects – the ones that need to be described – within the category relative to the criteria adopted. The generation of clusters within the constraints of being the most internally compact, and the most distantly spaced from each other as possible are obtained through the similarity of values derived by the interaction between the attributes selected for the analysis, and those of the items analyzed. This method furnishes both a structural description of the group of data and the precise allocation of each item analyzed. It also makes it possible to measure the degree of internal homogeneity of each cluster and to find the attributes that best distinguish one cluster from the other. The MDA results may be usefully compared to other classification proposals, and allow us to abandon solutions based on quantitatively uncontrollable assumptions.

MDA is a useful instrument to cope with the continuous character of empirical data and with the necessity to explicitly define and articulate the categories used in scientific description. MDA, furthermore, permits us to validate the classification by testing it with other objects that are not present in the sample of items examined, which in turn permit us to overcome the problem of predictivity (cf. Section 7), a dilemma for prototype classification.

MDA generates continuous metrics, allowing us to quantitatively estimate the relative similarity between items. In order to obtain a satisfactory description it is necessary to arm oneself with a sufficiently detailed grid that allows us precisely to assign a pattern of discrete values to any item that is analyzed. The convenience of a classification that takes into consideration various attributes simultaneously is highlighted by the fact that the number of possible distinct configurations exponentially increases with the number of variables by the power of 2 (usually, possible patterns with n qualitative categories are equal to 2^n, therefore, with two categories we have 4 (2^2) possible configurations, with three 8 (2^3) etc.). Thus with a given number of not overly elevated categories – such as the 21 that we used in our analysis – we arrive at a distribution that can be considered as continuous, having 2^{21} = 2,097,152 distinct possible configurations. This allows us to generate a fully quantitative metrics by means of the simple multiplicity of the analyzed features, without any unjustified quantification assumption. This continuous, highly dimensional, metric space can be analyzed with classical MDA methods:

Principal Component Analysis or Cluster Analysis. This analysis enables us to find relevant aggregations of CFs (cf. Section 5).

When applying an MDA-based approach, the linguist must still respond to the tasks required in inductive classification: pinpoint and select the descriptive characteristics, describe the data examined in relation to the various analytical criteria, evaluate and interpret the results of statistical analysis. But at the same time the linguist has both the advantage of obtaining a quantitative appreciation of the structural value of the systematization in terms of the statistical properties of the classification, and the possibility to insert new elements into the proposed classification by means of objective and verifiable criteria.

4. Linguistic criteria used for the classification of CFs

As mentioned in the introduction, we chose the CFs as a case study for the application of MDA, because their classification is difficult and still debated. CFs display distributional, semantic, positional, compositional characteristics, which, from a categorial point of view, fluctuate between the extremes of a lexeme on one hand and an affix on the other hand, and from the point of view of word-formation processes, fluctuate between compounding and derivation (CFs are also involved in other word-formation processes such as blending, clipping and secretion).

There is no agreement in literature on whether CFs form a distinct category or whether they must be subsumed under the category of lexeme or affix (or in part one or in part another), nor whether it is opportune to make distinctions (and if so, how many) internal to the CFs. [4]

In Iacobini (2004a) we provide a detailed linguistic classification of Italian CFs and their ways of integration in words used in everyday speech. Our general conclusion is that CFs may be a convenient (even if often misleading) label in lexicographic practice, but they are not a theoretical notion neither from a morphological nor from a lexicological point of view. In line with the recent contribution of Kastovsky (2009), CFs are to be classified as stems, [5] and the analysis of their use and of their integration in native word-formation processes is to be considered along a scale of progressively less independent constituents ranging from words via curtailed words, stems, constituents of blends to affixes. The similarity between a restricted and easily definable number of CFs and derivational affixes (which has often been overstated) is the result of processes of grammaticalization which are very common in compound members. The great majority of CFs are stems displaying heterogeneous lexical meanings and are employed as members of compounds in technical-scien-

tific terms. Others CFs result from clipping and secretion (generally from compound words of current usage). Only for the restricted number of CFs on their way to becoming derivational affixes (identified by some scholars with the term affixoid) the border between lexical and affixal status may be rather arbitrarily defined depending on the weight given to the different structural and usage characteristics taken into account, and it may differ from language to language.

In this work, in line with the statistical methodology adopted, our initial linguistic task was the collection of linguistic variables employed in literature for the description of CFs, irrespective of any theoretical commitment. The 21 variables that we used are a result of a selection of all the most relevant criteria commonly accepted and utilized in the classification of CFs (cf. Iacobini 1999, Prćić 2008: 4-13). Table 1 presents the list of the variables and the values that can be assumed for each variable. A majority of the variables are dichotomous, thus assuming only two possible values: 1 for presence and 0 for absence; variables j) 'endo/exo-centricity', k) 'relation among constituents', m) 'combinability', t) 'register' may also have intermediate values; the values employed for variable u) correspond to the number of syllables of each CF – with a range from 1 to 4.

It is important to underline that the presence of redundant information does not hinder the recognition of clusters, but is considered a useful resource, in that MDA is based on the correlation of the values of the adopted variables.

The variables that we have taken into consideration regard both CFs and the complex-words that are involved in their formation. The variables regarding the CFs are the following: variable a) concerns the possibility of using CFs as independent elements as well as bound elements (e.g. *auto-*[2] with the meaning 'automobile, car', *foto-*[2] with the meaning 'photography'); variables b), c), h) concern the position of CFs in complex word forms; variables d), e), f) concern the etymon of CFs; variable g) makes evident CFs that derive from a shortening of the original lexeme (e.g. *porno-* from *pornografia* 'pornography', *socio-* from *sociologia* 'sociology'), some of these CFs have a number in superscript that permits us to distinguish homographic CFs (e.g. *bio-*[2] 'biology' in words like *bioarchitettura* 'bioarchitecture', cf. *bio-*[1] 'life, course or way of living' in words like *biodegradabile* 'biodegradable', *biografia* 'biography'); variables n), o), p), q), r), s) regard the part(s) of speech from which the CFs derive; variable i) concerns CFs that can be the base of a derivational process, i.e. are used as independent lexemes with the addition of a derivational affix (e.g. *dermato-* > *dermatico* 'dermatic', *thanato-* > *tanatosi* 'thanatosis'); l) distinguishes between denotative-lexical meaning vs. relational meanings; the values of variable u) correspond to the number of syllables of each CF. [6]

Table 1. Linguistic variables employed for the classification of CFs

Linguistic variables	*Values*
a) autonomy	always bound CF = 0; also free CF = 1
b) final position	yes = 1; no = 0
c) non-final position	yes = 1; no = 0
d) etymon: Greek	yes = 1; no = 0
e) etymon: Latin	yes = 1; no = 0
f) etymon: Modern language	yes = 1; no = 0
g) shortening	yes = 1; no = 0
h) position	fixed position = 0; more than one position = 1
i) derivability	yes = 1; no = 0
j) endo/exo-centricity	endocentric = 0; exocentric = 1; endo/ exocentric = 0.25
k) relation between constituents	coordinative = 1; subordinative = 0; coord./ subord. = 0.5
l) meaning	relational = 0; denotative-lexical = 1
m) combinability	CF = 0; CF/W = 0.25; CF-W = 0.5; W/CF = 0.75; W = 1
n) part of speech: Noun	yes = 1; no = 0
o) part of speech: Adjective	yes = 1; no = 0
p) part of speech: Adverb	yes = 1; no = 0
q) part of speech: Preposition	yes = 1; no = 0
r) part of speech: Numeral	yes = 1; no = 0
s) part of speech: Verb	yes = 1; no = 0
t) register	CL (current language) = 1; CL/TS = 0.75; TS/ LC = 0.25; TS (technical-scientific terms) = 0
u) number of syllables	1; 2; 3; 4

The variables concerning words formed by employing CFs are not formed with dichotomous values, rather with more than two values, so as to be able to register the proportions of these values in our corpus. In variable j), value 0 corresponds to (almost exclusively) endocentric words, value 0.25 to prevailing endocentric words, while value 1 to (almost) exocentric words. The coordinative relation between constituents in the complex word is expressed by value 1 in variable k), while value 0 corresponds to a subordinative relation, value 0.5 displays an almost equal proportion of the two kinds of structures; in variable m) value 0 indicates CFs that form words (almost) exclusively with other CFs, value 1 CFs that (almost) exclusively combine with independent words, were given intermediate values (0.25, 0.75, 0.5) indicating a prevalent proportion of combination ability with words, CFs, or both respectively.

In Iacobini & Giuliani (2001) the above mentioned criteria were utilized to describe 362 CFs. This corpus was put together through a manual selection of CFs present in the following dictionaries: Battaglia (1961-2002), Alinei (1962), Cortelazzo & Zolli (1979-1988), Ratti *et al.* (1988), Cortelazzo & Cardinale (1989), Quarantotto (1987), Devoto & Oli (1990), Forconi (1990), Lurati (1990), Garzanti (1993), DISC (1997), Zingarelli (1999). The reference corpus for complex words formed with CFs corresponds to the words indexed in DISC (1997), and in Zingarelli (1999). Publication of GRADIT made a much larger corpus of CFs available. GRADIT lemmatizes 2635 CFS (including all those in the 'old' set). We estimated that the selection of approximately a further 200 CFs should be sufficient to extend the sample, validate the method of analysis, and furnish a more stable classification of CFs. We therefore selected 1 CF in every 13 presented in GRADIT, passing on to the following CF in the case that the CF selected was already present in the 'old set'. In this way we obtained a 'new' set consisting of 201 CFs.

As in our previous study, we followed the categorial indications of the dictionaries. We then selected the 'new' items among those indicated in GRADIT with the term *confisso* 'confix'.

The 'new set' taken from GRADIT presents two characteristics that differentiate them in part from the 'old set'. 1) The CFs of the 'new set' are characterized by being mostly used in the formation of technical-scientific terms. This depends on the fact that the word list of GRADIT contains a high number of CFs and of technical-scientific terms, and on the fact that the 'old set' already contained all of the 'everyday' CFs (i.e. the ones present in complex-words of common usage, like *ecoturismo* 'ecotourism', *bioritmo* 'biorhytm', *psicosomatico* 'psychosomatic'; 2) GRADIT also classifies as *confisso* 'confix' items such as: *apri-* 'open'*, carica-* 'charge'*, pela-* 'peel'*, rompi-* 'break'*,* namely verbal themes used in the formation of everyday verb-noun compounds, such as *apriscatole* 'can / tin-opener, *caricabatteria* 'battery charger', *pelapatate* 'potato-peeler' *rompighiaccio* 'ice-breaker'. The characteristics of these items are markedly different from those of CFs used in neoclassical compounds, since they are unanimously classified as lexical elements employed as first member of native compounds (cf. Namer & Villoing 2007, Gaeta & Ricca 2009). However, we have not excluded the verbal elements of verb-noun compounds from our corpus, both because we wanted a randomly selected sample (with the only correction that we have indicated above) and because we wanted to observe the consequences on the clusters obtained due to the presence of a group of items with characteristics that were not present in the previous sample.

5. Statistical analysis

The CFs were represented in a 19 dimensional space derived by the 21 variables described in Table 1 with the exclusion of 'autonomy' and 'relation between constituents' features whose practically null variance prevented any meaningful analysis. Each CF is thus a 19-dimension vector in a metric space, whose metrics (pairwise CFs distance) is the Euclidean distance computed as the square root of the squared differences between the two corresponding vectors computed variable by variable. This allows for the straightforward computing of both the between-variable correlation (the percentage of similar values in corresponding positions – i.e. CFs – for the two variables as measured by the Pearson correlation coefficient), giving rise to principal component spaces spanned by the eigenvectors of the correlation matrix, and the consequent clustering (K-means procedure, a cluster is made by CFs the most similar among them and the most distantly spaced from the CFs of other clusters), with the formation of bottom-up natural classes of CFs.

Iacobini & Giuliani (2001), starting from a corpus of 362 CFs, obtained a six component / seven class clustering. In this work we focused on another set of CFs (consisting of 201 different items). The aim was to check the agreement between the results obtained with the 'old' and the 'new' set. The reaching of substantial consistency between the two sets allows us to apply firm quantitative foundations to the classification of CFs.

This consistency was evident from the significant correlation ($r = 0.64$, $p < 0.0001$) holding between the values of the 171 (19x18/2) and pairwise distinct Pearson correlations between the 19 variables describing the 'old' and 'new' sets (Fig. 1).

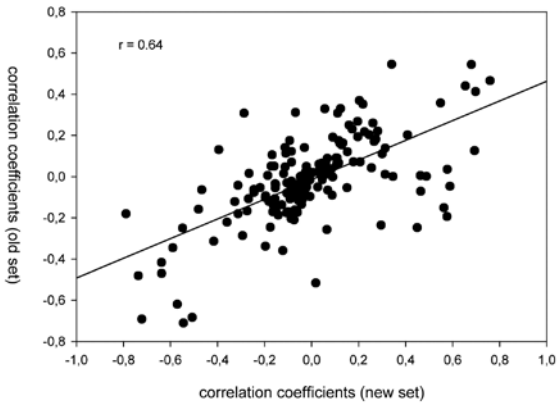

Figure 1. Between-variables correlation coefficients in the two data sets.

Figure 1 illustrates that in the two data sets, the 19 descriptors share very similar mutual relations between each other, which points to a strict structural resemblance between the two sets.

The correlation between the 'old' and 'new' sets is still more cogent, if we filter the two data sets through the agency of principal component analysis (henceforth **PCA**). [7] This technique can be considered a filter for correlated information in which the original space is decomposed into a 'correlated portion' that is retained by the first components and corresponds to the information content shared by different linguistic variables. The 'singular portion' that constitutes minor components (also called floor noise) collects the singularities of each feature and any spurious aspect of the data set.

It turns out that of fact both the 'old' and 'global' (old+new) sets, when submitted to PCA give rise to a six component solutions, well above the floor noise (cf. Broomhead & King 1986).

The most important proof of consistency, between the 'old' and 'global' sets in the component space, is the near to maximal correlation (Pearson r) between correspondent component scores of the 'global' and 'old' data sets, cf. (2).

(2) R(pc1old-pc1global) = 0.93, R(pc2old-pc2global) = -0.94, R(pc3old-pc3global) = 0.93, R(pc4old-pc4global) = 0.95, R(pc5old-pc5global) = 0.76

Figure 2 reports the relation between the two major components, the points correspond to the 'old' set words as described by the 'old' and 'global' components, according to the correlation coefficients sketched above, the concordance of the two representations is impressive. We again stress the fact the components arise in a purely automatic way, they correspond to the directions in a multidimensional space that maximize the variance between the objects studied (CFs).

Having stated the general concordance between the two spaces (i.e. 'old' and 'global' set) in terms of component meaning, we can go into greater depth and check if the clustering of the CFs remains invariant, moving from the 'old' to the 'global' data sets. The consistency between the automatic classification executed on the 'old' and on the 'global' space, point to the existence of natural classes of CFs that are independent of our specific ensemble.

With this objective in mind we clusterize the 'old' data set CFs by a k-means algorithm, thereby obtaining a 7-cluster solution, which explains globally 61% of variance, with a Pseudo F statistics of 93.64, neatly different from that expected by chance ($p < 0.0001$). The 'global' (old+new) data set was in turn analyzed by means of cluster analysis on

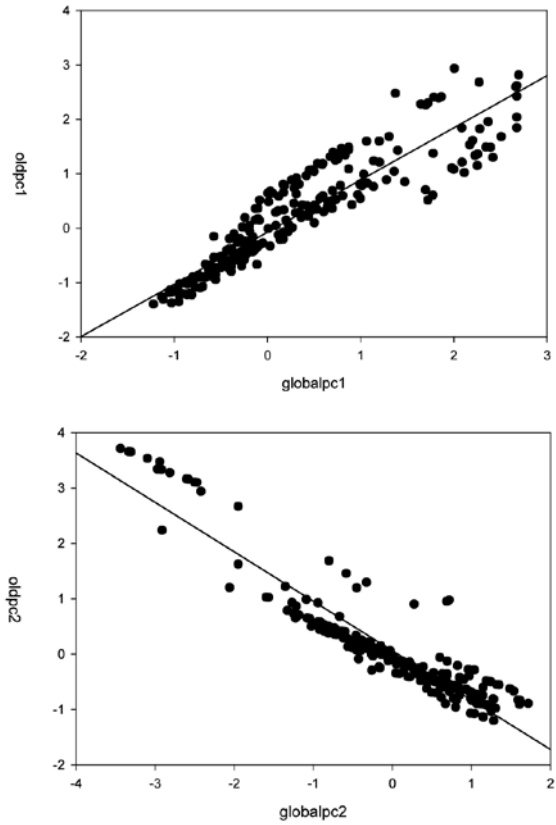

Figure 2. Component scores plane for the 'Old' and 'Global' sets: the points corresponds to CFs in common between the two sets.

the component space, obtaining again an optimal 7-cluster solution with similar values of explained variability (OVERALL R-square = 0.64,in the 'old' data set was R-square = 0.63). This is strong proof of the consistency of the two set classification structure, especially if we consider that the clustering is computed over a 563 data set, which is very different from the original 362 set.

Beside the general statistical resemblance of the two classification schemes, what is probably more interesting to us is the degree of superposition, between the classes based on the components extracted by the 'old' data set ('old' cluster) considered as such, and the classes (final clusters) extracted by the algorithm as applied on the 'global' data set. In order to do so a chi-square procedure is applied, in which the significance between the allocation of a CF in the 'old' classification and the allocation of the same CF in the 'new' scheme is computed. This significance is extremely high reaching a chi-square statistic near 900, and $p < 0.0001$. This implies that the classes are extremely stable, see Table 2, together with the chi-square statistics, the two-way contingency table linking the 'old' and 'global' classification of the 'old' data set words. The classes of the two partitions with the highest grade of superposition are in bold. Thus, cluster 1 of both the partitions are exactly the same, all the 11 words in cluster 1 of the 'old' partition are in cluster 1 of the 'global' partition, the same for cluster 2, while cluster 3 of the 'old' partition is mainly superimposed with cluster 5 of the 'global' partition, though the superposition is not perfect.

It is interesting to note how 'old' cluster 6 (an average cluster) is split by cluster 5 and 6, two newly formed classes, this implies that the addition of 'new' words gave an increased level of definition to the system, even if, as we have seen, the basic skeleton of the distribution of words into classes remained the same. This increased detail comes from the differences in composition (kind of CFs) present in the two ('new' and 'old') data sets, these statistically significant ($p < 0.001$) differences in composition make the high grade of superposition in both component meaning and cluster structure of the two sets even more remarkable.

In Table 3 the pertaining to 'old' and 'new' data sets is compared with the pertinence of the CFs to different classes. As evident from the table, cluster 4 is formed only by 'new' words, while cluster 3 is almost specific of the 'old'. These differences in composition will be commented on in the discussion, together with a linguistic interpretation of the extracted components and clusters. The allocation of CFs in clusters are reported in Appendix 1, while the method for allocating the 'new' CFs into classes by computing the correspondent principal components is reported in Section 7.

Table 2. Superposition between the 'Old set' and the 'Global set'.

old clusters		*FINAL CLUSTERS*							*Total*
		1	2	3	4	5	6	7	
1	Freq.	*11*	0	0	0	0	0	0	11
	Percent.	3.04	0.00	0.00	0.00	0.00	0.00	0.00	3.04
	Row Pct	100.00	0.00	0.00	0.00	0.00	0.00	0.00	
	Col Pct	22.45	0.00	0.00		0.00	0.00	0.00	
2	Freq.	0	*22*	0	0	0	0	0	22
	Percent.	0.00	6.08	0.00	0.00	0.00	0.00	0.00	6.08
	Row Pct	0.00	100.00	0.00	0.00	0.00	0.00	0.00	
	Col Pct	0.00	95.65	0.00		0.00	0.00	0.00	
3	Freq.	6	0	0	0	*62*	0	22	90
	Percent.	1.66	0.00	0.00	0.00	17.13	0.00	6.08	24.86
	Row Pct	6.67	0.00	0.00	0.00	68.89	0.00	24.44	
	Col Pct	12.24	0.00	0.00		38.04	0.00	100.00	
4	Freq.	*2*	0	0	0	0	0	0	2
	Percent.	0.55	0.00	0.00	0.00	0.00	0.00	0.00	0.55
	Row Pct	100.00	0.00	0.00	0.00	0.00	0.00	0.00	
	Col Pct	4.08	0.00	0.00		0.00	0.00	0.00	
5	Freq.	1	0	*39*	0	0	4	0	44
	Percent.	0.28	0.00	10.77	0.00	0.00	1.10	0.00	12.15
	Row Pct	2.27	0.00	88.64	0.00	0.00	9.09	0.00	
	Col Pct	2.04	0.00	97.50		0.00	6.15	0.00	
6	Freq.	29	1	0	0	53	61	0	144
	Percent.	8.01	0.28	0.00	0.00	14.64	16.85	0.00	39.78
	Row Pct	20.14	0.69	0.00	0.00	36.81	42.36	0.00	
	Col Pct	59.18	4.35	0.00		32.52	93.85	0.00	
7	Freq.	0	0	1	0	*48*	0	0	49
	Percent.	0.00	0.00	0.28	0.00	13.26	0.00	0.00	13.54
	Row Pct	0.00	0.00	2.04	0.00	97.16	0.00	0.00	
	Col Pct	0.00	0.00	2.50		29.45	0.00	0.00	
Total		49	23	40	0	163	65	22	362
		13.54	6.35	11.05	0.00	45.03	17.96	6.08	100.00

Statistic	*DF*	*Value*	*Prob*
Chi-Square	30	916.7350	<.0001
Likelihood Ratio Chi-Square	30	605.3269	<.0001
Mantel-Haenszel Chi-Square	1	17.9701	<.0001
Phi Coefficient		1.5914	
Contingency Coefficient		0.8467	
Cramer's V		0.7117	

Table 3. Allocation of 'Old' and 'New' CFs in the 'Global set' cluster classification.

		FINAL CLUSTERS							
		1	*2*	*3*	*4*	*5*	*6*	*7*	*Total*
New	Freq.	11	9	3	22	93	49	15	201
Set	Percent.	1.95	1.60	0.53	3.91	16.34	8.70	2.66	35.70
	Row Pct	5.47	4.48	1.49	10.95	45.77	24.38	7.46	
	Col Pct	18.33	28.13	6.98	100.00	36.08	42.98	40.54	
Old	Freq.	49	23	40	0	163	65	22	362
Set	Percent.	8.70	4.09	7.10	0.00	28.95	11.55	3.91	64.30
	Row Pct	13.54	6.35	11.05	0.00	45.03	17.96	6.08	
	Col Pct	81.67	71.88	93.02	0.00	63.92	57.02	59.46	
Total	Freq.	60	32	43	22	255	114	37	563
	Percent.	10.66	5.68	7.64	3.91	45.29	20.25	6.57	100.00

Statistic	*DF*	*Value*	*Prob*
Chi-Square	6	*66.7884*	*<.0001*
Likelihood Ratio Chi-Square	6	*77.6213*	*<.0001*
Mantel-Haenszel Chi-Square	1	13.9621	0.0002
Phi Coefficient		0.3444	
Contingency Coefficient		0.3257	
Cramer's V		0.3444	

6. Linguistic analysis and interpretation of the statistical results

The results of our analysis in principal components and the successive grouping clusters has demonstrated that there is a strong consistency between the classification developed by Iacobini & Giuliani (2001) and the one obtained for the corpus formed with the addition of another 201 CFs. We will now proceed with an illustration of the results reached in clustering of the 'global set' from the linguistic point of view. The linguistic description of the clusters (Section 6.3) will be proceeded by an illustration of the linguistic values significantly present in the principal components (Section 6.1.) and in the clusters (Section 6.2.).

6.1. Constitution of principal components

In Table 4 the matrix of the component loadings of the 'global set' is shown. The component loadings are the correlation coefficients between original variables and components. The loadings go from -1 (maximal negative correlation) to +1 (maximal positive correlation), while near zero values of the loadings point to the mutual independence between the component and the correspondent variable.

The analysis in principal components is able to obtain a reduction in the number of dimensions given the individuation of the correlation values (calculated for each item) between the original variables. In our case the analysis in 6 principal components is the representation that furnishes an optimal balance between compression of information and detail. The six components explicate approximately 80% of the total information. These are disposed in order of the variance that has been explained, so the first principal component is that which explains the largest portion of the original information, followed by the second, by the third, and so on. It is interesting to note that the first principal component explains approximately 26% of the information, and the first three principal components together explain more than 50% of the total data.

Table 4. Component loadings matrix of the 'Global set'.

Variable	*Global PC1*	*Global PC2*	*Global PC3*	*Global PC4*	*Global PC5*	*Global PC6*
final position	-0.50	-0.56	*0.57*	-0.02	-0.07	-0.24
non-final position	0.28	*0.67*	-0.22	*0.54*	0.01	-0.11
etymon: Greek	*-0.81*	0.01	-0.23	0.15	*-0.42*	0.11
etymon: Latin	0.30	-0.26	-0.15	-0.03	*0.85*	-0.27
etymon: Mod. lang.	*0.75*	0.22	0.41	-0.15	-0.23	0.10
shortening	*0.57*	0.27	0.35	-0.27	-0.33	-0.30
position	-0.35	-0.01	*0.50*	*0.55*	-0.08	*-0.42*
derivability	0.04	0.31	0.29	*0.56*	0.26	*0.44*
endo/exo-centricity	0.23	*-0.67*	0.19	0.36	-0.04	*0.40*
Meaning	-0.29	0.20	*0.67*	-0.10	0.23	0.29
Combinability	*0.81*	-0.06	0.16	0.10	-0.03	0.01
part of speech: Noun	-0.49	0.46	0.40	-0.23	0.28	-0.16
Register	*0.74*	-0.29	0.12	0.01	-0.09	-0.03
number of syllables	-0.17	0.13	0.11	*-0.53*	0.06	0.37
% Variance expl.	26	13	13	11	9	7

The linguistic variable values that characterize each of the principal components (PCs) are listed below in order of importance within each of the components. PC1: non-Greek etymon, strong tendency to combine with words from modern languages, tendency to be used in current language, to be a shortening, to occupy initial position; PC2: tendency to form endocentric complex words, to occupy non-final position; PC3: lexical meaning, final position, non-fixed position; PC4: derivability, non-fixed position, non-final position, low number of syllables; PC5: Latin etymon, non-Greek etymon; PC6: derivability, fixed position, tendency to form endocentric complex words.

6.2. Main linguistic attributes of clusters

The technique of cluster analysis aims at identifying similar items. Consequentially the items are ordered according to the criterion of minimal distance (or minimal internal variance), and the clusters are internally as compact as possible and the most distantly spaced between them (according to the criterion of maximum external variance).

The individuation of clusters is indicative of the possibility of item aggregation, and demonstrates the pertinence of variables that have been selected to construct efficient classification. The presence of strong aggregation of the items enables us to individuate the items with the name of the correspondent cluster, hence, in the case of the CFs we are examining the possibility of using seven 'pieces of information' instead of the corresponding number of the total CFs.

The distinction of seven clusters represents an optimal balance between detail and parsimony of classification: while a 6-cluster solution is too fuzzy (too high within-cluster variance), an 8-cluster solution generates classes that are not explicitly discriminated (too low between-cluster variance).

The composition of clusters (listed in Appendix 1) is ordered in terms of growing distance from the center, so the first CFs of each cluster are members that best represent it. Table 5 illustrates the cluster profile, specifically the number of CFs of each cluster (column n.) and the average of the principal components internal to the cluster.

Table 5. Number of CFs and position of cluster centroids in the component space (Global set).

Cluster	*n.*	*Global PC1*	*Global PC2*	*Global PC3*	*Global PC4*	*Global PC5*	*Global PC6*
1	60	0.71	-0.59	*-1.98*	0.39	-0.62	-0.76
2	32	0.13	*-2.79*	0.44	0.00	0.09	0.80
3	43	*2.00*	*1.05*	*1.22*	*-1.01*	*-1.00*	-0.91
4	22	*2.24*	-0.73	*1.04*	*1.31*	-0.16	*2.47*
5	255	-0.51	0.48	0.06	0.58	-0.04	-0.01
6	114	-0.66	-0.23	0.19	*-1.31*	-0.11	0.29
7	37	0.61	-0.00	-0.23	-0.21	*2.86*	-0.71

It is important to underline that the interpretation of the table components have a zero mean and unit standard deviation on all of the corpus, thus values greater than 1, or lower than -1, point to relevant components for the cluster mean interpretation, these components are indicated in bold in the Table. The linguistic attributes that characterize each cluster are those presented in the principal components and are presented with the highest values. To facilitate linguistic interpretation

following we list the main linguistic attributes for each cluster starting from the most relevant.

Cluster 1: relational meaning, initial position, fixed position, non-derivability, tendency to form endocentric complex words; tendency to combine with words, tendency to be used outside the technical-scientific register.

Cluster 2: final position; fixed position, derivability, tendency to form exocentric compounds.

Cluster 3: non-Greek etymon, strong tendency to combine with words from modern languages, tendency to be used in current language, to be a shortening, to occupy initial position; lexical meaning, tendency to form endocentric compounds, non derivability, fixed position.

Cluster 4: derivability, fixed position, tendency to form endocentric complex words; non-Greek etymon, strong tendency to combine with words, modern language etymons, tendency to be used in current language, to be a shortening, initial position; low number of syllables; lexical meaning.

Cluster 5: average values for every component (very slight preference for the following features: derivability, non-fixed position, non-modern language etymon, use in technical-scientific register, not be a shortening).

Cluster 6: very similar to C5 with a tendency towards derivability, final position, higher number of syllables.

Cluster 7: very similar to C5 with a preference for Latin etymon, initial position.

6.3. Linguistic description of clusters

Let us move on now to examine the linguistic characteristics of the CFs and the positions they occupy in clusters with the end of evaluating from a linguistic standpoint the reliability of the classification used with the automatic procedures of statistical analyses.

Cluster 1 includes among its most central members items which in many classifications are considered prefixes (cfr. Iacobini 2004b). Among the first ten members there are: *iper-, neo-, pseudo-, micro-, para-, poli-, mega-, macro-*. This is a cluster with well-identified attributes. The members of this cluster are all characterized by a relational meaning, the initial position (except in one case), and non-derivability. They are the best candidates to acquire a prefixal status if they are preposed to independent bases in formations of current usage. Two semantically homogeneous groups are formed by elements expressing quantity (e.g. *pluri-, multi, omni-*), and by numerals (e.g. *milli-, centi-, uni-, duo-*). The

percentage of combinability with words and the use in complex-words of everyday lexicon (e.g. *iperattivo* 'hyperactive', *neocolonialismo* 'neocolonialism', *pseudoscienza* 'pseudo-science', *microcomputer* 'microcomputer', *paramilitare* 'paramilitary') significantly decreases with the less central members (e.g. *oligo-, omeo-, enantio-, megalo-*), which tend to share characteristics with the typical neoclassical CFs grouped in cluster 5. The low number of items coming from the 'new' set is due to the fact that it is a rather closed class of elements, and the majority of its members had already been selected in the 'old set'. Among the items of the 'new' set prevail the numerals (e.g. *epta-, quadri-, bi-*), but there are also some items that are close to the center of the cluster (e.g. *epi-, omo-, peri-*). [8]

Cluster 2 is made up of CFs that are used exclusively in final position in the formation of both endocentric and exocentric compounds. These are elements of verb origin, that form nouns and adjectives with an agent or instrumental meaning (e.g. *-foro, -fago, -grafo*^{2}*, -crate, -grado, -voro, -cida, -colo, -fero*). These CFs give life to compounds that have some common features with *synthetic compounds* of the Germanic languages and the compounds of the V+N type of the Romance languages (cf. Engl. *meat-eater* and It. *carnivoro*, It. *portabandiera* and Late Lat. *vexilliferu(m)*): [9] the initial constituent is almost always interpreted as the argument of the final constituent (*fruttifero* 'fructiferous'*, insettivoro* 'insectivorous'), it can never be the subject, though may sometimes function as an adverb (*tardigrado* 'tardigrade'). Considering their position and the type of signification they are the items that come closest to derivational suffixes, though they have many characteristics in common with the typical CFs (cf. Iacobini 2004a: 89-95). Most of these CFs are employed in the formation of a high number of compounds. Such formations are often related one to another via suffixation (e.g. *geologo* 'geologist' / *geologia* 'geology' / *geologico* 'geologic') The 'new' entries are not many (the reasons are the same ones applying to cluster 1), but they are well integrated in the cluster (e.g. *-bolo, -bate, -fugo, -mane*).

Cluster 3 is distinguished by the typical CFs in that the items that characterize them are shortening of modern (compound and non-compound) words, and are employed in initial position mostly before words (e.g. *socio-, cine-, psico*^{2}*-, tele*^{2}*-, italo-, euro-, vibro-, fanta-, moto*^{2}*-, antero-*). The most peripheral members correspond to modified words, i.e. current words (usually nouns) used as the first member of the compound with the final vowel modified with an *o* or an *i* (e.g. *laringo-* cf. *laringe* 'larynx'*, musico-* cf. *musica* 'music'*, acqui-* cf. *acqua* 'water'), so as to uniform itself with the characteristic neoclassical CFs endings. [10] Shortenings are particularly adapt to be used as the initial element of a compound due to their phonological and prosodic characteristics (cf. Thornton 2007). We

can distinguish two types of CFs that are made by shortening. In the first type, the result of the shortening coincides with the formative element (e.g. *eco*[2]- from *ecologia* 'ecology', *foto*[2]- from *fotografia* 'photography'). Couples of homonyms are found among the prototypical CFs in this way (*eco*[1]- 'house, dwelling', *foto*[1]- 'light' that have been classified here with the central members of C5) and the shortened forms, that we may define as 'second generation', these last (as is normal for shortenings) assume the meaning of the word they originate from. In the second type, shortening does not coincide with the formative element of the original word, because the original word is not a compound (*socio*- from *sociale* 'social'), or because the original word is a compound, and the shortening does not respect the etymological or morphological segmentation (*bici*- from *bicicletta* 'bicycle').

Cluster 4 is the most compact cluster: it is made up of items that are very similar to each other and significantly different from the items in the other clusters. It is composed totally of the initial elements of V+N compounds (e.g. *apri*- 'open', *bacia*- 'kiss', *buca*- 'pierce', *pesa*- 'weigh', *carica*- 'charge'), the only distinction internal to the cluster is due to the number of syllables, the more typical items are disyllabic, the items with three syllables occupy a position that is slightly more external. [11] The cluster is made up totally of elements that belong to the 'new' set. As we have already said, this is a consequence of the GRADIT lemmatization criteria, which label with the same term *confisso* 'confix' this type of elements together with typical CFs and other bound elements. It is interesting to note that the procedure of automatic analysis recognized and grouped all these elements into a single cluster; they are characterized by fixed position, modern language origin, word combination, everyday language use, and the particular relation with the combining nouns.

Cluster 5 is definitely larger than the others (55% of the total corpus) and is the one with the most new CFs (45% of the new CFs go into this cluster). The typical neoclassical CFs group into this cluster. The neoclassical CFs manifest properly lexical characteristics and internally present a large variety of behavior, distributed so that none of their defining characteristics prevails significantly over the others. The most important characteristic of the cluster is exactly the absence of relevant properties. The CFs express meanings of the denotative-lexical type (principally originating from nouns), a strong tendency to be used in technical-scientific terminology, an almost exclusive combination with other CFs, and almost exclusive classical language origin with a strong preference for Greek, with few positional restrictions. Even though the most central members of the cluster (*emo-, ipso-, chimo-, etmo-, lasio-, meli-*) are used in initial position, it is the cluster with the least position

restrictions: in fact almost all of these CFs can be used in the initial or final position (e.g. *-onco-, -piro-, -chiro-, -gastro-, -xilo-, -andro-, -glotto-*) are part of this cluster. The CFs are almost exclusively disyllabic. The trisyllables occupy more peripheral positions.

Cluster 6 is second only to Cluster 5, both in the total number of items and in the number of new items. The items share the attributes of C5 (lexical meaning, technical-scientific use, combination with other CFs, classical language origin), differentiating only in the tendency to occupy the final position and to have the higher number of syllables of the two. Many of the most representative elements end with the suffix *–ìa* (e.g. *–plegia, –ragia, –termia, –algia, –emia, –dattilia, –cefalia, -tomia, -latria*), the corresponding not suffixed CF in majority belong to Cluster 5 (e.g. *algo-, emo-, -dattilo-, -cefalo-*), but some of them belong to Cluster 2 (*-tomo, -latra*). CFs ending with the suffix *-ia* occupy the final position exclusively, they are not further derivable by the adding of a suffix (affix substitution is very common: *filologia* 'philology' > *filologico* 'philologic'), and they normally form endocentric words. The cluster also keeps non-suffixed CFs (*-fima, -ftisi, -gipio, -nosi, -ptene, -rriza*) that occupy the final position exclusively. In the most external zone we also find a small number of initial trisyllable CFs of Greek origin (*onoma-, dolico-, cineto-, gefiro-*).

Cluster 7 is formed by CFs that are very similar to the two previous clusters, differentiated from Cluster 5 and Cluster 6 only because they are all of Latin etymon and the tendency to occupy the initial position (e.g. *ludo-, flessi-, igni-, olei-, scuti-, nivo-, audio-*). Even if the total number of items in the cluster is not large, the percentage of 'new' entries is among the highest. The influence of the Latin etymon criterion is probably the reason why this cluster hosts some items that - from a linguistic perspective - may be better allocated to other clusters (e.g. *ispano*, cf. *franco* and *italo* in C3, and three final items like *cidio, plano, ficio*); it is interesting to remark that all these CFs occupy the most peripheral positions of the cluster.

The numeric distribution of the CFs in the seven clusters is not very uniform at all: there are very crowded clusters (C5, 255 items), and another that is very rich in elements (C6, 114 items), the two correspond to 65% of the CFs. The CFs of C7 have very similar attributes to the two preceding clusters, but they are much inferior in number (37 members). C2 and C3 are formed by a number of items similar to C7. C1 is bigger (60 members) compared with the suffix-like elements grouped in Cluster 2 and the shortenings in Cluster 3, but counts less than a quarter of the items in C5. C4 is formed exclusively by 'new' items (first members of verb-noun compounds). The members of C4 clearly differentiate from those of

the other clusters, and they share with the member of the other clusters almost only the fact of being bound elements employed in compounds.

The percentage of the 'new' items in the various clusters is correlated with their attributes. The clusters formed by typical neoclassical CFs (C5, 6, 7) are those that have a greater increase in the percentage of 'new' items (about 40%); while the increase is less for C2 (28%), and is decisively inferior to C1 (18%) and C3 (7%).

The statistical comparison demonstrates a robust congruency between the analyses made on the 'old' and the 'global' set. The number of clusters remained unvaried, even though in the 'global' set a cluster (C4) formed that is made up of items that were not present in the 'old' set. This may be explained by the fact that a cluster in the 'old' set was made up of only two numeral elements (*ambi-, duo-*) and it was absorbed (together with other numerals not present in the 'old' set) by the cluster that contains initial elements that have strong similarity with derivational prefixes (C1). C5 grew in number compared to the analogous cluster of the 'old' set due to the contribution of some of the items that came from another 'old' cluster ('old' Cluster 6) that had very similar characteristics. The individuation of the 'new' C6 and C7 introduce some differences internal to the typical neoclassical CFs (related to preferred position and Latin vs. Greek etymon) which are of interest for an in-depth linguistic analysis, but beyond the scope of this work.

Beside the numerical data relative to the quantity and the percentage of increase, it is interesting to note how C5, together with C6 and C7, are virtually non-defined in the component space, in other words, they have all of the characteristics used to describe CFs in the medium range. This is a very relevant aspect; it tells us that the choice of the variables to represent the CFs is absolutely congruent with the scope. This aspect allows us to locate a large part of the corpus in a central position. In other words, all the chosen variables are equally represented by the nucleus of the corpus; and on the other hand, it indicates that the prototypical CFs are with C5 (and in keeping with the other two most similar clusters, C6 and C7). The absence of any positive connotation in C5 shows the impossibility of hypothesizing a different category for CFs that would be added to the lexeme and affix categories: the typical neoclassical CFs share the principal characteristics of lexemes (differing in being bound elements that can not be used as independent words through the addition of inflectional affixes, and in their technical-scientific terminological use). This type of interpretation may at first seem to be in conflict with what one would expect in a definition of prototypical nature, but it actually represents a totally legitimate example: that what the members in C5 have in common is that none of the characteristics prevail significantly over the

others. The more the other clusters distance from C5, that is, those with more defined values, (C1, C3, C4) highlight a shift in the concept of CFs toward other types such as: derivational prefixes (C1), the CFs result of current word shortening and modification (C2), the first elements of the verb-noun compounds (C4); and C2, while presenting some characteristics in common with derivational affixes, share some important defining characteristics of typical neoclassical CFs.

In conclusion, this linguistic analysis of the statistical grouping in clusters points out the effectiveness of the classification applied through a multidimensional analysis, and proves to be totally compatible with recently proposed linguistic classifications (cf. Iacobini 2004a, Kastovsky 2009) which consider the similarity between a limited and well-identified number of CFs and derivational affixes as the outcome of common processes of grammaticalization involving elements of complex words. The results of multidimensional analysis also argue against approaches giving a primary role to positional criteria.

7. Predictivity

The application of MDA techniques allows us to overcome the serious problem in prototype classification in linguistics, that is the absence of predictivity.

The possibility of using an explicit classification that is collocated in the space of its components (each cluster coordinates are the average of the components relative to the cluster) and a parametric function (each component is expressible in terms of the linear combination of the original variables) makes possible both the assignment of new items that have not been considered in the model construction of the preexisting clusters and the verification of the 'goodness' of the cluster classification that is obtained.

The procedure is quite simple: 1) calculate the value of each of the six principal components relative to the item that one wants to classify on the basis of the variable values listed in Table 1; 2) calculate the Euclidean distance of the corresponding vectors of the item to classify with the 7 pre-existent clusters; 3) assign the new item to the nearest cluster.

The evaluation of the model's predictive power is obtained in the following way: if the new items are found inside a cluster that already exists, this implies that the model is able to generalize, and therefore the underlying implicit theory is sufficiently powerful. On the other hand, if the new items do not have a privileged attractor or form a new cluster

occupying a same previously empty zone, this means that, probably due to a mistaken selection in the initial corpus, or a mistaken choice in the variables, the proposed model does not have heuristic capacities.

8. Conclusions

The comparison of the 'old set' and the 'global set' classification of CFs has demonstrated a marked stability of the clusters that have been identified, even with the input of a large number of new items.

The MDA has revealed to be a method that permits the classification of linguistic objects in clusters; where the distance between the elements is purely based on the structural characteristics of the corpus analyzed, that is, the correlation of the defining characteristics and the variables that have been considered.

From the methodological point of view, this study illustrates the utility of MDA use for linguistic analysis both in regard to the empiric results of the classification (that has a robust linguistic significance) and in regard to predictivity. The construction of clusters by means of explicit measures permits us to evaluate the relevance and the weight of the linguistic criteria used in the classification to determine the optimal number of clusters, to identify the most representative members of each cluster, and also to test the model through the analysis of new items.

From the linguistic point of view, the principal results consist in the identification and classification of the neoclassical CFs, and the distinction between typical neoclassical CFs and other types of bound elements with partially different characteristics. The fact that typical neoclassical CFs group in clusters (C 5, 6, 7), in which all of the principle descriptive variables congregate with an average value, demonstrates that it is not opportune to hypothesize a linguistic category for CFs different from the lexeme or affix. Neoclassical CFs are exogenous bound lexical elements (stems) used in technical scientific registers to form complex words with naming or classificatory functions; these constitute a large and open group. A contained number of neoclassical CFs with well identified attributes (cf. Cluster 1) has some features in common with derivational prefixes, the similarity with some of the neoclassical CFs and derivational suffixes are more fuzzy (cf. Cluster 2). The classification in clusters has pointed out other groups of bound elements that are increasingly differentiated from the typical neoclassical CFs (cf. Clusters 3 and 4). [12] The distinction between neoclassical CFs and the lexical elements of a particular natural language presents some practical problems in establishing boundaries due to greater or lesser integration of technical-scientific

terms and of some CFs in current linguistic usage, but this has no consequence on the opportuneness of attributing a special linguistic category to CFs. The resolution of these uncertainties concerns the classificatory customs of different languages and principally the languages that have a high number of Greek and Latin origin words in their lexicon.

We have demonstrated, with this study, the utility of MDA for an inductive type of linguistic classification. We hope that these analytic cluster techniques may be further extended into different sectors of linguistic analysis.

APPENDIX 1
Results of statistic classification in 7 clusters

CFs are ordered in terms of growing distance from the center: the first CFs of each cluster are the members that best represent it.

CLUSTER 1. iper, neo, pseudo, tele[1] 'distant', allo, meta, olo, tauto, micro, mono, para, poli[1] 'many, several, diverse, much', proto, mega, macro, anfi, penta endo, epta, idio, iso, epi, omo, peri, emi, ipo, auto[1] 'self, one's own, by oneself, independently', ecto, eso, meso, acro, panto, di, pan, opsi, eu, oligo, paleo, omeo, deutero, archeo, aniso, etero, enantio, megalo, protero, midi, teca, milli, centi, pluri, multi, equi, quadri, omni, uni, ambi, vetero, duo, bi.

CLUSTER 2. foro, fago, fobo, grafo[2] 'one that writes about specified material or in a specified way; instrument for making or transmitting records; something written', latra, logo[2] 'student, specialist', metro[2] 'instrument or means for measuring; measure', bolo, anche, bate, brico, maco, crate, tomo, geno, nomo[2] 'specialist', foria, dromo, scopo, ostio, fugo, mane, paro, grado, voro, fono[2] 'speaker of a specified language', coltore, cida, fero, colo, forme, filo.

CLUSTER 3. socio, cine, normo, demo[2] 'democracy, democratic' psico[2] 'psychology, psychological methods', eli, turbo, eco[2] 'ecological, environmental', moto[1] 'motion, motor', 'narco[2] 'of or relating to illegal narcotics', porno, tele[2] 'television, vibro, magneto, franco, elettro, italo, sovieto, austro[1] 'south, southern; Australia Australian', austro[2] 'Austria, Austrian', euro, fanta, indo, immuno, moto[2] 'motorbike, motorcycle', foto[2] 'photography', avio, bio[2] 'biology', fibro, vulvo, chemio, chemo, acqui, pancreo, latero, postero, antero, laringo, musico, farmaco, video, 'radio', auto[2] 'automobile, car'.

CLUSTER 4. apri, bacia, buca, frangi, leva, pela, pesa, piglia, premi, rompi, salpa, sbatti, spargi, strappa, stura, torci, tura, vendi, abbraccia, attacca, carica, infila.

CLUSTER 5. emo, ipso, chimo, etmo, lasio, meli, neso, reo, spodo, stauro, xifo, onco, eto, piro, chiro, crio, elio, pedo, tecno, biblio, cino, criso, mio, alo, copro, eco[1] 'household; economic; habitat or environment', embrio, ergo, isto, noso, oro, istio, fisio, melo, gipso, glico, foto[1] 'light', idro, demo[1] 'people, populace, popula-

tion', eno, etno, gastro, ipno, ittio, algo, bato, gero, gonio, sapro, scoto, xilo, calli, cleido, dino, ditto, elco, ezio, fico, icno, lacco, nicto, oo, orro, osmo, pluto, polio, scato, sfeco, tauro, tefro, teno, tio, zonio, pneumo, petro, neuro, nevro, zimo, cole, flebo, freno, lipo, mico, mielo, necro, nefro, onto, oto, pato, psico[1] 'soul, spirit; mind, mental processes and activities', sfigmo, spleno, narco[1] 'numbness, stupor; narcosis, narcotic; deep sleep; aided by drugs', radio[1] 'radial, radially; radiant energy, radiation; radioactive', cisti, igro, sarco, sema, semio, agri, botrio, antropo, cinesi, calamo, echino, elminto, folide, galeo, ostraco, masto, rizo, cheilo, bradi, tachi, xero, auxo, orto, xeno, sito, brachi, tiflo, adro, anfo, dasi, drio, pachi, pauro, picno, plesio, stilpno, trachi, cefalo, argiro, astero, metopo, onfalo, aplo, iero, oftalmo, dattilo, odonto, leuco, sclero, xanto, opto, steno, caco, callo, cripto, liso, tassi, blasto, ippo, dendro, lito, carpo, gino, artro, cheto, gnato, tamno, telo, procto, mero, belo, cebo, cerco, conio, erio, glifo, grapto, lemo, lofo, meco, mene, mia, placo, ptico, rinco, stachi, taco, terio, toco, troco, andro, artro, baro, dermo, fillo, gamo, tropo, crono, glosso, glotto, grafo[1] 'writing', morfo, rino, topo, cirto, mizo, odo, soma, spermo, diplo, angio, logo[1] 'discourse, talk', nomo[1] 'usage, law', derma, trico, geo, trofo, cordo, oniro, fito, cromo, fono[1] 'sound, voice, speech, tone', cloro, cito, lisi, tipo, stato, feno, bari, aero, dermato, talasso, zoo, bio[1] 'life, course or way of living', cardio, aldo, crico, eroto, blefaro, adeno, metro[1] 'uterus', entero, epato, onico, osteo, sidero, tanato, cromato, emato, galacto, ornito, steato, crocido, stereo, eritro, termo, actino, melano, stomato, mascalo, lepto.

CLUSTER 6. onoma, algia, tomia, dolico, cineto, coleo, emido, faringo, gefiro, pireto, psammo, emia, patia, plegia, ragia, scopia, termia, speleo, gineco, fagia, filia, iatria, latria, logia, metria, tipia, cardia, cromia, dermia, dromia, fillia, opsia, penia, plasia, tecnia, trofia, cheiria, geusia, teutide, crazia, ampelo, climato, entomo, cele, stasi, ieria, meride, rea, biosi, cicla, cladio, cormo, fima, ftisi, gipio, nosi, ptene, rriza, sciuro, sepo, tropio, anemo, estesio, geronto, espero, ipero, mirmeco, porfiro, morio, podio, machia, sofia, stenia, gramma, iatra, scopio, urgia, grafia, mania, manzia, nomia, edro, agio, auxano, fobia, onimo, allelo, anoplo, apalo, entelo, omalo, schisto, comio, mante, nauta, rama, meccano, poli[2] 'city', cefalia, dattilia, estesia, megalia, mielia, onichia, allegro, canfo, meningo, dibromo, toraco, gengivo, acetil, addomino, amigdalo, meteoro.

CLUSTER 7. ludo, flessi, igni, olei, scuti, pluvio, fluvio, ovi, sono, cuni, silvi, labio, coxo, digito, sino, castani, avi, burso, radicolo, oleo, vasculo, anglo, maxillo, cerebro, paremio, arbori, nivo, nulli, balneo, ispano, audio, ovo, dotto, cidio, lirio, plano, ficio.

Addresses of the Authors

Claudio Iacobini, Dipartimento di Studi Linguistici e Letterari, Università di Salerno, Via Ponte Don Melillo, 84084 Fisciano (SA), Italy, <ciacobini@unisa.it>

Alessandro Giuliani, Environment and Health Department, Istituto Superiore di Sanità, Viale Regina Elena 299, 00161, Rome, Italy <alessandro.giuliani@iss.it>

Notes

[1] Cf. Amiot & Dal (2005: 324-327) for an overview of the divergent classification approaches, and Prćić (2008: 2), who speaks of an unsettled state of affairs about the (in)ability of modern morphological theory to work out "a principled and consistent way of distinguishing between affixes and combining forms".

[2] According to Ramat (1999: 158) the attribution of a lexeme to a part of speech happens following the identification of the correlation between the features: "A category is a set of objects that are considered as having common features. Some of these features can be shared by other objects, but not all of them at the same time (otherwise all the objects would belong to the same category)". On the use of multidimensional (or multivariate) statistical techniques in typological research cf. Kapatsinski (2008), for a review of their (under)use, see Cysouw (2007).

[3] Cf. Bolasco (1999), Benigni & Giuliani (1994), Sneath & Sokal (1973).

[4] The literature specifically dealing with the subject is not very rich: apart from the "classical" Hatcher (1951) and Marchand (1967), most publications are quite recent, a selected list includes: Schmidt (1987), Warren (1990), ten Hacken (1994), Bauer (1998), Lehrer (1998), Fradin (2000), Baeskow (2004), Booij (2005), Prćić (2005), Amiot & Dal (2007), Prćić (2008), Kastovsky (2009). As far as Italian language is concerned, readers may see: Tollemache (1945), Migliorini (1963), Iacobini & Thornton (1992), Antonelli (1995), Iacobini (1999), Iacobini & Giuliani (2001), Sgroi (2003), Iacobini (2004a, b).

[5] Kastovsky's (2009: 9) definition of stem is "a word-class specific lexeme representation which cannot occur on its own as a word but has to combine with additional derivational and/or inflectional morphemes to function as a word, i.e., it is a bound form. It may itself contain derivational affixes or so-called stem formatives, which determine the inflectional category".

[6] Since affixes tend to be shorter than words, CFs with a low number of syllables might be favoured in the shift toward the affix-like status.

[7] Cf. Bolasco (1999), Benigni & Giuliani (1994), Bartholomew (1984), Lebart, Morineau & Warwick (1984).

[8] These last items are not considered in the old set because they were either classified as affixes or not lemmatized as dictionary entries.

[9] Cf. Namer & Villoing (2007).

[10] The status of the final vowel of initial combining forms is questionable. Even if historically, vowels *-i* and *-o* are mostly stem formative of the first member of compound, from a synchronic point of view they may also be considered as linking elements (*Fugenelement*) triggered by the particular compound pattern.

[11] On the prosodic characteristics of such elements cf. Ricca (2005) and Thornton (2008).

[12] Other types of bound elements employed by Romance languages in compounding are described in Fradin (2000) and in Iacobini (2004a).

Bibliographical References

ALINEI Mario 1962. *Dizionario inverso italiano.* The Hague: Mouton.

AMIOT Dany & Georgette DAL 2007. Integrating Neoclassical Combining Forms into a Lexeme-Based Morphology. In Booij, Ducceschi, Fradin, Guevara, Ralli & Scalise 2007.

ANTONELLI Giuseppe 1995. Sui prefissoidi dell'italiano contemporaneo. *Studi di Lessicografia Italiana* 13. 253-293.

BAESKOW Heike 2004. *Lexical Properties of Selected Non-native Morphemes of English*. Tübingen: Gunter Narr.
BARTHOLOMEW David J. 1984. The foundation of factor analysis. *Biometrika* 71. 221-232.
BATTAGLIA Salvatore 1961-2002. *Grande dizionario della lingua italiana*. Torino: Utet.
BAUER Laurie 1998. Is there a class of neoclassical compounds, and if so, it is productive? *Linguistics* 36.3. 403-422.
BENIGNI Romualdo, Grazia GALLO, Francesco GIORGI & Alessandro GIULIANI 1999. On the Equivalence Between Different Descriptions of Molecules: Value for Computational Approaches. *Journal of Chemical Information and Computer Science* 39.3. 575-578.
BENIGNI Romualdo & Alessandro GIULIANI 1994. Quantitative modeling and biology: the multivariate approach. *American Journal of Physiology* 266. R1697-R1704.
BOLASCO Sergio 1999. *Analisi multidimensionale dei dati. Metodi, strategie e criteri d'interpretazione*. Roma: Carocci.
BOOIJ Geert 2005. Compounding and Derivation. Evidence for Construction Morphology. In DRESSLER Wolfgang U., Dieter KASTOVSKY, Oskar E. PFEIFFER & Franz RAINER (eds.). *Morphology and its Demarcation*. Amsterdam: Benjamins. 109-131.
BOOIJ Geert, Luca DUCCESCHI, Bernard FRADIN, Emiliano GUEVARA, Angela RALLI & Sergio SCALISE (eds.) 2007. *Proceedings of the Fifth Mediterranean Morphology Meeting* (MMM5) Fréjus (France) 15-18 September 2005. University of Bologna. URL: <http://mmm.lingue.unibo.it/>.
BROOMHEAD David S. & George Peter KING 1986. Extracting Qualitative Dynamics from Experimental Data. *Physica D* 20. 217-236.
CORTELAZZO Manlio & Ugo CARDINALE 1989. *Dizionario di parole nuove 1964-1987*. Torino: Loescher.
CORTELAZZO Manlio & Paolo ZOLLI 1979-1988. *Dizionario etimologico della lingua italiana*. Bologna: Zanichelli.
CYSOUW Michael 2007. New approaches to cluster analysis of typological indices. In KÖHLER Reinhard & Peter GRZBEK (eds.). *Festschrift für Gabriel Altmann*. Berlin: Mouton. 61-75.
DEVOTO Giacomo & Gian Carlo OLI 1990. *Nuovo vocabolario illustrato della lingua italiana*. Firenze: Le Monnier.
DISC 1997= *Dizionario italiano Sabatini Coletti*. Firenze: Giunti.
FORCONI Augusta 1990. *Dizionario delle nuove parole italiane*. Milano: SugarCo.
FRADIN Bernard 2000. Combining forms, blends and related phenomena. In DOLESCHAL Ursula & Anna M. THORNTON (eds.). *Extragrammatical and marginal Morphology*. München: Lincolm Europa. 11-59.
GAETA Livio & Davide RICCA 2009. *Composita solvantur*: Compounds as lexical units or morphological objects? *Rivista di Linguistica* 21,1. 35-70.
GARZANTI 1993= *Il grande dizionario della lingua italiana*. Milano: Garzanti.
GRADIT 1999= *Grande dizionario italiano dell'uso*. Torino: UTET.
GROSSMANN Maria & Franz RAINER (eds.) 2004. *La formazione delle parole in italiano*. Tübingen: Max Niemeyer Verlag.

HATCHER Anna Granville 1951. *Modern English word-formation and Neo-Latin*. Baltimore, MD: Johns Hopkins Press.

IACOBINI Claudio 1999. Distinguishing derivational prefixes from initial combining forms. In BOOIJ Geert, Angela RALLI & Sergio SCALISE (eds.). *Proceedings of First Mediterranean Morphology Meeting*, Mytilene (Grecia), 19-21 settembre 1997. 132-140.

IACOBINI Claudio 2004a. Composizione con elementi neoclassici. In Grossmann & Rainer 2004. 69-95.

IACOBINI Claudio 2004b. Prefissazione. In Grossmann & Rainer 2004. 97-163.

IACOBINI Claudio & Alessandro GIULIANI 2001. Sull'impiego di metodi quantitativi nella classificazione degli elementi che prendono parte ai processi di formazione delle parole. In ALBANO LEONI Federico, Eleonora STENTA KROSBAKEN, Rosanna SORNICOLA & Carolina STROMBOLI (eds.). *Dati empirici e teorie linguistiche. Atti del XXXIII Congresso internazionale di studi della SLI*. Roma: Bulzoni. 331-359.

IACOBINI Claudio & Anna M. THORNTON 1992. Tendenze nella formazione delle parole nell'italiano del ventesimo secolo. In MORETTI Bruno, Dario PETRINI & Sandro BIANCONI (eds.). *Linee di tendenza dell'italiano contemporaneo. Atti del XXV Convegno internazionale di studi della SLI*. Roma: Bulzoni. 25-55.

KAPATSINSKI Vsevolod 2008. Principal components of sound systems: An exercise in multivariate statistical typology. *IULC Working Papers Online* 8. URL: https://www.indiana.edu/~iulcwp/.

KASTOVSKY Dieter 2009. Astronaut, astrology, astrophysics: About Combining Forms, Classical Compounds and Affixoids. In MCCONCHIE Rod W., Alpo HONKAPOHJA & Jukka TYRKKÖ (eds.). *Selected Proceedings of the 2008 Symposium on New Approaches in English Historical Lexis (HEL-LEX 2)*. Somerville, MA: Cascadilla Proceedings Project. 1-13.

LAKOFF George 1987. *Women, fire and dangerous things: What categories reveal about the mind*. Chicago, IL: University of Chicago Press.

LEBART Ludovic, Alain MORINEAU & Kenneth M. WARWICK 1984. *Multivariate descriptive statistical analysis*. New York: Wiley.

LEHRER Adrienne 1998. Scapes, holics, and thons: The semantics of English combining forms. *American Speech* 73. 3-28.

LURATI Ottavio 1990. *3000 Parole nuove: La neologia negli anni 1980-1990*. Bologna: Zanichelli.

MARCHAND Hans 1967. Expansion, Transposition and Derivation. *La Linguistique* 1. 13-26.

MIGLIORINI Bruno 1963. I prefissoidi (il tipo aeromobile, radiodiffusione). In Id., *Saggi sulla lingua del Novecento*, III ed.. Firenze: Sansoni. 9-60 (1935[1] with the title "Il tipo *radiodiffusione* nell'italiano contemporaneo". *Archivio Glottologico Italiano* 27. 13-39).

NAMER Fiammetta & Florence VILLOING 2007. Have Cutthroats Anything to Do with Tracheotomes? Distinctive Properties of VN vs. NV Compounds in French. In Booij, Ducceschi, Fradin, Guevara, Ralli & Scalise 2007.

PRĆIĆ Tvrtko 2005. Prefixes vs initial combining forms in English: A lexicographic perspective. *International Journal of Lexicography* 18. 313–334.

PRĆIĆ Tvrtko 2008. Suffixes vs Final Combining Forms in English: A Lexicographic Perspective. *International Journal of Lexicography* 21. 1-22.
QUARANTOTTO Claudio 1987. *Dizionario del nuovo italiano*. Roma: Newton Compton.
RAMAT Paolo 1999. Linguistic categories and linguists' categorizations. *Linguistics* 37.1. 157-180.
RATTI Daniela, Lucia MARCONI, Giovanna MORGAVI & Claudia ROLANDO 1988. *Flessioni, rime, anagrammi: l'italiano in scatola di montaggio*. Bologna: Zanichelli.
RICCA Davide 2005. Al limite tra sintassi e morfologia: i composti aggettivali V-N nell'italiano contemporaneo. In GROSSMANN Maria & Anna M. THORNTON (eds.). *La formazione delle parole. Atti del XXXVII congresso internazionale di studi della SLI*. Roma: Bulzoni. 465-486.
SCHMIDT Günter Dieterich 1987. Das Affixoid: Zur Notwendigkeit und Brauchbarkeit eines beliebten Zwischenbegriffs der Wortbildung. In Gabriele HOPPE (ed.). *Deutsche Lehnwortbildung*. Tubingen: Narr. 53-101.
SGROI Salvatore 2003. "Per una definizione di confisso": composti confissati, derivati confissati parasintetici confissati *vs* etimi ibridi e incongrui. *Quaderni di semantica* 24.1. 81-153.
SNEATH Peter H. & Robert SOKAL 1973. *Numerical Taxonomy*. Dordrecht: Springer.
TAYLOR John R. 1989. *Linguistic categorization*. Oxford: Clarendon.
TEN HACKEN Pius 1994. *Defining Morphology: A Principled Approach to Determining the Boundaries of Compounding, Derivation, and Inflection*. Hildesheim: Olms.
TEN HACKEN Pius 2000. Derivation and Compounding. In BOOIJ Geert, Christian LEHMANN & Joachim MUGDAN (eds.). *Morphologie / Morphology*. Berlin / New York: Mouton de Gruyter. 349-359.
THORNTON Anna M. 2007. Phenomènes de réduction en italien. In DELAIS-ROUSSARIE Élisabeth & Laurence LABRUNE (eds.). *Des sons et des sens. Données et modèles en phonologie et en morphologie*. Paris: Hermès Science / Lavoisier. 241-268.
THORNTON Anna M. 2008. Italian Verb-Verb reduplicative Action Nouns. *Lingue e linguaggio* VII.2. 209-232.
TOLLEMACHE Federigo 1945. *Le parole composte nella lingua italiana*. Roma: Rores.
VAN RYZIN John (ed.) 1977. *Classification and Clustering*. New York: Academic Press.
WARREN Beatrice 1990. *The importance of combining forms*. In DRESSLER Wolfgang U., Hans C. LUSCHÜTZKY, Oskar E. PFEIFFER & John R. RENNISON (eds.). *Contemporary Morphology*. *Proceedings of the 3rd International Morphology Meeting*. Berlin / New York: Mouton de Gruyter. 111-132.
ZINGARELLI 1999= *Vocabolario della lingua italiana di Nicola Zingarelli*. XII ed. Bologna: Zanichelli.