Outils de *Text Mining* pour l'analyse de structures lexicales à éléments variables

Sergio Bolasco[1], Rosanna Verde[2], Simona Balbi[3]

[1] Dip. Studi Geoeconomici, Linguistici, Statistici SAR – Università "La Sapienza" di Roma – Italia – sergio.bolasco@uniroma1.it

[2] Dip. di Strategie Aziendali e Metodologie Quantitative – Seconda Università di Napoli – Italia – rosanna.verde@unina2.it

[3] Dip. di Matematica e Statistica – Università "Federico II" di Napoli – Italia – sb@unina.it

Abstract

The paper aims at showing the advantages of formulating lexical structures with variable elements (LSVE) in terms of symbolic objects (OS). The proposal main consequence is the collapse of the huge information usually present in a text into matrices of complex data. Thus, it is possible to use statistical tools in order to analyse the underlying properties of lexical structures by taking into account their different component distributions. Describing lexical structures with variable elements as symbolic data improves the effort of text mining by putting in a strict relation the knowledge extraction and the statistical analysis steps. The study of such lexical structures is performed by applying a factorial analysis on complex data. An application deals with a very large corpus extracted from Italian newspaper "La Repubblica" during the Nineties. In this study, we analyse the relations among the different components of some lexical structures with variable elements and the temporal evolution of some peculiar semantic traits, related to geographical-historical-political contexts.

Keywords: lexical structures, text mining, factorial analysis on symbolic data, fuzzy coding, corpus linguistics.

Résumé

Dans ce papier on donne un cadre théorique différent au système de détection des structures lexicales à éléments variables (SLEV), en se servant - du point de vue statistique - de la notion de "donnée symbolique". Une formalisation des SLEV en termes d'objets symboliques (OS), présente des avantages tels que: recueillir sous une même classe diverses occurrences d'hapax qui autrement resteraient dispersées, rassembler l'information présente dans un texte en la comprimant considérablement dans des matrices de données complexes et, au même temps, analyser certaines propriétés des SLEV en fonction de la distribution des modalités de ses composantes. Après avoir donné des exemples de SLEV, on applique une méthode d'analyse factorielle pour données symboliques à une opération de text mining sur un corpus de très grandes dimensions, formé des collections d'articles de presse des années 1990, en étudiant l'évolution d'un trait sémantique géo/historique-politique.

Mots-clés : linguistiques de corpus, text mining, analyse factorielle sur données symboliques, codage flou

1 Introduction

Ce travail[1] s'inscrit dans un cadre de procédures d'intégration statistique-linguistique (Bolasco 2000, 2001) dont l'objectif est de montrer comment la présence de certaines unités lexicales, spécifiquement classifiées, permet de découvrir des éléments d'ensemble pour l'évaluation de l'*imprinting* d'un texte mesurable par l'incidence de quelques caractéristiques du discours (catégories grammaticales, classes dérivationnelles des mots, traits sémantiques) rapportées à d'autres textes de référence.

[1] Travail développé grâce au fonds CNR n° 99.03526.CT10

Ces caractéristiques sont souvent cachées dans des unités lexicales de basse fréquence ou dispersées sous forme d'hapax. D'ailleurs, beaucoup de ces unités font partie de "classes naturelles", soit morphologiques soit sémantiques. Il s'agit d'envisager, selon un objectif particulier, les structures lexicales qui sont à la base de ces classes, par exemple, les processus de formation des mots (par suffixation ou préfixation et, plus généralement, distinguer les affixes et la racine d'un mot), la typologie des adverbes composés (temps, lieu, quantité...), les classes des verbes, les types de temps et/ou de personnes dans les flexions verbales (en distinguant le morphème lexical du morphème grammatical), ou les pronoms personnels (clitiques verbaux inclus), des noms propres ou autre.

Ces classes, constituées par des données complexes, sont des structures à "éléments variables" (SLEV). Ce type de structures lexicales peut se présenter non seulement sous forme de lexies complexes (<Dnum Ntmp fa>) mais aussi sous forme de lexies simples (<antiterrorismo>) ou composées (<mille|nove|cento|quindici>). Les effectifs des formes différentes, pour chaque structure, peut aller de quelques dizaines d'éléments jusqu'à un nombre théoriquement infini de termes, comme dans le cas des chiffres écrits en lettres.

Dans cette perspective, grâce à une formalisation générale capable d'identifier de telles structures, dans ce papier on entend montrer l'utilité, pour une analyse textuelle, de: a) exécuter de simples opérations de *text mining*; b) recueillir sous une même classe (structure) diverses occurrences d'hapax qui autrement resteraient dispersées ; c) rassembler l'information présente dans un texte en la comprimant considérablement dans des matrices de données complexes; d) analyser certaines propriétés des SLEV en fonction de la distribution des modalités de ses composantes.

Une formalisation linguistique des structures lexicales telles que les classes des verbes et les mots composés est connue depuis longtemps (Elia 1984, 1996), alors que leur définition en tant que automates finis a été donné par les graphes d'Intex (Silberztein 1999, Gross 1999).

Il est possible de donner, cependant, un cadre théorique différent au système de détection des SLEV, en se servant - du point de vue statistique - de la notion de "donnée symbolique". Une telle formalisation présente des avantages et produit des utilités, du fait qu'elle constitue une étape essentielle pour la construction des matrices d'information (Balbi & Giordano 2000) utiles à comprimer les grandes matrices de données ou pour obtenir des méta-informations (Bolasco 1998) à associer à la matrice de base pour l'étude du corpus.

Ensuite, après avoir donné des exemples de SLEV, nous appliquerons la méthode proposée (analyse factorielle sur données symboliques) à une opération de *text mining* sur un corpus de très grandes dimensions, formé des collections complètes d'articles de presse des années 1990, en étudiant l'évolution d'un trait sémantique particulier de type géo/historique-politique.

2. Exemples de structures lexicales à éléments variables

On montre ici quelques exemples de structures lexicales qui peuvent être formalisées en tant qu'objets symboliques (OS). Les structures sont exprimées sous forme de lemme, mais elles sont saisies dans les textes à travers leurs flexions possibles.

1. On considère, par exemple, un adverbe composé de type temporel: expressions comme *<due mesi fa>* (il y a deux mois), *<dieci giorni or sono>* (il y a dix jours) constituent des occurrences spécifiques d'une structure plus générale "à éléments variables" définie par trois composantes: un déterminant numéral Dnum + un mot indiquant une unité temporelle Ntmp + un mot que renvoie au passé. Cet adverbe peut être formalisé dans la structure (Elia, 1996) de "forme canonique"

<Dnum Ntmp fa >

où les variables Dnum et Ntmp possèdent comme éléments de base: Dnum={*nombres naturels*}, Ntmp={*millennio, secolo, decennio, lustro, anno, stagione, mese, ..., attimo*}. La troisième composante peut comprendre plusieurs éléments comme {*fa, prima, or sono*}[2].

Il est évident que ces expressions -dans leur singularité- ne sont pas faciles à identifier dans un texte, ni très fréquentes. Au contraire, il est probable que, dans l'ensemble, leur "lemmatisation sémantique", dans le sens de la structure décrite, produira une quantité d'occurrences non négligeable.

2. Un deuxième exemple est fourni par les lexies "composées". On aura intérêt à relever dans un texte la présence de nombres écrits en lettres (dates, chiffres bancaires, etc.), tels que: *mille*|*nove*|*cento*|*settanta*|*quattro, un*|*milione*|*due*|*cento*|*cinquant*|*uno*|*mila*. L'objet est ici plus complexe, car il comporte un effectif variable de composantes (également variables). On peut le voir dans la structure récursive du type <[Dnum][Dnum]> où l'ensemble Dnum est exhaustivement défini par 63 éléments {un, uno, una, due, tre, quattro, quattr, ..., dieci, undici, undic, ..., venti, vent, trenta, ..., cento, cent, mille, mill, mila, milione, milioni, miliardi, miliard, *Nul*[3]}.

Avec quelques variantes, on peut ainsi saisir les ordinaux <[Dnum][Dnum]*esimo*> (*tre*|*cento*|*venti*|*sett*|*esimo*) ou les expressions qui décrivent des cohortes ou des classes d'âge <[Dnum][Dnum]*enne*> (*sessanta*|*cinqu*|*enni*)[4]. On peut noter que le troisième élément présente autant de valeurs que ses flexions (*esimo/a/i/e*, *enne/i*).

3. Un exemple, sur lequel nous reviendrons, est celui relatif à une classe dérivationnelle. On considère ici la classe des mots en *-ismo* (-isme). Cet ensemble d'unités lexicales caractérise un trait sémantique très large, constitué par des vocables, généralement abstraits, concernant des "doctrines et mouvements religieux, sociaux, philosophiques, littéraires, artistiques (par ex.: *islamismo, socialismo, empirismo, realismo, impressionismo* etc.), des attitudes, caractères collectifs ou individuels (*eroismo, radicalismo, dispotismo, ottimismo*), des tendances littéraires et artistiques (*futurismo, cubismo*), des comportements ou actions (*disfattismo, ostruzionismo*), des conditions ou qualités, et aussi des défauts moraux ou physiques et habitudes nuisibles (*egoismo, daltonismo, alcolismo*), ou des états physiques (*parallelismo, magnetismo*), activités sportives (*ciclismo, podismo,* qui sont connectés avec des substantifs en *-ista* comme *ciclista*, etc.) ...[5] " Cette amplitude ouvre la possibilité de caractériser des sous-classes d'intérêt qui peuvent engendrer des données symboliques, correspondant à autant d'objets dans une définition plus générale (§§ 3 e 4).

Dans une perspective exhaustive de *text mining*, il sera utile d'identifier, à l'intérieur de chaque classe, les mots qui se terminent en **iano*/*ista*/*istico* en plus des mots en **ismo.* C'est souvent le cas des dérivations à partir d'un nom (Marx: *marxismo, marxiano, marxista, marxistico*).

Si l'on veut centrer l'attention seulement sur certaines de ces sous-classes, en particulier celles concernant les contextes historico-politiques, le choix d'un petit nombre de préfixes

[2] Un tel objet peut être étendu en incluant, dans l'ensemble Dnum, des adjectifs indéfinis comme: <tanto>, <poco>, <qualche>, ou en incluant dans Ntmp la forme <tempo> comme par exemple <qualche tempo fa>.
[3] Le *Nul* indique la possibilité que l'ensemble soit vide.
[4] On trouve les formes coupées des Dnum avec les suffixes *-esimo* et *-enne*. Ce dernier peut être associé au préfixe <*ultra*> (ultra|sessanta|cinqu|enni).
[5] En outre, "avec un signifié concret, le suffixe *-ismo* sert à dénoter certaines caractéristiques du langage et du style (*arcaismo, grecismo*), quelques dérivés de noms de personne, comme *dantismo, craxismo*); ou a désigner des systèmes et des dispositifs (*meccanismo, organismo*). Dans quelques cas, ce suffixe se change en *-esimo* (*cristianesimo*)." (Vocabolario della Lingua Italiana, Istituto della Enciclopedia Italiana, Roma, 1987, vol. II, p. 997)

(*anti*/*ex*/*filo/neo/post/pro/ultra*) caractérise des mots, identifiant dans la plupart des cas, seulement des noms abstraits de contextes historico-politiques et ceux relatifs à des courants de pensée. Si l'on ajoute, aux mots communs, les noms de célébrités, cela permet d'identifier presque à coup sûr, le trait sémantique désiré (en comprenant, toutes les formes qui se terminent en **ismo/iano/ista/istico*). Cet OS peut être formalisé par une structure du type:

<Préfixe|Base|Suffixe>

où Préfixe={*anti, ex, filo, neo, post, ultra, Nul*[6]}, Base={NomsPropres/Cat(*x*)[7]}, Suffixe={*ismo/i, ista/i/e, iano/a/i/e, istico/a/i/he*}. Une requête sans contrainte sur l'ensemble des vocables de Base comprendrait aussi beaucoup de mots communs (non compris dans l'OS en question): d'où la possibilité de dresser un dictionnaire fondamental (un centaine de termes) capables de saisir le contexte d'intérêt. Pour le contexte politique, les mots de ce dictionnaire fondamental sont facilement associés automatiquement au trait sémantique concerné, en sélectionnant un sous-ensemble de préfixes; il s'agit en particulier de *anti/ex/filo/neo/post/pro/ultra* ... plus d'autres assez rares comme *euro/ pseudo/ pre/ proto/ multi/ nazional/ social/ contro/ vetero*)[8].

Cela confirme que la combinaison préfixe|base|suffixe "fige le sens" et permet de définir un champ sémantique bien précis. Les préfixes mêmes sont désambiguïsés, si on les considère en liaison avec des suffixes. On prends le cas du suffixe <*anti*> intrinséquement ambigu puisqu'il peut signifier <avant> (antiméridien) et <contre> (antinévralgique). Mais dans notre OS "-isme", ce suffixe est présent toujours avec le sens de <contre>.

3. Définition des données symboliques

Une définition de l'unité statistique en termes de donnée symbolique a été proposée par Diday en 1987. Cette définition trouve un domaine d'application, lorsque les unités élémentaires d'analyse ne sont pas de simples individus, caractérisés par une seule modalité de chaque descripteur, mais des unités complexes (par exemple: les différentes espèces d'insectes, les structures grammaticales, etc.), généralement issues d'une requête sur une base des données.

Ces données complexes et structurées peuvent être difficilement reproduites dans une matrice individus × variables pour être analysées par les techniques classiques de l'analyse de données. Ce type de transformation entraînerait la perte d'une grande quantité d'informations, relatives aux valeurs multiples que chaque unité statistique peut présenter soit par rapport à ses descripteurs, soit par rapport aux relations (de dépendances logique et fonctionnelle) qui peuvent exister entre ces descripteurs.

La représentation de l'unité complexe comme une donnée symbolique permet de sauvegarder au mieux les caractéristiques de la structure de la donnée, aussi bien lors de la représentation de l'information dans un tableau, qu'au cours de l'analyse ultérieure.

Un *objet symbolique* (OS) *s* est défini par la triplette (a, R, d) et il permet la modélisation d'un *concept*, selon l'approche proposée par Diday (1989) et récemment formalisée dans Bock et Diday (2000).

[6] Le *Nul* existe pour saisir aussi les mots sans préfixe.
[7] On renvoie au § 4 pour les Catégories relatives aux divers enregistrements de l'OS.
[8] En général, dans les préfixes il peut se produire des pertes d'information du fait des graphies multiples : les éléments d'un mot peuvent être séparés (*ex ministro*), ou liés avec un tiret (*anti-nucleare*) ou tronqués dans leur association avec un mot qui commence avec la même vocale (*ant*|*italiano*). Le premier cas est insoluble au niveau de la numérisation des formes simples; pour les autres on insère toutes les graphies possibles dans la requête (ex.: *anti anti- ant*).

L'élément "d" dans la triplette représente la *description* de l'objet *s* par un ensemble de variables y_1, …, y_p dans le domaine $D=\{D_1,\ldots,D_p\}$. Cette description est un sous-ensemble de D: $d=\{d_1,\ldots,d_p\}\subseteq D$ et contient toutes les valeurs des descripteurs $y_1,\ldots y_p$ que l'objet peut prendre. Ces descripteurs sont définis comme des *variables symboliques* pour les distinguer des variables classiques, puisqu'ils assument des valeurs multiples dans la description de chaque OS. Ces descripteurs peuvent être de différents types: variables à intervalles; variables multi-catégorielles ou variables modales[9].

Dans la description "d" d'un OS il est possible, en plus, de retenir des relations logiques et hiérarchiques entre descripteurs. Ces relations s'expriment par règles de dépendance logique (*si* $y_j=d_j$' *alors* $y_k=d_k$' ou *si* $y_j=d_j$' *alors* y_k=*Non applicable*[10]; avec $d_j'\subseteq d_j$ et $d_k'\subseteq d_k$) ou selon des taxonomies définies sur les modalités des descripteurs. De telles contraintes sur les descripteurs réduisent l'espace de description des objets: cela correspond à une spécialisation des objets cohérente avec les règles définies sur leurs descripteurs.

Soit Ω un ensemble d'éléments $\{w_1,\ldots,w_n\}$ extraits d'une base de données ou, plus généralement, observés sur le même ensemble de descripteurs caractérisant un OS *s.* Grâce à la relation "R" (définie dans la triplette), il est possible de comparer la description de chaque élément w avec la description de l'objet: y(w)Rd. En général, le type d'opérateur R dépend de la nature du descripteur. La comparaison systématique entre les caractéristiques de w et de *s* est effectuée par rapport à chaque caractère :

$$y_1(w)R_1d_1\,;\quad y_2(w)R_2d_2\,;\ \ldots\ ;\ y_p(w)R_pd_p$$

Donc, au moyen de la fonction "a", présente dans la définition de l'OS *s*, il est possible de calculer l'*extension* de cet objet, donnée par l'ensemble des éléments de Ω qui présentent des caractéristiques cohérentes avec la "d" de *s*. La fonction "a" joue donc le rôle de *fonction de reconnaissance ou d'affectation* d'un élément à un *concept* défini par l'*objet symbolique*. Si "a" est une fonction booléenne (*vrai* / *faux*), l'ensemble d'extension d'un objet *s* est donné par:

$$\text{Ext}\,(\Omega|s)=\{w\in\Omega|\ a(w)=y(w)Rd=\text{vrai}\}$$

Nature des descripteurs des structures lexicales selon l'approche symbolique

Dans le domaine particulier des données textuelles, l'idée d'utiliser une représentation des structures lexicales par données symboliques (description de l'OS[11]) est née de l'exigence de

[9] Les variables à intervalles sont quantitatives et présentent, dans la description de chaque objet, un ensemble de valeurs de $\Re$; les variables multi-catégorielles sont qualitatives ou ordinales et, pour chaque objet, elles définissent l'ensemble des modalités que le caractère peut prendre; les variables modales renvoient aux descripteurs multi-catégoriels auxquels est associée une distribution de probabilité ou de fréquence (ou, en général, un système de poids).

[10] Dans le cas choisi dans notre application, la relation consiste dans le fait d'associer un sous-ensemble de préfixes (*anti, neo* etc.) à un ensemble sélectionné de suffixes (*-ismo, -ista, -istico, -iano*).

[11] Exemple. Soit *s* un OS Booléen, caractérisé par des descripteurs multi-catégoriels. En particulier, soit *s* l'objet "sinistra storica" («gauche historique») décrit par l'ensemble: d={Préfixe={Nul_, anti, ex, filo, neo, post, ultra}; Base={... comun, lenin, marx, ... social, soviet, ...}; Suffixe={iana, iane, iani, iano, ismi, ismo, ista, iste, isti, istica, istici, istico}}. La forme graphique w="postcomunista" est classifiée par la fonction "a" comme appartenant à l'extension de *s*="sinistra storica" puisque: Pref(postcomunista)=**post**$\in$ {Nul_, anti, ex, filo, neo, post, ultra}; Base(postcomunista)=**comun**$\in$ {... comun, lenin, marx,social, soviet, ..} et Suff(postcomunista)= **ista**$\in$ { iana, iane, iani, iano, ismi, ismo, ista, iste, isti, istica, istici, istico}; donc: a_s(postcomunista) = vrai. On note que la relation R est l'opérateur '$\in$' qui compare un élément (par exemple, le préfixe 'post-') et un ensemble (es. l'ensemble des prefixes de *s*). Dans cette application, R sera toujours l'opérateur '$\in$', par rapport aux ensembles ({Préf.}, {Base}, {Suff.}) de description de *s*, puisqu'il contient tous les descripteurs multi-catégoriels.

comprimer de façon efficace l'information extraite d'un *corpus* de très grandes dimensions. Mais aussi de trouver une description qui puisse sauvegarder l'information sur certaines caractéristiques structurales de la donnée textuelle et en utilisant également l'information experte, fournie au moyen d'une première classification conceptuelle élaborée sur les formes lexicales, à partir des traits sémantiques ou du contexte analysé..

On utilise des variables modales pour sélectionner des descripteurs des formes lexicales en fonction du contexte. Par exemple, chaque objet - représentant la SLEV à trois éléments <Préfixe|Base|Suffixe> - est décrit par les distributions de fréquence de l'ensemble des préfixes, des bases et des suffixes, qui peuvent être présents dans chaque contexte analysé.

De plus, on a voulu tenir compte, dans la description de données lexicales symboliques, des relations entre les composantes. En particulier dans ce cas, la composante Base a été analysée par rapport à la Catégorie grammaticale attribuée au terme entier[12]. Ce choix a permis de créer une relation intrinsèque entre la Base et la Catégorie, par rapport aux Préfixes et Suffixes.

4. La procédure de "*text mining*" et la construction du tableau de données textuelles

Comme nous avons déjà dit, l'objectif d'une synthèse, en termes de données symboliques, de la structure lexicale à éléments variables consiste à rendre plus efficace une opération de *text mining* liant l'extraction des données à leur analyse. En effet, l'avantage de notre méthodologie est de traduire le résultat d'une requête sur un *corpus* en un tableau de données complexes directement analysable par la méthode proposée.

L'extraction des données symboliques du *corpus* a été réalisée en définissant tout d'abord les formes appartenant à la Base pour chaque thématique considérée (10 dans notre analyse) et donc en réalisant des requêtes par rapport à chaque contexte. Après avoir sélectionné des termes pour les différents contextes, on a compté les Préfixes et Suffixes en tant qu'éléments variables des structures lexicales (en gardant leur différentes flexions), ainsi que la Catégorie des formes grammaticales (A-"adjectif", N-"substantif", J-"adjectif/substantif",).

On a successivement effectué un comptage de la distribution des différents Bases de chaque contexte par rapport aux trois formes grammaticales considérées.

La structure lexicale a été donc représentée en termes de donnée complexe, considérant les trois variables Préfixe, Base, Suffixe et la Catégorie grammaticale comme des descripteurs modaux. Les distributions de fréquence de tels descripteurs sont obtenues directement par des requêtes sur le corpus. Dans notre application, nous avons considéré 10 thématiques différentes relatives aux contextes géo/historique-politique[13].

Chaque thématique est représentée par un OS, qui occupe une ligne dans le tableau lexical complexe. Ce tableau, dans ses colonnes, contient les descripteurs susmentionnés, en associant leur distributions respectives de fréquence.

La Base est caractérisée par les racines des termes qui définissent le contexte thématique (par ex.: "religions"). Les termes de la Base sont différents pour chaque objet. En revanche,

[12] Par exemple, une expression avec le suffixe *–ismo* appartient à la catégorie N, tandis qu'une expression en *-istico* appartient à la catégorie A. La catégorie grammaticale n'identifie donc pas la nature de la Base qui dans les deux cas peut être dérivée d'un adjectif.

[13] Les 10 thématiques concernent les contextes: centre - libéral, droit, gauche-historique, religions, occident, orient, terrorisme, mouvements philosophiques, célébrités, séparatisme.

Données symboliques	***y_1 Préfixe***	***y_2 Base***	***y_3 Suffixe***	***y_4 Catégorie Grammaticale***
1) sinistra storica (gauche historique)	anti-($f_{1.1.1}$)[1] neo-($f_{1.1..2}$) ex-($f_{1.1.3}$) filo-($f_{1.1.4}$) post-($f_{1.1.5}$) Null_($f_{1.1.6}$)	marx ($f_{1.2.1}$), stalin ($f_{1.2.2}$), comun ($f_{1.2.3}$), social ($f_{1.2.4}$), labur ($f_{1.2.5}$), crax ($f_{1.2.6}$), soviet ($f_{1.2.7}$), statal ($f_{1.2.8}$), solidar ($f_{1.2.9}$), sindacal ($f_{1.2.10}$), collettiv ($f_{1.2.11}$), egualitar ($f_{1.2.12}$), pacif ($f_{1.2.13}$), castr ($f_{1.2.14}$), bolscev ($f_{1.2.15}$), brigat ($f_{1.2.16}$), ...	-ismo ($f_{1.3.1}$) -ismi ($f_{1.3.2}$) -ista ($f_{1.3.3}$) -isti ($f_{1.3.4}$) -iste ($f_{1.3.5}$)	A ($f_{1.4.1}$) J($f_{1.4.2}$) N($f_{1.4.3}$)
…	…	…	…	…
4) religioni (religions)	anti-($f_{4.1.1}$) neo-($f_{4.1.2}$) ex-($f_{4.1.3}$) ...	cristian ($f_{4.2.1}$), protestant ($f_{4.2.2}$), musulman ($f_{4.2.3}$), arab ($f_{4.2.4}$), ebra ($f_{4.2.5}$), ...	-ismo ($f_{4.3.1}$) -ismi ($f_{4.3.2}$) -ista ($f_{4.3.3}$)	A ($f_{4.4.1}$) J($f_{4.4.2}$) N($f_{4.4.3}$)

[1] $f_{i.j.k}$= fréquence relative pour chaque contexte thématique, avec: *i*= donnée symbolique, *j*=variable, *k*=modalité.

les Suffixes et Préfixes se répètent dans les descriptions des OS[14].

La première étape de l'analyse de ce type de données est une transformation du tableau de données structurées en une matrice numérique de codage flou. Les valeurs de codage associées aux différentes modalités de chaque variable sont les fréquences relatives f_{ijk} qui respectent les propriétés suivantes:

$$1.\ f_{i.j.k.} \geq 0; \qquad 2.\ \Sigma_{\kappa}\, f_{i.j.k} = 1$$

La décomposition de chaque donnée complexe dans ses composantes : Préfixe, Base et Suffixe, permet d'étudier la structure du texte (pour chaque thématique) en relation avec les différents éléments constituants, en tenant compte des liens mutuels. L'idée de décrire les données par les distributions de fréquences associées à chacun de ses composants, permet de garder l'information sur l'occurrence des différentes modalités de chaque composante et donc de reconnaître les modalités qui sont les plus caractéristiques des différentes composantes de la structure lexicale[15].

La Catégorie grammaticale, retenue à côté des trois éléments de la structure de la donnée, joue un rôle d'information supplémentaire qui lie entre elles ces composantes.

Ce type de représentation permet de synthétiser simplement l'information extraite sur les composantes de la structure à éléments variables à l'aide des OS. En outre, il rassemble en un seul tableau, les relations entre les composantes et leurs différences dans les 10 contextes. Ce type d'information, généralement, n'est pas mise en évidence dans les tableaux des formes simples, sauf à l'aide d'une classification appropriée.

Une mesure de la caractérisation (*imprinting*) du texte peut donc déjà être déduite à partir d'une description des différentes thématiques sur la base de la distribution de leur propres composantes. Au niveau comparatif, une caractérisation du texte, en fonction des différentes thématiques, peut être effectuée par comparaison entre tous les couples de données symboliques. De telles mesures permettent d'évaluer la proximité entre les structures relevées dans les différents contextes. Pour atteindre cet objectif, on peut utiliser les mesures de similarité proposées dans l'analyse de données symboliques (De Carvalho et al., 1998).

[14] Dans la description proposée, il est possible d'introduire des règles de compatibilité internes à la structure, soit de nature grammaticale soit liées au contexte, en forme de règles logiques qui réduisent l'espace de description.

[15] On peut introduire un élément supplémentaire qui retient le poids de chaque modalité sur la base des occurrences. Par ex.: *anti-* est présent dans 13 expressions, avec un "poids" global de 641 occurrences.

Au niveau global, la caractérisation du texte peut être effectuée par une analyse factorielle sur données complexes, qui permet de considérer les relations entre les différentes modalités des composantes des données, en relation avec la Catégorie grammaticale et avec le contexte en question.

5. Corpus et base de données de référence

Le corpus analysé est constitué par dix années du quotidien "La Repubblica" (1990-1999) déchargées à partir des Cd-rom respectifs. La numérisation des textes a été faite une année à la fois par Lexico2 (Salem, 1999), puis on a réuni les vocabulaires afin de construire la matrice des sous-occurrences annuelles. Ce faisant, on a constitué une immense base de données de vocables, qui dépasse 270 millions d'occurrences avec environ 300.000 formes graphiques différentes, dont quelques noms propres (10.675) et à l'exclusion des <non mots> (erreurs d'orthographe, sigles, ...).

A partir de la matrice *mots x ans*, on a extrait une première base de données de référence contenant tous les mots définis par les préfixes pré choisis (*anti/ex/filo/neo/post/ultra*) ou par les suffixes du trait sémantique étudié (*ismo/ista/istico/iano*). Cette base de données a produit un vocabulaire de 31.627 formes diverses, mais elle contient aussi des expressions telles que <antimeridiano, filosofia, posteriore, ciclismo>, qui n'appartiennent pas aux OS recherchés. Sur cette base de données, on a développé la deuxième étape de *text mining* afin d'extraire, selon l'approche des données symboliques, les structures lexicales concernent les 10 thématiques à caractère géo/historico-politique. A partir d'autant de requêtes que d'OS, on a extrait 2.428 expressions différentes et 10 séries de 4 distributions de fréquence (Préfixe, Base, Suffixe, Catégorie), sur lesquelles on a développé des analyses factorielles non standard.

6. Analyse factorielle sur données complexes

L'objectif de l'analyse factorielle sur données complexes de nature textuelle est de représenter sur un plan les relations entre les composantes de la structure lexicale (dans notre cas: Préfixe, Base, Suffixe), par rapport aux Catégories grammaticales de l'expression constituant la structure, aussi bien qu'aux différentes thématiques considérées.

Ceci permet, dans une perspective de *text mining*, une visualisation des différents contextes thématiques et une synthèse de la composition des SLEV, fondée sur la distribution des composantes à l'intérieur de chaque contexte considéré.

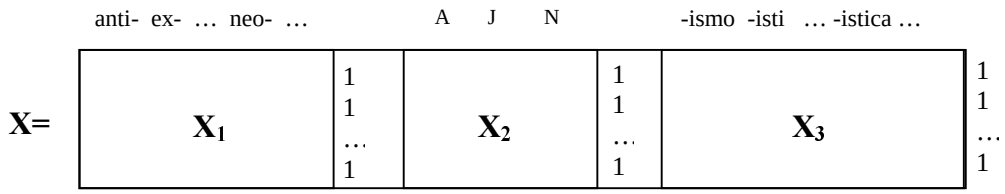
La technique d'analyse proposée ici fait partie d'une méthode d'analyse numérique pour le traitement des données symboliques. Cette méthode est fondée sur une transformation numérique de la description symbolique obtenue à l'aide de systèmes de codage adaptés.

L'analyse est conduite sur la matrice **X**, obtenue par juxtaposition des matrices de codages flous des composantes: Préfixe, Suffixe et Catégorie grammaticale des lemmes.

Pour limiter les dimensions du tableau de données, les éléments faisant partie de la Base seront projetés dans le sous-espace constitué par l'analyse sur trois composantes (cf. § 6.2). Les valeurs de codage en **X** sont les fréquences relatives associées aux différentes modalités des composantes retenues.

La matrice **X** peut être construite en juxtaposant les 3 matrices $\mathbf{X_1}$, $\mathbf{X_2}$ et $\mathbf{X_3}$, respectivement des Préfixes, des Catégories grammaticales et des Suffixes. Ces matrices de codage présentent la caractéristique d'être à valeurs positives ou nulles et la somme des modalités respectives est égale à l'unité. Divisant la fréquence marginale des colonnes de chaque matrice $\mathbf{X_j}$ par le

nombre d'objets (en ligne), on obtient une sorte de distribution moyenne pour chaque composante.



Pour les différents *termes* de la Base à l'intérieur d'une même donnée symbolique, on calcule les fréquences de la catégorie grammaticale. Les tableaux $\mathbf{Z_1,\ldots,Z_{10}}$ sont les matrices de codage des bases, caractéristiques pour chacun des 10 contextes retenus, par rapport aux catégories grammaticales.

Les valeurs de codage sont les fréquences relatives mêmes:

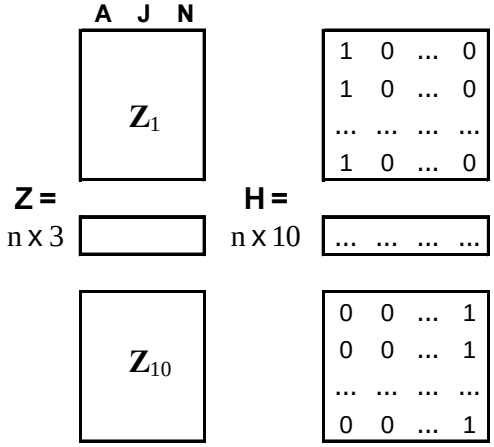
$\mathbf{Z_1}$=

BASE	CAT GRAMM		
	A	J	N
solidar	0.7	0.2	0.1
sindacal	0	0.5	0.5
collettiv	0.2	0.4	0.4
….	…	…	…

La matrice **Z** (n × 3) est constituée par la concaténation verticale des 10 matrices $\mathbf{Z_1},\ldots,\mathbf{Z_{10}}$, sur les colonnes des mêmes modalités du descripteur: Catégorie grammaticale. La dimension *n* de la matrice **Z** équivaut au total des différentes bases considérés dans l'analyse, pour tous les contextes: $n=n_1+\ldots+n_{10}$, où n_i (i=1,…,10) est la cardinalité de l'*i-ème* contexte.

Les marges des matrices $\mathbf{Z_i}$, normalisées par les respectifs n_i, coïncident avec les lignes de la matrice $\mathbf{X_2}$.

Soit: H ($n \times 10$) la matrice indicatrice (0/1) d'appartenance des bases à chaque contexte et $\mathbf{D}=\mathbf{H'}\square\mathbf{H}$ la matrice diagonale des n_i (i=1,…,10); la matrice $\mathbf{X_2}$ peut, donc, être écrite comme: $\mathbf{X_2}=\mathbf{D}^{-1}\,\mathbf{H'Z}$



Pour représenter sur un plan factoriel, les différentes composantes de la structure lexicale, on utilise l'Analyse des Correspondances Multiples, dans le sens d'une Analyse Canonique Généralisée – ACG – (Volle, 1985), généralisée au cas de matrices en codage flou (Verde, 1994). Cette approche a été déjà proposée dans le contexte de l'Analyse de Données Symboliques (Verde, 1999; Lauro *et al*. 2000).

6.1 – La technique proposé recherche les axes d'inertie maximale comme solutions de l'équation caractéristique:

$$1/(3\times n)\ \mathbf{X}'\Delta^{-1}\mathbf{X}\mathbf{u}_\alpha=\lambda_\alpha\mathbf{u}_\alpha \qquad \text{(avec: } \mathbf{u}_\alpha'\mathbf{u}_\alpha=1\ \mathbf{u}_\alpha'\mathbf{u}_{\alpha'}=0 \text{ ; où } \alpha\neq\alpha')$$

La matrice Δ^{-1} est une matrice diagonale en blocs, où chaque bloc est: $(\mathbf{X}_j'\mathbf{X}_j)^{-1}$ (j=1,2,3).

On introduit la matrice Δ^{-1} pour résoudre le problème de la singularité de la matrice $\mathbf{X}'\mathbf{X}$. De plus, en choisissant les matrices extra-diagonales $(\mathbf{X}_j'\mathbf{X}_h)^{-1}$ égales à des matrices nulles, on a imposé, implicitement, l'indépendance entre les descripteurs.

La matrice $\mathbf{X}'\Delta^{-1}\mathbf{X}$, peut être aussi écrite comme la matrice somme des opérateurs de projection orthogonaux $\mathbf{P}_{\mathbf{X}j} = \mathbf{X}_j(\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'$ sur le sous-espace engendré par les colonnes des $\mathbf{X}_j$:

$$\mathbf{X}'\Delta^{-1}\mathbf{X}= \Sigma_{j=1,2,3}\,\mathbf{P}_{\mathbf{X}j.}$$

C'est pour cette relation que l'analyse peut être interprétée comme une extension de l'ACG: les vecteurs orthogonaux u_α sont des synthèses globales des vecteurs de synthèse maximale de chaque sous-espace vectoriel E_j, engendré par les colonnes des matrices $\mathbf{X}_j$ (j=1,2,3) (Verde, 1999). Le nombre de solutions possibles (λ_α et $\mathbf{u}_\alpha$) de l'équation caractéristique dans notre analyse, est égal à 9, une fois éliminée la solution avec la valeur propre banale (λ_1=1).

Les coordonnées des OS sont directement obtenues par :

$$\psi_\alpha = \sqrt{\lambda_\alpha}\,\mathbf{u}_\alpha$$

Les coordonnées des modalités des composants "Préfixe, Suffixe, Catégorie grammaticale", sont obtenues par la relation qui lie les valeurs propres des deux sous-espaces ($\Re^{10}$ et $\Re^{\mathbf{K}}$, où *K* est le nombre des modalités totales des 3 composantes analysées):

$$\varphi_\alpha = \Delta^{-1/2}\mathbf{X}'\mathbf{u}_\alpha$$

où $\Delta^{-1/2}$ est la matrice en blocs diagonaux d'éléments $(\mathbf{X}_j'\ \mathbf{X}_j)^{-1/2}$.

De la même façon, il est possible d'obtenir les coordonnées des modalités de chaque composante comme:

$$\varphi_{\alpha,\mathbf{j}} = (\mathbf{X}_\mathbf{j}'\mathbf{X}_\mathbf{j})^{-1/2}\mathbf{X}_\mathbf{j}'\mathbf{u}_\alpha$$

6.2 – A partir de la construction de la matrice H des distribution des catégories grammaticales des bases codifiées dans la matrice $\mathbf{X}_2$, on peut projeter les racines présentes dans les bases des différents objets analysés, en calculant les coordonnés de la manière suivante:

$$\varphi_{\alpha,2} = (\mathbf{X}_2'\mathbf{X}_2)^{-1/2}\mathbf{X}_2'\mathbf{u}_\alpha = (\mathbf{X}_2'\mathbf{X}_2)^{-1/2}\mathbf{Z}'\mathrm{H}\mathbf{D}^{-1}\mathbf{u}_\alpha$$

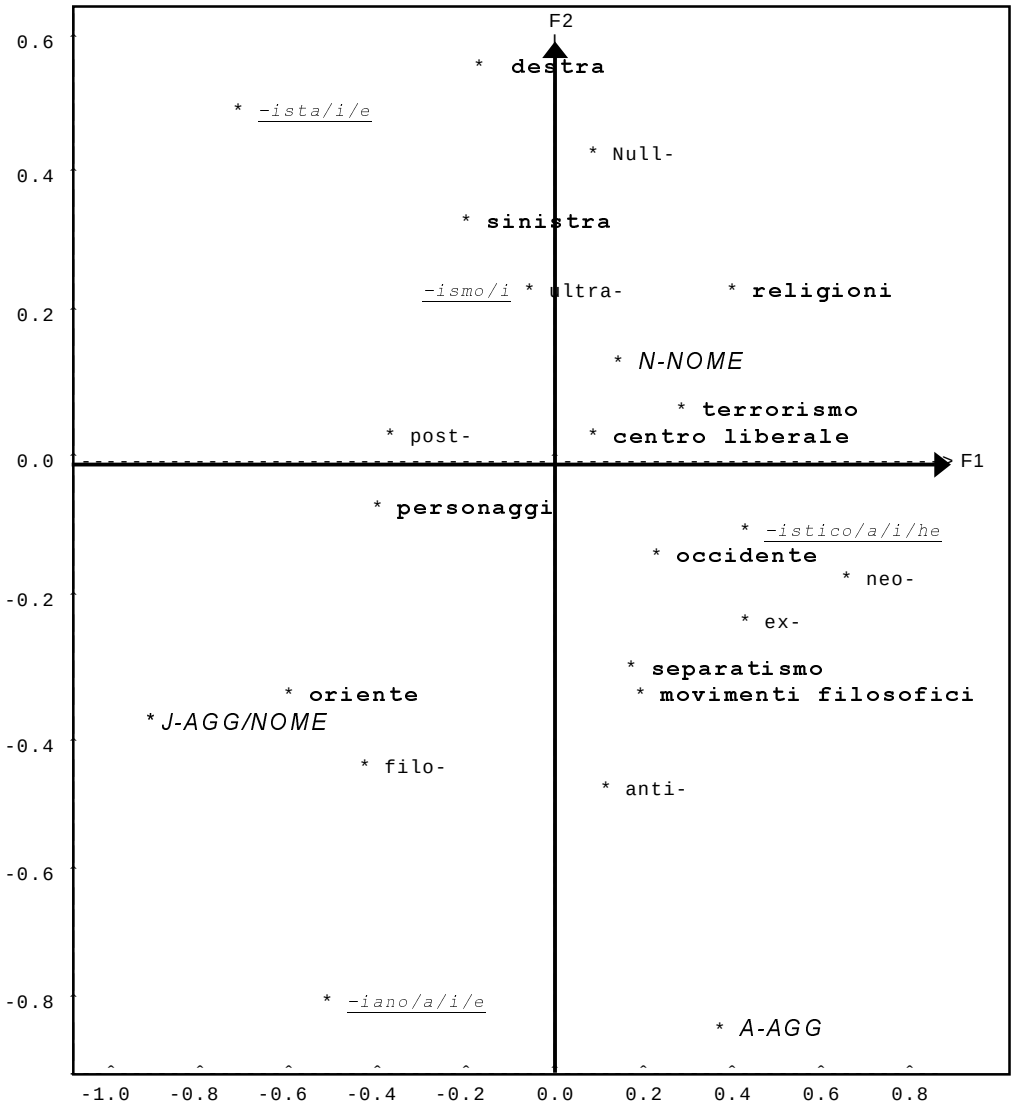
d'où: $\xi_\alpha = \mathbf{Z}\,(\mathbf{X}_2'\mathbf{X}_2)^{1/2}\varphi_{\alpha,2} = \mathbf{Z}\mathbf{X}_2'\mathbf{u}_\alpha$

7. Application

L'analyse développée sur les matrices de codage flou associées aux représentations des OS relatifs aux dix thématiques tirées du corpus "La Repubblica", révèlent une nette différence sur le premier plan factoriel, entre les suffixes *-iano*, *-istico* et *-ista*, et une similitude entre *-ista* e *-ismo*. Ces derniers apparaissent corrélés sur le premier axe avec *-iano*. Ils sont opposés tous les trois à *-istico*. On note que le premier facteur oppose, sur le versant négatif : Droite, Gauche,

Célébrités et Orient (OS se référant tous à des courants de pensée: voir les préfixes *filo-* et *post-*), à Religions, Terrorisme, Occident, Séparatisme et Mouvements (OS moins liées à des idéologies spécifiques) sur le versant positif.

Le deuxième axe sépare au contraire l'AGG (*-iano*) du NOME (*-ismo*, *-ista*), en opposant respectivement sur le versant négatif, les caractères géopolitiques (Orient, Occident, Séparatisme) ainsi que les Mouvements philosophiques, et sur celui positif, des éléments plus étroitement politico-religieux (Droit, Gauche, Religions). Pour ces derniers, on peut noter des préfixes comme *Nul-* (l'absence est due aux termes tels que fascisme, communisme, bouddhisme) o *ultra-* (préfixe emphatique).



Les 10 OS sont bien séparés sur le plan factoriel, en opposant les thèmes plus fortement liés à la "politique" (droite, gauche, centre-liberal, terrorisme et religion) et les autres "géo" (orient, occident, séparatisme) ou miscellanées (célébrités et mouvements philosophiques).

Les préfixes sont également bien distingués: on peut noter, parmi d'autres, l'association *filo-* et Orient, *anti-* et Séparatisme, *ex-/neo-* et Occident. Dans notre exemple, il faudrait prendre quelques précautions avant d'approfondir les différences entre OS, parce que les formes élémentaires qui les définissent, présentent quelques ambiguïtés: fascisme et antifascisme appartiennent au même objet, même si, en réalité, ils représentent des points de vue opposés.

Références

Balbi S., Giordano G.(2000). "Un'analisi dei dati testuali con informazioni esterne: le definizioni di qualità", in: M. Rajman & J. C. Chappelier (eds.) *JADT 2000 5es Journées internationales d'Analyse statistique des Données Textuelles*, Ecole Polytechnique Federale de Lausanne, Lausanne.

Bock, H. H., Diday, E. (eds.), (2000). *Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data*. Studies in Classification, Data Analysis and Knowledge Organisation, Springer-Verlag.

Bolasco S. (1998). "Meta-data and Strategies of Textual Data Analysis: Problems and Instruments", in Hayashi et al. (eds.) *Data Science, Classification and Related Methods*, (proceedings V IFCS - Kobe, 1996) Springer-Verlag Tokio, pp. 468-479.

Bolasco S. (2000). Déclarations et répliques gouvernementales dans le discours parlementaire italien, deux genres discoursifs. *Mots*, 64, p. 97-112

Bolasco S. (2001). Integrazione statistico-linguistica nell'analisi del contenuto in B. Mazzara (a cura di) *Metodi qualitativi in psicologia sociale*, Carocci Ed. Roma.

Chouakria, A., Verde, R., Diday, E., Cazes, P. (1996). Généralisation de l'analyse factorielle des correspondances multiple à des objects symboliques. In Proc. *Quatriemes Journées de la Societé Francophone de Classification,* Vannes.

De Carvalho, F.A.T., Souza, R. M. C. (1998). Statistical proximity functions of Boolean symbolic objects based on histograms. In: Rizzi, A., Vichi, M., Bock, H.-H. (Eds.): *Advances in Data Science and Classification*, Springer-Verlag, Heidelberg, 391-396.

Diday, E. (1987). Introduction à l'approche symbolique en analyse des données. *Première Journées 'Symbolique-Numerique'.* CEREMADE, Université Paris IX Dauphine, 21-56.

Diday, E. (1989). Knowledge representation and Symbolic Data Analysis. In Proc. *2nd Inter. Workshop on Data, Expert Knowledge, and Decision*. Hamburg.

Elia A. (1984). *Le verbe italien*, Schena-Nizet, Paris

Elia, A. (1996) "Per filo e per segno: la struttura degli avverbi composti" in E. D' Agostino (ed.) *Sintassi e semantica*, ESI, Napoli, pp. 167-263

Gross M. (1999). Nouvelles applications des graphes d'automates finis à la description linguistique, *Linguisticae Investigationes*, Tome XXII Vol. Spécial 1998/1999, p. 249-262.

Lauro N. C., D'Ambra L. (1984). "L'analyse non symétrique des correspondances", E. Diday et al. (eds.), *Data Analysis and Informatics*, III, Amsterdam, North-Holland, 433-446.

Lauro, N.C., Verde R., Palumbo F. (2000). "Factorial Discriminant Analysis on Symbolic Objects" (2000) in *Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data*. *Studies in Classification, Data Analysis and Knowledge Organisation*. Bock, H.H., Diday, E. (Ed.), Springer Verlag. Heidelberg, 212-233.

Salem A. (1999). *Lexico2* - Guide on line.

Silberztein M. (1999). Les Graphes Intex, *Linguisticae Investigationes*, Tome XXII Vol. Spécial 1998/1999, p. 3-29.

Verde R. (1994). *Funzioni B-spline e codifica delle variabili quantitative nell'analisi non lineare dei dati* – Collana del DMS, Università Federico II - Serie di ricerca – Rocco Curto editore, Napoli .

Verde, R. (1999). "Generalised Canonical Analysis on Symbolic Objects" In *Classification and Data Analysis, Theory and Application.* Vichi M., Opitz O. (eds), Springer-Verlag, Heidelberg, p. 195-202.

Volle M. (1985), *Analyse des données,* ed. Economica, Collection "Economie et statistique avancées", Paris.