TALTAC: un environnement pour l'exploitation de ressources statistiques et linguistiques dans l'analyse textuelle. Un exemple d'application au discours politique.

Sergio Bolasco
DSGSSAR - Università di Roma " La Sapienza " - bolasco@scec.eco.uniroma1.it

Abstract

In this paper we present a strategy for textual analysis. We show differents steps as: a) text preliminary normalization, b) finding of polyforms useful to identify the principal lexias and to disambiguate some simple words, c) comparison with frequency dictionaries to find out the peculiarities of corpus vocabulary; d) lexicalization of frozen expressions and lemmatization of principal verbs; e) analysis of latent syntagms by means of some typical lexico-textual units of corpus. The integrated use of software TALTAC and INTEX realize this strategy. Un example from a study of political discourse is illustrated.

Résumé

Dans cette communication on esquisse une stratégie statistique pour l'analyse textuelle, en montrant les apports: a) de la normalisation préliminaire du corpus; b) de l'identification des polyformes pour une reconnaissance correcte de certaines lexies et la désambigüisation de certaines formes simples; c) de la comparaison avec les dictionnaires de fréquence pour extraire le langage caractéristique du corpus; d) de la lexicalisation des expressions figées typiques et de la lemmatisation opératoire des principaux verbes; e) de l'étude des syntagmes latents à travers des unités mixtes (formes textuelles) significatives. L' utilisation intégrée des logiciels TALTAC et INTEX poursuit cette stratégie, décrite ici à l'aide d'une application au discours politique en vue de comparer les langages écrits et parlés.

Mots-clés: stratégie d'analyse textuelle, unités mixtes, dictionnaires de fréquence, connexion lexicale, logiciels

1. Introduction

Malgré l'enrichissement récent des fonctions présentes dans les logiciels pour l'analyse statistique de données textuelles[1], jusqu'à présent on n'a pas tracé une stratégie d'action pour le traitement des textes, qui puisse assurer une bonne désambiguation des termes et, en même temps, une sélection du langage caractéristique dans le vocabulaire d'un corpus. La solution à ce problème, comme je le soutiens depuis quelques années, se trouve dans la mise en correspondance directe d'instruments de types différents, c'est-à-dire: mesures statistiques et ressources linguistiques.

C'est ce que nous avons réalisé dans TALTAC (Bolasco et al., 1999), en regroupant dans un seul environnement: opérations de normalisation, calculs lexicométriques, sélection de polyformes, exploitation du tagging morphologique lexico-grammatical, connexion lexicale. En particulier dans TALTAC on sélectionne les termes caractéristiques du vocabulaire à l'aide d'un e comparaison avec un langage de référence, on lexicalise certaines expressions et phrases figées considérées comme cruciales pour le texte et qui ont un fort degré d'absorption par rapport aux mots composants. De cette manière, on procède à la désambiguation de certaines

[1] Parmi les logiciels les plus connus, on peut citer: Alceste, ATLAS, DBT, Hyperbase, INTEX, LEXICO, NUDIST, SPAD_T, SPHINX, TROPES, VIVO.

formes simples grâce à la reconnaissance des mots composés et on peut disposer, pour l'analyse de contenu, d'unités lexico-textuelles (ULT) de type mixte.

2. Intégration statistico-linguistique

On sait bien que dans le domaine de la recherche qualitative, il est difficile d'identifier une démarche unique pour l'analyse, valable en toutes circonstances. Au contraire, chaque étude requiert des choix, tout au long de la recherche, avec de fréquents effets de feed-back. Néanmoins, il est nécessaire de respecter la hiérarchie qui existe entre certaines étapes de l'analyse textuelle et de considérer les conséquences qui peuvent en découler si l'on ignore ces contraintes.

Dans ce cadre, il peut être intéressant d'utiliser plusieurs logiciel "en parallèle", c'est-à-dire ouverts simultanément[2] sur le même corpus: ce qui permet une forte intégration entre les aspects statistiques et linguistiques, et une optimisation des choix évoqués ci-dessus, tout au long de l'étude.

Le but final, du point de vue statistique, c'est une analyse de contenu sur les tableaux de fréquences, sans seuil de fréquence préétabli pour la sélection des ULT: même des hapax peuvent entrer dans ces tableaux. Concrètement, ces unités de basse fréquence appartiendront à des regroupements lexico-grammaticaux ou sémantiques (par exemple, en tant que modaliseurs spatio-temporels). Ceci permet la construction de matrices compactes obtenues en opérant en même temps une classification des fragments du corpus et un regroupement des ULT selon ces catégorisations. Une nouvelle méthode pour l'analyse de ces matrices a été récemment proposée par Balbi et Giordano (1999).

2.1 Logique propre à TALTAC

Le logiciel TALTAC représente un environnement homogène pour le Traitement Automatique Lexico-Textuel pour l'Analyse de Contenu, à travers l'étude du corpus selon des unités de type mixte (ULT). Dans ce cadre, on peut intervenir sur le corpus même, en disposant de son vocabulaire sous formes graphiques et d'un inventaire, éventuellement partiel, de ses segments répétés (obtenus par Lexico ou Spad).

À partir de quelques statistiques de base sur le vocabulaire du corpus (gammes de fréquence, richesse lexicale, seuil), en suivant une démarche ciblée, TALTAC met à disposition un corpus segmenté dans des unités mixtes, importantes pour une représentation du texte et il produit un vocabulaire, transparent quant aux transformations effectuées.

La logique qui nous a amené à concevoir ce programme a été de ne pas se substituer aux logiciels statistiques déjà disponibles, mais de réaliser la meilleure intégration possible entre informations statistiques sur le corpus à étudier et ressources externes (linguistiques et statistiques) disponibles "a latere", afin d'optimiser les opérations que chacun est obligé de développer dans le pré-traitement du texte d'une manière approximative et subjective, sans garantie d'être exhaustif et sans connaissance linguistiques. Au contraire, nous voudrions rendre automatique la plus part de ces opérations tout au long de l'analyse.

Dans notre exemple, outre TALTAC, nous nous sommes servis de INTEX et LEXICO. INTEX a permis de valider les mots du corpus à l'aide des dictionnaires linguistiques (corrections des erreurs d'orthographe et comptage des occurrences des locutions grammaticales déjà connues grâce au système DELACF, cf. par. 2.2), et de réaliser un étiquetage morphologique des

[2] Notre expérience montre qu'utiliser en même temps INTEX (www.ladl.jussieu.fr; Rebooah et Silberztein, 1999), TALTAC et Lexico produit un enrichissement remarquable de l'analyse textuelle.

principales parties du discours. LEXICO a servi à l' indexation des formes graphiques et des segments et à l' analyse des concordances de certains lexèmes.

TALTAC a été conçu sans format propre et travaille avec les standards les plus courants de traitement automatique des informations: textes en mode <.txt>, listes en mode <.txt avec tabulation>, dictionnaires de fréquence en format <MS/Access> (avec des requêtes dans SQL). Les résultats obtenus sur le corpus et sur les listes sont transparents (.txt) afin d' en rendre immédiat l' échange des données avec les autres logiciels.

2.2 Logique propre à INTEX

INTEX contient, parmi ses ressources, un système intégré de dictionnaires électroniques DELAxx (à deux niveaux : formes et polyformes) et un ensemble de graphes-automates, utiles à la définition de grammaires locales, et de transducteurs à états finis (FST), utiles à la normalisation et à la catégorisation du texte[3]. L'u tilisation de chaque dictionnaire ou graphe produit différentes lectures du texte au niveau de l' analyse documentaire.

Les dictionnaires des formes simples permettent, par désambigüisation, une lemmatisation préliminaire du texte (avec ou sans reconnaissance des polyformes déjà détectées). Cette opération peut être nettement améliorée à l' aide de quelques grammaires locales personnalisées[4]. INTEX privilégie le silence par rapport au bruit, en étiquetant seulement les termes non ambigus.

Les dictionnaires des polyformes facilitent la levée d' ambiguïtés des certains mots simples ambigus et permettent le relevé de lexies importantes pour l' analyse textuelle[5]. On peut attribuer un triple étiquetage aux polyformes: i) morphologique (catégorie grammaticale de la structure et celle de ses composantes: par exemple, <carte de crédit>=N+NPN; informations sur la flexion: invariable, genre, nombre ou autre), ii) sémantique (humaine, locatif, temporel ou autres modalisations), iii) statistique (usage). Ces catégorisations peuvent constituer des critères pour la mise au point des matrices compactes.

La création des FST permet d' identifier et d' étiqueter quelques concordances complexes (*quasi-segments*) dont la plupart proviennent d'h apax dans le texte, mais qui, une fois étiquetées par les FST, peuvent être prises en compte dans l' analyse de contenu, du fait de leur appartenance à une classe préalablement définie.

Grâce à INTEX il est donc possible de saisir certains ensembles de *segments* (ou *quasi-segments*) *non-répétés*, sans choisir un seuil de fréquence préalable.

3. Une stratégie pour l'analyse textuelle

Nous allons décrire quelques étapes parmi les plus usuelles de la procédure suivie par TALTAC pour une meilleure analyse textuelle.

A) La première étape, dont dépend la qualité de l' étude automatique des textes, estla *normalisation préliminaire* du corpus. Cette étape ne peut être généralisée dans ses détails, parce qu' elle est organisée en fonction des objectifs de la recherche et du type de textes. Parmi les opérations d' intérêt général on peut citer: le relevé des erreurs orthographiques, la standardisation des graphies (noms, abréviations, sigles, nombres et dates), l' éventuelle

[3] "Formellement, un transducteur est un automate dont les transitions sont étiquetées par des couples de symboles (*Sr* / *Sp*), où *Sr* est un symbole reconnu et *Sp* un symbole produit" (Silberztein 1993, p. 16).
[4] Pour le français, cf. Anne Dister (1999).
[5] Un dictionnaire, administré en première étape, permet d' optimiser certaines levées d' ambigüités. Par exemple les toponymes, les sigles et les noms (propres ou de célébrités) figurent parmi les dictionnaires le plus souvent utilisés en priorité afin d' exclure les cas d' homographie avec des mots communs.

transformation des majuscules au commencement de phrase, la gestion des accents et de l' apostrophe, la définition des caractères alphabet/délimiteur[6].

B) Les mesures lexicométriques de base sur le corpus nous permettent d' avoir une idée concrète de l' étendue du corpus (N) et du vocabulaire (V), de la gamme des fréquences, du seuil éventuellement utile à l' analyse de contenu. Une remarque, à propos du seuil: pour des corpus comportant plus de 20.000 occurrences, en considérant le premier décile des plus basses fréquences (Bolasco, 1990) comme limite pour obtenir un taux satisfaisant de recouvrement du texte, on peut vérifier qu'un tel choix, dans la majorité des cas, correspond à la fréquence moyenne N/V dans l' ensemble du corpus (arrondie à l' entier le plus proche: Bolasco 1999, p. 205), ce qui représente un critère non trivial de choix.

C) L'*identification des polyformes* pour une reconnaissance correcte de certaines lexies et une désambigüisation de certaines formes simples, constituent la troisième étape de la stratégie.
Une telle identification permet la levée automatique des ambiguïtés pour une proportion importante des occurrences des mots polysémiques très fréquents. Les dictionnaires électroniques d'INTEX pour les polyformes contiennent, comme nous l' avons déjà dit, la classification de la polyforme et celle de ses composants. La première, qui implique un nouveau calcul des occurrences du corpus, peut être utilisée pour l' analyse textuelle, la seconde pour l' analyse lexicale. En général, la localisation des polyformes peut s' effectuer à deux niveaux de complexité: phrases figées (segments) et structures avec éléments variables (quasi-segments).

a) Les phrases figées peuvent être repérées grâce aux dictionnaires de fréquence. Le choix d' identifier préalablement les expressions qui appartiennent au lexique fondamental des polyformes, FdP (Bolasco et Morrone, 1998), permet à TALTAC d'obtenir des fréquences nettes pour les mots communs dont les occurrences seraient sinon surévaluées dans le corpus. Ainsi peut-on approcher au mieux le vrai signifié des termes dans la logique d'une analyse textuelle. La comparaison avec le FdP permet de repérer également les segments non répétées déjà connus. Parmi ces derniers, figurent les hapax retrouvés par INTEX (à travers le transfert de FdP dans un dictionnaire de type DELACF) qui représentent, en général, entre 50% et 70% du total, selon le type et la taille du corpus. Grâce à leur regroupement, en classes prédéfinies (par exemple, les adverbes en tant que modaliseurs de quantité, temps, espace, manière etc.), on gagne en information et l'on enrichit l' analyse qui couvre une proportion plus importante du texte.

Pour ce qui concerne les groupes nominaux, TALTAC grâce à l' indice IS permet de sélectionner les quelques dizaines de polyrhèmatiques[7] parmi les centaines ou milliers de segments répétés répertoriés par les logiciels statistiques. Rappelons que l' indice IS mesure l' absorption des composantes dans une polyforme, ce qui met en évidence normalement les structures nominales les plus caractéristiques du corpus (Morrone, 1993).

Le calcul de cet indice permet aussi de choisir les polyformes à lexicaliser, aussi bien grammaticales (d' après FdP) que de contenu (groupes nominaux), comme l'on verra plus bas au paragraphe 5.
b) Les structures à éléments variables (quasi-segments) peuvent être identifiés dans INTEX, grâce à des transducteurs ad hoc[8].

---

[6] Cette opération élimine les espaces: chaque mot est un enregistrement; le texte est lisible verticalement, comme un index.

[7] Expressions ou lexies dont le signifié est de type non compositionnel.

[8] Dans ce logiciel, le repérage des expressions est indépendant de la complexité de la structure qui les caractérise. Certaines expressions, particulièrement importantes pour l' analyse textuelle, peuvent être détectées par des graphes *ad hoc* conçus autour des mots "pivot". On passe du cas particulier (niveau des mots) au cas général (niveau des catégories), par substitution des éléments correspondants dans le graphe. On peut aussi

D) A ce stade, il peut être utile de réaliser une *lemmatisation* grammaticale du texte. Elle peut être faite par INTEX, travaillant en parallèle avec TALTAC, à l' aide du module *Disambiguation*. Sans utilisation de graphes particuliers (les grammaires locales), une telle catégorisation est par définition non exhaustive[9]. Cependant, cela ne perturbe pas beaucoup les analyses de contenu, du fait qu' il s' agit d'un e étape subsidiaire, d'un ou til intermédiaire. Nous nous servons surtout de cette information pour la phase ultérieure de comparaison des listes. On notera cependant qu'un e lemmatisation partielle porte à sous-estimer les lemmes S+.

Parmi d' autres catégorisations morphologiques, celle concernant la suffixation peut se révéler particulièrement intéressante (Bolasco, 2000). Elle s' ajoute à l' étiquetage des flexions des parties du discours (comme les genre et nombre pour les noms et adjectifs, ou les temps et personnes pour les verbes). On dispose pour cela de l' indexation *inverse* des listes importées dans TALTAC.

E) La comparaison avec les *dictionnaires de fréquence* permet d' extraire le vocabulaire caractéristique du corpus. C' est surtout intéressant pour les verbes, mais aussi important pour les autres catégories. Du fait que les dictionnaires de fréquence portent sur des mots simples, ce niveau d' analyse lexicale doit être effectué avant la reconnaissance des polyformes, mais il serait souhaitable de pouvoir le réaliser aussi sur ces polyformes dans le futur.

En d' autres termes, dans les dictionnaires de fréquence disponibles aujourd'hu i, il existe beaucoup d' ambiguïtés, grammaticale et sémantique. Lorsqu'on aura admis l'h abitude de reconnaître les phrases figées, ces comparaisons pourront être effectuées après lexicalisation.

4. Les dictionnaires de fréquence dans TALTAC

Dans notre stratégie, une importance particulière est accordée aux lexiques (généralistes ou sectoriels), actualisées dans les dictionnaires de fréquence. En général, ces derniers sont à retenir en tant que paradigmes de divers codes linguistiques, utiles à la sélection du langage caractéristique (Bolasco, 1996b). Cela est possible en comparant de telles "distributions théoriques" avec le vocabulaire du corpus analysé. Dans ce but, on dispose également des mesures de connexion lexicale (Muller, 1992) et des écarts réduits sur les différences d'usage normé.

Ces dictionnaires sont à la disposition de la communauté scientifique depuis l'o rigine, mais malheureusement, dans la plupart des cas, seulement sur papier! Il n' est pas encore courant d' extraire les informations de ces bases de connaissance sur l'u sage des mots. Par exemple, cela permet de sélectionner les mots suremployés (S^+) ou sous-employés (S^-) dans notre corpus par rapport au modèle, c' est-à-dire le vocabulaire caractéristique du champ auquel appartient le corpus. Une fois sélectionnés, par exemple, les mots (S^-), il est important de revenir au texte pour interpréter dans quels contextes se trouvent ces mots *rares*. Cela permet d' envisager également les hapax.

Nous nous sommes efforcés de rassembler en un seul endroit les principaux dictionnaires de fréquence que la tradition linguistique italienne a produit durant les 20 dernières années sur notre langue observée empiriquement.

TALTAC contient, parmi d'autres dictionnaires de fréquence, le VdB: Vocabulaire de Base (De Mauro, 1980); le LIF: Lexique de Fréquence de l'Italien, (Bortolini *et al.* 1971); le LIP:

---

augmenter le degré de généralisation en cherchant les expansions et/ou les insertions possibles dans la structure de base.

[9] La plus part des flexions verbales et adverbiales sont bien identifiées, ce qui représente un intérêt prioritaire pour les étapes ultérieures. D' ailleurs, une analyse des concordances effectuée en parallèle par exemple avec Lexico, permet de repêcher certaines flexions de fréquence élevées, théoriquement ambiguës, mais pas dans le corpus, repérant ainsi une partie non négligeable d' occurrences.

Lexique de l'Italien Parlé (De Mauro *et al.*, 1993); le VELI: Vocabulaire Electronique de la Langue Italienne (IBM, 1989); le LE: Lexique Elémentaire (Marconi *et al.*, 1993); le VFLI: Vocabulaire Fondamental de la Langue Italienne (Sciarone, 1995). Aussi bien que les résultats de nos travaux précédents, tels que le FdP: Lexique Fondamental de Polyformes (Bolasco et Morrone, 1998); le DPG: lexique du Discours Programmatique de Gouvernement (Bolasco, 1996a); le LEF: Lexique Economique et Financier, (Bolasco, en préparation).

Le logiciel peut aussi inclure n'importe quel nouveau lexique (soit en formes simples, soit en polyformes).

5. La construction opérationnelle des formes textuelles

Une phase ultérieure de la stratégie TALTAC consiste à étudier la *lexicalisation des expressions* figées, c' est-à-dire à considérer chaque lexie complexe, comme une nouvelle seule occurrence. Cette opération est limitée aux lexies les plus typiques, extraites par comparaison avec le lexique des polyformes, FdP, ou sélectionnées grâce à l' indice IS.

Dans cette phase opératoire il peut être aussi utile de construire des "équivalences" qui concernent souvent les diverses flexions d'un v erbe, parmi ceux précédemment sélectionnés.

Ces deux dernières étapes comportent une *re-tokenisation* du texte et une re-indexation du corpus. A partir de ce moment, le texte est optimisé et prêt pour l' analyse de contenu.

6. Un cas d'étude: différences entre langage écrit et parlé

Nous avons appliqué certaines étapes de cette stratégie dans une étude sur le discours politique italien. À partir de l'analyse des Déclarations parlementaires des Présidents du Conseil des Ministres, en particulier du lexique du discours programmatique gouvernemental (Bolasco, 1996a), on met en relation la partie fondamentale de ce langage, avec son équivalent oral: les Répliques gouvernementales au débat parlementaire (Bolasco, 2000).

Pour des raisons d' espace, on se limite ici à illustrer quelques résultats obtenus sur les polyformes caractéristiques de ces deux types de langage. On isole quelques catégorisations sur les structures adverbiales et nominales qui caractérisent certains paradigmes du discours gouvernemental.

La comparaison de l' inventaire des segments (au seuil 5) dans les Déclarations et dans les Répliques, réalisé avec FdP, permet d' isoler clairement les deux genres de discours. Parmi les 16.000 segments des Déclarations et les 9.400 des Répliques, on reconnaît respectivement 930 et 700 polyformes appartenant au langage fondamental courant (le bilan avec INTEX en texte intégral serait beaucoup plus élevé). Le calcul de l' indice IS sur les mêmes inventaires originaux des segments permet de filtrer les expressions candidates à la lexicalisation, dont la majorité sont des groupes nominaux. Le tableau 1 montre la répartition des 100 premières polyformes avec les plus forts écarts réduits par rapport à FdP et avec les valeurs les plus grandes de IS. On peut noter que seulement une petite proportion de segments sont communs aux deux outils de filtrage; cela confirme une capacité différente de sélection entre les deux: le FdP sélectionne les locutions l' IS, les groupes nominaux. Simplement à partir du bilan quantitatif des effectifs classés, on peut déjà découvrir certaines "distances" entre les deux discours[10]: la référence dominante aux personnes et l' abondance des locutions dans les Répliques par rapport aux contenus des programmes ou aux référents parlementaires cités dans les Déclarations gouvernementales (Programmes).

[10] Cette distance pourrait être mesurée à l' aide d' une récente proposition de connexion intertextuelle (Labbé et Monière, 2000).

Tab. 1 - Filtrage des 100 segments plus importants à l'aide de la comparaison avec le lexique FdP et de l'indice IS

	Programmes		*Répliques*	
Catégories	IS	FdP	IS	FdP
Partis politiques	8	2	9	3
Personnes	5	2	49	6
Entités parlementaires	38	16	17	11
Contenus de programme	43	44	16	31
Locutions grammaticales	-	35	-	45
Verbes idiomatiques	6	1	9	4
	100	100	100	100
Polyformes en commun	16		11	

Un deuxième apport aux diversités des discours vient de l' étude des adverbes (tab. 2). La comparaison avec le FdP a fait ressortir un nombre remarquable d' adverbes composés: deux fois plus que les adverbes simples, toujours au seuil de fréquence 5; cette différence serait encore plus nette si l'on considérait tous les adverbes, jusqu' aux hapax, et certains quasi-segments. On peut noter que la majorité des adverbes simples sont de manière, alors que parmi les composés on retrouve plutôt de modaliseurs temporels, spatiaux et quantitatifs.

La connexion lexicale des adverbes entre les Déclarations et les Répliques est de 72%, mais les vocabulaires propres, concernant les adverbes, sont aussi bien différents: dans les Répliques, on trouve surtout ceux de manière ou de jugement (autre), dans les Déclarations ceux de temps et lieu.

Tab. 2 – Classes de modalisation des adverbes et connexion lexicale entre Répliques et Programmes

Adverbes	Manière	Lieu	Temps	Quantité	Autre	Total	% Total
% Adv. simples	89.8	3.9	2.9	1.5	1.9	100	35.6
% Adv. composés	42.2	12.1	28.2	8.1	9.4	100	64.4
% Total	59.2	9.2	19.2	5.7	6.7	100	100.0
Effectifs	342	53	111	39	33	578	
Programmes	37.4	15.2	38.4	5.1	4.0	100	17.1
Répliques	54.2	6.8	22.0	5.1	11.9	100	10.2
En commune (P∩R)	65.0	8.1	14.3	6.0	6.7	100	72.7

Pour l' étude des verbes nous renvoyons à une autre contribution (Bolasco, 2000) où l'on montre, grâce à des sélections ciblées de ULT, que le discours des Répliques comporte de nombreuses références à l' Assemblée parlementaire et que l'on y p arle plus volontier à la première personne.

Références

Balbi S., Giordano G. (1999). A Factorial Technique for Analysing Textual Data with External Information. In Book of short papers of *CLADAG99*, Classification and Data Analysis Group Italian Statistical Society. CNR, Roma, p. 285-288.

Bolasco S. (1990), Sur différentes stratégies dans une analyse des formes textuelles: une expérimentation à partir de données d' enquêteIn Bécue M., Lebart L., Rajadell N. editors, Actes des Premières journées *JADT*, UPC, Barcelone, 1992, pp. 69-88.

Bolasco S. (1996a). Il lessico del discorso programmatico di governo. In Villone, M., Zuliani A. editors, *L'attività dei governi della repubblica italiana (1947-1994)*. Bologna, Il Mulino.

Bolasco S. (1996b). Meta-data and Strategies of Textual Data Analysis: Problems and Instruments. In Hayashi et al. editors, *Data Science, Classification and Related Methods*, Springer-Verlag Tokio, 1998, pp. 468-479.

Bolasco S. (1999). *Analisi multidimensionale dei dati*, Carocci ed., Roma.

Bolasco S. (2000). Les Répliques aux Déclarations gouvernementales: une particularité du discours parlementaire italien, *Mots*, n. 62 (à paraitre).

Bolasco S., Morrone A. (1998). La construction d'un lexique fondamental de polyformes selon leur usage. In Mellet S. editor, *JADT 1998*. Nice 19-21 febbraio 1998, Univ. Sophie Antipolis de Nice, pp. 155-166.

Bolasco S., Morrone A., Baiocchi F. (1999), A Paradigmatic Path for Statistical Content Analysis Using an Integrated Package of Textual Data Treatment. In Vichi M., Opitz O. editors, *Classification and Data Analysis. Theory and Application*. Springer-Verlag, Heidelberg.

Bortolini N., Tagliavini C., Zampolli A. (1971). *Lessico di frequenza della lingua italiana contemporanea*. Garzanti, Milano.

De Mauro T. (1980). *Guida all' uso delle parole*Roma: Editori Riuniti.

De Mauro, T., Mancini, F., Vedovelli, M., Voghera, M. (1993) *Lessico di frequenza dell' italiano parlato*. Milano, EtasLibri.

Dister A. (1999) De l' étiquetage traditionnel au transducteur du texte: Intex et la levée d' ambiguité par grammaires locales. *Revue Informatique et Statistiques dans les Sciences Humaines*, n. 1-4.

IBM (1989). *VELI: Vocabolario Elettronico della Lingua Italiana*. Centro di Ricerca, Roma.

Labbé D., Monière D. (2000). La connexion intertextuelle. Application au discours gouvernemental québécois. In *JADT2000*, Actes des 5e Journées Internationales d' Analyse Statistique des Données Textuelles, LIA, EPFL, Lausanne.

Marconi L, Ratti D. et al. (1993). *Lessico Elementare. Dati statistici sull' Italiano Scritto e Letto dai bambini delle elementari*. Bologna, Zanichelli.

Morrone A. (1993). Alcuni criteri di valutazione della significatività dei segmenti ripetuti. In S. J. Anastex editors, *JADT 1993*, ENST-Telecom, Paris, pp. 445-53.

Muller Ch. (1992), *Principes et méthodes de statistique lexicale*, Champion, Paris (Réimpression: de l' édition Hachette, 1977).

Rebboah C., Silberztein M., (1999) *Intex*. ASSTRIL.

Salem A. (1987). *Pratique de segments répétés*. Klinchsieck, Paris.

Sciarone A. G. (1995). *Il vocabolario fondamentale della lingua italiana*. Guerra Ed., Perugia.

Silberztein M. (1993). *Dictionnaires électroniques et analyse automatique de textes : le système INTEX*. Masson, Paris.