Electronic Dictionaries and Linguistic Analysis of Italian Large Corpora

Simonetta Vietri and Annibale Elia
24 March, 1999

Word Count 2 429

Abstract

In this paper we will show how Italian electronic dictionaries have been built within the methodological framework of Lexicon-grammar We will see the structure of electronic dictionaries of simple and compound words, and we will show how to analyse texts employing these linguistic tools within INTEX, a morphological analyser Finally, we will show how electronic grammars (built with INTEX) interact with dictionaries and allow recognition of sequences of simple and compound words in large corpora

0. Introduction

We present the system of Italian morphological dictionaries (the DELI system) which has been developed at the Department of Communication Science of the University of Salerno We will see how these dictionaries can be employed in order to index a text Finally, we will examine the construction of local grammars which, interacting with dictionaries, allow precise tagging of sequences of words

1. The Italian Electronic Dictionaries

The DELI system contains several electronic dictionaries of simple words and of compound words The electronic dictionary of simple words, named DELAS, contains about 100 000 Italian entries to which an alphanumerical code has been assigned This code refers to the grammatical category of the word and to its inflectional paradigm What follows is an example of the DELAS dictionary

dottore N80
cortese A79
amare V3
di PREP
lentamente AVV

the noun (**N**) *dottore* is given above in masculine singular canonic form The adjective (**A**) *cortese* is in the masculine singular canonic form Verbs (**V**) are listed in the infinitive form, as *amare* Those items which do not inflect are assigned a code indicating only the grammatical category, as above shown for the preposition *di* and the adverb *lentamente* The numerical code refers to specific inflectional algorithms For example, code 80, associated to nouns, corresponds to the endings

N80

ms	**fs**	**mp**	**fp**
-e	-essa	-i	-esse

thus indicating that all nouns as *dottore*, i e *campione*, *professore*, etc , inflect by adding to the root *-e* for the masculine singular (**ms**), *-essa* for the feminine singular (**fs**), *-i* for the masculine plural (**mp**) and *-esse* for the feminine plural (**fp**) On the other hand, adjectives encoded A79, as *cortese* but also *tribale*, etc , can be described by the following inflectional model

A79

ms	**fs**	**mp**	**fp**
-e	-e	-i	-i

in which the masculine and feminine singular (**ms** and **fs**) on one side, and the masculine and feminine plural (**mp** and **fp**) on the other, correspond to the homographic forms *cortese* and *cortesi*

The algorithm for verbs is more complicated since it has to refer to 40 forms referring to simple tenses Therefore, verbs like *amare*, *abbandonare*, *imparare*, etc , are encoded as V3 which corresponds to the following endings and grammatical values

V3
ind/pr(3o,3i,3a,3iamo,3ate,3ano)
imp(3avo,3avi,3ava,3avamo,3avate,3avano)
pass r(3ai,3asti,3ò,3ammo,3aste,3arono)
fut s(3erò,3erai,3erà,3eremo,3erete,3eranno)
imperat(-,3a,3i,3iamo,3ate,3ino)
cong/pr(3i,3i,3i,3iamo,3iate,3ino)
imp(3assi,3assi,3asse,3assimo,3aste,3assero)
cond/pr(3erei,3eresti,3erebbe,3eremmo,3ereste,3erebbero)
part/pr(3ante,3anti)
pass(3ato,3ata,3ati,3ate)
ger/pr(3ando)

For example, the first line of the list above indicates that in order to form the Indicative Present (ind/pres) of the verb *amare,* it is necessary to delete the last three characters of the infinitive form of the verb and add *o* for the first singular person, *i* for the second singular person, *a* for the third singular person, *iamo* for the first plural, *ate* for the second plural, and finally *ano* for the third plural

Using DELAS and its inflection codes, a software allows to automatically generate all inflected forms The result will be the electronic dictionary of inflected forms of Italian simple words, named DELAF, which has the following structure

ama,amare V3 Imper2s
ama,amare V3 IndPres3s
amai,amare V3 IndPass1s
amammo,amare V3 IndPass1p
amando,amare V3 Ger
amano,amare V3 IndPres3p
amante,amare V3 PartPres ms fs
amanti,amare V3 PartPres mp fp
amare V3 Inf
amarono,amare V3 IndPass3p
amasse,amare V3 CongImp3s
amassero,amare V3 CongImp3p
amassi,amare V3 CongImp1s
amassi,amare V3 CongImp2s
amassimo,amare V3 CongImp1p
amaste,amare V3 CongImp2p
amaste,amare V3 IndPass2p
amasti,amare V3 IndPass2s
amata,amare V3 PartPass fs
amate,amare V3 Imper2p
amate,amare V3 PartPass fp
amate,amare V3 IndPres2p
amati,amare V3 PartPass mp
amato,amare V3 PartPass ms
amava,amare V3 IndImp3s
amavamo,amare V3 IndImp1p
amavano,amare V3 IndImp3p
amavate,amare V3 IndImp2p
amavi,amare V3 IndImp2s
amavo,amare V3 IndImp1s
amerà,amare V3 IndFut3s
amerai,amare V3 IndFut2s
ameranno,amare V3 IndFut3p
amerebbe,amare V3 CondPres3s
amerebbero,amare V3 CondPres3p
amerei,amare V3 CondPres1s
ameremmo,amare V3 CondPres1p
ameremo,amare V3 IndFut1p
amereste,amare V3 CondPres2p
ameresti,amare V3 CondPres2s
amerete,amare V3 IndFut2p
amerò,amare V3 IndFut1s

ami,amare V3 Imper3s
ami,amare V3 CongPres1s
ami,amare V3 CongPres2s
ami,amare V3 CongPres3s
ami,amare V3 IndPres2s
amiamo,amare V3 Imper1p
amiamo,amare V3 CongPres1p
amiamo,amare V3 IndPres1p
amiate,amare V3 CongPres2p
amino,amare V3 Imper3p
amino,amare V3 CongPres3p
amò,amare V3 IndPass3s
amo,amare V3 IndPres1s
cortese,cortese A79 fs
cortese A79 ms
cortesi,cortese A79 fp
cortesi,cortese A79 mp
di,di PREP
dottore N80 ms
dottoressa,dottore N80 fs
dottoresse,dottore N80 fp
dottori,dottore N80 mp
lentamente,lentamente AVV

The DELAS-DELAF dictionaries contain simple words which are formally defined as *sequences of characters between blank spaces or separators* On the other hand, dictionaries of compound words contain those words which are formally defined as *sequences of words, they contain spaces or separators* Compound words are constrained sequences of words which can have either a metaphorical meaning as *cavallo di battaglia*, which means "something at which somebody particularly excels, somebody's favourite piece" or a "neutral" or technical meaning as *carta di credito*, in English *credit card* Compound words are constrained sequences of words, since the substitution of lexical elements within the sequence with synonyms most of the time produces unacceptable compounds, as the following examples show

(cavallo + *puledro) di battaglia
cavallo di (battaglia + combattimento)

(carta + *tessera) di credito
carta di (credito + *fido)

Furthermore, it is not possible to change the second occurrence of the noun in the plural form

*cavallo di battaglie
*carta di crediti

The only morphological variation can be applied to the head of the whole compound, that is the first noun occurrence

cavalli di battaglia
carte di credito

The electronic dictionary of compound words, named DELAC, contains about 50 000 Italian entries to which grammatical codes have been assigned These codes refer to the grammatical category to which the item belongs and to its internal structure and morphological behaviour In the following examples of the DELAC dictionary

cavallo di battaglia,N+NPN ms-+
carta di credito,N+NPN fs-+
pesce spada,N+NN ms-+
casa madre,N+NN fs-+
agente speciale,N+NA ms++
alta carica N+AN fs-+

the compound items are followed by a symbol of part of speech (N), the separator "+" is followed by the internal structure of the compound The first two items are formed by a Noun, a Preposition and a Noun (NPN), the third and fourth items are formed by two nouns (NN), the fifth item is formed by a Noun and an Adjective (NA), while the last item is formed by an Adjective and a Noun (AN) Columns are followed by the gender and number the examples are either masculine singular (ms) or feminine singular (fs)

Finally, two marks indicate above morphological variations in gender and number If the variations are accepted then the mark is "+", if they are not accepted the mark is "-" The internal structure defines the element of the compound which inflects compounds which belong to the class NPN inflect the first noun, compounds which belong to the classes NA and AN inflect both elements, while compounds which belong to the NN class can either inflect both nouns, as *lingua madre*(fs) - *lingue madri*(fp), or only the first noun, as *pesce spada*(ms) – *pesci spada*(mp) Once we codify the morphological behaviour of compound nouns in such a way, a set of computational routines allows to automatically generate the DELACF, that is the electronic dictionaries of inflected forms of Italian compound nouns, which has the following format

agente speciale N+NA fs++
agente speciale N+NA ms++
agenti speciali,agente speciale N+NA fp++
agenti speciali,agente speciale N+NA mp++
alta carica N+AN fs-+
alte cariche,alta carica N+AN fp-+
carta di credito N+NPN fs-+
carte di credito,carta di credito N+NPN fp-+
casa madre N+NN fs-+
case madri,casa madre N+NN fp-+
cavalli di battaglia,cavallo di battaglia N+NPN mp-+
cavallo di battaglia N+NPN ms-+
pesce spada N+NN ms-+
pesci spada,pesce spada N+NN mp-+

2. The Automatic Morpho-lexical Analysis

Once dictionaries of simple and compound words are built, it is possible to apply them to texts by means of the programme INTEX of morpho-lexical analysis This software has been developed by Max Silberztein and allows to load electronic dictionaries of simple and of compound words structured in the way shown above INTEX applies both dictionaries to a text and builds the dictionary of that text which will contain not only simple words but also all compound nouns present in the text This step allows to recognize and highlight within a text all compounds

- Che storie dicono? - chiedo - Io non so niente So che lei ha un negozio, senza l'**insegna luminosa** Ma non so nemmeno dov'è
Me lo spiega E' un negozio di pellami, valige e **articoli da viaggio** Non è sulla piazza della stazione ma in una via laterale, vicino al **passaggio a livello** dello **scalo merci**

and also to build a frequency list for them

1 articoli da viaggio
1 insegna luminosa
1 passagio a livello
1 scalo merci

Such an indexation is extremely reliable for the management of technical and scientific documentation Technical documents contain a lot of terminology which includes mostly compound nouns INTEX gives us the possibility of loading more than one dictionary, so, the user can build not only a DELACF for generic compounds but also specialized dictionaries of compounds belonging to various fields such as Economy, Engineering, Computer Science, and so on It is then possible to analyze technical texts on the base of such dictionaries The following text in which compounds have been highlighted is an example

drawn fron an article of the Italian economics newspaper *il Sole 24 ore* (the whole article contains 84 lines)

Politica economica, anno zero E l'assenza di impegni precisi e credibili contro l'inflazione continua a tenere in tensione i mercati

Nei mesi che hanno preceduto la svalutazione della lira e' stato ripetutamente e autorevolmente affermato che la stabilità del cambio rappresentava l'asse portante di tutta la politica economica italiana La linea di condotta seguita nelle prime settimane dal Governo Amato era sembrata coerente con tale enunciazione
1 il riconoscimento dell'autonomia della Banca d'Italia,
2 l'affermazione della priorità assoluta dell'obiettivo anti-inflazionistico,
3 il blocco delle contrattazioni salariali nel settore pubblico,
4 l'accordo di fine luglio con i sindacati

I primi due punti erano fondamentali e complementari, in quanto l'autonomia della Banca centrale riceveva una precisa caratterizzazione (rafforzata dalla prospettiva di adesione all'unione monetaria europea) dalla priorità assoluta dell'obiettivoanti-inflazionistico E' bene sottolineare che tale priorità assoluta valeva anche nei confronti dell'obiettivo di risanamento della finanza pubblica Essa doveva in sostanza essere intesa nei seguenti termini a) la Banca centrale avrebbe rispettato target di crescita monetaria coerenti con gli obiettivi inflazionistici annunciati (cosa che non era fino allora avvenuta, nemmeno dopo l'adesione alla banda ristretta dello Sme, che pure avrebbe dovuto comportare un accresciuto rigore della politica monetaria), b) nel far ciò essa non si sarebbe curata degli effetti di breve periodo di tale condotta sui tassi di interesse, e quindi anche sulla finanza pubblica, c) il Governo avrebbe adottato con la massima urgenza provvedimenti di risanamento della finanza pubblica, senza tuttavia fare ricorso a misure che incidessero sull'indice dei prezzi al consumo Questi richiami hanno ormai sapore storico, ma sono utili per meglio mettere a fuoco la situazione attuale In particolare la prima affermazione (l'essere cioe' la stabilità del cambio asse portante di tutta la politica economica) ci sembra ancora pienamente valida l'asse portante non c'e' più e con esso e' sparita anche la politica economica Schematicamente sembra che esistano due alternative possibili di politica economica La prima (che riproduce in sostanza, pur nelle condizioni modificate, la linea pre-svalutazione tratteggiata sopra) potrebbe articolarsi nei seguenti termini a) aggiustamento fiscale come obiettivo della massima urgenza, b) riaffermazione del ruolo autonomo della Banca centrale, degli impegni assunti in vista del mercato unico (in particolare in materia di libera circolazione dei capitali), c) massimo contenimento delle spinte inflazionistiche derivanti dalla svalutazione della lira e riaffermazione di un obiettivo anti-inflazionistico preciso (tradotto in un target rigoroso e impegnativo, e quindi credibile, di crescita monetaria) Particolarmente importante quest'ultimo punto, in quanto non e'affatto inevitabile che gli effetti della svalutazione si traducano integralmente in una spinta inflazionistica aggiuntiva Se la crescita monetaria e' tenuta sotto controllo, e grazie agli effetti restrittivi della manovra di bilancio, la svalutazione può tradursi anziche' in un fattore inflazionistico, in uno di mutamento dei prezzi relativi La seconda linea punta invece a un abbassamento dei tassi di interesse e a impartire stimoli espansivi all'economia (entrambi gli obiettivi potrebbero ricevere una motivazione aggiuntiva grazie al sollievo che, nel breve periodo, potrebbero portare alla finanza pubblica) Una tale linea richiederebbe certamente l'accantonamento,almeno temporaneo, di qualsiasi obiettivo anti-inflazionistico E' anzi probabile che i suoi effetti più significativi e durevoli sarebbero quelli prodotti dall'inflazione sulla distribuzione del reddito, e soprattutto della ricchezza, e sul valore reale dello stock del debito pubblico Si noti che la seconda linea e' certamente incompatibile con un ritorno in tempi brevi a un regime di cambio fisso

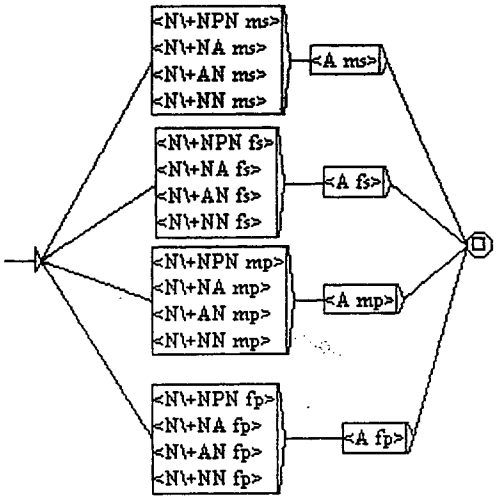
A frequency list which contains terminological compound nouns give us the possibility to immediately understand the specific content of this article Such an index provides a picture of the content of the text (the index which follows was built on the whole article)

8 politica economica
5 finanza pubblica
4 politica monetaria
4 crescita monetaria
3 priorità assoluta
3 tassi di interesse
3 Banca centrale
3 asse portante
2 valore reale
2 stabilità del cambio
2 svalutazione della lira
2 manovra di bilancio

2 tasso di inflazione
1 debito pubblico
1 abbassamento dei tassi
1 distribuzione del reddito
1 differenziali di interesse
1 titoli pubblici
1 cambio fisso
1 stabilizzazione del cambio
1 premio di rischio
1 settore pubblico
1 unione monetaria
1 contrattazioni salariali
1 Politica economica
1 linea di condotta
1 Banca d'Italia
1 libera circolazione
1 spinte inflazionistiche
1 spinta inflazionistica
1 mercato unico
1 indice dei prezzi
1 prezzi al consumo

3. Local Grammars

Electronic dictionaries give us the possibility of recognizing within texts words and sequences of words as defined by dictionaries INTEX allows to recognize combinations of simple and compound words, thanks to the interaction between dictionaries and grammars INTEX contains a tool which allows to construct local grammars on the model of finite state automata These grammars can be based not only on words but also on the non-terminal symbols contained in the dictionaries For example, in order to identify all compound nouns followed by an adjective which agrees in gender and number with them, we construct the following grammar



If we apply such a grammar to a text, INTEX will highlight all occurrences of this pattern and subsequently construct concordances for that pattern

trovare espressione sia in accresciuti differenziali di interesse reali, sia in un deprezzamen
utta la politica economica italiana La linea di condotta seguita nelle prime settimane dal Gov
ppresentava l'asse portante di tutta la politica economica italiana La linea di condotta segui
anti-inflazionistica Se questa e' la politica economica italiana si sarebbe tentati di dire
zione si traducano integralmente in una spinta inflazionistica aggiuntiva Se la crescita monet
apitali), c) massimo contenimento delle spinte inflazionistiche derivanti dalla svalutazione de
vi sufficientemente saldi per tollerare tassi di interesse penalizzanti a qualche asta,accompag
una scelta realmente impegnativa 2 Il tasso di inflazione programmato e' stato portato dal 3,
rzata dalla prospettiva di adesione all'unione monetaria europea) dalla priorità assoluta dell'

Hence, electronic dictionaries on one side, and the possibility of constructing grammars which interact with dictionaries on the other give us the possibility of automatically analyzing large corpora, considering not only words but also sequences of words

Bibliography

Gross M (1988), *Grammaire transformationelle du francais Syntaxe de l'adverbe* Paris Cantilène

Gross, M & M Silberztein (edd), *Actes des Premières Journées INTEX* (21-22 mars 1996), Paris L A D L., Université de Paris 7

Elia, A (1984), *Le verbe italien Les completives dans les phrases a un complement*, Fasano-Paris Schena-Nizet.

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