Electronic Dictionaries and Linguistic Analysis of Italian Large Corpora

Simonetta Vietri and Annibale Elia

Dipartimento di Scienze della Comunicazione, Università di Salerno
elianni@tin.it

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.

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	cortese,cortese.A79:fs
ama,amare.V3:IndPres3s	cortese.A79:ms
amai,amare.V3:IndPass1s	cortesi,cortese.A79:fp
amammo,amare.V3:IndPass1p	cortesi,cortese.A79:mp
amando,amare.V3:Ger	di,di.PREP
amano,amare.V3:IndPres3p	dottore.N80:ms
amante,amare.V3:PartPres:ms:fs	dottoressa,dottore.N80:fs
amanti,amare.V3:PartPres:mp:fp	dottoresse,dottore.N80:fp
amare.V3:Inf	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 (carta + *tessera) di credito
cavallo di (battaglia + combattimento) 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 **N**oun, a **P**reposition and a **N**oun (NPN); the third and fourth items are formed by two nouns (NN), the fifth item is formed by a **N**oun and an **A**djective (NA), while the last item is formed by an **A**djective and a **N**oun (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):

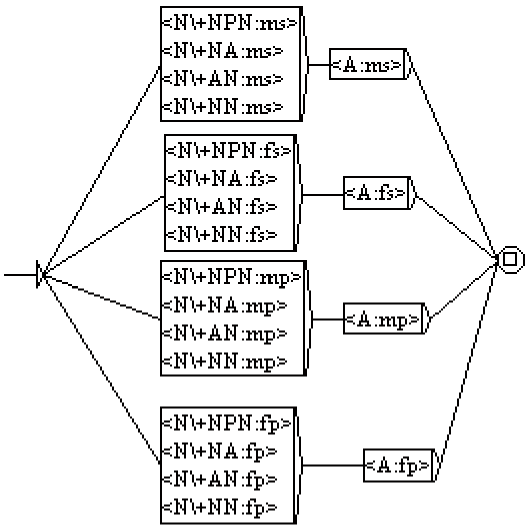
> <u>Politica economica</u>, anno zero. E l' assenza di impegni precisi e credibili contro l' inflazione continua a tenere in tensione i mercati.
>
> Nei mesi che hanno preceduto la <u>svalutazione della lira</u> e' stato ripetutamente e autorevolmente affermato che la <u>stabilità del cambio</u> rappresentava l'<u>asse portante</u> di tutta la <u>politica economica</u> italiana. La <u>linea di condotta</u> seguita nelle prime settimane dal Governo Amato era sembrata coerente con tale enunciazione:
>
> 1 il riconoscimento dell'autonomia della <u>Banca d'Italia</u>;
> 2 l'affermazione della <u>priorità assoluta</u> dell'obiettivo anti-inflazionistico;
> 3 il blocco delle <u>contrattazioni salariali</u> nel <u>settore pubblico</u>;
> 4 l'accordo di fine luglio con i sindacati.
>
> …

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
...

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.

4. Other samples

Other electronic dictionaries and local grammars give us the possibility of recognizing complex insertion of simple or compound words into idiomatic sentences.

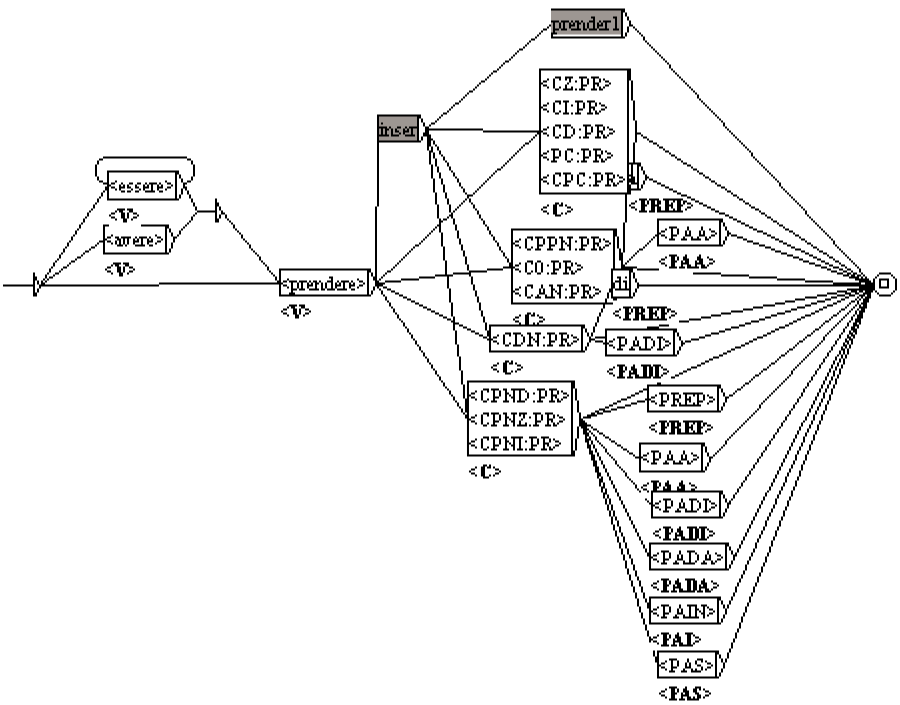
If we observe the frozen sentences with the verb prendere (170 sequences) we can construct a classification of 10 classes:

Class	Sample
CZ	*prendere aria*
CD	*prendere il largo*
CI	*prendere un granchio*
CPC	*prendere la palla al balzo*
PC	*prendere N a calci*
CDN	*prendere atto di N*
CAN	*prendere il cuore a N*
CPND	*prendere le mosse da N*
CPNZ	*prendere gusto a N*
CPPN	*prendere la palla di mano a N*

We construct a dictionary of noun phrases

alla/rete,alla/rete.PC:PR
animo,animo.CZ:PR
aria,aria.CZ:PR
armi/e/bagagli,armi/e/bagagli.CPC:PR
atto,atto/di.CDN:PR
bianco/per/nero,bianco/per/nero.CPC:PR
…

and a local grammar:



IdiomPrendere

If we take in account the following discourse:

Maria, dopo aver preso la sua amica con le mani nel sacco, ha deciso che fosse necessario prendere immediatamente la situazione di petto.

we have a morpho-syntactic analysis of the text:

amica	*amicare*	V8:Q2s
amica	*amicare*	V8:X3s

amica	*amico*	A87:fs
amica	*amico*	N87:fs
avere	*avere*	N608:ms
avere	*avere*	V1:I
con	*con*	PREP
con	*con*	PX
con le mani nel sacco	*con le mani nel sacco*	PC:PR
deciso	*decidere*	V34:Ums
deciso	*deciso*	A88:ms
di	*di*	N602:fp
di	*di*	N602:fs
di	*di*	N601:mp
di	*di*	N601:ms
di	*di*	PREP
di	*di*	PX
di petto	*di petto*	PC:PR
dopo	*dopare*	V3:X1s
dopo	*dopo*	AVV
dopo	*dopo*	CONG
dopo	*dopo*	N608:ms
dopo	*dopo*	PREP
ha	*avere*	V1:X3s
ha	*ha*	ESC
ha	*ha*	N601:mp
ha	*ha*	N601:ms
immediatamente	*immediatamente*	AVV
la	*la*	DET41:fs
la	*la*	N601:mp
la	*la*	N601:ms
la	*lo*	PRON88:fs
le	*la*	DET41:fp
le	*le*	PRON607:fp
le	*lo*	PRON88:fp
mani	*mani*	N606:mp
mani	*mano*	N49:fp
nel	*nel*	PA306:ms
petto	*petto*	N7:ms
prendere	*prendere*	V25:I
preso	*prendere*	V25:Ums
preso	*preso*	A88:ms
sacco	*sacco*	N10:ms
situazione	*situazione*	N46:fs
sua	*suo*	A115:fs
sua	*suo*	PRON115:fs

If we apply the local grammar *IdiomPrendere* we have immediately a good analysis:

> *Maria, dopo* ***aver preso la sua amica con le mani nel sacco****, ha deciso che fosse necessario* ***prendere immediatamente la situazione di petto****.*

The labels of the parsing are the followings:

{Maria,.?}
({dopo,dopare.V3:X1s} + {dopo,.AVV} + {dopo,.CONG} + {dopo,.N608:ms} + {dopo,.PREP})
({avere,.N608:ms} + {avere,.V1:I})
({preso,prendere.V25:Ums} + {preso,.A88:ms})
({la,.DET41:fs} + {la,.N601:mp} + {la,.N601:ms} + {la,lo.PRON88:fs})
({sua,suo.A115:fs} + {sua,suo.PRON115:fs})

({amica,amicare.V8:Q2s} + {amica,amicare.V8:X3s} + {amica,amico.A87:fs} + {amica,amico.N87:fs})
{con/le/mani/nel/sacco,.PC:PR}
({ha,avere.V1:X3s} + {ha,.ESC} + {ha,.N601:mp} + {ha,.N601:ms})
({deciso,decidere.V34:Ums} + {deciso,.A88:ms})
({di,.N601:mp} + {di,.N601:ms} + {di,.N602:fp} + {di,.N602:fs} + {di,.PREP} + {di,.PX})
{prendere,.V25:I}
{immediatamente,.AVV}
({la,.DET41:fs} + {la,.N601:mp} + {la,.N601:ms} + {la,lo.PRON88:fs})
{situazione,.N46:fs}
{di/petto,.PC:PR}
{S}

Bibliography

Chanod, J.P, P.Tapanainen (1994), *Statistical and constraint-based taggers for French*. Technical Report MLTT-016, Rank Xerox Research Centre, Grenoble, France.

Chanod, J.P, P.Tapanainen (1995), *Creating a tagset, lexicon and guesser for a French tagger*, Technical report, Rank Xerox Research Centre, Grenoble, France.

Chrobot, A. (1997), *Tagging statistique sans ambigüités ou tagging partiellement ambigü mais sans erreurs ?*, Technical Report, LADL, Université Paris 7.

D'Agostino E. (1992) *Analisi del discorso*, Napoli: Loffredo.

De Bueriis, G., M.Monteleone (1997), *Lessicografia e dizionari elettronici. Dal dato cartaceo alle basi di dati linguistiche*, in *Annali dell'Università degli Studi di Basilicata* (in stampa), Potenza.

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

Elia, A. (1984a), *Lessico-Grammatica dei verbi a completiva*, Napoli: Liguori.

Elia, A. (1995), *Per filo e per segno: la struttura degli avverbi composti*, in D'Agostino (a cura di), *Tra sintassi e semantica. Descrizioni e metodi di elaborazione automatica della lingua d'uso*, Napoli: ESI.

Elia, A., M. Martinelli, E. D'Agostino (1981), *Lessico e strutture sintattiche*, Napoli:Liguori.

Gross, M. (1975), *Méthodes en syntaxe*, Paris: Hermann.

Gross, M. (1982), *Une classification des phrases figées du français*, in *Revue québécoise de linguistique*, vol. 11, n°2, Montréal: Presses de l'Université du Quebec à Montréal.

Gross, M. (1993), *Local Grammars and their Representation by Finite Automata*, in Hoey M. (a cura di) *Data, Description, Discourse. Papers on the English language in honour of John McH Sinclair*, London: Harpers Collins

Karttunen, L., R. Kaplan, A. Zaenen (1992), *Two-level morphology with composition*, in *Proceedings of Coling 92. The fourteenth International Conference on Computational Linguistics*, Vol. I, Nantes.

Karttunen, L. (1994) *Constructing Lexical Transducers*, in *Proceedings of Coling 94. The fifteenth International Conference on Computational Linguistics*. Vol. I, Kyoto.

Kleene, S.C. (1956), *Representation of events in nerve nets and finite automata*, in C.E. Shannon e J.McCarthy (a cura di), *Automata Studies*, Princeton of University Press.

Koskenniemi, K. (1983), *Two-level morphology. A general computational model for word-form regnition and production*. University of Helsinki.

Lo Cascio, V. (1970), *Strutture pronominali e verbali italiane*, Bologna: Zanichelli.

Markov, A.A. (1913), *Essai d'une recherche statistique sur le texte du roman "Eugène Oneguine"*. Bull. Acad. Imper. Sci. St. Petersburg, 7.

Monteleone, M. (1996), *Survey and testing of Finite State Methods for the Recognition of Agglutinated Sequences in Italian with the Aid of Lexical Information about Possible Combinations of Verbs and Clitics*. Gramlex Project - Technical Report R3C1.

Perrin, D. (1994), *Les Débuts de la Theéorie des Automates*, Gaspard Monge Institut, Université de Marne la Vallée: Noisy le Grand.

Shannon, C.E. (1948), *A mathematical theory of communication*, in C.E. Shannon e W. Weaver (a cura di), *The Mathematical Theory of Communication*.

Shützenberger, M.P. (1965), *On monoids having only trivial subgroups*, in *Information and Control*, 8.

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

Silberztein M. (1997), *INTEX 3.4. Reference Manual, LADL*, Université Paris 7.

Turing, A.M. (1936), *On computable numbers with an application to the Eintscheidungsproblem*, in *Proceedings of the London Mathematical Society*.

Vietri, S. (1985), *Lessico e sintassi delle espressioni idiomatiche*, Napoli: Liguori.

Vietri, S. (1990), *On some comparative frozen sentences in Italian*, in *Linguisticae Investigationes* 14.1, Amsterdam/Philadelphia: John Benjamins.

Vietri, S. (1990a), *La sintassi delle frasi idiomatiche*, in *Studi italiani di linguistica teorica e applicata* XIX:1.

Vietri, S. (1992), *La struttura morfologica elettronica*, Rapporto tecnico n°10 (Ricerca EU 524-Genelex).

Vietri, S. (1994), *Il dizionario elettronico Genelex*, Rapporto tecnico n°7 (Ricerca EU 524-Genelex).

Vietri, S. (1995), *I risultati del testing e aggiornamento del Dizionario di economia*, Rapporto tecnico n°22 (Ricerca EU 524-Genelex).

Vietri, S. (1997), *The syntax of the Italian verb* essere, in *Linguisticae Investigationes* (in stampa), John Benjamins, Amsterdam.