Università degli Studi di Salerno

Dipartimento di Scienze della Comunicazione

# Descrizioni lessicali e dizionari elettronici delle polirematiche di alcuni domini specialistici dell'informazione comunitaria


## Gruppo di ricerca "*Maurice Gross*"

Annibale Elia, Federica Marano, Mario Monteleone, Simona Sabatino, Daniela Vellutino


Presentazione di Daniela Vellutino

# Argomenti

- Metodologia linguistica: Lessico-Grammatica
- Attività del gruppo di ricerca “Maurice Gross”: lingware per la comunicazione pubblica
- Studio del lessico dell’informazione comunitaria
- Codifica lessicografica computazionale per la creazione di dizionari elettronici di dominio
- Prime conclusioni

# Metodologia linguistica: Lessico-Grammatica

- La metodologia linguistica descrittiva di riferimento per il nostro studio è il Lessico-Grammatica (LG), introdotta nella comunità scientifica da Maurice Gross, Università Parigi 7, dopo il 1960 e basata sull'approccio trasformazionale e analitico-combinatorio di Z.S. Harris.

- Scopo principale del Lessico-Grammatica è descrivere dettagliatamente tutti i meccanismi combinatori (morfosintattici, distribuzionali e trasformazionali) indissolubilmente legati alle concrete entrate lessicali.

- Sui principi LG sono stati sviluppati lingware e software per l'analisi testuale automatica: INTEX, UNITEX, NOOJ® e CATALOGA.

# Lessico-Grammatica - LG

Per LG l'unità minima grammaticale e sematica da cui far partire gli studi sul linguaggio naturale è la frase semplice.

**La frase semplice è un contesto grammaticale e semantico formato da un solo elemento predicativo (verbo/nome/aggettivo) e dai suoi complementi essenziali per la costruzione del significato.**

L'analisi delle frasi semplici di una lingua avviene attraverso le regole di restrizione, selezione e distribuzione indotte dai predicati e dalle proprietà trasformazionali di ogni singolo tipo di frase semplice.

# Tipi di combinazioni lessicali: libere e ristrette

LG ha individuato le condizioni di *continuum* per le combinazioni lessicali in sintagmi o frasi, che possono essere di quattro tipi:

° **Combinazioni a distribuzione libera** con un grado elevato di variabilità di co-occorrenza fra le parole con significato composizionale e denotato.

Es. verbo > *guardare* (panorama, pallone, forchetta, Max)
Es. nome > *acqua* (sporca, pulita, torbida)

° **Combinazioni a distribuzione ristretta** con un grado ridotto di variabilità di co-occorrenza fra le parole

Es. verbo > *stendere* (i panni, gambe)
Es. nome > *acqua* (minerale, gassata, naturale)

# Tipi di combinazioni lessicali: fisse e invariabili

° **Combinazioni a distribuzione fissa** con un grado nullo o quasi nullo di variabilità di co-occorrenza fra le parole.

Es. verbo > *alzare il gomito (≠ sollevare il gomito)*
Es. nome > *lingua madre, colletto bianco, casco blu*

Sono combinazioni lessicali a distribuzione fissa le **polirematiche e le frasi idiomatiche. Entrambe sono sintagmi con atomicità semantica (unità di significato).**

° **Combinazioni a struttura invariabile** senza alcuna variabilità di co-occorrenza fra le parole, come nel sintagma dei **proverbi**

Es. *Chi rompe paga e i cocci sono i suoi*

# Polirematiche

Nell'ambito del Lessico-Grammatica lo studio del lessico terminologico riguarda l'uso delle unità lessicali superiori classificabili come polirematiche, vale a dire un "gruppo di parole che ha un significato unitario, non desumibile da quello delle parole che lo compongono, sia nell'uso corrente sia nei linguaggi tecnico-specialistici", come indicato dal dizionario di De Mauro (2000).

Nella lingua comune una parte di queste combinazioni fisse, polirematiche e idiomatismi, molto probabilmente è il residuo di operazioni metaforico-metonimico ormai cristallizzate.

Nel lessico terminologico le combinazioni fisse sono spesso il risultato di un processo neologico denominazionale. In particolare, nel lessico comunitario sono frequenti per designare istituzioni e principii.

# Lessico comunitario

La costruzione di unità lessicali superiori è il meccanismo di formazione che maggiormente caratterizza i termini del lessico comunitario (Nystedt, 2000).

Nel lessico dell'informazione comunitaria le polirematiche sono presenti in tutti i domini di conoscenza repertoriati in forma di glossari specialistici ed altre risorse linguistiche, quali banche dati terminologiche multilingui e thesauri.

Le parole terminologiche tecnico-specialistiche dell'informazione comunitaria sono spesso elencate in glossari istituzionali presenti nei siti web delle PA che, molte volte, le registrano più in virtù della loro maggiore frequenza nei testi piuttosto che tenendo conto di modalità lessicografiche.

Il nostro studio riguarda la descrizione formale utile alla costruzione di dizionari elettronici delle polirematiche dei domini specialistici: *Pari Opportunità*, *Società dell'Informazione* e il macrodominio *Fondi Strutturali*

# Gruppo "Maurice Gross"

- Svolgiamo attività di ricerca per progetti finalizzati alla creazione e allo sviluppo di lingware in cui sono maggiormente utilizzati i metodi di formalizzazione del linguaggio naturale indicati dal Lessico-Grammatica.
- Il lingware sviluppato è costituito da:
  - **Sistema dei dizionari elettronici DELA**
  - **Grammatiche locali**
- Tali risorse linguistiche sono funzionali ai software per l'analisi testuale automatica, per le operazioni di Information Retrieval, utili anche per il trattamento e gestione delle informazioni per la comunicazione pubblica.
- In particolare per gli scopi:
  - Semplificazione del linguaggio amministrativo
  - Traduzione assistita
  - Gestione delle informazioni per i servizi di e-government

# Dizionari elettronici

- I dizionari elettronici sono basi di dati lessicali formalizzate (Machine-Readable-Form) [1], strutturate omogeneamente, in cui le caratteristiche morfo-grammaticali delle entrate (genere, numero e flessione) sono indicate da etichette alfanumeriche univoche e non ambigue.

- **Il sistema dei dizionari elettronici DELA:**
  - DELAS: circa 135 mila parole semplici
  - DELAF: circa 930 mila parole semplici flesse
  - DELAC: circa 154 mila parole composte
  - DELACF: circa 270 mila parole composte flesse (dizionari di dominio)

[1] Il termine "basi di dati" è inteso nella più comune accezione informatica, vale a dire una collezione di dati gestita da un sistema software.

# NOOJ ®

Il Gruppo “Maurice Gross” sta sviluppando il modulo per la lingua italiana di NOOJ ®, sviluppato da Max Silberztein.

NOOJ ® è un software in grado di leggere automaticamente testi digitalizzati, localizzare ed estrarre da essi pattern linguistici in forma di concordanze.

NOOJ ® basa il suo motore sul sistema di dizionari elettronici DELA, nonché sulle tavole sintattiche e sugli automi a stati finiti del Lessico-Grammatica in forma di grafi.

Il sistema DELA per NOOJ ® non prevede più la distinzione tra forme flesse e canoniche perchè il paradigma di flessione è generato da grammatiche morfologiche in cui sono descritte le regole di flessione.

# CATALOGA

CATALOGA è un software per l'analisi testuale automatica ed opera come parser sui testi per rilevare i termini di dominio. Si basa sul matching fra testi e dizionari elettronici di parole composte realizzati in base alle indicazioni linguistico-teoriche del Lessico-grammatica.
In questo modo consente di "catalogare" testi estraendo lemmi codificati lessicograficamente e sematicamente per domini specialistici.
Opera in 4 fasi:

1. Lettura automatica di un testo;

2. Rilevamento e computo dei lemmi repertoriati (entrate canoniche, flesse e varianti)

3. Associazione ad ogni testo che analizza di uno o più domini semantici etichettati;

4. Costruzione di liste di frequenza delle occorrenze e attribuzione ad ogni entrata lessicale riconosciuta un indice d'incidenza statistica che rileva le parole più frequenti e le distribuzioni di frequenza (permette di confrontare i vocabolari dei testi, riscontrando analogie e specificità)

# Costruzione dei dizionari di dominio

- Estrazione polirematiche di dominio da un corpus di 75 testi non tipologizzati (115.557 occorrenze tokenizzate).
- Matching con lemmi registrati in 2 glossari istituzionali.

**Iter 1**
Società dell'Informazione
989 entrate lessicali

- Codifica lessicografica dei termini del glossario comunitario "*100 parole per la parità*".

**Iter 2**
Pari Opportunità
216 entrate lessicali

- Codifica lessicografica dei termini tratti da 15 glossari istituzionali.
- Testing sui corpora tipologizzati e supervisionati da esperti di dominio (3.071.610 occorrenze tokenizzate).

**Iter 3**
Fondi strutturali
1949 entrate lessicali

# Fasi di ricerca

## Prima fase

***Lexical acquisition***, vale a dire individuazione e selezione dei lemmi attraverso i glossari istituzionali contenenti le polirematiche di dominio, registrando anche tutte le varianti lessicali.

## Seconda fase

***Tagging morfo-grammaticale e sintattico***, vale a dire descrizione formale delle polirematiche, flesse nel genere (maschile/femminile) e nel numero (singolare/plurale).

## Terza fase

***Testing sui corpora***, vale a dire che i dizionari elettronici sono testati sui corpora per verficarne la completezza attraverso prima la lettura automatica con i software di analisi testuale e poi la lettura diretta dei testi con esperti di dominio.

# Glossari delle fonti istituzionali comunitarie e nazionali

| | Glossari | LEMMI |
|---|---|---|
| 1 | European Convention Glossary | 59 |
| 2 | Regional Development Glossary | 59 |
| 3 | Environment Glossary | 520 |
| 4 | Europa Glossary | 233 |
| 5 | Common Agricultural Policy | 106 |
| 6 | EU Competition Policy | 130 |
| 7 | European Judicial Network in civil and commercial matters | 34 |
| 8 | CORDIS Research and Development | 110 |
| 9 | Manuale interistituzionale di convenzioni redazionali | 98 |
| 10 | Dipartimento per le Politiche di Sviluppo e Coesione | 233 |
| 11 | Gergo europeo | 80 |
| 12 | "100 parole per la parità" | 100 |
| 13 | Glossario CNIPA | 411 |
| 14 | Glossario Europa Lavoro | 203 |
| 15 | Glossario dell'E-Government di Borruso et al. | 83 |

**TOTALE 2459**

# Varianti lessicali

Nei glossari istituzionali è presente un alto grado di disomogeneità: è stato rilevato l'uso frequente di varianti lessicali che abbiamo registrato nei nostri dizionari di dominio.

| Varianti Ortografiche | Varianti Protogrammi | Varianti d'uso |
|---|---|---|
| Caratteri tutti minuscoli<br><br>Caratteri tutti maiuscoli<br><br>Carattere iniziale in maiuscolo<br><br>Alternanza: caratteri iniziali in maiuscolo e caratteri in minuscolo<br><br>Uso dell'apostrofo | Acronimi<br><br>Esempio: **AdG** variante della poliremática **Autorità di Gestione**<br><br>Esempio: **Programma Operativo Regionale** può avere come varianti l'inizialismo con o senza marca distintiva **POR/P.O.R.** | Europeismi vs. Equivalenti italiani<br><br>Esempio:<br>gender mainstreaming<br><br>prospettiva di genere<br><br>ottica di genere |

# Esempio di formalismo dei dizionari

Stringa con il procedimento di formalizzazione morfo-grammaticale e sintattico per il sistema dei dizionari DELA

carte d'identità elettroniche,carta d'identità elettronica,N+NPNA+f+p+DIE

- Descrizione della codifica lessicografica computazionale
  1. N È il formalismo che indica la funzione grammaticale della poliremática
  2. NPNA sono gli elementi costituenti la struttura sintagmatica = Nome+Preposizione+Nome+Aggettivo
  3. f / p indicano il genere e il numero, in questo caso femminile e flessione plurale
  4. DIE indica l'etichetta terminologica che denomina il nostro dizionario di dominio (Dizionario Informazione Europea)

# Esempio di un estratto da un dizionario elettronico di dominio

```
accordo/sulla/applicazione/delle/misure/sanitarie/e/fitosanita
rie,.N+NPNPNACA:ms+-;DIE
accordo/verticale,.N+NA:ms-+;DIE
acquis/comunitario,.N+NN:ms--;DIE
adesione/di/un/nuovo/stato/all'/unione,.N+NPDANPN:fs-+;DIE
adesione/di/un/nuovo/stato/alla/Unione,.N+NPDANPN:fs-+;DIE
adesioni/di/un/nuovi/stati/all'/unione,adesione/di/un/nuovo/st
ato/all'/unione.N+NPDANPN:fp-+;DIE
adesioni/di/un/nuovi/stati/alla/Unione,adesione/di/un/nuovo/st
ato/alla/Unione.N+NPDANPN:fp-+;DIE
affare/economico/e/finanziario,.N+NACA:ms-+;DIE
affare/finanziario,.N+NA:ms-+;DIE
affare/marittimo,.N+NA:ms-+;DIE
affare/marittimo/e/pesca,.N+NACN:ms-+;DIE
affare/sociale,.N+NA:ms-+;DIE
affari/economici/e/finanziari,affare/economico/e/finanziario.N
+NACA:ms-+;DIE
```

# Osservarzioni sul dominio "Società dell'Informazione"

**Iter 1 - Dominio " Società dell'Informazione"**

Le entrate, estratte manualmente dai corpora, sono state sottoposte al matching con 3 glossari istituzionali specialistici per verificare la percentuale di poliremiatiche comuni.

**I risultati del matching indicano una scarsa sovrapposizione tra le unità estratte manualmente dai testi ed i lemmi dei glossari istituzionali.**

E' stata calcolata una ***precision*** [1] dell'1,61% e una ***recall*** [2] dell'1,62%.

[1] ***precision*** è la misura della correttezza e pertinenza dei termini presenti nel dizionario elettronico specialistico
[2] ***recall*** è la quantità di termini corretti sono in comune con i glossari istituzionali.

# Osservazioni sul dominio "Pari opportunità"

## Iter 2 - Dominio " Pari Opportunità"

Le entrate del glossario istituzionale "100 parole per la parità" sono state codificate per creare un dizionario per l'uso del software CATALOGA.

Attraverso CATALOGA sono state testate su un corpus di testi appartenenti ad uno specifico tipo di testo, il bilancio di genere, che, per le sue finalità pragmatiche, avrebbe dovuto contenere tutte le voci repertoriate nel glossario comunitario.

Corpus BdG contiene 1.523.423 occorrenze tokenizzate, tratte da 22 testi di bilancio di genere di Comuni, Province, Regioni.

Nel glossario istituzionale non sono presenti molti termini rilevati attraverso la lettura dei testi del corpus. La stessa denominazione *bilancio di genere* (1.150 occorrenze); *lavoro di cura* (95 occorrenze), *differenze di genere* (185 occorrenze) e *differenza di genere* (30 occorrenze).

# Osservazioni sul dominio "Fondi Strutturali" - *Work in progress*

**Iter 3 - Dominio " Fondi Strutturali"**

Sono state individuati 15 glossari istituzionali da cui sono stati estratti i termini di dominio che sono stati codificati nel formalismo del sistema dei dizionari elettronici DELA.

Attualmente stiamo usando i software NOOJ e CATALOGA per la fase di testing dei termini estratti sui corpora costituiti da testi scritti, in lingua italiana e in formato digitale, tipologizzati secondo design pertinente ai processi di comunicazione pubblica.

Contestualmente stiamo verficando la completezza dei glossari istituzionali attraverso la lettura diretta dei testi con esperti di dominio. I termini rilevati nei testi e non presenti nei glossari istituzionali saranno formalizzati per il dizionario elettronico.

In seguito sarà eseguito un matching tra termini estratti dai glossari istituzionali e i termini estratti con il supporto degli esperti di dominio per trarre indicazioni sulla completezza dei glossari istituzionali.

# Design corpora per la comunicazione pubblica

Il design corpora è stato definito attraverso una classificazione tipologica dei testi scritti di documenti di domino (in lingua italiana e in formato digitale) basata sulle finalità pragmatiche dei differenti processi di informazione e comunicazione pubblica.

E' monitor corpora di testi classificati in base alle forme di comunicazione pubblica e alle relative tipologie testuali:

- **Comunicazione normativa** → tipo di testo giuridico (fonte normativa primaria o secondaria) → Esempio: Testo "Trattato di Lisbona";

- **Comunicazione amministrativa** → tipo di testo documento di programmazione economica → Esempio: Testo "Complemento di Programmazione";

- **Comunicazione per la pubblica utilità** → tipo di testo FAQ → Esempio: Testo "Richieste d'informazione sui finanziamenti dei Fondi strutturali".

# Patrimonio lessicale del monitor corpora

Il nostro studio, volto ad indagare e repertoriare con modalità lessicografiche computazionali le polirematiche del lessico comunitario, ha previsto la creazione di un monitor corpora di testi scritti, in lingua italiana e in formato digitale, della dimensione complessiva di 3.071.610 occorrenze tokenizzate (aggiornamento aprile 2010).

- **Corpus per la comunicazione normativa contiene 674.746 occorrenze tokenizzate**

- **Corpus per la comunicazione amministrativa contiene 2.390.093 occorrenze tokenizzate**

- **Corpus per la comunicazione di pubblica utilità contiene 6.771 occorrenze tokenizzate**, estratte da 345 testi di frasi di FAQ estratte da 51 siti istituzionali di dominio, relative alle richieste d'informazioni sugli interventi di finanziamento comunitario

# Pilot study: testing sui corpora

- I termini riconosciuti da NooJ nei corpora rappresentano il 3,30% di quelli repertoriati nei glossari istituzionali. Questo dato avvalora la nostra ipotesi:
  - I glossari istituzionali non soddisfano il principio di completezza.

| Tipi di token | Frequenza | Polirematiche dei glossari |
|---|---|---|
| Misura | 9.818 | misura minima<br>misura minima di sicurezza |
| Interventi | 9.691 | criterio d'intervento |
| Progetti | 7.204 | progetti combinati di rst e di dimostrazione<br>progetti Integrati territoriali |

Le parole più frequenti nei corpora sono poco frequenti nelle polirematiche dei glossari istituzionali. Attraverso **l'analisi delle concordanze** con NOOJ, ad esempio, il termine "misura" può candidarsi come costituente nelle polirematiche "misura di politica attiva del lavoro" e "misura di assistenza tecnica" presenti nei corpora, ma assenti nei glossari.

# Base di dati lessicografica dell'Informazione Europea (DIE)

E' stata creata una base di dati lessicografica dell'Informazione Europea che contiene le seguenti informazioni per ogni entrata lessicale:

- Codifica lessicografica computazionale dei termini di dominio. Il dizionario elettronico DIE che comprende i domini osservati e che contiene 3.147 entrate in forma canonica e flessa.

- Definizioni terminologiche con il riferimento alle fonti di attestazione e alla data di rilevazione nei glossari istituzionali

- Riferimenti al testo del corpus in cui l'entrata lessicale è attestata

- Varianti lessicali

# Prime conclusioni

Dalla lettura dei testi dei corpora sono emersi elementi lessicali stabili che designano concetti di dominio che possono per questo diventare termini candidati ad entrare in repertori terminografici istituzionali comunitari (*thesauri*, banche dati terminologiche).

Le risorse linguistiche comunitarie, in particolare i glossari delle fonti istituzionali comunitarie e nazionali dovrebbero essere strumenti di comunicazione verticale volti a favorire la conoscenza dei termini di dominio ad un pubblico più ampio.

Dal nostro studio emergono carenze e la necessità di realizzare un processo di armonizzazione dei termini basato anche su indagini teoriche sulle modalità di creazione lessicale accompagnate a studi sociolinguistici che attestino i concreti usi lessicali nei diversi processi di comunicazione pubblica.

La politica linguistica dell'Unione europea per essere concretamente indirizzata al multilinguismo deve maggiormente sviluppare il lavoro terminologico, integrandolo a quello lessicografico computazionale, per poter sviluppare lingware applicabili a software utili allo sviluppo di servizi di e-Government di primo, secondo e terzo livello basati su IR e NLP.

# Riferimenti bibliografici


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

Elia A., "*Discorso scientifico e linguaggio settoriale. Un esempio di analisi lessico-grammaticale di un testo neuro-biologico*" in Cicalese A., Landi D. (a cura di) «*Simboli, linguaggi e contesti*», Carocci, Roma, 2002

Zellig S. Harris, *From Morphemes to Utterances*, in Language, 22 n°2, 1946

Gross M., *Méthodes en syntaxe, régime des constructions complétives*, Hermann, Paris, 1975

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

*Monteleone M., Lessicografia e dizionari elettronici. Dagli usi linguistici alle basi di dati lessicali. Fiorentino & New Technology, Napoli, 2002*

Vietri S., *Navigare nei testi. Teorie e applicazioni informatiche per la linguistica testuale.* Editoriale Scientifica Italiana, Napoli, 2001

Vietri S., *Dizionari elettronici e grammatiche a stati finiti*, Plectica, Salerno, 2008

*Vellutino D., "La comunicazione pubblica per la promozione delle pari opportunità", in V. D'Antonio e S. Vigliar (a cura di), Studi di Diritto della Comunicazione. Persone, Società e Tecnologie dell'Informazione, CEDAM, Padova, 2009*

# Progetto S.O.L.C.O

Il lavoro presentato è stato realizzato per il progetto S.O.L.C.O. - Soluzioni per la Società della Linguistica Computazionale.

Progetto cofinanziato dalla Regione Campania, misura 3.17 POR Campania 2000-2006, nell'ambito dell'Accordo di Programma Quadro in materia di E-government e Società dell'Informazione. Progetto Metadistretto del settore ICT.

Il Gruppo "Maurice Gross" ha realizzato la base dati lessicografica per l'informazione europea (DIE) che comprende i dizionari elettronici di dominio descritti.

Ha sviluppato circa 500 grammatiche locali ed automi e trasduttori a stati finiti per il riconoscimento delle FAQ del dominio "Fondi strutturali".

# Gruppo di ricerca "Maurice Gross"

| Progetto S.O.L.C.O. | Componenti |
|---|---|
| **Annibale Elia**<br>elianni@tin.it | Emilio D'Agostino<br>Simonetta Vietri<br>Alberto Postiglione<br>Caterina D'Elia<br>Giustino De Bueriis<br>Johanna Monti<br>Simona Messina<br>Alberto Maria Langella<br>Daniela Guglielmo<br>Antonella Napoli<br>Fabiola Bocchino |
| **Mario Monteleone**<br>mmonteleone@unisa.it | |
| **Daniela Vellutino**<br>dvellutino@unisa.it | |
| **Federica Marano**<br>fmarano@unisa.it | |
| **Simona Sabatino**<br>sabatino.simona@libero.it | |