Strutture lessicali delle informazioni comunitarie all'interno di domini specialistici

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

Università di Salerno, Dipartimento di Scienze della Comunicazione

Riassunto

L'informazione comunitaria è veicolata attraverso parole terminologiche tecnico-specialistiche, spesso elencate in glossari istituzionali. Dal punto di vista morfo-grammaticale, tali parole si presentano come repertori di termini registrati in virtù della loro maggiore frequenza nei testi e non in rapporto ai tratti distintivi delle loro realizzazioni lessicografiche. Questo studio riguarda la descrizione delle strutture compositive delle unità lessicali superiori rilevate come fisse nei glossari istituzionali comunitari e, successivamente, l'etichettatura morfo-grammaticale per la costruzione di dizionari elettronici di dominio. Le entrate lessicali dei dizionari elettronici di dominio sviluppati sono state testate attraverso un corpus di testi tipologizzati pertinenti l'informazione comunitaria, al fine di trarre indicazioni sulle distribuzioni di frequenza utili per indagare i processi di polirematicizzazione.

Abstract

European Community information is released using technical and specialized terminological words often are included in institutional glossaries that are repositories of terms morpho-grammatically collected only considering their statistical frequency in texts and not also distinctive lines of lexicographical expression. The study concerns description of composition structure of lexical units acquired as fixed collocations in European Community glossaries and, successively, it concerns morpho-grammatical tagging in order to build specific domain electronic dictionaries. Lexical units of domain electronic dictionaries have been verified using a structured corpus of texts concerning European Community information in order to derive frequency collocations helpful to investigate multiword expressions.

Keywords: computational linguistics, electronic dictionaries, multiword expressions, Italian lexicography, natural language processing

1. Introduzione

Il presente studio [1] descrive le unità lessicali superiori presenti nei testi e nei glossari istituzionali comunitari in lingua italiana, nonché la loro etichettatura morfo-grammaticale per la costruzione dei dizionari elettronici di domini specialistici pertinenti le informazioni comunitarie.

In questo lavoro, il sintagma nominale complesso "informazione comunitaria/informazioni comunitarie" denomina le informazioni diffuse dalle istituzioni dell'Unione Europea attraverso

[1] Riguardo i contenuti di questo articolo, Annibale Elia ha realizzato le sezioni 1 e 2. Mario Monteleone ha realizzato la sottosezione 2.1 e la sezione 4. Daniela Vellutino ha realizzato la sezione 3, la sottosezione 3.1 e insieme a Simona Sabatino la sottosezione 3.4. Federica Marano ha realizzato la sottosezione 3.2.

vari tipi di testo oggetto di differenti processi di comunicazione. Gli organismi comunitari sono fonte di diritto e di politiche di sviluppo socioeconomico, pertanto molteplici sono i domini specialistici delle informazioni comunitarie. In particolare, questo studio riguarda tre domini specialistici, o meglio:

- due domini di conoscenza che rappresentano i cardini del modello di sviluppo europeo definito dalla Strategia di Lisbona, vale a dire i termini dei lessici di specialità pertinenti le aree semantiche "Società dell'Informazione" e "Pari Opportunità";
- un dominio specialistico che li comprende poiché relativo all'area semantica "Fondi strutturali", gli strumenti finanziari della politica regionale dell'Unione europea. Questi lessici di specialità sono stati integrati in un unico dizionario elettronico che abbiamo denominato Dizionario Informazione Europea (DIE).

A partire dal 1988, il Dipartimento di Scienze della Comunicazione dell'Università di Salerno ha sviluppato il sistema di dizionari elettronici DELA (parole semplici e composte). Il DIE costituisce una ulteriore risorsa linguistica funzionale all'addestramento delle applicazioni di analisi testuale, in particolare per Nooj, il programma per il trattamento automatico del linguaggio naturale, sviluppato da Max Silbertztein a partire dal 2005.

2. Quadro teorico di riferimento

La metodologia linguistica descrittiva di riferimento per il nostro studio è il Lessico-Grammatica, originariamente introdotto nella comunità scientifica europea da Maurice Gross all'Università Parigi 7 dopo il 1960. Scopo principale del Lessico-Grammatica è descrivere dettagliatamente tutti i meccanismi combinatori indissolubilmente legati alle concrete entrate lessicali. Uno dei risultati più recenti ed importanti della descrizione lessico-grammaticale è proprio la costruzione di tre software e lingware orientati all'analisi testuale automatica, ovvero UNITEX [2] INTEX ed ora NOOJ [3].

Oggi, il quadro metodologico lessico-grammaticale è applicato da gruppi di ricercatori che lavorano in diverse università, laboratori e strutture di ricerca mondiali, e che si riconoscono nella rete di ricerca scientifica che va sotto il nome di RELEX.

Il Dipartimento di Scienze della Comunicazione dell'Università di Salerno ha avuto una lunga collaborazione scientifica con il LADL (Laboratoire d'Automatique Documentaire et Linguistique) dell'Università Parigi 7, diretto per oltre venti anni da Maurice Gross. Questa stretta collaborazione scientifica ha sempre avuto il fine di raffinare il lingware prodotto nel corso degli anni e di adattarlo ad applicazioni software.

Il presente lavoro è parte dello studio delle microlingue nell'ambito del Lessico-Grammatica (Gross, 1975; EMDA, 1983; Elia, 1984; D'Agostino, 1984; D'Agostino et al., 1985; Vietri 2000; Elia, 2008) e, in particolare, riguarda l'uso delle unità lessicali superiori che possono essere 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).

In tutte le lingue del mondo esiste una stretta relazione di necessità tra terminologia e parole composte polirematiche. Ciò è testimoniato dalla presenza nei lessici di specialità di un numero molto elevato di composti, in alcuni casi superiore al 90% di tutto l'insieme lessicale repertoriato (Monteleone, 2008). Pertanto, l'analisi delle strutture lessicali delle informazioni comuni-

[2] Per ulteriori indicazioni sulle funzionalità di UNITEX, si veda http://www-igm.univ-mlv.fr/~unitex/.
[3] Per ulteriori indicazioni sulle funzionalità di NOOJ, si veda http://www.nooj4nlp.net/pages/nooj.html.

tarie ha l'obiettivo di trarre indicazioni sui comportamenti distribuzionali, sulle co-occorrenze delle unità lessicali superiori repertoriate come fisse nei glossari istituzionali per acquisire informazioni utili per indagare i processi di poliremat icizzazione nelle lingue tecnico-speciali osservate.

2.1. Il sistema dei dizionari elettronici DELA

I dizionari elettronici per la lingua italiana sviluppati nel sistema DELA, a cui si integra DIE, sono basi di dati [4] lessicali formalizzate, strutturate omogeneamente ed in cui le caratteristiche morfo-grammaticali delle entrate (genere, numero e flessione) sono indicate da etichette alfanumeriche univoche e non ambigue. I dizionari elettronici hanno la funzione di motori linguistici basici per i software di analisi testuale automatica già citati, e sono di due tipi, ovvero:

- di parole semplici (denominati DELAS-DELAF), ed includono tutte quelle parole semanticamente autonome e composte da sequenze di lettere non interrotte e delimitate da spazi bianchi (ad esempio le parole *casa*, *battello*);
- di parole composte (denominati DELAC-DELACF), ed includono tutte quelle sequenze formate da due o più parole e che costruiscono congiuntamente singole unità di significato (ad esempio le sequenze *casa di cura*, *battello a vapore*). Il dizionario elettronico specialistico del settore delle informazioni comunitarie – DIE –, utilizzato insieme ai DELA per lo studio in oggetto, è stato configurato come dizionario di parole composte.

Tale suddivisione è necessaria sia dal punto di vista formale e morfologico, sia da quello semantico [5]. Infatti, in fase di compilazione di un software di query, recupero informazioni o analisi testuale automatica, come ad esempio i già citati INTEX e UNITEX, l'assenza di separatori all'interno delle parole semplici, e la presenza di questi all'interno delle parole composte, saranno fattori discriminanti e comporteranno impostazioni e scelte differenti per quanto riguarda gli automatismi nel trattamento dei dati. Nel caso delle parole semplici, sarà infatti necessario prevedere e trattare solo dati alfabetici o numerici; con le parole composte, la presenza di separatori inserirà un livello aggiuntivo di dati, ai quali si dovranno assegnare funzioni univoche, non ambigue, e di valore diverso da quelli alfabetici e numerici. Diverso è invece il caso di NOOJ, il più recente dei software precedentemente citati, compilato in modo da gestire congiuntamente basi di dati di struttura diversa; in NOOJ, infatti, è possibile realizzare un unico dizionario elettronico contenente sia parole semplici che composte.

3. Il lessico dell'informazione comunitaria

In questi anni i concetti dei principi comunitari e delle politiche pubbliche europee di sviluppo socioeconomico hanno acquisito sempre maggiore rilevanza nelle attività istituzionali degli Stati membri. Pertanto, soprattutto nella prima fase di applicazione del diritto comunitario ai diritti nazionali, è stata molto importante l'attività terminologica finalizzata a selezionare e designare, sul piano interlinguistico, unità lessicali utili a comunicare non solo la conoscenza specialistica del diritto comunitario, ma anche quella specifica dei vari domini scientifici e tecnici.

[4] Il termine *basi di dati* va qui inteso nella più comune accezione informatica, sia dal punto di vista teorico che da quello di strutturazione pratica. Inoltre, un dizionario elettronico non può essere usato come un dizionario cartaceo, né può sostituirne il ruolo.

[5] Aggiungiamo che le parole semplici hanno sempre un tasso di polisemia (leggi ambiguità) più elevato di quelle composte, che ne hanno uno molto prossimo allo zero. Per quanto riguarda il recupero informazioni e l'analisi testuale automatica, ciò implica indirettamente la necessità di prevedere ed impostare modalità analitiche diverse per i due tipi di parole.

Il meccanismo della formazione che maggiormente caratterizza i termini del lessico comunitario è la costruzione di unità lessicali superiori (Nystedt, 2000; Cosmai, 2007). Ciò è quanto si osserva consultando i vari repertori terminografici in forma di glossari specialistici, nonché altre risorse linguistiche, quali banche dati terminologiche multilingui e *thesauri* relativi ai temi dell'agenda politica e dei settori di competenza dell'attività giuridico-istituzionale dell'Unione Europea.

3.1. La metodologia d'indagine

Per il nostro studio, rivolto ad indagare e repertoriare con modalità lessicografiche computazionali le unità lessicali superiori del lessico comunitario, siamo partiti dall'analisi di alcuni domini specialistici attraverso due iter di ricerca.

Per il dominio "Società dell'Informazione" la prima fase di ricerca è consistita nell'individuazione delle unità lessicali superiori di dominio attraverso la lettura diretta di un corpus di 75 testi (115.557 occorrenze tokenizzate); in seguito, i lemmi rilevati sono stati confrontati con quelli registrati in glossari istituzionali.

Per il dominio "Pari Opportunità", invece, il procedimento è stato inverso: l'indagine è partita dalla descrizione delle entrate lessicali registrate nel glossario comunitario "*100 parole per la parità*" [6] annotate secondo il procedimento di formalizzazione morfo-grammaticale per la creazione di un dizionario elettronico di dominio da integrare al sistema dei dizionari DELA ed utilizzabile dai software per il trattamento testuale automatico sopracitati. Il dizionario elettronico di dominio "Pari Opportunità" sviluppato contiene 216 entrate in forma canonica e flessa [7].

Successivamente sono stati raccolti testi appartenenti ad una tipologia testuale, il bilancio di genere, che avrebbe dovuto contenere tutte le voci repertoriate nel glossario comunitario. Sono stati raccolti 22 testi di bilancio di genere e, dopo un primo esame manuale che ha portato alla depurazione degli elementi non verbali, è stato costruito un corpus della dimensione complessiva di 1.523.423 occorrenze tokenizzate. Il corpus è stato analizzato con il software di analisi testuale automatica CATALOGA [8] a cui è stato integrato il dizionario elettronico di dominio sviluppato.

Attraverso entrambi gli iter di ricerca, nei testi comunitari è stata rilevata la presenza di unità lessicali superiori di dominio assenti nei glossari istituzionali.

Successivamente l'interesse di ricerca è stato rivolto al macro-dominio "Fondi strutturali". In questo caso, l'indagine ha previsto al contempo la creazione di un corpus di testi scritti della dimensione complessiva di 3.071.610 occorrenze tokenizzate nonché la raccolta e il confronto di 16 glossari istituzionali.

Allo scopo di costruire un corpus utile per rilevare le unità lessicali terminologiche superiori, è stata definita una classificazione tipologica dei testi dei documenti di domino in base alle

[6] Il glossario "100 parole per la parità" è stato sviluppato nel 1998 dall'Unità "Pari opportunità tra le donne e gli uomini" della Commissione europea con il supporto del Servizio di Traduzione. È un repertorio lessicale elaborato in 11 lingue (Italiano, Inglese, Francese, Tedesco, Spagnolo, Greco, Danese, Finlandese, Olandese, Portoghese, Svedese) che a quel tempo erano gli idiomi degli Stati membri dell'Unione Europea.

[7] Il dizionario elettronico di dominio sviluppato a partire dal glossario comunitario "100 parole per la parità" è in Vellutino (2009a).

[8] CATALOGA è un software di analisi testuale automatica realizzato da Alberto Postiglione e Mario Monteleone nell'ambito delle attività di ricerca linguistico-computazionali dirette da Annibale Elia presso il Dipartimento di Scienze di Comunicazione dell'Università di Salerno. CATALOGA si basa sul matching fra testi e dizionari elettronici di parole composte, realizzati in base alle indicazioni linguistico-teoriche del Lessico-grammatica.

finalità pragmatiche dei differenti processi di informazione e comunicazione pubblica [9]. I testi del corpus sono perciò classificati rispetto alle forme di comunicazione e alle relative tipologie testuali. Ad esempio:

- per la comunicazione normativa → tipo di testo giuridico (fonte normativa primaria o secondaria) → testo "Trattato di Lisbona";
- per la comunicazione per la trasparenza amministrativa → tipo di testo documento di programmazione economica → testo "Complemento di Programmazione";
- per la comunicazione pubblico-istituzionale finalizzata a diffondere informazioni di pubblica utilità → tipo di testo FAQ → testo "Richieste d'informazione sui finanziamenti dei Fondi strutturali" (per il nostro caso specifico).

I testi delle fonti primarie e secondarie per la comunicazione normativa contengono 674.746 occorrenze tokenizzate; i testi per la comunicazione per la trasparenza amministrativa contengono 2.390.093 occorrenze tokenizzate; i testi per la comunicazione pubblico-istituzionale comprendono 345 frasi di FAQ estratte da 51 siti istituzionali in materia relative alle richieste d'informazioni sugli interventi di finanziamento comunitario (6.771 occorrenze tokenizzate).

Contestualmente sono stati repertoriati e confrontati 16 glossari istituzionali prodotti da varie amministrazioni pubbliche a livello europeo e nazionale. Questa operazione è stata effettuata allo scopo di individuare entrate comuni ed eventuali variazioni lessicali attestate per ogni parola terminologica. Queste informazioni linguistiche terminologiche e lessicografiche sono state registrate in una base di dati. Pertanto, ogni parola è stata descritta attraverso:

- la definizione terminologica (o le diverse definizioni terminologiche) con la relativa fonte (o fonti) di attestazione;
- i riferimenti ai testi del corpus in cui è presente;
- le varianti lessicali per sinonimia, per iponimia, per iperonimia;
- le varianti ortografiche (caratteri tutti minuscoli, tutti maiuscoli, con l'iniziale maiuscola, e con la presenza sia di maiuscole che di minuscole);
- gli acronimi (tipo: protogramma "AdG" che è una variante della sequenza lessicale fissa "Autorità di Gestione");
- altre varianti (ad esempio la sequenza lessicale fissa Programma Operativo Regionale che può avere le varianti POR e P.O.R.);
- le varianti per marche sociolinguistiche, in particolare gli internazionalismi sono stati registrati con i relativi corrispondenti italiani (tipo: "gender mainstreaming", "prospettiva di genere", "ottica di genere").

Solo dopo tale operazione di schedatura terminologica e lessicale è iniziato il lavoro di annotazione morfo-grammaticale, strutturale e flessionale delle entrate. Tale lavoro è stato effettuato in base al procedimento di formalizzazione già utilizzato per i dizionari DELA, che per ogni termine prevede, attraverso un sistema di etichette alfanumeriche univoche e non ambigue, l'indicazione di genere, numero, funzione grammaticale e struttura compositiva interna.

Il dizionario del macro-dominio "Fondi strutturali" contiene quindi 1.949 lemmi etichettati.

Dall'integrazione di questi domini specialistici investigati nasce infine il **dizionario elettronico DIE** che conta 3126 lemmi.

Nel paragrafo che segue sono descritte in modo analitico le fasi di costruzione di un dizionario elettronico di dominio specialistico ("Società dell'Informazione") e nel successivo paragrafo è riportato un estratto del dizionario elettronico DIE.

[9] Per una descrizione della classificazione dei testi di comunicazione pubblica si rimanda a Vellutino (2009b).

3.2. Creazione del dizionario elettronico del dominio specialistico "Società dell'Informazione"

La creazione e gestione del dizionario elettronico del dominio "Società dell'Informazione" (ora in avanti detto SI) è avvenuta attraverso quattro fasi principali.

Nella prima fase di *lexical acquisition* il lavoro si è concentrato soprattutto sul reperimento dei lemmi a partire dai testi. Sono stati selezionati 75 testi dai siti web istituzionali (www.europa.eu.int, www.campaniasi.it, www.mininnovazione.it). Da ogni singolo documento, sono state estratte le unità lessicali superiori con caratteristiche terminologiche di dominio. Infine, è stata creata una lista costituita da circa 270 lemmi in forma canonica.

In una seconda fase, ogni entrata è stata esaminata nella sua funzione linguistica. Sono state registrate le varietà lessicali (ortografiche, sinonimiche, iponimiche/iperonimiche) che sono state anch'esse formalmente descritte in base alla loro struttura compositiva.

Nella fase successiva, per ogni entrata terminologica, già precedentemente etichettata con un paradigma flessionale, sono state prodotte le corrispondenti entrate flesse sia nel genere (maschile/femminile), sia nel numero (singolare/plurale). Ad esempio la stringa:

carte/d'identità/elettroniche,carta/d'identità/elettronica.N+NPNA:fp-+;SI

presenta il procedimento di formalizzazione morfo-grammaticale per il sistema dei dizionari DELA. Pertanto, gli elementi costituenti la struttura interna morfo-sintattica *NPNA* (nome+preposizione+nome+aggettivo) sono preceduti dal formalismo *N+*. Tale elemento indica la funzione grammaticale dell'unità lessicale superiore, la quale è inoltre di genere femminile e, in questo caso, flessa al plurale.

Il nucleo di base del **dizionario SI** risulta quindi composto da 495 stringhe, di cui 455 sono forme che prevedono una flessione, mentre 40 non la prevedono perché attestate soltanto nella forma singolare o soltanto in quella plurale (Marano, 2005).

Di seguito alcuni esempi:

es 1. Autorità Nazionale per la Sicurezza
(non prevede flessioni)

es 2. Carta d'identità Elettronica
carta d'identità elettronica – carta di identità elettronica – Carta di identità Elettronica
(varianti ortografiche)
Carte d'identità Elettroniche – carte d'identità elettroniche – carte di identità elettroniche – Carte di identità Elettroniche
(flessioni)

La lista delle entrate estratte manualmente è stata poi sottoposta ad un matching con glossari istituzionali specialistici [10] per verificare la percentuale di unità lessicali superiori in sovrapposizione tra testi e glossari. A seguito di questa operazione, il dizionario elettronico specialistico "Società dell'Informazione" è in totale composto da 989 stringhe, di cui 919 sono forme che prevedono una flessione, mentre 70 non la prevedono.

I risultati del matching indicano una scarsa sovrapposizione tra le unità estratte manualmente dai testi ed i lemmi all'interno dei glossari. Nello specifico, si calcolano una *precision* e una *recall* nell'ordine rispettivamente dell'1,61% e dell'1,62% [11].

[10] I glossari utilizzati sono stati il *Glossario dell'informatica nelle norme italiane* curato dal Centro Nazionale per l'Informatica nella Pubblica Amministrazione (CNIPA), il *Glossario di diritto delle nuove tecnologie e dell'e-government* (2008) di Borruso, Riem, Sirotti Gaudenzi, Vicenzotto.

[11] Nel caso specifico la misura di *precision* indica la correttezza e pertinenza dei termini presenti nel dizionario

Le percentuali sono molto basse, i termini co-presenti sono soprattutto quelli meno tecnici e più d'uso comune (es. *amministrazione pubblica*). Nei glossari non sono state comunque repertoriate le seguenti varianti:

- varianti per relazione iponimica/iperonimica, come ad esempio *rete pubblica di comunicazione* → *rete via cavo* (quest'ultima può essere un tipo più specifico di *rete pubblica di comunicazione*);
- varianti per sinonimia es. *firma digitale = firma elettronica*; variante ortografica es. *carta d'identità = Carta d'identità*).

Nella quarta fase, attraverso il corpus costruito per il macro-dominio "Fondi Strutturali", è stata verificata la pertinenza terminologica delle entrate del dizionario SI e la completezza della base di dati. Questa fase di revisione è stata attuata manualmente, con un controllo formale della correttezza dell'etichettatura morfo-sintattica e flessionale, e automaticamente attraverso i software sopracitati per l'analisi automatica di testi.

Al fine di acquisire informazioni sui processi di poliremat icizzazione nel dizionario SI sono state individuate 55 tipologie di unità lessicali superiori: 4 *bi-gram* (strutture costituite da 2 parole), 10 *tri-gram*, 14 *fourth-gram*, 9 *fifth-gram*, etc. fino a strutture molto più lunghe costituite da 10, 12 o addirittura 13 elementi nella struttura.

In Tab. 1 sono illustrati alcuni esempi.

Numero elementi dell'unità lessicale	*Struttura interna*	*Esempi*
bi-gram	AN	pubblica amministrazione
	NA	documento digitale
	NN	interfaccia utente
	NV	soggetto abilitato
tri-gram	NAA	posta elettronica certificata
	NPN	servizio di autenticazione
	…	…
fourth-gram	NAPN	Carta Nazionale dei Servizi
	…	…
fifth-gram	NAPNA	gestione elettronica del flusso documentale
	…	…
più di 10 elementi	NCNAPNPNPNACA	detenzione e diffusione abusiva di codici di accesso a sistemi informatici o telematici

Tabella 1

Grazie a questa classificazione è stato possibile individuare quali sono le strutture lessicali più produttive nei processi di poliremat izzazione del lessico di specialità di dominio. Quelle più frequenti nel **dizionario "SI**" sono: NA e NPN che contano rispettivamente 295 e 263 stringhe; seguono le strutture NN (62 stringhe), NPNA (47 stringhe), NPNPN (40 stringhe), NAPN (33 stringhe), AN (31 stringhe), NAA (27 stringhe), NAPNA (16 stringhe), alle restanti 46 strutture appartengono unità lessicali superiori meno frequenti nel lessico specialistico perché hanno meno di 10 stringhe nel dizionario di dominio.

elettronico specialistico, mentre la *recall* indica quanti di quei termini corretti sono in comune con i glossari istituzionali.

3.3. Estratto del dizionario elettronico DIE

Diamo qui di seguito un estratto del dizionario elettronico DIE, in cui sono evidenziate diverse forme di etichettature formale delle parole terminologiche raccolte:

accordo/sull'/applicazione/delle/misure/sanitarie/e/fitosanitarie,.N+NPNPNACA:ms+-;FONDSTRUT

accordo/sulla/agricoltura,.N+NPN:ms-+;FONDSTRUT

accordo/sulla/applicazione/delle/misure/sanitarie/e/fitosanitarie,.N+NPNPNACA:ms+-;FONDSTRUT

accordo/verticale,.N+NA:ms-+;FONDSTRUT

acquis/comunitario,.N+NN:ms--;FONDSTRUT

adesione/di/un/nuovo/stato/all'/unione,.N+NPDANPN:fs-+;FONDSTRUT

adesione/di/un/nuovo/stato/alla/Unione,.N+NPDANPN:fs-+;FONDSTRUT

adesioni/di/un/nuovi/stati/all'/unione,adesione/di/un/nuovo/stato/all'/unione.N+NPDANPN:fp-+;FONDSTRUT

adesioni/di/un/nuovi/stati/alla/Unione,adesione/di/un/nuovo/stato/alla/Unione.N+NPDANPN:fp-+;FONDSTRUT

affare/economico/e/finanziario,.N+NACA:ms-+;FONDSTRUT

affare/finanziario,.N+NA:ms-+;FONDSTRUT

affare/marittimo,.N+NA:ms-+;FONDSTRUT

affare/marittimo/e/pesca,.N+NACN:ms-+;FONDSTRUT

affare/sociale,.N+NA:ms-+;FONDSTRUT

affari/economici/e/finanziari,affare/economico/e/finanziario.N+NACA:ms-+;FONDSTRUT

In questa fase, e solo per indagare le strutture compositive interne delle unità lessicali superiori presenti nel dizionario elettronico del macro-dominio "Fondi strutturali", è stato utilizzato il software di trattamento automatico dei testi NOOJ e sono state rilevare le strutture più produttive. Analogamente al dizionario elettronico "SI", sono più produttive, le strutture lessicali NA (513 stringhe) e NPN (448 stringhe) ed a seguire NPNA (137 stringhe) e NPNPN (75 stringhe). In Figg. 1 e 2 sono riportate le strutture lessicali presenti nel dizionario elettronico sviluppato.

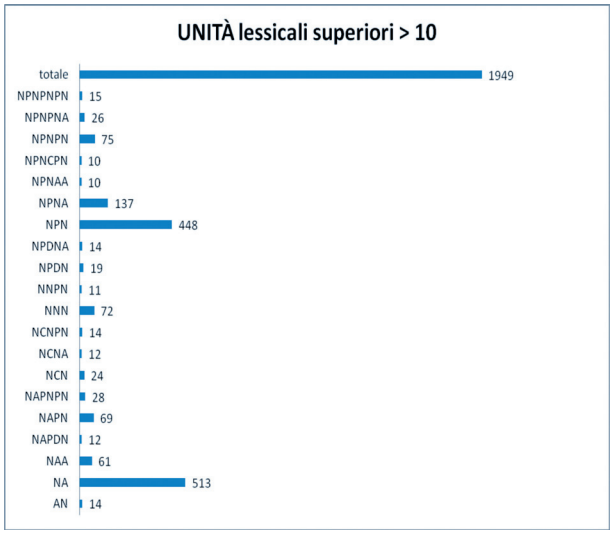

Figura 1

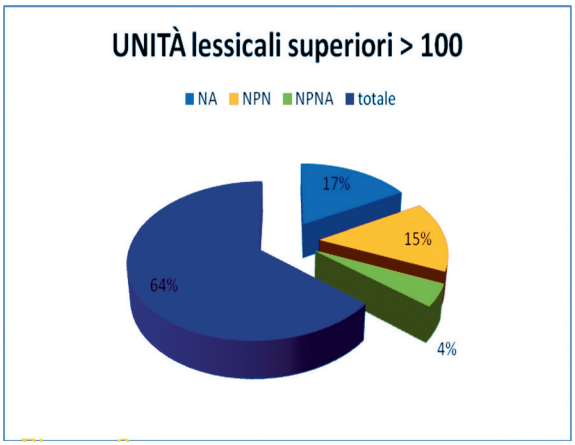

Figura 2

4. Sviluppi futuri

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 nella realizzazione di software per servizi di e-Government di primo, secondo e terzo livello.

Tali software potranno così essere basati su sistemi per il trattamento automatico del linguaggio finalizzati al riconoscimento e all'estrazione di conoscenze morfo-grammaticali e lessicali.

Il nostro lavoro sul corpus è quindi ancora in una fase iniziale. Sarà soprattutto necessario, in futuro, definire le misure statistiche per la stratificazione, aggiornando costantemente i risultati, proseguendo le indagini sulle modalità di creazione lessicale ed affiancando a queste studi sociolinguistici che attestino i concreti usi lessicali nei diversi processi di comunicazione pubblica.

Riferimenti bibliografici

Cosmai D. (2007). *Tradurre per l'Unione Europea*. Milano: Hoepli.

De Bueriis G., Di Maio F., Elia A. and Monteleone M. (2008). Le poliremathiche dell'italiano. In De Bueriis, G. and Elia, A., editors, *Lessici elettronici e descrizioni lessicali, sintattiche, morfologiche ed ortografiche*, Plectica: Salerno, pp. 15-68.

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

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

D'Agostino E. (1984). Les compléments de lieu comme compléments de verbe dans les constructions transitives italiennes. In Guillet, A. and La Fauci, N., editors, *Lexique-Grammaire des langues romaines*.

D'Agostino E. (editor) (1995). *Tra sintassi e semantica.Descrizioni e metodi di elaborazione automatica della lingua d'uso*. Napoli; Edizioni Sceintifiche Italiane.

De Mauro T. (editor) (2000). *Il dizionario della lingua italiana*. Torino: Paravia.

Elia A., Monteleone M., De Bueriis G. and Di Maio F. (2008). Le poliremathiche dell'italiano. In De Bueriis, G. and Elia, A., editors, *Lessici elettronici e descrizioni semantiche, sintattiche e morfologiche. Risultati del Progetto PRIN 2005 Atlanti Tematici Informatici - ALTI*, Collana "Lessici & Combinatorie", n. 2, Dipartimento di Scienze della comunicazione dell'Università degli Studi di Salerno, Salerno: Plectica, pp. 11-65.

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

Gross M. (1989). La construction de dictionnaires électroniques. *Annales des Télécommunications*, vol. 44, 1-2 : 4-19, CENT, Issy-les-Moulineaux/Lannion.

Grossmanm M. and Rainer F. (2004). *La formazione delle parole in italiano*. Tübingen: Max Niemeyer.

Magris M., Musachio M.T., Rega L. and Scarpa F. (2002). *Manuale di terminologia. Asoetti teorici, metodologici, applicativi*. Milano: Hoepli.

Marano F. (2005): Lessico-Grammatica della Società dell'Informazione. Tesi di Laurea. Relatori Proff. Annibale Elia, Simonetta Vietri.
Available on line from http://www.assinform.it/download/tesi/tesi_marano.PDF.

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

Nystedt J. (2000). L'italiano dei documenti CEE: le sequenze di parole. In *Atti del Convegno di studi «Linguistica giuridica italiana e tedesca: obiettivi, approcci, risultati»*, Bolzano 1-2 ottobre, Padova: Unipress.

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

Vellutino D. (2009a). *Lessico delle donne di «potere» nei processi di comunicazione istituzionale*. Salerno: Plectica.

Vellutino D. (2009b). La comunicazione pubblica per la promozione delle pari opportunità. In D'Antonio, V. and Vigliar, S., editors, *Studi di Diritto della Comunicazione. Persone, Società e Tecnologie dell'Informazione*, Padova: CEDAM, pp. 267-297.

Vietri S. and Elia A. (2000). Electronic Dictionaries and Linguistic Analysis of Italian Large Corpora. In Rajman, M. and Chappelier, J.C., editors, *JADT 2000 – Actes des 5es Journées internationales d'Analyse statistique des Données Textuelles*, 9-11 Mars 2000, Ecole Polytechnique fédérale de Lausanne, Suisse, pp. 181-196.

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

Vietri S. and Elia A. (2001b). Analisi automatica dei testi e dizionari elettronici. In Burattini, E. and Cordeschi, R. (2001). *Intelligenza artificiale*. Roma: Carocci.

Vietri S. (2008). *Dizionari elettronici e grammatiche a stati finiti*. Salerno: Plectica.

Sitografia

Common agricultural policy (http://ec.europa.eu/agriculture/glossary/glossary_it.pdf).

CORDIS Research and Development (http://cordis.europa.eu/guidance/glossary_it.html).

Dipartimento per le Politiche di Sviluppo e Coesione (http://www.dps.tesoro.it/glossario.asp).

Environment glossary (http://ec.europa.eu/atoz_it.htm).

EU Competition Policy (http://ec.europa.eu/competition/publications/glossary_it.pdf).

Europa glossary (http://europa.eu/scadplus/glossary/index_it.htm).

European Convention Glossary (http://european-convention.eu.int/glossary.asp?lang=IT).

European Judicial Network in civil and commercial matters (http://ec.europa.eu/civiljustice/glossary/glossary_it.htm).

Eurovoc thesaurus multilingue che copre tutti i settori d'attività delle Comunità europee http://europa.eu/eurovoc/sg/sga_doc/eurovoc_dif!SERVEUR/menu!prod!MENU?langue=IT.

Gergo europeo (http://europa.eu/abc/eurojargon/index_it.htm).

IATE banca dati terminologica multilingue dell'UE. http://iate.europa.eu/iatediff/switchLang.do?success=mainPage&lang=it.

Manuale interistituzionale di convenzioni redazionali (http://publications.europa.eu/code/it/it-390500.htm).

Regional Development Glossary (http://ec.europa.eu/regional_policy/glossary/glossary_it.htm).