AIDAinformazioni, *Rivista di Scienze dell'informazione,* è il periodico ufficiale dell'Associazione Italiana per la Documentazione Avanzata (AIDA). Pubblica articoli di carattere professionale sul mondo dell'informazione e delle tecnologie ed aggiorna sulla vita dell'Associazione.

Associazione Italiana Documentazione Avanzata
c/o CASPUR
via dei Tizii, 6
I-00185 Roma
aida@aidaweb.it
www.aidaweb.it

Associazione Italiana
Documentazione Avanzata

AIDA informazioni

Anno 26 – Numero 1-2/2008

AIDA informazioni

AIDAinformazioni

Numero 1-2/2008
Trimestrale
Anno 26
gennaio-giugno 2008
ISSN: 1121-0095

Associazione Italiana
Documentazione Avanzata

Poste Italiane SpA
Spedizione in AP 70% - DCB Roma

segue in quarta di copertina...

AIDAinformazioni

Associazione Italiana
per la Documentazione Avanzata
Presidente
Piero Cavaleri
Vice presidente
Serena Sangiorgi
Segretario tesoriere
Domenico Bogliolo
Consiglio direttivo
Piero Cavaleri, Marco Calvo, Alessandra Cornero, Augusta Franco, Lucia Maffei, Giulio Marconi, Serena Sangiorgi

AIDAinformazioni
Edizione a stampa: ISSN 1121-0095
Edizione elettronica: ISSN 1594-2201
Trimestrale
Anno 26, numero 1-2, gennaio-giugno 2008
Edizione elettronica
www.aidainformazioni.it
Direttore e responsabile
Mario De Gregorio
Direttore scientifico
Ferruccio Diozzi
Responsabile dell'edizione elettronica
Domenico Bogliolo
Comitato scientifico
Dunia Astrologo, Anna Baldazzi, Michèle Battisti, Marco Calvo, Maria Pia Carosella, Claudio Gnoli, Mariella Guercio, Susanna Mornati, Denis Reidy, Gino Roncaglia, Serena Sangiorgi
Segreteria di redazione
redazione@aidainformazioni.it

AIDAlampi
Supplemento elettronico
www.aidalampi.it
Caporedattore
Andrea Marchitelli
Redazione
Bonaria Biancu, Domenico Bogliolo, Francesca Cagnani, Maria Pia Carosella, Maria Cassella, Elisabetta Di Benedetto, Gabriele Gatti Elena Giglia, Perla Innocenti, Giulio Marconi, Vittorio Ponzani, Roberta Valente

Finito di stampare nel mese di dicembre 2008
Autorizzazione del Tribunale di Roma
n. 408/86 del 2/9/86
Iscrizione al Registro Nazionale della Stampa
n. 4656 del 16/6/94
© 2008 Associazione Italiana per la Documentazione Avanzata

Abbonamento gratuito per i soci, € 100,00 *per i non soci*
Modulo d'ordine: www.aidaweb.it/cgiabbonamento.html
con versamento sul c/c postale n. 73015000
oppure sul c/c BancoPosta IBAN IT45T0760103200000000730150002

Realizzato da
Cooperativa Nuova Cultura - Roma

AIDA in rete è...

AIDA*informazioni*, <www.aidainformazioni.it/>
la Rivista di Scienze dell'informazione
ora anche tutta online

AIDA*lampi*, <www.aidainformazioni.it/lampi/>
l'informazione in rete sul mondo della documentazione avanzata
Il supplemento elettronico di AIDA*informazioni* è disponibile su web
e inviato mensilmente per posta elettronica

AIDA*web*, <www.aidaweb.it>
l'organizzazione, i servizi, le pubblicazioni dell'Associazione
e inoltre...

AIDA*cornici*
Ambiente: le fonti dell'info-doc ambientale
Centri I&D italiani: *webrepertori*
Euroguida I&D: guida alla certificazione europea della professione
Pace: una risorsa documentaria per la coscienza critica
Reference: VRD per bibliotecari & documentalisti
Terminologia: siti di riferimento & lavori in corso

AIDA*formazione*
iniziative scientifiche, didattiche e formative
per documentalisti e specialisti dell'informazione

AIDA*job*
il cerca/trova lavoro per documentalisti & specialisti dell'informazione
inserzioni di richiesta/offerta di lavoro nel settore

AIDA*lavorincorso*
uno spazio libero per leggere e pubblicare prime idee,
stati di parziale avanzamento,
conclusioni provvisorie da confrontare e condividere

AIDAinformazioni

Rivista di Scienze dell'informazione

Anno 26, numero 1-2
gennaio-giugno 2008

Associazione Italiana per la Documentazione Avanzata

> «La terminologia fa parte dell'àmbito della lessicologia come l'informazione costituisce l'oggetto della documentazione, ma che, per svolgere determinate attività che riguardano l'informazione, i documentalisti, avvalendosi della terminologia, si muovono tra i termini reali, che trovano nei documenti che descrivono, e i termini normalizzati, che devono usare per rendere efficiente il loro lavoro. In tal modo, la terminologia reale e quella normalizzata convivono nel lavoro documentale. In quanto insieme di unità effettivamente usate nella comunicazione professionale, la terminologia costituisce parte della lessicologia, integrandosi in essa come una valenza delle unità lessicali, valenza che è alcune volte reale ma sempre potenziale. E la valenza normalizzata o meno dei termini si riduce a una informazione enciclopedica, della quale non si può dare spiegazione mediante le regole e i principi che governano le unità di cui si occupa una teoria del linguaggio» (M.T. Cabré).

Che terminologia e documentazione abbiano degli stretti legami di interdipendenza è noto ed evidente a chi si occupa di uno dei due àmbiti di conoscenza e numerosi sono i lavori scientifici che – da diversi punti di vista – analizzano e definiscono questo rapporto. *Dal terminologia y documentación* di Maria Teresa Cabré al *Terminologie et documentation* di Maryvonne Holzem, al *Knowledge transfer by computer-assisted terminology documentation* di C Galinsky e W. Nebodity, per citarne solo alcuni.

L'assenza di significativi contributi italiani è, dal un lato, lo specchio delle difficoltà ad affermarsi autonomamente che entrambe le discipline incontrano nel panorama culturale del nostro paese e, dall'altro, la conseguenza della mancanza di reali spazi di lavoro transdisciplinari capaci anche di ipotizzare momenti formativi comuni.

In questo contesto, l'annuale convegno dell'Associazione Italiana per la Terminologia (Assiterm), tenuto nel giugno 2008 all'Università della Calabria sul tema "Terminologia Analisi Testuale e Documentazione", ed i cui contributi sono oggi qui pubblicati, ha l'ambizione di rappresentare l'avvio o la ripresa di un cammino comune.

Programma

5 giugno 2008

9.00	Registrazione Partecipanti
9.30	Apertura lavori - Riccardo Gualdo (Presidente Ass.I.Term) - Presiede Giovanni Adamo (ILIESI - CNR - Roma)
10.00	Saluti - Giovanni Latorre (Magnifico Rettore - Università della Calabria) - Roberto Guarasci (Direttore Dipartimento Linguistica - Università della Calabria)
10.45	Relazione di apertura - *Realidad, cognición y lenguaje: la poliedricidad como principio* - Maria Teresa Cabré (Universitat Pompeu Fabra - Barcellona)
11.30	Coffee-break
11.45	*Repérage de la référence à partir du thésaurus, de la terminologie et de la sémantique lexicale* - Laurence Kister; Evelyne Jacquey; Bertrand Gaiffe (Nancy-Université - CNRS)
12.15	*Terminologia, modelli terminologici e reti* - Fóris Ágota (University of West Hungary)
12.45	Dibattito
13.00	Colazione di lavoro
14.30	*Processi di terminologizzazione e determinologizzazione nel dominio della diffusione e distribuzione del libro* - Franco Bertaccini; Claudia Lecci; Valentina Bono (SSLMIT Forlì - Università di Bologna)
15.00	*Terminologia e classificazione nel centro di documentazione della Democrazia Cristiana* - Roberto Guarasci (Università della Calabria)
15.30	*Archiwordnet, un thesaurus di settore integrato nel* wordnet *della lingua generica: compilazione e applicazioni* - Andrea Bocco; Enrica Bodrato; Antonella Perin (Politecnico di Torino)
16.00	Dibattito
17.00	*La terminologia: un capitale da non sottovalutare* - Donatella Pulitano (Cancelleria di Stato del cantone di Berna - Università di Ginevra)

17.30 *La memoria in rete: parole per ricordare* - Madel Crasta (BAICR - Roma)
18.00 *Esafety box* - Angela Aceti; Nunzia Bellantonio (ISPESL Roma)
18.30 Chiusura lavori
18.45 Assemblea Ass.I.Term.

6 giugno 2008

9.00 Inizio lavori - Presiede Riccardo Gualdo (Presidente Ass.I.Term) Relazione di apertura - *Terminologia dalla parte del ricevente* - Claudio Giovanardi (Università Roma Tre)
9.30 *La question des normes (numériques) de traitement documentaire (audiovisuelle) et en particulier les problématiques développées au sein du web sémantique* - Roger Roberts (RTBF/Titan)
10.00 *Energie rinnovabili: proposte di interventi terminologici* - Maria Teresa Zanola (Università Cattolica Sacro Cuore - Milano)
10.30 *Il trattamento linguistico dell'informazione al consumatore: lessico dell'energia e applicazioni terminologiche* - Sonia Piotti (Università Cattolica Sacro Cuore - Brescia)
11.00 Coffee-break
11.15 Proiezione del film sulla vita e l'opera di Paul Otlet: *L'Homme qui voulait classer le monde* - Introduzione Mauro Caproni (Università di Udine). Presenta il regista e autore Françoise Levie
12.45 Dibattito
13.00 Colazione di lavoro
14.30 *Linee di problema per il trattamento terminologico di documenti amministrativi elettronici* - Domenico Bogliolo (AIDA - Roma)
15.00 *I glossari dei siti della PA: la semplificazione del linguaggio per una comunicazione digitale interattiva e accessibile* - Claudia Rosa Pucci (Ministero delle Comunicazioni)
15.30 *Produzione e conservazione dei contenuti digitali: i requisiti tecnologici* - Stefano Pigliapoco (Università di Macerata)
16.00 *Verso un formato standard nelle intercettazioni e una proposta per l'archiviazione e la conservazione delle registrazioni* - Giuseppe A. Cavarretta (URT - CNR); Luciano Romito; Maria Tucci (Università della Calabria)
16.30 *L'attribuzione di testi con metodi quantitativi: riconoscimento di testi gramsciani* - Chiara Basile (Università di Bologna); Maurizio Lana (Università degli Studi del Piemonte Orientale "A. Avogadro" - Vercelli)

17.00 *La terminologia descrittiva di prodotti a stampa da ambiente analogico ad ambiente digitale* - Piero Innocenti (Università della Tuscia - Viterbo)
17.30 Dibattito
18.30 Partenza per Altomonte
21.00 Cena di gala

7 giugno 2008

9.00 Inizio Lavori: Presiede Maria Teresa Zanola - *Dal Testo alla conoscenza e ritorno: estrazione terminologica e annotazione semantica di basi documentali di dominio* - Simonetta Montemagni (Istituto di Linguistica Computazionale - CNR Pisa)
9.30 *Prospettive di relazioni fra linguistica del testo e descrizione Archivistica* - Paolo Franzese (Ministero per i beni e le attività culturali)
10.00 *Terminologia e Documenti per la formalizzazione standardizzata della conoscenza tacita* - Elena Cardillo; Antonietta Folino (Università della Calabria)
10.30 *Tra Terminologia e Documentazione: Voci indice ed estrazione terminologica in corpora documentali* - Maria Taverniti (Università della Calabria)
11.00 *Forma e contenuto nella terminologia della gestione documentale* - Piera Belcastro (Università della Calabria)
11.30 Coffee Break
12.00 *Strutturazione dell'informazione e integrazione della conoscenza* - Anna Rovella (Università della Calabria); Giovanni Marrè (It-Consult)
12.30 Dibattito
13.00 Chiusura lavori

I.Ter.An.Do
«Terminologia, analisi testuale e documentazione nella città digitale»
Convegno nazionale Ass.I.Term.

Indice

Realidad, cognición y lenguaje: la poliedricidad como principio

MARIA TERESA CABRÉ

The author makes a careful analysis of the process of lexicalization or terminologization of knowledge used to represent the complex cognitive system of reality. Firstly, she analyzes the point of confluence between the terminology and the documentation then switching to illustration of the learning process, conceptualization and verbalization, in linguistic units, of the objects of reality. She concludes by presenting some examples of relations between object-concept-term, highlighting the principle of poliedricity. In short, the author stresses the interdisciplinary plan of the terms understood as "objects" of knowledge, integrated for variants and facets, which corresponds to a different level of analysis. This principle, in effect, allows us to give an account of the complex relationship between a concept and its various terminological realisations.

Keywords: poliedricity, terminology, documentation, terminology unit.

Quiero agradecer en primer lugar a los organizadores de este congreso, y muy especialmente al profesor Roberto Guarasci, la confianza que han depositado en mi persona ofreciéndome la conferencia de apertura de la reunión de Ass.I.Term en la Universidad de Calabria. Espero no defraudarles con el tema que he elegido, que intenta abordar algunos aspectos de la intersección entre la Terminología y la Documentación en tanto que materias implicadas íntimamente en la información y el conocimiento.

Situaré mi intervención en la terminología y me centraré específicamente en el proceso de lexicalización o terminologización del conocimiento adquirido con la conceptualización de la realidad. Este proceso es relevante en Terminología por cuanto las unidades terminológicas no son sino el resultado de la verbalización de un concepto que ha sido construido a partir de operaciones de categorización de objetos de la realidad. Pero concierne también a la documentación porque un documento es una unidad compleja que concierne y es concernida por los mismos componentes involucrados en la terminología: el conocimiento, el lenguaje y la sociedad.

Iniciaré mi exposición presentando de manera breve los puntos de confluencia entre la documentación y la terminología para mostrar que se trata en primer lugar de objetos y materias interdisciplinares por naturaleza, y en segundo lugar de interdisciplinas que se han construido a partir de la articulación de elementos procedentes de

las mismas ciencias: las ciencias cognitivas, las ciencias del lenguaje y las ciencias sociales.

Entraré después en el proceso que se inicia en la aprehensión y discriminación de los objetos de la realidad, pasa después a su conceptualización y finalmente termina en su verbalización en forma de unidades semióticas, la gran mayoría de las cuales son unidades lingüísticas.

En tercer lugar presentaré algunos casos de relación entre objeto-concepto-término para mostrar que el principio de poliedricidad puede explicar de manera homogénea hechos aparentemente diferentes resaltando su unidad de partida y al mismo tiempo su diversidad.

1. Documentación y terminología

1.1. Puntos de confluencia

En ocasiones anteriores hemos tratado ya las similitudes entre la Terminología y la documentación como campos de conocimiento, tanto desde el punto de vista de sus características constitutivas, como de su evolución y ubicación entre los campos del saber. Veamos a continuación muy rápidamente algunas de estas similitudes que nos permitirán más tarde justificar que el principio de poliedricidad puede ser útil y adecuado para dar cuenta tanto de fenómenos terminológicos como de casos que conciernen a las unidades documentales.

Para empezar, un primer punto de similitud que, no por anecdótico, deja de ser interesante: Observemos que la unidad "documentación", al igual que la unidad "terminología", es una unidad polisémica. Utilizamos el término "terminología" para denominar tanto la materia que se ocupa de los términos como el conjunto de los términos, conjunto que constituye el objeto de la Terminología como disciplina. Este caso es idéntico al que tenemos en relación al término "documentación", que sirve para denominar tanto el campo de conocimiento como su objeto.

En segundo punto que hace similares la Terminología y la Documentación concierne al hecho de que ambos campos de conocimiento son disciplinas de reciente estabilización que han surgido de la práctica. Y ha sido a partir de la práctica que se han ido consolidando como campos científicos y haciéndose un hueco entre los campos del saber. Actualmente podemos decir que tanto la terminología como la documentación son materias reconocidas académica y profesionalmente como campos de conocimiento y de actividad necesarios en la sociedad del conocimiento y la información. El camino de su consolidación disciplinar.

Un tercer punto de similitud y tal vez el más relevante es que, aunque el objeto de ambas disciplinas sea distinto (en un caso las unidades terminológicas y en otro las unidades documentales), los dos objetos presentan la misma interdisciplinariedad constitutiva: se trata de unidades de conocimiento, de lenguaje (en sentido amplio) y de circulación social. Tanto la terminología como la documentación son campos de conocimiento interdisciplinarios constituidos por elementos procedentes de las ciencias cognitivas, de las ciencias del lenguaje y de las ciencias sociales. La unidad terminológica, objeto de la Terminología como disciplina, al igual que la unidad documental, objeto de la Documentación, son unidades interdisciplinarias constituidas a partir de elementos procedentes de estas tres vertientes del saber. Tomada esta interdisciplinariedad de las unidades en sentido geométrico, podemos decir que se trata de unidades poliédricas, es decir, de unidades que, como los poliedros, son figuras de distintas caras, aun tratándose de una sola unidad.

Y un último punto de similitud entre la Terminología y la Documentación que deseamos destacar es el hecho de que se trata en ambos casos de transdisciplinas. Ambas materias son transdisciplinares, en el sentido que están presentes en todas las demás materias de especialidad. No existe disciplina alguna sin terminología ni se concibe disciplina alguna sin documentación.

1.2. Documentación y terminología: ¿el mismo objeto u objetos diferentes?

Sin poner en cuestión los puntos que acabamos de presentar con el objetivo de mostrar las similitudes entre la Terminología y la Documentación como campos de conocimiento interdiscipliarios centrados en la descripción, análisis y tratamiento de objetos poliédricos, lo cierto es que el objeto de la documentación y el de la terminología no son el mismo objeto desde el punto de vista disciplinar, aunque puedan coincidir en los mismos materiales de análisis. El objeto de la terminología son las unidades terminológicas, y estas aparecen en los textos en tanto que construcciones lingüísticas; el objeto de la documentación, en cambio, lo constituyen las unidades documentales en tanto que unidades de información, aunque desde el punto de vista lingüístico puedan considerarse textos.

Ahora bien, el objeto de estudio de la Terminología y el de la documentación, dada su interdisciplinariedad coincidente, comparten las mismas perspectivas de análisis. Las unidades terminológicas y los textos en los que aparecen, son unidades al mismo tiempo cognitivas, sociales y lingüísticas, y, en consecuencia, pueden definirse desde estas tres perspectivas.

Definidas desde el punto de vista lingüístico, son unidades del léxico de las lenguas que adquieren un sentido preciso en una situación discursiva marcada pragmáticamen-

te. Como tales unidades del léxico pueden analizarse formalmente, semánticamente y funcionalmente. Como estructuras formales, se trata de unidades constituidas por uno o más morfemas léxicos, de los que por lo menos uno de ellos debe corresponder a un radical o raíz. Como unidades semánticas, poseen a un sentido preciso que puede analizarse a partir de combinaciones de rasgos o predicaciones. Y como unidades funcionales poseen una categoría gramatical que les permite estructurar a su alrededor unidades más amplias o participar en estructuras gramaticales de rango superior.

Definidas desde el punto de vista conceptual, las unidades terminológicas se conciben como unidades de conocimiento mínimas y autónomas ubicadas precisa y explícitamente en una estructura de conocimiento asociada a un campo del saber o a un campo de actividad, o bien a alguna de las perspectivas de un campo del saber o de una actividad. Dentro de esta estructura, representan un nodo de conocimiento pertinente. En tanto de unidades conceptuales pueden analizarse por su constitución interna, por su posición en la estructura y por las relaciones que mantienen con otras unidades conceptuales del mismo campo.

Y, como unidades sociales, se trata de unidades discursivas que, integradas en un texto, identifican a los individuos como miembros de grupos sociales científicos o profesionales, o de alguna corriente de opinión o pensamiento, y que les permiten expresar, transferir e interactuar socioprofesionalmente y adaptar su discurso a las condiciones pragmáticas en las que se lleva a cabo la transferencia.

Como puede deducirse fácilmente las unidades terminológicas son piezas mínimas, aunque centrales, en la articulación textual y discursiva, fundamentales cuando tratamos de discurso especializado. Las unidades terminológicas se combinan entre si y, junto a otras unidades lingüísticas, constituyen textos especializados. Y estos textos, tomados como unidades de información, constituyen el objeto de la documentación. El papel pues que juegan las unidades terminológicas en los documentos es muy preciso, aunque crucial: representan formalmente nodos de conocimiento específico dentro de una unidad informativa más compleja, por cuanto son las piezas más representativas del conocimiento especializado de un texto. Es por ello que el trabajo documental las toma como base para la descripción del contenido de los documentos. Así, las unidades terminológicas devienen en el ámbito de la documentación unidades de información, de importancia capital para representar el contenido de los documentos.

2. Realidad, representación, conceptualización y verbalización del conocimiento especializado

Un texto podría analizarse como el resultado de un proceso que arranca en la aprehensión de la realidad, continúa con la construcción mental de esta realidad a

través de la categorización y desemboca en la construcción de un discurso que se traduce en un texto.

2.1. El proceso

Trataré de analizar a continuación el proceso que presuntamente realiza un hablante desde la percepción de la realidad hasta su verbalización, sin entrar ni en los postulados neuropsicológicos que subyacen en este proceso ni en el tema de las identidades y diferencias entre los procesos de adquisición de conocimiento general en contraste o no con el especializado. Me interesa únicamente aquí establecer un marco en el que situar dos objetivos:

a) el primero, presentar el principio de poliedricidad como eje vertebrador de lo cognitivo, lo social y lo lingüístico en terminología,
b) y el segundo, explicar algunas de las relaciones que se establecen entre estos aspectos.

Para avanzar en esta línea, es preciso que distingamos tres planos de análisis de la terminología, en tanto que resultado:

- el plano referencial, que comprende el ámbito de los objetos y de la formación de clases de objetos
- el plano cognitivo, que incluye la formación de conceptos y su ubicación en la mente
- el plano lingüístico, o más ampliamente semiótico si integramos los signos de naturaleza artificial en la denominación "término", que comprende la realización de cada concepto en una o más unidades terminológicas.

Como han descrito ampliamente las ciencias cognitivas, la segmentación del contínuum de la realidad depende de mecanismos psicocognitivos muy complejos que no son ajenos a los valores culturales interiorizados por los hablantes de toda comunidad.

No tratamos de entrar a fondo en este punto, pero vamos a presentar muy someramente algunas fases del proceso de conceptualización-verbalización del conocimiento especializado con la finalidad de entender posteriormente cómo la categorización de la realidad conduce a la formación de un concepto de estructura compleja, lo que permite explicar el comportamiento y la variación en terminología.

Se sabe que los seres humanos "perciben" la realidad a través de filtros interiorizados. Estos filtros, de carácter psicológico, antropológico y sociológico, actúan de mediadores entre una supuesta "realidad por encima de los individuos" y la realidad percibida por los seres humanos que pertenecen a comunidades y grupos determinados y se ubican en contextos históricos y sociales. La realidad pura, pues, está fuera de nuestra comprensión como seres sociales.

Si asumimos este punto de partida, podemos decir que nos situamos en una realidad percibida socioculturalmente y discriminada psicológicamente en forma de "clases objetos", es decir, conjuntos de objetos individuales que se han agrupado en una clase por el hecho de, según unos, compartir determinadas características que actúan como catalizadoras en función de su centralidad en cada grupo humano, y, según otros, por el hecho de asociarse analógicamente a modelos previamente construidos.

2.2. Dos escenarios

Partiendo del supuesto de que los mecanismos de percepción de la realidad y de construcción de clases de objetos en las especialidades no son distintos de los que sirven para explicar el proceso de adquisición del conocimiento en general, vamos a tomar dos escenarios de conocimiento distintos con la finalidad de dar cuenta del proceso de adquisición, estructuración y transmisión de conocimiento especializado:

- un primer escenario de producción de conocimiento nuevo (ESC1);
- un segundo escenario de adquisición de conocimiento producido anteriormente, a través de la recepción de información vía discurso (ESC2).

En el primer escenario de producción de conocimiento nuevo (ESC1), lo que el individuo percibe, discrimina y clasifica está condicionado por sus conocimientos previos y por sus necesidades, en el marco del rol que se atribuye el individuo en cada proceso. Precisando un poco más, en los escenarios en los que un individuo, cuyo rol es el de un especialista en la materia, se acerca a la realidad lo hace desde una competencia previa (conocimientos adquiridos anteriormente desde este rol) y movido por la necesidad de descubrir nuevos fenómenos o de percibir nuevas características o relaciones de fenómenos ya anteriormente percibidos, que le permiten avanzar en su conocimiento.

En un contexto de "descubrimiento" [1] los individuos "perciben" por primera vez un fenómeno nuevo o una propiedad nueva de un fenómeno ya observado, o establecen una nueva relación entre dos fenómenos; es decir, perciben una nueva parcela del contínuum de la realidad sobre la base de la observación de los datos empíricos. Esta observación les conduce a una constatación de que se trata de algo nuevo; y así, en el largo proceso de observación y análisis, van caracterizando y estabilizando dicho conocimiento a través de operaciones cognitivas consistentes en la abstracción y jerarquización de las características esenciales y periféricas y usando a menudo el discurso como herramienta de progresión. Una vez estabilizada la discriminación de la nueva unidad de conocimiento, se categoriza en la mente como un nuevo concepto asociado a este fragmento de realidad percibido.

Observemos que hasta aquí hemos establecido dos variables esenciales. La primera

variable es el esquema o marco situacional del que el individuo parte, que incluye sus condiciones inherentes (sexo, edad, época, grupo social, grupo profesional, especialidad, escuela de pensamiento, nivel de competencia en el tema, función social, etc.), lo que determina su rol en este proceso de adquisición de conocimiento, en forma de activación de filtros. La segunda variable la constituyen los mecanismos psicocognitivos que explican cómo un individuo dentro de un marco situacional es capaz de crear unidades de conocimiento en forma de conceptos a partir de una realidad que es un contínuum.

El proceso que se produce en el segundo escenario, que hemos caracterizado como de adquisición de conocimiento producido anteriormente a través de la recepción de información vía discurso (ESC2), es bastante distinto del anterior. En este caso no se trata de la adquisición de conocimiento nuevo a partir de la observación de la realidad, sino de la "construcción" de conocimiento a partir de conocimiento ya estabilizado, expresado discursivamente. Lo que el individuo activa en este contexto son los mecanismos que le permiten "detectar", a través siempre de filtros individuales y culturales, conocimiento contenido en el texto y formulado a través de estructuras lingüísticas. En este proceso de detección hay que tener en cuenta las condiciones previas inherentes de quien lo lleva a cabo, sus intenciones y las finalidades para las que lo lleva a cabo. Todo este conjunto de condiciones las hemos definido en el escenario anterior como el marco en que el individuo se sitúa, que, como también hemos visto en el primer escenario, mediatiza sus posibilidades y actividades. Lo que diferencia claramente este segundo escenario del anterior es que el individuo no "construye" un nuevo conocimiento a través de la cognición y el discurso, sino que trata de detectar unidades de conocimiento contenidas como tales en un texto. Se basa pues en el análisis del texto, y no en el análisis de la realidad, para adquirir conocimiento. Trata de "descubrir" las unidades de conocimiento que presumiblemente "están" en el texto. Se trata pues de un proceso claramente semasiológico, de un mecanismo de adquisición de conocimiento a partir de un proceso de descodificación mediatizado.

Este nuevo conocimiento se ubica en la mente humana en forma de unidades conceptuales más o menos complejas. El proceso y la forma de ubicación son temas que escapan a nuestra propuesta, pero de los que podemos decir que están condicionados por las hipótesis psicocognitivas o neurocognitivas de las que se parte. Desde el punto de vista neurocognitivo se han lanzado dos hipótesis: la primera afirma que cada tipo de conocimiento está localizado en un lugar de la mente, la segunda propone que el conocimiento no se agrupa por tipos sino que se encuentra distribuido en la mente.

Se abren así una serie de interrogantes sobre cómo se estructura en la mente el conocimiento adquirido y qué modelo puede dar cuenta más apropiadamente de esta estructuración: un modelo jerárquico o un modelo en red. No es nuestro objetivo aquí responder a estos interrogantes, sino únicamente dar cuenta de que la información

extraída de la realidad en las condiciones que hemos descrito, y ubicada como clase o categoría en la mente, establece relaciones con el resto de categorías también interiorizadas, y que estas relaciones no son bidimensionales sino multidimensionales, de forma que un concepto interiorizado establece relaciones de mayor o menor intensidad o de mayor o menor cercanía con la totalidad de los conceptos de la mente.

3. Hacia una explicación a través del principio de poliedricidad

3.1. La descripción de los conceptos

Uno de los puntos que han inquietado y sigue inquietando a quienes se ocupan del conocimiento es cómo se describen los conceptos. Es evidente que desde nuestra posición lingüística no podemos abordar los conceptos directamente y extraer conclusiones que sean empíricamente adecuadas, pero sí podemos hacer inferencias sobre las características que presumiblemente poseen y las condiciones que cumplen, a través de la observación de las unidades terminológicas que representan conceptos en el discurso especializado.

Pongamos pues a la vista algunos de los datos que pueden guiar nuestro análisis del concepto:

- Sabemos que en el discurso un término lo es porque tiene un sentido preciso en un ámbito de conocimiento determinado y es usado significativamente en el discurso especializado de este ámbito.
- Sabemos también que este término se corresponde con un concepto en la mente del especialista en el tema, porque cuando usa un término lo asocia a una unidad conceptual precisa dentro de su ámbito de especialidad.
- Sabemos además que alguien es especialista cuando conoce una especialidad, y que conocer una especialidad presupone tener una estructura conceptual general de cada saber en la mente, reconocida, aunque con posibles variantes o alternativas, por la comunidad experta, de forma que el individuo considerado especialista es capaz, por un lado, de expresar conocimiento sobre la materia de acuerdo con esta estructura variada aunque estabilizada, y, por otro lado, reconocer discursos especializados no solo por su expresión y formato, sino sobre todo por su contenido. Asimismo es capaz de detectar transgresiones conceptuales que un no experto nunca percibiría.
- También sabemos que existe variación terminológica para representar un mismo concepto, lo que se conoce genéricamente como sinonimia cuando se trata de una variante léxica. Esta variación no solo la reconocen los especialistas sino que son ellos mismos quienes la producen.

- Sabemos que una misma unidad de conocimiento puede aparecer en las producciones de los especialistas no solo denominada de manera diferente (sinonimia) sino también "explicada" o expuesta discursivamente de manera diferente, ya sea en bloque, ya sea introduciendo progresivamente aspectos distintos de la unidad conceptual.
- Finalmente sabemos que una unidad formal puede usarse asociada a más de un sentido, dentro de un mismo ámbito temático o en distintos ámbitos temáticos, lo que se conoce genéricamente como polisemia u homonimia en función de la posibilidad o imposibilidad de explicar la variación semántica desde una sola unidad [2].

3.2. Conceptos y términos en el proceso de adquisición, estructuración y erbalización

Lo que podemos inferir de los datos del discurso sobre las relaciones entre realidad, objeto, concepto y término, puede resumirse en los siguientes puntos:

a) En primer lugar, tenemos cuatro planos de análisis interrelacionados:
b) El plano de los objetos individuales
c) El plano de las clases de objetos
d) El plano de los conceptos
e) El plano de los términos

A partir de los objetos individuales, siempre plural, construimos clases de objetos. En este proceso nos interesa destacar dos posibilidades:

a) que de un conjunto de objetos construyamos una clase de objetos
b) que de un conjunto de objetos construyamos más de una clases de objetos.

Situados en el plano de los conceptos, se nos abren dos posibilidades:

a) que a partir de una clase de objetos construyamos un concepto
b) que a partir de más de una clase de objetos construyamos más de un conceptoun concepto unitario construyamos más de una clase.

Y situados en el plano de los términos, se nos ofrecen tres posibilidades:

a) que a partir de una clase de objetos construyamos un concepto y este se exprese a través de una sola unidad terminológica
b) que a partir de más de una clase de objetos construyamos más de un concepto, cada uno de los cuales se exprese a través de una unidad terminológica distinta, aunque formalmente idéntica (homonimia)
c) que a partir de una clase de objetos construyamos un concepto y este se exprese a través de más de una unidad terminológica (variación denominativa o sinonimia).

Y, finalmente, en este último caso, estas variantes denominativas pueden conducirnos a dos situaciones:

a) que la variación denominativa no tenga consecuencias cognitivas
b) que la variación denominativa tenga consecuencias cognitivas.

Veamos en el siguiente cuadro un resumen ejemplificado de este panorama:

Plano de los objetos individuales	**Plano de las clases de objetos**	**Plano conceptual**	**Plano terminológico**
Conjunto de objetos individuales	un objeto discriminado	un concepto "seno del complemento de un ángulo o de un arco"	un término *coseno*
Conjunto de objetos individuales	varios objetos discriminados	varios conceptos "estado patológico que se caracteriza por la pérdida de la conciencia, la sensibilidad y la capacidad motora voluntaria" "signo ortográfico que indica una pausa breve"	varios términos distintos o formalmente iguales (homonimia): *coma* (signo ortográfico) *coma* (estado patológico)
Conjunto de objetos individuales	un objeto discriminado	un concepto "afección cutánea caracterizada por vesículas rojizas y xudativas, que dan lugar a costras y escamas"	varios términos (sinónimos) sin consecuencias cognitivas *eccema* *eczema* *erupción*
Conjunto de objetos individuales	un objeto discriminado	un concepto "difusión mundial de modos, valores o tendencias que fomenta la uniformidad de gustos y costumbres"	varios términos (sinónimos) con consecuencias cognitivas *mundialización* *globalización*

3.3. La variación denominativa

A pesar de la definición de concepto dada por las mormas ISO [3], los avances en ciencia cognitiva han mostrado que un concepto no puede definirse como una estructura cerrada de rasgos definitorios, sino como una categoría que se enmarca en un esquema de rasgos que la definen. Nuestro propósito es, por un lado, analizar las relaciones que se pueden establecer entre unidades de categoría conceptual y unidades terminológicas y así poder explicarnos cómo un concepto puede representarse en el plano lingüístico a través de más de un término.

Si retomamos el cuadro del apartado anterior, podemos decir que de los cuatro casos presentados lo siguiente:

El primer tipo de caso es muy poco frecuente: el término *coseno* es un término puramente matemático cuyo significado representa un concepto en su globalidad y corresponde a un objeto "construido" desde la realidad en un esquema determinado (el de la Matemática).

El segundo caso no merece más atención que decir que representa la totalidad de los casos de homonimia, es decir, de realidades verdaderamente distintas y no relacionadas entre sí, que han generado conceptos claramente distintos cada uno de los cuales sí que se representa mediante un término, presentando estos distintos términos una coincidencia formal entre si. La única relación que guardan estos casos es en el plano del término y aun únicamente por el hecho de que haya una coincidencia formal, y solo formal, entre las denominaciones.

Los casos que merecen más atención son los dos últimos. Ambos intentan ejemplificar cómo un objeto puede representarse verbalmente a través de distintos términos [4]. En esta relación pueden darse dos casos:

a) que este objeto corresponda a un solo concepto, expresado por diferentes signos semánticamente coincidentes, aunque formalmente diferentes (caso nº 3),
b) que este objeto corresponda a un solo concepto, expresado por diferentes signos semánticamente y formalmente diferentes (caso nº 4).

En el caso número 3 pueden incluirse ejemplos de índole distinta: por una parte variantes puramente gráficas (*eczema-eccema*), por otra, variantes morfológicas (*alboroto-alborotamiento; condiciones climáticas-condiciones climatológicas*) y en tercer lugar variantes léxicas (*bencina-esencia-gasolina*). La característica común que comparten es el hecho de que usar una u otra variante no cambia la manera cómo se representa el concepto, o dicho de otro modo, la manera cómo el concepto se proyecta en la denominación. En todos los casos tan opaca o transparente es una variante como otra, y, desde el punto de vista cognitivo, el uso de una u otra variante no introduce cambios en la proyección del concepto en el discurso ni presupone tampoco una intención diferente (a no ser que sea estilística o extralingüística) por parte del locutor.

Muy distinto es en cambio el caso número 4, que podría complementarse con otros ejemplos como:

agricultura ecológica-agricultura biológica
mundialización-globalización
residuos biosanitarios-residuos hospitalarios
contaminación del aire-contaminación atmosférica

En todos estos ejemplos podemos ver que el uso de una variante u otra para designar un mismo objeto podría estar condicionado por una intención cognitiva por parte del locutor y, evidentemente, tiene consecuencias cognitivas en el receptor. Se trata de dos signos formados sintagmáticamente a partir de un núcleo común (residuos) y cuyo significado coincide parcialmente. Lo que se trata de saber, sin embargo, es si

esta sinonimia parcial puede explicarse a partir de un único concepto o si cada sentido de un término remite a un concepto distinto aunque haya tantos puntos en común entre ambos

3.4. La poliedricidad en el trasfondo de la variación denominativa y como principio explicativo

Retomando ahora la concepción de la unidad terminológica como un signo compuesto de forma (denominación) y contenido (significado), lo que cabe preguntarnos es cómo es la relación entre los conceptos y los términos a la vista de que una misma categoría conceptual puede proyectarse en distintos términos, cada uno de los cuales puede transmitir un sentido distinto. Estos términos son *grosso modo* sinónimos, pero no en su totalidad, ya que a través de la denominación nos hacen percibir una parte o una faceta distinta del mismo concepto que representan.

En el trasfondo de esta digresión, si asumimos la concepción de la unidad terminológica como signo compuesto de forma y contenido, se esconde una cuestión clave para explicarnos la naturaleza de la unidad terminológica y su papel en todo el proceso de construcción y verbalización del conocimiento: mientras que la denominación se entiende claramente como la unidad formal que verbaliza el concepto, no es tan clara la relación que se da entre el significado y el concepto. ¿Se trata de una misma noción o de nociones diferentes? ¿Son lo mismo el significado y el concepto?

La concepción del concepto como una estructura compleja, desarrollada dentro de la teoría del prototipo, deja fuera de la explicación la multidimensionalidad del concepto, en el sentido de que las características que lo describen pueden agruparse en conjuntos, cada uno de los cuales puede representar una faceta del mismo y a menudo cada faceta puede estar determinada por un criterio o dimensión. Nuestra propuesta puede sintetizarse en los siguientes puntos:

- el concepto es una estructura compleja, plural en cuanto a características y facetada en cuanto a sus dimensiones;
- el concepto puede proyectarse globalmente en un término, pero también puede proyectarse en distintos términos y resultar todos ellos sinónimos;
- el concepto puede proyectarse en uno o más términos no motivados, pero puede proyectarse también en distintos términos motivados, que presentan entre si variación formal y pueden tener sentidos distintos, ya que la proyección de cada faceta de un concepto en un término comporta un cambio de sentido;
- la ocurrencia de un término en un enunciado puede entenderse como una "instanciación" del concepto;

- la variación denominativa se repara en el discurso; la variación denominativa con consecuencias cognitivas puede expresarse a través de la forma denominativa en sí o por pistas discursivas que actúan de indicios de variación conceptual;
- estos indicios pueden ser de muchos tipos y pueden expresarse mediante distintas formas gramaticales.

El principio de poliedricidad, según el cual los términos son unidades interdisciplinarias integradas por vertientes o facetas distintas, cada una de ellas correspondiente a un plano de análisis, nos permite dar cuenta de la compleja relación entre un concepto y sus diferentes realizaciones terminológicas. Hasta ahora sólo habíamos aplicado este principio a las unidades terminológicas, pero creemos que su alcance es mucho más amplio y puede aplicarse también al concepto, por cuanto es desde su carácter multifacético o poliédrico que podemos explicar los casos de variación denominativa con consecuencias cognitivas. Estos casos no son más que proyecciones en el plano lingüístico de la poliedricidad conceptual.

La distinción de planos de análisis, aparte de resultar metodológicamente potente, es coherente con el postulado del carácter interdisciplinario de la terminología y de la posibilidad de abordarla desde diferentes ópticas, como ha sido representado a través del modelo de las puertas (Cabré 2000, 2003).

Note

[1] Hablamos de "descubrimiento" sólo en sentido metafórico.

[2] En la lingüística más ortodoxa, la homonimia se reserva para las unidades que proceden de formas etimológicas distintas. No serían homónimas por tanto las unidades cuyo sentido se hubiera generado por restricción, ampliación o cambio de un significado de base.

[3] Unit of knowledge created by a unique combination of characteristics. (ISO-1987-1 2000).

[4] Queremos recordar que en este trabajo consideramos término o unidad terminológica aquella unidad sígnica dotada de forma (que corresponde a la denominación) y contenido (que corresponde al significado).

Bibliografia

Adelstein A. (2007). *Unidad léxica y significado especializado: modelo de representación a partir del nombre relacional "madre"*. Barcelona: Institut Universitari de Lingüística Aplicada. [Tesis doctoral]

Cabré M.T. (2000). "Terminologie et linguistique: la théorie des portes". *Terminologies nouvelles. Terminologie et diversité culturelle* 21. 10-15.

Cabré M.T. (2003). "Theories of terminology. Their description, prescription and explanation". *Terminology* 9 (2). 163-200.

Cabré 2005: Cabré M.T. (2005). "La Terminologia, una disciplina en evolución: pasado, presente y algunos elementos de futuro". *Debate Terminológico* 1 (1).

Freixa J. (2002). *La variació terminològica: anàlisi de la variació denominativa en textos de diferent grau d'especialització de l'àrea de medi ambient*. Barcelona: Institut Universitari de Lingüística Aplicada. [Tesis doctoral]

ISO 1087-1: (1997). *Part 1: theory and applications*.

ISO 1087-1: (2000). *Part 1: theory and applications*.

ISO 1087-2 (1997). *Part 2: computer applications*.

ISO 1087-2 (2000). *Part 2: computer applications*.

Kuguel I. (2007). *La semántica del léxico especializado: los términos en textos de ecología*. Buenos Aires: Facultad de Filosofía y Letras, Universidad de Buenos Aires. [Tesis doctoral]

Lakoff G. (1987). *Women, fire, and dangerous things: what categories reveal about the mind*. Chicago [etc.]: University of Chicago Press.

Leech G.N. (1990). *Semantics: The Study of Meaning*. 2ª edición. Londres: Penguin.

Rosch E. (1978). "Principles of categorization". En Rosch, E.; Lloyd, B.B. (ed.). "Cognition and Categorization". Hillsdale (NJ): Lawrence Erlbaum Associates. 27-48.

Tebé C. (2005). *La representació conceptual en terminologia: l'atribució temàtica en els bancs de dades terminològiques*. Barcelona: Institut Universitari de Lingüística Aplicada. [Tesis doctoral]

Temmerman, R. (2000). *Towards New Ways of Terminology Description: The Socio-cognitive-Approach*. Amsterdam, Philadelphia: John Benjamins.

Wüster E. (1979). *Einführung in die Allgemeneine Terminologielehre und Terminolische Lexicographie*. Viena: Springer. (Versión española: Cabré, M.T. (dir.) (1998). *Introducción a la teoría de la terminología y a la lexicografía terminológica*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra).

Repérage de la référence à partir du thésaurus, de la terminologie et de la sémantique lexicale

LAURENCE KISTER, EVELYNE JACQUEY, BERTRAND GAIFFE

These paper seeks to draw a parallel between the thematic structure of specialized text and the structure of a thesaurus or a terminology. We try to annotate semantically, in an automatic manner, the thesaurus in order to track automatically the reference chains of texts. For this purpose, we suggest the annotation of the thesaurus, the terminology and the text. The semantically annotated thesaurus and the terminology will be used to annotate the lexical items of reference chains and to determine the thematic structure of the documents. The thesaurus and the terminology are used as representations of the knowledge of one particular domain and are associated to lexical semantics to realize the annotation and the analysis of the text's contents.

Keywords: thematic structure – reference chain – thesaurus – terminology – lexical semantic

1. Introduction

Les travaux présentés se trouvent à l'intersection de la terminologie et de la sémantique lexicale: ces disciplines peuvent entretenir des collaborations fructueuses dans les domaines de spécialité et dans l'étude des textes qui relèvent de ces domaines. Une terminologie est une représentation des connaissances propre à un domaine défini et délimité. Nous la considérons comme une liste de contrôle qui permet de repérer les termes désignant les concepts dans les textes. La sémantique lexicale, lorsqu'elle s'attache à l'analyse d'un domaine de spécialité, s'intéresse à l'intersection entre les unités lexicales signifiantes et les concepts spécifiques au domaine.

A l'interface entre sémantique lexicale, thésaurus et terminologie, nous cherchons à déterminer si les concepts utilisés pour structurer un domaine de spécialité sont utilisés pour structurer les textes de ce domaine et, si c'est le cas, de quelle manière la structuration s'opère. Les travaux entrepris portent sur l'existence de liens entre la structure du discours scientifique et technique (mention des concepts) et la hiérarchie établie par le thésaurus et ou la terminologie dans la mesure où ceux-ci s'apparentent à des représentations des connaissances dans un domaine fini à l'aide d'un langage de spécialité.

Dans les textes, les chaînes de références contribuent à l'identification des formes que prennent les désignations des concepts. Parmi elles, apparaissent des formes nomi-

nales correspondant aux termes, au sens où l'entend la terminologie (Cabré, 1998; Béjoint et Thoiron, 2000): un terme désigne un et un seul concept dans un domaine déterminé. La confrontation de ces formes et de leur organisation dans les textes avec la structure du thésaurus vise à identifier en quoi la lexicalisation des concepts dans les textes s'appuie sur la distinction générique/spécifique comme c'est le cas dans les thésaurus. Cette confrontation repose sur l'hypothèse que plus un terme est à un niveau élevé de la hiérarchie dans le thésaurus ou correspond à un domaine ou un sous-domaine de premiers niveaux en terminologie, plus il est générique et plus les formes nominales qui lui correspondent sont abstraites du point de vue de la distinction [+concret]/[+abstrait] en sémantique lexicale.

La mise en oeuvre de cette confrontation passe par l'annotation des formes nominales en [+concret] et [+abstrait]. Cette annotation repose sur l'exploitation d'un lexique d'emploi de substantifs étiquetés [+concret] issu d'une acquisition automatique à partir du Trésor de la Langue Française informatisé (TLFi) (pour la procédure d'étiquetage voir Kister et Jacquey 2006). Munis des formes nominales qui correspondent aux termes présents dans un thésaurus, et de l'annotation de ces termes selon la distinction [+concret]/[+abstrait], il est possible de mettre en œuvre la confrontation, dans un domaine de spécialité, entre la structure thématique d'un texte – telle qu'elle peut être identifiée par le biais des chaînes de référence – et la structure des termes dans le thésaurus ou la terminologie.

La confrontation des unités linguistiques des textes et des termes issus des terminologies et/ou des thésaurus vise à révéler les similitudes et les distorsions entre l'ordre d'apparition dans le texte et le thésaurus. A terme, nous cherchons à déterminer si les textes sont structurés selon une logique similaire à celle des thésaurus et des terminologies, si les points communs ou les divergences dans les deux types de documents facilitent ou entravent les activités d'analyse et d'indexation.

2. Sémantique lexicale, terminologie et thésaurus: des approches complémentaires

Bien qu'elles abordent la désignation des concepts et la représentation des connaissances d'un domaine de manières différentes la sémantique lexicale, la terminologie et les techniques documentaires qui permettent la création de thésaurus ne doivent pas être perçues comme antinomiques. Elles doivent être considérées comme trois approches qui se complètent et qui méritent d'être confrontées afin de tirer le meilleur parti de chacune d'elles.

La différence entre la terminologie et le thésaurus bien qu'elle paraisse évidente est souvent relativement minime. Le thésaurus parfois défini comme un lexique hiérarchisé ou un vocabulaire normalisé propre à un domaine n'a pas pour fonction de définir les

termes qu'il organise. C'est un outil documentaire d'indexation qui permet de faire une description du contenu d'un document en texte intégral. Le thésaurus est un ensemble structuré de termes destiné à faciliter la description d'un domaine, à harmoniser la diffusion, la gestion et le traitement de l'information relative à un domaine. Les termes ou *descripteurs* doivent être le moins polysémiques possible ce qui a pour conséquence d'écarter au maximum la synonymie et la quasi-synonymie. Le thésaurus peut aussi être présenté comme un vocabulaire contrôlé: il dresse un inventaire des termes et les hiérarchise pour rendre compte de l'organisation du domaine. Cette organisation se matérialise par la présence de termes de différentes natures. D'une part, les génériques désignent les concepts principaux tandis que les spécifiques précisent les concepts particuliers propres à un concept principal. D'autre part, les équivalents correspondent à des variantes de certains spécifiques dans certains domaines ou sous-domaines et permettent de tenir compte de l'usage sur le terrain tout en préservant l'utilisation des spécifiques du thésaurus. Enfin, les associés rendent compte des liens de causalité, de localisation, de temps, etc. et servent à affiner la description ou à préciser les requêtes lors de la recherche d'information. Une telle organisation hiérarchique se retrouve en partie dans la structure d'une terminologie.

La terminologie peut se définir comme la discipline scientifique qui étudie les vocabulaires spécialisés et qui en analyse les conditions d'utilisation. Elle est souvent perçue comme l'étude des vocabulaires de spécialité donc des termes propres à des domaines de connaissance délimités. Certains termes peuvent être communs à la langue courante et à la langue du domaine de spécialité et, de ce fait, faire l'objet d'une entrée tant dans un dictionnaire de langue que dans une base de données ou banque de données terminologique. La terminologie est souvent perçue comme une discipline au service d'autres disciplines: traduction, enseignement des langues étrangères, normalisation, rédaction scientifique et technique, etc. Elle partage avec l'ontologie la notion de concept: un terme est l'association d'une désignation (entité linguistique) et d'un concept qui en exprime la signification.

Elle se présente généralement sous forme de fiches contrairement au thésaurus et à l'ontologie qui prennent souvent la forme d'arborescences. Les concepts font l'objet d'une description systématique qui recense différents types d'informations dont les plus fréquentes sont: le terme utilisé pour désigner le concept, la définition du concept via celle du terme utilisé, le domaine et les sous-domaines d'emploi du terme, les éventuelles variantes (orthographiques ou synonymes de type géographique, par exemple) et les équivalents dans d'autres langues. Parfois, une note et des exemples d'emploi apportent des informations complémentaires.

En sémantique lexicale, certaines caractéristiques du sens des lexèmes peuvent être représentées par des traits sémantiques usuels ([+animé], [+inanimé], [+humain, [+abstrait], [+concret], etc.). Ainsi, un lexème comme *transcription* peut être codé de

différentes manières: il peut rendre compte de l'action de transcrire et être marqué [+processus] ou correspondre à ce qu'on obtient à l'issue du processus de transcription et être marqué [+résultat]. Dans le premier cas, il sera porteur du trait [+abstrait] et dans le second du trait [+concret]. La sémantique lexicale est aussi l'analyse des relations sémantiques entre les lexèmes d'un lexique d'une langue. Elle s'intéresse alors à des relations telles la méronymie, la synonymie, l'hyperonymie, l'hyponymie qui servent de fondement à la hiérarchie du thésaurus et à l'organisation en domaines et sous-domaines en terminologie.

Les rappels que nous venons d'effectuer font émerger l'intérêt commun de la terminologie, du thésaurus et de la sémantique lexicale pour la description des concepts via les unités lexicales qui les désignent. Les traits de la sémantique lexicale se retrouvent en partie dans les domaines et les sous-domaines de la terminologie et la structure hiérarchique du thésaurus même s'ils ne sont pas matérialisés en tant que tels. La sémantique lexicale s'intéresse au sens du terme, à sa granularité en fonction des différents emplois tandis le thésaurus s'intéresse au sens du terme lorsqu'il est fils d'un générique (un "super" générique et des génériques "intermédiaires") et la terminologie au sens du terme dans un domaine et dans un ou plusieurs sous-domaines. Pour la sémantique lexicale, le sens découle de l'emploi qui est fait du terme dans un contexte. Pour la terminologie et le thesaurus le sens est le résultat de la mise en place d'un contexte spécifique dans un domaine particulier.

3. Structures thématiques des documents primaires et des documents secondaires

Avant de comparer les structures thématiques des différents types de documents, il nous paraît intéressant de rappeler en quoi consistent les chaînes de références, le thésaurus et les micro-thésaurus.

3.1. Chaîne de référence

Une définition rapide de la chaîne de référence consiste à la présenter comme une succession d'unités linguistiques qui désignent un même référent dans un document primaire (lexème, pronoms sujet et objet, relatifs, etc.). Parmi ces unités figurent, comme signalé ci-dessus, des termes du thésaurus ou de la terminologie. Une autre partie des chaînes de référence correspond aux relations anaphoriques qu'elles soient nominales ou pronominales (Amsili et Roussarie, 2005; Boudreau et Kittredge, 2005; Kleiber, 1994; Mitchell, Cuetos et Zagar, 1990; De Mulder et Schnedecker, 2001; Kleiber , Schnedecker et Tyvaert, 1997; Schnedecker, 1998). Afin d'illustrer ce qu'est une chaîne de référence nous examinerons l'exemple d'une chaîne de référence autour

du terme *phonétique* extrait du Cours de linguistique générale de Saussure (Kister et Jacquey 2007b).

La physiologie des sons est souvent appelée "**phonétique**" **CE TERME** nous semble impropre; nous **LE** remplaçons par **CELUI DE** phonologie. Car **phonétique** a d'abord désigné et doit continuer à désigner **l'étude des évolutions des sons**; l'on ne saurait confondre sous un même nom ***deux études*** absolument distinctes. **La phonétique** est **une science historique**; **elle** analyse des événements, des transformations et **se** meut dans le temps. La phonologie est en dehors du temps, puisque le mécanisme de l'articulation reste toujours semblable à lui-même. Mais non seulement ***ces deux études*** ne se confondent pas, ***elles*** ne peuvent même pas s'opposer. **La première** est **une des parties essentielles de la science de la langue**.

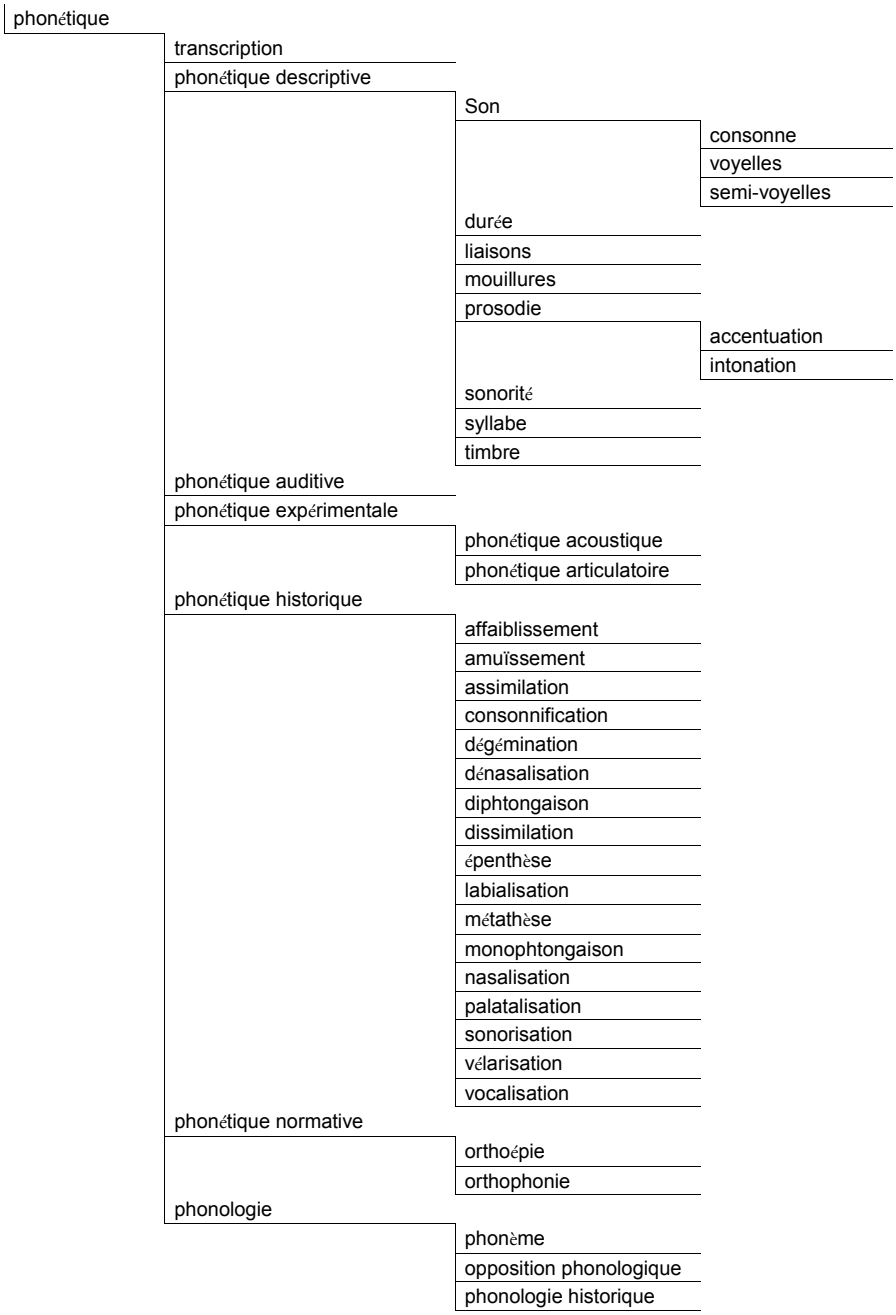
Cet exemple nous montre que le terme *phonétique* – présent dans Thésaulangue (thésaurus mis au point et utilisé par le centre de documentation de l'ATILF) – constitue le terme générique du micro-thésaurus de la phonétique et qu'il est présent dans la chaîne de référence. Il commute avec d'autres entités linguistiques ou syntagmes nominaux tels *physiologie des sons*, *terme, étude*, *une des parties essentielles de la science du langage*, *science historique* mais aussi avec des anaphoriques comme *le*, *celui de*, *elle, la première*. Cet exemple illustre aussi le fait que le terme générique n'est pas le point de départ de la chaîne mais qu'il correspond à la seconde occurrence de la chaîne. Le terme *phonétique* est présent trois fois et il alterne avec des unités lexicales et des groupes nominaux ainsi que des pronoms personnels et des relatifs. Pour l'analyse qui nous intéresse nous nous attachons dans un premier temps aux termes présents dans le micro-thésaurus de la phonétique et aux substantifs avec lesquels ces termes alternent.

3.2. Micro-thésaurus de la phonétique

Pour examiner les similitudes et les divergences de structuration des concepts dans les documents en texte intégral et dans les documents secondaires (thésaurus et terminologie), il est nécessaire de s'intéresser à l'apparition des concepts dans les chaînes de référence et à leur structure hiérarchique dans le thésaurus et la terminologie. A cet effet, nous proposons ci-dessous le micro-thésaurus de la phonétique tel qu'il apparaît dans Thésaulangue.

Bien que le micro-thésaurus utilisé pour les tests se limite à quatre niveaux de profondeur hiérarchique, il est suffisant pour débuter et évaluer la faisabilité des comparaisons que nous souhaitons effectuer.

Fig 1 - Micro-thésaurus de la phonétique de Thésaulangue



3.3. Confrontation de la structure thématique du texte intégral et des thésaurus ou terminologies

L'expérimentation que nous avons réalisée consiste à identifier en quoi la lexicalisation des concepts présents dans un texte s'appuie sur la distinction générique/spécifique. Elle repose sur une série d'hypothèses:

a) les chaînes de référence du texte intégral en matérialisent la structure thématique et contiennent des unités linguistiques lexicales dont certaines sont des termes de Thésaulangue,
b) le thésaurus et la terminologie ont la propriété de hiérarchiser les termes,
c) chaque terme correspond à un seul concept,
d) un terme générique du thésaurus correspond à un domaine de la terminologie et à une forme lexicale marquée [+abstrait] en sémantique lexicale.

L'expérimentation vise à examiner si les termes présents dans le texte, ou plus précisément dans les chaînes de références, apparaissent dans un ordre identique, un ordre proche ou un ordre différent de celui dans lequel ils apparaissent dans Thésaulangue. Afin d'examiner la structure thématique, nous avons choisi d'annoter sémantiquement les termes du thésaurus en [+concret] et [+abstrait] afin de vérifier si les [+concret] correspondent plutôt à des spécifiques, c'est-à-dire à des termes situés à droite dans l'arborescence et si les [+abstrait] s'apparentent aux génériques donc aux termes présents dans la partie gauche de l'arborescence. Pour tester l'annotation semi-automatique nous avons annoté le micro-thésaurus de la phonétique de Thésaulangue en utilisant un lexique d'emploi de substantifs [+concret] issu d'une acquisition automatique à partir du TLFi (32200 emplois pour 13300 substantifs) réalisé au cours de travaux antérieurs (Kister et Jacquey, 2006).

Les résultats de l'annotation semi-automatique ont été supervisés afin d'en vérifier la validité. Celle-ci est satisfaisante malgré certaines erreurs dues à:

- la présence de substantifs ambigus conduisant à la sélection d'un sens inadéquat lors de la constitution du thésaurus comme c'est le cas pour *accentuation* qui peut être interprété comme un processus ou un résultat,
- l'existence d'indices lexicaux introuvables en raison de leur position éloignée de l'entrée dans la définition comme pour *consonne* où l'indice *mot* est le dernier élément de la définition alors que nous cherchons les indices dans les trois premiers mots de la définition,
- la présence d'indices lexicaux dans l'article sans que ceux-ci se situent dans la définition comme le montre l'exemple d'*assimilation* pour lequel on trouve les indices dans la remarque,
- la présence d'indices lexicaux ambigus car polysémiques comme *radical,*
- l'existence de vedettes catégoriellement ambiguës codées en adjectif et substantif comme *dérivatif* et *complétif*.

Les termes marqués [+abstrait] sont selon nos hypothèses dans la partie gauche du thésaurus et correspondent à des génériques tandis que les [+concret] sont présents dans la partie droite du thésaurus et s'apparentent à des spécifiques.

3.3.1. Répartition des termes du micro-thésaurus phonétique dans les documents primaires

Afin de repérer les termes du micro-thésaurus et d'en évaluer la position dans les chaînes de référence et la place dans la structure thématique des documents en texte intégral, nous avons sélectionné trois documents disponibles dans la base Frantext catégorisée: *Cours de linguistique générale* (Saussure, 1968), *La linguistique* (Perrot, 1957) et *Le langage et la vie* (Bally, 1925).

Les résultats des comptages montrent que 22 termes présents dans le micro-thésaurus *phonétique* produisent 557 occurrences ce qui nous permet d'affirmer que les termes du thésaurus sont employés dans les documents primaires. L'examen de la répartition des termes montre que les termes les plus représentés ne sont ni les plus génériques ni les plus spécifiques. Dans le tableau ci-dessous le niveau 1 correspond au terme *phonétique*, le niveau deux à ses fils, le 3 à ses petits-fils et le 4 à ses arrières petits-fils.

Tab. 1 - Nombres d'occurrences des termes de phonétique en fonction de leur position dans le thésaurus

Niveau 1	Niveau 2	Niveau 3	Niveau 4
26	31	403	95

La répartition est identique pour les micro-thésaurus de la morphologie, de la sémantique et de la syntaxe (Kister et Jacquey 2007a et b). Ces résultats nous permettent de conclure que certains termes de Thésaulangue sont présents dans les documents examinés. Ils nous permettent de constater que les termes des niveaux intermédiaires sont majoritaires par rapport aux génériques et aux spécifiques. Une question se pose alors quant à l'influence de l'indexeur sur l'indexation et celle du document en texte intégral sur l'indexeur.

3.3.2. Descriptions des chaînes de référence

L'examen des chaînes de référence paraît indispensable à la détermination de la ou des structures thématiques présentes dans les textes. Ainsi, les tests sur la phonétique nous permettent de faire les premières observations qui demanderont à être vérifiées sur un *corpus* plus étendu avec un thésaurus non réduit à l'un de ses micro-thésaurus.

Le générique *phonétique* est présent dans la chaîne de référence mais les occurrences de ces descendants, c'est-à-dire des termes ayant un niveau de spécificité supérieur, sont plus nombreuses. Certains termes reprennent des termes de même niveau appartenant à une autre sous-branche de la représentation arborescente. Ainsi, *phonème* reprend *son* qui est de même niveau. D'autres termes reprennent des termes ayant un niveau de spécificité supérieur et appartenant à une autre branche de l'arborescence. C'est le cas de *phonème* lorsqu'il reprend *voyelle*. Certains termes apparaissent au pluriel dans les documents primaires où ils reprennent un terme singulier de même niveau: *phonèmes* reprend *syllabe.*

Les termes du micro-thésaurus constituent préférentiellement les têtes des chaînes de références tandis que les reprises lexicales sont effectuées soit par un autre terme du thésaurus soit par un substantif qu'on peut définir comme un quasi-synonyme.

Tab. 2 - Répartition des types de reprises dans les chaînes de références (Kister et Jacquey, 2007b)

Type de reprise	%
Même déterminant + même terme	5.5
Autre déterminant + même terme	5
Autre déterminant + autre terme ou substantif	58
Déterminant + autre(s) ou même(s)	2
Pronom sujet	10
Pronom objet	3
Pronom relatif	10
Autre relatif	3
Autre forme de reprise	3.5

Parmi les autres substantifs qui entrent dans la composition des chaînes de références, nous avons observé la présence de substantifs utilisés pour rendre compte des relations sémantiques exprimées dans le thésaurus: employé pour, voir aussi, etc. Nous voyons aussi apparaître des substantifs que nous considérons comme des candidats termes destinés à affiner certaines branches et certaines ramifications du thésaurus.

4. Evolutions récentes

La phase de test présentée ci-dessus, nous a permis de déterminer certaines perspectives à moyen et à plus long terme. Tout d'abord, il nous paraît nécessaire de nous

intéresser aux raisons qui font que les termes les plus présents en texte intégral ne sont ni les plus génériques, ni les plus spécifiques. Les résultats encourageants en ce qui concerne le lien entre les [+concret] et les spécifiques et celui qui unit les [+abstrait] et les génériques nous conduit à envisager une annotation plus fine des termes. La prise en compte de l'ensemble des termes présents dans le thésaurus, c'est-à-dire des termes génériques et spécifiques, mais aussi des relations sémantiques exprimées dans le thésaurus et de leurs variantes demandent qu'on s'intéresse à l'ensemble des formes linguistiques que peuvent prendre les concepts: références lexicales, pronoms personnels, relatifs, etc. afin de traiter de manière exhaustive les différentes formes de références. Le travail sur les chaînes de référence demande aussi que soient prises en compte des informations plus complexes que celles prises actuellement en compte: relations de quasi-synonymie entre des termes appartenant à des sous-domaines distincts, usage de certaines unités linguistiques dans des contextes particuliers. Il paraît aussi intéressant d'élargir les annotations à des informations de types encyclopédiques pour envisager une annotation efficace en vue du repérage des structures thématiques du texte. On peut alors s'interroger sur l'apport que pourrait avoir l'utilisation d'ontologies dans le processus d'annotation. Ces informations moins sémantiques peuvent aussi contribuer à l'analyse des divergences d'emploi entre thésaurus et texte intégral et à celles liées aux types d'interprétation et à la saisie d'un type particulier d'emploi par l'indexeur.

L'étape sur laquelle nous travaillons actuellement pour effectuer l'annotation la plus précise possible consiste à utiliser les définitions des termes de Thésaulangue présents dans le TLFi. La première étape consiste à transformer le thésaurus en terminologie en lui associant les définitions extraites automatiquement du TLFi. Le contenu sémantique des définitions sera à l'origine de l'annotation automatique des unités lexicales contenues dans le texte intégral. La procédure de repérage automatique des définitions présentes dans le TFLi pour le domaine des Sciences du langage qui nous intéresse (grammaire, lexicographie, lexicologie, linguistique, philologie, phonétique, phonologie, rhétorique, sémiologie, sémiotique, terminologie, toponymie) fait apparaître 2402 entrées et syntagmes nominaux susceptibles de constituer des termes du domaine de la linguistique alors que Thesaulangue contient 872 termes. Cette différence peut-être considérée comme l'opportunité de préciser, de compléter le thésaurus. L'apport terme/définition est certes important d'un point de vue quantitatif mais moindre d'un point de vue qualitatif puisque les termes repérés et définis ne sont si structurés, ni hiérarchisés et demanderaient pour être insérés un travail sur la structure du thesaurus. En parallèle, nous avons commencé à repérer les unités lexicales qui constituent les chaînes de références dans les textes intégraux et sur lesquelles seront reportées les informations sémantiques.

Références bibliographiques

Amsili Pascal, Denis Patrice, Roussarie Laurent (2005). *Anaphores abstraites en français: représentation formelle,* in J. Busquets et D. Hardt (eds), *Modèles et algorithmes pour la résolution d'anaphores*, Traitement automatique des langues, vol. 46/1, pp. 15-39.

Bally Charles (1925). *La linguistique et la vie*, Genève: Droz, Société de publications romanes et françaises.

Béjoint Henri, Thoiron Philippe (2000), *Le sens en terminologie*, Lyon: Presses universitaires de Lyon.

Boudreau Sylvie, Kittredge Richard (2005), *Résolution des anaphores et détermination des chaînes de coréférences: différences entre variétés de textes*, in J. Busquets, d. Hardt (eds), *Modèles et algorithmes pour la résolution d'anaphores*, Traitement automatique des langues, vol. 46/1, pp. 41-69.

Cabré Maria Teresa (1998), *La terminologie: théorie, méthode et applications*, Paris: Armand Colin.

De Mulder Walter, Schnedecker Catherine (2001). *Les référents évolutifs entre linguistique et pholosophie*, Recherches linguistiques, Paris: Klincksieck, n° 24.

Kister Laurence, Jacquey, Evelyne (2006). Traits sémantiques et anaphores pronominales, 4ème Rencontres de sémantique et de pragmatique, Orléans, 13-15 juin 2006.

Kister Laurence, Jacquey Evelyne (2007a). Acquisition lexicale sémantique à partir de données lexicographiques au service de la comparaison entre des structures thématiques de textes spécialisés et de thésaurus, *Terminologie: approches transdisciplinaires, Gatineau, Canada, 2-4 mai 2007*. (Actes à paraître en ligne <www.uqo.ca/terminologie2007>)

Kister Laurence, Jacquey Evelyne (2007b). Comparaison des structures thématiques de textes spécialisés et de thésaurus ou de terminologie, *Terminologia e mediazione linguistica: approcci e metodi a confronto, Assiterm/Realiter, Bertinoro (Forli), Italie, 8-9 juin 2007*. (Actes à paraître en ligne <www.realiter.net>)

Kleiber Georges (1994), *Anaphores et pronoms*, Champs linguistique, Louvain-la-Neuve: Editions Duculot.

Kleiber Georges, Schnedecker Catherine, Tyvaert Jean Emmanuel (1997). *La continuité référentielle*, Recherches linguistiques, Paris: Klincksieck, n° 20.

Mitchell D.C., Cuetos F., Zagar Daniel (1990). Reading in Different Languages: is there a Universal Mechanism for Parsing Sentences, in Balota, D.A., Flores d'Arcais, G.B., et Rayner, K., (eds.), *Comprehension Processus in Reading*, Hillsdale: Lawrence Erlbaum.

Perrot Jean (1957). *La linguistique,* Paris: PUF, Que sais-je?

Saussure Ferdinand (1968). *Cours de linguistique générale*, Paris: Payot.
Schnedecker Catherine (1998). *Les corrélats anaphoriques*, Recherches linguistiques, Paris: Klincksieck, N° 22.

Terminologia, modelli terminologici e reti

ÁGOTA FÓRIS

This paper provides a brief summary of the basic information regarding networks and surveys previous findings of linguistic research that were interpreted with the help of a) network models and b) Cabré's theory of doors. The paper elaborates some aspects of the application of network theory in terminology. The aim of these studies is to draw attention to the application of the theory of scale-free networks as a model in terminology.

Keyword: network – scale-free network – terminological unit – "theory of doors" – model of terminological networks

1. Introduzione

In alcune nostre pubblicazioni apparse negli ultimi anni abbiamo riassunto le nozioni basilari relative al funzionamento delle reti, per poi passare in rassegna i risultati delle ricerche linguistiche, analizzati con l'ausilio di vari modelli di reti (terminologia, linguistica quantitativa, linguistica generativa, psicolinguistica) (p. es. Fóris, 2005, 2007a). Abbiamo inoltre esaminato la possibilità che questi risultati possano trovare una collocazione all'interno della teoria delle reti e, per giustificare questa nostra supposizione, abbiamo fatto delle rilevazioni servendoci di precedenti risultati quantitativi. Siamo partiti dal presupposto che la teoria delle reti possa svolgere un ruolo importante anche nelle ricerche terminologiche. Con i risultati delle nostre ricerche vorremmo illustrare le possibilità di applicazione della teoria delle reti a invarianza di scala, relativamente alle ricerche terminologiche.

Le ricerche linguistiche prendono in esame sia la struttura della lingua che il rapporto dei segni linguistici con la realtà, sia il ruolo svolto dalla lingua nel processo di trasmissione delle informazioni che il rapporto tra lingua e pensiero. Le leggi della linguistica esprimono concettualmente tutte le nozioni che in questo sistema complesso possono essere illustrate relativamente alle proprietà ed alle relazioni dei diversi elementi. Gli elementi formali e sostanziali della lingua svolgono un ruolo determinante nei processi comunicativi, e le singole unità linguistiche possono essere classificate secondo sistemi ben definiti. I risultati che sono stati raggiunti nei vari campi della linguistica, hanno di volta in volta rivelato la struttura a rete delle diverse unità della lingua, nonché la possibilità che il funzionamento della lingua sia descrivibile secondo modelli correlati con le reti. Alcuni esempi di grafi già utilizzati in passato, provengono da vari

campi della linguistica: pensiamo ai grafi terminologici, a quelli utilizzati nella grammatica generativa, ovvero ai modelli a nucleo atomico ed a ragnatela utilizzati nella psicolinguistica per la descrizione dell'organizzazione del "dizionario mentale" (Aitchinson, 2003). Le ricerche hanno portato ad asserire che le relazioni di rete svolgono un ruolo fondamentale nei sistemi complessi della natura, dell'economia, della società e di moltissimi altri campi (cfr. Barabási, 2002, in italiano: Barabási, 2004; Csermely, 2005; in inglese: Csermely, 2006).

Gli alberi (grafi) terminologici di cui ci si è sinora serviti generalmente, rendono accessibili numerose informazioni relativamente alle relazioni tra i termini, come nel caso delle relazioni di subordinazione e di coordinazione. La rete linguistica però, ed all'interno di essa la rete terminologica dei diversi domini, sono più complesse rispetto alle capacità rappresentative della forma dei grafi terminologici sinora utilizzata, pertanto la nostra ipotesi è che per la loro descrizione ci si debba servire dei modelli delle reti a invarianza di scala. Abbiamo denominato questo nuovo modello delle reti a invarianza di scala "*modello terminologico delle reti"*.

I due punti di partenza della nostra relazione sono 1) il modello della teoria delle reti utilizzato con successo in vari campi della natura, della società e della tecnica, e 2) la teoria detto della "teoria delle porte" (*theory of doors*) di Cabré, che esamina le questioni della terminologia in maniera nuova, da tre punti di vista, cognitivo (concettuale), linguistico e comunicativo (Cabré, 2003): il termine viene considerato come unità costituita da tre componenti, e rispetto ai due metodi di approccio utilizzati in precedenza (quello basato sul significato, ovvero semasiologico, e quello basato sul significante ovvero onomasiologico), ne introduce un terzo, ovvero l'approccio comunicativo-pragmatico.

2. I modelli nella terminologia

2.1. Importanza dei modelli nella ricerca scientifica

Nel corso delle ricerche può verificarsi sovente che l'oggetto stesso della ricerca in questione, per determinati motivi, non sia analizzabile direttamente. Non possiamo osservare direttamente, per esempio, i processi microcosmici nel corso delle ricerche di fisica, chimica, biologia, ingegneria e medicina, così come gli astronomi non possono avere dati diretti sulle stelle lontane. Se da un lato sono le moderne tecnologie di misurazione a fornirci un aiuto per superare questo genere di difficoltà, lì dove non siamo coadiuvati dalle strumentazioni nell'ampliamento delle nostre possibilità di osservazione, possiamo far ricorso a mezzi di altra natura. Un altro genere di difficoltà nel processo della conoscenza può essere il fatto che il sistema preso in esame, oltre alle

caratteristiche che sono oggetto dell'analisi, dispone di numerose altre caratteristiche che si sovrappongono ai segni che desideriamo conoscere, in questo modo ostacolando il procedere della ricerca. Possiamo servirci di modelli per superare entrambi i generi di difficoltà.

Il *modello* può significare un campione, una riproduzione in piccolo di un oggetto, un individuo, un oggetto, un evento e così via, che utilizziamo in qualche modo come campione di riferimento. I vari campi scientifici modellano i sistemi che intendono conoscere servendosi non soltanto di oggetti (modelli fisici), ma anche di insiemi numerici, di formule (un modello fisico può essere creato relativamente a sistemi concettuali astratti, come per esempio la lingua).

Il modello è un prodotto del pensiero umano, nonché un mezzo indispensabile per la ricerca. La creazione di un modello è possibile sempre per mezzo dell'uso di un processo d'astrazione, indipendentemente dalla forma in cui il modello si presenta: acquista una forma fisica, ovvero una espressione matematica, o si presenta come concetto astratto. Papp definisce il modello in senso lato: «Definiamo modello ogni struttura da noi creata per poter meglio osservare, servendoci di essa, determinati gruppi di fenomeni della realtà» (Papp, 2006/1965: 45, traduzione di Á.F.); mentre Melcuk sottolinea l'importanza fondamentale dei modelli in qualsiasi scienza, e quindi anche nella linguistica, affermando che «la scienza altro non è che creazione di modelli» (Melcuk, 2001:141, traduzione di Á.F.).

Ed ora vediamo come possiamo risolvere, con l'aiuto di modelli che si basino sull'applicazione di astrazioni, i problemi di osservazione di fatti che restano celati al ricercatore.

- Prendiamo in esame il caso in cui l'osservazione è limitata per un determinato motivo (grandi distanze, piccole dimensioni, etc.): se pure è dato un certo grado di conoscenza dell'oggetto della ricerca, ciò non è sufficiente alla sua conoscenza completa, dunque sono necessari altri dati. Sulla base delle nozioni a nostra disposizione e con l'aiuto delle altre conoscenze da noi acquisite in precedenza, mettiamo insieme un modello che disponga delle caratteristiche dell'entità che desideriamo conoscere. Il modello ci permette di giungere a determinate conseguenze logiche, su di esso possiamo eseguire delle analisi analoghe, e tutto ciò ci permette di avvicinarci sempre di più alla soluzione del compito che costituisce l'essenza della nostra ricerca.
- Nel caso in cui l'oggetto della ricerca sia estremamente complesso, possiamo eliminare gli elementi di disturbo creando un modello astratto che sostituisca il sistema realmente in esame, senza comprendere i dettagli riferibili alle caratteristiche secondarie, così da poterci concentrare soltanto sulle parti determinanti. Questo processo di creazione di un modello, nel corso del quale ci liberiamo delle parti che dal punto di vista della ricerca sono superflue, così che rimangono

soltanto le parti essenziali, è un processo analogo a quello di formazione dei concetti. Il modello semplificato, che contiene tutti gli elementi fondamentali dal punto di vista dell'analisi, rende possibile un'osservazione scevra da elementi di disturbo. Melcuk parla, a questo proposito, di *modello funzionale* (il modello che illustra perfettamente le funzioni del funzionamento, mentre – per esempio – non contiene precisamente tutti gli elementi strutturali). Tali modelli sono le "scatole nere" utilizzate nel campo della matematica, della logica o della didattica, che sono adatte all'esecuzione di determinate operazioni logiche. La creazione di un modello funzionale è giustificata nel momento in cui desideriamo creare un modello del *funzionamento* dell'oggetto rappresentato dal modello – nel nostro caso, della lingua. Un modello simile è, per esempio, il *modello del senso-testo* creato da Melcuk, Jolkovski e Apresian (cfr. Melcuk 2001).

Non è necessario, dunque, che il modello corrisponda in tutto e per tutto alla realtà che rappresenta. Infatti possono esistere, ad esempio, notevoli differenze di dimensioni tra il modello e la realtà a cui esso si riferisce. Nei casi di modelli di macchine, di elementi di analisi effettuate per mezzo di modelli (esperimenti di conduttività) e così via, è molto spesso proprio la differenza di dimensioni (modelli atomici, sistemi planetari) a rendere possibili l'osservazione e l'analisi. Il materiale di cui è costituito il modello fisico, può variare a seconda dei casi. Alcune parti della realtà descritta con un modello, possono non apparire nel modello, poiché è proprio la loro assenza a rendere più facile l'esame. Inoltre, le differenze possono essere anche tali da permettere al modello di non offrire alcuna informazione sulla forma di manifestazione esterna dell'oggetto, purché il modello contenga, della realtà che si ripromette di descrivere, tutte le caratteristiche che sono importanti dal punto di vista dell'esame da effettuare. La semplificazione del modello deve dunque essere eseguita solo in misura tale da non modificare le caratteristiche principali del sistema da analizzare.

Nel caso della lingua, se è possibile raccogliere dati, esaminarne i parlanti ed i prodotti (ovvero i testi), non è però possibile esaminarla direttamente, poiché è impossibile indirizzarci direttamente alla lingua, né "coglierla" fisicamente. Per questo motivo è essenziale, nella linguistica, ricorrere a modelli, come nella matematica e nelle scienze naturali. In una prospettiva storica, i modelli presentano numerose differenze, a seconda dei singoli campi della linguistica: nella psicolinguistica, per esempio, hanno riscosso successo soprattutto i modelli a ragnatela e a nucleo atomico. Anche nella terminologia i modelli (e le teorie) hanno grande importanza, ed i più noti sono quelli di Wüster e Cabré: quest'ultimo esamina dettagliatamente le teorie terminologiche, arrivando alla conclusione che la scienza progredisce per mezzo del confronto e dell'interazione, grazie al confronto di ipotesi con oggetti empirici, grazie alla proposta di modelli e teorie alternative, alla nostra capacità di apprezzare l'accettabilità di tali teorie (Cabré 2003).

2.2 Importanza dei modelli di rete

Per la creazione di modelli generalmente utilizzabili in ogni campo scientifico moderno, possiamo riferirci agli esempi di elaborazione della teoria delle reti complesse.
Il modello a rete è un modello matematico. L'impiego di modelli matematici è particolarmente indicato perché tali modelli rendono ben descrivibili le interrelazioni astratte tra i processi ed i fenomeni concettualizzati con l'ausilio del formalismo matematico. È inoltre utile perché tali processi divengono esprimibili anche da parte dei calcolatori. Due sono i casi possibili di impiego di modelli matematici.

- Accade sovente di avere a nostra disposizione il sistema matematico necessario alla strutturazione del modello, e che nel corso della creazione di quest'ultimo dobbiamo far sì che le caratteristiche dell'oggetto preso in esame siano coerenti con il formalismo matematico. I rapporti esistenti tra le diverse entità si possono ben modellare con l'aiuto di metodi matematici, una parte dei quali è esprimibile per mezzo di numeri, con l'aiuto di relazioni quantitative che prendono il nome di *aritmetica* (come la quantificazione della legge di Zipf nel linguaggio, cfr. p. es. Carloni, 2005). La matematica, inoltre, analizza anche altre relazioni del mondo circostante, che possono essere espresse non solo con l'aiuto di concetti numerici, e rispetto alle quali non si manifesta la numericità, ovvero si manifesta con un altro contenuto, come nel caso della commensurabilità quantitativa. Si tratta per esempio – se non di concetti aritmetici – dei concetti risultanti dall'analisi degli insiemi (e anche dei simboli introdotti per la loro individuazione): la parte comune degli insiemi, l'elemento dell'insieme, l'insieme vacante, etc. In maniera simile, la teoria degli operatori – che viene applicata, per esempio, anche nella linguistica formale – non rientra nel campo dell'aritmetica. La scienza combinatoria analizza le possibilità di relazioni che possono venire a crearsi tra insiemi diversi. In natura, nella società, nella lingua e così via, sono molte le possibilità di relazione che possono crearsi tra i componenti, e le strutture che in tal modo vengono a formarsi determinano le caratteristiche delle entità createsi. La ricerca ha per compito la descrizione della relazione tra le caratteristiche degli elementi, quelle della struttura creatasi, e la natura dell'entità. Quando diviene necessaria la creazione di un modello di questi gruppi, è utile applicare all'analisi della questione i modelli generici – già esistenti – offerti dalla matematica.
- Si verifica però anche il caso in cui non disponiamo di un apparato matematico adatto alla descrizione dell'oggetto della nostra analisi, ed in questi casi è necessario effettuare delle ricerche matematiche, che spesso portano allo sviluppo di nuovi campi della matematica. L'evoluzione del calcolo differenziale ed integrato venne stimolata dalle questioni intervenute nel corso della creazione di modelli per la descrizione di processi meccanici. In conseguenza di tali esigenze venne

formulata, ad esempio, la teoria dei grafi, o la teoria ludica. La realizzazione del modello di rete a invarianza di scala si inserisce tipicamente in questa casistica, infatti sono numerosi gli argomenti che dimostrano come le reti siano presenti in ogni parte della natura e della società. Dopo aver preso atto di questo, si è verificata l'elaborazione del modello che rende possibile un'interpretazione unitaria dei fenomeni.

Similmente alla matematica, anche i risultati della logica e della filosofia sono ben utilizzabili per la creazione di modelli. La difficoltà è rappresentata il più delle volte dal grado di corrispondenza tra la realtà ed il modello matematico (o di altra natura).

Il concetto di *rete* è noto da tempo. Potremmo descriverlo genericamente dicendo che i *nodi* rappresentano gli elementi di un insieme, mentre gli *spigoli* descrivono le relazioni esistenti tra gli elementi. I risultati delle ricerche dimostrano che le reti, nel corso dell'evoluzione, si sono da tempo formate nella natura e nella società, mentre solo nel corso degli ultimi decenni ne abbiamo scoperto l'esistenza, le caratteristiche e il ruolo fondamentale. È stata verificata, per esempio, l'esistenza di una breve serie di relazioni tra due punti apparentemente molto distanti reciprocamente, a dimostrare che il rapporto, l'interdipendenza di cose diverse, sono di natura ben diversa rispetto a quanto finora credevamo. Negli ultimi tempi si è verificato, in una serie di campi scientifici, che a ricoprire un ruolo fondamentale nel nostro mondo è una rete di tipo speciale, la *rete a invarianza di scala* ("*scale-free network*"). Le caratteristiche fondamentali di questa rete sono nell'esistenza dello "small world", delle relazioni forti e deboli, e nel fatto che la distribuzione dei nodi sia descrivibile non con una curva a campana, ma con una curva di potenza (v. Barabási 2002). Nelle reti a invarianza di scala, oltre a numerosi *nodi* che dispongono di poche relazioni, esistono i cosiddetti *nodi centrali* ("hube"), che dispongono di numerosissime relazioni. Questi nodi centrali, dotati di molte relazioni, hanno un ruolo particolare nella formazione e nel funzionamento delle reti. Il gruppo delle reti complesse denominato delle reti a invarianza di scala, è quello in cui la distribuzione, secondo una determinata caratteristica, dei nodi centrali è descritta da una curva di potenza, e per questo i nodi centrali non possono essere raggruppati secondo un criterio interno. È un fatto dimostrato che i sistemi complessi della natura, della società, delle tecniche, dell'economia e via dicendo, sono delle reti a invarianza di scala di questo tipo, e che gli elementi esistenti, per esempio, in queste reti, costituiscono degli "*small world*" facilmente attraversabili. La caratteristica di "small world" vuol dire che in queste reti esistono delle relazioni il cui allacciamento, veloce e a breve termine, viene assicurato da nodi centrali molto distanti tra loro nello spazio o nel tempo, generalmente con una media di sei passaggi (ciò vuol dunque dire che gli elementi dello "*small world*" non sono un gruppo correlato di elementi vicini l'uno all'altro). La struttura della rete www è un tipico esempio di rete a invarianza di scala, in cui i siti web sono i nodi centrali, i link sono invece gli spigoli, e la caratteristica

di "*small world*" assicura la possibilità di attraversamento rapido, attraverso alcuni link, da un sito web a un altro (l'ontologia informatica ci porta ormai in questo speciale campo scientifico) (cfr. Ontologie, 2003).

L'esame delle reti che si possono trovare in natura e nella società, dimostra che qualsiasi entità – a causa delle sue numerose caratteristiche – è allo stesso tempo elemento di più reti, da cui consegue che le reti sono intercorrelate tra loro in maniera complessa. Le varie reti, dunque, sono interrelate in maniera complessa, e da ciò consegue che tutto è collegato a tutto, come si vede chiaramente, ad esempio, nelle interdipendenze esistenti all'interno della rete dell'economia, e che si verificano in maniera complessa. (Sui modelli delle curve di potenza nella lingusitica e sulle reti vedi p. es Köhler, 2002, Ferrer i Cancho e Solé, 2001).

2.3 Il modello terminologico delle reti

Secondo il modello della "teoria delle porte", formulato da Cabré, le questioni terminologiche devono essere analizzate da tre punti di vista: cognitivo (concettuale), linguistico e comunicativo (Cabré 2003). L'unità di queste tre componenti determina il valore comunicativo del termine. L'elemento complesso che si forma intorno al termine in questo triplice contesto, viene denominato da Cabré *unità terminologica* (*terminological unit*). Nel corso del processo comunicativo (che si tratti di creazione di un testo (codificazione) o di comprensione di un testo (decodificazione) è necessario l'orientamento complesso in queste tre reti, per un impiego felice dell'unità terminologica.

Le reti a invarianza di scala sono ritenute adatte alla creazione di un modello per la descrizione della lingua, perché si tratta di un modello dinamico, ed anche l'uso della lingua è un processo dinamico. Con ciò intendiamo dire che questo modello rende possibile anche la creazione di un modello per i processi temporali dell'uso della lingua. Partendo dal modello di Cabré, immaginiamo il termine come unità complessa costituita da tre componenti. Supponiamo che i componenti cognitivi, linguistici e comunicativi, costituiscano separatamente delle reti a invarianza di scala, e che il processo della comunicazione sia descrivibile con un modello costituito dal collegamento di queste tre reti. La rete terminologica è quella rete dalla struttura complessa e multidimensionale, le cui parti sono queste reti stesse. Ad un determinato livello della lingua le reti cognitive, linguistiche, pragmatico-comunicative, costituiscono una determinata sezione dell'intera rete, e queste sezioni, nel loro complesso, assicurano il funzionamento ottimale della lingua. Le caratteristiche principali del *modello terminologico delle reti* sono le seguenti: 1) la rete terminologica contiene numerosi nodi centrali e spigoli; 2) la rete, che ha un'estensione d'infinita grandezza, è costituita da formazioni spaziali che s'intertangono in maniera complessa; 3) la lingua è scomponibile in ben più parti

essenziali che nelle tre sezioni qui citate, e in virtù di questa proprietà, anche il modello usato per la descrizione della struttura della rete, può essere più complesso. Non si possono non prendere in considerazione, per esempio, le relazioni fonetiche, oppure le parole grammaticali accanto ai termini, anche se le nostre esemplificazioni non ci permettono di prenderle in considerazione.

Guardando questa unità triplice dal punto di vista cognitivo (concettuale), il ruolo comunicativo dell'unità terminologica dipende dal contesto del tema, anche se ha un posto ben determinato nella rete concettuale. Il posto occupato dalle unità terminologiche nella rete concettuale è la parte determinante del ruolo che esse svolgono nella comunicazione. Nella complessa rete cognitiva sono presenti, come semplici reti parziali, i grafi terminologici già utilizzati in passato. Per esempio, la struttura della rete a invarianza di scala del "dizionario mentale", si rispecchia anche nella struttura della rete concettuale. I nodi centrali della rete svolgono un ruolo determinante nel funzionamento della rete a invarianza di scala: tale ruolo determinante è svolto nella rete terminologica dalle *parole chiave* e dai *termini dei concetti di base*.

Dal punto di vista della componente linguistica le unità terminologiche sono nello stesso tempo unità lessicali, dotate di struttura lessicale e sintattica, mentre la creazione della struttura lessicale può avvenire nei modi usuali della formazione di parole. Naturalmente, il segno delle unità terminologiche può essere non solo un segno linguistico, ma qualsiasi altro genere di codice, segno, etc. La categoria grammaticale può essere quella del sostantivo, del verbo, dell'aggettivo, dell'avverbio, oppure la struttura nominale, predicativa, aggettivale. La combinatorietà sintattica si può restringere secondo i principi combinatori di tutte le unità lessicali della lingua. Le differenze che passano tra le strutture delle lingue naturali, dimostrano in vari modi la codificazione dell'informazione, ragion per cui – ad esempio – le traduzioni di testi non possono essere effettuate esclusivamente con operazioni di corrispondenza linguistica.

Sulla base della componente comunicativa, l'unità terminologica è caratterizzabile come segue: si manifesta nella comunicazione, si adatta formalmente alle caratteristiche tematiche e funzionali del discorso, e nel discorso che di volta in volta presenta diversi contenuti specifici, le unità si riferiscono a sistemi diversi. Le unità terminologiche svolgono un ruolo decisivo nella formazione dei segni alla base (coerenza, coesione) delle caratteristiche tipiche del testo, ma nello stesso tempo il testo nella sua integrità può avere un effetto riflesso sull'impiego del termine. Per esempio, le norme terminologiche determinano l'uso delle unità terminologiche, oltre le regole semantiche e linguistiche.

Nel corso dell'uso della lingua, dunque, l'unità terminologica viene determinata contemporaneamente dai sistemi delle tre reti. Durante l'uso della lingua (che si tratti di codificazione o decodificazione dell'informazione), la complessa unità terminologica si costituisce, secondo le relazioni assicurate dallo "*small world*", nella rete che risulta dall'interrelazione delle tre reti. Le unità terminologiche, dunque, svolgono un ruolo

decisivo nei processi comunicativi, è a dire che – per esempio – l'analisi di testi non può prescindere dalle relazioni cognitive, tanto che un errore tipico che da ciò deriva è il giudizio formatosi relativamente all'esame dei libri di testo, più precisamente intorno alla questione della comprensione dei testi. Sono sempre più numerose le analisi delle strutture testuali dei libri di testo, in cui i segni della comprensibilità dei dati testi si ricercano esclusivamente nelle unità linguistiche, come accade quando vengono illustrati la struttura della frase (numero e lunghezza delle parole) o il suono delle parole (nel caso dei forestierismi) e così via, mentre mancano completamente gli esami delle caratteristiche cognitive e comunicative. Si tratta di esami quantitativi, poiché i dati vengono quantificati (ridotti a numeri), e da essi si estrapolano conseguenze di ordine statistico. La questione della comprensione del testo viene ricondotta a deficienze delle competenze di lettura, le cui cause vengono ricercate nella formazione linguistica dei testi. Le nostre ricerche, però, hanno chiaramente dimostrato (Fóris 2006) che le difficoltà di comprensione dei testi non dipendono esclusivamente dalla formazione linguistica dei testi, poiché nella comprensione dei testi un ruolo fondamentale è svolto dalle unità concettuali e dalle unità comunicative, e che quindi in questo caso sarebbe bene esaminare la rete complessa delle tre componenti – cognitiva, linguistica e comunicativa.

3. Conclusioni

Il *modello terminologico delle reti* da noi presentato in questa sede è insieme quantitativo e qualitativo: tra gli esami quantitativi si possono impiegare procedimenti statistici, mentre sul lato qualitativo si può creare un modello per la descrizione del processo di creazione dei testi, con l'aiuto di esso.

Secondo la nostra concezione, dunque, il ruolo svolto dai termini nel processo della comunicazione, è determinato da diversi fattori, tra i quali hanno un ruolo notevole la componente cognitiva ed il suo sistema di relazioni. È per questo che riteniamo particolarmente giustificato l'isolamento della prospettiva terminologica tra le varie possibilità di approccio multilaterale alla lingua, infatti il concetto individuato dal termine determina il significato del termine. Il fatto che i particolari del significato del termine vengano differenziati dal posto che esso occupa nella rete linguistica, dimostra le relazioni di cui l'unità terminologica è dotata nella rete linguistica stessa.

Siamo naturalmente coscienti del fatto che un tale modello presenta non solo vantaggi, ma anche svantaggi. Rispetto ai modelli precedenti, esso indica un approccio da un altro punto di vista, e siamo fiduciosi che con l'aiuto di ulteriori riflessioni e discussioni sulla questione, si potrà dare forma ad un modello dinamico e che si presta all'at-

tività analitica, grazie al quale si potranno meglio descrivere sia le unità terminologiche che i sistemi terminologici.

Bibliografia

Aitchinson J., (2003), *Words in the mind. An introduction to the mental lexicon.* Malden, Basil Blackwell.

Barabási A.-L., (2002), *The New Science of Networks*. Cambridge MA, Perseus.

Barabási A.-L., (2004), *Link. La scienza delle reti*. Torino, Einaudi.

Cabré Castellví M.T., (2003), *Theories of terminology. Their description, prescription and explanation*, «Terminology» 9, n. 2, p. 163-200.

Carloni F. (2005), *La legge di Zipf sul numero dei significati in italiano e inglese, in: Parole e numeri. Analisi quantitative dei fatti di lingua*. Roma: Aracne: De Mauro e Chiari, p. 355-370.

Csermely P. (2005), *A rejtett hálózatok ereje.* Budapest, Vince.

Csermely P., (2006), *Weak Links*. (The Universal Key to the Stability of Networks and Complex Systems) Heidelberg, Springer. <www.weaklink.sote.hu/weakbook.html>.

Ferrer i Cancho R., Ricard V. Solé (2001), The small world of human language, in: *Proceedings of the Royal Society of London, B 268*. p. 2261-2266.

Fóris Á., (2005), *Hat terminológia lecke*. Pécs, Lexikográfia.

Fóris Á., (2006), *A terminológiai szemlélet a tankönyvek minöségi megítélésében*. "Iskolakultúra", XVI, n. 5, p. 79-88.

Fóris Á., (2007a), A skálafüggetlen hálók nyelvészeti vonatkozásai. "Alkalmazott Nyelvtudomány", VII, n. 1-2, p. 105-124.

Fóris Á., (2007b), *Terminology and the Theory of Scale-free Networks,* in: *Current Trends in Terminology*. Proceedings of the International Conference on terminology. Szombathely, Hungary, 9th-10th of November, 2007. Szombathely, BDF: Fóris e Pusztay. <http://termik.bdf.hu>

Köhler R., (2002), *Power law models in linguistics: Hungarian*. «Glottometrics», n. 5. p. 51-61.

Melcuk I., (2001), *Egy értelem-szöveg nyelvészet felé*, in: A moszkvai szemantikai iskola, Budapest, Corvina: Papp, p. 139-187. [fonte: Igor Mel'cuk (1997), *Vers une linguistique sens-texte*. Leçon inaugurale faite le Vendredi 10 janvier 1997. Collège de France, Chaire Internationale. Paris, Collège de France, p. 1-78.]

Papp F., (2006/1965), *Modell*, in: Papp Ferenc olvasókönyv. Budapest, Tinta: Klaudy, p. 45-52.

Ontologie (2003), "Sistemi Intelligenti", XV, n. 3.

Processi di terminologizzazione e determinologizzazione nel dominio della diffusione e distribuzione del libro

FRANCO BERTACCINI, CLAUDIA LECCI, VALENTINA BONO

The present article analyses the aspect of "terminologization" and "determinologization" of lexical units, that is, the passage of a term from special language to common language. To this purpose, the terminology of the macro-domains of the diffusion and the distribution of the book has been investigated. The terms are part of a list of 172 concepts, suggested and validated by the OQLF, and represent well the domain of investigation. This study is involved in the project "Lexique panlatin de la diffusion et de la distribution du livre", with the purpose of aiding the development of the neo-Latin languages.

Keywords: terminologization – determinologization – general language – specialist languages – book diffusion and distribution

Il dominio di applicazione e il progetto

Il progetto all'interno del quale abbiamo approfondito i fenomeni oggetto del presente articolo è denominato "Lexique panlatin de la diffusion et de la distribution du livre" e si inserisce in un insieme di altri progetti creati e realizzati con il medesimo obiettivo: favorire lo sviluppo armonizzato delle lingue neolatine (francese, catalana, spagnola, galiziana, italiana e rumena), in virtù della loro origine comune e del fatto che utilizzano modalità di formazione lessicali simili. Questi nascono dalla collaborazione tra l'OQLF e *Realiter* (Rete panlatina di terminologia). Il *Lessico panlatino della diffusione e della distribuzione del libro* è un documento multilingue (è importante evidenziare anche la presenza dei termini in lingua inglese) che raccoglie i termini più utilizzati nel *set*tore.

Il nostro contributo rientra in una collaborazione tra l'OQLF e la sezione di terminologia della Scuola Superiore di Lingue Moderne per Interpreti e Traduttori di Forlì, ormai consolidata, che ha visto la realizzazione di altri progetti concernenti altri ambiti disciplinari: il *Lessico panlatino della geomatica*, il *Lessico panlatino dei carrelli per movimentazione* e, infine, il *Lessico panlatino della nanotecnologia*.

Nel caso specifico del dominio sulla diffusione e distribuzione del libro, il nostro apporto al progetto appena descritto è consistito nell'identificare gli equivalenti in lingua italiana di un insieme di termini legati alla diffusione e alla distribuzione del

libro. Tali termini fanno parte di una lista di 172 concetti individuati e validati dall'OQLF, rappresentativi del dominio indagato.

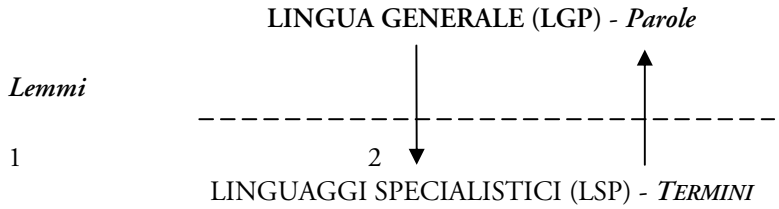
La terminologia indagata, dunque, presenta termini che appartengono al macro-dominio della **diffusione di libri,** l'insieme di attività legate alla promozione e alla commercializzazione di libri, e al macro-dominio della **distribuzione di libri,** l'insieme delle attività logistiche messe in atto per distribuire i libri ai diversi luoghi di vendita.

L'oggetto del dominio specialistico in questione, il libro, costituisce un'entità di grande interesse, non solo per gli esperti quali tipografi, distributori, stampatori, ma anche e soprattutto per i "fruitori" di tale oggetto, ossia la gente comune. Tale puntualizzazione si è resa necessaria durante il lavoro da noi condotto per l'identificazione degli equivalenti in lingua italiana dei termini fornitici dall'OQLF. Il reperimento di tali equivalenti si è svolto a partire dalla costruzione di un *corpus* specialistico, utilizzato per l'analisi e l'estrazione della terminologia in lingua italiana. A questo punto del lavoro, tuttavia, ci si è resi conto che non tutte le unità lessicali che costituivano il *Lessico panlatino della diffusione e distribuzione del libro* potevano essere definiti, in senso stretto, "termini", perché gran parte di essi sono utilizzati con un'elevata frequenza nella lingua generale. Non approfondire tale aspetto avrebbe dunque potuto compromettere la qualità del *corpus* da noi creato e, di conseguenza, la qualità dell'intero progetto. Dunque, i due processi che analizzeremo qui di seguito, in particolare la determinologizzazione, sono stati fondamentali per l'analisi dei termini in preparazione della costruzione del *corpus* di riferimento. Si è proceduto, infatti, alla suddivisione dei *termini* dalle unità lessicali "a rischio di determinologizzazione" al fine di ottenere un *corpus* piuttosto eterogeneo e in grado di comprendere il più possibile le unità del lessico in questione.

Terminologizzazione e determinologizzazione [1]

Il patrimonio lessicale si arricchisce sempre più per adattarsi alle nuove e crescenti esigenze comunicative della società e deve dunque ricorrere a procedimenti vari, dal mutamento semantico, al prestito linguistico e/o alla formazione di nuove parole: si hanno così parole che cadono in disuso, altre che vengono "rinnovate" e altre ancora che cambiano (in parte o del tutto) il significato o l'ambito di impiego (Aprile 2002: 161).

In questo contesto, si inseriscono due processi fondamentali della formazione e dell'uso dei lemmi, la *terminologizzazione* e la *determinologizzazione*. Prima di addentrarci nell'analisi di tali fenomeni all'interno del dominio da noi indagato, riteniamo utile riportare uno schema esplicativo del passaggio a cui un'unità lessicale soggiace durante i due processi:



Lo schema esemplifica in modo chiaro le componenti coinvolte nei due processi che andremo ad analizzare: il valore *1* indica il processo di *terminologizzazione*, mentre *2* quello di *determinologizzazione*. A metà tra i due processi, parole e termini sono definiti *lemmi* in quanto costituiscono delle unità ancora "senza identità". La linea tratteggiata dei lemmi sottolinea quanto sia labile il confine tra parole e termini, tra lingua generale e linguaggi specialistici: è una linea di confine molto sfumata, non netta, poiché «le terminologie condividono con le lingue naturali alcune proprietà qualificanti che possono essere interpretate come spie di un relativo ancoraggio endocentrico dei concetti, e in particolare la presenza di casi di anisomorfismo, omonimia, polisemia e sinonimia» (Bertaccini, Prandi e al. 2004).

La terminologizzazione

> «La *terminologizzazione* è un processo per cui una parola o un'espressione di uso generale o comune viene trasformato in un termine che designa un concetto particolare in un linguaggio speciale» [2].

I linguaggi specialistici hanno dovuto fare ricorso, tra gli altri, anche a questo processo poiché il sapere specialistico cresce sempre più e assume maggiore interesse, non solo per gli esperti di determinati *set*tori, ma anche e soprattutto per l'intera società (*knowledge society*): «selon ses besoins, la langue de spécialité élargit, rétrécit ou modifie le sens des mots de la langue générale» (Dubuc 1992:26).

Tale processo è ormai al centro di numerosi studi; oltre Rega, Cabré e Scarpa, anche altri studiosi hanno trattato le diverse procedure di terminologizzazione. Riediger include tale fenomeno nei processi di formazione dei termini: «*terminologizzazione*: a una parola della lingua comune viene attribuito un nuovo significato speciale» (Riediger, s.d.). Anche Aprile accenna al medesimo concetto, ma utilizzando una denominazione differente, ossia *rideterminazione semantica*: «si usano parole che sono già della lingua comune ma le si specializza attraverso una rideterminazione semantica, cioè attraverso l'acquisizione di un nuovo significato, proprio di quel *set*tore». Aprile ag-

giunge che una parola del linguaggio comune può, se rideterminata semanticamente, diventare un termine specialistico appartenente a più *set*tori: «in ciascun *set*tore avrà il suo significato specifico, senza che siano possibili ambiguità o confusioni da parte di chi recepisce la parola» (2005:52-53).

All'interno del lessico da noi analizzato si possono trarre diversi esempi di questa modalità di formazione di termini. Oltre a *cerniera*, *dorso* e *piatto,* ai quali faremo riferimento in seguito, esempi significativi sono *fascetta*, *collana* e *coperta*. Tali unità lessicali, appartenenti alla lingua comune, hanno subìto un processo di terminologizzazione al fine di designare concetti e oggetti di ambiti specialistici differenti, come nel caso del dominio in questione. Nella sua accezione comune, ad esempio, *fascetta* designa una «piccola fascia, solitamente ad anello, di materiale vario, usata per avvolgere, imballare e/o tenere insieme più cose». L'accezione "terminologizzata" di tale lessema, all'interno del sottodominio pubblicitario del lessico da noi analizzato, è «striscia di carta applicata trasversalmente alla copertina del libro, utilizzata per riportare slogan pubblicitari destinati a sottolineare il successo del libro». Tale lessema è stato rideterminato semanticamente anche in altri ambiti disciplinari: infatti, indica anche la «striscia di tessuto chiusa ad anello recante l'indicazione dei gradi e dello scudetto del reparto» (dominio militare) o, nel linguaggio meccanico, l'«anello metallico che stringe e blocca le due parti di un raccordo».

La determinologizzazione

Il breve *excursus* sul processo di terminologizzazione consente di affrontare in modo più chiaro il fenomeno contrario, la *determinologizzazione*, il quale costituisce un passaggio fondamentale nello studio della terminologia moderna, in particolare all'interno di alcuni domini che si situano a metà tra la conoscenza prettamente specialistica e l'interesse sempre più forte della gente comune verso determinati ambiti disciplinari.

Di contro ai numerosi studi sulla terminologizzazione, il fenomeno della determinologizzazione è stato studiato in dettaglio solo recentemente da pochi studiosi, nonostante «the migration of terms into general language is by no means a new phenomenon: it is well known that some terms, originally used only by a community of specialists, are later taken up by a broader language community» (Meyer e Mackintosh 2000). Tra i pochi studiosi, troviamo Aprile, il quale accenna a tale processo (benché lo nomini in modo differente):

> «possono essere considerati neologismi anche gli elementi del lessico che nascono come termini specialistici di un determinato ambito *set*toriale e poi si diffondono, di solito con un allargamento del significato, presso la generalità dei parlanti» (2005:59, sottolineato nostro).

Lo studioso sostiene che, passando da una scienza all'altra, alcuni termini propri ad un dominio specialistico possono acquisire un significato più ampio ed entrare a far parte della lingua comune, secondo dinamiche a volte molto complesse legate alla formazione delle parole.

Solo di recente si è dunque constatato come tale fenomeno linguistico abbia registrato un incremento non indifferente nel corso degli ultimi venti anni, motivo che ha spinto ad una maggiore attenzione e ad uno studio mirato del fenomeno. Molti lessicografi hanno deciso di riservare un'attenzione particolare a tale processo. All'interno dei loro dizionari, hanno deciso di non limitarsi ad un accenno marginale alle unità lessicali che provengono da un linguaggio tecnico, in quanto sono divenuti sempre più consapevoli dell'importanza che un termine specialistico può assumere per la *knowledge society* (società della conoscenza) in cui viviamo entrando così gradualmente a far parte della lingua comune, perdendo inoltre quella connotazione monosemica e tecnica che lo rendeva una *fixed entity* (entità fissa).

Ingrid Meyer e Kristen Mackintosh [3] sono state le prime ad utilizzare il termine *de-terminologization* per indicare il fenomeno mediante cui «a lexical item that was once confined to a fixed meaning within a specialized domain is taken up in general language» (2000:112) (un'unità lessicale che originariamente possedeva un significato fisso in un determinato dominio terminologico è utilizzato dalla lingua generale), puntualizzando in una nota che la scelta di tale termine per denominare il processo è

> «loosely based on the French *dé-spécialisation* (Mazière 1981: 84). We prefer not to translate by *de-specialization,* since this term could apply to any lexical item (including non-terminological) where a lexical meaning becomes more general» (2000:136).

Le componenti sociologiche

La spinta ad una maggiore attenzione verso il fenomeno della determinologizzazione si spiega innanzitutto a partire da comportamenti di natura sociologica: viviamo in una società toccata da profondi cambiamenti ed evoluzioni, dovuti in gran parte alla creazione e allo sviluppo di sempre nuove idee e tecnologie. Questo vortice di innovazioni e cambiamenti ha travolto e continua a travolgere tutti, esperti e non, anche se in modi e a livelli differenti; anche la gente comune sente il bisogno di avvicinarsi ai domini specialistici spinta soprattutto dalla necessità di arricchire le proprie conoscenze e la cultura personale, al fine di conoscere a fondo il "micro-mondo" in cui si vive (il quotidiano) ma anche il macro-mondo.

> «In una società sempre più tecnologica come quella odierna la comunicazione tecnico-scientifica viene infatti sempre più indirizzata anche alla gente comune, che la utilizza per

> soddisfare il bisogno di aumentare il proprio sapere scientifico e fare scelte informate nella vita di tutti i giorni» (Scarpa, 2002:35).

Questo, naturalmente, si traduce in un susseguirsi e concatenarsi di approfondimenti della vita sociale, della vita privata ma anche e soprattutto della vita professionale. Essere a conoscenza di tutti i meccanismi e nozioni di un dominio specialistico equivale dunque ad essere parte di una società produttiva, in cui ognuno vuole (e probabilmente, in qualche modo, **deve**) divenire un *knowledge worker* (lavoratore della conoscenza) produttivo. Esempio lampante di questo è l'invenzione del *computer* e la sconvolgente rapidità con cui è entrato a fare parte della vita di tutti: è proprio dal dominio informatico che si possono trarre gli esempi più rappresentativi dei meccanismi in atto durante il processo di determinologizzazione, esempi come *virtual* e *virtually* ampiamente descritti nell'articolo *When terms move into our everyday lives: an overview of de-terminologization* di Meyer e Mackintosh:

> «computing is by far the best example of a specialized domain that is becoming critical in our everyday lives. However, other domains of expertise such as economics, environmental studies, genetics, and healthcare are also of great interest to the general public» (2000:127).

Alla curiosità e all'attenzione verso fenomeni e nozioni tecniche e scientifiche e dunque verso il linguaggio che li designa si aggiunge il fattore *trendiness* (l'essere di tendenza): «general-language sense is sometimes "revived" as a result of the **trendiness** of the de-terminologized word»: ad esempio, «it appears that because of its **buzzword status**, people are using *mega* in contexts where they previously might have preferred other words such as *large»* (2000: 130; gras*set*to nostro). Esiste dunque anche un fattore "modaiolo" nella scelta dei termini da utilizzare; si usa un determinato lemma nuovo o meno utilizzato piuttosto che una parola già ampiamente in uso perché sembra essere in voga in quel momento: ripescando dal nostro *lessico*, è il caso di *poster* utilizzato al posto di *manifesto* o di *copyright* invece di *diritti di autore*.

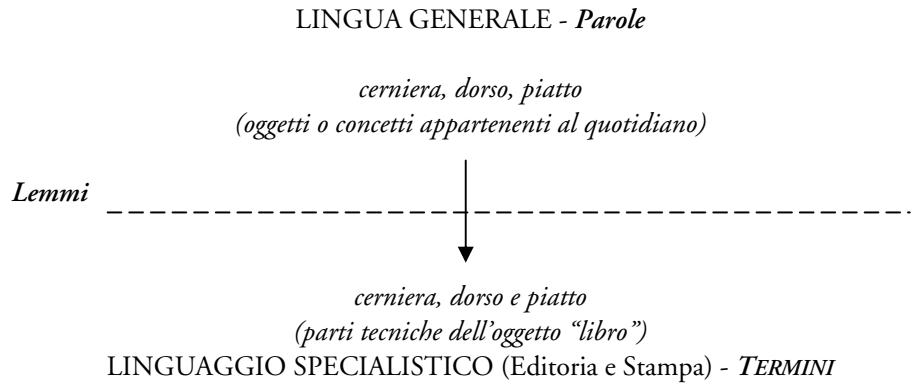
Parimenti, l'incremento del processo di determinologizzazione nel corso degli ultimi anni può essere giustificato anche dal ruolo fondamentale che i *media* hanno assunto all'interno della comunicazione, in particolare a partire dall'era della diffusione di Internet: nella lista dei termini oggetto del lavoro di tesi da noi affrontato, è il caso, ad esempio, di *best seller* o di *merchandising*: due termini specialistici, originariamente utilizzati in lingua inglese per indicare rispettivamente un "successo di libreria" e una particolare modalità di vendita. Oggi, essi sono entrati a far parte non solo della lingua italiana a tutti gli effetti (si tratta infatti di prestiti accettati *in toto* e compaiono anche nei dizionari della lingua italiana), ma soprattutto sono di uso comune anche tra i non esperti: nessuno direbbe "il libro di Tolkien è un successo di libreria", ma piuttosto "il libro di Tolkien è divenuto un *best seller*". Questo è dovuto, in gran parte, allo sviluppo

della modalità di vendita online di qualsiasi bene e servizio, tra cui appunto l'oggetto "libro". Dunque, «almost any event that obtains extensive media coverage may trigger de-terminologization» (Meyer e Mackintosh, 2000:127), puntualizzando che il mezzo che consente maggiormente una *extensive media coverage* è il Web.

"Predisposizione" alla determinologizzazione

I *media* contribuiscono, come abbiamo visto, in modo non indifferente a far sì che un termine sia determinologizzato ed entri a far parte della lingua generale. Tuttavia, Meyer e Mackintosh ritengono che alcuni termini posseggano, per natura, una certa "predisposizione" ad essere determinologizzati: «items which are **simple** (and "**user-friendly**" in other ways) probably have a higher likelihood of de-terminologizing than those which are not» (2000:127-128; gras*set*to nostro). Termini come *mouse* o *desktop* (per rimanere in tema informatico) sono "predisposti" al processo di determinologizzazione perché costituiscono delle metafore trasparenti semplici da utilizzare per i non-esperti. In particolare per questi due termini, il processo di determinologizzazione è stato immediato poiché unità lessicali originariamente appartenenti alla lingua comune, le quali hanno subito un processo di terminologizzazione con l'avvento del *computer* per poi rientrare nella lingua generale con un'altra connotazione.

Questo è anche il caso di alcuni lemmi oggetto del presente lavoro: *cerniera*, *dorso* e *piatto*, ad esempio, sono stati presi in prestito dalla lingua generale e "terminologizzati" per indicare oggetti specialistici, ossia parti del libro. Essi hanno subito il seguente processo:



La loro originaria appartenenza alla lingua comune li rende più "predisposti" di altri ad un processo di determinologizzazione mediante il quale rientrano nell'uso generale della lingua con una nuova connotazione. Subiscono il processo rappresentato nello schema sopra indicato, ma al contrario, ovvero con la freccia rivolta verso la lingua generale o, addirittura, entrano a far parte di un altro dominio specialistico.

Tale "predisposizione" naturale desta in Meyer e Mackintosh il sospetto che, in alcuni casi, gli esperti designino gli oggetti specialistici con nomi, in un certo senso, già predisposti ad un'eventuale determinologizzazione («certain terms appear to be *deliberately* formed in a user-friendly way», 2000:129). *Mouse* e *windows* possono essere stati presi in prestito dalla lingua generale appositamente per creare quel gioco metaforico che consente alla mente di un non-esperto di memorizzarli immediatamente: «software developers have become keenly aware of the marketing potential of metaphors» (2000:128).

Tipologie di determinologizzazione

Nella "migrazione" da un dominio terminologico alla lingua generale, secondo Meyer e Mackintosh, un termine può subire due tipi di modifiche che interessano il piano semantico: il significato specialistico originario può essere mantenuto (sebbene ci siano lievi mutamenti) o il significato del termine originale si trasforma in modo sostanziale inducendo ad una "diluizione" della nozione terminologica. Vediamo ora in dettaglio i passaggi che sottostanno alle due tipologie di determinologizzazione riportando anche alcuni esempi tratti dai lemmi oggetto del nostro lavoro.

Significato terminologico inalterato

> «When a term starts to be used in general language, the *essence* of the concept perceived by laypersons is similar to that perceived by experts. In other words, when laypersons refer to the concept, they are still evoking its basic *domain sense*» (Meyer e Mackintosh, 2000: 113-114).

In questi casi, la connotazione originale del termine è presente anche nell'uso determinologizzato della lingua generale, in quanto i «non-esperti colgono l'essenza del concetto allo stesso modo in cui avviene per gli esperti». Tuttavia, cogliere l'"essenza" di un concetto non vuol dire che il parlante "profano" utilizza tale concetto con la medesima intenzione e con le medesime sfumature sottostanti ad esso: tali significati possono quindi subire dei piccoli mutamenti considerati inevitabili poiché si tratta, in

ogni caso, di un passaggio da un sistema linguistico ad un altro. La conseguenza più ovvia di tale passaggio risiede nel fatto che il significato della parola determinologizzata diviene più superficiale e dunque di grado leggermente più generico rispetto al significato del termine originale.

Nel nostro caso, un esempio di tale fenomeno è costituito dal sintagma *cessazione della vendita*: un esperto del *set*tore commerciale è a conoscenza di tutte le trafile burocratiche, dei documenti, della tempistica che tale denominazione include; al contrario, un non-esperto è consapevole che la conclusione di un'attività commerciale implica una serie di azioni burocratiche e pratiche da portare a termine, ma non conosce in dettaglio cosa esse comportino o come siano denominate, riducendo così il sintagma "cessazione della vendita" al significato più superficiale di, ad esempio, "fine della vendita di un libro" oppure, a volte, di "chiusura di un'attività commerciale". Anche il lemma *cliente* può essere portato ad esempio per questa tipologia di determinologizzazione: una persona non esperta ritiene un cliente chiunque acquisti un libro presso una libreria e non, ad esempio, una libreria che acquista da un distributore un insieme di libri da esporre nel proprio punto vendita. Questo implica un uso più colloquiale del lemma in seguito al processo di determinologizzazione: «many determinologized lexical items are used more colloquially than they are in specialized discourse» (Meyer e Mackintosh, 2000: 124).

"Diluizione" del significato terminologico

Questa tipologia di determinologizzazione prevede un "allontanamento" del significato della parola determinologizzata dal significato del termine specialistico originale, arrivando a non designare più unicamente il medesimo oggetto o la stessa nozione:

> «the de-terminologized word has "loosened" so much that it no longer designates the same concept that the original term did. In other words, when laypersons use the word, it is not with the intention of designating the basic domain sense of the original term» (Meyer e Mackintosh, 2000:115).

Un discostamento dal significato terminologico originale implica gradi differenti di cambiamento semantico a seconda dei casi: un'estensione del senso originale a più oggetti che rimangono comunque legati ad esso (ad esempio, un lemma che, nel dominio originale, designa un oggetto non concreto e, nell'uso determinologizzato, può designare una persona o una sua caratteristica) oppure un distacco più o meno netto della parola determinologizzata dal significato originale; questa infatti può arrivare a non connotare più nessun elemento del dominio originale di appartenenza ma ad assumere

un significato, in parte o del tutto, diverso. La perdita parziale o totale del significato originale di un lemma può essere anche la conseguenza dell'aggiunta di elementi connotativi che il significato originale non possedeva: le due autrici riportano l'esempio di *download* che dal dominio informatico è passato a designare anche un «trasferimento di responsabilità da un livello alto ad uno più basso», arrivando dunque a far parte di un dominio completamente differente da quello di partenza: «*download:* computing → general-language → politics» (Meyer e Mackintosh, 2000:134).

Le parole determinologizzate, appartenenti ad una o all'altra tipologia appena descritte possono subire, a volte, dei cambiamenti sul piano grammaticale: alcuni nomi, ad esempio, una volta determinologizzati, possono essere utilizzati come verbi o viceversa; in lingua inglese, è frequente che un gerundio venga usato come un nome (all'interno del nostro lessico, è il caso di *merchandising*) o i verbi determinologizzati siano utilizzati come aggettivi. Inoltre, Meyer e Mackintosh individuano un "rapporto" differente tra lemmi determinologizzati e preposizioni rispetto a quello tra lemmi e preposizioni in un contesto "non-determinologizzato": accade, infatti, che le preposizioni associate ad un verbo specialistico non siano le stesse associate allo stesso verbo dopo il processo di determinologizzazione. Anche in questo caso, gli esempi più rappresentativi vengono dalla lingua inglese [4].

L'impatto della determinologizzazione

È necessario sottolineare che il processo di determinologizzazione non è un semplice passaggio di un'unità lessicale da un sistema specialistico (*starting point*) a quello generale (*end-point*), bensì una procedura piuttosto complessa ricca di tante sfumature ed eccezioni che coinvolgono, come è stato descritto sopra, i diversi livelli della comunicazione, quali la grammatica, il registro, il significato del lemma prima e dopo la determinologizzazione e infine la presenza e il grado di polisemia. Questo crea conseguenze non trascurabili sia per la lingua generale sia per i domini specialistici. Un caso di complessità di tale fenomeno è quello del *va-et-vient* di un lemma dal sistema generale ad uno specialistico per poi far ritorno in quello generale:

> «Il y a un mouvement de va-et-vient constant entre les processus de lexicalisation et de terminologisation. Une unité lexicale qui a été terminologisée peut se relexicaliser si elle est utilisée en tant que terme générique et perd de ce fait sa spécificité dans le domaine (par exemple *ordinateur, puce, frein*)» (Meyer e Mackintosh, 2000:137).

Si tratta di un processo molto diffuso che tuttavia può creare confusione poiché, il più delle volte, il termine che ritorna a far parte della lingua generale non ha le stesse

connotazioni iniziali (ossia le medesime sfumature di significato che aveva prima di subire il processo di terminologizzazione), e vengono a crearsi casi di polisemia e dunque problemi di comprensione per i non-esperti. Di atti, «the new de-terminologized sense *co-exists* in general language with an older, general-language sense of the same lexical item». Dunque, «the end-point of de-terminologization may be not only the de-terminologized sense, but also the original general-language sense» (Meyer e Mackintosh, 2000:130-131).

Tale processo può avere anche degli effetti "terminologici": in primo luogo, può accadere che una parola determinologizzata venga reinserita nel dominio specialistico a cui apparteneva prima della determinologizzazione; in secondo luogo, una parola determinologizzata può tornare ad essere usata non solo nel dominio di partenza, ma anche in altri domini specialistici. Nel primo caso,

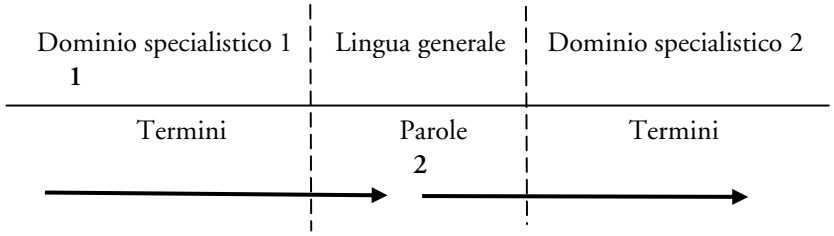
> «when a lexical item starts to become widely used in general language, it can be very tempting for experts in the original domain to "cash in" on its popularity by re-applying it within the domain, but to concepts that are broader than the original terminological concept» (Meyer e Mackintosh, 2000:131-132).

Questo comporta un ritorno del termine al dominio di partenza, ma con una "svalutazione" del significato, ossia con una connotazione più colloquiale e generica assunta in conseguenza al processo di determinologizzazione. Il "gioco" di terminologizzazione e determinologizzazione a cui si assiste in questi casi provoca fenomeni di polisemia *intradomain*: il *va-et-vient* da e verso uno stesso dominio specialistico attribuisce ad un'unità specialistica uno *status* polisemico in contraddizione con il proprio carattere monosemico. Allo scopo di favorire la monosemia *intradomain*, Meyer e Mackintosh suggeriscono: «it would be helpful for terminographers to develop a greater sensitivity to this phenomenon, ideally describing it explicitly in their term records» (2000:133-134).

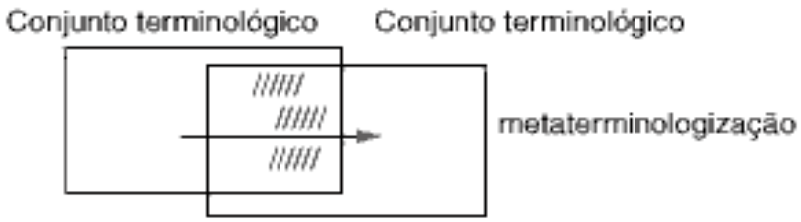
Nel caso di un "ritorno" ad un dominio specialistico, ma non obbligatoriamente quello di partenza,

> «when a term becomes well-known, the general public begins to use it. This *general* public, however, includes *experts* in a variety of domains. Whether consciously or unconsciously, these experts may cash in on the word's popularity and familiarity by using it to designate new concepts in their domains of expertise» (Meyer e Mackintosh, 2000:132).

Tale processo è stato denominato dalle due autrici ***re-terminologization*** (*riterminologizzazione*) (nello schema seguente, è il processo contrassegnato dal valore *2*, mentre il valore *1* indica la determinologizzazione).



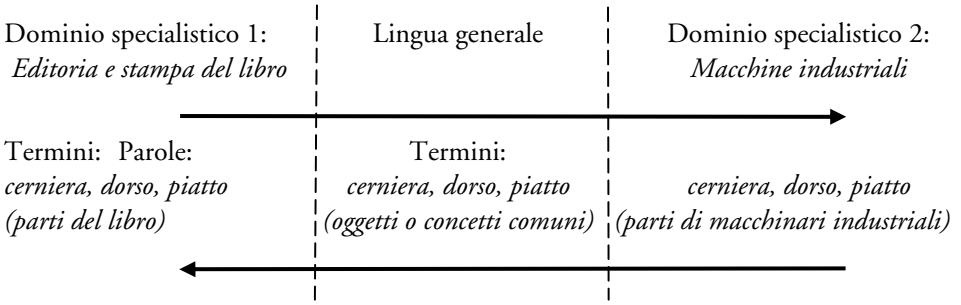
È il caso, ad esempio, di *best seller*: nella sua accezione originale, il termine indica un libro che ha riscosso particolare successo; successivamente, è entrato nel sistema della lingua generale (grazie ai meccanismi mediatici che lo hanno reso una *buzzword*); oggi, tale lemma è utilizzato anche nel dominio informatico per indicare un *software* o applicazione molto diffuso oppure, nel *set*tore musicale, dove indica un successo musicale o un CD di successo. In questo caso, si sviluppa un tipo di polisemia denominata dalle autrici *interdomain*, poiché un medesimo termine può indicare oggetti e nozioni appartenenti a più domini specialistici, conservando tuttavia alcuni tratti semantici comuni. È un caso di *metaterminologização* (metaterminologizzazione [5]) come è definito da Barbosa nel suo articolo *Delimitação do conceito e da definição do termo técnico e científico: percursos epistemológicos e metodológicos*.



Conjunto terminológico: spazio in cui il termine conserva tratti semantici comuni.

Il lemma ha dunque conservato un nucleo semantico comune a tutti i domini specialistici in cui viene utilizzato.

Riprendiamo ad esempio i tre lemmi già nominati in precedenza, *cerniera*, *dorso* e *piatto*, i quali, possono anche designare parti di macchinari industriali:



Un altro esempio significativo di questo "trasferimento interdominiale" è costituito dai lemmi *dittico* e *trittico.* Originariamente, entrambi designavano rispettivamente una coppia e una tripletta di tavolette esposte nelle chiese durante le sacre celebrazioni recante l'immagine dei santi venerati oppure un dipinto costituito da due o tre raffigurazioni distinte chiuse in una cornice, quindi concetti e oggetti appartenenti al dominio dell'arte sacra. Successivamente, l'uso di questi termini in più domini ha fatto sì che essi subissero un processo di determinologizzazione mediante un allargamento di significato. Nella lingua generale, infatti, questi lemmi indicano insiemi composti rispettivamente da due e tre oggetti uguali. A partire da questa connotazione più generica, questi lemmi sono stati sottoposti nuovamente ad un processo di terminologizzazione divenendo anche termini specialistici del dominio della pubblicità e in particolare del sottodominio della cartellonistica pubblicitaria. Infatti, il *dittico* e il *trittico* sono «pannelli pubblicitari autoreggenti a due o tre ante a facce convergenti o divergenti».

Come è possibile notare, dunque, il percorso che un lemma opera mediante la determinologizzazione non ha un unico *end-point* fisso, ma una serie di "mete" che spaziano da un linguaggio altamente specialistico ad uno altamente generale, questo poiché «determinologization is [also] controlled by an extra-linguistic factor, namely the degree of diffusion of domain concepts into the consciousness of the general public» (Halskov, 2005). Tale eterogeneità di fenomeni "determinologizzanti" rappresenta perfettamente l'eterogeneità dei lemmi che costituiscono l'oggetto del nostro lavoro.

Conclusioni

Lo studio dei processi descritti nel presente articolo, in particolare quello di determinologizzazione, hanno consentito la costruzione di un *corpus* eterogeneo. L'utilizzo di tale *corpus* si è rivelato uno strumento molto efficace per l'individuazione e la verifica degli equivalenti in lingua italiana dei termini in lingua francese. Infatti, come anche Halskov (2005) ha evidenziato nel suo articolo, lo studio approfondito di questo fenomeno in continua espansione, si rivela notevolmente importante per ottimizzare la ricerca terminologica. La conoscenza di questi principi è fondamentale in particolar modo quando si utilizza la rete, considerata oggi la più importante ed esaustiva fonte di ricerca per terminologi/terminografi. La ricerca in rete, infatti, realizzata attraverso "parole chiave", pone una serie di problemi causati dalla natura dei lessemi e dalla loro predisposizione o grado di determinologizzazione e/o terminologizzazione:

> «The extensive usage of terminology from a domain like IT by vast numbers of non-experts in a variety of communicative *set*tings complicates the automatic extraction task. Although determinologized usage can be avoided by using corpora which have been manually compiled and are known to represent specialized communication between experts, such corpora are expensive to come by and age swiftly» (Halskov, 2005).

Come è emerso dai paragrafi precedenti, dunque, il fenomeno della determinologizzazione deve essere approfondito e divenire parte integrante del lavoro di un terminologo/terminografo, ma anche di altri esperti quali i lessicografi, in virtù del suo alto potenziale nel generare polisemia all'interno di un dominio (*intradomain*) e anche tra più domini (*interdomain*) e delle conseguenze sulle procedure tradizionali di creazione e armonizzazione dei termini.

Note

[1] È necessario far notare che il termine *determinologizzazione* è un neologismo, poiché non è stato riscontrato nella lingua italiana, ma solo in lingua francese e in lingua inglese. Si è deciso di "coniare" questo nuovo termine principalmente per una questione di chiarezza nei confronti del lettore italiano. Tale termine è un adattamento dei termini in lingua francese (*déterminologisation*) e inglese (*de-terminologization*), scelta giustificata anche dall'esistenza nella lingua italiana della denominazione del processo contrario, ossia *terminologizzazione*.

[2] Tratto da ASS.I.TERM, Associazione Italiana per la Terminologia, online: <www.assiterm91.org/it/index.php?option=com_content&task=view&id=11&Itemid=25>.

[3] Si vuole sottolineare che l'articolo delle due studiose ha costituito un punto di riferimento di notevole importanza per lo svolgimento del lavoro al progetto sopra descritto, in quanto,

attualmente, costituisce lo strumento più esaustivo per l'approfondimento del processo di determinologizzazione.

[4] All'interno del lessico da noi analizzato, non sono stati riscontrati altri lemmi appartenenti a questa tipologia di determinologizzazione.

[5] Processo di trasposizione di un *termine* da un ambito specialistico ad un altro senza che si attui nessuna modifica totale del significato o, perlomeno, con la conservazione di alcuni tratti semantici.

Bibliografia

Aprile M. (2005). *Dalle parole ai dizionari*. Bologna: Il Mulino.

ASS.I.TERM, Associazione Italiana per la Terminologia. Online: <www.assiterm91.org/it/index.php?option=com_content&task=view&id=11&Itemid=25> (consultato il 21/05/07).

Barbosa M.A. (1998). *Delimitação do conceito e da definição do termo técnico e científico: percursos epistemológicos e metodológicos*. Universidade de São Paulo, Brasil. Simpósios de RITerm - Atas 1988-2002.

Bertaccini F., Prandi M., Sintuzzi S., S. Togni (2004). *Tra lessico naturale e lessici di specialità: la sinonimia*. Università di Bologna - SSLMIT, Forlì.

Dubuc R. (1992). *Manuel pratique de terminologie* (3ª ed.). Brossard: Linguatech.

Halskov J. (2005). *Probing the Properties of Determinologization - The DiaSketch*. I: Lambda bd. 29, Copenhagen Business School.

Magris M. et al. a cura di (2002). *Manuale di terminologia: aspetti teorici, metodologici e applicativi*. Milano: Hoepli.

Meyer I e K. Mackintosh (2000). *When terms move into our everyday lives: an overview of de-terminologization*. "Terminology", 6(1): 111-138.

Riediger H. (s.d.). *Cos'è la terminologia e come si fa un glossario*. <erm-minator.it/corso/doc/mod3_termino_glossa.pdf> (consultato il 30/04/07).

Terminologia e Classificazione nel centro di Documentazione della Democrazia Cristiana

ROBERTO GUARASCI

This article presents the activity of the center of Documentation of the Christian Democrat party from 1945 up to the collapse of the first Republic. Its function was that of collecting, organizing, sorting and analyzing records information within the party. In particular, the article focuses on the activities of Charles Dané who, from 1954 onwards, was responsible for information management within the party. Dané also used Paul Otlet's (UDC) Universal Decimal Classification System and terminological information extraction systems that he efficaciously associated with the activity of microfilming of documents related to the Christian Democrat party.

Keywords: Classification system – Democrazia Cristiana – information management

> «Se è vero che la verità è unica, non possiamo pretendere che una affermazione offerta alla mente di un uomo libero che sente da altri pulpiti altre affermazioni, anche se vera, sia automaticamente accettata. La conquista della verità è fatta attraverso il libero ragionare, la possibilità di confrontare e particolarmente di documentarsi. Pertanto la nostra propaganda che ha per fine ultimo la verità, deve stimolare la discussione, fornire una ricca documentazione per dimostrare quanto siano buoni i nostri programmi e le nostre opere, quanto siano fallaci i programmi avversari» [1].
>
> «Per combattere una buona battaglia occorre non solo avere buone fanterie e buone artiglierie, ma conoscere bene anche il terreno su cui si combatte e l'efficienza nonché i piani dei nemici. … Anche a questo deve provvedere in gran parte la SPES, - [Studi, Propaganda e Stampa] la quale oltre ad essere un organo di formazione e di propaganda deve essere un centro di informazioni». I dati, da raccogliere localmente e trasmettere alla sede centrale, erano sia «Dati di pubblico dominio» e «comunque tali da interessare il centro che direttamente non potrebbe conoscerli», sia «Dati riservati relativi a partiti, amministrazioni, organizzazioni sindacali, enti, ecc.». La trasmissione avrebbe dovuto avvenire con cadenza quindicinale utilizzando un questionario appositamente predisposto con l'accortezza che «per i dati riservati le comunicazioni dalle provincie devono essere riservatissime e, se necessario, fatte anche personalmente» [2].

L'idea di far nascere una struttura specificatamente destinata alla raccolta e analisi delle informazioni documentali all'interno del neonato partito della Democrazia Cristiana è di Giuseppe Dos*set*ti, all'epoca vice segretario nazionale, che Il 22 *set*tembre del 1945 invia alle segreterie dei comitati provinciali e regionali una circolare per l'imme-

diata costituzione degli uffici Studi Propaganda e Stampa. «L'ufficio studi, propaganda e stampa (Ufficio S.P.E.S.) dovrà coordinare e, in certo modo, riassumere le attività che sino ad ora venivano, un po' frammentariamente e separatamente, svolte dalle Segreterie delle eventuali commissioni di studio, dalla redazione dei periodici locali, dai dirigenti della propaganda» [3]. La stessa circolare definisce la struttura organizzativa dell'ufficio che risulta suddiviso in: servizio inchieste, servizio studi, servizio attività culturale, servizio propaganda, servizio raccolta ed emissione informazioni, servizio stampa periodica del Partito.

La sanzione formale della costituzione della struttura si avrà nella riunione della Direzione Nazionale del 1 maggio 1946, il cui resoconto sintetico pubblicato sul Il Popolo del 3 maggio recitava, tra l'altro: "Si è provveduto anzitutto ai due uffici fondamentali della Segreteria Centrale, cioè l'ufficio Organizzazione e l'ufficio Studi, Propaganda e Stampa (Spes). È stata affidata a Giuseppe Dos*set*ti la coordinazione generale dei due uffici al primo dei quali è stato preposto Giulio Pastore e al secondo Amintore Fanfani" [4].

La funzione di Documentazione e raccolta delle informazioni è, quindi, fin dall'origine inclusa nella struttura di stampa e propaganda voluta da Dos*set*ti anche se, fino ai primi anni cinquanta, le esigenze elettorali faranno si che la stragrande maggioranza delle risorse e delle attività vengano destinate alla redazione di testi con finalità squisitamente propagandistiche. La struttura – in quei primi anni – continua incessantemente a crescere e a, lentamente, diversificarsi. Nel 1951 il vice segretario nazionale Giorgio Tupini ha alle proprie dipendenze un Servizio Articoli e Documentazione, detto anche "Super Redazione" – diretto da Giuseppe Sala – che ha, essenzialmente, il compito della redazione della terza pagina della stampa quotidiana del partito; dallo stesso vice segretario dipendono, inoltre, l'ufficio centrale Spes, l'ufficio propagandisti, l'ufficio attivisti, "popolo e libertà" e l'ufficio cinema. In totale, al 31 luglio 1951, risultano assegnati 39 dipendenti di ruolo, numero più alto tra tutti i vari comparti della direzione nazionale [5].

L'Ufficio Documentazione acquista una sua autonoma connotazione qualche anno più tardi, nell'aprile 1954, «con lo scopo di creare e rendere veramente funzionale un archivio documentativo delle attività del Partito affiancato alla ricerca di elementi di documentazione su fatti politici di particolare importanza» [6]. Centrato sull'idea della microfilmatura massiva della documentazione l'ufficio si configurò rapidamente come un vero e proprio centro di documentazione e di *intelligence* con la missione strategica di ridurre i volumi dei documenti cartacei mediante l'uso di nuove tecnologie rendendoli così più facilmente consultabili. Il neonato ufficio "documentazione-microfilmico" viene tenuto a battesimo dai rappresentanti degli uffici della direzione nazionale – primo fra tutti Benigno Zaccagnini – che ne decidono i metodi di «attua-

zione e sviluppo» [7]. Da questo momento la sua storia è inscindibilmente legata alla figura di Carlo Danè [8] – che lo reggerà, di fatto, fino alla fine [9].

Il non ancora trentenne Carlo Danè proviene dai ranghi della DC di Savona ed è da qualche anno nei ruoli di partito. Inviato dal Dirigente Nazionale Spes dell'epoca, Valdo Fusi, a coordinare la propaganda in Sicilia, in occasione delle amministrative del 1952, aveva avuto come viatico una sintetica lettera «*O vinci o crepa. Cordialmente. Valdo Fusi*». Tornato a Roma lavora all'interno della Spes con Vincenzo Sangalli, Dino Del Bo e Raimondo Manzini fino a quando con Mariano Rumor assume un ruolo di maggiore responsabilità nel nascente ufficio di documentazione.

Sotto la gestione Danè ed in assonanza con l'efficientismo della nuova segreteria Fanfani l'ufficio diventa il centro di raccolta, organizzazione e smistamento delle informazioni operando in maniera discreta e riservata e con un profilo volutamente molto basso tanto che ne ignoreranno l'esistenza anche molti addetti ai lavori confondendolo – spesso e volentieri – con il più evidente servizio Stampa e Propaganda dal quale originava. Ma i compiti erano sostanzialmente diversi. Nei *dossier* riservati dell'ufficio di documentazione la trascrizione della lettera del generale egiziano Mohamed Neguib al comandante Junio Valerio Borghese per la costituzione della "Federazione di combattenti del Mediterraneo" per riunire a Roma «le forze militari che durante l'ultima guerra combatterono contro l'egemonia inglese» corredata dell'appunto dattiloscritto «sembra che la cosa sia destinata a morire a meno che non intervenga l'ex maresciallo Graziani». Tra l'altro i due divergevano sull'impostazione dell'associazione. L'uno voleva accentuare l'aspetto anti-inglese, l'altro quello anti-comunista e anti-sovietico. O, ancora, la lettera del dirigente provinciale Spes di Napoli a Franco Maria Malfatti con allegata copia dell'assegno a riprova di finanziamenti occulti nella campagna elettorale dei partiti avversari: «la faccenda è di effetto – dirà l'avv. Clemente – più di quanto possa sembrare: abbiamo sempre parlato di queste cose nei nostri comizi ma non abbiamo mai potuto offrire la prova di quanto dicevamo» [10].

«Il Partito – dirà Danè – ha quasi sempre bisogno più che [del]la congerie di dati e notizie, di dati e fatti già orientati su tesi» [11]. L'ufficio documentazione si deve occupare di costruire una rete di contatti e relazioni per potere, in ogni momento, reperire i dati necessari a supportare l'attività decisionale dei vertici del Partito. Dati che vanno strutturati ed elaborati in relazioni alle specifiche finalità alle quali sono destinati.

Parallelamente all'ufficio documentazione – in quegli anni – viene anche avviata la costituzione di una agenzia di stampa, inizialmente denominata SADI (Servizio Articoli Documentazioni Internazionali) e poi, successivamente, Sviluppo Democratico (S.D.) - Agenzia di commento e informazione.

Nella prima stesura dell'atto di nascita di questa agenzia, firmata da Marcello Capitanio, vengono delineate le motivazioni della proposta: «La maggior parte dei giornali

sovvenzionati dalla D.C., dall'A.C. e da Enti governativi o comunque vicini al partito, non dimostra sempre di avere una chiara ed uniforme impostazione politica e propagandistica sui problemi fondamentali, impostazione che esigerebbe invece un preciso coordinamento, soprattutto quando l'univocità della stampa sarebbe strettamente necessaria. Per raggiungere lo scopo si dovrebbe approntare questo strumento che permetterebbe di intervenire ed esercitare una diretta influenza ed un più stretto controllo sulla stampa collegata mediante la costituzione di una Agenzia che possa contribuire con materiale di scelta all'orientamento ed alla compilazione dei quotidiani fiancheggiatori».

Nella successiva bozza del progetto di costituzione viene indicato come Direttore dell'Agenzia Antonio Di Raimondo e fanno parte della direzione Fabrizio Abbate (movimenti giovanili), Pino De Rosa (università-tecnici), Carlo Danè (Documentazione), Giuseppe Fornaro (Associazionismo - Stampa Giovanile), Paolo Guerra (Movimenti studenteschi), Roberto Chiodi (magistrati - quotidiani), Carlo Napoli (Rai-TV - Mondo cattolico), Cristiano Zironi (Movimenti universitari) [12].

Da metodi empirici finalizzati all'episodicità della propaganda – emblematiche le uova di pasqua con il simbolo della Spes nelle elezioni del 18 aprile 1948 [13] – si passa ad una raccolta sistematica di informazioni la cui strutturazione viene più volte rivista in funzione del mutare delle esigenze organizzative e funzionali. Gli stabili contatti con le ambasciate degli Stati Uniti e della Gran Bretagna servono non solo ad acquisire informazioni riservate ma anche a studiare e perfezionare i metodi di classificazione che, dopo pochi mesi, nel novembre 1954, portano all'avvio della raccolta e schedatura di tutto il materiale edito dalla Spes e dal partito, e alla costituzione di una fototeca che raccoglieva tutte le foto giacenti presso i vari uffici del partito. Saranno censite 3600 immagini e definite 111 voci di classificazione [14]. La fototeca è realizzata con un sistema di conservazione del negativo e del positivo legati da un indicatore numerico e dotati di uno schedario di corredo per soggetti [15].

Nel maggio 1956 l'Ufficio Documentazione comincia ad applicare diffusamente la Classificazione Decimale Universale, dopo aver scritto alla Federazione Internazionale di Documentazione (F.I.D.) per chiedere informazioni ed acquistare copia dei testi necessari. L'applicazione di questo sistema di classificazione era comunque stato avviato quasi contestualmente alla nascita dell'ufficio e parallelamente ad altri sistemi inizialmente in uso. Difatti in una minuta anonima datata 11 giugno 1954 si legge: «L'ufficio tecnicamente svolge il lavoro in tre fasi: 1. Studio e ricerca, 2. Riproduzione in microfilm, 3. Classificazione e per quest'ultima si è adottato il sistema internazionale della Classificazione Decimale Universale» [16].

«Si sottolinea in particolare – scrive Danè – che si intende adottare per la classificazione del materiale il sistema decimale universale (schede) e per la raccolta del medesimo la riproduzione micro filmica delle fonti e dei documenti» [17].

Dopo un primo periodo di utilizzo massivo questo sistema di classificazione sarà abbandonato perché ritenuto troppo macchinoso e inadatto alle specifiche finalità dell'ufficio. Al suo posto viene costruito uno schema di classificazione a più livelli. «Il sistema della Classificazione Decimale Universale – affermerà Danè – che era stato in un primo tempo prescelto come l'optimum dei metodi di classificazione venne poi abbandonato perché troppo complesso e di difficile applicazione. La sistemazione del materiale raccolto avveniva perciò in base a duecento voci, sufficienti a dare una suddivisione logica per materia» [18].

Le voci complessive del primo livello di classificazione sono inizialmente 86, poi 98 e – nel 1958 – appunto 200 con quasi 1500 sottovoci. La voce "Biografie" – ad esempio – è articolata in "capi", "maggiori esponenti" e "personaggi vari". La prima comprende *dossier* su Togliatti (PCI), Nenni (PSI), Pacciardi (PRI), Saragat (PSDI), Malagodi (PLI), Carandini (PR), Fanfani (DC), Covelli (PNM), Lauro (PMP). De Marsanich (MSI). Le biografie non si limitano alle sole informazioni di pubblico dominio ma raccolgono anche particolari riservati e notizie fornite da informatori anonimi dei quali viene indicato anche il grado di affidabilità. Il sistema di selezione delle informazioni attraversò diverse fasi. Inizialmente – prima dell'adozione della CDU – si trattò di ritagliare articoli e testi a stampa incollandoli su schede di diverso colore a seconda della macro-voce o argomento trattato. In questa fase le macro-voci erano solo due: Italia (verde), Esteri (rosa). Successivamente si aggiungeranno Pubblicazioni (bianco) e Biografie (giallo) e le voci cresceranno fino a 252 suddivise in "Interni n. 90 voci, Esteri n. 99 voci, Biografie n. 60 voci, Manifesti n. 3 voci". Sarà anche predisposto un siglario/elenco delle abbreviazioni convenzionalmente usate per la redazione delle schede "mod. 1099 Buffetti" [19] evidenziando – contestualmente – la priorità della creazione di un "bollettino delle radioricezioni" per la memorizzazione, schedatura e classificazione delle informazioni con quel mezzo veicolate. Il sistema dei colori continuerà a sussistere anche quando si cercherà di utilizzare la CDU e quando poi, abbandonata questa, si passerà al sistema auto costruito per voci, esemplato su quello in uso presso la Presidenza del Consiglio dei Ministri per i "Documenti di Vita Italiana". In assenza di strumenti di elaborazione dei testi la costituzione delle schede e dei *dossier* tematici era un lavoro dispendioso ed immane che venne realizzato con un personale relativamente modesto e, a diversità di esperienze similari, con l'assillo dei tempi della politica e non con la tranquillità dell'azione culturale di lungo respiro. «Un'altra difficoltà da noi incontrata – dirà molti anni dopo Danè – è consistita nella schedatura dei documenti che per anni abbiamo realizzato attraverso l'elaborazione dei cosiddetti abstract più o

meno ampi. … Questo metodo però richiedeva l'impiego di diverse persone sia nella fase preparatoria delle schede sia in quella della loro quotidiana sistemazione manuale nello schedario generale, cosicchè la mancanza di un numero adeguato di addetti a questi compiti ci costrinse a rinunciare. …Si passò allora dove possibile, naturalmente, distruggendo almeno parzialmente le fonti, all'operazione del ritaglio …» [20]. Massiccia era comunque l'attività di microfilmatura a proposito della quale è appena il caso di ricordare che questo tipo di supporto di memorizzazione, utilizzato per la prima volta durante la guerra franco-prussiano e poi teorizzato nell'uso documentale da Otlet e Goldschmidt [21] nel primo decennio del XX secolo, era stato diffusamente usato per scopi di intelligence durante la seconda guerra mondiale ma la sua utilizzazione massiva come strumento di duplicazione e memorizzazione documentale era – negli anni cinquanta – di là da venire. Non a caso la prima norma che consentirà – utilizzando appunto i microfilm – la duplicazione sostitutiva dei documenti amministrativi è il D.P.C.M. 11 *set*tembre 1974 *Norme per la fotoriproduzione sostitutiva dei documenti di archivio e di altri atti delle pubbliche amministrazioni*.

Nel 1956 l'ufficio attiva una propria specifica sezione microfilm dotata di «riproduttore su pellicola 35 mm per riprese fino alla misura massima di cm 50x80, tre schedari metallici kardex per l'archivio fotografico e microfilmico, ingranditore, smaltatrice e sviluppatrice per le riproduzioni su carta». Si microfilmava non solo il tradizionale materiale archivistico o a stampa ma anche tipologie di fonti all'epoca trascurate come i rapporti interni o i manifesti di propaganda. «Si chiede, inoltre, – si legge in una richiesta al dirigente nazionale SPES Arnaldo Forlani – di integrare detta attrezzatura "in sé apprezzabile" con ulteriori apparecchiature per la microfilmatura e con una fotocamera Rolleiflex "per le riproduzioni all'esterno delle manifestazioni della propaganda avversaria» [22].

I primi anni di attività saranno intensi ma non privi degli inevitabili problemi di una struttura siffatta, non ultima la carenza di personale specificatamente dedicato. «L'ufficio documentazione era composto, in partenza, da otto elementi: quattro operatori, due dattilografe in esclusiva (di cui una per i soli ritagli), un archivista ed un fotografo. Attualmente gli addetti sono due, più l'archivista ed il fotografo. Sono necessari, per la ripresa e la continuità di un lavoro serio e sistematico, giorno per giorno: un responsabile, due collaboratori, una dattilografa fissa, l'archivista, il fotografo» [23].

Di fatto, l'ufficio Documentazione possiede un autonomo archivio [24] che, unitamente a quello della Spes con il quale Danè proporrà, senza riuscirci, di unificarlo [25], costituisce, almeno quantitativamente, parte rilevante del più generale archivio della Democrazia Cristiana [26]. Si tratta, comunque, di un archivio in senso lato in quanto composto dai *dossier* del centro di documentazione e, in misura minore, dell'archivio propriamente detto dello stesso ufficio. Nonostante la presenza di archivisti nell'orga-

nico della direzione nazionale fin dagli anni 50 [27], a differenza di quanto accade in molte realtà periferiche, non sembra riscontrarsi – in quegli anni – grande cura verso la conservazione di alcune tipologie di documenti tanto che Carlo Danè, rispondendo all'on. Ottorino Momoli, presidente del comitato provinciale di Mantova che chiedeva di «esaminare i documenti e gli atti anche riservati della direzione centrale dal congresso di Napoli 1947 al 1949 (Congresso di Venezia)» [28], dirà: «Non mi risulta che gli atti e i documenti della Direzione Centrale siano conservati negli Archivi del Partito oltre una certa data. Ritengo invece che essi (purtroppo, dico, e questo è un mio giudizio strettamente personale) vadano in parte dispersi, in seguito all'avvicendarsi delle varie segreterie e, in parte, distrutti. Il rinvio, da parte del Segretario politico, al dirigente Spes, è in proposito significativo: vuol dire che nulla esiste in quella sede, ma nulla esiste neppure qui all'ufficio documentazione-spes. L'unica cosa che rimane, fortunatamente, è la collezione del quotidiano della D.C. Il Popolo, e questa noi l'abbiamo per intero. Neppure le pubblicazioni ufficiali si trovano con facilità: praticamente infruttuosa sarebbe la ricerca, per esempio, dei vecchi numeri di Popolo e Libertà, o di Civitas e di Libertas» [29].

Molti anni dopo, nel giugno 1991, Gabriella Fanello Marcucci, responsabile dell'archivio storico della Democrazia Cristiana, scrivendo a Carlo Danè solleverà proprio il «problema della distinzione tra archivio e documentazione che credo sia vivo – affermerà – all'attenzione di tutti coloro che, per motivi diversi, affrontino il riordino, la sistematizzazione, la classificazione o l'utilizzo dei fondi costituiti da carte politiche» [30].

Nel 1958 viene proposto un «piano di lavoro per il rilancio dell'attività dell'ufficio» e per «la ripresa di un sistematico lavoro di schedatura di tutto il materiale già raccolto e da raccogliere; la sistemazione delle schede secondo una classificazione che, senza avere la razionalità, l'espansività e l'internazionalizzazione della C.D.U. poggi su un valido criterio empirico, e cioè: una voce generale per l'Italia, con suddivisione per ciascuna regione, per ciascun partito, per ciascun genere di problemi interni (politici, economici, sindacali, religiosi, ecc.), una voce per ciascun paese del mondo con le necessarie sottovoci, una voce per i personaggi. La schedatura sarà duplice: su ogni scheda piccola (9x12) gli abstract degli articoli (dei libri, dei dattiloscritti), con le indicazioni necessarie ad individuare la fonte e la data e a localizzare l'originale; su grandi schede Buffetti (cm. 17x24) l'elenco puro e semplice del materiale raccolto».

Un paragrafo specifico del testo che dettaglia il piano di lavoro è dedicato alla schedatura che «deve evitare segnalazioni generiche, approssimative o di questioni di assoluto carattere generale: in questi casi la scheda diviene inutile. La scheda deve portare in alto a destra, sottolineata, la voce; immediatamente sotto la voce la sottovoce che specifica l'argomento schedato. Quindi l'argomento, con dicitura breve ma chiaramente definita, ed in calce la fonte da cui è tratta la materia della scheda. In basso a sinistra

la sigla di chi scheda. … La scheda compilata deve essere passata all'archivista allegata alla fonte, per poter poi essere da questi collocate, l'una e l'altra, secondo lo specchio di archivio» [31].

Con il crescere del numero delle schede e dei *dossier* e l'articolarsi del reticolo dei termini di indicizzazione, Danè e i suoi collaboratori approcciano artigianalmente il problema dell'estrazione terminologica, della validazione dei nuovi termini e della costruzione dei descrittori.

Le carte Danè abbondano di lunghi elenchi di voci indice più volte aggiornate, disambiguate con l'indicazione di una rudimentale nota d'ambito che ne delinea il campo di applicazione, riscritte e confrontate con i precedenti in uso per costruire rimandi e tabelle capaci di evitare perdite di informazioni. La sistematica lettura della stampa quotidiana offre l'occasione per l'individuazione di ulteriori termini candidati a diventare descrittori che vengono verificati tra i vari redattori e con la lettura comparata di più testi in ambiti cronologicamente definiti. Vengono costruiti elenchi di sigle per indicare sinteticamente ed univocamente le varie testate e si verifica costantemente la disponibilità di elenchi già costruiti da altri soggetti in maniera tale da standardizzare le funzioni e velocizzare le ricerche.

«Nella definizione delle sottovoci bisogna evitare l'uso di sinonimi – si afferma a più riprese – che potrebbero ingenerare confusioni. Pertanto la compilazione della scheda verrà fatta dagli operatori a ciò incaricati previa reciproca consultazione» [32].

Nella «presentazione e meccanismo dell'archivio microfilmico» vengono chiaramente esplicitati alcuni di questi concetti. Oltre agli articoli stralciati da giornali, ai dispacci d'agenzia, ai "dattiloscritti vari", quella che oggi chiameremmo letteratura grigia, agli opuscoli, ai manifesti ed a «fotografie di particolare valore documentale» si intendeva procedere ad una microfilmatura integrale di «cinque giornali base» provvedendo anche alla loro «raccolta dal '46 ad oggi che dovrà costituire l'ossatura principale in fatto di informazioni e ricerche d'archivio. (…) Bisogna anche chiarire che la filmazione delle annate dei giornali base avviene, comunque, dopo una ricerca delle notizie su ogni pagina e della loro classificazione. Le sezioni base sulle quali sarà suddiviso l'archivio sono: Interni, Esteri, Biografie, Pubblicazioni. Le cartelline riferentesi ad ognuna di queste sezioni avranno diverso colore che renderà più facile la ricerca oltre che ad essere divise in diversi raccoglitori. Ed i colori potrebbero essere per gli Interni: verde, per gli Esteri: Rosso, per le Pubblicazioni: Bianco, per le Biografie: Giallo. Qui di seguito scriverò una nomenclatura base delle diverse sezioni tenendo, però, presente che, per alcune di queste voci, sarà necessario fare delle sottosezioni. … È chiaro come tale elenco resta veramente solo un indicazione e gli aumenti e le aggiunte saranno continue… La ricerca di qualsiasi argomento con uno schedario siffatto sarà effettuata in poco tempo con queste operazioni. Ad es. volendo ricercare una documentazione

che riguarda l'artigianato, inizieremo consultando il raccoglitore dai cartellini verdi. … Quindi alla lettera A troveremo Artigianato e la notizia sarà così scritta (f. 125 - b. 10 - P. 2/4/46), il che significa fotogramma 125, bobina n. 10 dal Popolo del 2/4/46. … Risulterà chiaro come la sigla data al giornale Il Popolo fa parte di una nomenclatura data preventivamente a tutti i giornali sui quali si intende operare la ricerca».
Nella visione di Carlo Danè l'Ufficio Documentazione ha, comunque, due funzioni basilari: la raccolta dei dati e l'ascolto.

> «I funzione: raccolta dati.
> L'Ufficio Documentazione, in primo luogo, si occupa di reperire il dato, la notizia, il documento, la bibliografia, l'emerografia che necessitano agli uffici del partito. Non già che raccolga, per questa funzione, ne' suoi scaffali tutto quanto può occorrere all'attività multiforme, spesso analitica, del Partito. … Si preoccupa, invece, di stendere una rete di contatti con gli enti vari, Pubblica Amministrazione, Parlamento, Governo, Associazione di Categoria, ecc. , in modo tale da poter utilizzare, volta a volta, queste fonti dirette e specializzate».

> «2 funzione dell'Ufficio Documentazione: la funzione di ascolto.
> L'Ufficio Documentazione è l'orecchio del partito su quel che avviene di significativo per l'azione politica nel Paese e nelle sue organizzazioni: decisioni delle categorie economiche; espressioni dei gruppi di interesse; espressione dei partiti; posizioni dei gruppi parlamentari; vita degli Enti pubblici e privati di maggior peso. L'ufficio documentazione non è la redazione di un giornale di informazioni: seleziona le notizie importanti per una presa di posizione del partito, ideologica o legislativa. Svolge perciò una prima selezione, sia pur rudimentale, tra i fatti che accadono nel paese, scegliendo quelli che interessano il Partito. L'ufficio documentazione non è un ufficio stampa. Non è l'Eco della Stampa. È l'ufficio che tiene aggiornati gli altri uffici del partito su quanto viene riferito sui quotidiani, sulle riviste specializzate, relativamente alla vita dei *set*tori di cui essi si occupano. Egli segnala tali notizie; non le trasmette per intero».

L'Ufficio è organizzato in due sezioni: una politica e una economica. La prima è organizzata in direttrici di lavoro. Una nazionale ed una locale su base provinciale tendente a «determinare la statica e la dinamica politica di ogni singola provincia e si prefigge la precisazione delle modalità della propaganda secondo linee di sociologia politica» [33].

Questa struttura articolata e complessa resterà in piena attività – seppur con alterne vicende – fino agli anni Ottanta quando sarà parte attiva del nascente sistema informativo della Democrazia Cristiana che collegherà tutte le sedi periferiche alla Direzione Nazionale con circa 300 connessioni al 14 *set*tembre 1989 [34], anno dell'avvio a regime. La nuova infrastruttura ovviamente recuperava, in chiave tecnologica, alcune delle principali finalità dell'ufficio documentazione. Essa – oltre alla gestione delle anagrafiche – mirava a «controllare la situazione reale corrente, rilevare tempestivamente fenomeni emergenti ed eventuali anomalie, misurare il grado di presenza del partito nella

pubblica amministrazione o a qualunque livello territoriale al fine di ottimizzare la programmazione politica ed organizzativa» [35].

Danè, in rappresentanza dell'ufficio Documentazione, farà parte fin dall'avvio della commissione – coordinata dall'on. Fiorenzo Maroli – per il "progetto di sistema informativo D.C." concorrendo – unitamente agli altri uffici – a definire le specifiche funzionali con un ruolo esplicitamente evidenziato anche dagli stessi partecipanti [36]. Il sistema informativo, con una strutturazione progettuale all'avanguardia per il periodo, avrebbe rappresentato la necessaria infrastruttura tecnologica per i contenuti dell'ufficio di documentazione. Non ci sarà però tempo e tutto, travolto dal crollo della cosiddetta prima repubblica, sarà consegnato alle cronache dei quotidiani ed ai tentativi – come il nostro – di riscoprire esperienze poco note ma forse non meritevoli dell'oblìo.

Note

[1] Livio Olivieri, *L'Importanza della propaganda per orientare la pubblica opinione*, dattiloscritto della lezione tenuta all'Ateneo di Studi Politici della Democrazia Cristiana, Canazei, 22 agosto 1952, Archivio Privato di Carlo Danè. L'Archivio di Carlo Danè è – di fatto – la sedimentazione documentale dell'Ufficio di Documentazione della Democrazia Cristiana e contiene materiale databile dal 1946 al 1990 circa. È composto da circa venti scatole ed è detenuto da un privato. In esso sono identificabili due parti principali: i materiali ed i *dossier* del centro di documentazione e l'archivio propriamente detto contenente tutta la documentazione, ordinata cronologicamente, sulla nascita e l'attività dell'ufficio. Tale seconda parte – di consistenza modesta – è stato acquistato dal Dipartimento di Linguistica dell'Università della Calabria. Con la dizione "Archivio Privato di Carlo Danè" si identifica tale complesso documentario. Materiale librario e documentale dello stesso Danè è stato anche donato dalla famiglia all'Istituto Luigi Sturzo.

[2] Livio Olivieri *L'Ufficio Provinciale SPES centro di informazioni*, dattiloscritto della lezione tenuta all'Ateneo di Studi Politici della Democrazia Cristiana, Canazei, 30 agosto 1952, Archivio Privato di Carlo Danè.

[3] *Democrazia Cristiana - Bollettino della Direzione Centrale*, Anno I, n. 5 del 30 settembre 1945.

[4] «Mi pare di capire dal tono stesso del comunicato che abbiamo citato, che la Spes doveva esistere, con questo suo nome, anche prima di allora; e del resto le sue funzioni erano pur svolte da qualcuno e da un ufficio della Direzione già da molto tempo…» Archivio Privato di Carlo Danè, *Appunti per una storia della Spes,* dattiloscritto senza data.

[5] *Riepilogo* di Alessandro Gedda, direttore di segreteria, del marzo 1951. Istituto Luigi Sturzo, Archivio Privato Guido Gonella, busta 24, fascicolo 8.

[6] Archivio Privato di Carlo Danè, *minuta dattiloscritta* 11 giugno 1954.

[7] Archivio Privato di Carlo Danè, *Appunto sulla riunione dei rappresentanti degli uffici della*

Direzione riuniti per decidere i metodi di attuazione e sviluppo del nuovo ufficio documentazione-microfilmico. Dattiloscritto, 1954. All'incontro risultano presenti «On. Zaccagnini per l'ufficio problemi del lavoro, dott. Signorello per l'ufficio legislativo, dott. Fucili per la Spes, dott. Schneider per la Discussione, dott. Faldella per l'ufficio Culturale, comm. D'Amato direttore degli uffici». p. 1.

[8] Carlo Danè nasce a La Spezia l'8 ottobre 1927 da Maurico, impiegato delle Regie Poste e da Luigia Zino, casalinga. Rimasto orfano di padre dopo pochi anni milita giovanissimo nell'Azione Cattolica e poi della Democrazia Cristiana di Savona occupandosi, fin dal 1946, dell'attività di propaganda del partito. Nel 1952 entra a far parte della struttura organizzativa diventando funzionario prima ad Agrigento, poi a Savona e, dal 1953, presso la Direzione Nazionale della Spes. Collabora attivamente alla stampa di partito e cura la pubblicazione, per le edizioni Cinque Lune, degli atti di molti dei congressi nazionali della Democrazia Cristiana oltre a numerose altre pubblicazioni tra le quali merita di essere ricordata "Parole e immagini della Democrazia Cristiana" che contiene una raccolta delle immagini dei manifesti realizzati dalla Democrazia Cristiana oltre alla storia dell'attività di propaganda del partito [informazioni fornite da Loredana Zino, congiunta di Carlo Danè].

[9] Il Prof. Aldo Morinello risulta dirigente dell'Ufficio Documentazione, inserito nell'Ufficio Centrale Spes, al 1953. Nell'Archivio Danè è presente una lettera di assunzione di Carlo Danè, datata 16 ottobre 1952, che lo destina a prestare temporaneamente servizio presso il comitato provinciale di Savona. Secondo quanto scrive lo stesso Danè in *Parole e immagini della Democrazia Cristiana*, Roma, tip. Clame, 1985, tale destinazione sarà solo formale in quanto nello stesso periodo lo troviamo prima in Sicilia e poi presso la sede centrale del Partito. *Lettera* di Manfredo de Vita, capo di gabinetto di Guido Gonella a Livia Zanzotto, Roma 20.1.1953. Istituto Luigi Sturzo, Archivio Privato Guido Gonella, Busta 28, fascicolo 7.

[10] Archivio privato di Carlo Danè, *Lettera* di Nando Clemente, dirigente provinciale Spes, a Franco Maria Malfatti, dirigente nazionale Spes. Napoli 7 marzo 1958.

[11] Archivio Privato Carlo Danè, *Ufficio di Documentazione*, dattiloscritto senza data.

[12] Archivio Privato Carlo Danè, *progetto di una agenzia di stampa*, dattiloscritto in due versioni, seconda bozza, [1954].

[13] «In occasione delle feste pasquali partirono da Roma alcuni inviati speciali della SPES con l'incarico di intavolare trattative con i maggiori produttori di uova del Nord, per far applicare su ogni uovo un piccolo rombo con i colori nazionali e l'incitamento a votare. I compagni e le compagne, a loro dispetto, le dovettero acquistare non essendovene altre sulla piazza». Giorgio Tupini, *18 aprile 1948: Metodo e azione della propaganda DC*, in "Cronache sociali", nn. 11-13 del 15 luglio 1948, p. 19.

[14] «Contatti con i centri di documentazione delle ambasciate degli Stati Uniti e di Gran Bretagna le quali ci forniscono regolarmente i loro bollettini riservati di documentazione

ed altro eventuale materiale» (…) Si è deciso, pertanto, di esaminare giornalmente queste fonti classificandole poi in tre rami principali: interni (con 83 voci) esteri (con 93 voci) biografie (con 75 voci) ed anche i documenti riservati. (Particolarmente si è quindi deciso di ritagliare articoli o incollare documenti su fogli di diverso colore inseriti poi in cartelline ugualmente distinte a seconda dei tre rami principali e quindi Interni gialle, Esteri verdi, Documenti riservati rosa). Archivio Privato Carlo Danè, *relazione di lavoro dell'Ufficio Documentazione*, 18.11.1954.

[15] «Da ciò si comprende in maniera evidentissima, che, data la celerità che il lavoro deve per sua stessa natura avere, presupposto di un buon funzionamento della fototeca medesima, sarà la razionalità della classificazione e registrazione delle foto. A tale scopo è stato elaborato il seguente progetto».

- Le fotografie formato 6x9 dovrebbero venire applicate nella scheda *kardex* che ne contiene 8.
- Le schede dovrebbero avere una numerazione dal numero 1 all'infinito ed una divisione letterale che iniziasse dalla lettera a) in ogni scheda non oltrepassando quindi ma la lettera h. Questa divisione letterale dovrebbe avere la funzione di indicare posizioni fisse sulla scheda.
- Su di un apposito indice grafico per soggetti sarà indicata la posizione delle foto: in formato 6x9 sulla scheda *kardex*; in formato grande nella fototeca; in negativo nell'archivio microfilmico". Archivio Privato Carlo Danè, *Progetto di organizzazione per la Fototeca dell'Ufficio Documentazione* [1954].

[16] Archivio Privato Carlo Danè, *appunto sull'ufficio di documentazione*, 11 giugno 1954.

[17] Archivio Privato Carlo Danè, *ufficio di Documentazione* [memo interno].

[18] Archivio Privato Carlo Danè, *appunto sull'ufficio documentazione* [1958] p. 1

[19] Archivio Privato Carlo Danè, *schedario base*, [minuta manoscritta, 1959-1960].

[20] Carlo Danè, *Gli Archivi della DC*, in *Gli Archivi dei Partiti Politici*, Roma, Ministero per i Beni Culturali, 1996, p. 120. «Si è deciso pertanto di esaminare giornalmente queste fonti classificandole poi in tre rami principali: interni (con 83 voci), esteri (con 93 voci), biografie (con 75 voci) ed anche i documenti riservati. Particolarmente si è quindi deciso di ritagliare articoli o incollare documenti su fogli di diverso colore inseriti poi in cartelline ugualmente distinte a seconda dei tre rami principali e quindi Interni gialle – Esterni verdi – Documenti riservati rosa» Archivio Privato Carlo Danè, *Relazione di Lavoro sull'Ufficio Documentazione*, dattiloscritto datato 18 novembre 1954.

[21] Robert Goldschmidt, Paul Otlet, *Sur une forme nouvelle du livre - le livre microphotographique,* L'Institut international de bibliographie, Bulletin, 1907.

[22] Archivio privato di Carlo Danè, *minuta dattiloscritta* con indicazione di inoltro "Forlani" [1955-56].

[23] Archivio Privato di Carlo Danè, *Appunto sull'ufficio documentazione - Piano di lavoro - Personale, dattiloscritto*, [1958] p. 4.

[24] In alcune circolari viene indicato anche come Archivio di Documentazione Politica.

[25] «A tal fine, invece di fare un archivio nuovo, sarà meglio unificare quello dell'Ufficio Documentazione con l'altro della Spes, usando anche per quest'altro materiale la classificazione indicata». Archivio Privato Carlo Danè, *Appunto sull'ufficio documentazione - Piano di lavoro - Archivio*, [1958] p. 3.

[26] La documentazione relativa alla Spes è conservata – unitamente agli archivi della Democrazia Cristiana – nell'Archivio dell'Istituto Sturzo. Ritengo – personalmente – che tale documentazione sia stata fortemente rimaneggiata.

[27] Nell'elenco del personale addetto alla Direzione Nazionale negli anni '50 compaiono sempre uno o più addetti al Protocollo, Archivio, Archivio storico; Dattiloscritto *Uffici di Direzione Centrale*, datato 16 marzo 1951. Istituto Luigi Sturzo, Archivio Privato Guido Gonella, busta 24, fascicolo 8. Cfr. anche Roberto Guarasci, *Archivi e Sistema informativo nella Democrazia Cristiana 1943-1993*, in: Atlanti, vol. 17(2007) n. 1-2, pp. 239-245.

[28] Archivio Privato di Carlo Danè, *Lettera* di Ottorino Momoli a Gian Carlo Arnaud, Dirigente Nazionale Spes, Roma 18 maggio 1971.

[29] Archivio Privato di Carlo Danè, *Lettera* di Carlo Danè a Ottorino Momoli, Roma, 16 luglio 1971.

[30] Archivio Privato Carlo Danè, *Lettera* di Gabriella Fanello Marcucci a Carlo Danè. Roma 17 giugno 1991. Nel 1992 sarà approvato dal Consiglio Nazionale della Democrazia Cristiana l'art. 95 bis dello statuto relativo all'istituzione dell'archivio storico ed alle procedure di versamento degli atti correnti. L'anno prima l'Ufficio di Presidenza della Camera dei Deputati aveva approvato l'inserimento della voce archivio nel modello di bilancio per i partiti politici.

[31] Archivio Privato Carlo Danè, *Appunto sull'ufficio documentazione - Piano di lavoro*, [1958] p. 2.

[32] *Ibidem*.

[33] Archivio Privato di Carlo Danè, *Ufficio Documentazione*, dattiloscritto senza data.

[34] Il sistema informativo sarà realizzato, negli anni 1986-1989, da Olivetti e Informatica Friuli Venezia Giulia. Cfr. Roberto Guarasci, *Archivi e sistema informativo* … cit.

[35] Sistema informativo DC, *Progetto,* raccolta di slides a stampa.

[36] Il Dirigente del Dipartimento Autonomie Locali, Antonio Belfiore, scrivendo a Fiorenzo Maroli il 15 novembre 1986 in ordine proprio alla definizione delle specifiche del sistema operativo ravviserà la necessità di uno specifico rapporto con l'ufficio documentazione. Archivio privato Carlo Danè, *lettera dattiloscritta*, sub data.

Archiwordnet, un *thesaurus* di settore integrato nel *wordnet* della lingua generica: compilazione e applicazioni

ANDREA BOCCO, ENRICA BODRATO, ANTONELLA PERIN

Linguistic resources with domain-specific coverage are crucial for the development of concrete application systems, especially when integrated with domain-independent resources. In this paper, we present our experience in the creation of ArchiWordNet, a specialized WordNet for the architecture and construction domain which is being created according to the WordNet model and integrated with WordNet itself. Problematic issues related to the creation of a domain-specific WordNet and its integration with a general language resource are discussed, and adopted practical solutions are described. Moreover, two examples of using AWN in describing documents are explained.

Keywords: thesaurus – lexical database with domain-specific – architecture, database – archives

1. Introduzione

Il progetto ArchiWordNet è nato dalla collaborazione tra la Fondazione Bruno Kessler (già Istituto Trentino di Cultura) e il Politecnico di Torino finalizzata alla costruzione di un *thesaurus* dei termini architettonici ed edilizi da utilizzare nell'ambito di una banca dati di immagini fotografiche attinenti l'architettura, Still Image Server (SIS), e di fondi archivistici di architettura. Il contributo illustra gli sviluppi di quanto presentato alla Second International WordNet Conference del 2004 presso l'Università di Brno (Repubblica Ceca) [5].

Nella fase di catalogazione, il contenuto di ogni immagine viene descritto con parole, e a ciascuna vengono assegnate parole chiave; perché l'uso delle parole sia sistematico e possa facilitare la ricerca sulle immagini è necessario legare le parole utilizzate tanto dai catalogatori quanto dagli utenti finali attraverso un *thesaurus*. Non essendo disponibile alcun *thesaurus* esaustivo per il dominio dell'architettura, si decise di creare ArchiWordNet, un *thesaurus* bilingue (inglese/italiano), integrato nel WordNet della lingua generica.

In questo testo presentiamo la nostra esperienza nella creazione di ArchiWordNet. Nel paragrafo 2 sono descritte le motivazioni che sottostanno alla scelta di costruire un *thesaurus* "WordNet-like" e le sue caratteristiche; nel paragrafo 3 si affrontano alcuni dei principali problemi connessi alla redazione di una risorsa linguistica di settore inte-

grata in una risorsa per la lingua generica e si descrivono le soluzioni adottate; nel paragrafo 4 sono illustrati due esempi applicativi. Infine, nel paragrafo 5, si accenna agli sviluppi futuri e ad un nuovo campo di applicazione

2. ArchiWordNet: un *thesaurus* WordNet-like

La caratteristica principale di ArchiWordNet è che, sebbene faccia riferimento, il più possibile, a *thesauri* dell'architettura esistenti e ad altre fonti specialistiche, è strutturato secondo il modello WordNet realizzato presso l'Università di Princeton [2] e pienamente integrato al suo interno.

ArchiWordNet si differenzia dai *thesauri* tradizionali tanto dal punto di vista concettuale quanto dal punto di vista delle relazioni [1]. Nei *thesauri*, infatti, i concetti sono solitamente rappresentati con l'uso di un vocabolario controllato nel quale molti sinonimi sono omessi e si fa uso di poche relazioni (iperonimia, iponimia: cioè di tipo ISA) la cui semantica è piuttosto informale. Al contrario, i concetti in WordNet sono rappresentati da insiemi di termini sinonimi propri del linguaggio corrente e le relazioni sono esplicite e omogeneamente codificate in modo da garantire la transitività e l'ereditarietà delle relazioni stesse. Date queste differenze, abbiamo deciso di adottare il modello WordNet per un insieme di ragioni: una struttura più rigorosa permette di ottenere, in fase di ricerca, risposte più efficaci e complete e al contempo rende ArchiWN maggiormente idoneo a finalità didattiche in quanto fornisce contesti concettuali che possono sostenere l'apprendimento. Le gerarchie, ben strutturate, possono essere percorse tanto allo scopo di farsi un'idea generale del dominio dell'architettura, quanto per indagare nel dettaglio un tema specifico.

Le differenze tra ArchiWordNet e i *thesauri* tradizionali non sono riscontrabili solo nella struttura, ma anche nel fatto di integrare pienamente una risorsa linguistica di settore in WordNet. Da un punto di vista teorico WordNet offre al sapere specialistico presente in ArchiWordNet un contesto generale e potenzialmente multilingue; dal punto di vista pratico, la possibilità di un accesso integrato permette maggiore flessibilità nel reperimento delle informazioni. Tuttavia, considerato il grande sforzo e costo in termini di risorse umane, necessario alla costruzione di una risorsa linguistica di questo tipo, non è da trascurare il fatto che l'integrazione si rivela particolarmente utile in quanto le informazioni già presenti nel WordNet generico possono essere utilizzate per la creazione di quello specialistico.

Durante la fase di impostazione di ArchiWordNet abbiamo dovuto confrontarci con le tensioni derivanti dai diversi punti di vista di due discipline: la linguistica computazionale e l'architettura. Più in specifico abbiamo dovuto trovare una mediazione tra la necessità di costruire una risorsa linguistica formalizzata e idonea per le applica-

zioni che trattano il linguaggio naturale e la necessità di costruire uno strumento orientato a soddisfare le necessità pratiche degli "specialisti di dominio". Questa cooperazione interdisciplinare si è rivelata di grande interesse. ArchiWordNet infatti ha il vantaggio di avere una struttura formalizzata e di ereditare dal WordNet generico informazioni linguisticamente strutturate.

Un'ulteriore caratteristica che distingue ArchiWordNet da altre risorse lessicali anch'esse costruite su modello WordNet è che i sinonimi sono ordinati sulla base della loro rappresentatività in rapporto al concetto espresso: dato un gruppo di sinonimi ("synset"), il primo è quello che viene più comunemente utilizzato dagli esperti di settore per identificare quel concetto. Nella creazione di ArchiWordNet abbiamo dovuto affrontare un certo numero di problemi derivanti tanto dall'adozione del modello WordNet, quanto dall'integrazione all'interno di WordNet.

3. Adottare e adattare il modello WordNet

Nella costruzione di ArchiWordNet sono stati adottati due criteri di base. In primo luogo ci siamo riferiti quanto più possibile alle risorse linguistiche specializzate già disponibili e maggiormente riconosciute nel campo dell'architettura e dell'edilizia; in secondo luogo abbiamo utilizzato le informazioni già contenute in WordNet ogniqualvolta è stato possibile.

Per quanto riguarda le risorse linguistiche specialistiche, si è fatto riferimento a diverse fonti per la creazione tanto dei *synset*, quanto delle strutture gerarchiche; tra queste citiamo l'*Art and Architecture thesaurus* (AAT) [3], il *Construction Indexing Manual* del CI|SfB [4], le norme di standardizzazione internazionali e nazionali (ISO, CEN, UNI), il *Lessico per la descrizione delle alterazioni e degradazioni macroscopiche dei materiali lapidei* messo a punto dalla Commissione NORMAL, e altra letteratura di settore, inclusi i dizionari tecnici. Le risorse linguistiche utilizzate sono tanto in lingua inglese quanto in italiano così da trovare le corrispondenze tra le due lingue e poter formare i *synset* bilingui. Tali risorse linguistiche molto spesso non sono risultate compatibili con il modello WordNet o perché non strutturate secondo le relazioni ISA o perché presentano gerarchie miste nelle quali i livelli non sono omogenei e le relazioni tra concetti sono ambigue; al contrario, in WordNet le relazioni sono esplicite e le informazioni sono codificate in modo omogeneo. È stato così necessario riorganizzare queste fonti per renderle compatibili con il modello WordNet. Un esempio, come illustrato in figura 1, è la riorganizzazione della gerarchia del termine "metallo", presente nell'AAT.

Per rendere AAT compatibile con ArchiWordNet abbiamo dovuto interpretare le sue relazioni disambiguando il tipo di relazione che connetteva concetti sopraordinati a

concetti subordinati e decidendo come trattare i "nodi artificiali" intermedi che non sono rilevanti dal punto di vista della gerarchia ISA. Come si può vedere nella figura, i nodi artificiali sono stati eliminati e sono state mantenute le sole gerarchie ISA. I concetti originariamente collegati a *metallo* attraverso una relazione "per forma" sono stati modificati e collocati nella gerarchia ISA appropriata, quindi connessi a *metallo* attraverso la relazione associativa "per materia".

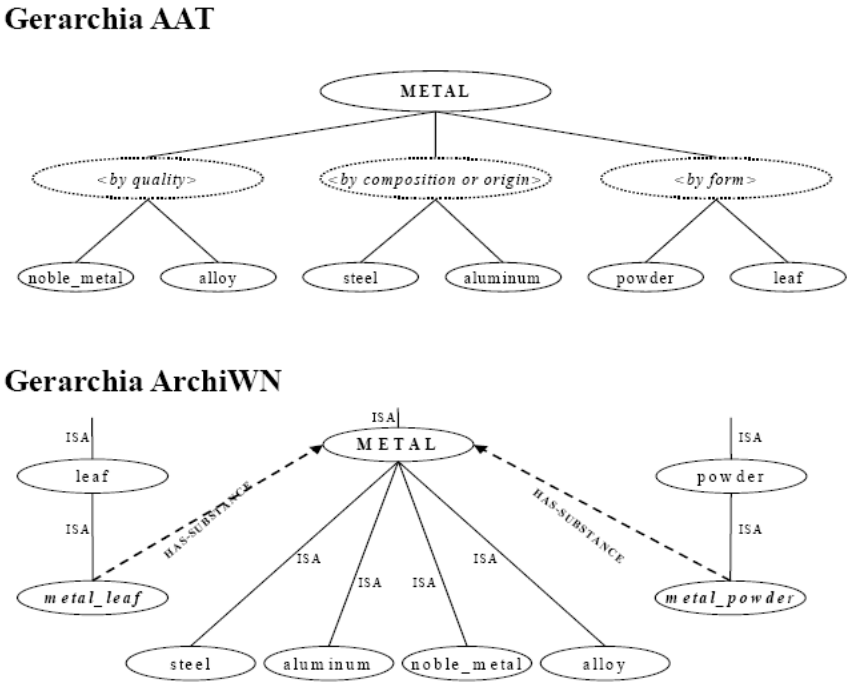


Figura 1 - Riorganizzazione della gerarchia presente in AAT per il termine "metallo" secondo il modello WordNet

La seconda risorsa, ampiamente usata per la creazione di ArchiWordNet, quando non è disponibile una terminologia specifica e ben strutturata, è lo stesso WordNet. I *synset* già presenti in WordNet e considerati appropriati dagli esperti di dominio vengono inclusi in ArchiWordNet. Questa operazione non può sempre essere applicata rigidamente. Infatti, mentre i *synset* di WordNet rappresentano la lingua generica, ArchiWordNet rappresenta un linguaggio di settore ed è possibile che i *synset* e/o le relazioni di WordNet non siano compatibili con l'esigenza di rappresentare il campo

dell'architettura e dell'edilizia. Quando inclusi in ArchiWordNet, i *synset* di WordNet possono subire tre diversi tipi di trasformazione:

a) quando i sinonimi disponibili in WordNet risultano inadeguati per ArchiWordNet, è possibile aggiungere o eliminare sinonimi dal *synset* originario. Ciò può accadere quando termini considerati sinonimi nel linguaggio quotidiano non possono essere considerati tali nel dominio dell'architettura. I concetti di *parete* e *muro*, sinonimi nella lingua generica, non possono esserlo nell'architettura; la *parete*, infatti, è un elemento verticale che delimita uno spazio, il *muro* è un tipo di *parete* formata dalla sovrapposizione ordinata di elementi legati o no tra loro;
b) quando le definizioni del linguaggio generico non sono compatibili con le definizioni tecniche, è possibile modificare la definizione generica già presente nel *synset*. Nel caso dell'esempio precedente la scomposizione del *synset* originario in due nuovi *synset* legati gerarchicamente comporta la modificazione delle definizioni;
c) è possibile sia eliminare sia creare nuove relazioni tra *synset*. Quando viene incluso in ArchiWordNet, infatti, un *synset* può mantenere tutte, alcune o nessuna delle sue relazioni originali presenti in WordNet e questo dipende dal loro essere o meno appropriate al dominio dell'architettura. È inoltre possibile aggiungere nuove relazioni in grado di fornire ulteriori informazioni rilevanti dal punto di vista del linguaggio specifico. Rifacendoci ancora al caso di *parete* e *muro* la modificazione del *synset* originario che li vedeva sinonimi ha comportato la creazione di una relazione ISA tra i due termini e l'aggiunta di relazioni "parte/tutto" per esempio con i concetti di *mattone* o *concio*.

Per integrare ArchiWordNet con WordNet è stata creata una prima lista composta di 5.000 termini, basata sulle fonti specialistiche citate in precedenza e sull'esperienza diretta degli esperti di dominio. La maggior parte di questi termini è stata raggruppata in 13 aree semantiche, come da tabella 1. Queste aree semantiche corrispondono ai principali nodi gerarchici presenti in ArchiWordNet.

A seguito dell'identificazione dei nodi gerarchici di WordNet, all'interno dei quali inserire le gerarchie di ArchiWordNet, le procedure di integrazione richiedono l'inclusione delle gerarchie di ArchiWordNet in WordNet e la gestione delle sovrapposizioni tra termini presenti tanto in ArchiWordNet quanto in WordNet. Quest'ultima esigenza è dovuta al fatto che, contrariamente ad altri domini caratterizzati da una terminologia molto specialistica, l'architettura utilizza un numero significativo di termini presenti nella lingua generica.

Tabella 1

Nodi gerarchici di ArchiWN	Schede lessicali compilate
Architectural styles	
Materials	2.075
Construction products	
Techniques	
Tools	
Components of buildings	530
Single buildings and building complexes	1.250
Physical properties	
Conditions	
Disciplines	
People	
Documents	
Drawings and representations	

Per quanto riguarda il popolamento delle gerarchie, come da tabella 1, sono stati compilati un totale di 3.855 *synset*, per la maggior parte completi dei sinonimi italiani e inglesi e di un'accurata definizione.

4. Esempi di applicazioni

La terminologia già sviluppata è sottoposta a test con l'applicativo Guarini Archivi nell'ambito di progetti di catalogazione di fondi archivistici di architettura, condotti presso il Laboratorio di Storia e Beni Culturali del DICAS, ed è inoltre sperimentata nella BDIS - Banca Dati degli Insediamenti Storici, dal settore della Tecnologia dell'Architettura del DICAS. La possibilità di condurre in parallelo alla redazione del *thesaurus* la sua applicazione in contesti di catalogazione, ci permette, da un lato, di cominciare a utilizzare lo strumento per l'indicizzazione, dall'altro, di verificare, in contesti applicativi, la completezza e l'efficacia delle gerarchie lessicali già completate. Molto spesso è proprio l'applicazione a consentirci di arricchire il patrimonio di termini presenti in ArchiWordNet.

Esempio 1.

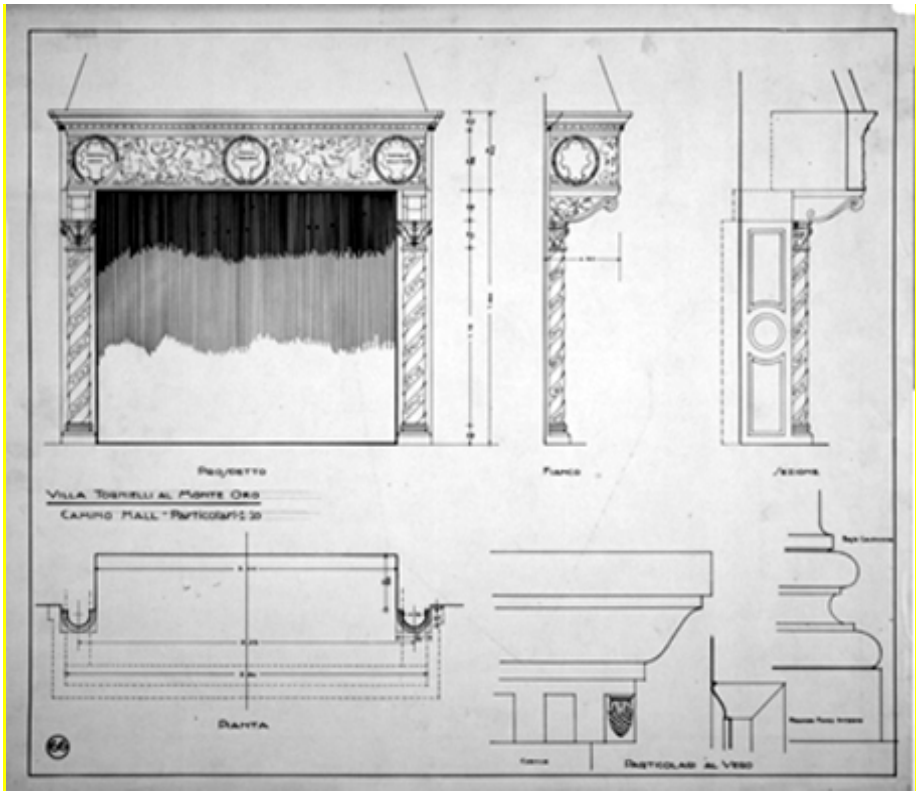


Figura 2 - Politecnico di Torino, Dicas-LSBC, Fondo Melis, Villa Tornielli al Monte Oro, 66

Il disegno proposto è tratto dall'inventario del Fondo Melis de Villa, realizzato con Guarini Archivi. Il Fondo raccoglie i documenti grafici prodotti dall'architetto e urbanista Armando Melis de Villa attivo in Italia e nelle colonie tra il 1925 e il 1956. Il disegno scelto appartiene al progetto per la Villa Tornielli realizzata tra il 1924 e il 1930 ad Armeno (VB).

Attualmente, l'applicativo Guarini Archivi non prevede il collegamento con un *thesaurus*; le sole risorse di controllo terminologico sono i vocabolari e i lemmari. Tra questi, nel caso del paragrafo "Indicizzazione argomento" è prevista la presenza di 2 campi, che attingono ad un lemmario, legati da una relazione gerarchica (Argomento, Argomento specifica). Il catalogatore dunque non potendo accedere direttamente al

thesaurus lo utilizza in parallelo, quale fonte dalla quale estrarre la terminologia che verrà utilizzata e inserita nel lemmario. Così facendo, il termine scelto per l'indicizzazione perde i propri legami tanto di tipo orizzontale, la sinonimia, quanto di tipo verticale, la relazione ISA. Abbiamo tentato di salvaguardare, almeno in parte, quest'ultima, sfruttando la struttura gerarchica della descrizione archivistica; così facendo nelle schede di inventario di livello alto (fondo, serie) vengono selezionati termini generali che nel *thesaurus* stanno ad un livello alto, per scendere poi nel dettaglio della descrizione nelle schede di livello inferiore: l'unità archivistica e l'unità documentaria.

Nel caso del disegno presentato, per quanto riguarda l'indicizzazione "Argomento", nella scheda Fondo è presente tra gli altri il termine *edificio*, nella scheda dell'unità archivistica sono presenti i termini *edificio residenziale* e *villa,* nella scheda dell'unità documentaria i termini che identificano il soggetto rappresentato sul disegno e la sua localizzazione: *androne, camino, colonna, trabeazione, fregio, cornice, mensola, capitello, base, toro, scozia, listello*. La scelta dei termini proposti in questo esempio si attiene a una descrizione architettonica del documento, trascurando quella che potrebbe essere la descrizione di tipo artistico-decorativo da cui sicuramente emergerebbero altri termini.

Nell'affrontare l'indicizzazione di un documento d'archivio, inoltre, ci si deve confrontare con le parole scritte sul documento. È infatti possibile trovare termini storici, termini di uso non comune o ancora in uso con accezione semantica diversa da quella scelta come termine preferito nel *thesaurus*. Nell'esempio che presentiamo compaiono 3 casi come quelli citati: *fianco* utilizzato per indicare un *prospetto laterale*, *hall* per indicare un *androne* e *al vero* per indicare la scala di rappresentazione *1:1*. In questo caso i termini presenti sul disegno vengono ricondotti ai *synset* del *thesaurus* nei quali compaiono come sinonimi; il termine *hall*, per esempio, viene ricondotto nell'indicizzazione al termine preferito *androne*.

La descrizione del documento qui presentato non si limita però alla identificazione del contenuto e dunque alla compilazione del paragrafo Argomento, ma interessa anche le caratteristiche estrinseche che richiedono, negli appositi campi, di accedere ad altri rami del *thesaurus*: l'identificazione del supporto e della tecnica esecutiva (*carta burro, penna a china nera*); della tecnica di rappresentazione (*pianta, prospetto, sezione*); della scala di rappresentazione (*1:10, 1:1*).

Esempio 2

L'immagine proposta fa parte di un fondo di oltre 2.000 fotografie scattate negli ultimi 8 anni durante sopralluoghi in borgate alpine in oltre 60 comuni da docenti, collaboratori e laureandi in Tecnologia dell'Architettura. Oggetto dell'attenzione è stato principalmente l'architettura rurale, messa in relazione sia con l'ambiente di cui è parte

integrante sia con le tecniche esecutive e la cultura materiale. Tale fondo e un'altra quarantina, che raccolgono documenti fotografici su altri temi, afferiscono all'archivio fotografico di cui al § 1; le informazioni relative ai documenti ivi contenuti sono organizzate secondo cartelle (una per ogni edificio) e schede (una per ogni immagine) strutturate appositamente come descritto in [6] e gestite da un database relazionale.
L'analisi del contenuto dell'immagine (campo Descrizione), che rappresenta il *fronte sud* di un *edificio rurale*, privilegia il riconoscimento di tre soluzioni di *involucro*:

- *muratura portante* di forte spessore, di *pietrame* di varia pezzatura, *intonacato* in *malta di calce* con ruolo anche *legante*;
- *struttura lignea a telai*, *tamponata* con piccole pietre;
- struttura lignea a telai, tamponata con *tavole* lignee.

Figura 3 - Politecnico di Torino, DICAS, Eco-Disegni Fotografie Parole dell'Ambiente costruito, casa nella borgata Joussaud (Pragelato, TO), eco_037_002_AC_1446, foto di Diego Cappellazzo,

La Descrizione è intenzionalmente lasciata "aperta" per consentire integrazioni e annotazioni successive, anche da parte di persone diverse. La scelta di consentire al compilatore, nel rispetto della correttezza terminologica disciplinare, l'espressione verbale a testo libero consente schede composte da un limitato numero di campi, ma richiede da

parte del sistema la capacità di riconoscere, in fase di recupero dell'informazione, non solo le flessioni dei termini ma anche le equivalenze logiche (l'uso del participio passato con funzione di aggettivo qualificativo, "tamponato", equivale a dire che si sta parlando di un *tamponamento*; l'uso dell'aggettivo "ligneo" che ci si riferisce a un elemento di *legno*, ecc.). Il tutto col fine di aumentare le possibilità di interrelazione, anche casuale, tra immagini diverse nella cui Descrizione siano stati adoperati termini di medesimo significato, o termini comunque relazionati sematicamente secondo il modello Archi-WordNet.

Un altro aspetto specifico del fondo cui appartiene questa fotografia è la raccolta, dove è stato possibile, di denominazioni locali relative agli elementi costruttivi. Queste non solo rappresentano un patrimonio lessicale minacciato dal rischio di estinzione, ma portano con sé un modo locale di stare al mondo, vale a dire a denominazione dialettale specifica spesso corrisponde modalità specifica di realizzare una parte dell'edificio [7]. Per il momento, i termini dialettali raccolti non sono stati integrati in ArchiWordNet (cosí come, del resto, i termini italiani desueti): attività di grande impegno che potrà essere affrontata sistematicamente solo nel medio-lungo periodo.

5. Conclusioni e sviluppi futuri

Abbiamo qui presentato la nostra esperienza nella creazione di ArchiWordNet e due delle sue possibili applicazioni. L'analisi dei problemi che emergono e gli sviluppi e le integrazioni a cui siamo giunti finora dimostrano tanto che è possibile integrare ArchiWordNet con WordNet, quanto che WordNet può esso stesso essere considerato una risorsa utile per la formazione delle gerarchie di ArchiWordNet.

Per quanto riguarda i prossimi sviluppi del lavoro, proseguiremo nel popolamento delle gerarchie non ancora affrontate e nel perfezionamento di quelle già in corso.

In aggiunta alle applicazioni esemplificate, un importante risultato è rappresentato da un progetto in fase di avvio, approvato e finanziato dalla Regione Piemonte - Direzione Beni culturali, che prevede l'uso di ArchiWordNet nella validazione della banca dati delle schede di censimento del patrimonio architettonico regionale (L.R. 35/1995).

Gruppo di ricerca

Il progetto è condotto sotto la direzione scientifica del prof. Gianfranco Cavaglià (Politecnico di Torino) e del dott. Fabio Pianesi (Fondazione Bruno Kessler) e collaboratori (Emanuele Pianta, Luisa Bentivogli, Christian Girardi).

Note

[1] Clark P., Thompson J., Holmback H., Duncan L. (2000), *Exploiting a thesaurus-Based Semantic Net for Knowledge-Based Search*. In: *Proceedings of AAAI/IAAI 2000*, Austin, Tex;

Fellbaum C. (ed.) (1998), *WordNet: an Electronic Lexical Database*, The MIT Press, Cambridge, Mass;

Petersen T. (1994), *Art and Architecture thesaurus*, Oxford University Press, New York-Oxford <www.getty.edu/research/tools/vocabulary/aat/>;

Ray-Jones A., Clegg, D. (1991), *CI|SfB. Construction Indexing Manual 1976*, RIBA Publications, London;

Bentivogli L., Bocco A., Pianta E., (2004), *ArchiWordNet: Integrating WordNet with Domain-Specific Knowledge*. In: *Proceedings of the Second International WordNet Conference 2004*, Masaryk University, Brno;

Gianfranco Cavaglià (2001), *L'analisi fotografica e la comprensione del costruito. Dalle patologie edilizie al progetto tecnologico*, Celid, Torino;

Andrea Bocco, Gianfranco Cavaglià (2008), *Flessibile come di pietra. Tattiche di sopravvivenza e pratiche di costruzione nei villaggi montani*, Celid, Torino.

La terminologia: un capitale da non sottovalutare

DONATELLA PULITANO

Terminology is the tool for knowledge transfer and therefore a knowledge management tool. A terminological database is a knowledge management system. Thus, terminology is an important asset for an institution or a firm since the knowledge of the institution or firm can only be interesting and productive if it is shared among all actors. The goal of this paper is to present:

- *the role of terminology within Probst's knowledge model*
- *the value of terminology as a factor of production*
- *the added value gained through terminological activities*

Keywords: terminology – knowledge management – terminological database – terminological activities

1. Introduzione

Da un paio di anni "buzz words" come *web semantico*, *ontologie* o *data mining* suggeriscono che siamo in presenza di novità legate alla conoscenza, quando invece si tratta di attività che, in maniera più o meno latente (e cosciente), vengono svolte da più di mezzo secolo.

Ciò che invece costituisce in qualche modo una novità è la consapevolezza di un legame diverso tra terminologia e conoscenza, per cui la terminologia non è più un *mezzo per descrivere* la conoscenza quanto *conoscenza* vera e propria. Una raccolta terminologica deve dunque essere considerata a pieno titolo un *sistema di gestione della conoscenza* e il terminologo svolge un *lavoro della conoscenza* ("knowledge worker"), come è stato sottolineato anche durante il congresso del Deutscher Terminologie-Tag del 2004 intitolato, appunto, *Terminologie und Wissensmanagement* (Mayer, Schmitz *et alii*, 2004).

2. Terminologia e accesso alla documentazione

In un'accezione molto riduttiva la "terminologia" è intesa unicamente come "terminologia normativa, prescrittiva", quindi spesso considerata uno strumento per la descri-

zione di contenuti documentari. Non raramente, quando si parla di *attività terminologiche*, in realtà si intendono attività documentarie che sboccano in particolare nella preparazione di *thesaurus*.

Un *thesaurus* è una lista di parole naturali organizzata, standardizzata e quindi disambigua, per descrivere e formalizzare i contenuti di documenti e per poterli ritrovare facilmente. Il linguaggio documentario è quindi artificiale e costruito, anche se basato sul linguaggio naturale. La terminologia invece, mira a rilevare le unità linguistiche che costituiscono i linguaggi specialistici e i relativi concetti, con finalità descrittive che, in determinati contesti, possono diventare finalità normative. Se è vero che un termine può essere un descrittore in un *thesaurus*, non tutti i descrittori sono termini o dànno luogo a terminologia («Ainsi, les articles sur la néologie abondent, mais on chercherait en vain une terminologie de la néologie» [de Bessé, 2000]).

La terminologia non solo dà accesso alla conoscenza, ma produce conoscenza in quanto denominatrice e descrittrice di concetti.

3. La conoscenza

Tradizionalmente, nella macroeconomia classica, i fattori di produzione sono la *terra*, il *lavoro* e il *capitale*; recentemente, a questi fattori se ne è aggiunto un nuovo, il *capitale intellettuale*, ovvero la *conoscenza*, che ha acquisito un'importanza strategica e finanziaria per le aziende e le istituzioni (in seguito: organizzazioni).

Bacone in effetti, già nel XVI secolo, affermava che «Sapere è Potere», ma il passaggio alla *società della conoscenza* è avvenuto nella seconda metà del XX secolo, quando le organizzazioni si sono rese conto che la conoscenza è un fattore concorrenziale e che nessuno si può permettere di lasciarla inutilizzata o di perderla: «La conoscenza è diventata la vera risorsa chiave a cui imprese, persone, Stati e sistemi locali si appoggiano per produrre valore economico e generare vantaggi competitivi nel confronto con i concorrenti» (Azzariti e Mazzon, 2005).

Solo l'organizzazione che ha una visione completa delle sue conoscenze può usarle in modo ottimale e trarne profitto: «The goal of knowledge management is a practical one: to improve organizational capabilities through better use of the organization's individual and collective knowledge resources» (Probst, 1998).

4. Tipologia della conoscenza e ruolo della terminologia

Le conoscenze di un'organizzazione diventano interessanti e produttive nel momento in cui sono condivise da tutti gli attori; a questo fine sono indispensabili un linguaggio unico ("corporate language") e soprattutto una comprensione unica dei concetti. Chi vende prodotti vende anche la terminologia che li spiega e li descrive. Il trasferimento della conoscenza passa inevitabilmente dalla terminologia. La terminologia ha di conseguenza un ruolo fondamentale all'interno di un'organizzazione.

Idealmente, la divisione responsabile delle attività terminologiche è il *fulcro* dell'organizzazione, il crocevia per il quale transitano informazioni, persone e strumenti. In sintesi, conoscere e sapere ciò che significano i termini, cosa comprendono i concetti, quali relazioni sussistono e quale documentazione è rilevante, significa sapere in modo sostenibile ciò che è importante nell'organizzazione. La terminologia presenta dunque gli stessi vantaggi del *knowledge management*.

Le attività terminologiche intervengono nell'ambito della

- *conoscenza individuale*: la terminologia evita la perdita di conoscenze dovuta all'uscita di un collaboratore dall'organizzazione e incita i singoli a esplicitare la loro conoscenza individuale "affidandola" alla banca dati terminologica;
- *conoscenza esplicita*: la banca dati terminologica è il "contenitore" della conoscenza esplicita dell'organizzazione;
- *conoscenza organizzativa*: tramite le attività terminologiche si memorizza e si trasmette il sapere collettivo dell'organizzazione, la banca dati terminologica è uno strumento pedagogico per i nuovi collaboratori;

e in modo minore della

- *conoscenza tacita*: grazie alla loro esperienza, i terminologi hanno un ruolo importante nella creazione di nuova conoscenza, perché sanno come e dove cercare e conoscono le regole per la creazione di nuovi termini.

5. Il modello della conoscenza di Probst e le attività terminologiche

Gilbert J. B. Probst, professore ordinario all'università di Ginevra nonché "Managing Director and Dean of the Global Leadership Programme" del World Economic Forum ha elaborato un modello della conoscenza basandosi sui lavori di Nonaka.

Analizzando i vari "componenti" (*building blocks*) del modello, si nota subito il parallelismo tra attività della conoscenza e attività terminologiche.

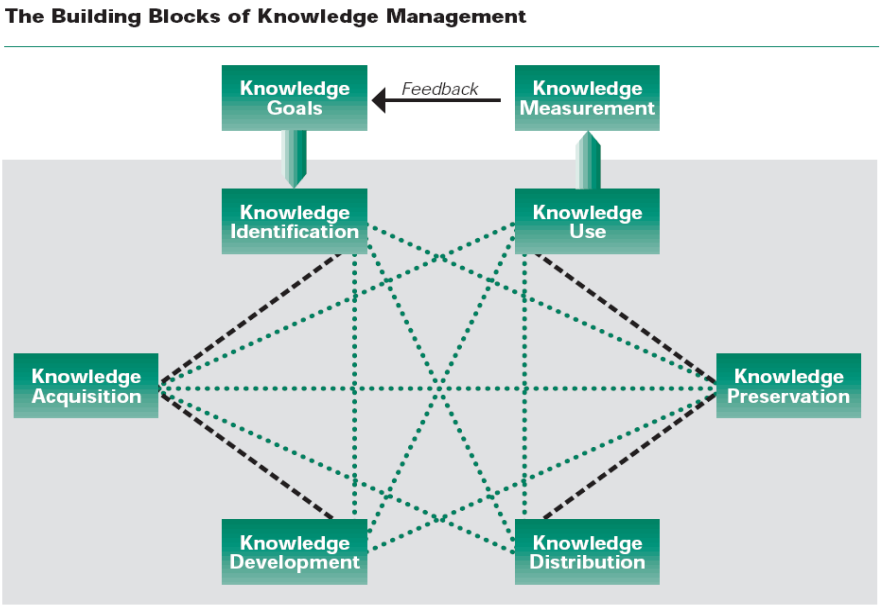


Grafico 1 - Modello della conoscenza di Probst (Probst, 1998)

5.1. Knowledge goals

Gli obiettivi della conoscenza definiscono i livelli di conoscenza importanti nel presente e in futuro per l'organizzazione. Gli obiettivi sono normativi, strategici e operativi e determinano in particolare le politiche aziendali relative alla conoscenza.

Applicato alla terminologia: decisioni inerenti alla *corporate language*, politica linguistica, possibilità di intervento della divisione di terminologia.

5.2. Knowledge identification

L'identificazione della conoscenza serve ad analizzare la conoscenza di cui già è provvista l'organizzazione e soprattutto a evidenziare le lacune. A questo scopo, è utile preparare una "mappatura della conoscenza" (Capitani 2006) che permette di capire quali conoscenze esistono, dove sono, chi ne ha e chi ne avrebbe bisogno, ecc.

Applicato alla terminologia: inventario dei glossari e delle raccolte terminologiche "grigie" disponibili all'interno dell'organizzazione, inventario della documentazione utilizzabile per la terminologia, evidenziazione di campi e settori non ancora descritti nella banca dati terminologica.

5.3. Knowledge acquisition

L'acquisizione della conoscenza si fa soprattutto all'esterno dell'organizzazione, avvalendosi di esperti o acquistando prodotti oppure implicando i potenziali clienti nello sviluppo ("stakeholder knowledge").

Applicato alla terminologia: rilevamento di documentazione, creazione o ampliamento del centro di documentazione, organizzazione di contatti (esperti del settore, revisori, collaboratori free-lance).

5.4. Knowledge development

I dirigenti devono promuovere lo sviluppo e la creazione di conoscenza individuale e collettiva; dall'altro lato, tutti i membri dell'organizzazione devono vedere lo sviluppo come processo innovativo, effettuato per la collettività. Le ricerche e le spiegazioni sono fatte una volta sola e sono poi a disposizione di tutta l'organizzazione, il che permette di usare in modo ottimale tutti i potenziali di conoscenza.

Applicato alla terminologia: arricchimento della banca dati terminologica in particolare in seguito a mandati delle altre divisioni, chiarimenti terminologici nell'àmbito di progetti.

5.5. Knowledge distribution

La distribuzione della conoscenza si effettua tramite un'infrastruttura adeguata che permette a tutti i membri dell'organizzazione di ottenere la conoscenza di cui necessitano nel momento in cui ne hanno bisogno.

Applicato alla terminologia: diffusione della terminologia all'interno dell'organizzazione tramite contatti interpersonali (richieste individuali) e banca dati terminologica.

5.6. Knowledge preservation

L'"immagazzinamento" (MIPA, 2006), il "radicamento" della conoscenza è vitale per evitare ciò che gli esperti chiamano "corporate amnesia", ovvero la perdita della memoria d'impresa (Minguzzi, 2006). Ogni elemento di conoscenza deve essere salvato in un sistema di gestione della conoscenza e quest'ultimo deve essere flessibile e aggiornato in continuazione.

Applicato alla terminologia: banca dati terminologica, il contenitore ('repository') per eccellenza delle informazioni terminologiche.

5.7. Knowledge use

Per poter fruire della conoscenza bisogna accertarsi che ogni membro dell'organizzazione sia in grado di accedere alla conoscenza e che la usi in maniera produttiva, sempre.

Applicato alla terminologia: promozione dell'uso della terminologia all'interno dell'organizzazione, ad esempio nel dipartimento traduzioni o nel dipartimento marketing, controllo qualità per verificare che vengano usati i termini corretti all'interno di tutta l'organizzazione.

5.8. Knowledge measurement

Come per ogni attività produttiva, è necessaria un'attività di controllo per verificare che l'*output* corrisponda all'*input* richiesto. La misurazione e la valutazione della conoscenza sono un processo continuo per verificare se la conoscenza corrisponde alle attese degli utenti e se sono necessarie modifiche.

Applicato alla terminologia: verificare che la terminologia elaborata sia quella richiesta dagli utenti, ottimizzare il tempo di risposta alle richieste SOS.

6. Terminologia e processo di produzione

Come il *knowledge management*, anche le attività terminologiche rientrano in tutti i livelli del processo di produzione: dalla concezione di un nuovo prodotto alla descrizione per gli ingegneri, dalla pubblicità al *marketing* e all'ufficio vendite, dalla preparazione del catalogo alla gestione del magazzino.

Le attività terminologiche devono iniziare, al più tardi, al momento dello sviluppo. È essenziale che i termini scelti o creati in quel momento siano usati in maniera coerente fino a quando il prodotto finale arriva al cliente. Se coesistono terminologie diverse nei *dépliant*, nel manuale utente e sul prodotto stesso (ad es. *ripiano* e *mensola* o *menu di scelta rapida* e *menu contestuale*), non solo si creano malintesi, ma si sprecano anche risorse importanti per la disambiguazione.

Società come Daimler o Geberit hanno implementato le attività terminologiche nella catena di produzione, per cui lo sviluppo di un prodotto viene ritardato se non ne è stata stabilita la corrispondente terminologia.

Secondo i critici, il lavoro terminologico è:

- un mero problema di traduzione,
- un fattore di costo importante,

- complicato e richiede molto tempo,
- un lusso,
- contrario alle esigenze degli autori, che temono di perdere la loro creatività a causa della terminologia "imposta".

A queste argomentazioni si può rispondere che:

a) la terminologia non è legata unicamente alla traduzione e quindi non è un'attività riservata ai contesti multilingui, bensì alla *comunicazione* in genere, in quanto i malintesi terminologici si creano anche quando si parla la *medesima* lingua;
b) ciò che costa di più è la *mancanza* di terminologia: D. Gust ha calcolato che la mancanza di terminologia varia tra il 5 e il 15% dei costi complessivi di un progetto (Gust, 2006) mentre, per un caso concreto, A. Ottmann ha mostrato che il lavoro supplementare generato dalla mancanza o dall'errata terminologia era di 5,5 mesi/uomo - per *una sola lingua* (Ottmann, 2005);
c) se si "fa" terminologia in modo corretto, non è molto complicato e non richiede molto tempo;
d) gli autori in genere si rendono conto che la creatività è legata alla redazione e non ai termini, che invece devono essere univoci.

7. Tipologia del lavoro terminologico e impatto sulla conoscenza

La terminologia permette di usare al meglio il *potenziale di conoscenze*, a tutti i livelli di lavoro terminologico; ogni tipo di lavoro terminologico è all'origine di un valore aggiunto strategico per l'organizzazione.

7.1. Lavoro terminologico tematico ⇔ lavoro terminologico ad hoc

7.1.1. Lavoro terminologico tematico

Il lavoro terminologico tematico consiste nell'elaborazione completa ed esauriente del vocabolario di un dato settore. Un caso simile è dato dal lavoro terminologico sistematico in cui si effettua lo spoglio di un *corpus* ben preciso.

Valore aggiunto strategico: l'organizzazione può disporre di una terminologia consolidata e coerente che permette di accrescere la qualità dei testi e di evitare ambiguità onerose e perdite di tempo per la ricerca dei termini "corretti".

7.1.2. Lavoro terminologico ad hoc

Il lavoro terminologico *ad hoc* serve a cercare, attestare ed eventualmente creare un termine per un caso particolare. È il lavoro in genere effettuato dai traduttori mentre

traducono o dai redattori mentre redigono; può essere facilmente delegato alla divisione che si occupa di terminologia (cosiddetto *servizio SOS* o *RSVP*).

Valore aggiunto strategico: l'organizzazione non deve ricorrere a risorse esterne, ma può ottenere un aiuto interno rapido e privo di burocrazia, guadagnando quindi tempo e personale per gli altri processi.

7.2. Lavoro terminologico descrittivo ⇔ lavoro terminologico normativo

7.2.1. Lavoro terminologico descrittivo

Il lavoro terminologico descrittivo permette di costituire l'"inventario" dei termini usati all'interno dell'organizzazione e di definire i concetti inerenti.

Valore aggiunto strategico: con la definizione dei concetti si accresce la sicurezza giuridica dei testi redatti all'interno dell'organizzazione. Grazie al rilevamento di eventuali sinonimi aumenta la comprensibilità dei testi.

7.2.2. Lavoro terminologico normativo

Il lavoro terminologico normativo (o prescrittivo) ha come scopo l'uniformazione del vocabolario usato all'interno dell'organizzazione. Si evitano così denominazioni molteplici per lo stesso pezzo di ricambio o la stessa sigla per due unità aziendali diverse.

Valore aggiunto strategico: ridurre le ambiguità e i malintesi.

7.3. Lavoro terminologico a posteriori ⇔ *lavoro terminologico* a priori

7.3.1. Lavoro terminologico a posteriori

Il lavoro terminologico *a posteriori* consiste nell'elaborare e quindi preservare la terminologia già esistente all'interno dell'organizzazione. Questo lavoro consente in particolare di fissare i risultati di un lavoro concettuale, in quanto le riflessioni a proposito dei concetti e delle loro denominazioni sono già state svolte. Il vantaggio in questo caso è che la documentazione già esiste ed è a disposizione per lo spoglio.

Valore aggiunto strategico: il riutilizzo ("reusability") e la recuperabilità ("retrievability") della terminologia in raccolte terminologiche; in particolare, in banche dati terminologiche.

7.3.2. Lavoro terminologico a priori

Il lavoro terminologico *a priori* consiste nella preparazione di raccolte terminologiche a monte di nuove realtà.

Valore aggiunto strategico: preparando il "materiale linguistico" in anticipo è possibi-

le evitare i malintesi a garanzia di una buona comunicazione grazie all'introduzione e alla diffusione tempestiva di nuovi termini.

8. Conclusione

L'istituzione di attività terminologiche in un'organizzazione assicura che tutte le conoscenze – tacite, individuali, esplicite e organizzative – siano trasferite in maniera univoca e con il consenso di tutti gli attori coinvolti. A questo fine, le attività terminologiche devono essere parte integrante della catena della conoscenza ed essere presenti a tutti i livelli dell'organizzazione, mentre i risultati del lavoro terminologico devono essere di facile accesso, controllabili, di qualità elevata e sempre aggiornati.

Gli effetti positivi che ne conseguono sono l'univocità, la consistenza, la comprensione e la sicurezza giuridica.

La terminologia è capitale intellettuale e come tale deve essere trattato all'interno di un'organizzazione.

Bibliografia

Azzariti Ferdinando, Mazzon Paolo (2005). *Il valore della conoscenza - Teoria e pratica del knowledge management prossimo e venturo*. Etas, Milano.

Capitani Paola, (2006), *Il Knowledge Management - Strumento di orientamento e formazione per la scuola, l'università, la ricerca, il pubblico impiego, l'azienda*. FrancoAngeli, Milano.

de Bessé Bruno, (2000), "*Le domaine*", in: H. Béjoint/Ph. Thoiron (dir.), *Le sens en terminologie*, Presses universitaires de Lyon, Lione.

Gust Dieter, (2006), "Wirtschaftliche Terminologiearbeit in der Technischen Dokumentation - denn Verzicht auf Terminologie kommt Sie teuer zu stehen" in: *eDITion* 2/2006, 16-20.

Mayer Felix, Schmitz Klaus-Dirk, Zeumer Jutta (Hrsg.), (2004), *Deutscher Terminologie-Tag e.V. Terminologie und Wissensmanagement: Akten des Symposions*. Köln: 26.-27. März 2004.

Minguzzi Paolo, (2006). La gestione della conoscenza nelle organizzazioni - Il contributo della memoria d'impresa. Franco Angeli s.r.l., Milano.

MIPA, Consorzio per lo sviluppo delle metodologie e delle innovazioni nelle pubbliche amministrazioni (2006). Capitale intellettuale e amministrazioni pubbliche - Riferimenti metodologici e studi di caso per la gestione e la valorizzazione. <www.istat.it/dati/catalogo/20060907_01/mipa_vol_12.pdf>; letto il 26 maggio 2008.

Ottmann Angelika, (2005), "Ist Terminologiearbeit wirtschaftlich? Unterlassung von Terminologiearbeit bei der Softwareentwicklung als Kostenfaktor - ein Erfahrungsbericht" in: eDITion 1/2005, 12-13.

Probst Gilbert J.B., (1998), "Practical Knowledge Management: A Model That Works" in: Prism 2/1998, 17-29. <know.unige.ch/publications/Prismartikel.PDF>, letto il 26 maggio 2008.

La memoria in rete: parole per ricordare

MADEL CRASTA

This contribution emphasizes the potential of the Web as a tool for the dissemination and preservation of historical and cultural heritage in historical archives. The author affirms, however, that the potential value added in terms of knowledge will be achieved only through the development of databases and search engines that allow greater access to knowledge heritage in the "deep" web. Stressing the importance of indexing systems for the recovery of information, she presents the Thesaurus "Le parole del Novecento", produced in conjunction with the "Direzione Generale degli Archivi del Mibac"

Keywords: Historical Archives – indexing system, web

Come rappresentante di un Consorzio di Istituti e fondazioni culturali, esprimo l'impegno dedicato dagli anni '90 alla emersione del patrimonio conoscitivo stratificato nelle istituzioni culturali, la cui specificità nell'epoca delle grandi banche dati catalografiche, risulta poco leggibile immersa com'è nell'immenso serbatoio del web. Questa vaghezza di contorni costituisce un freno indiscutibile alla circolazione dei contenuti la cui conoscenza, come oggetti, testi, immagini e suoni frutto della ricerca e della elaborazione culturale, deve essere largamente condivisa per poter svolgere una funzione creativa nel vivo della produzione culturale.

Dal nostro punto di vista, l'aspetto cognitivo è fortemente legato alle condizioni comunicative e allo sviluppo di un linguaggio efficace per la trasmissione della memoria attraverso logiche e strumenti adeguati allo scenario multimediale e digitale. Conosciamo tutti il ruolo di Internet come "gate" di accesso ad una sperimentata stratificazione di contenuti digitali; ciò di cui si è meno consapevoli è che noi produttori di contenuti per le banche dati e gli ambienti digitali nel web, pensiamo per forza di cose con logiche e modelli concettuali formatisi in un secolare processo di elaborazione di linguaggi e metodologie descrittive applicati agli oggetti della memoria. I contenuti immateriali sono trasportati dalla fisicità degli oggetti nelle teche destinate alla conservazione, separati da tipologie e classificazioni anche se concettualmente affini, contigui o complementari.

Il sistema delle teche – biblioteche, musei, archivi, fototeche… – ha certamente garantito la conservazione e la trasmissione delle memorie ma ha anche separato rigida-

mente oggetti e contenuti, così come separati e distanti sono stati i saperi specialistici accumulati intorno ad essi.

L'esplosione dell'ICT, della multimedialità e del web hanno reso disponibili nuove modalità di conoscenza e favorito la connessione semantica a prescindere dai luoghi e dalle mura di conservazione. È possibile oggi aggregare dati, testi, immagini, suoni di diversa provenienza, ricostruire negli ambienti digitali i nessi di significato che collegano i documenti e ridisegnare il contesto che li ha espressi. Tuttavia il potenziale valore aggiunto in termini di conoscenza non si realizzerà se non attraverso l'evoluzione delle basi dati e dei motori di ricerca verso una più ampia accessibilità al patrimonio stratificato nel "deep" web. Le parole, i termini, i metodi di indicizzazione continuano a sembrare secondari rispetto al prodotto, alla quantità e qualità dei dati, mentre costituiscono la garanzia che le risorse digitali producano quel risultato di conoscenza per cui vengono realizzate impegnando considerevoli risorse economiche e professionali.

La scelta delle parole ha a che fare con la chiarezza sui risultati attesi, sui destinatari dell'informazione e sulla "community" che si vuol creare, sul grado di specializzazione dei contenuti e della terminologia. In sintesi, se non si vuole evocare il messaggio nella bottiglia lanciata nell'oceano, senza sapere dove, quando e chi lo leggerà, è necessario focalizzare l'attenzione sulla progettualità culturale, sui contenuti di beni e documenti e della loro efficace comunicazione oltre che sulle metodologie catalografiche. La recente convenzione Unesco che riconosce come bene culturale patrimonio dell'Umanità i significati immateriali delle tradizioni e della memoria, rende ancora più significativo l'impegno a far emergere dagli archivi digitali la ricchezza dei contenuti e della rete semantica che li collega. Ciò significa nel concreto l'introduzione sistematica di descrittori e di *thesauri* anche in quegli ambiti come gli archivi storci, in cui si è finora cercato di farlo per tante e condivisibili ragioni.

Un esempio concreto in questo senso è la rete <www.archividelnovecento.it> cui partecipano oggi 70 istituzioni con i loro archivi storici. Poiché il sistema *software* consente la ricerca integrata nell'intero archivio digitale anche attraverso termini contenuti nelle schede, si è affermata progressivamente la necessità, ma anche l'opportunità, di rafforzare la ricerca terminologica per valorizzare tutti gli incroci fra i documenti e far emergere come un tessuto coerente la trama della vita culturale del Novecento. È nato così il lavoro di redazione del *Thesaurus* "le parole del Novecento", realizzato d'intesa con la Direzione Generale degli Archivi del Mibac. Il fondamento teorico del progetto si è basato sull'attento esame di esperienze analoghe già realizzate e sull'adattamento degli standard operanti nell'ambito biblioteconomico in vista delle esigenze proprie del settore archivistico. È stato quindi sviluppato un prototipo di *thesaurus*, nato dall'interconnessione tra riflessione metodologica, confronto con le esperienze nel campo archivistico e indicizzazione applicata alla banca dati di Archivi del Novecento. Sulla base di questo risultato, si è aperta la fase di sperimentazione, che

fungerà da test sulla funzionalità del *thesaurus*, sia nella fase di indicizzazione che in quella di ricerca, oltre a essere finalizzata a ottenere proposte per nuclei tematici non presenti attualmente nel vocabolario controllato. Il Coordinamento di Archivi del Novecento sta inoltre progettando un'interfaccia di interrogazione on line del thesaurus per integrare le attuali modalità di ricerca del sito.

Concluderei ricordando a tutti noi che la scelta e il valore delle parole non sono freddi tecnicismi per addetti ai lavori, ma materia calda che tocca la possibilità stessa di ricordare e riconoscersi.

Terminologia dalla parte del ricevente

CLAUDIO GIOVANARDI

This paper presents the results of a survey conducted by the author on the spread and the level of knowledge of anglicisms in contemporary Italian. This contribution analyses xenisms, in the perspective of the final user in daily communication.

Parole chiave: Anglicisms – Specialist terminology – Tecnicisms

Desidero riallacciarmi idealmente alle conclusioni del mio intervento al Convegno Ass.I.Term. dello scorso anno, nelle quali rivendicavo la necessità di far uscire il dibattito sugli anglicismi in italiano dalle aule universitarie e dalle colonne dei giornali per portarlo metaforicamente nelle piazze, ovvero per raccogliere le determinanti indicazioni da parte di utenti reali, di coloro che realmente, nella vita di tutti i giorni, fanno i conti con la comprensione dei termini (stranieri o meno) che assediano la nostra quotidianità [1].

Nel titolo del mio contributo risuona volutamente quello di un convegno della Società di Linguistica Italiana i cui atti videro la luce giusto vent'anni fa: *Dalla parte del ricevente: percezione, comprensione, interpretazione*, in cui, per la prima e unica volta, a mia conoscenza, ci si poneva dall'altra parte della barricata, dalla parte di chi è il terminale e non il punto d'avvio del processo comunicativo [2]. Ho sfogliato l'indice di quel volume e mi pare che scarsa attenzione sia stata data allora al problema della ricezione della terminologia; mi piace però ricordare la parziale eccezione rappresentata dal contributo del sottoscritto insieme con Maurizio Dardano e Adriana Pelo, intitolato *Per un'analisi del discorso divulgativo: accertamento e studio della comprensione*, nel quale, analizzando i vari procedimenti di riformulazione "che accompagnano i termini tecnici e i forestierismi non integrati presenti, a diversi livelli, in alcuni settori della nostra stampa" [3] si coglievano due aspetti interessanti: da un lato la necessità di individuare un drappello di informatori cui chiedere conto dei meccanismi della comprensione e delle strategie di riformulazione; dall'altro la doverosa attenzione verso il fenomeno montante (già vent'anni or sono) degli anglicismi non adattati.

Da allora molto è stato fatto nel campo dello studio delle terminologie specialistiche. E molto di quel che è stato fatto lo si deve all'impulso di un'associazione come Ass.I.Term. in particolare sotto la presidenza di Giovanni Adamo, prima, e di Riccardo Gualdo, poi. Buona testimonianza di quel che dico è il volume curato da Maria Teresa

Zanola, *Terminologie specialistiche e tipologie testuali*, che raccoglie gli atti del convegno di Milano del 2006 [4]. In quel volume sono ospitati contributi di notevole valore, accomunati, seppure secondo prospettive diverse, dall'attenzione rivolta all'impatto tra i termini specialistici e fasce diverse di utenti, culturalmente più o meno attrezzati per l'interpretazione degli stessi (penso in particolare ai saggi di Serianni, Gualdo e Zanola).

La crescente importanza dei linguaggi tecnici e scientifici nella cultura linguistica odierna è ormai un fatto acclarato. Maurizio Dardano, uno dei massimi studiosi dei linguaggi scientifici in Italia, considera "la penetrazione di vocaboli ed espressioni specialistiche nei livelli medi della lingua" uno dei fenomeni più vistosi dell'italiano contemporaneo [5]. Tutte le raccolte di neologismi e gli aggiornamenti lessicografici ci dicono che i tecnicismi rappresentano una fetta cospicua delle parole nuove del nostro lessico [6]. Afferma giustamente Gualdo: "Il rilievo e il prestigio del lessico tecnico-scientifico caratterizzano la lingua d'oggi; e si prevede che aumentino nell'era del digitale, in seguito alla tematizzazione e alla specializzazione dei canali di comunicazione" [7].

Gli studiosi si sono spesso interrogati sulle dinamiche sociolinguistiche che regolano l'enorme successo dei forestierismi nei media tradizionali ed elettronici. Talvolta il favore si fonda su presupposizioni almeno in parte stereotipate: l'alone di modernità e di espressività, di semplicità, di grande incisività che circonda le parole straniere e inglesi in particolare [8]. Da altre parti si è però insistito sulle ripercussioni negative che tale moda comporta nella comunicazione pubblica, laddove un elementare diritto di trasparenza e di democrazia linguistica imporrebbe il rifiuto degli xenismi all'interno di discorsi e testi destinati alla cosiddetta "società civile" [9].

Dardano ha più volte ricordato le tre modalità fondamentali nella creazione e nell'incremento dei vocabolari tecnico-scientifici: a) rideterminazione semantica di un vocabolo della lingua comune o di un termine già esistente; b) ricorso al prestito linguistico sia da lingue moderne sia da lingue classiche; c) uso dei vari procedimenti di formazione delle parole [10]. Ma oltre alle considerazioni di linguistica "interna" nel caso dei tecnicismi sono altrettanto importanti quelle di linguistica "esterna", che attengono cioè alla percezione e al prestigio sociale dei termini tecnico-scientifici. Da tempo si è notato che i sottocodici hanno riempito in qualche misura un vuoto storico che si è creato nel repertorio dell'italiano contemporaneo, vale a dire quello lasciato dal modello della lingua letteraria, per secoli varietà di prestigio e norma indiscussa della lingua scritta, ma oggi in rovinoso declino [11]. Il risultato di tale processo è che indubbiamente in Italia "tecnicismo è bello", e in virtù di tale slogan sembrano saldarsi in un corpo solo le pessime abitudini degli esperti e la voglia di sapere "superiore" dell'utente comune.

Mettendo a confronto i cosiddetti "bugiardini", ovvero i foglietti illustrativi dei farmaci italiani con i corrispettivi francesi e spagnoli, Luca Serianni, altro illustre studioso dei tecnoletti scientifici, ha dimostrato come lo stile comunicativo italiano sia inutilmente appesantito da numerosi tecnicismi, non sempre trasparenti e soprattutto non sempre necessari, al contrario dei foglietti francesi e spagnoli che si contraddistinguono per il tono medio, discorsivo, volutamente privo di asperità tecnicistiche [12]. Ma ciò che va a mio avviso aggiunto è che probabilmente tale stile comunicativo è in qualche misura gradito anche all'utente inesperto, il quale, seppure non capisce o capisce solo in parte il messaggio del bugiardino, vi respira tuttavia un'aria impettita e seriosa, tutto sommato rassicurante; ne coglie l'adeguatezza scientifica (vera o fittizia), ne avverte l'appartenenza ai registri più elevati della lingua.

Nella presente occasione darò conto di talune indagini che hanno investito alcuni termini tecnici sotto forma di anglicismi. Ciò è dovuto da un lato alla mia consuetudine col fenomeno della penetrazione di anglicismi in italiano, di cui ho in parte dato conto già lo scorso anno a Bertinoro [13], dall'altro al fatto che indubbiamente gli studi più recenti insistono sul fenomeno degli xenismi all'interno delle terminologie come ulteriore elemento di complicazione per quanto riguarda l'aspetto già di per sé complesso della ricezione e della interpretazione. Il rapporto inglese-italiano nelle terminologie tecnico-scientifiche non è uniforme, ma risponde a diversi livelli di interazione o di mutua esclusione. Ne ha scritto recentemente Maria Teresa Zanola a proposito del lessico finanziario, all'interno del quale la studiosa ha individuato tre diversi livelli di "convivenza" tra gli anglicismi e i corrispondenti italiani [14]. Il primo livello potremmo chiamarlo di convivenza pacifica, laddove i due termini coesistono più o meno alla pari; il secondo vede la prevalenza della soluzione inglese su quella italiana, senza che ciò abbia una motivazione evidente: si tratta del comportamento prevalente nella stampa che tende a privilegiare l'anglicismo; il terzo livello vede il dominio quasi assoluto dell'anglicismo, dovuto all'assenza o alla scarsa praticabilità dell'alternativa italiana [15].

In una situazione così poco stabilizzata, in cui gli equilibri tra la parola straniera e la corrispondente italiana appaiono ancora non definiti, è ancora più importante l'atteggiamento del parlante comune, che potrebbe in qualche misura essere l'arbitro che assegna il successo all'una o all'altra risorsa. Grazie all'aiuto fornitomi da alcune tesi di laurea da me dirette [16] potuto ragionare sui risultati di un questionario somministrato a un campione di informatori (più di cento, anche se non tutti interrogati sull'intero *corpus* di lemmi) di età compresa tra i 16 e i 55 anni, tutti di estrazione sociale medio-alta e in possesso di un livello di studi (in taluni casi, ovviamente, in corso di espletamento) anch'esso medio-alto. In particolare richiamo l'attenzione su un sotto-

campione composto da 20 informatori di età compresa tra i 30 e i 55 anni (10 donne e 10 uomini) impiegati presso il *World Food Programme*, un'agenzia delle Nazioni Unite, tutti parlanti italiani in grado di usare l'inglese come lingua di lavoro.

Riprendendo e adattando il sondaggio sulla diffusione degli anglicismi condotto alcuni anni or sono da Dardano e dai suoi allievi [17] e tenendo presenti i parametri e le voci analizzate in Giovanardi-Gualdo (2003), sono state predisposte una serie di interviste, fondate su questionari elaborati appositamente, nelle quali si sono testate le seguenti abilità e conoscenze da parte degli informatori (dopo averne fotografato la fisionomia sociolinguistica e il grado di conoscenza delle lingue straniere):

a) capacità di cogliere la collocazione diafasica del forestierismo, superando la secca contrapposizione tra lingua comune e linguaggio settoriale;
b) effettiva conoscenza e uso del forestierismo, stimolando al tempo stesso l'indicazione di un corrispondente italiano o, quanto meno, chiedendo una breve parafrasi per valutare quanto e in che modo il forestierismo sia penetrato realmente nell'uso;
c) possibilità dell'informatore di scegliere tra il forestierismo e la sua eventuale alternativa italiana all'interno di un contesto testuale sufficientemente ampio; in tal modo si eviterebbe un giudizio *in vacuo* consentendo di verificare all'interno di una sequenza discorsiva;
d) capacità metalinguistica di giudizio sulle caratteristiche del forestierismo, in particolare per quanto attiene alla comprensibilità, all'utilità, alla novità, alla facilità di pronuncia e di grafia;
e) capacità di produrre autonomamente enunciati nei quali, a giudizio dell'informatore, l'impiego del forestierismo è inevitabile; in questo caso si punta a stimolare un'abilità linguistica attiva da parte dell'informatore per evitare un atteggiamento puramente passivo nei confronti del fenomeno dei forestierismi.

Un'analisi ponderata dei risultati di test di tal genere, consente di conoscere l'"altra faccia della medaglia", cioè le reazioni degli utenti di fronte al fenomeno dei tecnicismi (*sub specie anglica*), la profondità di penetrazione, la predilezione (o la reiezione) per determinate categorie di vocaboli ed espressioni. Tale criterio consente, tra le altre cose, di misurare quanto sia grande la distanza tra l'uso scritto e quello parlato in materia di forestierismi e quanta distanza passi tra la comunicazione degli e tra gli esperti e quella tra "comuni mortali".

Vengo a illustrare il questionario e preciso che esso ha riguardato i 150 anglicismi raccolti in Giovanardi-Gualdo (2003). In questa occasione ci si soffermerà prevalen-

temente su quelli più implicati nell'àmbito tecnico-scientifico. Il questionario si apre con una richiesta di autovalutazione del proprio livello di conoscenza dell'inglese. A parte il sottocampione degli impiegati alle Nazioni Unite, il livello di conoscenza presso le giovani generazioni risulta modesto: solo un'esigua minoranza dichiara una conoscenza buona o ottima, mentre la stragrande maggioranza si inserisce nel livello medio o scarso.

Il primo quesito chiedeva agli informatori di indicare per ciascun anglicismo l'àmbito di riferimento potendo scegliere tra le seguenti opzioni: *politico*, *letterario*, *nuove tecnologie*, *sport*, *quotidiano*, *medico-psicologico*, *finanziario*. I risultati di questo primo quesito sono molto interessanti, non tanto per le inevitabili oscillazioni che si trovano nelle risposte degli informatori, quanto perché da tali risposte emerge la tendenza del parlante a giudicare le parole in base alle proprie competenze individuali più che sulla scorta di considerazioni oggettive. Così, ad esempio, gli informatori che per motivi di lavoro o di studio hanno a che fare con l'informatica collocano termini come *cookie*, *banner*, *blog* nel linguaggio quotidiano, anziché, come ci saremmo aspettati, in quello delle nuove tecnologie. Talvolta la diversa percezione è legata a motivi di "genere": per le donne *push up*, ovvero il reggiseno volumizzante, fa parte della lingua quotidiana, mentre tra gli uomini alcuni non sanno rispondere, altri pensano all'àmbito medico-psicologico e altri ancora alle nuove tecnologie. La scelta è fatta in base alle suggestioni che il nome evoca. Il *cordless*, è dai più inserito nel linguaggio tecnologico, benché sia ormai un utensile comunissimo, forse perché è sentito come uno strumento ad alto grado di tecnologia; al contrario *car sharing*, l'auto condivisa, finisce nel linguaggio quotidiano perché si pensa all'uso che se ne fa, più che alle qualità della parola. Naturalmente più la parola risulta oscura, più la collocazione diafasica si fa incerta. Non sono poche le voci che sono state collocate in tutte le caselle proposte, compresa quella del "non so". In alcuni casi ciò non desta meraviglia, per esempio nel caso di *folder* o di *low impact*; ma quando troviamo un termine come *exit poll* spalmato sull'intero arco delle possibili collocazioni (anche se l'attribuzione al linguaggio politico è prevalente) deve sorgerci qualche dubbio sul grado di "digeribilità" anche di anglicismi semplici, se non addirittura banali.

Il secondo quesito è il classico test di conoscenza ("conosco e uso", "conosco ma non uso", "non conosco"), ma corredato della richiesta di indicare per ciascun anglicismo un corrispondente italiano oppure una breve parafrasi. In linea di massima possiamo dire che gli anglicismi più noti sono quelli che appartengono alla lingua quotidiana e al settore delle nuove tecnologie (fra questi ultimi ricordiamo *account*, *chat line*, *dolby surround*, *home page*), mentre i meno noti risultano quelli del linguaggio politico ed economico, come *advisor*, *bipartisan*, *job sharing*, *moral suasion*. Molto oscillanti i risul-

tati per quel che riguarda i termini del linguaggio sportivo e del linguaggio massmediatico. Ma al di là delle dichiarazioni di conoscenza, il nostro quesito consente di riflettere su due aspetti interessanti. Innanzi tutto non vi è una corrispondenza chiara e univoca fra le indicazioni di appartenenza diafasica (quesito 1) e la conoscenza del vocabolo. Un vocabolo prevalentemente assegnato alla lingua comune risulta poi maggioritariamente sconosciuto; viceversa un vocabolo di incerta collocazione diafasica risulta poi noto ai più. È il caso di *ambient music* e di *badge*, che pur identificati come parole della lingua comune, risultano sconosciuti alla maggior parte degli informatori. In secondo luogo le corrispondenze italiane indicate dagli utenti rivelano spesso una notevole approssimazione. Per *account* troviamo ad esempio "spazio dedicato ad un utente", "Indirizzo postale per il computer", "nome di accesso a internet", "connesso al computer e alla tecnologia", "indirizzo", "conto". Le varie definizioni risultano piuttosto imprecise e decisamente poco sintetiche, oppure, se sintetiche, troppo vaghe. Tuttavia gli utenti dichiarano di conoscere e utilizzare la parola *account*. Ciò porta a ipotizzare che gli utenti fanno un uso "meccanico" di molti anglicismi, assumendoli in una sorta di pacchetto "tutto compreso", di cui si apprezza la funzionalità e, perché no, l'essere di moda, ma di cui si ignora l'esatto significato. Una pragmatica forte, insomma, contrapposta a una semantica molto debole.

Il terzo quesito chiedeva, per alcuni anglicismi selezionati, di elencare tutte le espressioni (i sintagmi polirematici) conosciute in cui l'anglicismo ricorre. In questo caso l'obiettivo era ovviamente quello di stimolare la competenza attiva degli informatori, e inoltre di valutare l'effettiva diffusione dei composti sia totalmente inglesi sia ibridi italo-inglesi (del tipo *donna day*, *caldo record*, *uomo sandwich*). Ebbene i risultati sono come al solito oscillanti. Quel che colpisce è il forte influsso del mondo televisivo e della pubblicità. Penso all'indicazione di *Happy days*, titolo della famosa serie televisiva, oppure di *body lotion* e *body cream* locuzioni tipiche dei prodotti di bellezza. Molto poche sono le polirematiche formate a partire da due termini del linguaggio burocratico come *customer* e *format*, a riprova della loro scarsa diffusione nella lingua comune. È comunque interessante che tra le pochissime locuzioni indicate con *format* abbiamo appunto *format TV*, a riprova del forte potere modellizzante che la televisione esercita sulla lingua. Quanto a *customer*, la scarsa dimestichezza con il vocabolo sorprende soprattutto nelle giovani generazioni universitarie, dal momento che tale vocabolo vive i suoi momenti più fulgidi proprio all'interno della comunicazione aziendale universitaria, la quale, per nostra fortuna, è evidentemente meno convincente di quella televisiva.

Il quarto quesito è particolarmente interessante perché consiste nel porre in alternativa l'anglicismo e un corrispondente italiano non astrattamente, ma all'interno di un effettivo contesto frastico. Per ciascuna frase prodotta si chiede all'informatore di sce-

gliere tra l'anglicismo e l'italianismo sulla base dei criteri di appropriatezza ed efficacia. Riporto alcuni esempi in cui figurano termini tecnici: "L'azienda sta attraversando una crisi economica molto grave, il suo salvataggio è stato affidato ad un *advisor/consulente finanziario* molto conosciuto nel campo"; "Il *dolby surround/sistema dolby* rappresenta un'ulteriore evoluzione qualitativa nel campo della riproduzione sonora cinematografica"; "Secondo *l'exit poll/sondaggio a caldo* il partito X avrebbe totalizzato dal 50 al 55% dei voti"; "La settimana scorsa ho visto un film molto bello con George Clooney nel quale l'attore americano interpretava la parte di un *peacemaker/mediatore*". Bene, come si sono comportati gli intervistati di fronte a tale scelta? Secondo me è molto forte l'"effetto traino" che i mezzi di comunicazione di massa e la comunicazione pubblictaria operano sull'anglicismo. Nel caso di *advisor/consulente finanziario* prevale largamente la soluzione italiana, poiché l'anglicismo non gode di particolare fortuna nei media; viceversa *dolby surround* prevale quasi unanimemente su *sistema dolby* grazie alla potente amplificazione della pubblicità. Nel caso di *exit poll* il bombardamento televisivo in occasione di ogni tornata elettorale assegna la preferenza alla forma inglese, ma non in misura schiacciante; in questo caso è probabile che a sfavore dell'anglicismo giochi anche l'antipatia e la diffidenza che il pubblico ha nei confronti di tutto ciò che ha a che fare con la politica e che, quindi, il tecnicismo imposto dai giornali e dalla televisione risulti sgradito e rimpiazzato con l'assai poco diffuso *sondaggio a caldo*. Infine il successo quasi totale di *peacemaker* su *mediatore* è ascrivibile non certo alla perspicuità dell'anglicismo, ma alla popolarità di Clooney e al fatto che il film sia uscito nelle sale italiane con il titolo originale. Di fatto, con l'eccezione di *advisor*, nell'àmbito tecnico-specialistico il forestierismo sembra largamente prevalere. Il predominio è tanto più consistente quanto più l'anglicismo è appoggiato dai mezzi di comunicazione di massa e dalla pubblicità.

Il quinto quesito proponeva una valutazione "esterna", diciamo *grosso modo* di carattere sociolinguistico. Per ciascun anglicismo si sono proposti alcuni parametri che richiedono all'informatore un giudizio relativo alla comprensibilità, all'utilità, alla novità, alla diffusione, all'appartenenza diafasica, alla facilità di scrittura e pronuncia. Nel dettaglio le coppie dei parametri erano le seguenti: "Facile da capire/Difficile da capire", "Utile/Superfluo", "Nuovo/Datato", "Diffuso/Ristretto a un ambito tecnico-scientifico", "Facile da scrivere o da pronunciare/Difficile da scrivere o da pronunciare". Questo giudizio così articolato è un'utile risposta alla speranza di successo che il corrispondente italiano può avere rispetto alla voce inglese. Le tante voci degli informatori costituiscono "il giudizio degli altri", quel giudizio che nella nuova edizione di *Inglese-Italiano 1 a 1* inseriremo a complemento del nostro. In genere diciamo che si può intravedere una sorta di "scala di implicazione" per cui, per lo più, la risposta affermativa al para-

metro "utile" implica una risposta affermativa anche al parametro "nuovo" (e viceversa alla superfluità raramente si accoppia la novità); e ancora in genere il parametro "facile da capire" è strettamente collegato a "facile da scrivere o da pronunciare" (per converso "difficile da capire" va d'accordo con "difficile da scrivere o da pronunciare"). Per il resto la combinazione dei parametri è la più varia. Alcune considerazioni generali. Non vi è una linea coerente nell'attribuire un vocabolo all'àmbito tecnico-scientifico; la sensazione è che l'etichetta di tecnicismo sia attribuita al vocabolo sconosciuto, a prescindere dalle sue caratteristiche intrinseche. Generalmente i tecnicismi (per es. *advisor*, *counseling*, *day surgery*, *highlight*) sono giudicati di difficile comprensione ma si aggiudicano il beneficio della novità. La novità, peraltro, appare una caratteristica quasi costante degli anglicismi. È allora interessante chiederci quali sono le voci inglese sentite come datate. Anche in questo caso, è molto spesso la frequenza d'uso a dare una patina di "antichità" ai singoli vocaboli; è certamente questo il caso di *chat line* e di *day*, due anglicismi diffusissimi; è il caso di *flop*, effettivamente tra i più acclimati; e ancora di parole assai comuni quali *ticket*, *trendy*, *trolley*. Tutto ciò ci fa capire come nella valutazione del parlante i vari parametri si intreccino secondo volute difficilmente definibili a priori; e come di questo aspetto si debba tenere gran conto quando si avanza una prognosi relativa alla speranza di successo dell'anglicismo o del sostituente italiano.

L'ultimo quesito, infine, proponeva agli intervistati di scegliere dieci anglicismi fra quelli compresi nel *corpus* e di formare altrettante frasi con ciascuno di essi. È evidente, in questo caso, la volontà di scoprire quali anglicismi paiano più familiari agli informatori. In questo caso i risultati sono quelli attesi, ovvero la tendenza è quella di impiegare vocaboli inglesi con i quali si abbia una consuetudine acclarata; quindi, ancora una volta, il primato va alle voci che sono più legate all'universo televisivo e cinematografico, alla pubblicità, al mondo dello sport, soprattutto con riferimento alle pratiche ginniche della palestra. Ovviamente i termini tecnici sono evitati, con qualche eccezione, tra cui vorrei citare almeno *dolby surround* che un discreto numero di informatori (tra i più giovani) utilizzano per costruire frasi create *ad hoc*.

Che cosa ci suggerisce in conclusione il nostro sondaggio a proposito della diffusione e del livello di conoscenza degli anglicismi e in particolare dei tecnicismi presso un pubblico composito per età, ma abbastanza omogeneo per livello culturale? In primo luogo la conoscenza delle parole inglesi continua ad essere piuttosto limitata. Vi sono delle aree privilegiate, ovvero quelle che riguardano gli anglicismi mediatici e gli anglicismi legati al settore dell'informatica e delle nuove tecnologie. Assai più deficitaria risulta la conoscenza delle voci inglesi di àmbito, per dir così, "civile", ossia legate al mondo della politica, della pubblica amministrazione, del funzionamento della società

contemporanea. Se usiamo una nozione ampia di terminologia, in cui rientrino non solo i linguaggi ad alto grado di specializzazione, ma anche quelli meno caratterizzati in senso tecnico, eppure largamente necessari nel corredo intellettuale di ciascun parlante, ci rendiamo conto che la distanza tra il sapere degli specialisti e quello dei parlanti è assai profonda. Il problema principale che emerge dai dati che abbiamo discusso sin qui è non tanto costituito dalla percentuale aritmetica di vocaboli conosciuti rispetto a quelli sconosciuti, quanto dalla generale difficoltà manifestata dall'insieme degli informatori nell'indicare un'alternativa italiana agli anglicismi esaminati. La difficoltà maggiore nel "tradurre" dall'inglese all'italiano deriva non solo e non tanto dalla mancata conoscenza delle parole inglesi, quanto piuttosto dalla sensazione che passando dall'inglese all'italiano si perde una quota sostanziosa della carica connotativa dell'anglicismo; è come se, di per sé, il termine inglese possieda un *quid* di tecnicità intinseca che ne rende la versione italiana qualcosa di mutilato non tanto semanticamente quanto pragmaticamente. L'uso degli anglicismi risulta quindi frutto di un modello imposto dall'alto, un'abitudine quasi inconsapevole da parte del parlante che accetta in modo passivo ciò che gli viene proposto di dire. Non si tratta, dunque, di uno sfoggio snobistico, ma dell'impossibilità di rintracciare un'alternativa italiana altrettanto valida ed efficace. In un certo senso, se ci si passa la metafora turistica, potremmo dire che l'anglicismo fa parte di un pacchetto di offerte che comprende non solo la parola, ma anche il suo significato e il suo *status* sociolinguistico prestigioso.

Al termine del ragionamento, possiamo dire che quando ci avviciniamo ai meccanismi linguistici che regolano l'uso della terminologia presso i parlanti comuni, sia essa anglicizzata o autoctona, ci rendiamo conto che in Italia manca completamente una serie indispensabile di filtri che favoriscano la "digeribilità" del tecnicismo presso gli strati più ampi della popolazione. Mancano del tutto figure di mediazione linguistica. Mi è capitato recentemente di seguire alcune tesi in cui si analizzava il modello comunicativo prodotto da grandi strutture pubbliche, come ad esempio le banche o organismi di risonanza mondiali legati alle Nazioni Unite. Ebbene, la prima domanda che ho posto ai laureandi, che lavorano dentro tali istituzioni, è sempre stata la stessa: chi si occupa della veste linguistica della comunicazione? Le risposte sono state a dir poco sconfortanti: dirigenti di varia estrazione, capiufficio di buona volontà, stagisti reclutati chissà come, con risultati, spesso, inevitabilmente deludenti. E i linguisti? E i traduttori? E i mediatori linguistici? Non ve n'è traccia. Tempo addietro si era manifestata una certa sensibilità al problema. Ricorderete tutti la famosa riscrittura della bolletta dell'Enel patrocinata da De Mauro e i suoi allievi [18], ricorderete i manuali di stile di Cassese e Fioritto, vòlti a semplificare e ammodernare il linguaggio dell'amministrazione pubblica [19]. Ma oggi sembra sia passato un rapido colpo di spugna e la situazione pare

tornata alla solita, deludente sordità verso le ragioni di chi tenta di favorire un collegamento tra il necessario e sempre crescente bisogno sociale di terminologia da un lato, e l'esigenza democratica di ampliare le fasce di cittadini in grado di capire e usare la terminologia medesima. È proprio però da Convegni come questo e come quelli che lo hanno preceduto negli anni scorsi, che deve partire un segnale di rinnovata sensibilità al problema della terminologia, non più visto solo dalla parte di chi la produce, ma anche dalla parte di chi la riceve e anche, possibilmente, da quella di chi ha l'immane compito di trasferirla da un polo all'altro del processo comunicativo, ovvero linguisti, interpreti, traduttori e altri operatori nel campo della comunicazione.

Note

[1] Mi riferisco al mio intervento al Convegno Ass.I.Term. tenutosi a Bertinoro nel giugno del 2007.
[2] Cfr. De Mauro *et alii* (1988).
[3] Dardano-Giovanardi-Pelo (1988: 153).
[4] Cfr. Zanola (2007a).
[5] Cfr. Dardano (2006/2008: 8). Tra i saggi dedicati dallo studioso ai linguaggi settoriali e scientifici ricordo la pregevole sintesi Dardano (1994), nonché Dardano (1998) sul linguaggio dell'economia.
[6] Cfr. *GRADIT* (2003), De Mauro (2005), Adamo-Della Valle (2003) e (2005).
[7] Cfr. Gualdo (2007: 45).
[8] Si vedano le considerazioni svolte al riguardo da Frenguelli (2006).
[9] Mi sia consentito, al riguardo, di rinviare a Giovanardi-Gualdo (2003: 9-27) e a Giovanardi (2007).
[10] Da ultimo: Dardano (2006/2008: 19).
[11] Ne ha parlato, da ultimo, Dardano (2006/2008).
[12] Si veda Serianni (2007: 17-29).
[13] Si veda la nota 1.
[14] Cfr. Zanola (2007b: 117-119). Anche Frenguelli (2006: 225) ricorda come ormai si preferisca abbandonare termini italiani di antico impianto, come *consulente*, *drenaggio fiscale*, *pareggio di bilancio*, in favore dei corrispettivi inglesi *advisor*, *fiscal drag*, *breakeven*.
[15] Per gli esempi relativi al vocabolario finanziario rinvio a Zanola (2007b: 117-119). Anche Gualdo-Scarpino (2007) avevano più o meno analizzato alla stessa maniera altre tipologie di anglicismi.
[16] Mi riferisco a Dardano-Frenguelli-Perna (2000).
[17] Il riferimento è a De Mauro-Vedovelli (2001).

[18] Cfr., in particolare, Fioritto (1997), che riprende e rielabora precedenti considerazioni di Sabino Cassese.

Bibliografia

Adamo-Della Valle (2003) = G. Adamo, V. Della Valle, *Neologismi quotidiani. Un dizionario a cavallo del millennio 1998-2002*, Firenze, Olschki.

Adamo-Della Valle (2005) = G. Adamo, V. Della Valle, *2006 parole nuove*, Milano, Sperling & Kupfer.

Dardano (1994) = M. Dardano, *I linguaggi scientifici*, in *Storia della lingua italiana*, vol. II: *Scritto e parlato*, a cura di L. Serianni, P. Trifone, Torino, Einaudi, pp. 497-551.

Dardano (1998) = M. Dardano, *Il linguaggio dell'economia e della finanza*, in *Con felice esattezza. Economia e diritto tra lingua e letteratura*, a cura di I. Domenighetti, Bellinzona, Casagrande, pp. 65-87.

Dardano (2006/2008) = M. Dardano, *La lingua italiana d'oggi*, in *Siamo una nazione? Nationales Selbstverstandnis im aktuellen Diskurs über Sprache, Literatur und Geschichte Italiens*, a cura di S. Schwarze, Tübingen, Stauffenburg, pp. 119-142; poi ristampato col titolo *Tra innovazione e conservazione*, in *L'italiano di oggi*, a cura di M. Dardano - G. Frenguelli, Roma, Aracne, pp. 7-28.

Dardano-Giovanardi-Pelo (1988: 153) = M. Dardano, C. Giovanardi, A. Pelo, *Per un'analisi del discorso divulgativo: accertamento e studio della comprensione*, in De Mauro et alii (1988), pp. 153-164.

Dardano-Frenguelli-Perna (2000) = M. Dardano, G. Frenguelli, T. Perna, *L'italiano di fronte all'inglese alle soglie del terzo millennio*, in *L'italiano oltre frontiera*. Atti del V Convegno internazionale (Leuven, 22-25 aprile 1998), vol. I, a cura di S. Vanvolsem et alii, pp. 31-55.

De Mauro (2005) = T. De Mauro, *La fabbrica delle parole. Il lessico e problemi di lessicologia*, Torino, Utet Libreria.

De Mauro et alii (1988) = T. De Mauro et alii (a cura di), *Dalla parte del ricevente: percezione, comprensione, interpretazione*, Atti del XIX Congresso internazionale di studi SLI, Roma, 8-10 novembre 1985, Roma, Bulzoni.

De Mauro-Vedovelli (2001) = T. De Mauro, M. Vedovelli (a cura di), *Dante, il gendarme e la bolletta: La comunicazione pubblica in Italia e la nuova bolletta Enel*, Roma-Bari, Laterza.

Fioritto (1997) = A. Fioritto, *Manuale di stile. Strumenti per semplificare il linguaggio delle amministrazioni pubbliche*, Bologna, Il Mulino.

Frenguelli (2006) = G. Frenguelli, *Ricezione degli anglismi e mezzi di comunicazione di massa*, in *La "nuova Europa" tra identità culturale e comunità politica*. Atti del Convegno internazionale (Roma, Università La Sapienza, 21-22 ottobre 2005), a cura di F. Cabasino, Roma, Aracne, pp. 222-236.

Giovanardi (2007) = C. Giovanardi, *Sulla traducibilità in italiano degli anglicismi contemporanei: alcune proposte*, in Vanvolsem et alii (2007), pp. 241-255.

Giovanardi-Gualdo (2003) = C. Giovanardi, R. Gualdo (con la collaborazione di A. Coco), *Inglese-Italiano 1 a 1. Tradurre o non tradurre le parole inglesi?*, San Cesario di Lecce, Manni.

GRADIT (2003) = T. De Mauro, *Nuove parole italiane dell'uso*, Torino, UTET.

Gualdo (2007) = R. Gualdo, *Punti di vista su terminologia e lingua comune*, in Zanola (2007a), pp. 43-63.

Gualdo-Scarpino (2007) = Riccardo Gualdo, Cristina Scarpino, *Quanto pesa l'inglese? Anglicismi nella vita quotidiana e proposte per la coabitazione*, in Vanvolsem et alii (2007), pp. 257-281

Serianni (2007) = L. Serianni, *Terminologia medica: qualche considerazione tra italiano, francese e spagnolo*, in Zanola (2007a), pp. 7-29.

Vanvolsem et alii (2007) = S. Vanvolsem et alii (a cura di), *Identità e diversità nella lingua e nella letteratura italiana*, Atti del XVIII Congresso dell'AISSLI (Lovanio, Louvain-la-Neuve, Anversa, Bruxelles, 16-19 luglio 2003), vol. I: *L'italiano oggi e domani*, Firenze, Cesati.

Zanola (2007a) = M.T. Zanola (a cura di), *Terminologie specialistiche e tipologie testuali. Prospettive imterlinguistiche*. Atti del Convegno (Milano, Università Cattolica, 26-27 maggio 2006), Milano, Università Cattolica del Sacro Cuore.

Zanola (2007b) = M.T. Zanola, *Terminologia dell'economia e della finanza: prospettive di studio*, in Zanola (2007a), pp. 109-132.

Introduction aux technologies du Web Sémantique

ROGER ROBERTS

Starting from the theoretical basis of general linguistics, the author proposes an overview of the technologies used to manage information and the documentary content on the World Wide Web. The main focus of his attention is on the semantic Web, outlining its technologies and potentiality.

Keywords: HTML – Protegé – RDF – Semantic Web – XML

La sémantique est une question de distance…, distances géographique, temporelle et culturelle. Elle permet de construire des ponts entre mon interlocuteur et moi lorsque nous sommes éloignés géographiquement, entre celui que je serai dans 10 ans et moi quand il s'agit d'accéder à nouveau à ce que j'ai conservé, entre un interlocuteur inconnu et moi qui n'évoluons pas dans le même environnement culturel… Il est donc tout à fait normal de retrouvé cette composante majeure dans l'univers de la toile qui tisse des liens entre des individus qui échangent de l'information au sein d'un village virtuel!

Dans l'ensemble des langages véhicules par les humains, il en est un auquel Jacques Derrida accorde la primauté sur les autres: la langue. Il désigne ce système métaphysique comme *logocentrisme*. Derrida élabore une différence proche de celle qui, chez Ferdinand de Saussure <fr.wikipedia.org/wiki/Ferdinand_de_Saussure>, donne sens aux éléments signifiants, sous forme de trace. La "trace " <fr.wikipedia.org/w/index.php?title=Trace_(philosophie)&action=edit&redlink=1>, cependant, pleinement en eux-mêmes, il n'y a aucune vérité première, aucune différence transcendantale à poursuivre.

Or, la *différance* (qui du fait de sa représentation graphique différente par la présence d'un "a") est précisément le mouvement "producteur" de ces différences: elle est le "processus" par lequel diffèrent les concepts.

Ces analyses offrent un intérêt majeur pour tous les chercheurs qui sont à l'étude de théories cherchant à établir des bases opérationnelles pour une sémiologie accessible à des machines. Dans l'univers audiovisuel, l'hypothèse de Jacques Derrida est pleine de sens dans la mesure où pour communiquer, pour indexer un objet ou pour échanger des objets, le recours à la langue et plus précisément à sa représentation textuelle, est le seul outil praticable. Une représentation particulièrement accessible pour des outils informatique!

Au début du XXè siècle, les travaux menés dans le cadre de la linguistique ont ouvert des portes extrêmement précieuses à toutes les personnes essayant de comprendre l'univers particulièrement hermétique de la signification. On estime en Europe que c'est Ferdinand de Saussure qui a fondé la linguistique <fr.wikipedia.org/ wiki/ Linguistique> moderne et établi les bases de la sémiologie <fr.wikipedia.org/wiki/ Sémiologie>.

Dans son Cours de linguistique générale <fr.wikipedia.org/w/index.php?title= Cours_de_linguistique_générale&action=edit&redlink=1> (publié *post-mortem* en 1916), c'est lui qui a défini entre autre la distinction entre langage, langue <fr.wikipedia.org/ wiki/Langue> et parole <fr.wikipedia.org/wiki/Parole>, le caractère arbitraire du signe linguistique... <fr.wikipedia.org/wiki/Signe_linguistique>. L'idée fondamentale de Saussure est que le langage est un système clos de signes. Tout signe est défini par rapport aux autres, par pure différence (négativement), et non par ses caractéristiques propres ("positives"): c'est pourquoi de Saussure parle de «système»:

- Sémantique: l'étude du «sens» des symboles et expressions. Il s'agit de considérer le «sens» de façon opérationnelle
- Langage: un moyen de communication avec un ensemble de signes (vocaux, gestuels, graphiques, tactiles, olfactifs, etc.) doté d'une sémantique, et le plus souvent d'une syntaxe.
- Langue: un système de signes linguistiques, vocaux ou graphiques ou gestuels, qui permet la communication entre les individus doté d'une syntaxe précise et d'une grammaire.
- Métadonnées: une métadonnée est une donnée servant à définir ou décrire une autre donnée quel que soit son support (papier ou électronique).

La théorie linguistique de Ferdinand de Saussure est nettement sémiotique <fr.wikipedia.org/wiki/Sémiotique> dans la mesure où elle interprète le langage comme un ensemble de signes: le linguiste distingue dans le signe <fr.wikipedia.org/wiki/ Signe> deux éléments: le signifiant <fr.wikipedia.org/wiki/Signifiant> et le signifié <fr. wikipedia.org/wiki/Signifié>:

- **Le signifiant** désigne la représentation graphique d'un mot, d'une image, d'un son... Ce qui importe dans une représentation, ce sont les différences qui les distinguent les uns des autres. La valeur d'une représentation découle de ces différenciations. Chaque langage construit des lexiques à partir d'un nombre limité de caractères (phonèmes pour la langue), et d'une syntaxe qui définit l'ordre dans lequel ces caractères doivent être organisés. Pour l'univers informatique, le HTML est à tous points de vue, un signifiant d'une représentation en évidence!
- **Le signifié** désigne le concept, c'est-à-dire une représentation sémantique associée à un signifiant. Cette observation conduit Saussure à distinguer également **signification** et **valeur** puisque l'existence de langues différentes introduit néces-

sairement des significations et des valeurs différentes. Ainsi le signifié est un concept défini du fait de l'existence ou de l'absence dans une langue d'autres concepts qui lui sont opposables.

Pour certains psychanalistes, et notamment Jacques Lacan, tout l'intérêt de l'analyse de Saussure se situe dans cette ligne qui représente à la fois la différence mais également le lien entre le signifiant et le signifié. Pour le monde informatique, elle exprime la différence entre le HTML et les autres langages issus du XML et de ses dérivés!

Tous les travaux entrepris dans le cadre du «web sémantique» visent en fait à permettre de rapprocher au travers d'outils sophistiqués le signifiant du signifié afin de rendre les différences accessibles à des machines. Il ne faudrait pas oublier tous ceux qui ont contribué au premiers pas de l'informatique et rendu possible ce débat machine/sémantique. Comme Alan Turing, considéré comme un des pères fondateurs de l'informatique <fr.wikipedia.org/wiki/Informatique> moderne. Il est à l'origine de la formalisation des concepts d'algorithme <fr.wikipedia.org/wiki/Algorithmique> et de calculabilité <fr.wikipedia.org/wiki/Calculabilité>. La théorie des classes, qui permet de parler de collections d'objets qui ne sont pas nécessairement des ensembles chère à von Neumann ou encore Gödel avec le premier théorème d'incomplétude (dans n'importe quelle théorie récursivement axiomatisable, on peut construire un énoncé arithmétique qui ne peut être ni prouvé ni réfuté dans cette théorie).

La création et le développement rapide d'un outil de communication à vocation mondiale ne pouvaient que rebondir sur toutes les avancées proposées puisqu'en tant que outil gérant de l'information, elle sous-tend l'ensemble des activités liées à l'encodage d'information dans des bases de données et à des outils de recherche ouverts.

Le World Wide Web représente aujourd'hui une avancée technologique de première importance par l'influence qu'il exerce sur la majorité des aspects de nos économies et de nos sociétés. Cependant, en l'état actuel, il est peu satisfaisant en raison du nombre élevé d'activités souhaitées qui ne sont pas bien prises en charge par les outils automatiques. Ainsi, l'outil principal servant à la récupération d'informations est constitué de moteurs de recherche basés sur des mots clés. Ces outils, mêmes s'ils sont indispensables, présentent de fortes restrictions en termes de récupération, précision et contenu à partir de pages du web. Il est en effet assez aberrant de pratiquer au niveau de la recherche un outil basé sur le "protocole Gutenberg": en clair, le moteur de recherche analyse la syntaxe du signifiant en n'ayant aucun outil capable de resituer cette représentation dans un contexte de signifié. Pour la machine EBU signifie tout autant European Broadcast Union, European Boxing Union que European Bhuddist Union qui sont effectivement des organisations pleinement reconnues, mais que tout distingue dans leur raison sociale et morale ou sportive.

L'essentiel du contenu actuel est destiné à une interprétation par l'homme; la machine n'étant capable de le capturer et de le manipuler qu'au niveau syntaxique et donc de proposer un nombre incalculable de réponses à une requête non définie au niveau contextuel.

L'idée maîtresse du web sémantique est de rendre le contenu accessible et assimilable par la machine. Ceci ouvre la voie au développement d'outils sophistiqués susceptibles d'apporter un niveau bien supérieur de fonctionnalités pour assister les activités humaines sur le web.

Le web sémantique repose sur l'association des technologies suivantes:

- *les métadonnées explicites:* elles permettent aux pages web de comporter leur signification dans leurs balises.
- *les ontologies:* il s'agit des principes fondamentaux d'un domaine et leurs relations. *la logique:* elle permet de déduire des conclusions en combinant les métadonnées aux ontologies.

Brève introduction aux technologies du web sémantique:

HTML (*Hyper text Mark-up language*) est le langage à balises standard dans lequel sont écrites les pages web. Il repose sur une série de balises prédéfinies qui contrôlent l'édition d'une page (comme les caractères gras ou italiques d'une police, les listes numérotées ou non, les ruptures de ligne, etc.) Bref, dans le langage de Ferdinand de Saussure il s'agit bien de la représentation en évidence du signifiant.

Bien que le langage XML (*eXtensible Mark-up Language*) repose également sur des balises pour l'enrichissement du contenu web, ce langage permet aux utilisateurs de définir leurs propres balises. A cet égard, XML est donc un métalangage à balises indépendant du domaine (langage servant à définir un langage à balises). Les balises définies par l'utilisateur structurent la page qui ainsi devient assimilable.

Par contre, les balises XML ne décrivent pas la mise en forme des pages web. XML distingue donc le contenu de sa mise en forme, une caractéristique bien utile pour déterminer différentes présentations et vues sur la base des mêmes données et constitue la base pour une représentation du signifié.

XML fait en réalité partie d'une famille de langages destinés à diverses activités s'articulant autour du noyau du langage XML:

- DTD et XML Schema: deux langages permettant à l'utilisateur de définir son propre vocabulaire.
- XPath: langage fournissant l'accès à certaines parties des documents XML. Cet accès est une condition préalable et nécessaire pour adresser une requête de documents XML.

- XQuery: langage de requête destiné à XML.
- XSLT: langage déterminant les transformations de XML en HTML ou entre des représentations XML. On a ainsi XSLT comme outil essentiel pour la manipulation syntaxique des documents XML.

Dans l'élaboration du web sémantique, XML fournit la couche de base de la manipulation syntaxique. Bien que XML soit un langage universel pour définir des balises, il ne procure aucun moyen d'approcher la sémantique (le sens) des données. Il n'y a, par exemple, aucun signifié associée à l'encapsulation des balises. Il revient à chaque application d'interpréter l'emboîtage ce qui dans les faits, nécessite des outils supplémentaires pour pallier à ce manque.

Le RDF

RDF (*Resources Description Framework*) est un langage servant à décrire des ressources. Son élément constitutif de base est la formulation d'un triplet se composant d'une entité (appelée ressource en terminologie web), d'une propriété et d'une valeur (qui peut être une autre ressource). La formulation) est essentiellement la définition d'un fait P (a,b) où P est une propriété binaire, et (a,b) sont des ressources. Dans la perspective du web sémantique, RDF définit une couche située au-dessus de XML. De ce fait, RDF a été doté d'une syntaxe XML.

RDF est indépendant du domaine, en d'autres mots, il ne pose aucun présupposé quant à un domaine spécifique. C'est donc à l'utilisateur que revient le rôle de définir sa propre terminologie dans un langage schéma appelé RDF Schema (RDFS). Constitutivement, RDFS est un langage d'ontologies primitif (ou naturel) proposant les caractéristiques suivantes:

- Organisation des objets en classes (professeur, membre du personnel, étudiant, cours, cours pour étudiants) et propriétés binaires (enseigne, étudie, travaille).
- Sous-classes (tous les professeurs sont des membres du personnel) et relations des sous-propriétés (tout chef de département fait partie de ce département).
- Restrictions de domaine (seul le personnel académique peut enseigner) et d'étendue (une personne ne peut qu'enseigner) au niveau des propriétés.

RDF et RDFS fournissent les langages de base pour le web sémantique.

OWL

La puissance d'expression de RDF et de RDFS est volontairement très limitée: RDF est (en gros) limité à des attributs binaires et RDFS est (toujours en gros) limité aux

hiérarchies de sous-classes et de sous-propriétés, avec restrictions de domaine et d'entendue des propriétés.

Il existe cependant plusieurs cas particuliers d'utilisation du web sémantique qui nécessitent une plus grande expressivité. Ce type d'extensions comprend:

- la *disjonction* (par ex. une personne ne peut être à la fois professeur et membre du personnel administratif).
- les *combinaisons booléennes des classes* (par ex. le personnel est l'ensemble du corps académique, du personnel administratif et du personnel d'assistance technique).
- les *restrictions de cardinalité* (par ex. un service ne peut avoir qu'un seul chef).
- les *caractéristiques spéciales des propriétés* (par ex. "supérieur de" est transitif, "enseigne" et " reçoit des cours de " sont des propriétés inverses)
- l'*étendue locale des propriétés ("range")* définit la portée d'une propriété, par exemple "mange" pour toutes les classes. Dans certains cas, on peut souhaiter réduire la portée en fonction de la classe. Par exemple, les vaches ne mangent que de l'herbe tandis que d'autres animaux mangent également de la viande.

OWL a été élaboré comme nouveau langage d'ontologies standard pour le web. Il repose sur RDFS et tente de trouver un équilibre entre puissance d'expression et support logique efficace. La logique (le raisonnement) est un facteur important parce qu'elle permet de

(a) vérifier la cohérence d'une ontologie et des connaissances,
(b) vérifier la présence de relations non voulues entre classes
(c) classer automatiquement les instances en classes.

Logique

La création formelle du langage OWL est une partie de la représentation et du raisonnement des connaissances appelée logique descriptive. Cette création est riche de promesses; l'approche est différente pour la représentation et le raisonnement sur la base de règles. Ses principaux avantages sont:

- Les moteurs de règles existent et sont très puissants.
- Les règles sont bien connues et s'utilisent en informatique générale. Elles sont faciles à apprendre.

Les systèmes de règles peuvent être envisagés comme une extension ou une alternative à OWL. La première idée est de poursuivre les recherches actuelles en visant à intégrer les règles et la logique descriptive tout en maintenant un support logique assez efficace. Une idée plus récente étudie l'utilisation de RDF/S conjointement aux règles comme base d'un autre langage ontologique web.

Outre les systèmes classiques à base de règles, il est intéressant d'analyser ceux capables de prendre en compte des conclusions contradictoires. Ces systèmes sont utiles pour la modélisation des données ayant hérités de défauts et des règles comportant des exceptions. Ils sont aussi très pratiques pour l'intégration des connaissances où des incohérences peuvent évidemment se produire lorsqu'on assemble des connaissances de sources différentes.

Pour plus de plaisir (une partie du texte qui précède a été rédigé avec la complicité active de ce site web et de Wikipedia <www.ics.forth.gr/isl/swprimer>).

L'idée principale est de rendre le sens accessible et manipulable par un moteur de recherche et une organisation qui prennent en compte la dimension culturelle de l'activité humaine grâce aux nouvelles technologies développées au sein du Web sémantique.

L'industrie audiovisuelle a produit ces dernières années beaucoup de standards afin de contribuer à aider les diffuseurs dans leur quête sémantique:

1. Solutions disponibles: AAF / MXF / SMPTE / P-Meta/ TV anytime … (pragmatique mais …)
2. Solutions pour la représentation de metadonnées, structures and synchronisation: SMIL (*Synchronized Multimedia Integration Language*) / …
3. Modèles spécifiques: SMEF (BBC) / RAI / FARAO (ORF) / INA / IMMIX (Pays-Bas) / DR-M / MPEG-7/ MPEG 21 <www.enthrone.org>, …
4. Encapsuleurs: MXF / METS / PK-ZIP / SPK-ZIP / PDF/A
5. Ontologie and Sémantique: FRBR (*Functionnal requirements for Bibliographic records* - IFLA)
6. Ontologie: Dublin Core Metadata Initiative (DCMI) / RDF / MARC (*MAchine-Readable Cataloguing*).

Une ontologie décrit les concepts et les rapports qui sont importants dans un domaine particulier, fournissant un vocabulaire pour ce domaine aussi bien que des spécifications automatisées de la signification des termes utilisés dans le vocabulaire. Les Ontologies traitent de taxonomies et de classifications, schémas de base de données, et de théories entièrement axiomatisées. Ces dernières années, des ontologies ont été introduites dans beaucoup de communautés scientifiques de manière à partager, réutiliser et traiter la connaissance d'un domaine spécifique. Les Ontologies sont devenues vitales pour beaucoup d'applications telles que des portails de la connaissance scientifique, des systèmes de gestion d'information et d'intégration, du commerce électronique, et de services de Web sémantique. L'absence d'une véritable solution générique au niveau de la norme MPEG 7 de l'ISO à constitué un handicap majeur!

Cerise sur le gâteau, les scientifiques de certaines universités ont développé des technologies nouvelles qui sont de véritables moteurs d'ontologie qui vont grandement simplifier la vie de tous ceux qui rêvent de produire des outils en OWL, RDF, …

Un moteur d'ontologies est un processus applicatif qui possède une dimension sémantique (définition et mise en relation de concepts); une dimension logique (validation de la création de relations entre concepts ou déduction de relations non explicites entre concepts) et enfin une dimension usage avec la construction d'un vocabulaire («outil», «utilisateur», «traitement»), la construction d'une syntaxe (sujet, verbe, complément) et d'une grammaire: «l'utilisateur» «accorde» «l'outil» pour réaliser «un traitement».

La finalité du moteur d'ontologie est de:

- réduire la distance entre le langage de la machine (logique) et le langage de l'utilisateur (lisible) en intercalant entre l'une et l'autre un langage compréhensible par l'un et l'autre;
- réduire la distance entre des langages pratiqués dans différentes cultures (métier, région, temps, etc.) en intercalant entre l'une et l'autre un langage interprétable par l'un et l'autre;
- améliorer la qualité, l'efficacité et l'efficience des échanges entre les acteurs d'un projet via un socle sémantique commun;

Un moteur d'ontologies est un processus applicatif qui possède:

Une dimension sémantique:	Définir des concepts Mettre en relation des concepts
Une dimension logique:	Valider la création de relations entre concepts Déduire des relations non explicites entre concepts
Une dimension usage:	Construction d'un vocabulaire («outil», «utilisateur», «traitement») Construction d'une syntaxe (sujet, verbe, complément)
Construction d'une grammaire:	«l'utilisateur» «accorde» «l'outil» pour réaliser «un traitement».

Un exemple: «Protege»

«Protege» est une plate-forme ouverte développée par l'université de Stanford et qui fournit à une communauté d'utilisateur une série d'outils logiciels pour construire des modèles de domaine et des applications basées sur la connaissance des ontologies. En son sein, «Protege» met en application un ensemble riche de structures de «connaissance-modélisation et actions» qui soutiennent la création, la visualisation, et la manipulation des ontologies dans divers formats de représentation. «Protege» peut être

adapté aux besoins d'un client pour fournir l'appui logistique à la création de modèles de la connaissance de saisie de données. «Protege», basé sur Java, est extensible, et fournit un environnement prêt à l'emploi qui en fait une base flexible pour le prototypage et le développement d'applications rapides. De plus, «Protege» peut être utilisé en mode Plug-in avec une interface de programmation API pour construire des outils basés sur la connaissance et le développement d'applications.

La plate-forme Protégé propose deux manières principales de modeler des ontologies:

- L'éditeur «Protege - Frames» permet à des utilisateurs d'établir et populer des ontologies basées sur le protocole ouvert de connectivité de base de connaissance (OKBC). Dans ce modèle, une ontologie se compose d'un ensemble de classes organisées dans une hiérarchie pour représenter les concepts fondateurs d'un domaine, un ensemble de connecteurs associés aux classes pour décrire leurs propriétés et rapports, et un ensemble d'exemples de ces classes – différents exemplaires des concepts qui tiennent des valeurs spécifiques pour leurs propriétés.
- L'éditeur de «Protege-OWL» permet à des utilisateurs d'établir des ontologies pour le Web sémantique, en particulier dans la spécification du W3C (OWL).

La BBC a mis en application «Protege» pour contrôler une description RDF/OWL des contenus des nouveaux médias, pour éditer les sites Web, la TV interactive, et les plates-formes mobiles, etc. La BBC a créé une ontologie OWL décrivant le metamodèle de l'entreprise. Elle utilise «Protege» (avec quelques connexions faites sur demande) pour créer des schémas d'objet et des conceptions d'interaction en vue d'éditer ces objet et faciliter la saisie sur les systèmes de gestion de contenus.

L'aventure ontologique commence à l'adresse suivante: <protege.stanford.edu>.

Sitographie

Cover Pages, METS: <xml.coverpages.org/mets.html>.
Dublin Core: <dublincore.org/documents>.
Introduction au web sémantique: <www.ics.forth.gr/isl/swprimer>.
Metadata Principles and Practicalities, 2002, "The D-Lib Magazine", April 2002, Vol. 8 N. 4: <www.dlib.org/dlib/april02/weibel/04weibel.html>.
MPEG-7: <mpeg.tilab.comcselt.it>.
RDF' by W3C: <www.w3.org/TR/rdf-primer>.
The Metadata Encoding and Transmission Standard, (METS), <www.loc.gov/standards/mets/>.
Un moteur d'ontologie: <protege.stanford.edu>.

Remerciements

Je voudrais remercier toutes les personnes qui ont contribué directement ou or indirectement à cette présentation au Congrès de l'Association Italienne de Terminologie:

- SKEMA (UTC Compiègne), pour les contributions aux developpements sur les Ontologies et le projet MediaMap
- PROSI* and MEMNON*, in particulier M. Guy Marechal et M. Michel Merten pour les développements d'AXIS
- EBU, en particulier M. Jean-Pierre Evain

TITAN, pour l'organisation des "European Media Wrapper Conference and Round Table"

- SBS SIEMENS*, en particulier M. John Jordan (past manager of the BBC SMEF project)
- "ISO", pour la contribution à la normalisation de l'OAIS

Energie tradizionali e rinnovabili: proposte di interventi terminologici

MARIA TERESA ZANOLA

As public concern over renewable energy sources grows, there is a large amount of publications which focus on the issue, ranging from scientific discussions to institutional communication and mass media coverage of the issue itself. The different needs of this array of communicative situations are met by variations in terminology. This study investigates the terminology used by some electricity and gas companies and analyses their purposes of clarifying their public communications, both for experts in the field and for the general public. Our inquiry is focused on on-line glossaries and reports in order to examine the popularisation of technical terminology in this field.

Keywords: Terminology – specialized lexicon – institutional communication

Dinanzi al diffondersi sempre crescente del dibattito intorno alle energie rinnovabili, aumenta anche l'ampiezza descrittiva del fenomeno, vuoi per divulgazione scientifica specialistica, vuoi per divulgazione istituzionale di massa. Così le variazioni terminologiche delle forme di energia rinnovabile solare, geotermica, fotovoltaica, eolica, lasciano il discorso specialistico per essere accolte, con crescente rapidità, da programmi e orientamenti di società nazionali e private, da dichiarazioni programmatiche di linee politiche, dai linguaggi dei media.

Abbiamo ritenuto utile osservare la descrizione terminologica offerta da compagnie e società che operano nel settore e che si sono preoccupate di arricchire la loro comunicazione pubblica – agli specialisti del settore, così come agli utenti – di glossari consultabili on-line o nella stampa di settore, nell'intenzione di osservare i punti di incontro fra necessità di denominazione e divulgazione delle conoscenze.

1. Nuovi concetti, nuovi termini

Silvi (2005) sottolinea che «per indicare l'energia del sole e le fonti rinnovabili alcuni termini sono entrati nell'uso comune mentre altri […] sono stati completamente dimenticati». Nel 1951 Palmer Putnam parlava di *fonti di energia capitale* che distingueva dalle *fonti di energia rendita*, in base ad una diversa qualità economica delle fonti

stesse. Silvi spiega che «le fonti fossili sono un capitale, di cui disponiamo sul "libretto di risparmio" creato per noi dalla natura in milioni anni, mentre l'uso delle fonti di "energia rendita" non ne intacca la consistenza e quindi queste costituiscono una rendita a disposizione nostra e delle future generazioni». Capiamo il significato del sintema in questo contesto, ma non lo ritroviamo più tra le attuali pratiche d'uso.

È interessante allora analizzare il percorso di diffusione dei nuovi concetti e termini che si sono affermati nel settore energetico a partire dalla seconda metà del Novecento, ricordando gli accadimenti storici, sociali e scientifici più significativi che li hanno determinati.

Due saranno i campi tematici oggetto di osservazione: le nuove fonti di energia e le nuove categorie di utenti e clienti del settore energetico.

1.1. "Nuove" fonti di energia

Il 4 maggio 1956 il Consiglio economico e sociale delle Nazioni Unite aveva deciso che «le Nazioni Unite mostrassero per tutte le nuove forme di energia lo stesso interesse avuto per quelle convenzionali e per quella atomica» (Silvi, 2005).

Nell'agosto 1961, nel corso della conferenza internazionale su "New Sources of Energy and Energy Development", promossa dalle Nazioni Unite presso la FAO, viene riproposta la poliremática "nuove fonti di energia": l'aggettivo "nuove" sta ad indicare il carattere innovativo delle tecnologie tramite cui erano sfruttate le energie solare, geotermica ed eolica (*Ibidem*).

Silvi ricorda che Agostino Capocaccia poneva nel 1972 la distinzione fra *fonti a riserva finita* e *fonti a riserva infinita* di energia. A seguito dello shock petrolifero del 1973, i termini furono adombrati in favore dell'espressione *energie alternative*, iperonimo dell'intera categoria di forme di energia altre dal petrolio, quindi sia solare sia nucleare.

The United Nations Conference on New and Renewable Sources of Energy dell'agosto 1981 a Nairobi, a cui parteciparono 125 paesi, introdusse *energie rinnovabili*.

Energie alternative, *rinnovabili*, *assimilate*, *a basso tenore di carbonio* si presentano oggi come la gamma sinonimica delle rinnovabili, o energie rinnovabili, o fonti (di energia) rinnovabili.

È evidente la necessità di trasmettere informazioni chiare ai cittadini, in cui la terminologia del settore sia un riferimento univoco alla comprensione dei fenomeni anche complessi, oltre che una via di accesso chiara alle esigenze del mercato e della vita comune.

Allo scopo di «promuovere una migliore conoscenza per il cittadino consumatore delle nuove opportunità legate all'apertura del mercato europeo dell'energia», l'Autorità

per l'Energia Elettrica e il Gas e la Rappresentanza in Italia della Commissione Europea hanno promosso una pubblicazione (*Le novità nel mercato dell'energia e del gas,* 2008 – (luglio 2008) <www.autorita.energia.it/energia_semplice/opuscolo.pdf> –, nella quale, per esempio, si spiega che cosa sia l'offerta bioraria dell'energia elettrica. Quella di un orario bio? No, è un'offerta dell'energia elettrica bi-oraria o multi oraria, in base alla quale la tariffa varia in funzione delle fasce orarie. Il programma televisivo della RAI "Uno Mattina", *Energia semplice*, un'iniziativa di servizio pubblico lanciata dal programma della RAI, curata dagli esperti dell'Autorità per l'energia, ha promosso una serie di appuntamenti dal 28 febbraio 2008, che riprendono i temi della citata brochure, dedicati a temi quali: "Come si legge la bolletta elettrica", "Come orientarsi e scegliere al meglio fra le nuove offerte del mercato libero", "Cosa sono e come funzionano le offerte biorarie", "I diritti del consumatore per la qualità dei servizi commerciali - pagamenti, rimborsi automatici, appuntamenti".

Energia idroelettrica o *carbone bianco*? Figura retorica di sinonimia e di analogia o speranza di un futuro senza inquinamento? *Effetto serra*, *riduzione e commercio delle emissioni*, *sequestro della* CO^2 sono termini ed espressioni entrati a seguito del testo del protocollo di Kyoto: la scoperta di nuove tecnologie e le pressioni delle politiche ambientali ed energetiche influiscono così su un lessico che non può tuttavia variare la realtà delle energie disponibili sulla Terra, conclude Silvi (2005).

Non stupisce perciò ritrovare abbondanza di glossari in questo settore tematico, presenti nei siti delle principali società per la distribuzione dell'energia in Italia, che propongono agli utenti dei siti la descrizione delle energie rinnovabili e di quelle convenzionali, dai problemi della loro distribuzione a quelli della gestione economico-finanziaria.

Ne sia un primo esempio la definizione di "energie rinnovabili" riportata nel glossario disponibile nel sito del Gestore dei Servizi Elettrici (GSE) - (luglio 2008) <www.grtn.it/ita/glossario/glossario.asp>:

> *Fonti energetiche convenzionali*: Olio combustibile, carbone e gas naturale.
> *Fonti energetiche rinnovabili*: Il sole, il vento, le risorse idriche, le risorse geotermiche, le maree, il moto ondoso e la trasformazione in energia elettrica dei prodotti vegetali o dei rifiuti organici e inorganici.

Ricordiamo lo sfondo economico-politico, relativo alla liberalizzazione della domanda di energia, aperta dal 1° luglio 2007, in attuazione delle Direttive UE 54 (per l'elettricità) <eur-lex.europa.eu/LexUriServ/LexUriServ. do?uri=OJ:L:2003:176:0037:0055: IT:PDF> e 55 (per il gas) <eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:176: 0057: 0078: IT:PDF> del 2003. Le famiglie italiane hanno avuto la possibilità di rivolgersi a venditori di energia elettrica anche diversi da quello da cui sono state rifornite sino ad allora, scegliendo l'offerta ritenuta più interessante. La scelta libera della fornitura per il

gas era già disponibile dal gennaio 2003, avendo l'Italia anticipato le scadenze europee in merito.

1.2. Le nuove categorie di utenti e clienti del settore energetico

L'utente del settore energetico amplia la propria tipologia, e la sua descrizione, quale si presenta nei siti istituzionali, tratteggia la persona fisica o giuridica che ha re-sponsabilità dell'impianto o della connessione ad una rete più importante:

Utente (in <*Ibidem*>)

> Persona fisica o persona giuridica che gestisce, anche non avendone la proprietà, un impianto connesso alla rete di trasmissione nazionale; un utente può essere:
>
> a) diretto (o direttamente connesso), nel caso di connessione diretta dell'impianto alla rete di trasmissione nazionale;
> b) indiretto (o indirettamente connesso), nel caso di connessione indiretta dell'impianto alla rete di trasmissione nazionale; ove non specificato, per utente si intende l'utente diretto.

Utente della rete (in <*Ibidem*>)

> Persona fisica o giuridica che rifornisce o è rifornita da una rete di trasmissione o di distribuzione.

Utente del dispacciamento

> Soggetto che ha concluso con Terna S.p.A. <www.mercatoelettrico. org/It/Tools/Glossario.aspx#terna%23terna> un contratto per il servizio di dispacciamento. Per ciascun punto di offerta è l'unico soggetto tenuto alla presentazione dei margini a salire e a scendere sul MGP e sul MA e alla presentazione di offerte sul MSD, qualora il punto di offerta sia abilitato a tale mercato - (luglio 2008) <www.mercatoelettrico.org/It/Tools/Glossario.aspx>.

Si tratta di definizioni adeguate nella scrittura alla natura stessa del glossario, che riunisce soprattutto parole-chiave dei testi normativi di riferimento delle istituzioni citate. L'utente così descritto assume quindi gradi di specializzazione nello sfruttamento dei servizi, e possiamo anche immaginarlo a sua volta l'utente delle numerose sigle in uso, basti citare MGP - *Mercato del Giorno Prima*, MA - *Mercato di aggiustamento*, e MSD - *Mercato dei servizi di dispacciamento*.

Nella caratterizzazione dei clienti, spiccano le tre categorie di clienti, nell'ordine dato nel glossario del GSE – (luglio 2008) <www.grtn.it/ita/glossario/glossario.asp> –, quelli *idonei*, *vincolati* e *grossisti*.

Clienti

> Le imprese o Società di distribuzione, gli acquirenti grossisti e gli acquirenti finali di energia elettrica.

Cliente finale

È la persona fisica o giuridica che acquista energia elettrica esclusivamente per uso proprio.

Cliente idoneo

Persona fisica o giuridica che può stipulare contratti di fornitura con qualsiasi produttore, distributore o grossista, sia in Italia che all'estero. A partire dal 1° maggio 2003 è classificato cliente idoneo chi consuma più di 100.000 kwh all'anno.

Cliente vincolato

Cliente finale che, non rientrando nella categoria dei clienti idonei, può stipulare contratti di fornitura esclusivamente con il distributore che esercita il servizio nell'area territoriale dove è localizzata l'utenza.

Cliente grossista

Persona fisica o giuridica che acquista e vende energia elettrica senza esercitare attività di produzione, trasmissione e distribuzione nei Paesi dell'Unione Europea.

Le categorie sono due nel glossario del Gestore del Mercato Elettrico (GME) – (luglio 2008) <www.mercatoelettrico.org/It/Tools/Glossario.aspx> –, i *clienti idonei* e i *clienti vincolati*. Nella definizione diventano precisi i riferimenti giuridici: da un generico «a partire dal 1° maggio 2003», il cliente idoneo è individuato come persona fisica o giuridica che «a partire dal 1° luglio 2004, in base alla Delibera AEEG 107/04» si rende tale se acquista energia elettrica non destinata all'uso domestico, ma senza la precisione del quantitativo. La categoria è estesa a tutti dal 1° luglio 2007. Relativamente al cliente vincolato, si esplicita che si tratta dei clienti che acquistano energia solo per uso domestico, per il tramite dell'Acquirente Unico, e che non hanno quindi diritto ad acquistare energia direttamente in borsa se non dal 1° luglio 2007 – ma questa informazione si evince dalla definizione di "cliente idoneo" e non da quella di "cliente vincolato".

Cliente idoneo

Persona fisica o giuridica che ha facoltà di stipulare contratti di fornitura con qualsiasi fornitore di propria scelta (produttore, distributore, grossista). A partire dal 1° luglio 2004, in base alla Delibera AEEG 107/04, sono idonee tutte le persone fisiche o giuridiche che acquistano energia elettrica non destinata al proprio uso domestico, inclusi i produttori e i clienti grossisti. Dal 1° luglio 2007 tutti i clienti saranno idonei. A partire dal 1° gennaio 2005 i clienti idonei hanno diritto ad acquistare energia direttamente in borsa.

Cliente vincolato

Persona fisica o giuridica che, non rientrando nella categoria dei clienti idonei, può stipulare contratti di fornitura esclusivamente con il distributore che esercita il servizio nella propria area territoriale. Dal 1° luglio 2004 sono clienti vincolati i clienti che acquistano energia elettrica per il proprio consumo domestico, esclu-

se le attività commerciali o professionali. Gli acquisti dei clienti vincolati sono effettuati dall'Acquirente Unico <www.mercatoelettrico.org/It/Tools/Glossario.aspx#au%23au> sia in borsa sia tramite contratti bilaterali.

Non stiamo facendo uso di lessico poco trasparente, ma nuovi sono i contesti di applicazione e le caratteristiche funzionali di un soggetto tipicamente commerciale: le definizioni citate arricchiscono il carattere enciclopedico, che aiuta a riconoscere le idee chiave e portanti dell'argomento, con rimandi contestuali comunque utili per chi si accosta alla materia, e cercano di fissare l'attenzione del lettore prospettando una sintesi delle nozioni utili per passare alla costruzione di un insieme conoscitivo e operativo soddisfacente.

Rileviamo così l'importanza della diffusione assunta da questo lessico: le fonti di energia, la loro organizzazione di distribuzione, il loro valore commerciale diventano un prodotto di alto consumo e di alto impatto nella frequenza d'uso degli utenti, dei clienti e dei cittadini in generale.

2. Alcuni cenni allo sfondo legislativo

Occorrerebbe altresì ricordare lo sfondo giuridico che riunisce la legislazione in materia energetica. Le linee legislative illustrano gli obiettivi comunitari in materia di fonti di energia rinnovabile e le principali norme sulle fonti energetiche rinnovabili, sul risparmio energetico e sul libero mercato. Ricordiamo perciò che nel Libro Bianco *Una Politica Energetica per l'Unione Europea* - COM (1995) 682 Def., l'Unione Europea ha definito tre obiettivi per la propria politica energetica:

- la sicurezza negli approvvigionamenti, anche tramite la diversificazione;
- la competitività delle fonti;
- la tutela e il rispetto dell'ambiente.

La Commissione ha inoltre adottato il 26 novembre 1997 un Libro Bianco - COM (1997) 599 intitolato Energia per il futuro: le fonti energetiche rinnovabili. Libro bianco per una strategia e un piano di azione della Comunità.

Il protocollo di Kyoto del 1997 sui cambiamenti climatici ha rafforzato l'importanza della dimensione ambientale e dello sviluppo sostenibile nella politica energetica comunitaria, mentre la variabilità dei prezzi petroliferi osservati nell'ultimo decennio ha evidenziato i rischi per l'Unione Europea che derivano dalla sua dipendenza energetica dalle fonti fossili di altri Paesi. Seguono così le seguenti direttive ed indicazioni europee:

- la Direttiva 99/296/CE;
- *Libro verde sullo scambio dei diritti di emissione di gas ad effetto serra all'interno dell'Unione europea* (COM 2000 - 87);
- la comunicazione al Consiglio ed al Parlamento europeo *Verso un programma europeo per il cambiamento climatico (ECCP)*, che descrive le politiche e le misure dell'Unione europea per ridurre le emissioni di gas a effetto serra (COM 2000 - 88);

- la Direttiva 2003/87/CE del Parlamento Europeo e del Consiglio, che istituisce un sistema per lo scambio di quote di emissioni dei gas a effetto serra nella Comunità e che modifica la direttiva 96/61/CE del Consiglio;
- la promozione di diverse misure fiscali destinate alla protezione dell'ambiente. La proposta di Carbon Tax (COM - 92 - 226 e COM - 95 - 172) e i tentativi di armonizzazione delle accise sui prodotti energetici sono attualmente in fase di negoziazione e di accordo politico con gli stati membri (COM - 97 - 30);
- la Direttiva 2001/77/CE del Parlamento Europeo e del Consiglio, del 27 settembre 2001, sulla promozione dell'energia elettrica prodotta da fonti energetiche rinnovabili nel mercato interno dell'elettricità: la direttiva stabilisce che gli Stati membri adottino misure appropriate atte a promuovere l'aumento del consumo di elettricità prodotta da fonti rinnovabili perseguendo degli obiettivi indicativi nazionali rimodulati ogni due anni e compatibili con gli impegni nazionali assunti nell'ambito degli impegni sui cambiamenti climatici sottoscritti dalla Comunità ai sensi del protocollo di Kyoto.

Quest'ultima Direttiva è stata recepita in Italia con Decreto Legislativo del 29 Dicembre 2003 n. 387. Solo per dare una breve sintesi di sfondo di quanta scrittura legislativa, burocratica e amministrativa segue lo sviluppo e la diffusione delle fonti di energia rinnovabili.

L'apertura del mercato dell'energia è iniziata con le Direttive europee 96/92/CE e 98/30/CE, che stabilivano rispettivamente le regole comuni per il mercato interno dell'elettricità e del gas relativamente a produzione, trasporto e distribuzione, successivamente abrogate rispettivamente dalla Direttiva 2003/54/CE, relativa a norme comuni per il mercato interno dell'energia elettrica, e dalla Direttiva 2003/55/CE, relativa a norme comuni per il mercato interno del gas naturale.

Le principali norme italiane sulle fonti energetiche rinnovabili sul risparmio energetico e sul libero mercato sono le seguenti:

- la Legge 9 gennaio 1991, n. 9, *Norme per l'attuazione del nuovo Piano Energetico Nazionale: aspetti istituzionali, centrali idroelettriche ed elettrodotti, idrocarburi e geotermia, autoproduzione e disposizioni fiscali*, che ha introdotto l'aspetto significativo della parziale liberalizzazione della produzione dell'energia elettrica da fonti rinnovabili e assimilate;
- il provvedimento n. 6 del 1992, detto anche "CIP 6", ritirato nel 1996;
- la Legge 9 gennaio 1991, n. 10, *Norme per l'attuazione del Piano Energetico Nazionale in materia di uso razionale dell'energia, di risparmio energetico e di sviluppo delle fonti rinnovabili di energia*, in sostituzione della Legge 308/86 (nel Titolo I reca norme in materia di uso razionale dell'energia, di risparmio energetico e di sviluppo delle fonti di energia);

- il D.P.R. 26 agosto 1993, n. 412, *Regolamento recante norme per la progettazione, l'installazione, l'esercizio e la manutenzione degli impianti termici degli edifici ai fini del contenimento dei consumi di energia*, in attuazione dell'articolo 4/IV della Legge 9 gennaio 1991, n. 10, poi modificato ed integrato dal D.P.R. 21 dicembre 1999, n. 551, *Regolamento recante modifiche al Decreto del Presidente della Repubblica 26 agosto 1993, n. 412, in materia di progettazione, installazione, esercizio e manutenzione degli impianti termici degli edifici, ai fini del contenimento dei consumi di energia*, che ha introdotto norme precise sui rendimenti degli impianti termici nonché sulle modalità di controllo e verifica da parte delle Province e dei Comuni.

Ricordiamo infine che il riassetto del sistema elettrico italiano e la sua liberalizzazione è stato avviato dal Decreto Legislativo 16 Marzo 1999, n. 79, *Attuazione della Direttiva 96/92/CE, recante norme comuni per il mercato interno dell'energia elettrica* (detto anche Decreto Bersani).

L'attuazione di questa legislazione è spesso condizionata dall'emanazione di un numero considerevole di decreti successivi, che non stiamo a riportare in questa sede.

Segnaliamo, invece, l'entrata nel mercato di nuovi operatori e di altri interlocutori oltre all'Enel, quali:

a) l'AEEG (Autorità per l'Energia Elettrica ed il Gas), che fissa le condizioni atte a garantire l'imparzialità e la neutralità del servizio di trasmissione e dispacciamento e può autorizzare la costituzione di contratti bilaterali, in deroga al mercato elettrico, sulla base di criteri oggettivi, trasparenti e non discriminatori;
b) il GRTN (Gestore della rete di Trasmissione Nazionale), che esercita attività di trasmissione e dispacciamento dell'energia elettrica e che con proprie delibere fissa le regole del dispacciamento;
c) il GME (Gestore del Mercato Elettrico), che assume la gestione delle offerte di vendita e acquisto dell'energia elettrica e di tutti i servizi connessi;
d) l'AU (Acquirente Unico), che deve garantire, per i clienti vincolati, la fornitura dell'energia elettrica, la gestione dei relativi contratti e la tariffa unica a livello nazionale.

È tuttavia necessario avere sotto gli occhi la complessità disciplinare che l'utente si trova ad affrontare, per capire con quali modalità le istituzioni pubbliche e le società private hanno inteso offrire un aiuto alla comprensione terminologica di questa massa di dati e di conoscenze. Ampia la fioritura di glossari e di legende terminologiche, disponibili nei siti delle istituzioni che trattano di energia. Ne daremo illustrazione, limitandoci ad un commento sulle principali tipologie e caratteristiche.

3. Borsa e mercato elettrico: i glossari presso le istituzioni

Il settore del mercato elettrico vanta i glossari più rigorosi ed accurati, a cominciare dal *National Grid*, che presenta la seguente introduzione al glossario posto nel sito - (luglio 2008) <www.nationalgrid.com/uk/Electricity/SYS/glossary>:

> «This Glossary defines and explains some of the key terms and phrases used in the Statement. Other documentation in current use within the Electricity Supply Industry (ESI) in Great Britain (e.g. the GB Grid Code and the GB Transmission System Quality and Security of Supply Standard) may contain similar terms and phrases. Where such terms and phrases carry the same meaning in this Statement, the same definition has been used for consistency. Where the meaning differs in some respect, a modified or new definition, as appropriate, is used. In all cases, for the purpose of this Statement, the definitions set out in this Glossary take precedence».

È così precisata la fonte di provenienza del lessico repertoriato, che funge da illustrazione e raccoglitore di termini ed espressioni usati nello *Statement*, oltre che da fonte anteriore attestata.

Il Glossario presente nel sito del Gestore del Mercato Elettrico – (luglio 2008, <www.mercatoelettrico.org> – non dà invece esplicita indicazione delle fonti, ed è una versione arricchita e completata rispetto ad una precedente del 2006, pubblicata nel *Vademecum della Borsa elettrica italiana,* che riportava i riferimenti alla normativa e alla manualistica di riferimento.

Il Glossario del 2006 è costituito da 61 voci, mentre l'attuale versione on-line ne registra 99. L'attuale glossario on-line ha apportato aggiornamento delle informazioni, e qualche modifica alle voci, e segnala con rimandi ipertestuali nelle definizioni alle voci presenti nel Glossario, come di seguito elenchiamo.

a) È cambiato il nome del Gestore, da GRTN (Gestore della Rete di Trasmissione Nazionale) in GSE (Gestore dei Servizi Elettrici):

> Il Gestore dei Servizi Elettrici - GSE S.p.A. ha un ruolo centrale nella promozione, nell'incentivazione e nello sviluppo delle fonti rinnovabili in Italia. Azionista unico del GSE è il Ministero dell'Economia e delle Finanze che esercita i diritti dell'azionista con il Ministero delle Attività Produttive. Il GSE è capogruppo delle due società controllate AU (Acquirente Unico) e GME (Gestore del Mercato Elettrico).
>
> In seguito al trasferimento del ramo d'azienda relativo a dispacciamento, trasmissione e sviluppo della rete a Terna S.p.A, avvenuto il 1° novembre 2005 per effetto del DPCM dell'11 maggio 2004, il GSE si concentra sulla gestione, promozione e incentivazione delle fonti rinnovabili in Italia, attività in parte già svolte.

Il Gestore dei Servizi Elettrici - GSE S.p.A. svolge un ruolo fondamentale nel meccanismo di incentivazione della produzione di energia da fonti rinnovabili e assimilate, predisposto dal provvedimento CIP 6/92, e a gestire il sistema di mercato basato sui Certificati Verdi.

La sostituzione di GRTN con Terna s.p.a., o GSE, dove il caso, ricorre in tutto il testo del glossario.

b) Nel Glossario 2006 si parlava di *CVE (Coefficiente di Variazione Elettrico)*, definito come il "Rapporto tra la deviazione standard e il valore assoluto della media dei prezzi orari espresso in termini percentuali". Nell'attuale glossario si semplifica la polirematica in *Coefficiente di variazione* e la sigla attribuita in *CV*:

Indice di volatilità dei prezzi calcolato come rapporto tra deviazione standard e il valore assoluto della media dei prezzi orari espresso in termini percentuali. È un indice di dispersione relativo che permette di confrontare fenomeni con numerosità e unità di misura differenti, in quanto si tratta di un numero puro (ovvero senza unità di misura).

c) Sono registrate alcune varianti:

a) per diverso uso della preposizione*: esiti del mercato* (2006) si stabilizza nella variante grafica *esiti di mercato*;

b) per adattamenti morfologici: *fascia oraria* prende il plurale *fasce orarie*.

d) La voce Indice di operatore residuale (IOR) diviene più articolata:

È un indice relativo ai singoli operatori che offrono sul mercato e misura la presenza di operatori di mercato residuali, vale a dire necessari al fine del soddisfacimento della domanda. È definito, per ciascun operatore, [Esso viene calcolato per le singole zone geografiche e per ciascun utente del dispacciamento (UdD)] come rapporto tra le quantità complessivamente offerta dai concorrenti [dagli altri UdD]e la quantità complessivamente venduta [all'interno della zona]. L'indice assume valore <1 in presenza di un operatore residuale *e tanto più è prossimo allo 0 tanto maggiore è la quota della sua offerta che può essere venduta a prescindere dal prezzo di offerta. Lo IOR viene calcolato aggregando le quantità offerte dai singoli operatori, raggruppati sulla base dell'appartenenza di gruppo, ivi incluse le quantità oggetto di contratti bilaterali: le quantità relative a contratti CIP6 sono incluse in questo calcolo e sono assegnate all'operatore GSE. L'utilizzo della quantità accettata al denominatore consente di scontare l'effetto sulla domanda interna ad ogni zona dei transiti con le zone limitrofe. Mensilmente vengono pubblicati, per ogni macrozona: la percentuale di ore in cui c'è stato almeno un operatore necessario; la percentuale dell'energia venduta in condizioni di residualità sull'energia complessivamente venduta, pari alla media semplice delle quantità residuali orarie della macrozona (definite a loro volta come somma, su tutti gli*

operatori, della quantità offerta da ciascuno meno la quantità complessivamente offerta più la quantità complessivamente venduta); il numero di operatori necessari e la percentuale di ore per cui sono stati necessari.

La parte in corsivo era formulata solo come segue:

> Viene pubblicato, per ogni mese e per ogni zona, il numero di operatori necessari e la percentuale di ore per cui sono stati necessari. L'utilizzo della quantità accettata Al denominatore consente di scontare l'effetto sulla domanda interna ad ogni zona dei transiti con le zone limitrofe.

e) La voce *liquidità* si riduce nel sito alla definizione di "Rapporto tra le quantità di borsa <www.mercatoelettrico.org/It/Tools/Glossario.aspx#quantità%23quantità> e le quantità totali" <www.mercatoelettrico.org/It/Tools/Glossario.aspx#quantità%23quantità>, mentre nel 2006 è accompagnata da una glossa ulteriore:

> Rapporto tra le quantità di borsa e le quantità totali sul Mercato del Giorno Prima. Tale quota rappresenta il grado di utilizzo della borsa elettrica per l'approvvigionamento dell'energia rispetto al totale dei consumi nazionali.

f) Non compaiono più le voci *aggregato Borsa, aggregato Borsa + Bilaterali*, *indici di operatore marginale* (IOM), mentre sono inserite voci nuove, o perché trattate autonomamente rispetto ad un riferimento che era dato precedentemente all'interno di altri termini, o perché novità tematiche sopraggiunte.

g) La rassegna descrittiva delle offerte si articola nell'ingresso generico, poi trattato nel dettaglio:

Offerte

Sui mercati del GME possono essere presentate quattro tipologie di offerta.

a) *Offerte semplici*: sono offerte costituite da una coppia di quantità (espressa in MWh) e prezzo (espresso in €/MWh), dove la quantità rappresenta la massima disponibilità ad immettere o prelevare energia ed il prezzo rappresenta il prezzo massimo di acquisto o il prezzo minimo di vendita richiesto. Nel caso delle offerte di acquisto il prezzo "0" indica la disponibilità ad acquistare a qualunque prezzo.
b) *Offerte multiple*: sono offerte costituite al massimo da quattro coppie di offerte semplici. Ciascuna offerta costituisce a tutti gli effetti un'offerta distinta, tuttavia le quantità indicate nelle diverse coppie devono rispettare congiuntamente i margini a salire e a scendere dichiarati per ciascuna unità di produzione dai relativi utenti del dispacciamento.
c) *Offerte predefinite*: sono offerte che possono essere presentate in qualunque momento e che vengono utilizzate dal sistema informatico del mercato elettrico (SIME) ogni volta che non vengono presentate offerte valide più recenti. Le offerte predefinite possono essere presentate solo su MGP e MSD.
d) *Offerte bilanciate*: sono gruppi di due o più offerte, di cui almeno una di

acquisto ed una di vendita, riferite ad una stessa zona geografica e ad una stessa ora, complessivamente bilanciate in quantità ed aventi prezzo zero (sia in vendita che in acquisto). Tali offerte possono essere presentate solo sul MA, da uno o più operatori, e godono di massima priorità.

Offerta integrativa

Offerta di acquisto o di vendita presentata da Terna S.p.A. sul MGP, ai sensi dell'articolo 48 della delibera 168/03 e successive modifiche ed integrazioni, finalizzata a compensare la differenza tra ammontare del fabbisogno orario stimato da Terna S.p.A. per ciascuna zona per il giorno successivo e ammontare complessivo delle offerte di acquisto e dei programmi bilaterali di prelievo presentati sul MGP.

Offerta marginale

Per offerta marginale si intende, in ogni zona di mercato e in ogni ora, l'offerta accettata con il più alto ordine di merito.

h) Sono introdotte nuove voci, quali:

Piattaforma di Aggiustamento Bilaterale per la domanda (PAB)

La PAB – Piattaforma di Aggiustamento Bilaterale per la Domanda – è una piattaforma informatica, operativa dal 31 dicembre 2004, che consente la registrazione di scambi orari bilanciati di energia elettrica tra gli operatori che gestiscono i punti di offerta in prelievo appartenenti alla stessa zona geografica seguendo le disposizioni contenute nell'apposito Regolamento. Il GME verifica il rispetto di tale Regolamento e delle Disposizioni Tecniche di Funzionamento al fine di assicurare il regolare funzionamento della PAB secondo i criteri di neutralità, trasparenza, obiettività e concorrenza tra gli operatori. Gli scambi comunicati al GME tramite tale piattaforma, insieme agli impegni derivanti da contratti bilaterali o da acquisti sul mercato elettrico, determinano il programma vincolante di ciascun punto di offerta in prelievo.

Polo di produzione limitato

Insieme di unità di produzione connesse ad una porzione della RTN senza punti di prelievo, la cui produzione massima esportabile verso la restante parte della RTN è inferiore alla produzione massima possibile a causa di insufficiente capacità di trasporto.

4. Glossari di società italiane e estere

Illustriamo qualche esempio dei caratteri e dei contenuti dei numerosi glossari disponibili in rete nei siti di società italiane ed estere, che operano nei settori dell'energia. La descrizione dei *Certificati verdi,* introdotti nel 1999, è presente nei glossari delle sole società italiane, trattandosi di una realtà solo nazionale. Scorriamo gli esiti delle diverse definizioni:

Glossario del GME (luglio 2008) <www.mercatoelettrico.org/It/Tools/Glossario.aspx>

> Certificati che, ai sensi dell'art. 5 del Decreto del Ministro dell'Industria 11/11/99, attestano la produzione di energia da fonte rinnovabile al cui obbligo sono tenuti produttori ed importatori di energia elettrica da fonti non rinnovabili per una quantità superiore ai 100 GWh/anno. I Certificati Verdi sono emessi dal GSE <www.mercatoelettrico.org/It/Tools/Glossario.aspx#GSE%23GSE> per i primi otto anni di esercizio di un impianto ed hanno un valore pari a 50 MWh e possono essere venduti o acquistati sul Mercato dei Certificati Verdi (MCV) <www.mercatoelettrico.org/It/Tools/Glossario.aspx#mcv%23mcv> dai soggetti con eccessi o *deficit* di produzione da fonti rinnovabili.

Glossario del GSE (luglio 2008) <www.grtn.it/ita/glossario/glossario.asp>

> I Certificati Verdi sono titoli annuali emessi dal GSE che attestano la produzione da fonti rinnovabili di 50 MWh di energia. A partire dal 2002, in base al decreto 79/99, produttori e importatori hanno l'obbligo di immettere in rete energia da fonti rinnovabili, in quantità pari al 2% del totale dell'elettricità prodotta o importata l'anno precedente da fonti convenzionali (al netto di esportazioni, autoconsumi di centrale e cogenerazione).

Le due definizioni enciclopediche variano nell'ordine della descrizione che tocca l'illustrazione del ruolo dei Certificati verdi e delle loro caratteristiche, nonché nel dettaglio descrittivo dei riferimenti legislativi: nel glossario del GME la loro creazione è riportata nell'art. 5 del Decreto del Ministro dell'Industria 11/11/99, mentre in quello del GSE non si menziona il decreto, ma direttamente la fonte di emissione.

Proseguiamo nella lettura delle definizioni, riportate nei glossari di siti di società locali:

Le parole dell'energia: dall'adduzione alle zone di mercato
(luglio 2008) <www.edipower.it/energia/parole.asp>

> Sono una certificazione di produzione (della durata di 8 anni) che il Gestore della Rete di Trasmissione Nazionale (GRTN) emette a favore dei produttori di energia rinnovabile. Definiscono la quantità di energia rinnovabile prodotta da ciascun impianto. I produttori e gli importatori di energia elettrica hanno l'obbligo di immettere nel sistema elettrico nazionale, una quota di energia prodotta da impianti da fonti rinnovabili pari al 2% della loro produzione o importazione termoelettrica. Ciò può avvenire costruendo e gestendo impianti alimentati

da fonti rinnovabili oppure acquistando "certificati verdi" dalle aziende che ne dispongono. Il prezzo dei certificati è determinato dall'incontro tra la domanda e l'offerta degli stessi.

Glossario di EGEA S.p.A.
<www.egea.it/commerciale/strumenti/glossario.php>

Titolo annuale, oggetto di contrattazione nell'ambito della Borsa dell'Energia, che verrà attribuito dal Gestore della Rete di Trasmissione Nazionale (GRTN) all'energia elettrica prodotta mediante l'uso di fonti energetiche rinnovabili, attraverso impianti entrati in esercizio dopo il 1° Aprile 1999, per i primi otto anni di esercizio degli stessi. Il titolo è previsto dal Decreto Bersani quale strumento alternativo per soddisfare l'obbligo, imposto a decorrere dal 2002 ad ogni produttore/importatore di energia, di immettere in rete una quota minima di energia verde pari al 2% dell'energia non rinnovabile prodotta/importata nell'anno precedente. L'offerta di certificati verdi potrà pervenire da due categorie di soggetti: i produttori (nazionali ed esteri) e, per la parte di domanda non soddisfatta da questi ultimi, dal Gestore della Rete di Trasmissione Nazionale (GRTN).

Glossario di Hera S.p.A.
<www.gruppohera.it/energia/?sub=contesto_elettricita_glossario#protocollo_kyoto>

Titoli che attestano la produzione di energia da fonti rinnovabili, imposta in una certa percentuale minima per legge ai soggetti che importano o producono energia da fonti convenzionali (gas, petrolio, gas naturale) oltre una certa soglia. I Certificati verdi hanno sostituito il sistema precedente di incentivi denominato Cip 6 nato nel 1992. Anche per questi certificati esiste un mercato specifico, il Mercato dei certificati verdi.

Il contenuto del glossario delle società private o con partecipazione statale è vario e mescola nei termini selezionati l'insieme dei livelli disciplinari implicati, scientifico e tecnologico, economico-finanziario, politico-istituzionale, giuridico-amministrativo.

Si noti cosa ritroviamo nella lettera B del Glossario AMGA (Trieste) <www.amgaenergiaeservizi.com/CMS/main.php?elemId=113&classId=5>:

Bassa tensione

Tensione non superiore a 1000 V in corrente alternata e a 1500 V in corrente continua.

Bersani, Decreto

Decreto Legislativo n. 79 del 19 marzo 1999 (in vigore dal 1° aprile successivo), in recepimento della Direttiva Comunitaria 96/92/CE. Introduce la liberalizzazione nelle attività di produzione, importazione e vendita di elettricità ai clienti idonei. Scopo del decreto è creare un sistema di libera concorrenza, pur regolamentato da precise norme di tutela del consumatore finale, nel rispetto del principio di pubblica utilità dell'energia elettrica.

Bilanciamento

Attività finalizzata a mantenere l'equilibrio fra immissioni e prelievi di energia elettrica sulla Rete di Trasmissione Nazionale.

Un portale dell'installazione intitola così un suo "glossario", *Per capire la liberalizzazione*:

> Una volta bastava conoscere l'Enel, il contatore e il kilowattora, per potersi agevolmente destreggiare nel mondo del mercato elettrico nazionale. Oggi tutto questo non è più sufficiente, e a questi termini se ne sono aggiunti di nuovi che testimoniano un cambiamento epocale nella distribuzione elettrica.

Riporta di seguito la definizione, rilevata da "Repubblica" di *GRTN*, *AU*, *GME*, *cliente idoneo, cliente vincolato, AEEG.*

Conclusione

La terminologia delle energie rinnovabili si caratterizza per un lessico basato su costruzioni sintagmatiche, che sono spesso riportate ad acronimi, nonché sulla rideterminazione semantica di sostantivi di uso comune, e gode di una discreta variabilità sinonimica al suo interno. Si tratta di una terminologia aperta alle nuove designazioni, ma soprattutto rispondente alle esigenze della varietà di utenti, che accedono a questo campo della conoscenza per vie diverse: per le necessità della vita quotidiana, per la via giuridica e amministrativa, per la via scientifica e tecnologica, per la via economico-finanziaria, secondo gradi di specializzazione in ognuna delle discipline e degli ambiti citati.

La passione – o la moda – del glossario esplicativo è ben presente nella storia della divulgazione scientifica. Basti ricordare le paginette di Palissy a chiosa dei suoi *Discours admirables* (1580), nelle quali l'autore tratteggia con 62 voci i termini più difficili o più significativi della sua argomentazione di studio: l'*Explication des mots les plus difficiles* è una vera e propria guida nel campo della paleontologia, dell'idrologia, della geologia, della fisiologia vegetale (Zanola, 2007).

I professionisti stessi del settore si rendono conto dell'importanza della terminologia nella comunicazione, e lanciano la diffusione di glossari – designazione con cui sono chiamate queste raccolte di termini chiave – per una comunicazione non più tra esperti, per i quali la terminologia settoriale è un sapere condiviso. L'intento di questi glossari è quello di pensare alla divulgazione dei contenuti oggetto di interesse per capire le azioni dell'istituzione, per conoscere i prodotti energetici, per rassicurare della correttezza nell'informazione chi si accosta per la prima volta a queste conoscenze.

Ci chiediamo nei casi osservati quale sia il ruolo della terminologia in rapporto agli scopi della comunicazione. La diversità dei contesti di comunicazione può portare alla variazione delle rese espressive e alle conseguenti differenze nella scelta terminologica, così come è possibile osservare, nella diversificazione che la terminologia assume nel settore delle energie rinnovabili, lo sviluppo di filoni terminologici in funzione dei settori disciplinari cui si riferiscono. Mentre scriviamo, si riapre la possibilità dell'apertura all'energia nucleare, e si riversano nella stampa e nei programmi televisivi nuove argomentazioni che, dopo anni di silenzio in merito, riaprono una possibilità energetica considerata impensabile fin qui come energia di sicura affidabilità. Il lavoro preparatorio per la costituzione di lessici e/glossari delle energie pulite offrirà così contenuti esposti dalle voci e dalle loro definizioni che si riveleranno testimoni del tempo.

L'intento comunicativo di un glossario, le sue ragioni, i destinatari per i quali è pensato, ricordano che ogni scelta terminologica non vive in sé, ma nella storia e nella rete concettuale di riferimento di una tecnica, di una disciplina, di un campo del sapere e della realtà. Corrisponde perciò ad un documento che descrive e rende conto di uno stato linguistico che vive nello spazio e nel tempo della lingua che lo esprime.

Bibliografia

Silvi Cesare, (2005). *Il linguaggio dell'energia*, "Fotovoltaici", 5, p. 58-59.

Gestore del mercato elettronico, (2006). *Vademecum della Borsa elettrica italiana*, <www.mercatoelettrico.org/It/GME/Biblioteca/pubblicazioni.aspx>.

Zanola Maria Teresa, (2007), *Synonymie et vulgarisation scientifique: l'Explication des Mots plus Difficiles dans les 'Discours Admirables' de Bernard Palissy (1580)*, Colloque "La Synonymie", Université de la Sorbonne, Paris 29 Novembre-1 Dicembre 2007, Presses de l'Université de la Sorbonne, Paris (c.d.s.).

L'informazione al consumatore: la terminologia delle fonti energetiche e le variazioni negli usi testuali

SONIA PIOTTI

The present article investigates the strategies of information regarding energy services and their promotion in Italy and Great Britain. The article is a comparative study of texts addressing the general public especially, on the grounds that the layman needs to develop specialised knowledge in order to be able both to choose among and properly evaluate the range of offers, services and goods available. As some difficulties may arise from differences in the terminology used in the field of energy services, an overview of the types of terminological databases available is proposed.

Keywords: Lexicography – terminology – lexicology – energy services

Il presente lavoro si propone di osservare e analizzare le strategie di presentazione dell'informazione, di promozione e incentivazione delle fonti energetiche rinnovabili in prospettiva contrastiva fra Italia e Regno Unito, fra lingua italiana e lingua inglese. Particolare rilievo viene dato alla terminologia relativa a tali fonti energetiche e alle variazioni negli usi testuali, ripercorrendone il "viaggio" a partire dalle attestazioni interne alla comunicazione sia degli organismi istituzionali dell'Unione Europea sia delle pubbliche amministrazioni dei due stati comunitari, per giungere alle attestazioni negli scambi comunicativi tra istituzioni e cittadini e, infine, agli usi negli scambi comunicativi tra mezzi di informazione e cittadini.

Nel corso della seconda metà del 2006 e durante i primi mesi del 2007, in particolare, le tensioni geo-politiche internazionali e le nuove evidenze sul tema del cambiamento climatico hanno riportato alla ribalta, anche nell'agenda politica delle istituzioni europee, i temi dell'energia, e in particolare delle fonti energetiche *rinnovabili* o *alternative* a quelle tradizionali rappresentate da petrolio, carbone e gas naturale. Se ne sono occupati e continuano ad occuparsene organi politici, legislativi e amministrativi, le figure professionali che operano nel mercato energetico – produttori, trasportatori e distributori (vedi Allegato 1) –, studiosi e specialisti del settore energetico, scientifico e tecnologico, i mercati finanziari, gli organi di divulgazione di massa, con riferimento ad ampiezze concettuali diverse legate necessariamente ai diversi gradi di specializzazione. All'interno di questo scenario, il cittadino o consumatore finale rappresenta l'ultimo anello della catena e accede a questi temi principalmente per via delle necessità della vita quotidiana.

È soprattutto sulla categoria dei cittadini consumatori che si focalizza il presente lavoro, tenendo conto della complessità dei campi tematici, delle competenze tecniche, delle informazioni e della terminologia che l'utente si trova sempre più spesso a dovere acquisire e riconoscere per potere essere in grado di valutare i prodotti delle nuove tecnologie nonché i prodotti e i servizi offerti dal mercato.

Numerose sono le difficoltà derivate dalle variazioni terminologiche relative alle fonti energetiche osservate nei diversi testi, soprattutto in quelli maggiormente esposti e visibili dal punto di vista sociale e sociolinguistico. Al fine di verificare il grado di accesso a tali informazioni da parte del consumatore, il contributo si propone di osservare in che grado e modo quest'ampia gamma di variazioni terminologiche è presente o lemmatizzata nei repertori terminologici in essere in ambito energetico o nei repertori lessicografici.

1. L'informazione al consumatore: la complessità dei campi concettuali

Per tutelare attivamente gli interessi dei consumatori, affinché possano sfruttare appieno i vantaggi dell'apertura dei mercati energetici dopo la liberalizzazione del 1° luglio 2007, la Commissione Europea ha redatto una *Carta europea dei consumatori di energia*, che prevede alcuni diritti fondamentali: il diritto ad avere informazioni aggiornate sulla fornitura di energia, le condizioni contrattuali, i prezzi e le tariffe, le misure di efficienza energetica, nonché l'origine e le fonti di produzione dell'energia, quella elettrica in particolare.

Analogamente, allo scopo di promuovere una migliore conoscenza delle nuove opportunità legate all'apertura del mercato europeo dell'energia, l'Autorità per l'Energia Elettrica e il Gas (AEEG) e la Rappresentanza in Italia della Commissione Europea hanno promosso nel 2008 la pubblicazione dell'opuscolo *Energia semplice. Le novità nel mercato dell'energia e del gas* <www.autorita.energia.it/energia_semplice/opuscolo.pdf>, in cui il Commissario Europeo per l'Energia afferma:

> Garantire ai 480 milioni di consumatori dell'Unione Europea il diritto ad un approvvigionamento energetico sicuro ed a prezzi convenienti: è questo l'obiettivo della liberalizzazione del mercato dell'energia elettrica e del gas voluto dalla Commissione, con l'approvazione delle direttive n. 2003/54/CE e 2003/55/CE.
>
> Oggi tutti i consumatori sono diventati protagonisti attivi sul mercato dell'energia […] e possono scegliere in base a diverse offerte, alla qualità del servizio, eventualmente anche contribuendo a difendere l'ambiente comprando solo energia rinnovabile […].

Alle parole del Commissario Europeo, fanno eco le affermazioni del Presidente dell'Autorità per l'Energia Elettrica e il Gas:

> […] L'Autorità per l'energia elettrica e il gas è impegnata a far sì che […] i consumatori possano decidere sempre più liberamente, consapevolmente e convenientemente, in uno scenario di vera concorrenza ed efficienza di mercato [che] si tradurrà in benefici per i cittadini consumatori, come prezzi e qualità di servizi o forniture. […] Il primo vantaggio che i cittadini stanno toccando con mano, si radica nel cuore stesso della democrazia, ed è la libertà di scelta reale per tutti i consumatori, che […] possono cambiare fornitore se riscontrano un servizio di cattiva qualità, se considerano eccessivo il prezzo offerto, se vogliono partecipare alla lotta contro i cambiamenti climatici, scegliendo energia da fonti rinnovabili oppure a basso tenore di carbonio.
>
> *Questa guida "Energia semplice" vuol contribuire ad un'informazione più completa; aiutare il cittadino-consumatore a conoscere più approfonditamente le "liberalizzazioni" per il settore energetico e le opportunità che esse offrono, ad orientarsi nel nuovo scenario di mercato con una maggiore consapevolezza dei propri diritti, delle tutele previste e degli strumenti a disposizione.*

Risulta fondamentale per il consumatore il diritto ad avere informazioni in vari ambiti disciplinari: scientifico (fotovoltaico, elettrico, minerario), tecnologico (conoscenza di servizi e prodotti in commercio, offerte, tariffe, materiali e loro rendimento), ambientale (difesa dell'ambiente e lotta contro i cambiamenti climatici), economico-finanziario, politico-istituzionale, per giungere a informazioni di natura giuridica (consapevolezza dei propri diritti e delle tutele previste) (vedi Allegato 1). Tale ampiezza di saperi trova riscontro in alcuni glossari e repertori terminologici in essere, soprattutto quelli di società private o con partecipazione statale, che spesso includono e mescolano tra le voci i diversi ambiti disciplinari (Zanola 2008).

2. Le fonti energetiche: terminologia e variazioni negli usi testuali

«Dinnanzi al diffondersi sempre crescente del dibattito intorno alle energie, aumenta anche l'ampiezza descrittiva del fenomeno, vuoi per divulgazione scientifica specialistica, vuoi per divulgazione istituzionale di massa», osserva Zanola (2008). Numerose sono le tipologie testuali a cui ricorre il dibattito intorno all'energia: esse sono il risultato di diversità di bisogni comunicativi e situazioni espressive per un'ampia gamma di questioni tematiche. L'utilizzo strategico delle fonti energetiche, le nuove evidenze del cambiamento climatico, nonché la sostenibilità ambientale delle scelte energetiche, sono tematiche tipiche dei testi normativi e degli scambi comunicativi tra istituzioni e cittadini, per arrivare ai risvolti economico-finanziari e alle necessità pratiche della vita quotidiana che caratterizzano principalmente lo scambio informativo tra mezzi di comunicazione e cittadini.

Ai fini del presente lavoro si è raccolto un *corpus* documentale costituito da testi normativi politici e giuridici, sia a livello comunitario sia di ciascuno dei due stati membri, unitamente a testi informativi (Sabatini, 1999), in entrambe le lingue, inglese e italiano.

2.1. Terminologia in uso nei testi normativi comunitari e dei singoli stati membri

Fonti rinnovabili, alternative o energie pulite? Relativamente alle fonti energetiche, in tutte le direttive e discipline comunitarie recanti norme per il mercato interno dell'energia elettrica o relative alla promozione di energia elettrica prodotta da fonti energetiche non convenzionali, è sempre presente una sezione dedicata alle definizioni: la polirematica *fonti di energia rinnovabili* è l'unica espressione lemmatizzata. È significativo osservare, tuttavia, che il cotesto di tali direttive o discipline permette successivamente di recuperare e stabilire una rete di relazioni lessico-semantiche di sinonimia e di antonimia tra *fonti di energia rinnovabili* (*renewable energy sources*) e altre espressioni polirematiche. L'articolo 2. della *Direttiva 96/92/CE recante norme comuni per il mercato interno dell'energia elettrica* e la *Disciplina comunitaria degli aiuti di Stato per la tutela ambientale, 2008/C 82/01* sono riportati di seguito in quanto rappresentativi di questo fenomeno:

Direttiva 96/92/CE recante norme comuni per il mercato interno dell'energia elettrica	***Community Guidelines on the Internal Market of Electricity, 96/92/CE***
2.2. Definizioni Ai fini della presente disciplina, si applicano le seguenti definizioni: […] fonti di energia rinnovabili: le seguenti fonti energetiche rinnovabili non fossili: *energia eolica, solare, geotermica, del moto ondoso, maremotrice, delle centrali idroelettriche, energia derivata da biomasse, da gas di discarica, da gas residuati dai processi di depurazione e da biogas;* […].	2.2. Definitions For the purpose of these Guidelines the following definitions shall apply: […] renewable energy sources means the following renewable non-fossil energy sources: *wind, solar, geothermal, wave, tidal, hydropower installations, biomass, landfill gas, sewage treatment plant gas and biogases;* […].

Disciplina comunitaria degli aiuti di Stato per la tutela ambientale, 2008/C 82/01	***Community Guidelines on State Aid for Environmental Protection, 2008/C 82/01***
[…] Gli Stati membri possono concedere aiuti al funzionamento a favore di nuovi impianti di energia rinnovabile calcolati sulla base di un calcolo dei costi esterni evitati. […] I costi esterni evitati sono una quantificazione monetaria dell'aggiuntivo danno socio-ambientale che la società subirebbe se la stessa quantità di energia fosse prodotta da un impianto funzionante con energie convenzionali. Tali costi sono calcolati sulla base della differenza tra i costi esterni generati ma non pagati dai produttori di energia rinnovabile e i costi esterni generati ma non pagati dai produttori di energie non rinnovabili.	[…] Member States may grant operating aid to new plants producing renewable energy on the basis of a calculation of the external costs avoided. […] The external costs avoided represent a monetary quantification of the additional socio-environmental damage that society would experience if the same quantity of energy were produced by a production plant operating with conventional forms of energy. They will be calculated on the basis of the difference between, on the one hand, the external costs produced and not paid by renewable energy producers and, on the other hand, the external costs produced and not paid by non-renewable energy producers.

Sono varianti sinonimiche di fonti di energia rinnovabili (renewable energy sources) le seguenti espressioni: fonti energetiche rinnovabili non fossili (renewable non-fossil energy sources) ed energia rinnovabile (renewable energy).

Sono antonimi di *fonti di energia rinnovabili* (*renewable energy sources*) le polirematiche energie non rinnovabili (*non-renewable energy*) ed energie convenzionali (*conventional forms of energy*).

All'interno di ciascuna categoria di relazione lessico-semantica, sinonimia o antonimia, la natura dell'ampia gamma di varianti rilevate è dovuta a processi di natura morfologica quali:

- ellissi e sostituzione della testa del sintagma: energia rinnovabile per fonti di energia rinnovabile (*renewable energy per renewable energy sources*);
- sostituzione del determinante: energie convenzionali (*conventional forms of energy*) per energie non rinnovabili (*non-renewable energy*);
- espansione tautologica del determinante: fonti di energia rinnovabili (*renewable energy sources*) per fonti di energia rinnovabili non fossili (*renewable non fossil energy sources*).

Il Parere del Comitato economico e sociale europeo sul tema *Efficienza energetica*, 2006/C 88/13 (G.U. C 88 dell'11.4.2006), permette di stabilire ulteriori relazioni di sinonimia tra la famiglia lessicale delle fonti rinnovabili e quella delle alternative:

Parere del Comitato economico e sociale europeo sul tema Efficienza energetica, 2006/C 88/13	***Opinion of the European Economic and Social Committee on Energy Efficiency, 2006/C 88/13***
[…] Nel campo della produzione vengono continuamente introdotti dispositivi atti ad aumentare l'efficienza delle modalità di produzione. Ad esempio, la cogenerazione di calore ed elettricità si basa proprio sul principio del recupero di un'energia che altrimenti verrebbe sprecata. Esistono però anche delle nuove tecnologie che consentono di utilizzare fonti di energia alternativa.	*As far as production is concerned, efficiency gains are being regularly introduced into production methods. Accordingly, the cogeneration of heat and electricity seeks to recover energy that would otherwise have been wasted; new technologies are also applied, allowing sources of alternative energy to be used.*

Nel corso del 2006 e nei primi mesi del 2007, un utilizzo strategico delle fonti energetiche da parte di Paesi produttori, unitamente alle nuove evidenze del cambiamento climatico hanno ridato priorità anche ai temi della sostenibilità ambientale delle scelte energetiche. Nelle direttive che disciplinano in merito a questa tematica compare anche la famiglia lessicale delle energie pulite: la polirematica è sempre usata in costruzioni binomiali unitamente a rinnovabili, come nel caso di energie pulite e rinnovabili (*clean and renewable energy*), a sottolineare maggiormente la natura "eco-sostenibile" delle nuove forme di produzione energetica.

Analizziamo di seguito testi normativi di legislazione nazionale per la comunicazione interna alle istituzioni dei singoli stati comunitari, a cui i cittadini o consumatori finali possono avere accesso per le necessità soprattutto di natura giuridico-amministrativa della vita quotidiana.

Nei testi normativi di entrambi i paesi trovano attestazione solo le forme della famiglia lessicale delle energie rinnovabili: sorprende l'assenza totale di attestazioni della famiglia lessicale delle alternative. Numerose sono le varianti morfologiche ottenute a mezzo di variazione del numero (singolare vs. plurale), di ellissi – fonti rinnovabili (*renewable (re)sources*) –, o di sostituzione della testa del sintagma – energie/a rinnovabili/e (*renewable energy/ies*).

In inglese, particolarmente sfruttato e produttivo risulta essere il fenomeno dell'ellissi: mentre in italiano il fenomeno è parziale con l'omissione di uno solo degli elementi del determinato, in inglese l'ellissi da origine sia a forme con riduzione parziale del determinato – *renewable sources* –, sia a forme con omissione totale del determinato, *renewables*. È questa ultima, tuttavia, la forma maggiormente e quasi esclusivamente attestata nei testi normativi inglesi. L'equivalente italiano le rinnovabili risultato attestato una sola volta nel *corpus* selezionato.

Relativamente alla famiglia delle energie pulite, si osserva nei testi italiani la tendenza a non usare tale espressione. Nell'Energy White Paper 2007, l'unica occorrenza dell'espressione clean energy supplies è utilizzata ad indicare tecnologie di rifornimento e approvvigionamento energetico, con particolare attenzione a tematiche ambientali e di sostenibilità ambientale.

Riportiamo di seguito un paio di esempi.

Orientamenti per una Politica Nazionale in materia di energia, Sezione 4., La situazione delle riserve petrolifere, Consiglio Nazionale dell'Economia e del Lavoro (N. 78, 2005)	***Energy White Paper 2007, UK Government, Sec. 5.3., Renewables <www.berr.gov.uk/energy/whitepaper/page39534.html>***
Secondo molti esperti e Centri di Studio è necessario puntare su diverse opzioni facilmente raggiungibili per traghettare dal petrolio ad altri sistemi energetici efficienti e puliti. Quindi è necessario: • […] incoraggiare il rapido sviluppo delle fonti di energia rinnovabili, come il vento, il sole, la geotermia e le coltivazioni a fini energetici di biomasse; […].	*Renewables.* *The UK has some of the richest renewable resources in Europe - particularly in terms of our wind and marine (wave and tidal stream) resources […].*

2.2. Informazione al consumatore: terminologia in uso nei testi informativi di istituzioni nazionali e di diritto pubblico

Come già nella comunicazione interna alle istituzioni dei singoli stati comunitari, così anche negli scambi comunicativi tra istituzioni e cittadini si osserva la tendenza a una semplificazione informativa e terminologica. Le uniche fonti attestate sono quelle della famiglia delle rinnovabili, con una maggiore variazione morfologica nella lingua inglese in seguito ai fenomeni di ellissi.

Energia Semplice	***UK Energy in Brief 2007 <www.berr.gov.uk/files/file39881.pdf>***
[…] i cittadini […] possono cambiare fornitore se riscontrano un servizio di cattiva qualità, se considerano eccessivo il prezzo offerto, se vogliono partecipare alla lotta contro i cambiamenti climatici, scegliendo energia da fonti rinnovabili oppure a basso tenore di carbonio.	*Electricity.* *The mix of fuels used to generate electricity continues to evolve. Since 1990, the use of coal, oil, and hydro in electricity generation has fallen, while gas, nuclear and renewables other than hydro have risen*
Più trasparenza nella bolletta. […] A breve saranno indicate anche informazioni sul mix di fonti (rinnovabili, gas, carbone) utilizzate per produrre l'elettricità acquistata.	*Energy sources.* *In 2006, biofuels accounted for 82% of renewable energy sources used with most of the remainder coming from large-scale hydro and wind generation.*

2.3. Informazione al consumatore: terminologia in uso nei testi giornalistici

I dati osservati nei testi giornalisti, italiani in particolare, sono indicativi sia della rilevanza delle tematiche energetiche a livello nazionale e internazionale, sia della varietà e insieme del "ricambio" lessicale ad esse collegato.

A livello di campi tematici, oltre a questioni relative all'utilizzo strategico delle fonti energetiche, alle nuove evidenze del cambiamento climatico, nonché alla sostenibilità ambientale delle scelte energetiche, nei testi giornalistici si da ampio spazio anche ai risvolti economico-finanziari.

A livello lessicale, le tendenze osservate includono l'addensarsi di anglicismi e l'uso di polirematiche non attestate al di fuori di queste tipologie testuali. Nell'articolo *Un'energia da reinventare* ("Il Sole 24 Ore", 12 maggio 2008) colpisce l'accostamento di espressioni italiane ad anglicismi in riferimento sia a nuove tecnologie energetiche sia a nuove fonti energetiche: energie rinnovabili, *biofuel*, *carbon capture*, *sequestration*. È significativo osservare che differenti sono anche le tecniche definitorie utilizzate, che includono definizioni di tipo enumerativo per energie rinnovabili, definizioni di tipo lessicografico per i due anglicismi *carbon capture* e *sequestration*, ma totale assenza di definizione per *biofuel*.

Un'energia da reinventare
Insufficienti ma fondamentali i fondi pubblici.

Di fronte a energia scarsa e costosa, ogni settore dell'economia globale è a rischio. Così gli alimentari stanno salendo di concerto con il greggio anche a causa dell'aumento dei costi di produzione, ma anche perché i terreni agricoli, soprattutto negli Stati Uniti, vengono convertiti alla coltivazione di prodotti per il *biofuel*.

[…] Per permettere ai Paesi in via di sviluppo di proseguire la crescita e ai Paesi ricchi di evitare uno scivolone, si rende sempre più urgente lo sviluppo di nuove tecnologie energetiche. […] Forse la tecnologia singola più promettente in termini di efficienza energetica è quella ibrida a "plug in" per le automobili: l'idea è che le macchine vadano prevalentemente con batterie da ricaricare ogni notte sulla rete elettrica, con l'alternativa dell'ibrido a benzina […].

La tecnologia più importante per lo sfruttamento sicuro dal punto di vista ambientale del carbone è l'eliminazione dell'anidride carbonica derivante dalle centrali a carbone. Queste tecniche di "carbon capture" e "sequestration" (Ccs) sono necessarie nei principali Paesi consumatori di carbone […] Per tutte queste tecnologie promettenti i governi dovrebbero investire nella ricerca e nei *test* di *early stage*. Senza finanziamenti almeno in parte pubblici, queste nuove tecnologie avranno infatti un decollo lento e faticoso. […]

La situazione è ancora più scoraggiante se si guarda ai singoli particolari. I finanziamenti Usa per le energie rinnovabili (solare, eolico, geotermico, oceanico e bioenergetico) hanno raggiunto una misera quota di 239 milioni di dollari, tre ore di spese per la difesa. ("Il Sole 24 Ore", 12 maggio 2008)

Non di rado si osserva l'addensarsi, all'interno di un unico testo, delle diverse famiglie lessicali fonti alternative, fonti rinnovabili, energie pulite ed energie ambientali, anche quando il tema centrale non è né il risparmio energetico né la sostenibilità ambientale. Nell'articolo che di seguito viene riportato, energie alternative, energie rinnovabili ed energie pulite vengono utilizzate come sinonimi, benché l'autore utilizzi l'espressione energie pulite tra virgolette, quasi ad indicarne un uso improprio. Che dire però delle energie ambientali? Nuovo concetto? Iperonimo di fonti rinnovabili, alternative, pulite oppure loro sinonimo ma con valori connotativi diversi?

Adesso è ora di investire sul tempo di domani
Dalle risorse alternative alle tecnologie ambientali

Non è più una novità così glamour quella dei fondi che puntano su gruppi quotati in Borsa attivi nelle energie alternative. Società di risparmio gestito grandi e piccole hanno intuito da tempo le potenzialità di un business ricco, carico di innovazione e quindi di (eventuali) ghiotti ritorni. È chiaro però come una situazione congiunturale tanto sfavorevole per il petrolio spinga le energie "pulite".

[…] Union Investment ha lanciato da poco un prodotto di questo tipo che si chiama «Climate Change»: un fondo azionario che scommette sui titoli quotati sulle piazze finanziarie di tutto il mondo attivi prevalentemente nel business delle fonti rinnovabili, delle energie ambientali e del riciclo dei rifiuti.

[…] Ma dove punta, nelle specifico, questo strumento? Risponde il gestore, Thomas Deser: «Climate Change investe a livello globale in società che hanno al centro delle proprie strategie le energie alternative, eolica, solare e idrica, le tecnologie ambientali, dove si tiene conto

di fattori quali l'efficienza e la riduzione delle emissioni di anidride carbonica, e il riciclaggio, sia per quanto riguarda il riutilizzo di acqua potabile, sia per il riciclo dei rifiuti di vario genere o degli imballaggi».

("Il Sole 24 Ore", 12 maggio 2008)

Ancora un esempio e un commento sull'uso dell'espressione "energie pulite". È interessante osservare come, in data 13 maggio 2008, "Il Sole 24 Ore" annuncia il nuovo mensile Energia 24 nell'articolo Nasce il mensile «Energia24»: la rivista, si legge nel testo, si occuperà anche di temi legati alle nuove fonti alternative e spiegherà «cosa s'intende per energie "verdi"»: poiché il mensile è rivolto in modo particolare a tutti gli operatori del settore energetico, cosa dobbiamo inferire a proposito dell'uso che è stato fatto fino ad ora di questa espressione a livello "popolare", "divulgativo"?

Nasce il mensile "Energia24"

Nasce "Energia24", il nuovo mensile de "Il Sole 24 Ore Business Media" dedicato al mondo dell'energia. Una rivista che parla ai manager delle piccole e medie imprese italiane, con l'obiettivo di diventare uno strumento di lavoro per tutti gli operatori del settore. […]

Uno strumento per capire come cambierà il concetto stesso di approvvigionamento nell'era del caro-greggio, per trovare le soluzioni più adatte per aziende e privati, per capire cosa s'intende per energie "verdi" e quali risparmi possono portare.

[…] Ogni mese "Energia24" affronterà quindi il tema energia nella sua accezione più classica, parlando quindi di petrolio, benzina, gas, elettricità, carbone e nucleare, ma approfondendo anche tutti i temi legati alle nuove fonti alternative: dall'eolico al fotovoltaico alle biomasse. ("Il Sole 24 Ore", 13 maggio 2008)

Riportiamo di seguito la tabella riassuntiva delle variazioni terminologiche osservate nei diversi testi analizzati.

Tipologia testuale	*Italiano*	*Inglese*
Testi normativi comunitari	Fonti (energetiche)/energie rinnovabili; energie alternative; energie pulite e rinnovabili	Renewable energy (sources); (sources of) alternative and (renewable) energy; clean and renewable energy
Testi normativi di legislazione nazionale	Fonti (energetiche)/energie rinnovabili; (fonti di) energia rinnovabile;	Renewable energy (sources); renewables; renewable electricity/ technologies
Testi informativi da parte di enti nazionali o di diritto pubblico	Fonti (energetiche) rinnovabili	Renewable energy (sources) Renewable sources Renewables
Testi informativi giornalistici	Energie rinnovabili/alternative Energie verdi Energie pulite Energie ambientali *Biofuel* *Carbon capture* *Sequestration*	Renewable energy (sources); renewables

3. Le fonti energetiche: repertori terminologici e strumenti lessicografici di riferimento

Dinnanzi alle numerose variazioni terminologiche osservate nelle diverse tipologie testuali, soprattutto in quelle maggiormente esposte e visibili dal punto di vista sociale e sociolinguistico, è opportuno verificare in che grado e modo il cittadino ha accesso a quest'ampia gamma di informazioni. L'analisi è stata effettuata attraverso lo spoglio di repertori terminologici in essere in ambito energetico e di strumenti lessicografici di riferimento, diversi per l'ente o autore che li ha prodotti, per tipologia e per dominio di riferimento. Pur nell'ampiezza e varietà degli strumenti rilevati, si possono osservare alcune tendenze generali, sia in riferimento alle modalità di lemmatizzazione delle voci, sia alle modalità definitorie e ai domini utilizzati.

Quanto alle modalità di lemmatizzazione, pur nella vastità ed eterogeneità delle forme e voci lessicali rilevate, si osserva la tendenza generale a lemmatizzare solo le famiglia delle rinnovabili e delle alternative, con rarissime precisazioni lessico-semantiche all'interno delle voci definitorie: assai raramente le espressioni fonti rinnovabili e fonti alternative sono indicate come sinonimi, più frequentemente vengono lemmatizzate come voci separate senza alcuna indicazione di carattere stilistico-retorico. Per quanto riguarda le varianti morfologiche ad esse collegate, le forme sorte per ellissi totale, già pressoché inesistenti nella lingua italiana, non trovano alcuna lemmatizzazione nei glossari o thesauri selezionati in entrambe le lingue; sono invece ampiamente attestate le varianti sorte per ellissi parziale e per sostituzione della testa del sintagma.

Modalità definitorie includono descrizioni analitiche con iponimi, iperonimi e sinonimi, tipiche dei thesauri, definizioni estensionali, tipiche dei glossari di istituzioni o società che operano nel settore energetico, e definizioni relazionali, tipiche dei dizionari della lingua italiana e inglese. È facile, tuttavia, incontrare anche definizioni miste.

Riportiamo di seguito alcuni esempi di glossari e repertori terminologici in essere rappresentativi dei fenomeni sopra descritti.

3.1. Thesauri

Riportiamo di seguito le schede terminologiche rilevate nel *thesauro* multilingue di termini ambientali GEMET EIONET <www.eionet.europa.eu/gemet>. L'unica famiglia lessicale lemmatizzata è quella delle rinnovabili, di cui sono lemmatizzati anche gli antonimi. Le definizioni sono miste e includono dettagli descrittivi di carattere enciclopedico, analitico e relazionale:

Eng: **Renewable energy sources**

Concept definition:

Energy sources that do not rely on fuels of which there are only finite stocks. The most widely used renewable source is hydroelectric power; other renewable sources are biomass energy, solar energy, tidal energy, wave energy, and wind energy; biomass energy does not avoid the danger of the greenhouse effect.

Broader terms: *energy source*
Themes: *energy resources*
Groups: ENERGY
It: *fonte di energia rinnovabile*

Eng: **Non-renewable energy resource**
Concept definition:
Non-renewable resources have been built up or evolved over a geological timespan and cannot be used without depleting the stock and raising questions of ultimate exhaustibility, since their rate of formation is so slow as to be meaningless in terms of the human life-span. (Source: GOOD)

Broader terms: *non-renewable resource*
Related terms: *energy production, fossil fuel*
Themes: *energy resources*
Groups: RESOURCES (utilisation of resources)
It: *fonte di energia non rinnovabile*

Non-renewable resource
Concept definition:
A natural resource which, in terms of human time scales, is contained within the Earth in a fixed quantity and therefore can be used once only in the foreseeable future (although it may be recycled after its first use). This includes the fossil fuels and is extended to include mineral resources and sometimes ground water, although water and many minerals are renewed eventually. (Source: ALL)

Narrower term: *non-renewable energy resource*
Related terms: *mineral resource*
Themes: *energy resources*
Groups: RESOURCES (utilisation of resources)
It: *energie non rinnovabili*

Renewable resource
Concept definition:
Resources capable of being continuously renewed or replaced through such processes as organic reproduction and cultivation such as those practiced in agriculture, animal husbandry, forestry and fisheries. (Source: LANDY)

Related terms: *energy production*
Themes: *energy resources*
Groups: RESOURCES (utilisation of resources)
It: *risorse rinnovabili*

4. Glossari

a) Glossari di enti nazionali o locali/regionali e istituzioni di diritto pubblico

Le voci dei glossari di seguito riportati variano sia nel criterio di lemmatizzazione delle voci sia nell'ampiezza del dettaglio descrittivo e comunque sono caratterizzate principalmente da descrizioni di tipo analitico.

Energy Glossary, Department for Business, Enterprise & Regulatory Reform (BERR), UK Government
<www.dti.gov.uk/energy>:

Renewable energy
Renewable energy includes solar power, wind, wave and tide, and hydroelectricity. Solid renewable energy sources consist of energy crops, other biomass, wood, straw and waste, whereas gaseous renewables consist of landfill gas and sewage waste.

Sportello Energia, Provincia di Parma
<www2.provincia.parma.it/page.asp?>
IDCategoria=1257&IDSezione=12907&IDOggetto=18896&Tipo=
GENERICO():

Energia alternativa
Energia derivata da sorgenti diverse da quelle fossili (carbone, petrolio, gas) e da quella nucleare. Si tratta di fonti prevalentemente rinnovabili. v. Biogas, Biomassa, Cella combustibile, Fonti energetiche rinnovabili, Impianto fotovoltaico, Impianto a pannelli solari termici.

Fonti energetiche rinnovabili
Categoria di fonti energetiche in cui rientrano il sole, il vento, le maree, il moto ondoso, l'energia idraulica, le risorse geotermiche e la trasformazione di prodotti vegetali o dei rifiuti organici e inorganici. v. Energia alternativa.

b) Glossari di società e imprese nel settore energetico

Enereco, Glossario Allargato Elettrico - Fotovoltaico - Minerario
<www.enerecosrl.com>:

Fonti energetiche rinnovabili
Categoria di fonti energetiche in cui rientrano il sole, il vento, le maree, il moto ondoso, l'energia idraulica le risorse geotermiche e la trasformazione di prodotti vegetali o dei rifiuti organici e inorganici.

3.3. Strumenti lessicografici della lingua inglese e italiana

È significativo osservare che il dizionario della lingua italiana di Sabatini e Coletti registra contemporaneamente come unico lemma entrambe le famiglie lessicali delle

rinnovabili e delle alternative, di cui fornisce definizione di tipo analitico e relazionale. Nessuna definizione di tipo analitico è invece presente nell'Oxford English Dictionary.

Sabatini Coletti - Dizionario della Lingua Italiana

Energie rinnovabili / energie alternative: (s.f., pl.) fonti energetiche primarie che possono essere utilizzate senza limitazione e che non inquinano, come l'energia solare, quella eolica ecc.

Oxford English Dictionary

Alternative energy

Energy fuelled in ways that do not use up natural resources or harm the environment.

Conclusioni

Nel presente lavoro abbiamo voluto ripercorrere il "viaggio" delle informazioni sulle tematiche energetiche, con particolare riferimento al lessico relativo alle fonti di energia, partendo dalle attestazioni interne alla comunicazione sia degli organismi istituzionali dell'Unione Europea sia delle pubbliche amministrazioni dei due stati comunitari, per giungere alle attestazioni negli scambi comunicativi tra istituzioni e cittadini e, infine, agli usi negli scambi comunicativi tra mezzi di informazione e cittadini.

I testi giornalistici in particolare, primo veicolo di diffusione dell'informazione al cittadino, possono trasmettere ad un pubblico esterno una certa "indeterminatezza" e vaghezza proprio per la diversità dei campi tematici trattati e della terminologia utilizzata. Il confronto poi con i repertori terminologici e lessicografici in essere in ambito energetico dimostra che sono i testi gli unici veri depositari del ricambio lessicale di una lingua e della sua variazione sociolinguistica.

Bibliografia

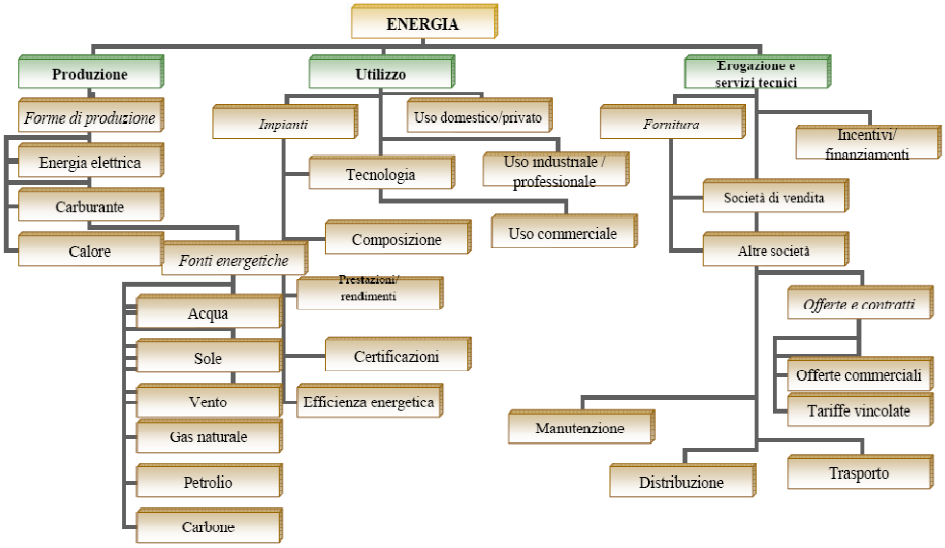
Oxford English Dictionary, 2nd ed. 1989, OED Online, O.U.P., Oxford <www.oed.com>.

Sabatini F., *"Rigidità-espilictezza" vs. "elasticità-implicitezza": possibili parametri massimi per una tipologia dei testi*, "Etudes Romanes", 42, 1999, pp. 141-172.

Il Sabatini Coletti - Dizionario della Lingua Italiana, Rizzoli Larousse, Milano, 2003.

Zanola M.T. *Energie tradizionali e rinnovabili: proposte di interventi terminologici*, Atti del Convengo Nazionale Ass.I.Term. I-TerAnDo, Università degli Studi della Calabria 5-7 giugno 2008.

Allegato 1: il campo concettuale dell'energia elettrica



Linee di problema per il trattamento terminologico di documenti amministrativi elettronici

DOMENICO BOGLIOLO

This paper presents a review of the strategic and operational objectives of AIDA (Italian Association for Advanced Documentation) related to the study of scientific and professional situations and their reflections on their terminological descriptors. The inter- and trans-disciplinary approach calls for new frontiers and for the participation of other professional as well as institutional bodies. Knowledge management is presented as a unifying domain of scientific and organizational activities to develop linguistic and computational tools for the achievement of dynamic innovation of record management indexing in large administrative organizations.

Keywords: AIDA – record management – administrative document – terminology

Collaborazione trasversale

L'Associazione Italiana per la Documentazione Avanzata, AIDA, che qui rappresento, dalla sua fondazione 25 anni fa ha cercato di affrontare in ampio spettro la varietà scientifica e professionale del mondo dell'informazione e della documentazione.

Si tratta di un contenitore di idee e di iniziative quasi generalista e quasi catalizzatore, sia per l'attenzione posta alla collaborazione con altre associazioni e realtà del settore – fino a farne un promotore di interscambi scientifici e organizzativi – sia per la visione globale di questo dominio, attenta a ogni declinazione, a ogni piega o rigagnolo suscettibile di innovare la comprensione di questa realtà e, per quanto possibile, di tracciarne rotte inedite e sviluppi futuri.

Due esempi soltanto. All'inizio della sua costituzione nel 1984, l'*Online Information Meeting* ha visto AIDA costante nell'organizzare i suoi seminari londinesi raccogliendo i partecipanti italiani alla manifestazione, e più volte i temi trattati sono stati stimolo per le edizioni successive dell'*Online* stesso.

Recentemente, AIDA si è fatta realizzatrice delle iniziative nate dalla "Dichiarazione di Roma" del 1995 in seno allo *European Council of Information Associations* (ECIA) per l'impegno inter-associativo a instaurare sistemi di certificazione professionale, facilitarne il mutuo riconoscimento e contribuire alla loro compatibilità nel contesto europeo, dapprima con la traduzione ufficiale italiana del referenziale europeo delle professioni

del settore (*Euroguida I&D: volume 1: competenze e attitudini dei professionisti europei dell'informazione-documentazione; volume 2: livelli di qualificazione dei professionisti europei dell'informazione-documentazione*) liberamente scaricabile dal nostro sito e, poi, fino alla promozione, insieme con altre Associazioni, della nascita, due anni fa, di "Certidoc Italia", organismo italiano di certificazione a norma ISO delle competenze professionali acquisite, e che ha da pochi giorni esaurito il primo bando di certificazione del 2008. Con "Certidoc Italia" gli organismi di certificazione europea pienamente attivi sono ora sette, e undici le versioni linguistiche nazionali del referenziale europeo del quale, tra l'altro, è appena iniziata la revisione quinquennale per adeguarlo al continuo divenire delle realtà professionali. Non si tratta, qui, di "riconoscimenti" ordinistici di queste professioni, che non consentono controlli di qualità nel tempo, ma dell'innesco di un processo di determinazione periodica delle competenze reali possedute dal professionista – al limite, indipendentemente dal titolo di studio – che vuole, così, offrirsi al mercato internazionale del lavoro garantendo al committente o al datore di lavoro certezze scientifiche e procedurali. E, nel panorama italiano, si tratta di una quasi-rivoluzione.

Il termine "documentazione"

L'impegno di AIDA nelle scienze dell'informazione non può disgiungersi da una costante attenzione all'evoluzione della stessa terminologia che caratterizza il campo e le professioni connesse: ne è un esempio il termine "documentazione" che caratterizza questa Associazione e una parte degli operatori professionali.

L'origine del termine è relativamente antica, risalendo almeno al *Traité de documentation* di Paul Otlet che nel 1934 indicava nella bibliografia scientifica, statistica e sociometrica uno degli strumenti principali contro la dispersione delle pubblicazioni e per la comunicazione e trasmissione della scienza. Studio, quindi, epistemologico e, insieme, prassi bibliologica, come scienza (anzi, metascienza) e tecnica generale del documento, visto come manifestazione diretta del pensiero depositata su qualsiasi supporto. Differenziandosi progressivamente dalla bibliologia, la documentazione ne diviene, prima, un suo sottoinsieme – così come la biblioteconomia e la bibliografia – e, successivamente, se ne emancipa fino a caratterizzarsi come scienza dei modi di utilizzo dei documenti. Illuminante, al riguardo, la metafora metallurgica impiegata dallo stesso Otlet per indicare l'estrazione dai documenti del loro contenuto puro per così separarlo dalla ganga, dalla scoria che ne appesantisce la comprensione e la diffusione.

Questo lavoro di scavo dei contenuti, impegno minerario, quasi, per l'estrazione della quintessenza, la loro rielaborazione in forme scientificamente significative e comunicative, caratterizza molto bene le attività dei professionisti variamente impegnati nelle scienze dell'informazione, al di là delle tecnologie impiegate.

È infatti frutto di un errore (in gran parte, tutto italiano) di prospettiva degli anni '70, per esempio, la caratterizzazione del documentalista come il professionista che utilizza le cosiddette "nuove tecnologie", di solito in fallace contrapposizione con il bibliotecario o il bibliografo che, invece, ne utilizzerebbero esclusivamente di tradizionali e antiche. La diffusione generalizzata degli strumenti elettronici fino alla pervasività di Internet ha, di fatto, unificato prassi e strategie di queste e di altre figure connesse con il trattamento dei documenti, al punto che possiamo imputare, genericamente, alle scienze dell'informazione un complesso professionale che parte dall'autore e finisce con l'utente, passando variamente per, che so? e alla rinfusa, l'archivista, il tipografo, il terminologo, il paleografo, il *webmaster* e, insomma e per farla breve, tutti i (finora, nell'edizione 2005) trentatré campi di competenza definiti nell'*Euroguida I&D*.

Ma quanto è, oggi, comprensibile il termine "documentazione"? Non dimentichiamo che la FID, *Fédération Internationale de Documentation* (e dal 1988 "*d'Information et de Documentation*"), è stata sciolta, sia pur accidentalmente, nel 2002. Fuori dall'Europa il nostro termine non ha, quasi, cittadinanza, se non nell'attività di colui che scrive il manuale d'uso di uno strumento, non importa se *hardware* o *software*, o che fornisce le cosiddette "pezze d'appoggio" per qualsiasi attività, non escluso il rendiconto delle spese. Altre ambiguità nascono se ci volgiamo al professionista della documentazione, "documentalista" o "documentarista" (confondibile con il creatore di documentari cinematografici – con buona pace dei colleghi della Camera dei Deputati) o "documentatore", come fu proposto contro la derivazione francofila del termine. Da tempo si pensa a nuovi termini, riconoscibili in àmbito internazionale, come "specialista (o scienziato) dell'informazione" ma così creando, a mio parere, nuove confusioni perché un medesimo termine si vedrebbe applicato ad àmbiti diversi e distinti.

E di quale informazione stiamo, poi, parlando? Pensiamo ai guasti provocati, per esempio, dal sig. Philippe Dreyfus, ingegnere della Bull e inventore nel 1962 del termine "informativity" e del suo parallelo "informatique" come fusione di "informazione" e "automatica", dando così il destro agli esperti di "computer science" di ritenersi operatori delle informazioni anziché dei dati...

È in contesti come questo odierno che, invece, si può percepire tutta la ricca varietà di contenuti di una professione, forse *senza nome* o *con troppi nomi*, ma alla quale non mancano i contenuti né gli àmbiti applicativi.

Attività "miste"

In attesa delle definizioni professionali e disciplinari che l'evoluzione delle cose (e, dopo di queste, la riflessione accademica) ci fornirà, non resta che considerare tutto ciò che riguarda i dati, l'informazione, la conoscenza, come un unico dominio ampio e

sfaccettato nel quale pescare per la promozione di attività "miste" – così come ci siamo ridotti a dire dopo le utili precisazioni distintive che Ingetraut Dahlberg ha proposto al congresso ISKO del 2007 per *interdisciplinarità* (lo studio di una disciplina dal punto di vista di un'altra), *transdisciplinarità* (l'applicazione dei metodi di una disciplina a un'altra), *multidisciplinarità* (lo studio di un fenomeno in diverse discipline), *pluridisciplinarità* (lo studio di una proprietà in diverse discipline), *sindisciplinarità* (la collaborazione tra discipline per un medesimo fine).

Sembra infatti questa la sede appropriata per AIDA per lanciare la proposta dell'attivazione – diciamo così, "sperimentale" e del tutto, per ora, esemplificativa – di un ampio gruppo "misto" di ricerca per professionisti del settore, per esempio per la gestione della conoscenza di grandi amministrazioni, in ordine all'integrazione delle attività connesse con la gestione documentale, dal protocollo all'archivio passando per il lavoro collaborativo.

Non nego un mio personale interesse nella proposta: da un certo tempo, infatti, mi occupo di coordinare progetti per la gestione dei flussi documentali – e quindi dell'intero sistema informativo – per l'Università di Roma "La Sapienza". Questi progetti sono diversificati e toccano punti e azioni diverse della macchina burocratica, ma tutti presuppongono ed esigono la soluzione di un problema unico: il rapporto fra forma e contenuto degli atti amministrativi, bisognosi di codifica terminologica, linguistica, semantica e anche sintattica, ai fini della trasparenza dell'informazione, dell'*information seeking* & *retrieval* dei documenti dell'Ente, delle politiche di scarto archivistico, della facilitazione dei *workflow*, dell'innovazione dinamica dei titolari al mutare delle forme e degli scopi dell'organizzazione.

Arricchimento professionale

È questo, per noi, un obiettivo unificante di professionalità diverse ma integrabili e quindi uno strumento per la realizzazione, nei fatti, di un arricchimento delle competenze professionali degli addetti all'intero sistema: così come le competenze giuridiche a amministrativo-contabili dell'impiegato si sono da tempo variamente integrate con quelle cosiddette informatiche, nel medesimo modo c'è necessità di ulteriori integrazioni con competenze archivistiche e genericamente documentalistiche, per poter giungere a una visione globale e finalizzata dei processi e dei procedimenti che vengono quotidianamente posti in essere, fino alle forme e ai contenuti della comunicazione all'utente finale. Non, qui e nel lavoro quotidiano, quindi, un gruppo di lavoro composto da professionisti diversi, ciascuno con la sua specialità, che affianchi – cosa improponibile se non in misura molto limitata – il singolo impiegato ma, piuttosto, l'acquisizione, da parte dell'impiegato stesso, di nuove e diverse competenze professionali: più

un arricchimento dell'individuo che il supporto dato da esperti di discipline diverse seppur complementari. Si tratterebbe, cioè, di instaurare un'attività formativa vera e propria oltre che informativa o di addestramento alla gestione "globale" del documento.

Per giungere a questo risultato in modo permanente c'è, però, necessità di approntare strumenti *software* che consentano all'impiegato di gestire procedimenti e processi in modo il più possibile facilitato, automatico per quanto possibile. Qui il gruppo "misto" acquisisce un ruolo essenziale, per l'analisi linguistica e terminologica del documento amministrativo fino a estrarne radici (alla Propp…) di significato e di funzione, per la costruzione dell'ontologia specifica del documento amministrativo, per la derivazione di tesauri funzionali alla descrizione dei metadati del contenuto, per la costruzione di indici sia documentari sia archivistici per l'arricchimento dinamico e (anche) *bottom-up* dei titolari.

Concorso e sinergia

Un impegno come quello qui velocemente esemplificato necessiterebbe, verosimilmente, del concorso, oltre che di professionalità diverse, anche della sinergia di istituzioni connesse con il cómpito: dalle associazioni professionali del caso alle istituzioni universitarie e di ricerca interessabili – un impegno di programma che andrebbe ben oltre l'àmbito settoriale e nazionale per coinvolgere *partner* europei e, comunque, internazionali perché il problema è generale e attraversa tutte le amministrazioni, qualunque siano la struttura organizzativa e la lingua e la cultura di riferimento.

AIDA può intanto mettere a disposizione alcuni degli strumenti di comunicazione del caso, come *forum*, *blog*, *wiki* e, insomma, l'apparato 2.0 di cui dispone, per la creazione del gruppo "misto" necessario alla definizione e alla promozione dell'impresa, oltre che alla creazione delle prime *liquid publication* per l'arricchimento cumulativo delle conoscenze via via conseguite.

Probabilmente, sfide del genere sono i soli meccanismi di reale unificazione operativa della molteplicità di oggetti, di metodi, di tecniche che caratterizzano l'oceano delle scienze dell'informazione e della documentazione: aggregazione e scambio (anche litigioso...) su un progetto finalizzato comune.

E questo è il nostro augurio di buon lavoro e, contemporaneamente, una chiamata alla partecipazione.

Verso un formato standard nelle intercettazioni: archiviazione, conservazione, consultazione e validità giuridica della registrazione digitale [1]

LUCIANO ROMITO, MARIA TUCCI, GIUSEPPE CAVARRETTA

Wiretapping or recording for forensic purposes is a way of researching evidence to determine the natura of a crime. It is a recording of spontaneous and natural speech, and it is for this reason that it is of great interest to linguists and the authors of this study. This interest is also stimulated by the fact that in Italy wiretapping does not exclusively regard low-level organized crime, where the linguists would find a low register of Italian or, more likely, the use of dialect, but on the contrary, the people tapped are ministers, princes, doctors, bank director, industrialists, journalists, teachers, judges, lawyers, priests, policemen, soccer players, referees, showgirls, employees, and so on, touching all social strata of the Italian population and all the possible registers of Italian or its dialects. All this is possible without the need of building lists of words or sentences and most of all without researching half-spontaneous speech through complex and expensive strategies. So much spontaneous material differentiated by production, age, sex, diafasic and diastratic variables presents just one problem: the quality of the recording and the instruments used. Due to this great attention paid to recordings produced by non-legal experts, throughout the years we can notice both a high degree of variance and superficiality in the techniques of acquisition, preservation, cataloguing and registration of the material recorded and the use made of it.

Keywords: wiretapping – cataloguing – preservation – original, copy

Le intercettazioni

Fin dall'inizio delle intercettazioni su larga scala e solo fino a pochi anni fa, le intercettazioni venivano effettuate con l'ausilio di un registratore marca UHER modello RT 2000 o RT 4000. Salvo alcune rarissime eccezioni, il formato era unico per tutte le Procure italiane. Le registrazioni avvenivano su supporto analogico ovvero su una bobina normalmente di marca BASF (cfr. Romito 2000). L'unica variabile riscontrabile era la velocità di scorrimento del nastro 2,38 cm/sec per le intercettazioni di tipo telefonico (su rete fissa) e 4,75 cm/sec per le intercettazioni ambientali.

Nell'era del digitale, il miglioramento sarebbe dovuto essere quasi scontato, invece si rileva la nascita e l'uso contemporaneo di una infinità di formati digitali (MP3, WAV,

formati proprietari e criptati molti dei quali compressi, ecc.) con supporti differenti (cassette analogiche, microcassette digitali DDS, dischi ottici, dischi ORB, ecc.), frequenze di campionamento differenti con oscillazioni che variano da 8000 a 44100 Hz con numero di *bit* diversi e registrazioni sia stereo che mono. Anche i tipi di intercettazione sono molto diversi tra loro: le intercettazioni telefoniche possono essere su rete fissa o mobile, su ponte radio, su centrale digitale o meccanica, le reti possono essere GSM, UMTS o VOIP. Anche l'operatore è molto diverso, infatti con l'avvento del digitale e la lentezza delle istituzioni nell'adeguarsi al cambiamento è divenuta prassi comune delegare al privato qualunque tipo di operazione a partire dal noleggio della microspia e della apparecchiatura per l'intercettazione alla sistemazione e installazione della stessa, fino al trattamento analitico della registrazione attraverso consulenza tecnica di trascrizione, comparazione, filtraggio ripulitura ecc.

I registratori analogici (RT2000 e 4000) sono stati soppianti dai registratori digitali RT6000 (con supporto analogico: cassetta DDS), RT8000 (con supporto digitale: CD) e RT10000 (senza alcun supporto; la registrazione avviene tramite server e memorizzata direttamente su Hard Disk). Tutto ciò è avvenuto in un tempo ridottissimo e, ancor più grave, in momenti differenti da Procura a Procura. Si hanno, quindi, contemporaneamente intercettazioni analogiche in una parte del paese e digitali in un'altra, formati e supporti completamente differenti tra loro ecc.

Anche l'informatizzazione degli uffici preposti è arrivata in momenti differenti nel paese, così non è affatto raro trovare bobine e CD ROM catalogati e conservati in ugual modo o archiviazioni che oscillano dal *libro rubrica* al *software* fatto in casa nato dalla volontà dell'operatore di gestire le registrazioni in maniera quanto meno più moderna.

Eppure, molte sentenze della Cassazione (cfr in seguito) riportano che, la **prova** in un processo, non deve essere intesa la trascrizione, la trasposizione su carta o l'analisi della registrazione sonora, ma bensì la **bobina originale**, sulla quale è avvenuta la registrazione stessa. È proprio per questo motivo che le bobine una volta registrate e munite oltre che di brogliaccio cartaceo anche di *striscetta* sulla quale ogni conversazione veniva identificata attraverso l'ora, il giorno, il mese e l'anno, il numero di telefono in entrata e in uscita, il numero progressivo e il numero di giri della bobina all'inizio e alla fine, veniva sigillata con ceralacca e il plico veniva firmato da un operatore responsabile.

Molte delle indagini dipendono dalla qualità degli strumenti utilizzati per l'intercettazione, per la conservazione dei supporti e dalle disponibilità economiche di ogni singola Procura. Il ricorso ai privati è aumentato esponenzialmente e spesso grossi investimenti economici non corrispondono ad alta qualità. È eclatante il caso presentato in una puntata della trasmissione televisiva "Chi l'ha visto" in onda su Rai 3 (12 febbraio 2007) nella quale si fa riferimento ad una inchiesta della Procura della Repubblica presso il Tribunale di Reggio Calabria denominata "Gioco D'azzardo". L'inchiesta basata (anche) su intercettazioni telefoniche e ambientali, ha portato all'arresto di 16 im-

portanti personaggi di Messina, tra cui imprenditori, poliziotti, magistrati e anche un ex sottosegretario al tesoro. In particolare, ciò che in queste sede interessa è che una tra le intercettazioni più importanti effettuata dalla DIA nel 2001, è avvenuta tramite un microregistratore analogico per appunti su una micro-cassetta a velocità LP. Su tale registrazione hanno lavorato e sono intervenuti decine di periti a livello nazionale. Come dire, a fronte di enormi investimenti e spinte verso una sempre crescente digitalizzazione e informatizzazione dei processi, una tra le prove più importante è stata fornita grazie ad un piccolo registratore analogico portatile dal costo di circa una decina di euro.

Lo scopo di questo lavoro quindi è quello di evidenziare la necessità di una formato unico e standard per le registrazioni sonore in ambito forense in tutte le procure italiane; la necessità di una creazione di un protocollo per l'archiviazione, la conservazione e la consultazione dei flussi sonori intercettati; l'utilizzazione di protocolli metodologici Standard come AES ad esempio; l'identificazione della registrazione originale e quindi della **prova** in un processo, nonché la necessità di accertare la validità giuridica della registrazione digitale oggi; la creazione di un documento sonoro informatico; la riservatezza e la sicurezza di tale *documento*; lo scambio di dati con formati simili e confrontabili su tutto il territorio nazionale con un notevole abbattimento dei costi e dei tempi.

Dall'analogico al digitale: il caso delle intercettazioni

Il processo di digitalizzazione è legato a uno dei più importanti e significativi sviluppi tecnologici del nostro tempo. È noto che ciò che avviene nella conversione di un documento da analogico a digitale comporta la suddivisione in unità discrete di qualcosa che in realtà è continuo. Il procedimento è molto semplice; infatti, in fase di digitalizzazione qualsiasi carattere viene tradotto in una sequenza numerica composta da otto cifre. Il *bit* come unità di informazione rappresenta una vera e propria rivoluzione tecnologica che inizia con l'invenzione dei primi *computer* e raggiunge la sua massima diffusione con la rete Internet. La potenzialità del digitale consiste nel far convivere codici e linguaggi differenti sulla stessa macchina mentre nella rappresentazione di tipo analogica informazioni diverse devono essere archiviate su supporti differenti e non possono essere decodificate dallo stesso dispositivo. Questo è quanto succedeva nel campo delle intercettazioni quando queste venivano effettuate utilizzando un registratore analogico. L'intercettazione avveniva in parallelo su due differenti registratori e le registrazioni venivano segnalate come originale PG e copia AG; di fatto, un secondo originale registrato contemporaneamente su un identico registratore (cfr §§ succ.). Attraverso un filtro posto in entrata, nella banda di frequenza intorno a 2138 Hz, venivano registrati contemporaneamente sul nastro e, attraverso una piccola stampante collegata al registratore, su una striscetta cartacea (tipo scontrino fiscale) tutte le informazioni

relative all'operazione in corso: numero progressivo della registrazione, anno, mese, giorno e ora di inizio della registrazione, numero telefonico composto, ora, minuto e secondo relativo alla fine della registrazione, numeri di giri della bobina e infine numero della pista utilizzata sul nastro per la registrazione.

Le medesime indicazioni venivano riportate su un brogliaccio o verbale cartaceo ad opera di un operatore. Ogni turno di servizio aveva quindi un verbale ed un responsabile. Alla fine della intercettazione, le bobine PG ed AG venivano sigillate con ceralacca e il reperto veniva firmato da un operatore-responsabile. Il formato, il supporto e le procedure, salvo rarissime eccezioni, erano uniche in tutta Italia. Il reperto veniva in séguito consegnato all'Ufficio Reperti dove avveniva la catalogazione e la conservazione.

Il consulente che abbia necessità di consultare la registrazione, dietro autorizzazione di un giudice, si reca presso l'Ufficio Reperti e preleva la copia AG della registrazione, lasciando l'originale PG presso gli uffici della Procura.

È capitato in alcuni casi che il tempo abbia logorato o rovinato parte della bobina registrata; l'usura, per esempio, può aver portato allo strappo o alla rottura in alcuni casi del nastro o del supporto. In tali casi, è stato necessario effettuare prima un restauro del supporto, recuperare le informazioni presenti e non ancora danneggiate sul nastro e confrontarle con quelle presenti nella bobina originale PG. Proprio come potrebbe accadere per un vecchio libro come una cinquecentina ad esempio. Nonostante le parti mancanti il contenuto della pagina non è completamente perso.

Figura 1 - Anche se solo in parte, il documento può ancora essere letto e ciò anche dopo 600 anni dalla sua produzione e una non ottima conservazione.

Nell'era del digitale la situazione si è capovolta completamente i registratori sono stati sostituiti da modelli differenti che utilizzano formati e supporti differenti. Si oscilla da una cassetta DDS analogica ma registrata in digitale che contiene fino a 54 ore di registrazione (al contrario di un massimo di 5 ore di una bobina registrata a velocità

bassa) RT6000 [2], fino alla registrazione tramite server su *hard disk*. In questo ultimo caso, alla fine dell'intercettazione viene creata un copia della registrazione su un CD o su un DVD. Il registratore durante la fase di registrazione crea un *file* TXT con tutte le informazioni relative alla registrazione: anno, mese, giorno e ora, inizio e durata della conversazione, cella utilizzata, progressivo della registrazione, ecc. Si ha anche un campo note che può essere riempito dall'operatore con un riassunto della conversazione o con appunti relativi all'indagine. Dopo un certo lasso di tempo l' *hard disk* viene formattato cancellando definitivamente tutti i segnali *originali* presenti. Il séguito della procedura è rimasto inalterata. Il CD viene confezionato e chiuso con ceralacca:

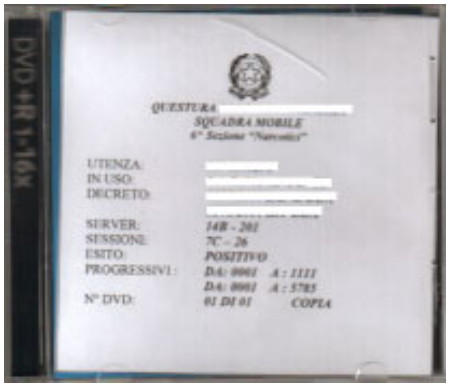


Figura 2 - Supporto digitale

Questo viene debitamente firmato e consegnato all'Ufficio Reperti. Come si può leggere dalla figura risulta essere una copia, in quanto l'originale è presente sull'*hard disk* che per ovvi motivi non può certo essere archiviato. Nonostante quindi la grande qualità, almeno in linea teorica, del digitale, non si ha alcuna certezza riguardo la correttezza della registrazione, della conservazione, del formato e, soprattutto, non sappiamo se le macchine costruite nei prossimi cinque o dieci anni saranno ancora capaci di leggere i supporti attuali. In meno di venti anni infatti questi si sono evoluti passando da schede forate (ormai inutili ed illeggibili) a memorie statiche o flash.

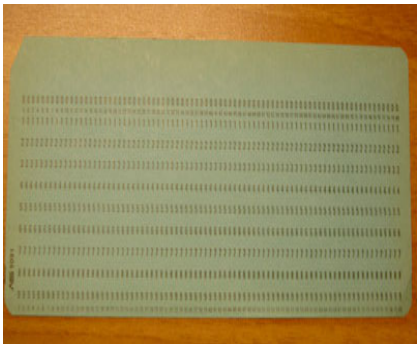


Figura 3 - Alcuni supporti digitali ormai in disuso

Ulteriori argomenti di discussione riguardano la durata ed il restauro. Al contrario del nastro magnetico o del vinile, la durata del supporto digitale (CD o DVD) non supera i dieci anni di vita e, ancora più drammatico, al contrario di una pagina scritta nel '500, un piccolo danneggiamento del supporto digitale rende tutto il supporto illeggibile con la conseguente perdita di tutti i dati contenuti nel CD o nel DVD.

La rapida evoluzione verso i sistemi digitali ha determinato una proliferazione di formati differenti soprattutto nell'ambito dei *files audio*. Riteniamo che le risorse del digitale dovrebbero essere sfruttate per uniformare e rendere trasparente l'intero intervento relativo alle intercettazioni. In questo momento, il sistema *analogico* risulta, al contrario, essere più "fedele" rispetto all'attuale sistema digitale.

Originale e copia nel sistema analogico e in quello digitale

Nel paragrafo precedente abbiamo ripercorso per grandi linee come in pochissimo tempo la procedura legata alle intercettazioni sia cambiata evolvendo verso il sistema informatizzato. Abbiamo anche accennato al concetto di supporto e registrazione originale e in copia.

La corte di Cassazione Penale Sez. V con una Nota del 11/03/2002, n° 9633 sancisce che «[…] la prova è costituita dalla bobina. […] La trasposizione su carta del contenuto delle registrazioni rappresenta solo un'operazione di secondo grado»

Nel sistema analogico, l'intercettazione avveniva *in parallelo* su due differenti registratori e le registrazioni venivano segnalate come *originale PG* e *copia AG*. Sul termine *copia* molte pagine sono state scritte, molte delle quali anche inutilmente. La *copia AG* di fatto non è mai stata una vera copia, bensì un *secondo* originale registrato in parallelo su un *secondo* registratore. L'unica differenza fra le due registrazioni o, meglio, fra i due originali, sta nel fatto che la registrazione AG avveniva di continuo su un'unica pista senza alcuna interruzione, mentre la registrazione PG (o di lavoro) veniva ascoltata anche durante la fase della intercettazione (durante i momenti in cui nessuna registrazione era in atto), quindi poteva accadere che tutte e quattro le piste del nastro venissero utilizzate.

Altra condizione per noi importante è che la copia di una registrazione analogica non ha mai la medesima qualità dell'originale e, comparando le due registrazioni, è sempre possibile identificare l'originale.

L'intercettazione in *era digitale* avviene direttamente sulla memoria di massa di un calcolatore (*hard disk*). Alla fine dell'intercettazione, o anche prima, l'addetto o l'operatore trasferisce i file su supporti digitali (CD-ROM/ DVD o altro). A questo punto si pone un quesito, almeno sotto il profilo formale: qual è la registrazione originale?

È noto che la *copia* informatica di un *file audio* ha caratteristiche identiche, in termini acustici e di qualità del segnale, e senza adeguati sistemi di protezione non è assolutamente possibile distinguere la *copia* dall'*originale*. Spesso però, ci si trova a lavorare su supporti la cui origine rimane ignota, ed è difficile, senza l'apporto di verbali o di indicazioni riguardo le caratteristiche della strumentazione utilizzata in fase di registrazione o di copiatura, stabilire se si tratti di un originale o di una copia. La maggior parte delle volte la *copia-originale* è un semplice CD ROM contenuto in una bustina in

plastica o in una busta da lettere senza alcun riferimento con una scritta effettuata con un pennarello indelebile del numero di RIT (Registro di intercettazione) o del Procedimento Penale (cfr figura 2). Nulla viene riportato riguardo al formato, all'attrezzatura utilizzata, alle metodiche di riversamento o di copiatura. L'unica informazione è la traccia informatica dell'avvenuta registrazione.

L'utilizzo delle tecnologie digitali pone il problema di come rendere inconfutabile la *prova* della registrazione. Mentre per il procedimento analogico le bobine e le relative striscette venivano prodotte in duplice copia (doppio originale), venivano inserite in un plico firmato dall'operatore responsabile e sigillato con ceralacca, e al consulente o al perito veniva consegnata solo uno dei due originali, nel mondo digitale la registrazione avviene su *hard disk* e le copie che vengono prodotte non hanno alcun sigillo o sistema di protezione o certificato di autenticità. Quindi, nel caso della registrazione analogica il problema della *copia* è un problema tecnico, mentre nel caso del digitale il problema della *copia* diventa un problema di **validità giuridica**.

Quanto detto ha portato ultimamente la Cassazione a assumere alcune posizioni come la seguente, ad esempio [3]:

Garanzia sul riscontro originale delle trascrizioni della polizia

Cassazione: "server " nelle Procure per memorizzare le registrazioni

ROMA Sulle intercettazioni – andando anche nel solco della previsione del ddl del governo che in maniera ancora più netta prevede centrali uniche di ascolto presso le Corti di Appello – la Cassazione ha messo a punto un piccolo assestamento «garantista» che, comunque, non mette in discussione i processi in corso in quanto – già adesso – in tutte le Procure c'è una «memoria» centrale che custodisce il materiale captato.
Gli ermellini delle Sezioni Unite penali di Piazza Cavour hanno stabilito che affinché le intercettazioni autorizzate dai pubblici ministeri siano utilizzabili in dibattimento, devono essere memorizzate dal «server» della Procura che non può solo fare da «ponte» per la trasmissione del segnale verso i centri d'ascolto esterni (quelli collocati negli uffici di polizia giudiziaria). Ma deve essere in grado di fornire, agli avvocati che lo richiedono, il riscontro originale alle copie dei dischetti contenenti le trascrizioni realizzate nei comandi di polizia e carabinieri.
Insomma l'indagato, se non si fida delle forze dell'ordine, deve poter comparare le intercettazioni che lo incastrano con la registrazione originale rimasta impressa nella memoria della centrale di registrazione dislocata in Procura.
Ad ogni modo, fanno presenti fonti della stessa magistratura, già adesso in ogni Procura c'è il server fisso e quindi gli avvocati che sospettano «taroccamenti», possono confrontare le registrazioni della polizia giudiziaria con quelle del disco rigido dei pm.
A sollevare il caso è stato il ricorso di un indagato di Comacchio (Ferrara), al quale era stata imposta la misura cautelare dell'obbligo di dimora nel comune di residenza col divieto di uscire nelle ore notturne. Il legale dell'uomo sosteneva che i resoconti delle intercettazioni, in base alle quali era stata emessa la misura cautelare, erano inutilizzabili in quanto fatte nel Comando dei carabinieri di Comacchio e non in Procura.
I supremi giudici – in sostanza – gli hanno risposto che se ha simili dubbi può benissimo bussare in Procura e consultare il server fisso alla ricerca di imprecisioni nelle trascrizioni delle intercettazioni che «danneggiano» il suo cliente.

Ovviamente, tutto ciò diventa molto costoso sia dal punto di vista delle attrezzature ma anche e soprattutto sotto il profilo della manutenzione.

Quanto appena detto accade, nonostante lo stesso Ministero per le Riforme e le Innovazioni nella Pubblica Amministrazione abbia pubblicato la "*Proposta di regole tecniche in materia di formazione e conservazione di documenti informatici*". La proposta è particolarmente interessante per quanto riguarda le intercettazioni soprattutto per gli articoli 5 e 7 del decreto. Nel primo relativo alla **Riproduzione di documenti informatici** si legge:

> «Il processo di riproduzione di documenti informatici ai fini della conservazione avviene mediante memorizzazione su diverso e adeguato supporto fisico e termina con l'apposizione sull'insieme dei documenti o su una evidenza informatica contenente una o più **impronte** dei documenti o di insiemi di essi del riferimento temporale e della **firma digitale** da parte del responsabile della conservazione che attesta il corretto svolgimento del processo»

e nel secondo, relativo ai **Formati per la conservazione**:

> «Il formato di conservazione dei documenti informatici assicura la conservazione del documento e delle sue caratteristiche nel rispetto della normativa vigente. Il formato di conservazione è un **formato standard** aperto, compreso tra quelli riconosciuti dagli organismi nazionali e internazionali preposti alla relativa normazione».

Se, dunque, il processo di realizzazione e riproduzione dei documenti informatici frutto di intercettazioni fosse così gestito dagli operatori del settore anche nel caso del digitale, non parleremmo più di originale e di copia in quanto tutte le copie sarebbero degli originali. Il contenuto del documento, della firma digitale e di tutte le informazioni di sicurezza sono riportate all'interno del documento stesso per cui, effettuando una copia informatica, di fatto viene riprodotto integralmente il documento e non esiste la possibilità di distinguere la copia dall'originale, mentre è sempre possibile verificare l'integrità del documento e la validità della firma digitale. Apporre la firma digitale sia sui dati di partenza sia su quelli copiati garantisce il servizio di intercettazione nel possedere sempre una registrazione in originale; bisogna ricordare, infatti, che il flusso di registrazioni su *hard disk* è temporaneo e dopo un certo periodo di tempo i dati acquisiti sul *server* centrale vengono cancellati, lasciando quindi esistere l'*originale* solo nella versione *copia*. La tecnologia digitale, con tutte le procedure di sicurezza che prevede, permette di garantire autenticità a qualsiasi tipo di documento.

L'impronta e la firma digitale

Alcuni dei passi da progettare sono: fare in modo, attraverso la firma digitale, che la registrazione effettuata con sistema digitale non venga manomessa e alterata durante la fase di copia; certificare l'originale, differenziarlo dalla copia e creare un documento informatico. Solo in questo modo si garantisce l'integrità del documento anche in caso di copia.

Per procedere è necessario introdurre il concetto di *impronta*. Dato un documento di lunghezza arbitraria, un algoritmo crittografico di Hash produce una stringa di lunghezza fissa detta anche *impronta*. Detta impronta ha le seguenti caratteristiche: resistenza alle pre-immagini (infatti è impossibile ricostruire il documento a partire dall'impronta); resistenza alle collisioni (l'impronta è unica per ogni documento); resistenza alle correlazioni (una piccola modifica in un documento genera una grande modifica nell'impronta).

Quindi, riassumendo, il *file* o onda sonora viene convertita in formato digitale con l'apposizione dell'impronta (Hash a 160 bit) a questo può essere applicata una K privata, una *firma digitale* che certifichi l'unicità e l'originalità, trasformando il formato digitale in *documento informatico*.

Di seguito vengono presentati alcuni schemi standard che riguardano la firma digitale, la creazione e il contenuto di un file formato P7M e una comparazione tra due impronte con conseguente ricerca della originalità di un documento informatico.

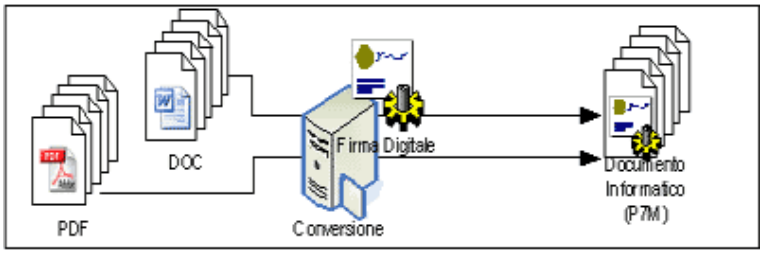


Figura 4 - Firma digitale

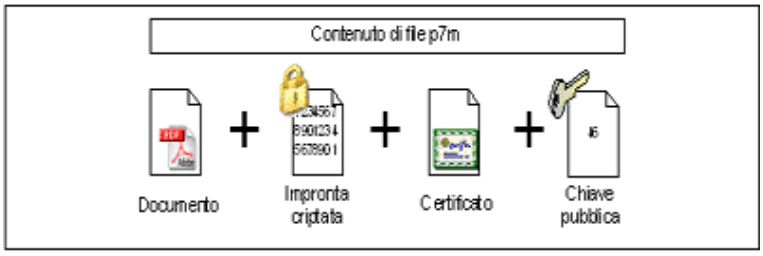


Figura 5 - Contenuto del file P7M

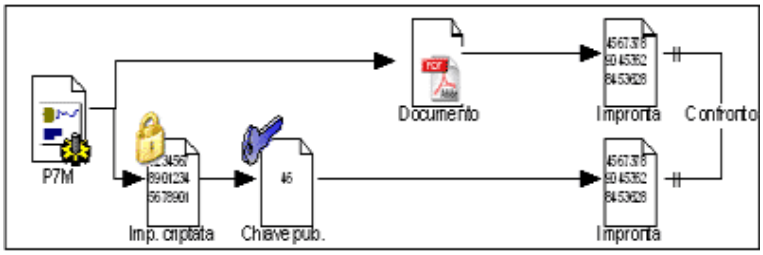


Figura 6 - Comparazione tra due impronte e quindi due documenti informatici

La struttura del *database*

Da uno studio condotto dall'istituto di legge criminale internazionale Max Planck in Italia si effettuano più intercettazioni che in tutti gli altri paesi d'Europa. Diventa quindi un problema la catalogazione e la conservazione di tutti questi documenti sonori o supporti.

Oggi, nella maggior parte dei casi il materiale *audio* finisce in scatoloni e in ambienti che non rispettano le misure necessarie per la conservazione di tali supporti. I fattori che influenzano maggiormente l'alterabilità dei supporti *audio* sono la polvere, le impurità, l'umidità, la temperatura e i campi magnetici; ma anche l'integrità del supporto è di fondamentale importanza: infatti una minima deformazione può risultare il fattore determinante per la perdita di informazioni. Il tutto viene complicato, come detto già in precedenza, dal fatto che i supporti da archiviare e catalogare sono sia digitali che analogici e i supporti sono di infinita diversità.

Riportiamo di seguito solo alcuni degli standard AES:

- AES7_2000 (r2005) *standard* AES per la preservazione la restaurazione delle registrazioni *audio*. Metodo per la misurazione del flusso registrato delle registrazioni sonore magnetiche a lunghezza d'onda media (Revisione dell'AES7-1982);
- AES22_1997 (r2003) raccomandazioni pratiche dell'AES per la preservazione e la restaurazione dell'*audio* – storaggio e manutenzione – storaggio di nastri magnetici basati su poliestere;
- AES22_1996 (r2002) raccomandazioni pratiche dell'AES per scopi legali. Trattamento del materiale *audio* registrato espressamente per essere sottoposto a esame;
- AES43_2000 (r2005) *standard* AES per scopi legali. Criteri per l'autenticazione di registrazioni analogiche *audio* su nastro.

È da questa constatazione dei fatti che nasce la necessità di pensare e progettare un *database dinamico* che dia la possibilità di archiviare il documento informatico frutto di intercettazione e di associare a esso alcuni metadati riguardanti tutte le informazioni

legate in qualche modo alla registrazione. I metadati dovrebbero costituire in qualche modo il *curriculum* o *la storia della registrazione*. L'archiviazione potrebbe essere organizzata in moduli fissi contenenti i dati ed in moduli dinamici contenenti i metadati. Questi ultimi conterrebbero informazioni analitiche e dettagliate sul *file* come il tipo di registrazione, se si tratta di registrazione ambientale o telefonica, il canale utilizzato, il numero di procedimento penale, l'ufficio richiedente (si può verificare il caso in cui la registrazione sonora sia stata già oggetto di valutazione in altro procedimento), la trascrizione dialettale e la sua eventuale traduzione/interpretazione, i commenti e le note di operatori e consulenti, le operazioni di filtraggio con il relativo *file* sonoro associato ecc. Una corretta archiviazione delle registrazioni e anche una corretta gestione dei dati potrebbe rappresentare un punto di partenza per la creazione di un *corpus* di voci anonime e voci note ai fini del riconoscimento del parlatore, nel caso di perizia di comparazione per la costituzione di una comunità linguistica geograficamente dettagliata, l'omologazione di alcuni metodi e procedure di analisi quali i filtraggi, le statistiche utilizzate, ecc. Il modello del *database* potrebbe essere il seguente:

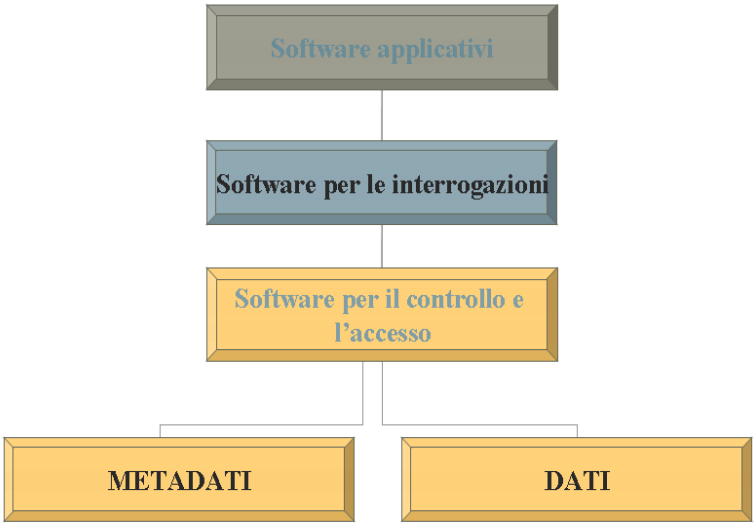


Figura 7 - Modello del Database

I dati sarebbero costituiti dalla registrazione sonora originale:

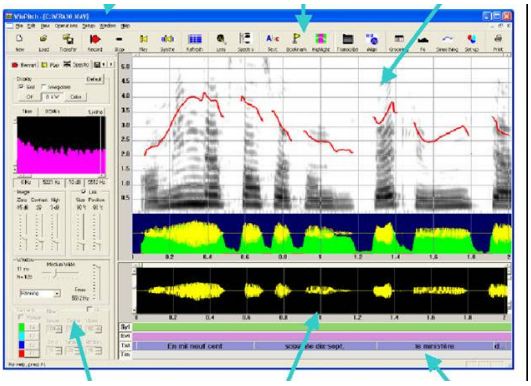


Figura 8

mentre invece i metadati avrebbero una sezione ovviamente fissa, certificata, immutabile e inalterabile e una sezione dinamica e aggiornabile.

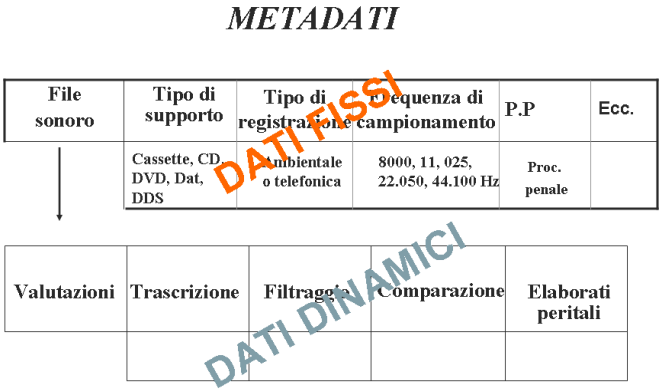


Figura 9 - Schema del modulo dei Metadati

Conclusioni

Il Codice di Procedura Penale precisa che le intercettazioni di conversazioni o comunicazioni (art. 226-271) rientrano tra i mezzi di ricerca della prova ai fini dell'accertamento della verità e la perizia (art. 220-223) rappresenta il mezzo di prova. Di conseguenza, le intercettazioni come mezzo di ricerca della prova e la validità giuridica

che assume la perizia diventano di cruciale importanza in un processo e *i risultati raggiunti sono pienamente utilizzabili dal giudice ai fine della decisione*.

Questo lavoro nasce dall'osservazione di tutto il processo di captazione in ambito forense: dalla installazione della microspia, alla registrazione e intercettazione, dalla copia all'archiviazione e alla catalogazione sia dei materiali che dei supporti.

I problemi riscontrati in questo campo, da un punto di vista meramente scientifico ovviamente, sono tanti, alcuni di questi potrebbero essere risolti solo attuando delle piccole procedure e rispettando dei protocolli già noti ed esistenti; in altri casi, invece, abbiamo proposto l'impronta e la firma digitale per la certificazione del *file* originale e quindi per garantire la validità giuridica della registrazione o la progettazione di un *database* che, oltre a sopperire a gravi carenze di archiviazione e catalogazione, risolverebbe il grave problema dell'infinità di formati diversi oggi presenti in Italia. Tale diversità nei formati rende le analisi non comparabili, l'accesso molto limitato e infine un grande dispendio di risorse economiche.

Riteniamo che qualsiasi documento, al di là del suo valore caratterizzante o del supporto sul quale è registrato, acquisisce valore in relazione alla sua utilizzabilità e fruibilità. La forza dei *database* sta nel fatto che riescono a contenere informazioni di ogni genere (testi, grafica, filmati, suoni, ecc.) e di renderli immediatamente disponibili su dispositivi di basso costo e di grande diffusione. Un *database* in questo settore così delicato potrebbe essere utile non solo per il recupero del materiale *audio* intercettato o per la conservazione, la standardizzazione dei formati e la veloce consultazione di una grande mole di dati archiviati ma anche per la loro relazione in ambito di indagine.

In questo preciso momento storico, in cui si parla molto di Giustizia e di spese relative all'intercettazione, riteniamo che andrebbe effettuata una più corretta analisi dell'attuale situazione.

Il problema è che il grande dispendio economico, a nostro avviso, non è da ricercare *nello strumento*, ma nel suo utilizzo. Un gruppo di studio che affrontasse seriamente il problema del digitale e dell'informatizzazione in ambito forense con la messa appunto di reti protette e di formati unici in modo da sfruttare le potenzialità del digitale – come per esempio lo scambio veloce di dati, l'archiviazione con più chiavi, ecc. – renderebbe tutto più facile ed economico. Ad oggi, invece, paghiamo lo scotto relativo alle spese del digitale senza sfruttarne le potenzialità. Tale gruppo di lavoro scoprirebbe che molte delle voci di spesa riguardano il noleggio di differenti attrezzature. Apparati strumentali di proprietà dello Stato eviterebbero il giogo dei privati, personale dipendente preparato attraverso percorsi formativi strutturati per l'occorrenza migliorerebbero le acquisizioni e quindi i risultati accorciando i tempi delle indagini e riducendo di molto i costi.

Note

[1] L'impostazione della ricerca è di L. Romito, la ricerca documentale è di M. Tucci mentre il § *Impronta e firma digitale* è di G. Cavarretta.

[2] Il Ministero di Grazia e Giustizia precisa che per quelle Procure che non superano i 200 bersagli annui non è conveniente evolvere verso il sistema informatizzato, potendosi conseguire già apprezzabili risultati con i sistemi tipo RT 6000.

[3] Articolo apparso sulla Gazzetta del Sud del 27 giugno 2008

Bibliografia

Bellucci P., (2002), *A onor del vero. Fondamenti di linguistica giudiziaria*, UTET.

Paoloni A., Zavattaro D., (2007), *Intercettazioni telefoniche e ambientali*, Centro Scientifico Editore.

Romito L., (2000) *Manuale di fonetica articolatoria, acustica e forense*, Università degli Studi della Calabria: Centro editoriale e librario.

Romito L., (2003), *Passato presente e futuro nelle analisi di* speaker recognition. In "Voce Canto Parlato", Zamboni A. (a cura di), Padova: Unipress, p. 237-246.

Romito L., Galatà V., (2004), *Towards a protocol in speaker recognition analysis* in "Forensic Science International", p. 105-113.

Romito L., (2004), *La misura dell'intelligibilità e il rapporto segnale-rumore*. "*Atti del convegno* AISV, Associazione Italiana di Scienze della Voce", Padova, 2004.

Romito L., (2005), *La competenza linguistica nelle trascrizioni f: l'intelligibilità, l'oggettività e il rapporto segnale/rumore*. "Detective And Crime".

L'attribuzione di testi con metodi quantitativi: riconoscimento di testi gramsciani [1]

CHIARA BASILE, MAURIZIO LANA

This contribution takes account of some theoretical considerations and experimental results developed by a quite heterogeneous group of scientists while studying the problem of quantitative authorship attribution (stylometry) in a selection of newspaper articles written by Antonio Gramsci and his co-workers. Methodological and historical issues of stylometry are discussed, concentrating particularly on the relatively recent spread of methods based on very simple indicators with no syntactical meaning. Two methods for the recognition of authorship are then described and experimented on the Gramscian corpus, both based on a mathematical model of texts and the author/text relationship, and both using similarity distances. The first method compares the statistics for sequences of n characters (n-grams) in the texts, while the second is based on the very deep concept of entropy of a symbolic sequence and on some techniques for data compression.

Keywords: Stylometry – attribution by quantative methods – stylistic fingerprinting – similarity differences – n-grams, data compression

Presentazione

La ricerca qui descritta trae origine da una scelta innovativa della Fondazione Istituto Gramsci, nelle persone del presidente, Giuseppe Vacca, e del curatore di una delle annate dell'"Edizione Nazionale delle opere di Antonio Gramsci", Leonardo Rapone: chiedere l'intervento di un gruppo di esperti in attribuzione di testi con metodi quantitativi per individuare gli scritti gramsciani all'interno di un *corpus* di articoli giornalistici pubblicati anonimi [2] negli anni 1913-1926 sui quotidiani ai quali Gramsci collaborava (*Il Grido del Popolo, Avanti!, La Città Futura*, …), allo scopo di offrire ai curatori delle varie annate degli scritti di Gramsci una nuova chiave di lettura per valutare l'inclusione (o esclusione) di testi nell'Edizione Nazionale. La ricerca è condotta da un gruppo formato da Dario Benedetto e Emanuele Caglioti, fisici matematici all'Università La Sapienza di Roma, Mirko Degli Esposti, fisico matematico dell'Università di Bologna, e dagli autori di queste pagine. Della ricerca nel suo complesso renderanno conto due contributi di imminente pubblicazione in un volume metodologico dell'Edizione Nazione delle Opere di Antonio Gramsci [3].

Note metodologiche e storiche sull'attribuzione di testi per mezzo di analisi quantitative

Stilometria e stylistic fingerprint

La stilometria, fondata sulla convinzione che sia possibile misurare le caratteristiche stilistiche di un testo, procede quantificando le caratteristiche del testo, scelte solitamente tra quelle ancorabili alle parole (frequenze di parole di specifici tipi, per esempio: congiunzioni, preposizioni, rapporto nomi/aggettivi, e così via): contando e misurando gli elementi caratteristici dello stile si spera di scoprire le caratteristiche di uno specifico autore. All'origine c'è infatti l'idea che ogni testo abbia uno stile caratteristico e che di conseguenza testi che hanno caratteristiche stilistiche molto simili siano del medesimo autore (il modo in cui questo criterio si attua nella realtà è ovviamente ben più complesso della semplice formulazione adottata qui). Il termine stile è quindi usato non nell'accezione di *qualità estetica distintiva*, o di norma espressivo-compositiva, ma nell'accezione di *caratteristica espressiva individuale*.

Se, come, quanto, gli aspetti di stile inconsci siano stabili nel corso del tempo, oppure siano soggetti ad evoluzione, è questione complessa e non risolta: quindi questione a cui dedicare molta attenzione. Inoltre si assume che le caratteristiche di stile che derivano da scelte inconsce non possano essere consapevolmente modificate (la cosa dovrebbe essere ovvia se si tratta di scelte inconsce; ma ciò che viene definito come inconscio potrebbe in realtà risultare frutto di scelte almeno in parte consce). L'attribuzione dei testi per mezzo delle caratteristiche di stile inconsce attira molta attenzione poiché *se* le caratteristiche di stile inconsce sono strettamente connesse con l'identità dell'autore, *e se* le caratteristiche inconsce reperite in una serie di testi di un autore si ripresentano in uno o più testi dubbi, *allora* si può concludere che i testi dubbi sono di quell'autore, con tutta la forza logica derivante dal procedimento. Un'altra questione delicata è quella dell'interferenza tra caratteristiche di stile autoriali e caratteristiche di stile derivanti dal contenuto (in primo luogo lessico e fraseologia), interferenza descritta in modo efficace da R. Clement e D. Sharp [4].

L'idea recente di una struttura matematica latente dei testi

Tra la fine del secolo scorso e l'inizio del secolo attuale, con lo sviluppo delle procedure di analisi qualitative inizia ad affermarsi la consapevolezza che esistono nei testi strutture matematiche descrivibili solo in termini quantitativi.

L'approccio tradizionale allo studio di un testo era (è) quello per cui di fronte ad una serie di problemi si ricorre ai consueti strumenti di tipo semantico, linguistico,

storico, cercando di ampliare e approfondire le conoscenze sul contenuto del testo, sulla lingua in cui è scritto, sulla storia della sua composizione e trasmissione. Può però succedere che non si facciano progressi significativi e si può allora decidere di ricorrere allo studio nel testo di caratteristiche di tipo differente. Quando ciò avviene è perché in primo luogo ci si rende conto che il modello di studio del testo basato sul contenuto non risponde alle domande che lo studioso sta ponendo; in secondo luogo perché si ipotizza, con un atteggiamento esplorativo, che per cercare le risposte si debba costruire un differente modello di ricerca in cui contenuto, lingua, storia del testo vengono messi in secondo piano e si sposti l'attenzione dagli oggetti ai numeri, dall'individuazione degli oggetti al loro conteggio; o, in termini più formalizzati, si cerca di passare da un sistema qualitativo ad un sistema quantitativo (Doležel [5]) trasformando un sistema di relazioni qualitative (il testo) in un sistema di relazioni quantitative (l'insieme dei dati che contiene le informazioni sugli oggetti dell'analisi) grazie ad una o più operazioni di *classificazione* del testo. Se cadono i vincoli semantici, linguistici, grammaticali, la scelta degli oggetti misurabili è potenzialmente illimitata [6].

Se i due sistemi si corrispondono perfettamente, la situazione risulta per certi versi insoddisfacente perché il passaggio dal sistema qualitativo a quello quantitativo non ha messo in luce nulla di *nuovo e rilevante*. Ma se tra i due sistemi appaiono delle discrepanze, se si percepiscono differenze, allora diventa necessaria una riorganizzazione delle conoscenze allo scopo di capire come si rapportano i dati dei due sistemi e che cosa significano le differenze tra l'uno e l'altro. Alla base di questa riorganizzazione della conoscenza, che è nuova conoscenza, sta *il riconoscimento che i dati quantitativi fungono da indicatori della presenza nel testo di proprietà qualitative che non appaiono in evidenza al livello semantico*. Il lavoro dello studioso mira ad individuare i fattori qualitativi, formali, stilistici, da cui dipendono i valori dell'indicatore quantitativo. La situazione si fa difficile quando si abbia motivo di ritenere che uno specifico indicatore quantitativo sia controllato da molteplici fattori qualitativi (la lunghezza delle frasi di un testo dipende da molteplici fattori: idiosincrasie dell'autore, pubblico di riferimento, argomento, testi e stili di riferimento, e altro ancora), che non sono necessariamente sempre i medesimi in testi differenti.

Si possono segnalare alcune ricerche recenti particolarmente significative in quanto coerenti con questa tendenza al riconoscimento dell'esistenza di strutture matematiche nei testi. Ciò non significa dimenticare che sono oggi preponderanti per numero le ricerche che utilizzano le tecniche statistiche multivariate per estrarre e mostrare l'informazione contenuta nelle matrici di dati. Ma il versante delle ricerche che riconoscono l'esistenza di strutture matematiche nei testi pare particolarmente interessante; e lo studio di attribuzione di testi con metodi quantitativi effettuato per individuare gli scritti gramsciani anonimi si colloca in questa linea.

Nel 2001 Dmitri Khmelev (che aveva già pubblicato nel 2000 un articolo di argomento simile [7]) e Fiona Tweedie pubblicarono un articolo [8] in cui mostrarono i risultati che si potevano ottenere con una tecnica di attribuzione di testi basata su catene di Markov.

Nel 2002 D. Benedetto, E. Caglioti e V. Loreto, nell'articolo *Language Trees and Zipping* [9], proposero di utilizzare un programma di compressione dati come strumento per misurare la similarità tra sequenze differenti, e quindi come base per la classificazione ed il riconoscimento di sequenze simboliche [10]. Questo lavoro è di particolare importanza per l'attribuzione degli scritti gramsciani, come si vedrà più avanti.

Nel 2003 R. Clement e D. Sharp pubblicarono un articolo intitolato *Ngram and Bayesian Classification of Documents for Topic and Authorship* [11] in cui mostrarono di ottenere attribuzioni di testi molto soddisfacenti utilizzando *n*-grammi e *naive Bayes classifiers*.

Ad hoc Authorship Attribution Competition

Un'altra testimonianza significativa della progressiva crescita di importanza degli approcci matematici all'attribuzione dei testi è costituita dalla "Ad hoc Authorship Attribution Competition" (AAAC) bandita nel 2003 da Patrick Juola, matematico della Duquesne University di Filadelfia che da tempo si interessava di problemi di attribuzione [12], e dai risultati di rilevo che al termine di essa furono ottenuti con metodi di attribuzione di tipo matematico.

La gara si svolse su 13 set testuali scritti in varie lingue o varianti: inglese contemporaneo, del XIX secolo, di scrittori elisabettiani; *middle English*; francese; serbo-slavonico; latino; olandese; e formati da campioni appartenenti a 2 o più autori. Per ogni autore c'erano 2 o più campioni che si dovevano classificare correttamente individuando quelli del medesimo autore. In alcuni set era presente 1 campione testuale di argomento simile ma che non apparteneva a nessuno degli autori di cui erano forniti 2 o più campioni: un campione spurio che serviva a verificare la finezza operativa del metodo di analisi. La composizione dei set ovviamente non era nota ai partecipanti.

Coloro che ottennero i migliori risultati furono Moshe Koppel e Jonathan Schler (Università Bar-Ilan, Israele), Vlado Kešelj (Dalhousie University, Canada), David Hoover (New York University, USA) e Patrick Juola stesso. Tanto Kešelj quanto Juola avevano adottato un metodo di classificazione basato su n-grammi; Koppel e Schler avevano invece adottato un metodo differente, nel quale assumevano come dato di partenza le tradizionali misure stilometriche, sulla base delle quali un algoritmo di apprendimento costruiva un modello delle relazioni tra i testi del *corpus* in esame,

modello di cui si verificava la robustezza a fronte di degradazioni introdotte intenzionalmente [13].

Riflessioni dopo l'AAAC

Alla fine dell'800 Lutosławski [14] riteneva che i marcatori stilistici si collocassero in strutture grammaticali e sintattiche complesse; all'inizio del 1900 Mendenhall [15] costruiva le curve di frequenza della lunghezza delle parole; e in generale fin verso la fine del 1900 oggetto delle misurazioni stilometriche erano le parole e le unità più ampie composte di parole (sintagmi, segmenti, frasi) [16]. In anticipo sui tempi Ledger [17] nel 1989 indicò la tendenza degli anni di fine secolo per cui vengono studiate caratteristiche del testo non descrivibili in termini qualitativi, di significati. Infine, nel 2003, in occasione dell'AAAC, tra i metodi di analisi complessivamente più efficaci ci furono quelli che operano sui testi al livello degli *n*-grammi: il testo "diventa" un'unica lunghissima sequenza alfanumerica letta a blocchi di *n*-grammi (metodo che richiama l'analisi del DNA) [18].

Da Lutosławski all'AAAC si manifesta quindi una tendenza per cui dall'analisi di caratteristiche sintattiche complesse, individuabili solo da parte di una mente umana molto competente, si passa all'analisi di caratteristiche sempre più elementari fino ad arrivare agli *n*-grammi. Non mancano esempi di ricerche che non rientrano – in parte o in tutto – in questo schema interpretativo (il lavoro di Koppel e Schler ne è un esempio), ma non al punto da invalidarlo.

L'attribuzione degli scritti giornalistici di Gramsci

La fase preliminare di scelta e messa a punto dei metodi di analisi: l'esperimento controllato

Il problema dell'attribuzione degli scritti gramsciani anonimi si caratterizza per il fatto che si deve "solo" decidere se un testo è o non è di Gramsci, non attribuire al giusto autore ciascuno dei testi anonimi, il che facilita il lavoro; ma anche per il fatto che i testi incogniti sono omogenei sotto vari punti di vista (tipo di fonte, collocazione cronologica, argomento, presumibile e ampia condivisione di linguaggio e idee da parte degli autori); e che spesso questi testi sono molto brevi (poche centinaia di parole): tutti aspetti che rendono difficile e complesso il lavoro. Non si aveva garanzia che i metodi noti potessero operare efficacemente su questo problema così specifico (per omogeneità e brevità dei testi), per non parlare del fatto che non c'è consenso riguardo a quali siano le caratteristiche stilistiche di un autore da impiegare in uno studio quantitativo e se

esse siano stabili nel tempo. Si dovette quindi procedere in via sperimentale e per questo ad una fase preliminare di messa a punto di metodi e procedure seguì un test cieco con lo scopo di verificare l'adeguatezza dei metodi sviluppati.

Nella fase di messa a punto (maggio-giugno 2006) fu realizzato un esperimento controllato utilizzando 100 testi, 50 di Gramsci e 50 di altri autori. I due metodi per l'attribuzione dei testi messi a punto durante la fase preliminare sono descritti più avanti e portarono alle attribuzioni descritte in Figura 1. Furono attribuiti a Gramsci i soli testi che entrambi i metodi identificavano come gramsciani, cioè 43 su 50. In Figura 1 il quadrante in alto a destra contiene i testi gramsciani (punti rossi) correttamente attribuiti. I punti rossi negli altri quadranti (g07, g35, g14, …) rappresentano i testi gramsciani non correttamente attribuiti. Nel quadrante in alto a destra non ci sono testi non gramsciani (punti blu), cioè il sistema non genera falsi positivi (testi attribuiti a Gramsci benché non siano di Gramsci).

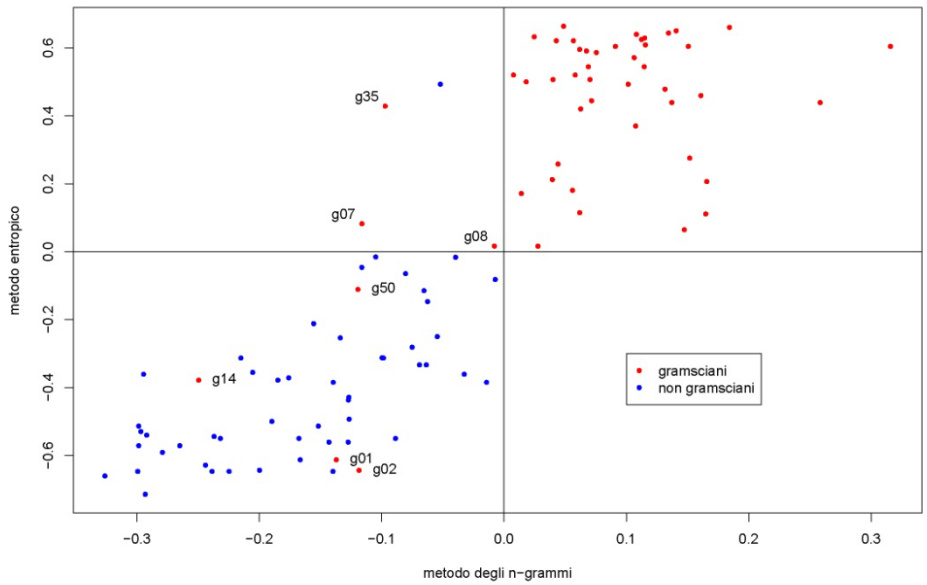


Figura 1 Le attribuzioni dei 100 testi di messa a punto, al termine della fase preliminare

Il test cieco

Il test cieco, in cui il gruppo di ricerca non conosceva le corrette attribuzioni dei testi, fu effettuato su 40 articoli gramsciani e non gramsciani utilizzando come riferimento i 100 articoli di attribuzione nota impiegati nell'esperimento controllato. I

risultati sono mostrati in Figura 2, costruita con il medesimo procedimento usato per Figura 1 (nel quadrante in altro a destra ci sono i testi che entrambi i metodi attribuiscono a Gramsci).

Come si può osservare, vennero correttamente individuati e attribuiti 18 testi gramsciani su 20, pari al 90%, senza falsi positivi (testi non riconosciuti: *Due inviti alla meditazione*, "La Città futura", 11 febbraio 1917, e *I monaci di Pascal* ,"Avanti!", 26 febbraio 1917).

Ottenere un risultato così chiaro e significativo andò al di là delle aspettative del gruppo di ricerca, consapevole della complessità del dominio in cui operava; ma d'altra parte impone prudenza il fatto che l'attribuzione dei testi gramsciani anonimi, essendo un problema reale e non un caso di studio, superata la fase del test cieco non potrà più avere prove o controprove sperimentali. Di qui una questione – irrisolvibile – di verificabilità. Resterà pertanto responsabilità dei curatori delle varie annate dell'Edizione Nazionale decidere se accogliere le attribuzioni emerse dall'analisi quantitativa.

Poiché anche alle attribuzioni di testi con metodi quantitativi potrebbe applicarsi in qualche misura il principio espresso da A. Clarke per cui "any sufficiently advanced technology is indistinguishable from magic" [19], il che non vorremmo accadesse, è necessario a questo punto descrivere gli aspetti matematici dei metodi utilizzati per l'attribuzione.

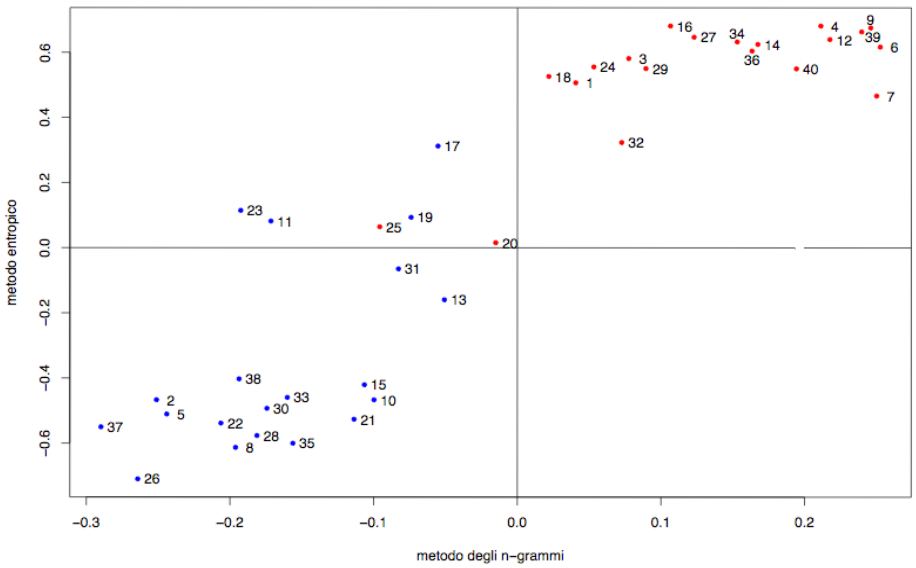


Figura 2 - Attribuzioni del test cieco

Un modello astratto per la scrittura dei testi

Il punto di partenza per lo sviluppo dei nostri metodi consiste nel considerare un testo "semplicemente" come una sequenza di simboli in un dato alfabeto; come già accennato precedentemente, si tratta di un approccio più astratto e generale rispetto a quello abitualmente utilizzato in ambito linguistico, dal momento che tutte le strutture "naturali" insite in un testo sono trascurate. Consideriamo infatti un insieme $A = \{a_1,...,a_N\}$ di N simboli (caratteri), che chiameremo d'ora in avanti *alfabeto*, e l'insieme A^* delle sequenze (o *stringhe*) costituite da un numero finito di simboli di A. Ad esempio, se A è formato dalle 21 lettere dell'alfabeto italiano maiuscole e minuscole, dal punto, dalla virgola e dallo spazio separatore (e dunque contiene $N = 42 + 3 = 45$ simboli), un possibile testo in A^* è:

State leggendo un testo in italiano, chiaramente

oppure anche:

A,hs heb,p.Fr ., eem.baDAqfgbzp.i,tan

Nel caso del *corpus* gramsciano l'alfabeto è costituito da 84 simboli: le lettere della lingua italiana (minuscole e maiuscole, accentate e non), con qualche lettera degli alfabeti stranieri; i più comuni simboli di interpunzione; lo spazio separatore; le cifre da 0 a 9.

Adottando questo punto di vista, non ha senso concentrarsi sulla struttura grammaticale o sintattica del testo, né distinguere il ruolo delle lettere da quello dei segni di interpunzione. Avendo completamente destrutturato il testo, le parole perdono il loro ruolo di elementi base del testo stesso; tale ruolo sarà invece assunto dagli *n-grammi*, ovvero le sequenze *qualsiasi* di *n* caratteri, con *n* parametro fissato, presenti nella stringa. Può essere utile fare qualche esempio:

- un *monogramma* (1-gramma) è un qualsiasi carattere dell'alfabeto;
- un *bigramma* (2-gramma) è una sequenza di due caratteri dell'alfabeto, come **la** o anche **a,**;
- un *trigramma* (3-gramma) è una sequenza di tre caratteri dell'alfabeto, come **int** o anche **e l**;
- **l pr** è un 4-gramma, **l prole** è un 7-gramma, **l prolet** è un 8-gramma...

A questo punto, utilizzando un modello proveniente dalla teoria dell'informazione, si può considerare un autore come una *sorgente* che genera stringhe in un certo alfabeto; la sorgente è un'entità astratta, del cui funzionamento non ci si interessa, e che si pensa completamente determinata dalle *regole* secondo le quali essa genera le sequenze simboliche. Ad esempio, un autore seguirà in generale le regole grammaticali della lingua in cui scrive, così che un autore italiano non genererà il testo **questo sono un testi sbagliate**; è d'altra parte evidente che le sole regole grammaticali non sono sufficienti ad

identificare un singolo autore come sorgente: dovranno invece intervenire altri tipi di "regole", che siano caratterizzanti dello stile di quel particolare autore.

Conoscere le regole (o, per dirlo in modo più formale, la *distribuzione di probabilità* per la generazione delle stringhe) per un dato autore/sorgente significherebbe essere in grado di stabilire se un certo testo è stato generato o meno da quell'autore. D'altra parte, però, anche ammettendo che questo modello sia effettivamente utile a descrivere il rapporto tra un autore ed i testi che egli genera, non è possibile conoscere a priori tali regole; quello che invece abbiamo a disposizione sono i testi di riferimento, ovvero alcune realizzazioni particolari della sorgente/autore (generate secondo regole che rimangono incognite). Quello che cerchiamo sono essenzialmente dei modi per ricostruire le regole della sorgente/autore a partire da misure effettuate su alcune sue realizzazioni, che considereremo come fossero *esempi casualmente generati* dalla sorgente.

Distanze di similarità

Per un matematico l'idea di "vicinanza" tra testi di una stessa classe (nel nostro caso di uno stesso autore) si esprime in modo naturale definendo una funzione *distanza* che sia in grado di misurare la *similarità*, o meglio la *dissimilarità*, tra ogni data coppia di stringhe (testi) [20]. L'esempio più semplice di distanza tra sequenze simboliche è la cosiddetta distanza di Hamming [21] che conta semplicemente il numero di caratteri diversi nelle due stringhe; ad esempio se x = **unasequenza** e y = **duesequenze** sono due stringhe nelil medesimo alfabeto A la loro distanza di Hamming è $d_H(x, y) = 4$, perché x e y si differenziano nei primi tre caratteri (rispettivamente **una** e **due**) e nell'ultimo (rispettivamente **a** e **e**).

Le metriche che utilizzeremo per l'attribuzione saranno un po' più complicate di quella di Hamming, e saranno definite in base alle misure che si vogliono effettuare sui testi; in base cioè alle quantità che si ritengono importanti per ricostruire le "regole" dell'autore/sorgente a partire dai suoi testi. Sottolineiamo anche che spesso tali metriche non saranno vere e proprie distanze dal punto di vista matematico; tralasceremo però in questa sede ogni considerazione a questo proposito, continuando a chiamare "distanze" anche funzioni che non soddisfano tutte le proprietà della definizione matematica di distanza.

La distanza degli *n*-grammi

Una prima misura che è naturale prendere in considerazione è la statistica degli *n*-grammi; si può cioè pensare che la frequenza con cui un autore utilizza i diversi

n-grammi nel processo di scrittura sia in qualche modo caratterizzante di quell'autore [22]. Modificando un'idea proposta originariamente da Vlado Kešelj nel 2003 [23] e poi da lui stesso utilizzata per la AAAC del 2004 [24], definiamo la distanza tra due testi *x* e *y* scritti nelil medesimo alfabeto *A* con la seguente formula:

$$d_n(x,y) = \frac{1}{|D_n(x)| + |D_n(y)|} \sum_{\omega \in D_n(x) \cup D_y(y)} \left(\frac{f_x(\omega) - f_y(\omega)}{f_x(\omega) + f_y(\omega)} \right)^2,$$

dove $D_n(x)$ è il *dizionario degli n-grammi* di *x*, ovvero l'insieme di tutti gli *n*-grammi contenuti nel testo *x*, $|D_n(x)|$ è il numero di elementi di $D_n(x)$ e $f_x(\omega)$ è la frequenza nel testo *x* dell'*n*-gramma ω. Considereremo sempre *n*-grammi *con sovrapposizione*, cioè ad esempio D_4 (**stringa**) = {**stri**, **trin**, **ring**, **inga**}.

Si osservi che nella formula qui sopra i termini della sommatoria:

- sono minori quando le frequenze di ω in *x* e in *y* si somigliano: testi con statistiche degli *n*-grammi simili sono dunque più "vicini" per questa distanza;
- sono maggiori quando $f_x(\omega)$ e $f_y(\omega)$ sono piccoli, per via del denominatore $[f_x(\omega) + f_y(\omega)]^2$: ciò significa che i termini corrispondenti ad *n*-grammi più rari danno un contributo maggiore alla distanza.

Il fattore moltiplicativo $1/(|D_n(x)| + |D_n(y)|)$ ha lo scopo di eliminare la dipendenza della distanza dalla dimensione dei dizionari, e quindi in particolare dalle lunghezze dei testi confrontati.

La versione di Kešelj della distanza è un po' diversa; le modifiche che sono state apportate derivano da considerazioni sulla particolarità del *data set* in esame (testi brevi, di molti autori diversi) e dalle indicazioni ottenute dai risultati del primo esperimento controllato, che hanno suggerito l'utilizzo di *n*-grammi piuttosto lunghi, *n* = 7, 8, 9.

La distanza entropica

Una quantità standard, definita per la prima volta da C. E. Shannon nel 1948 [25], è l'*entropia* di una stringa simbolica. Sarebbe fuori luogo in questa sede darne una definizione matematica precisa; diremo semplicemente che l'entropia misura, in un qualche senso, la *quantità di informazione* contenuta in una stringa: sequenze più regolari avranno un valore dell'entropia minore, mentre essa sarà massima per sequenze casuali, in cui cioè le regolarità sono inesistenti. Facendo un esempio semplice: la sequenza

0101010101010101010101010101010101010101010101,

generata semplicemente ripetendo 25 volte il bigramma **01**, avrà entropia molto minore della sequenza

01011110101010110111110111101100011011100001000110,

ottenuta lanciando 50 volte una moneta e scrivendo **0** quando esce testa e **1** quando esce croce. Si può pensare l'entropia come la "sorpresa" che la stringa è in grado di generare in un ipotetico "lettore": per conoscere la prima delle due sequenze riportate qui sopra è sufficiente leggerne i primi 2 caratteri, dopodiché la lettura della parte rimanente della stringa non porta "nessuna novità"; nella seconda stringa, invece, l'assenza di regolarità fa sì che non si possano fare previsioni ragionevoli circa il simbolo che seguirà nemmeno ricordando tutto quanto si è letto fino a quel momento. Esistono programmi ideati precisamente per sfruttare le regolarità all'interno delle sequenze simboliche in modo da codificarle occupando la minor quantità possibile di memoria all'interno di un computer: sono i compressori di dati, come *gzip, Winzip* o *Winrar*.

A. Lempel e J. Ziv pubblicarono nella seconda metà degli anni '70 una serie di interessantissimi articoli [26] in cui proposero alcuni algoritmi per la compressione di dati, che sono alla base dei moderni programmi di compressione. Essi dimostrarono anche che il rapporto tra lunghezza della sequenza compressa e lunghezza della sequenza originaria (*rate di compressione*) è un buon indicatore dell'entropia di una stringa. L'idea che sta alla base degli algoritmi di Lempel e Ziv (LZ) è abbastanza semplice: la sequenza viene letta dall'inizio alla fine e ogni volta che si trova un n-gramma che è già apparso precedentemente esso viene sostituito nella sequenza compressa da due indici che fanno riferimento al punto in cui tale n-gramma era apparso. In questo modo le regolarità della stringa sono utilizzate per ridurre la quantità di memoria necessaria per archiviarla.

Se l'entropia misura quanto è difficile codificare una stringa per mezzo delle sue regolarità, l'*entropia relativa* esprime (ancora una volta, senza essere troppo rigorosi) la "sorpresa" che si ha leggendo la seconda stringa dopo aver letto la prima. Facendo un esempio semplice: testi in lingua inglese ed in italiano contengono entrambi delle regolarità, date ad esempio dalla ripetizione di articoli, congiunzioni e parole particolarmente comuni. La conoscenza di un modo molto efficiente per codificare l'inglese utilizzando le ridondanze che lo caratterizzano, però, non permette di codificare efficacemente un testo in italiano, perché i tipi di regolarità (sequenze più frequenti, struttura delle frasi...) delle due lingue sono diversi.

Già nel 1993 J. Ziv e N. Merhav [27] dimostrarono che i compressori di dati potevano essere usati per stimare l'entropia relativa tra due stringhe simboliche; successivamente Benedetto, Caglioti e Loreto [28] svilupparono un metodo efficace per la stima

dell'entropia relativa, che illustriamo ora brevemente. Supponiamo di comprimere il testo *A*+*X*, cioè il testo che si ottiene mettendo in sequenza il testo *A* e il testo *X*. L'algoritmo di compressione, che è sequenziale, codificherà prima tutti i caratteri di *A* e poi inizierà a codificare i caratteri di *X*, cercando le stringhe nella parte già letta, cioè dentro il testo *A*. Tanto più i due testi saranno simili tanto più lunghe saranno le stringhe di *X* trovate in *A*, e quindi più efficacemente sarà compresso il file complessivo. Il compressore infatti può in questo caso utilizzare non solo la ridondanza all'interno dei singoli testi, ma anche la ridondanza tra i due testi, migliorando il rapporto di compressione. La differenza di lunghezza tra la versione compressa del testo *A*+*X* e del testo *A*, divisa per la lunghezza di *X*, è una misura dell'entropia relativa del testo *X* rispetto ad *A*. Tale numero è tanto più piccolo quante più parti di *X* vengono trovate in *A* o, in modo più suggestivo, quanto più facile è esprimere il contenuto del testo *X* conoscendo il contenuto di *A* [29].

La seconda distanza utilizzata per gli esperimenti sul *corpus* gramsciano sfrutta proprio questa idea di Benedetto, Caglioti e Loreto. In seguito al primo esperimento controllato si è deciso in questo caso di lavorare su testi tutti della medesima lunghezza, accorpando e poi tagliando a lunghezze uguali tutti i testi di Gramsci (e, rispettivamente, non gramsciani) del *corpus*. La ragione di tale scelta è il fatto che il metodo entropico è molto sensibile alla lunghezza dei testi di riferimento. In generale, infatti, tutti i metodi che confrontano singoli testi tendono a scegliere tra quelli di dimensione maggiore il testo più vicino al testo incognito: testi più lunghi sono relativamente più ricchi di statistica e informazione, e quindi hanno maggiore possibilità di avere caratteristiche in comune con i testi incogniti. D'altra parte, mentre per gli *n*-grammi i testi che appaiono come primi vicini nelle attribuzioni hanno una dimensione che è circa 1,5 volte quella media dei 100 testi, per il metodo entropico questo rapporto è superiore a 2.

Gli indici di gramscianità e non gramscianità

Una volta che si è calcolata con qualche formula (ad esempio quella degli *n*-grammi) la distanza tra tutti i testi di riferimento e tutti i testi di test (che chiameremo anche *incogniti*), occorre decidere in che modo attribuire un testo incognito. L'idea più semplice è naturalmente quella di attribuire un testo all'autore del testo di riferimento a distanza minima da esso, il suo *primo vicino*. Questo metodo fornisce già alcuni risultati incoraggianti ma, data la ricchezza di testi del *corpus* gramsciano, che già nel primo esperimento consta di ben 100 testi di riferimento, di cui 50 gramsciani e 50 non gramsciani, ci è sembrato più opportuno tenere conto di tutta l'informazione contenuta nel *data set*. Per fare ciò abbiamo calcolato per ogni testo incognito *x* un *indice di gramscianità*

$i_G(x)$ e un *indice di non gramscianità* $i_{NG}(x)$, che esprimono la maggiore o minore vicinanza "in media" tra *x* e i due sottoinsiemi del *corpus* di riferimento, rispettivamente quello formato dai testi gramsciani e quello formato dai testi non gramsciani. Mettiamoci ad esempio nel caso del test cieco; l'indice di gramscianità è in questo caso definito da:

$$i_G(x) = \sum_{j=1}^{50} \frac{k_j}{j},$$

dove k_j è la posizione, nell'elenco (*classifica*) dei testi di riferimento ordinati per distanza crescente da x, del *j*-esimo testo di Gramsci. Un'analoga definizione vale per $i_{NG}(x)$. Si osservi che i termini k_j / j della sommatoria sono sempre maggiori o uguali a 1 (non è possibile che il terzo testo gramsciano in classifica si trovi in una posizione migliore della terza!) e l'indice di gramscianità è tanto più piccolo quanto più i testi di riferimento gramsciani occupano le prime posizioni della classifica.

Una opportuna combinazione dei due indici fornisce poi per ogni testo un indice globale, che ha valori nell'intervallo [−1, 1] ed è positivo se il metodo attribuisce il testo a Gramsci, negativo se lo attribuisce alla classe dei testi non gramsciani. Il valore assoluto di tale indice fornisce inoltre un'indicazione circa l'attendibilità dell'attribuzione: quanto più esso è vicino a zero, infatti, tanto più l'attribuzione ottenuta è da considerarsi incerta.

Per gli esperimenti sul *corpus* gramsciano abbiamo calcolato per ogni testo incognito *x* due versioni di tale indice, una per ognuna delle due distanze definite precedentemente:

- $i_n(x)$ ottenuto con la distanza degli *n*-grammi (con *n* fissato) calcolando le somme negli indici di gramscianità e non gramscianità su tutti i 50 testi gramsciani (rispettivamente, non gramsciani) dell'insieme di riferimento;
- $i_e(x)$ ottenuto con la distanza entropica calcolando le somme negli indici di gramscianità e non gramscianità solo sui primi 3 testi gramsciani (rispettivamente, non gramsciani) in classifica.

Commento ai risultati e sviluppi futuri

Torniamo ora ai risultati del test cieco mostrati in Figura 2. Per ogni testo incognito x, rappresentato nel grafico da un punto di colore corrispondente alla sua reale attribuzione (che ci è stata comunicata solo a test concluso), sull'asse orizzontale è rappresentato il valore di $i_n(x)$ per $n = 8$, mentre sull'asse verticale si trova il valore di $i_e(x)$. La rappresentazione bidimensionale sul piano cartesiano permette di visualizzare i risultati in modo particolarmente efficace, intrecciando e contemporaneamente mante-

nendo separati i contributi delle due distanze al metodo globale di classificazione. Abbiamo ad esempio osservato che quasi tutti i punti corrispondenti a testi gramsciani si trovano nel quadrante in alto a destra: questo corrisponde a dire che entrambi gli indici *in* e *ie* sono positivi. Al contrario, il testo contrassegnato con il numero 25, che si trova nel quadrante in alto a sinistra, ha indice i_n negativo e indice i_e positivo, perciò uno dei due metodi (in questo caso quello degli *n*-grammi) dà un'attribuzione scorretta. Inoltre osserviamo ancora che la distanza dall'origine dà un'indicazione circa la forza dell'attribuzione: il testo indicato col numero 20, per il quale entrambi gli indici hanno valori vicini a zero, è da considerarsi non attribuibile con queste tecniche (ed in effetti è un testo molto breve).

La scelta di utilizzare per la distanza degli *n*-grammi il valore $n = 8$ è stata motivata dai risultati del test controllato effettuato in fase di messa a punto, che hanno indicato questo valore (insieme a quelli ad esso vicini, $n = 7$ o $n = 9$) come il più efficace per l'attribuzione; l'esito del test cieco (confrontato con quello che si sarebbe ottenuto per valori diversi di *n*) ha poi confermato tale scelta.

Vogliamo sottolineare ancora una volta che il *data set* gramsciano ha delle caratteristiche di omogeneità tali da mettere davvero alla prova la capacità dei metodi utilizzati di classificare i testi *per autore*: poiché infatti i testi appartengono tutti ad un medesimo genere, sono stati scritti nella medesima epoca, da persone con un retroscena culturale simile e rivolgendosi al medesimo tipo pubblico, ed inoltre gli argomenti trattati sono trasversali a più autori, il metodo di attribuzione sperimentato su tale *data set* non può "accontentarsi" di distinguere i testi rispetto ad una qualsiasi di queste altre categorie, e si può perciò essere certi che quel che si sta facendo è un *effettivo* esperimento di attribuzione d'autore.

I risultati sperimentali ottenuti nel 2006 hanno convinto sia noi sia la Fondazione Gramsci dell'efficacia dei metodi basati su distanze, spingendoci ad applicarli a nuovi corpora testuali. Ad oggi sono state attribuite diverse centinaia di articoli anonimi, tutti scritti nel periodo 1916-17; le attribuzioni ottenute per questi testi sono per gran parte in accordo con quelle proposte dagli esperti di filologia gramsciana che stanno lavorando insieme a noi al progetto dell'Edizione Nazionale. Attualmente stiamo utilizzando come insieme di riferimento i cento testi del primo *corpus*, ma il *data set* necessiterà a breve di essere aggiornato, non appena si renderanno disponibili per l'attribuzione testi di anni successivi al 1919, di cui non si hanno esempi in quel primo *corpus*.

È possibile fermarsi qui, all'attribuzione corretta degli scritti di Gramsci? Crediamo di no, perché a questo punto occorre capire perché i metodi adottati per questa ricerca funzionino: «Computational models, however finely perfected, are better understood as temporary states in a process of coming to know rather than fixed structures of knowledge» [30].

Dal punto di vista teorico molte questioni rimangono in effetti aperte ad un'indagine ulteriore, in particolar modo per quanto riguarda gli *n*-grammi. Se infatti esistono risultati consolidati che giustificano l'utilizzo dell'entropia relativa come indicatore della distanza tra due sorgenti e dei programmi di compressione dati per approssimarla, il metodo degli *n*-grammi necessita invece di essere compreso più a fondo. Innanzitutto, suddividere il testo in *n*-grammi e non in parole significa cogliere aspetti quantitativi di molti tipi diversi nel testo stesso: lunghezza delle parole, frequenze dei caratteri, delle parole brevi, dei segni di interpunzione, delle lettere maiuscole, e così via. Se da una parte questa eterogeneità è probabilmente la ragione dell'efficacia del metodo, d'altro canto essa rende più difficile comprendere a fondo quali fattori sono responsabili della maggiore o minore "vicinanza" di due testi secondo la distanza degli *n*-grammi.

È molto interessante poi che il valore di *n* che ha dato i risultati migliori per l'attribuzione sia $n = 8$. Ci si potrebbe infatti aspettare in prima analisi che un valore così grande di *n* sia più indicativo dell'argomento di un testo, piuttosto che del suo autore; d'altra parte questo non può essere vero per i testi del *corpus* gramsciano, che, come già discusso, sono piuttosto omogenei dal punto di vista dell'argomento.

Si osservi in ultimo che per $n = 8$ il metodo degli *n*-grammi non può essere considerato un metodo statistico; per illustrare questo punto si può fare il confronto tra bigrammi ed 8-grammi. I bigrammi presenti nei diversi testi sono praticamente gli stessi, a parte qualche rara eccezione; inoltre ogni bigramma compare più volte nei singoli testi, e dunque la misura della loro frequenza è statisticamente significativa. Al contrario, nel confronto tra due testi qualsiasi (ad esempio, il 26esimo e il 27esimo testo gramsciano dell'insieme di riferimento finora in uso), l'87% degli 8-grammi compare una sola volta e il 95% compare in uno solo dei due testi. In altre parole: i dizionari degli 8-grammi dei due testi hanno solo una piccola porzione in comune, e la frequenza del singolo 8-gramma non è in genere statisticamente significativa. Anche se si considera l'insieme di tutti i testi del *data set*, poi, molti 8-grammi compaiono più volte, ma sempre abbastanza al di sotto della soglia di significatività statistica. È dunque in un certo senso sorprendente che una distanza per cui la statistica è così "povera", come nel caso degli 8-grammi, dia risultati migliori rispetto al caso $n = 2$, in cui la statistica è molto più corposa. Questo fatto, insieme agli altri sopra discussi, è stato e sarà occasione di un'approfondita riflessione metodologica che crediamo debba procedere parallelamente allo sviluppo delle tecniche sperimentali.

Note

[1] In questo testo, frutto del lavoro comune di entrambi gli autori, sono di Maurizio Lana le prime sei pagine, di Chiara Basile le successive.

[2] Di qui in avanti: "scritti gramsciani anonimi".

[3] M. Lana, *L'attribuzione di testi gramsciani e i metodi quantitativi*; D. Benedetto, E. Caglioti, M. Degli Esposti, *L'attribuzione dei testi gramsciani: metodi e modelli matematici*.

[4] Clement, R., Sharp, D., *Ngram and bayesian classification of documents for topic and authorship*, in: "Literary and linguistic computing", 2003, 18(4):423-447; pag. 426

[5] Doležel L., *A note on quantification in text theory*, in: "Text Processing. Text Analysis and Generation. Text Typology and Attribution. Proceedings of Nobel Symposium 51", a cura di S. Allén, Almqvist & Wiksell International, Stockholm, 1982, pp. 539-552; qui pp. 540-42

[6] A titolo di esempio ipotetico, si potrebbe pensare di contare e vedere come si distribuiscono nel testo le sequenze di caratteri che iniziano con una **a** e che hanno una **z** a distanza di 4 caratteri dalla **a**.

[7] Khmelev, D. V., *Disputed authorship resolution through using relative entropy for Markov chains of letters in human language texts*, in: "Journal of quantitative linguistics", 7, 2000, pp. 115-26, all'URL: <www.philol.msu.ru/~lex/khmelev/published/jql/khmelev.html>.

[8] D. Khmelev, Tweedie F., *Using Markov chains for identification of writers*, in: "Literary and linguistic computing", 16, 4, 2001, pp. 299-307

[9] D. Benedetto, E. Caglioti, V. Loreto, *Language Trees and Zipping.* "Physical review letter", 88, n. 4, 048702-1, 048702-1 (2002).

[10] Sul tema dell'uso dei compressori per misurare l'entropia informativa si veda più avanti.

[11] Clement, R., Sharp D., *Ngram and bayesian classification of documents for topic and authorship*, in: "Literary and Linguistic Computing", 18, 4 2003, pp. 423-447.

[12] Si veda per esempio: Juola, P. and Baayen, H., *A controlled corpus experiment in authorship attribution by crossentropy*, in: "Proceedings of ACH/ALLC 2003", Athens, GA.

[13] M. Koppel, J. Schler (2004), *Authorship verification as a one-class classification problem*, in: "Proceedings of 21st International Conference on machine learning, July 2004, Banff, Canada", pp. 490-91.

[14] Lutosławski, Wincenty, *The origin and growth of Plato's logic*, London-New York- Bombay: Longmans, Green & Co. 1897 (Repr., Hildesheim: Georg Olms, 1983).

[15] T.C. Mendenhall, *The characteristic curves of composition*, "Science", 11, Marzo 1887, ns-9: 237-246.

[16] Cfr. Kenny, A. (1982), *The computation of style: an introduction to statistics for students of literature and humanities*. Oxford & New York: Pergamon Press.

[17] G.R. Ledger, *Re-counting Plato: a computer analysis of Plato's tyle*, Oxford, Clarendon Press; New York, Oxford University Press, 1989.

[18] Sugli n-grammi si veda questo stesso contributo più avanti.

[19] A. Clarke, *Hazards of prophecy: the failure of imagination*, in: "Profiles of the future. An inquiry into the limits of the possible", New York, Harper & Row, 1962.

[20] È chiaro che il significato del concetto molto astratto di similarità dipende fortemente dal contesto che si prende in esame. Le distanze (o *metriche*) di similarità sono in effetti utilizzate in ambiti anche molto diversi tra loro per trattare vari problemi di *classificazione* di stringhe simboliche; a seconda dei casi raggruppare stringhe *simili* può dunque significare suddividere per genere un certo numero di brani musicali oppure distinguere i pazienti sani da quelli malati conoscendo delle sequenze simboliche estratte dai loro elettrocardiogrammi; si vedano a questo proposito, ad esempio:

- M. Degli Esposti, C. Farinelli, M. Manca, A. Tolomelli. *A similarity measure for biologic signals: new applications to HRV analysis*. JP J Biostat., vol. 1, n. 1, pp 53-78 (2007).
- M. Degli Esposti, C. Farinelli, G. Menconi. *Sequence distance via parsing complexity: Heartbeat signals*. Chaos, Solitons and Fractals (2007), in corso di stampa.

Nel nostro caso la similarità cercata è quella tra lo stile del testo da attribuire e quello di uno scrittore che potrebbe esserne l'autore.

[21] R.W. Hamming. *Error detecting and error correctiong codes.* Bell Systems Technical Journal 26(2): 147-160, 1950.

[22] È possibile giustificare questa idea con considerazioni legate alla teoria delle sorgenti markoviane; per questa ed altre questioni di carattere più fisico-matematico si vedano il già citato contributo: D. Benedetto, E. Caglioti, M. Degli Esposti. *L'attribuzione dei testi gramsciani: metodi e modelli matematici* e anche C. Basile, D. Benedetto, E. Caglioti, M. Degli Esposti. *An example of mathematical authorship attribution.* In pubblicazione su Journal of Mathematical Physics, .vol. 49, n. 12 (dicembre 2008).

[23] V. Kešelj, F. Peng, N. Cercone, C. Thomas. N-gram based author profiles for authorship attribution. In *Proceedings of the Conference Pacific Association for computational linguistics, PACLING'03*, pp. 255-264, Dalhousie University, Halifax, Nova Scotia, Canada, 2003.

[24] V. Kešelj, N. Cercone. CNG Method with Weighted Voting. In P. Joula*, Ad-hoc authorship attribution competition. In Proceedings 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2004)*, Göteborg, Sweden, 2004.

[25] C.E. Shannon. *A mathematical theory of communication.* The Bell System Technical Journal 27 (1948), 379-423, 623-656.

[26] A. Lempel, J. Ziv. *On the complexity of finite sequences.* IEEE Transactions on Information Theory, IT-22, n.1, 75-81(1976).

- J. Ziv, A. Lempel. *A universal algorithm for sequential data compression*, IEEE Transactions on Information Theory, IT- 23, n. 3, 337-343 (1977).
- J. Ziv, A. Lempel. *Compression of individual sequences via variable-rate coding.* IEEE Transactions on Information Theory, IT-24, n. 5, 530-536 (1978).

[27] J. Ziv, N. Merhav. *A measure of relative entropy between individual sequences with application to universal classification.* "IEEE transactions on information theory", 39, n. 4, 1270-1279 (1993).

[28] D. Benedetto, E. Caglioti, V. Loreto. *cit.*

[29] Per un'analisi dettagliata di cosa succede quando si comprime un file seguito da un altro si veda: A. Puglisi, D. Benedetto, E. Caglioti, V. Loreto, A. Vulpiani. *Data compression and learning in time sequences analysis.* "Physica D", 180*, n. 1-2, 92-107 (2003).*

[30] W. McCarthy. *Modeling: A study in words and meaning*, in: *A companion to digital humanities*, a cura di Susan Schreibman, Ray Siemens, John Unsworth, Blackwell, London, 2005, pag. 257.

Bibliografia

Basile C., Benedetto D., Caglioti E., Degli Esposti M., *An example of mathematical authorship attribution,* in pubblicazione su Journal of Mathematical Physics, vol. 49, n. 12 (dicembre 2008).

Benedetto D., Caglioti E., Degli Esposti M., *L'attribuzione dei testi gramsciani: metodi e modelli matematici* (in corso di stampa nell'Edizione nazionale delle opere di Antonio Gramsci).

Benedetto D., Caglioti E., Loreto V., (2002), *Language Trees and Zipping,* "Physical review letter" 88, n. 4, 048702-1, 048702-1.

Clarke A., (1962), *Hazards of prophecy: The failure of imagination*, in: «Profiles of the future. An inquiry into the limits of the possible», New York, Harper & Row.

Clement R. e Sharp D., (2003), *Ngram and Bayesian classification of documents far topic and authorship*, in: "Literary and linguistic computing".

Degli Esposti M., Farinelli C., Manca M., Tolomelli A., (2007), *A similarity measure for biologic signals: new applications to HRV analysis*, JP J Biostat., vol. 1, n. 1.

Degli Esposti M., Farinelli C., Menconi G., (2007), *Sequence distance via parsing complexity: Heartbeat signals*, "Chaos, solitons and fractals".

Doležel L., (1982), *A note on quantification in text theory*, in: «Text processing. Text analysis and generation. Text typology and attribution», Proceedings of Nobel Symposium 51, a cura di S. Allén, Almqvist & Wiksell International, Stockholm.

Hamming R.W., (1950) *Error detecting and error correctiong codes.* "Bell Systems technical journal", 26(2): 147-160.

Juola P. and Baayen H.,(2003), *A Controlled corpus experiment in authorship attribution by crossentropy*, in: "Proceedings of ACH/ALLC 2003", Athens, GA.

Kenny A. (1982), *The computation of style: an introduction to statistics for students of literature and humanities*. Oxford & New York: Pergamon Press.

Kešelj V.F. Peng N., Cercone C. Thomas, (2003), *N-gram based author profiles for authorship attribution*. In "Proceedings of the Conference Pacific Association for Compu-

tational Linguistics, PACLING '03", Dalhousie University, Halifax, Nova Scotia, Canada.

Kešelj V., Cercone N., (2004), *CNG method with weighted voting*, In P. Joula, *Ad-hoc authorship attribution competition*. In "Proceedings 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2004)", Göteborg, Sweden;

Khmelev D.V., (2000), *Disputed authorship resolution through using relative entropy for Markov chains of letters in human language texts*, in: Journal of quantitative linguistics, 7, accessibile all'URL: <www.philol.msu.ru/~lex/khmelev/published/jql/khmelev.html>.

Khmelev D., Tweedie F., (2001), *Using Markov chains for identification of writers*, in: "Literary and linguistic computing", 16, 4.

Koppel M., Schler J., (2004), *Authorship verification as a one-class classification problem*, in: "Proceedings of 21st International Conference on machine learning, July 2004", Banff, Canada.

Lana M., *L'attribuzione di testi gramsciani e i metodi quantitativi* (in corso di stampa nell'Edizione nazionale delle opere di Antonio Gramsci).

Ledger G.R., (1989), *Re-counting Plato: a computer analysis of Plato's style*, Oxford, Clarendon Press; New York, Oxford University Press.

Lempel A., Ziv J., (1976), *On the complexity of finite sequences.* "IEEE transactions on information theory", IT-22, n.1.

Lutosławski Wincenty, (1983), *The origin and growth of Plato's logic*, London-New York-Bombay: Longmans, Green & Co. 1897 (Repr., Hildesheim: Georg Olms);

McCarthy W., (2005), *Modeling: a study in words and meaning*, in: "A companion to digital humanities", a cura di Susan Schreibman, Ray Siemens, John Unsworth, Blackwell, London, pag. 257.

Mendenhall T. C., (1887),*The characteristic curves of composition*, "Science" 11, Marzo 1887, ns-9: 237-246.

Puglisi A., Benedetto D., Caglioti E., Loreto V., Vulpiani A., (2003), *Data compression and learning in time sequences analysis.* "Physica D", 180, n. 1-2, 92-107.

Shannon C.E., (1948), *A mathematical theory of communication,* "The Bell System technical journal", 27.

Ziv J., Lempel A., (1977), *A universal algorithm for sequential data compression*, "IEEE transactions on information theory," IT- 23, n. 3.

Ziv J., Lempel A., (1978), *Compression of individual sequences via variable-rate coding.* "IEEE Transactions on information theory", IT-24, n. 5.

Ziv J., Merhav N., (1993), *A measure of relative entropy between individual sequences with application to Universal Classification.* "IEEE transactions on information theory", 39, n. 4, 1270-1279.

Dal testo alla conoscenza e ritorno: estrazione terminologica e annotazione semantica di basi documentali di dominio

FELICE DELL'ORLETTA, ALESSANDRO LENCI, SIMONE MARCHI,
SIMONETTA MONTEMAGNI, VITO PIRRELLI, GIULIA VENTURI

The paper focuses on the automatic extraction of domain knowledge from Italian legal texts and presents a fully-implemented ontology learning system (T2K, Text-2-Knowledge) that includes a battery of tools for Natural Language Processing, statistical text analysis and machine learning. Evaluated results show the considerable potential of systems like T2K, exploiting an incremental interleaving of NLP and machine learning techniques for accurate large-scale semi-automatic extraction and structuring of domain-specific knowledge.

Keywords: Natural Language Processing – Machine Learning – knowledge extraction from texts – ontology learning – legal ontologies

1. Introduzione

La necessità quotidiana di accedere a grandi quantità di conoscenza digitale disponibile in forma *non strutturata*, cioè convogliata attraverso testo libero all'interno di basi documentali anche molto vaste e variegate per stile e argomento, ha dato grande impulso allo sviluppo di tecnologie per l'acquisizione, la classificazione e la gestione automatica dell'informazione testuale e al loro sempre più diffuso impiego in una miriade di contesti applicativi (si vedano a questo proposito, tra gli altri, Brin, 1998; Kogut e Holmes, 2001; Vargas-Vega *et alii*, 2002; Dingli, Ciravegna e Wilks, 2003; Popov *et alii* 2003; Dill *et alii*, 2003; Giovannetti *et alii*, 2007). Nonostante gli evidenti successi conseguiti, tuttavia, le tecnologie esistenti si scontrano ancora oggi con un limite fondamentale: l'insufficiente attenzione prestata all'analisi linguistica dei testi. Un reale salto tecnologico verso l'accesso avanzato all'informazione testuale richiede di superare i limiti rappresentati da una capacità solo rudimentale di accedere al contenuto semantico codificato dalla struttura linguistica di un testo. Affinare questa capacità significa dotare i sistemi per l'estrazione di informazione da testi di un'adeguata *intelligenza linguistica*. Questa può variare dalla semplice integrazione di conoscenze lessicali e terminologiche all'annotazione di livelli più avanzati di informazione sintattico-semantica, essenziali per aumentare la quantità e la qualità dell'informazione recuperata e per filtrare il "rumore di fondo" dell'informazione irrilevante.

A nostro avviso, l'esigenza di analizzare grandi repertori di documenti in cui la lingua si manifesta in tutta la sua complessità e variabilità d'uso è destinata ad avere un notevole impatto non solo sulle fasi di progettazione e sviluppo delle tecnologie per il Trattamento Automatico del Linguaggio (TAL), ma anche sul nostro stesso modo di studiare il linguaggio umano e di concettualizzare i rapporti tra linguaggio, cognizione e contesti comunicativi concreti. Da una parte, si va sempre più avvertendo la necessità di affiancare ai tradizionali componenti di analisi linguistica (grammatiche formali sviluppate su base introspettiva) nuove tipologie di strumenti per l'acquisizione dinamica di conoscenza, basati sull'impiego di algoritmi di *machine learning*, in grado di adattarsi con rapidità ed efficienza a diversi domini applicativi e terminologici e alla variabilità linguistica offerta da tipologie testuali anche radicalmente differenti. D'altro canto, gli attuali sistemi di rappresentazione formale della conoscenza di dominio (ontologie) offrono l'inedita opportunità di sviluppare procedure di integrazione dinamica tra livelli di conoscenza linguistica ed extra-linguistica, consentendo di superare gli attuali limiti delle rispettive tecnologie. Ad esempio, l'estrazione di informazioni di dominio da testi liberi può utilmente integrare le attività di sviluppo e popolamento manuale di un modello ontologico, per loro natura lente, ripetitive e soggette a errori. Per converso, la natura fortemente implicita, personalizzata e fraseologizzata dell'informazione estraibile da testi può beneficiare in larga misura dell'insieme di concetti, proprietà e relazioni offerte da una rappresentazione formale e strutturata di uno specifico dominio di conoscenza.

Text-to-Knowledge (T2K), una piattaforma *software* sviluppata congiuntamente dall'Istituto di Linguistica Computazionale (CNR) e dal Dipartimento di Linguistica dell'Università di Pisa finalizzata all'acquisizione di tipi diversi di informazione semantico-lessicale da documenti testuali, ci consente di illustrare concretamente la portata di questi cambiamenti metodologici e di delinearne le prospettive future. Attraverso l'uso combinato di tecniche statistiche e di strumenti avanzati per il TAL, T2K è in grado di analizzare il contenuto linguistico dei documenti, individuare i termini potenzialmente più significativi, ricostruire una "mappa" multidimensionale dei concetti espressi da questi termini, sviluppare un'ontologia del dominio di interesse.

2. Il paradosso dell'acquisizione

La costruzione semi-automatica di ontologie di dominio, intese come *repertori strutturati di concetti rilevanti per la descrizione e organizzazione di un certo dominio di conoscenza* (Gruber, 1995), è un ambito di ricerca particolarmente attivo in diversi settori specialistici quali la bio-informatica, il campo della pubblica amministrazione, quello della gestione documentale aziendale e giuridico-legislativa. In questo senso sono state

proposte diverse metodologie finalizzate all'estrazione automatica di informazione da basi testuali e alla loro strutturazione in ontologie di dominio (per una rassegna, cfr. Buitelaar et alii, 2005). Gli sforzi in tale direzione, tuttavia, sono tipicamente inficiati da un classico paradosso. Stabilire una corrispondenza adeguata tra la rappresentazione linguistica del dominio di conoscenza fornita da un insieme di documenti rilevanti e la rappresentazione ontologica sottostante dello stesso dominio *presuppone* la disponibilità di una notevole quantità di conoscenza rilevante rispetto all'ambito trattato. Ad esempio, acquisire conoscenza relativa agli obblighi a cui è soggetta una particolare entità giuridica (ad es. il "datore di lavoro") richiede la capacità preliminare di identificare il "datore di lavoro" come un tipo di entità giuridica, nonché la capacità di localizzare, nel contesto di un testo legislativo, gli obblighi a cui tale entità è soggetta (ad es. garantire la sicurezza personale del lavoratore). Estrarre informazione da un testo richiede dunque altra informazione. Più tecnicamente, "popolare" i nodi di un'ontologia a partire da informazione testuale richiede che la struttura dell'ontologia sia già in piedi.

È nostra convinzione che l'unione sinergica di tecnologie linguistiche e tecniche di *machine learning* possa rappresentare una strategia metodologica vincente per far fronte a questo paradosso. Da questo punto di vista, T2K offre un esempio interessante di strumento *ibrido* per l'analisi automatica del contenuto testuale, al cui interno si integrano aspetti tradizionali di analisi linguistica e funzionalità per l'accesso a livelli sempre più astratti e strutturati di conoscenza. Riteniamo che questa strategia possa portare a coniugare in modo non banale due esigenze complementari del mondo della comunicazione: l'esigenza di una rappresentazione esplicita, normalizzata e condivisa del contenuto, cui fanno fronte i modelli più o meno recenti di rappresentazione formale della conoscenza, e quella derivante dal bisogno di personalizzare questo contenuto, secondo prospettive soggettive condizionate dal contesto e dal punto di vista dell'utente, rispetto alla quale il linguaggio rappresenta uno strumento insostituibile. Ritorneremo su questo aspetto metodologico al paragrafo 4 del presente contributo.

Il funzionamento di T2K sarà illustrato con i risultati di alcuni esperimenti di estrazione e strutturazione di terminologia condotti su due corpora di testi giuridici italiani in materia di legislazione ambientale e di protezione del consumatore. Ad oggi, il dominio giuridico costituisce un'area attiva di studio, particolarmente aperta alla necessità di dotare le tecnologie di gestione dell'informazione di un'adeguata "intelligenza linguistica". Le difficoltà intrinseche al linguaggio naturale, in generale, e al Trattamento Automatico del linguaggio dei testi giuridico-amministrativi, in particolare, hanno infatti fatto sì che le ricerche in materia di costruzione di ontologie giuridiche siano state condotte per lo più *in modo manuale* da esperti del dominio e in una prospettiva *top-down* rivolta soprattutto alla strutturazione della dottrina giuridica (per uno stato dell'arte cfr. Valente, 2005). Riteniamo, pertanto, che la metodologia ibrida seguita da T2K nell'acquisizione di conoscenza di dominio a partire da categorie di analisi interne

ai testi possa costituire un promettente punto di partenza per sviluppare in modo semi-automatico un'ontologia giuridica in una prospettiva *bottom-up*. Finora, pochi tentativi sono infatti stati fatti in questa direzione per indurre in modo automatico ontologie giuridiche a partire da corpora testuali (per uno stato dell'arte cfr. Lenci *et alii*, 2008).

3. Obiettivi

Obiettivo generale di T2K è trasformare le conoscenze implicitamente codificate all'interno di un corpus di testi in conoscenza esplicitamente strutturata. Il risultato finale è un glossario terminologico arricchito con informazione semantico-concettuale. Per arrivare a identificare i concetti rilevanti e più caratterizzanti i documenti di un certo dominio di interesse, T2K si avvale di strumenti di analisi in linea con lo stato dell'arte nella ricerca linguistico-computazionale partendo da un'ipotesi di lavoro molto semplice: i concetti e i temi rilevanti nel testo sono veicolati dai termini statisticamente più significativi. Questi ultimi possono essere unità lessicali monorematiche come *accordo*, *produttore* o *presidente* oppure unità lessicali polirematiche come *procedimento amministrativo*, *Ministro dell'ambiente*, *incenerimento dei rifiuti pericolosi*, *assistenza reciproca*, *contratto di multiproprietà*, ecc. La compilazione di un repertorio di terminologia di dominio sulla base delle concrete attestazioni nei testi costituisce il risultato della prima fase operativa di T2K sulla base del quale è possibile condurre un'indicizzazione terminologica dei documenti.

I termini che formano il glossario terminologico acquisito possono essere a loro volta raggruppati secondo diverse relazioni di similarità. Ad esempio, *tutela ambientale*, *tutela dei consumatori*, *tutela dell'ozono stratosferico*, *tutela del paesaggio* e simili condividono il concetto più generale di TUTELA cui sono tutti ricondotti attraverso la relazione di iponimia (o *ISA*). Oltre questa strutturazione concettuale di tipo gerarchico, T2K è anche in grado di identificare classi di termini semanticamente correlati come ad esempio {*disposizioni, norme, decisione, atto, prescrizioni*}, {*legge, regolamento, protocollo, accordo, statuto*}, {*inquinamento, danno ambientale, effetti nocivi, conseguenza*}. L'organizzazione e la strutturazione dei termini secondo relazioni gerarchiche e di quasi-sinonimia rappresenta il risultato della successiva fase operativa di T2K sulla base della quale è possibile condurre un'indicizzazione concettuale dei testi.

Un sistema di conoscenza non è tuttavia costituito solo da concetti che si riferiscono a entità del dominio, ma anche da processi, azioni ed eventi che vedono coinvolte queste entità secondo ruoli e funzioni diverse. Ad esempio, un *decreto legislativo* così come un *articolo* o un *comma* sono tipicamente *abrogati*, *sostituiti*, *modificati* così come possono essere *emanati*, *recepiti*, *applicati*; la *qualità dell'aria* insieme al *livello di protezione ambientale* e all'*ecosistema* sono *garantiti*, *protetti* e *salvaguardati* così come posso-

no essere *pregiudicati* e *inquinati*. Gli sviluppi più recenti di T2K vanno nella direzione appena delineata, cercando di identificare le relazioni più tipiche che legano le entità e i concetti identificati con il fine ultimo di arrivare a ricostruire dai testi una "mappa" semantica del dominio esplorato.

Nel suo complesso, il risultato finale di T2K si presenta dunque come una rete multidimensionale di unità terminologico-concettuali con significativi punti di contatto sia con l'architettura classica dei thesauri sia con reti semantico-lessicali come WordNet (Fellbaum, 1998). La rete di conoscenza acquisita da T2K è articolata sui seguenti livelli:

- *glossario terminologico*, costituito da una lista di termini mono o polirematici, estratti automaticamente dai testi arricchiti con vari livelli di annotazione linguistica, in particolare morfo-sintattica e sintattica. I termini vengono selezionati con criteri statistici come i più significativi o salienti per la caratterizzazione dei documenti. Il glossario terminologico finale include anche indicazione delle varianti ortografiche, morfologiche e strutturali associate ad ogni unità terminologica acquisita;
- *tassonomia concettuale* – i termini del glossario sono strutturati attraverso relazioni gerarchiche di iponimia/iperonimia ricostruite a partire dalla loro struttura linguistica interna;
- *famiglie di termini concettualmente affini* – i termini del glossario sono organizzati in classi di termini semanticamente simili sulla base della loro distribuzione all'interno di contesti lessico-sintattici;
- *rete semantico-concettuale* – i concetti che si riferiscono ad entità del dominio sono messi in relazione attraverso le azioni e gli eventi che li vedono tipicamente coinvolti secondo ruoli e funzioni diverse. A partire dal testo viene ricostruita una rete semantica costituita dalle relazioni più tipiche che coinvolgono i concetti selezionati.

Ad oggi, il risultato di T2K è circoscritto a quanto descritto ai punti 1-3, mentre la costruzione della rete semantico-concettuale (4) rappresenta un'area attiva di studio e di sperimentazione.

4. Architettura funzionale di T2K

L'architettura e i processi di elaborazione di T2K sono organizzati in due fasi successive, strettamente legate ai livelli di organizzazione dell'ontologia descritti nel paragrafo 3:

Fase 1: creazione del glossario dei termini;

Fase 2: strutturazione concettuale del repertorio terminologico acquisito. Il diagramma in Figura 1 schematizza l'architettura e i processi di elaborazione

di T2K. Nella colonna centrale del diagramma sono riportate le varie fasi del processo estrattivo, dove livelli di analisi linguistica si alternano con fasi di elaborazione statistica del testo linguisticamente annotato. Il risultato di questo processo è costituito dall'ontologia di dominio (rappresentata nel riquadro a destra) articolata su diversi livelli: glossario terminologico, tassonomia concettuale, famiglie di termini semanticamente affini e rete semantico-concettuale. I componenti di analisi linguistica integrati in T2K sono parte di AnIta, una batteria di strumenti ad ampio spettro per il trattamento automatico dell'italiano (si veda il diagramma in Figura 2), disponibile anche in versione web <foxdrake.ilc.cnr.it/webtppls>. Ritorneremo più avanti sul loro ruolo all'interno di T2K.

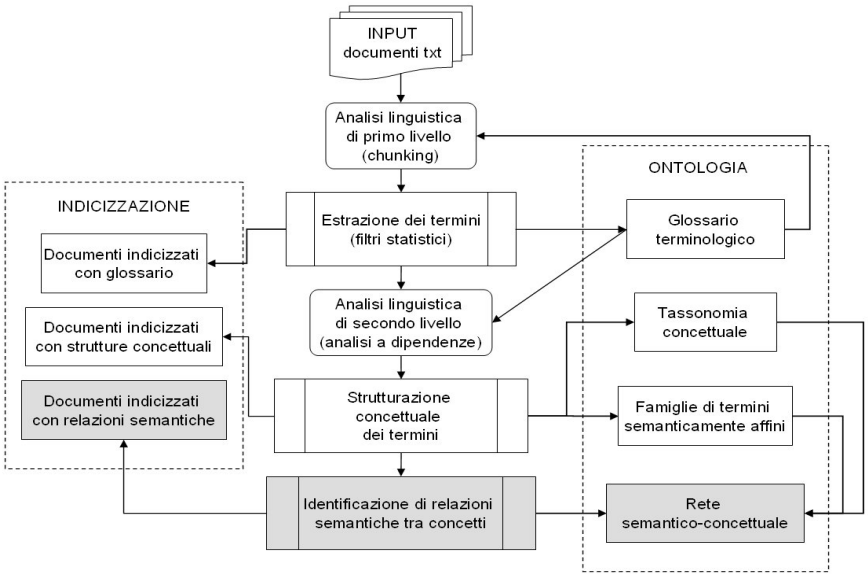


Figura 1 - L'architettura e i processi di elaborazione di T2K

Vale la pena di sottolineare che alcuni livelli di informazione ontologica estratti dal testo vengono reintegrati nel testo di partenza, stratificati come livelli di annotazione o glosse, per arricchirne il contenuto. Il testo così arricchito viene a sua volta usato come input nelle fasi successive del processo di acquisizione di conoscenza. Ad esempio, il glossario terminologico estratto dal testo viene riproiettato sul testo stesso in vista di una seconda fase di analisi linguistica a dipendenze. È a questo livello di analisi che diventa possibile recuperare le relazioni che ciascuno dei termini acquisiti intrattiene

contestualmente con altri termini all'interno della base documentale di partenza. In questo modo, si attiva un ciclo virtuoso di annotazione-acquisizione-annotazione, in cui il testo mantiene il suo ruolo centrale di deposito di informazioni progressivamente strutturate a livelli di crescente astrazione e complessità. Questa metodologia consente a nostro avviso di integrare al meglio i contributi di diverse tecnologie dell'informazione, affrontando in modo dinamico il problema di come acquisire la conoscenza implicitamente contenuta in basi documentali di domini specialistici, dove le strutture linguistiche del testo si intrecciano strettamente con aspetti di conoscenza del mondo reale. Soprattutto in campo giuridico, infatti, applicazioni sviluppate su ampie basi di conoscenza necessitano di ontologie che comprendano conoscenza di dominio continuamente aggiornata. Un processo testo-conoscenza organizzato in modo dinamico può dunque risultare estremamente efficace in questo ambito operativo. Da una parte, la necessità di disporre di rappresentazioni altamente strutturate, compatte e decontestualizzate dell'informazione acquisita è soddisfatta attraverso la costruzione di repertori informativi strutturati quali glossari, ontologie e reti semantiche. Usando queste strutture come chiavi di accesso al testo (in forma di "metadati" testuali) è possibile indicizzarne il contenuto per termini, concetti e relazioni. D'altra parte, la stratificazione di questi livelli di informazione sul testo è funzionale alla creazione di ulteriori livelli di analisi, attraverso un processo incrementale che da una parte distilla nuova conoscenza e dall'altra la mette a disposizione per successive elaborazioni. Torneremo su questi aspetti metodologici nelle conclusioni dell'articolo.

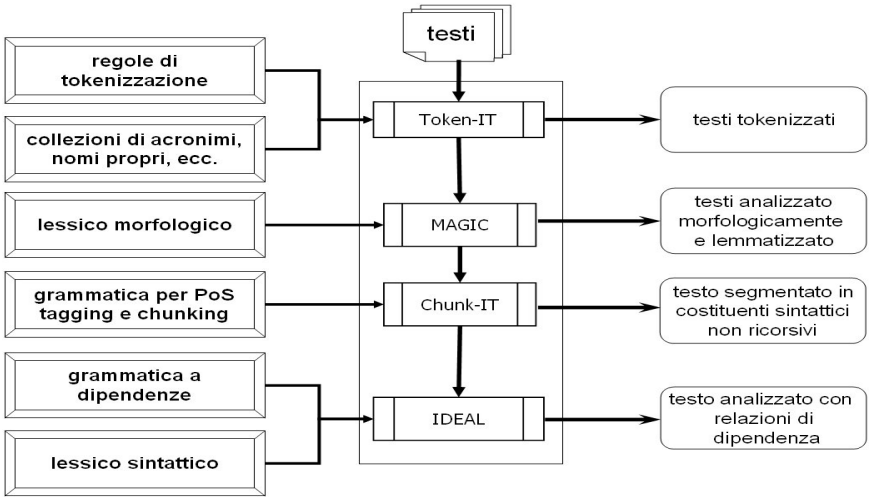


Figura 2 - Architettura di AnIta

5. Creazione del glossario terminologico

In T2K il processo di estrazione terminologica riguarda unità terminologiche monorematiche e polirematiche. Nel caso delle prime, il processo di acquisizione opera sul testo lemmatizzato ed etichettato a livello morfo-sintattico (Battista e Pirrelli, 2000) e avviene sulla base della frequenza dei lemmi nei documenti. Ciò equivale a dire che la frequenza di ogni termine del glossario è la somma delle frequenze delle diverse forme flesse riconducibili allo stesso lemma. Una volta definita la lista dei potenziali termini semplici, i termini candidati sono filtrati sulla base della loro frequenza all'interno del corpus di acquisizione.

Diverso è il caso delle unità terminologiche polirematiche la cui acquisizione si articola in più fasi. Innanzitutto, i potenziali termini complessi sono identificati all'interno di testi segmentati sintatticamente in costituenti sintattici elementari non ricorsivi detti "chunks" (Abney, 1991; Federici et alii, 1996; Lenci et alii, 2003). Un chunk è identificato come una sequenza continua di parole del testo che va dalla prima unità grammaticale (tipicamente, una preposizione, un ausiliare, un (pre)determinatore o un ausiliare) incontrata nel testo (detta "iniziatore di chunk") alla prima unità lessicale piena selezionata dall'iniziatore di chunk. Esempi di chunk nominale (N_C) sono *la mia prima casa* (con *la* iniziatore di chunk e *casa* testa lessicale piena selezionata) e *questo difficile problema* (con *questo* iniziatore di chunk e *problema* testa lessicale piena selezionata). *Per il medico* e *da sotto il tavolo* sono tipici chunk preposizionali (P_C), con *per* e *da* iniziatori di chunk e *medico* e *tavolo* teste lessicali. Infine nei chunks verbali di modo finito (FV_C) *è stato evidenziato* e *ho ripetutamente osservato* gli ausiliari *è* e *ho* iniziano il chunk e i participi passati *evidenziato* e *osservato* rappresentano le teste lessicali selezionate dagli ausiliari.

Il testo così segmentato viene analizzato da una mini-grammatica deputata al riconoscimento delle strutture linguistiche che formano potenziali termini complessi: ad esempio, una sequenza di chunk nominale (N_C) seguito da un chunk aggettivale (ADJ_C) (es. *organizzazione internazionale*), oppure una sequenza di chunk nominale seguito da un chunk preposizionale (P_C) (es. *presidente della repubblica*). L'assunto di base è che se due o più parole formano un termine complesso in un certo dominio, è molto probabile che nell'uso linguistico relativo a quel dominio esse tendano a ricorrere insieme in maniera statisticamente significativa. La soglia di significatività statistica è determinata a partire da una stima della probabilità che le parole in questione *possano ricorrere insieme per caso*, probabilità esprimibile in funzione della frequenza delle parole stesse all'interno di un corpus rappresentativo. Se una coppia di parole ricorre nel testo più spesso di quanto saremmo autorizzati ad aspettarci in base alle leggi del caso, allora il test di significatività è soddisfatto.

Per ciascuno dei potenziali termini complessi identificati dalla mini-grammatica viene dunque quantificata su base statistica la forza di associazione tra le parole che lo compongono. In T2K, questa forza è stimata applicando la misura associativa detta "log-likelihood" (Dunning, 1993). Una serie di esperimenti preliminari ha infatti evidenziato che tale misura produce risultati sensibilmente migliori rispetto ad altre misure statistiche, quali "pointwise mutual information", "t-score", ecc., in quanto appare più robusta nel caso di dati linguistici con bassa frequenza di occorrenza. Infatti, a differenza di altre misure associative quali la "mutual information", la log-likelihood non privilegia l'associazione di parole rare (Manning e Schütze, 1999). Non escludiamo tuttavia che risultati diversi potrebbero essere ottenuti lavorando su corpora di grandi dimensioni, dove il problema della "sparsità" dei dati risulta notevolmente ridimensionato.

La lista dei potenziali termini complessi viene poi ordinata sulla base del grado di significatività della loro associazione all'interno del dominio dei documenti. Vale la pena qui ricordare che in T2K la misura di associazione si applica alle teste lessicali piene dei costituenti sintattici che lo compongono. Ad esempio, nel caso di un'unità terminologica poliematica come *adempimento dell'obbligo* viene misurata la forza di associazione tra le teste lessicali dei due chunks *adempimento* (N_C) e *dell'obbligo* (P_C): rispettivamente *adempimento* e *obbligo*. Il vantaggio derivante da questo approccio è duplice: da un lato permette di condurre il processo di estrazione terminologica facendo astrazione da variazioni di natura ortografica, morfologica così come strutturale, dall'altro rende possibile l'acquisizione delle varianti terminologiche associate a ciascun termine acquisito (vedi infra).

La lista dei termini candidati così ottenuta può essere ulteriormente estesa con termini complessi di "ordine" superiore al primo. Infatti, a questo livello i termini acquisiti sono tipicamente caratterizzati da una struttura sintattica binaria: ad es. N_C-ADJ_C, N_C-P_C, N_C-N_C, ecc. Per arrivare ad acquisire termini più complessi, la procedura di estrazione di termini appena descritta viene applicata iterativamente riproiettando sul testo segmentato in chunks i termini composti identificati durante la prima fase di estrazione. Ad esempio, se al primo passo è stato estratto il termine composto *comunità economica*, al secondo passo è possibile estrarre un nuovo termine composto, *comunità economica europea*, che include il termine individuato al passo precedente. Riportiamo di seguito alcuni esempi di unità terminologiche poliematiche ottenute mediante l'applicazione di questo processo di acquisizione incrementale. In grassetto sono evidenziati i termini complessi acquisiti allo stadio di analisi precedente:

- ***acquisizione di un diritto** di **godimento a tempo parziale***
- *violazione delle **disposizioni nazionali***
- *direttiva del **parlamento europeo***
- ***dispositivo automatico** di chiamata.*

Abbiamo verificato che questo approccio incrementale all'estrazione terminologica presenta numerosi vantaggi rispetto a un'acquisizione condotta in un passo unico: innanzitutto la tipologia di strutture sintattiche da tenere in considerazione viene ridotta a un numero contenuto di strutture di base. L'acquisizione di termini più complessi è vincolata al fatto che almeno uno dei componenti del termine più complesso sia già stato selezionato come tale al passo precedente di acquisizione, con il risultato di ridurre il potenziale rumore nei risultati. In sintesi, questo processo incrementale di acquisizione terminologica, che rappresenta uno dei tratti caratterizzanti di T2K, fornisce maggiori garanzie sull'effettiva rilevanza dei termini acquisiti (Bartolini et alii, 2005).

Il processo di acquisizione terminologica produce due insiemi di termini candidati, ovvero una lista di potenziali unità terminologiche monorematiche ordinate per frequenze decrescenti, e una lista di potenziali unità terminologiche polirematiche ordinate per forza di associazione decrescente. Il glossario terminologico finale è costruito stabilendo diverse soglie riguardanti a) la frequenza di occorrenza minima dell'unità terminologica nella collezione documentale, e b) la percentuale di termini selezionati nelle liste ordinate di termini potenziali. Tali soglie possono essere definite in modo interattivo dall'utente sulla base delle caratteristiche della collezione documentale (tipicamente la sua estensione) e del tipo di risultato atteso (in termini di accuratezza e copertura). I termini così selezionati vengono a comporre il glossario di base, di cui riportiamo un estratto in Tabella 1 dove la colonna "termine" riporta la forma del termine che è stata selezionata come prototipica all'interno del corpus di acquisizione, la colonna "valore" indica la frequenza di ciascun termine nell'intera collezione di documenti analizzati e la colonna "lemma" riporta i lemmi delle teste lessicali dei chunks corrispondenti al termine in questione. Infine, la colonna "stop" è usata per segnalare i termini potenzialmente spuri che sono contrassegnati da un valore diverso da NULL (per maggiori dettagli sul ruolo di questo campo v. infra). Alcune precisazione sono necessarie per una corretta interpretazione delle colonne "termine" e "lemma" della Tabella 1. Per quanto riguarda la prima, abbiamo deciso di rappresentare il termine acquisito attraverso la sua forma prototipica attestata all'interno del corpus piuttosto che mediante un esponente lessicale o lemma selezionato con criteri astratti (ad es. la forma maschile singolare); questa scelta deriva dal fatto che spesso un termine di dominio è legato a una forma specifica, ad esempio quella plurale come nel caso del termine *operazioni di smaltimento* riportato nel frammento di glossario in Tabella 1. La colonna "lemma" riporta invece una codifica astratta del termine formulata come sequenza delle teste lessicali dei chunks che lo costituiscono (es. *ordinanza presidente* come codifica astratta del termine *ordinanza del presidente*); questa codifica è condivisa da tutte le varianti morfologiche e strutturali del termine in questione.

KWID	Termine	Valore	Lemma	Stop
544	OPERATORE ECONOMICO	18	OPERATORE ECONOMICO	NULL
920	OPERAZIONE DI RECUPERO	8	OPERAZIONE RECUPERO	NULL
888	OPERAZIONE DI RICICLAGGIO	8	OPERAZIONE RICICLAGGIO	NULL
201	OPERAZIONI	109	OPERAZIONE	NULL
1098	OPERAZIONI DI SMALTIMENTO	5	OPERAZIONE SMALTIMENTO	NULL
1443	ORDINAMENTI DEGLI STATI MEMBRI	5	ORDINAMENTO STATO MEMBRO	NULL
240	ORDINAMENTO	91	ORDINAMENTO	NULL
360	ORDINAMENTO GIURIDICO	39	ORDINAMENTO GIURIDICO	NULL
1252	ORDINAMENTO GIURIDICO INTERNO	11	ORDINAMENTO GIURIDICO INTERNO	NULL
609	ORDINAMENTO NAZIONALE	15	ORDINAMENTO NAZIONALE	NULL
1043	ORDINAMENTO OLANDESE	6	ORDINAMENTO OLANDESE	NELL'
288	ORDINANZA	69	ORDINANZA	NULL
730	ORDINANZA DEL PRESIDENTE	11	ORDINANZA PRESIDENTE	NULL
921	ORDINANZA DI RINVIO	8	ORDINANZA RINVIO	NULL
825	ORGANI AMMINISTRATIVI	10	ORGANO AMMINISTRATIVO	NULL
398	ORGANI GIURISDIZIONALI	32	ORGANO GIURISDIZIONALE	NULL

Tabella 1 - Un estratto del glossario di base automaticamente acquisito da T2K

Un'ulteriore precisazione è necessaria in relazione al campo "stop" che, quando diverso da NULL, contiene la preposizione che introduce il termine in questione in tutte le sue attestazioni nel corpus di apprendimento. Questa informazione è stata introdotta per poter identificare semi-automaticamente la presenza di termini spuri, corrispondenti a locuzioni preposizionali (del tipo *ai sensi di*, *in materia di*) o locuzioni avverbiali (come *in linea di massima* o *in tempo utile*). Si è deciso di segnalare piuttosto che di eliminare i termini potenzialmente spuri in quanto essi possono corrispondere a termini caratterizzati da una bassa frequenza, come nel caso del termine *ordinamento olandese* in Tabella 1. La Tabella 2 riporta un insieme di termini spuri correttamente segnalati come tali:

KWID	Termine	Valore	Lemma	Stop
12	SENSI	691	SENSO	AI
42	SENSI DELL' ART.	332	SENSO ART.	AI
380	SENSI DELL' ARTICOLO	35	SENSO ARTICOLO	AI
377	SENSI DELLA DIRETTIVA	37	SENSO DIRETTIVA	AI
1366	SENSI DELLE DISPOSIZIONI DEL CAPITOLO	6	SENSO DISPOSIZIONE CAPITOLO	AI

Tabella 2 - Un estratto del glossario di T2K con termini marcati come potenzialmente spuri

Nell'estratto riportato, la colonna "stop" registra la preposizione *ai* che sistematicamente introduce le varie attestazioni di questa famiglia di termini spuri. Sulla base di questa informazione è possibile rimuovere dal glossario tali locuzioni preposizionali in modo automatico o – preferibilmente – attraverso una selezione e valutazione manuale. Il glossario terminologico finale è affiancato da una tabella che registra le varianti terminologiche attestate nel corpus di acquisizione. L'acquisizione di varianti terminologiche rappresenta una questione centrale nel processo di costruzione di risorse terminologiche (Nenadic et alii, 2004). Secondo Jacquemin (2001), in media un terzo delle occorrenze di un termine sono varianti. Ne consegue che un compito di riconoscimento automatico di terminologia debba tenere in considerazione non solo la forma attestata più frequente, ma la debba mettere in relazione alle varianti terminologiche corrispondenti. Ciò se da un punto di vista teorico permette di far luce sulla natura stessa di un termine (ad esempio scoprendo se si tratta di un'unità polirematica "monolitica" o flessibile), da un punto di vista applicativo consente di migliorare i risultati di compiti di indicizzazione e recupero di documenti. La tipologia di varianti acquisite da T2K nell'ambito del processo di estrazione terminologica include:

- varianti ortografiche: *tasso [**d'**interesse* vs ***di** interesse]*
- varianti morfologiche:
  - singolare vs plurale: *bene* vs *beni*
  - preposizione semplice vs preposizione articolata: *esercizio [**del** diritto* vs ***dei** diritti* vs ***di** un diritto]*
- varianti strutturali: *fornitore [**per** il servizio finanziario* vs ***di** servizi finanziari]*
- varianti con modificatori: *ostacoli [al funzionamento* vs *al **buon** funzionamento]*
- varianti che combinano diversi tipi di variazione: *contratt**o** [**di** fornitura* vs *contratt**i** **di** fornitura* vs *contratt**i** **per** la fornitura]*

Per ogni unità terminologica, mono- o polirematica, T2K estrae le varianti che coprano almeno il 5% delle occorrenze del termine nel corpus di partenza. Il risultato di questo processo è esemplificato in Tabella 3 dove ogni riga rappresenta una variante: la colonna "var_ID" contiene un identificativo della variante, la colonna "Termine" specifica il termine a cui si riferisce la variante, la colonna "Forma_variante" contiene la variante stessa, e la colonna "Frequenza" riporta la frequenza di occorrenza della variante descritta nella base documentale.

Alla tipologia di varianti discussa sopra va aggiunta un'altra classe di varianti terminologiche che sono acquisite da T2K nel corso della fase successiva di strutturazione concettuale dei termini (descritta nel paragrafo 6): si tratta di varianti lessicali che possono includere, oltre a sinonimi, acronimi e abbreviazioni, questi ultimi forme di varianti molto comuni in linguaggi settoriali, come ad esempio *PBC* abbreviazione di *policlorobifenile* o *l.r.* acronimo di *legge regionale*.

var_ID	Termine	Forma_variante	Frequenza
680	ACQUISTO DEL BENE IMMOBILE	ACQUISITO DELL'IMMOBILE	5
681	ACQUISTO DEL BENE IMMOBILE	ACQUISTO DEL BENE IMMOBILE	12
682	ACQUISTO DEL BENE IMMOBILE	ACQUISITO DI UN BENE IMMOBILE	10
1050	ADEMPIMENTO DELL'OBBLIGO	ADEMPIMENTO DI OBBLIGHI	2
1051	ADEMPIMENTO DELL'OBBLIGO	ADEMPIMENTO DEL DETTO OBBLIGO	2
1052	ADEMPIMENTO DELL'OBBLIGO	ADEMPIMENTO DEGLI OBBLIGHI	6
1053	ADEMPIMENTO DELL'OBBLIGO	ADEMPIMENTO DELL'OBBLIGO	6
1631	ADOZIONE DI MISURE	ADOZIONE DELLA MISURA	2
1632	ADOZIONE DI MISURE	ADOZIONE DI MISURE	6
1629	ADOZIONE DI MISURE	ADOZIONE DI TUTTE LE MISURE	2
1630	ADOZIONE DI MISURE	ADOZIONE DELLE MISURE	3
1633	ADOZIONE DI MISURE	ADOZIONE DI UNA MISURA	3
937	AGENTE	AGENTE	87
938	AGENTE	AGENTI	67
378	CONTRATTI DI FORNITURA	CONTRATTO DI FORNITURA	3
379	CONTRATTI DI FORNITURA	CONTRATTI PER LA FORNITURA	4
380	CONTRATTI DI FORNITURA	CONTRATTI DI FORNITURA	17
680	ACQUISTO DEL BENE IMMOBILE	ACQUISITO DELL'IMMOBILE	5
681	ACQUISTO DEL BENE IMMOBILE	ACQUISTO DEL BENE IMMOBILE	12
682	ACQUISTO DEL BENE IMMOBILE	ACQUISITO DI UN BENE IMMOBILE	10

Tabella 3 - Varianti terminologiche acquisite automaticamente da T2K

6. Strutturazione concettuale dei termini

Dopo la fase di estrazione terminologica, T2K procede alla fase di organizzazione concettuale. Essa consiste nell'identificazione a) di relazioni di iponimia e iperonimia e b) di relazioni di affinità semantica tra i termini del glossario. A partire dai frammenti di strutture concettuali identificate, è possibile procedere alla loro validazione e riorganizzazione in base alle relazioni individuate.

Ad oggi, l'identificazione delle relazioni tassonomiche tra termini si basa su una relazione di inclusione lessicale. Due unità terminologiche polirematiche che condividano la medesima testa lessicale (e talora anche gli stessi modificatori) al livello della loro rappresentazione in chunks vengono interpretati come iponimi del termine corrispondente alla struttura condivisa. Da queste singole relazioni iponimiche è possibile ricostruire catene tassonomiche articolate su diversi livelli di profondità come esemplificato di seguito:

PROTEZIONE
    PROTEZIONE DEI CONSUMATORI
    PROTEZIONE DEGLI INTERESSI
        PROTEZIONE DEGLI INTERESSI ECONOMICI
    PROTEZIONE DEI MINORI

È in corso di valutazione il potenziamento di T2K con un componente per l'estrazione di relazioni tassonomiche tra termini che non condividano la testa lessicale (ad esempio, relazioni iperonimiche come quella che lega *ministro dell'ambiente* a *autorità*) secondo tecniche già ampiamente sperimentate per l'acquisizione di informazione tassonomica da dizionari (cfr. Calzolari, 1984; Montemagni, 1996).

L'identificazione di relazioni di affinità semantica tra i termini acquisiti segue un approccio completamente diverso. Si parte dall'assunto di base che la semantica di una parola si correla alle sue proprietà distribuzionali nel testo, ovvero due parole sono semanticamente simili se sono reciprocamente sostituibili in un numero significativo di contesti sintattici. Questo assunto è condiviso da un fertile filone di studi della letteratura linguistico-computazionale che a partire dalla lezione di Firth (1957) e di Harris (1968) cerca di indurre automaticamente aspetti del contenuto semantico delle parole sulla base della loro distribuzione nel testo (si veda, tra gli altri, Pereira et alii, 1993; Grefenstette, 1994; Lin, 1998; Rooth et alii, 1999; per una rassegna dei metodi e delle tecniche di acquisizione automatica di rappresentazioni semantico-lessicali a partire da testi cfr. Lenci et alii, 2005).

In T2K, l'acquisizione di famiglie di termini semanticamente affini è condotta sulla base di misure di similarità semantica basate su proprietà distribuzionali come illustrato in Allegrini et alii (2000a e 2000b). Secondo questo approccio, due termini sono correlati semanticamente se sono reciprocamente sostituibili all'interno di un numero significativo di relazioni di dipendenza sintattica tra teste lessicali. Ad esempio, l'evidenza fornita da contesti come *abrogare un decreto* e *abrogare una direttiva* così come di *integrare un decreto* e *integrare una direttiva* suggerisce che *decreto* e *direttiva* rappresentano termini semanticamente simili perché si correlano con la stessa funzione sintattica a due verbi, *abrogare* e *integrare*. Ovviamente, non tutti i verbi che co-occorrono con i

termini sono ugualmente significativi; si pensi a contesti del tipo *scrivere un decreto* dove il termine *decreto* si accompagna con un verbo che è poco informativo riguardo al suo significato in quanto la classe dei possibili oggetti di *scrivere* è alquanto vasta. Ne consegue che la similarità tra termini che ricorrono con verbi più selettivi (nei riguardi dei loro complementi) deve essere intuitivamente più alta di quella tra termini che dipendono da verbi meno selettivi. *Decreto* e *direttiva* sono dunque molto più simili tra loro (per il fatto di essere entrambi *abrogati* o *integrati*) di quanto non lo siano – mettiamo – *decreto* e *libro* a causa del fatto che entrambi possono essere *scritti*.

Il tipo di similarità semantica catturata attraverso il meccanismo inferenziale delineato sopra è basato sul reale contesto di uso delle parole. Come illustrato in Allegrini et alii (2002, 2003), sono possibili diverse misure di similarità semantica: in particolare, viene fatta distinzione tra una misura relativizzata di similarità semantica e una assoluta (dove la scelta tra le due dipende essenzialmente dall'informazione disponibile in partenza e dal tipo di risultato atteso). Mentre nel primo caso la similarità semantica di due parole w_1 e w_2 viene valutata in relazione a specifici contesti di uso (che costituiscono la prospettiva rispetto alla quale viene formulato il giudizio di prossimità), nel secondo caso la similarità semantica di w_1 e w_2 viene valutata in termini assoluti, ovvero senza alcuna indicazione riguardante i contesti d'uso rispetto ai quali deve essere valutata la similarità delle due parole.

L'induzione di classi di termini semanticamente affini in T2K avviene sulla base di una misura di similarità semantica "ibrida" che combina le due nozioni di similarità semantica relativizzata e assoluta appena enunciate. Il corpus viene prima analizzato da un componente per l'analisi delle relazioni di dipendenza funzionale (IDEAL, Bartolini et alii, 2002, 2004). All'interno del risultato di questo processo di analisi sintattica a dipendenze, vengono rintracciate le dipendenze grammaticali principali (soggetti, oggetti) che interessano i termini del glossario all'interno del corpus di acquisizione, ad esempio, ogg_d(*garantire, protezione dell'ambiente*) oppure sogg(*garantire, ministero*). L'insieme delle coppie verbo-termine funzionalmente annotate rappresenta la base di conoscenza distribuzionale sulla base della quale vengono identificati i raggruppamenti di termini semanticamente affini all'interno del dominio analizzato. Per evitare il rumore introdotto da contesti verbali poco selettivi (es. *essere*, *prendere*, *fare*, ecc.) e al contempo per focalizzarsi sugli eventi più salienti a cui ciascun termine si correla all'interno del dominio esaminato, per ciascun termine del glossario è stato selezionato un sottoinsieme della base di conoscenza generale ritagliato a partire dall'identificazione dei verbi più rilevanti rispetto al dominio ed escludendo a priori verbi semanticamente "vuoti" o semplicemente più "leggeri" quali gli ausiliari o i verbi supporto. I verbi più salienti in relazione a ciascun termine sono stati identificati ricorrendo nuovamente alla misura associativa della log-likelihood (vedi sopra); ad esempio, nella base di conoscenza distribuzionale acquisita, tra i verbi più salienti associati al termine *decreto* troviamo

abrogare, *istituire*, *applicare*, *regolare*, *stabilire* e simili. In questo modo, per ciascun termine del glossario è stata estratto lo spazio delle sue distribuzioni sul piano sintagmatico, a partire dalla quale vengono ricostruite le sue relazioni di similarità a livello semantico-paradigmatico. Il risultato finale di questa fase di elaborazione si presenta in forma tabellare, come esemplificato in Tabella 4 dove per ciascun termine del glossario (prima colonna) viene riportato l'insieme dei termini identificati come semanticamente affini (seconda colonna), ordinati secondo valori decrescenti di similarità (vale a dire che, ad esempio, il termine *norma* è più strettamente correlato a *regolamento* di quanto non lo sia *art.*).

KWID	Termine	Termine correlato
40	REGOLAMENTO	NORMA
41	REGOLAMENTO	MODIFICA
42	REGOLAMENTO	PARAGRAFO
43	REGOLAMENTO	DIRETTIVA
44	REGOLAMENTO	ART.
676	REGOLE	NORMA
677	REGOLE	PRINCIPIO
678	REGOLE	REQUISITI
679	REGOLE	DIVIETO
680	REGOLE	PROCEDURA COMUNITARIA

Tabella 4: Ragruppamenti di termini semanticamente affini acquisiti da T2K

7. Un esperimento di acquisizione terminologica a partire da corpora di testi giuridici

In questa sezione riportiamo i risultati di un esperimento condotto su due corpora di testi giuridici diversificati rispetto al dominio legislato (legislazione ambientale e tutela del consumatore) e all'ente emittente (Unione Europea, Stato italiano, Regione Piemonte).

7.1. I Corpora

Abbiamo applicato T2K a due *corpora*, un *Corpus* Ambientale (AMB) e un *Corpus* sulla Tutela del Consumatore (CONS). In dettaglio, AMB contiene 824 testi normati-

vi (es. legge, decreto, direttiva, ecc…) e amministrativi (es. ordinanza, deliberazione, circolare, ecc…), emanati dall'Unione Europea, dallo Stato italiano e dalla Regione Piemonte, per un totale di 1.399.617 parole, reperiti dal Bollettino Giuridico Ambientale edito dall'Assessorato all'ambiente della regione Piemonte. CONS è invece costituito dalla versione italiana di 16 direttive comunitarie e 42 sentenze, per un totale di 292.609 parole.

I due *corpora* sono stati pre-analizzati con il duplice obiettivo a) di estendere il lessico morfologico ad ampia copertura (*general purpose*) sottostante ad AnIta, aggiornandolo con i termini dei domini d'interesse, e b) di adattare la mini-grammatica per il riconoscimento di unità polirematiche alle specificità del linguaggio dei testi giuridico-amministrativi.

7.2. Valutazione del glossario terminologico acquisito

Dato l'ambito settoriale dei due *corpora* di partenza, in entrambi gli esperimenti di acquisizione il glossario terminologico estratto comprende termini che appartengono sia al dominio giuridico sia al dominio legislato. Dal momento che le due tipologie di termini hanno frequenze di distribuzione piuttosto diverse nelle rispettive basi documentali di partenza, abbiamo deciso di condurre la valutazione dei glossari acquisiti da T2K tenendo in considerazione il fatto che a) a basse soglie percentuali di acquisizione corrispondeva un incremento di precisione nell'estrazione di terminologia appartenente al dominio legislato e b) viceversa, imponendo soglie di selezione più alte, la precisione generale del glossario terminologico acquisito aumentava a discapito della terminologia ambientale e in materia di tutela del consumatore.

La valutazione è stata condotta su un glossario di 4.685 unità terminologiche mono- e polirematiche estratte da AMB e su un glossario di 1.443 termini estratti da CONS confrontando i risultati ottenuti con risorse terminologiche di riferimento sia di tipo giuridico sia specifiche dei domini legislati. In considerazione della natura eterogenea dei glossari acquisiti, le risorse di riferimento selezionate come "gold standard" (GS) di riferimento sono state: sul versante giuridico, il *Dizionario Giuridico* (Edizioni Simone), *JurWordNet* (JWN) e la lista di parole chiave usate per la ricerca in rete dell'*Archivio DoGi* (Dottrina Giuridica); sul versante ambientale, il *Glossario dell'Osservatorio Nazionale sui Rifiuti* (rilasciato dal Ministero dell'Ambiente) e il *Thesaurus EARTh* (Environmental Applications Reference Thesaurus). La valutazione è stata condotta in termini di precisione, calcolata come la percentuale di termini acquisiti in modo corretto da T2K rispetto a tutti i termini estratti. A causa della diversa copertura delle risorse di riferimento selezionate rispetto alle basi documentali di partenza, una valutazione in termini di "recall" (calcolato come la percentuale di termini acquisiti corretta-

mente rispetto a tutti i termini presenti in GS) è stata condotta solo in relazione a un sottoinsieme di concetti selezionati come particolarmente rilevanti (v. infra).

Per quanto concerne i criteri di valutazione, oltre ai casi di corrispondenza piena tra i termini estratti automaticamente da T2K e quelli presenti nella risorsa di riferimento, sono stati considerati diversi tipi di corrispondenza parziale (per una descrizione dettagliata dei criteri di valutazione cfr. Montemagni et alii, 2007 e Lenci et alii, 2008) distinti nei seguenti casi:

- termine che nei due glossari si presenta con diverse forme prototipiche, ad esempio:
  - *accantonamento* (T2K) vs *accantonamenti* (GS_giuridico);
  - *acquisizione dati* (T2K) vs *acquisizione di dati* (GS_ambientale);
  - *abbandono di rifiuti* (T2K) vs *abbandono dei rifiuti* (GS_ambientale);
- termine che si presenta con diversa estensione di significato nei due glossari:
  - la risorsa di riferimento contiene il termine nella sua accezione più vasta, mentre T2K ha acquisito uno dei suoi iponimi, ad esempio *abrogazione di norme* (T2K) vs *abrogazione* (GS_giuridico);
  - la risorsa di riferimento contiene il termine nella sua accezione più ristretta, mentre T2K ha acquisito il suo iperonimo, ad esempio *agente di polizia* (T2K) vs *agente di polizia giudiziaria* (GS_giuridico);
- il termine estratto da T2K è un co-iponimo di un termine presente nel *gold standard*; è il caso ad esempio del termine estratto automaticamente *direttiva del consiglio*, i cui co-iponimi in *DoGi* e *JurWordNet* sono *direttiva comunitaria* e *direttiva amministrativa*.

Con l'esperimento condotto su AMB si è raggiunta una precisione del 75,4% considerando come risorse di riferimento sul versante giuridico il *Dizionario Giuridico* e le parole chiave dell'*Archivio DoGi* (Dottrina Giuridica) e sul versante ambientale il *Glossario dell'Osservatorio Nazionale sui Rifiuti* e il *Thesaurus EARTh*. Si noti che al 75,4% di precisione hanno contribuito sia i casi di corrispondenza piena sia i casi di corrispondenza parziale di cui ai punti 1) e 2) sopra. Per i rimanenti termini estratti è stata condotta una valutazione manuale che ne ha stabilito la rilevanza rispetto ai domini trattati: ad esempio, sono stati recuperati come rilevanti termini come *anidride carbonica* per il dominio ambientale, oppure *beneficiari* per il dominio legale. In questo modo, la percentuale di termini rilevanti è salita a 83,7%. Ancora migliori sono stati i risultati della valutazione del glossario estratto da CONS in relazione a *DoGi* e *JurWordNet*: i casi di corrispondenza sia totale sia parziale (casi 1), 2) e 3) sopra) costituiscono l'85,38%.

Nel caso degli esperimenti condotti con CONS, è stato inoltre possibile condurre una valutazione in termini di "recall" rispetto a un sottoinsieme di 56 concetti EULG (*European Union Legal Concepts*) selezionati come rilevanti rispetto al dominio legislato (cfr. Peters et alii, 2005 per la lista completa): in relazione a questo sottoinsieme si è raggiunto un *recall* dell'80,69%.

8. Conclusioni

T2K si presenta come uno strumento versatile e personalizzabile a vari livelli (sulla base delle caratteristiche quantitative dei repertori documentali così come delle peculiarità linguistiche dei testi e delle finalità dell'utente) per l'individuazione di tipologie diverse di informazione testuale, che vanno dall'estrazione di terminologia tecnica di dominio all'acquisizione di diversi tipi di strutture concettuali per l'indicizzazione di banche dati documentali. L'interesse di T2K non è a nostro avviso limitato al versante puramente applicativo della gestione documentale (Bourigault et alii, 2001; Jacquemin e Bourigault, 2002), ma rappresenta un importante strumento di ausilio esplorativo per l'indagine terminologica e per la strutturazione ontologica di un dominio di conoscenze.

I risultati degli esperimenti di estrazione condotti su corpora di testi giuridici dimostrano infatti che uno degli aspetti più innovativi dell'architettura di T2K sia proprio l'interazione tra livelli di annotazione della struttura linguistica della base documentale di partenza e livelli di strutturazione semantico-lessicale della conoscenza di dominio acquisita. Tale visione dinamica e incrementale del processo di accesso al contenuto dimostra quanto il preteso confine tra conoscenza linguistica e conoscenza di dominio sia inesistente in reali contesti d'uso, laddove strutture linguistiche e aspetti di conoscenza del mondo sono uniti in modo inestricabile. Il ciclo annotazione-estrazione-annotazione alla base di T2K rappresenta, a nostro avviso, una sfida metodologica importante: estraendo basi di conoscenza di dominio direttamente dal testo, utilizzando strumenti di analisi linguistici "poveri" a priori di tale conoscenza, possiamo raggiungere livelli di rappresentazione e strutturazione del contenuto progressivamente sempre più "ricchi".

Bibliografia

Abney S. (1991), *Parsing by chunks*. In: R.C. Berwick et al. (a cura di), *Principle-based Parsing: Computation and Psycholinguistics*, Kluwer, Dordrecht.

Allegrini P., Montemagni S., Pirrelli V. (2000a), *Controlled Bootstrapping of Lexico-semantic Classes as a Bridge between Paradigmatic and Syntagmatic Knowledge: Methodology and Evaluation*. In: *Proceedings of Conference on Language Resources & Evaluation (LREC 2000)*, Atene, Grecia.

Allegrini P., Montemagni S., Pirrelli V.(2000b), *Learning Word Clusters from Data Types.* In: *Proceedings of International Conference on Computational Linguistics (Coling 2000)*, Saarbruecken, Germania: 8-14.

Allegrini P., Lenci A., Montemagni S., Pirrelli V. (2002), *Le Forme del Significato. Acquisizione e Rappresentazione dell'Informazione Semantica*. In: *Actas del Segundo*

Seminario de la Escuela Interlatina de Altos Estudios en Linguistica Aplicada. Matematica y Tratamiento de Corpus, Fundaciòn San Millàn de la Cogolla, Logroño: 245-268.

Allegrini P., Montemagni S., Pirrelli, V. (2003), *Example-Based Automatic Induction Of Semantic Classes Through Entropic Scores*. In "Linguistica Computazionale", XVI-XVII: 1-45.

Bartolini R., Lenci A., Montemagni S., Pirrelli V. (2002), *Grammar and Lexicon in the Robust Parsing of Italian. Towards a Non-Naïve Interplay*. In: *Proceedings of International Conference on Computational Linguistics (Coling 2002-Workshop on Grammar Engineering and Evaluation)*, Taipei.

Bartolini R., Lenci A., Montemagni S., Pirrelli V. (2004), *Hybrid Constraints for Robust Parsing: First Experiments and Evaluation*. In: *Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, 26-28 May 2004, Centro Cultural de Belem, Lisbon, Portugal: 795-798.

Bartolini R., Giorgetti D., Lenci A., Montemagni S., Pirrelli V. (2005), *Automatic Incremental Term Acquisition from Domain Corpora*. In: *Proceedings of 7th International conference on Terminology and Knowledge Engineering (TKE2005)*, Copenhagen Business School, 17-18 August 2005, Copenhagen, Denmark.

Battista M., Pirrelli V. (2000), *Una piattaforma di morfologia computazionale per l'analisi e la generazione delle parole italiane*. Rapporto Tecnico ILC-CNR-2000.

Bourigault D., Jacquemin C., L'Homme M.C. (a cura di) (2001), *Recent Advances in Computational Terminology*. John Benjamins Publishing Company, Amsterdam-Philadelphia.

Brin S. (1998), *Extracting Patterns and Relations from the World Wide Web*. In: *WebDB Workshop at 6th International Conference on Extending Database Technology*.

Buitelaar P., Cimiano P., Magnini B. (Eds.) *Ontology Learning from Text: Methods, Evaluation and Applications.* Frontiers in Artificial Intelligence and Applications Series, Vol. 123, IOS Press, July 2005.

Calzolari N. (1984), *Detecting Patterns in a Lexical Database*. In: *Proceedings of the 10th International Conference on Computational Linguistics (COLING-84)*, Stanford, California: 170-173.

Dill S., Gibson N., Gruhl D., Guha R., Jhingran A., Kanungo T., Rajagopalan S., Tomkins A.,Tomlin J.A., Zien J.Y. (2003), *SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation*. In: *Twelfth International World Wide Web Conference*. the Semantic Web, 2005.

Dingli A., Ciravegna F., Wilks Y. (2003), *Automatic Semantic Annotation using Unsupervised Information Extraction and Integration*. In: *K-CAP 2003 Workshop on Knowledge Markup and Semantic Annotation*.

Dunning, T. (1993), *Accurate Methods for the Statistics of Surprise and Coincidence*. In "Computational Linguistics", 19(1).

Federici, S. Montemagni, S. Pirrelli, V. (1996), *Shallow Parsing and Text Chunking: a View on Underspecification in Syntax*. In: *Proceedings of the Workshop On Robust Parsing*, tenuto nell'ambito della European Summer School on Language, Logic and Information (ESSLLI-96), Praga, Repubblica Ceca, 12-16 Agosto 1996.

Fellbaum C. (a cura di) (1998), *WordNet. An Electronic Lexical Database*. MIT Press, Cambridge MA.

Firth J.R. (1957), *A synopsis of linguistic theory 1930-55*. In: *Studies in Linguistic Analysis (special volume of the Philological Society)*, Oxford, The Philological Society: 1-32.

Giovannetti E., Marchi S., Montemagni S., Bartolini R. (2007), *Ontology-based Semantic Annotation of Product Catalogues*. In: *Proceeding of the 6th International Conference in Recent Advances in Natural Language Processing* (RANLP 2007).

Grefenstette G. (1994), *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Boston.

Gruber T.R. (1995), *Toward principles for the design of ontologies used for knowledge sharing*. In "International Journal of Human and Computer Studies", XLIII, 1995: 907-928.

Harris Z.S. (1968), *Mathematical structures of language.* Wiley.

Jacquemin C. (2001), *Spotting and Discovering Terms through NLP*, MIT Press, Cambridge MA.

Jacquemin C., Bourigault D. (2002), *Term extraction and automatic indexing*. In: R. Mitkov (a cura di), *Handbook of Computational Linguistics*. Oxford University Press, Oxford.

Kogut P., Holmes W. (2001), *AeroDAML: Applying Information Extraction to Generate DAML Annotations from Web Pages*. In: *First International Conference on Knowledge Capture*.

Lin D. (1998), *Automatic Retrieval and Clustering of Similar Words*. In: *Proceedings of COLING-ACL'98*, Montreal, Canada.

Lenci A., Montemagni S., Pirrelli V. (2003), *Chunk-It. An Italian Shallow Parser For Robust Syntactic Annotation*. In "Linguistica Computazionale", XVI-XVII, 2003: 353-386.

Lenci A., Montemagni S., Pirrelli V. (2005), *Acquiring and Representing Meaning: Computational Perspectives*. In: A. Lenci S., Montemagni V., Pirrelli (a cura di), *Acquisition and Representation of Word Meaning. Theoretical and computational perspectives*, Istituti Editoriali e Poligrafici Internazionali, Pisa/Roma, Italia: 19-66.

Lenci A., Montemagni S., Pirrelli V., Venturi G., (2008), *Ontology learning from Italian legal texts*, in Breuker J. et al. (Eds.), *Legal Ontologies and the Semantic Web*, IOS-Press, (in corso di pubblicazione).

Manning C.D., Schütze H. (1999), *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge MA.

Montemagni S. (1996), *Architecture and Functioning of a System for the Acquisition of Taxonomical Information from Dictionary Definitions*. In: *Proceedings of the 4th Conference on Computational Lexicography and Text Research* (*COMPLEX '96*), Budapest, Ungheria, 15-17 Settembre 1996.

Montemagni S., Marchi S., Venturi G., Bartolini R., Bertagna F., Ruffolo P., Peters W., Tiscornia, D. (2007), *Report on Ontology learning tool and testing.* Progetto Europeo DALOS (Drafting Legislation with Ontology-Based Support), Deliverable 3.3, Dicembre 2007.

Nenadic G., Ananiadou S., McNaught J. (2004), *Enhancing Automatic Term Recognition through Term Variation*, in *Proceedings of 20th International Conference on Computational Linguistics (Coling 2004)*, Geneva, Switzerland.

Pereira F., Tishby N., Lee L. (1993), *Distributional Clustering Of English Words*. In: *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, 1993: 183-190.

Peters W., Sagri M-T., Tiscornia D. (2005), *The Structuring of Legal Knowledge in LOIS*, In *Proceedings of 10th International Conference of Artificial Intelligence and Law, (ICAIL 2005)*, Bologna, Italy, June 6th-11th.

Popov B., Kiryakov A., Kirilov A., Manov D., Ognyanoff D., Goranov M. (2003), *KIM - Semantic Annotation Platform in 2nd International Semantic Web Conference*. In: *ISWC2003*.

Rooth M., Riezler S., Prescher D., Carroll G., Beil F. (1999), *Inducing a Semantically Annotated Lexicon via EM-Based Clustering*. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, Maryland, USA, June 1999: 104-111.

Valente A. (2005), *Types and Roles of Legal Ontologie,* In Benjamins, V. R. et alii (eds.) *Law and the Semantic Web*. Springer: Berlin/Heidelberg, DE, pp. 65-76.

Vargas-Vera M., Motta E., Domingue J., Lanzoni M., Stutt A., Ciravegna F. (2002), *MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup*. In *the 13th International Conference on Knowledge Engineering and Management*.

Prospettive di relazioni fra linguistica del testo e descrizione archivistica. Il problema della denominazione

PAOLO FRANZESE

Il nome è un mezzo suscettibile di insegnare
e di farci cogliere l'essenza
(Platone, *Cratilo)*

The article discusses the problem of archival description and the objective problem linked to the denomination of archival units. With the development of the Web and the diffusion of research tools in digital format, people need to attribute a meaningful and effective denomination to records with the purpose of helping a remote consumer in searching and in the recovery of information needed.

Keywords: Archival unit – Archival description – denomination – ISAD – record management – analysis of the text

Prospettive di relazioni fra linguistica del testo e descrizione archivistica. Il problema della denominazione

Gli oggetti della vita quotidiana possono essere identificati e riconosciuti attraverso il loro nome proprio, individuale, associato al loro nome comune, di specie o di genere. Questo criterio si mostra utile anche per alcuni tipi di beni culturali, ma risulta inadeguato per i documenti, la cui identità è definibile solo in relazione con il contesto di appartenenza. Perché un documento possa essere identificato, non è sufficiente che abbia una propria denominazione [1], ma occorre far riferimento anche a quella dell'archivio, della serie e dell'unità archivistica (in genere il fascicolo) di cui fa parte. All'interno di una struttura caratterizzata da relazioni gerarchiche, il singolo documento costituisce un livello meno significativo e qualificante del fascicolo, l'unità di senso più elementare dell'archivio, che rappresenta le relazioni logiche esistenti fra i documenti. Anche l'identità di un fascicolo coincide con il posto e con il senso che esso occupa all'interno dell'archivio. Questo suo carattere relazionale si riflette nella denominazione, che esprime la sintesi dei suoi elementi costitutivi e significativi e costituisce un imprescindibile elemento d'identificazione. È soprattutto al livello del fascicolo che com-

porre la denominazione implica la definizione di criteri rigorosi per esprimere la molteplicità di elementi e di riferimenti al concreto.

Le norme ISAD (G) pubblicate dal Consiglio Internazionale degli Archivi nel 1999, che forniscono istruzioni sulla descrizione degli archivi, danno un grande rilievo alla denominazione [2], collocandola fra gli elementi compresi nell'area dell'identificazione [3]. L'argomento è trattato però da un'angolazione strettamente contenutistica, che non fornisce indicazioni sulla forma e sulla struttura logica della denominazione. Di conseguenza l'archivista impegnato nella descrizione degli archivi non dispone ancora di modelli per intitolare le unità documentarie secondo criteri corretti e condivisi.

I problemi derivanti dalla mancanza di schemi risultano enfatizzati dalla pubblicazione degli strumenti di ricerca in formato digitale sul web, per la cui consultazione l'utente remoto non può contare su alcun aiuto esterno allo strumento stesso, come invece avviene nelle sale di studio degli Archivi di Stato, dove l'assistenza diretta supplisce alle carenze e alle incoerenze della descrizione.

In questo àmbito uno dei problemi più rilevanti per l'archivista è quello di attribuire una denominazione significativa e efficace a ciò che descrive. Il titolo di un'unità documentaria svolge, come si è detto, la funzione di unificare il molteplice dei dati distintivi delle sotto-unità. Esso costituisce in sostanza un'informazione cataforica o con-densativa del documento e rappresenta un'istruzione macrolinguistica all'aspettativa, che determina nel destinatario un orizzonte d'attesa e che tende ad organizzarne il comportamento e le capacità d'interpretazione. È dunque un'operazione intellettuale delicata, che implica un'assunzione di responsabilità. In primo luogo quella di offrire le istruzioni sull'utilizzazione dei documenti all'utente, il quale, nel consultare uno strumento di ricerca, ha bisogno in un certo senso di capire prima di comprendere. Titoli non pertinenti o non esaurienti possono deviare l'attenzione dell'utente in altre direzioni e far sì che documenti riguardanti il suo studio risultino irraggiungibili. Il titolo di un'unità documentaria riflette le relazioni con l'entità gerarchica superiore (la serie, il fondo) e con le funzioni e con le attività del soggetto produttore. Gli uffici di protocollo delle amministrazioni pubbliche moderne adoperano nella titolazione dei documenti uno stile nominale simile a quello utilizzato dai giornalisti per gli articoli di stampa. "Il titolo, emblema della sintesi, è un testo brevissimo che, come tale, non segue gli stessi criteri che regolano la produzione di testi più lunghi e complessi" [4]. Fra i suoi requisiti ci sono perciò la "condizione della brevità e la condizione della coerenza" [5], rispetto al contenuto dell'entità di cui il titolo costituisce in genere la principale chiave d'accesso. Ciò può implicare l'attivazione di "meccanismi di riduzione di frase", come l'ellissi del verbo e la selezione e la messa in evidenza degli elementi informativi fondamentali [6]. "Un testo risulta più comprensibile se le sue informazioni sono organizzate in modo tale che quelle principali emergano subito rispetto a quelle secondarie. Chi legge deve cioè poter individuare immediatamente il nucleo fondamen-

tale delle informazioni presenti nel testo" [7]. Il titolo è necessariamente il risultato di una scelta fra varie possibilità, purtroppo raramente operata in base a criteri regolamentati o convenzionali e più spesso invece realizzata sulla base della tradizione o delle situazioni concrete.

Relativamente al destinatario del messaggio, il titolo non deve essere soltanto correttamente comprensibile, ma anche accettabile, cioè rilevante, e informativo, capace cioè di trasmettere nuovi elementi di conoscenza.

Una denominazione ben formata non implica che si debba puntare soltanto sulla correttezza archivistica dell'informazione e quindi sulla sua capacità di riflettere e di riassumere il contenuto dell'unità documentaria e sulla coerenza del testo. Perché un titolo sia efficace sotto il profilo della comunicazione, occorre che il suo autore tenga conto anche del destinatario, delle sue esigenze e del suo punto di vista. Questo è particolarmente importante nel caso di contenuti culturali, che comportano determinate capacità di comprensione e di interpretazione da parte dell'utente. Poiché gli strumenti di descrizione fanno parte di un servizio pubblico, la loro leggibilità costituisce un requisito richiesto dalla normativa sui servizi e sulla comunicazione. I provvedimenti prodotti nell'ambito del lungo processo di riforma della Pubblica Amministrazione hanno riconosciuto al cittadino sia un ruolo centrale come "destinatario dell'attività di erogazione dei servizi" [8], sia il diritto di accesso agli atti amministrativi e quindi hanno attribuito agli uffici pubblici il dovere della trasparenza. Al centro dell'azione amministrativa la nuova normativa ha posto i principi di efficienza, tempestività, economicità e efficacia.

A garanzia del rispetto di questi principi, si è via via imposto agli istituti l'obbligo di emanare una *Carta della qualità dei servizi* [9], sorta di patto con gli utenti e strumento di comunicazione e d'informazione. Essa nasce dall'esigenza di fissare principi e regole nel rapporto fra gli uffici che erogano servizi e i cittadini utenti.

La *Carta della qualità dei servizi* degli Archivi di Stato (2007) comprende la disponibilità di strumenti di ricerca e di banche dati a corredo degli archivi fra gli indicatori di qualità dei servizi.

Fra le condizioni dell'intesa, è da notare l'impegno della Pubblica Amministrazione di utilizzare opportuni accorgimenti per ridurre al minimo i rischi derivanti dalla distanza che la separa dall'utente.

Nella stessa direzione si pone il *Manuale per i siti web delle amministrazioni pubbliche* pubblicato dal progetto Minerva che dichiara che queste applicazioni devono essere accessibili e usabili [10].

Descrivere gli archivi dunque significa comunicare con l'utente e metterlo in condizione di accedere alle informazioni sul patrimonio documentario riguardanti le sue ricerche, in condizioni e secondo criteri di cui viene messo a conoscenza. L'archivista

può riconoscersi pertanto a pieno diritto come professionista della società dell'informazione.

Il problema della denominazione non riguarda soltanto il lavoro di attribuzione del titolo ai documenti, ma anche quello di analisi della documentazione e quindi di unità archivistiche già dotate in origine di un titolo. L'individuazione e l'interpretazione di questi titoli, che può contribuire in modo significativo a ricostruire il significato attribuito ai documenti dal soggetto che li ha prodotti, implica lo studio delle tecniche di testualizzazione adottate.

Negli elenchi e nelle rubriche degli archivi di uffici degli Stati preunitari gli incartamenti degli affari sono indicati secondo criteri che spesso riflettono le modalità con cui chi era addetto alla conservazione e alla gestione delle carte pensava di poter reperire i documenti, senza preoccupazioni di obiettività e di trasparenza nei confronti di terzi. L'archivario produceva i documenti e formava l'archivio in modo autoreferenziale, essendone lui stesso il principale se non l'unico fruitore. Poiché competeva a lui rispondere delle carte affidategli, le registrava e le rubricava secondo criteri da lui stesso definiti. Quando cambiava l'archivario, potevano cambiare anche i metodi di registrazione e di archiviazione dei documenti e quindi anche il procedimento con cui si componeva il titolo dei fascicoli. Nella mia esperienza di archivista a Napoli ho potuto verificare che, nell'ambito di questi strumenti di ricerca, la voce con cui si rinviava alla documentazione e che in genere ne rappresentava anche il titolo consisteva a volte nel nome della persona che chiedeva l'avvio del procedimento, nello scopo per cui si era presentata l'istanza e nel nome della località a cui faceva riferimento l'affare.

Nelle antiche rubriche o pandette si trovano spesso riferimenti a fascicoli espressi con intitolazioni rappresentate in forma di regesto, sorta di riassunto del contenuto dei documenti, reso attraverso una proposizione, comprendente il gruppo verbale e eventuali "partecipanti" o "ruoli sintattici". A volte l'elemento indicizzato (spesso un toponimo), posto all'inizio del titolo, è seguito da un'espressione, in parte anch'essa analoga al regesto, la quale, generalmente in forma di espansione causale o finale, fa riferimento all'ente implicato, all'antefatto, alle ragioni o agli obiettivi del procedimento.

Nelle rubriche dei dispacci della Segreteria di Stato degli affari ecclesiastici di epoca borbonica (1737-1806) si assiste al passaggio da voci costituite dal nome del luogo a cui si riferisce l'affare, seguito da quello dell'ente, a rimandi formati da una sorta di intestazione del fascicolo, composta da vari elementi, disposti in modo regolare e non casuale: il toponimo, il nome dell'ente, un'espressione che sintetizza l'oggetto dell'affare. Ne risultava un nuovo procedimento per la formazione del titolo, attraverso una proposizione di tipo finale, espressa in forma esplicita o implicita (con il predicato verbale quindi coniugato in un modo indefinito), introdotta dai connettivi "perché", "per", o attraverso un complemento di argomento, introdotto da "circa". Non sono infrequenti titoli costituiti da espressioni nominali senza predicato verbale, eventualmente seguite

da una o da più espansioni o complementi. Dell'uso di questo procedimento si hanno testimonianze anche in epoca più recente [11].

Negli uffici di polizia dello Stato italiano, a partire dall'unità, si riscontrano, nel nuovo contesto determinato dalle classifiche del titolario d'archivio, introdotto nel 1887, nomi di fascicoli realizzati con modalità di testualizzazione che riflettono gli strumenti concettuali e operativi con cui l'ufficio perseguiva le sue finalità e i suoi obiettivi. I fascicoli della Questura di Napoli [12] portano un titolo formulato con uno stile nominale e costituito in genere da tre elementi, luogo, soggetto e attività, con i quali l'ufficio di Gabinetto metteva a fuoco il significato e la funzione del fascicolo. L'ordine di questi elementi (il luogo precede il soggetto, seguito a sua volta dall'attività) ne riflette le priorità e contribuisce a dare un determinato senso al documento.

Infatti gli addetti agli archivi preparavano con attenzione le informazioni utili per la ricerca dei documenti per conto dei funzionari del proprio ufficio. Mettevano in evidenza, anticipandola, l'informazione di maggior rilievo e distribuivano in ordine d'importanza decrescente gli altri elementi informativi, separandoli con il punto, senza esprimere quindi esplicitamente relazioni sintattiche. Nell'ambito di uno stesso elemento, i dati erano disposti secondo un ordine di astrazione decrescente, dal generale al particolare [13]. Focalizzando l'informazione principale e nuova, la ricerca del fascicolo risultava più agevole. L'inventario o l'elenco del materiale documentario funzionava così da strumento di selezione dei dati, allo scopo di escludere quelli non pertinenti gli obiettivi della ricerca e di puntare invece sugli altri, analizzandone compiutamente gli elementi costitutivi. Predisposti così i dati, il reperimento del documento doveva implicare un non impegnativo lavoro di decodifica, attraverso il quale l'addetto all'archivio poteva individuare le informazioni contenute nel documento, secondo un ordine di priorità. È evidente l'analogia di questa tecnica di formazione del testo con un noto procedimento linguistico che, rispetto all'ordine normale o oggettivo degli elementi dell'enunciato, tende a privilegiare un ordine invertito o soggettivo, il cui punto di partenza è costituito da un'informazione nuova o non nota o comunque non immediatamente deducibile dal contesto, che contribuisce allo sviluppo della comunicazione avviata dalla descrizione di livello superiore. Questo modo di costruire il testo si fonda sulla distinzione fra tema e rema, rispettivamente base e nucleo dell'enunciato. Nella formulazione della denominazione, il rema del titolo di livello superiore diventa il tema, non necessariamente espresso, del titolo delle singole unità, mentre l'informazione nuova può trovarsi all'inizio del testo.

L'archivista che deve descrivere l'archivio deve valutare la pertinenza del titolo originale con il contenuto dei documenti. Quel titolo costituisce una traccia per interpretare i documenti, che tuttavia può risultare inadeguata a esprimere compiutamente e univocamente il significato dell'unità archivistica. Riutilizzarlo all'interno di un nuovo titolo nella forma di una citazione, eventualmente integrandolo, dipende dunque da

una valutazione intellettuale da parte dell'archivista, che si pone il problema di attribuire un titolo appropriato [14].

La definizione dei criteri con cui formulare la denominazione di unità archivistiche costituisce un problema che l'archivista storico condivide con chi lavora all'intitolazione dei documenti da protocollare. Rendere efficiente un sistema di registrazione significa stabilire regole e principi per formulare titoli appropriati, coerenti e informativi, in grado di identificare univocamente e di fornire idonee chiavi di ricerca.

Le modalità per l'elaborazione del testo riflettono in genere la strategia con cui l'archivista intende impostare la trasmissione dei contenuti e predisporne la corretta comprensione da parte del destinatario. Ad un'impostazione generale fondata su criteri di essenzialità e di forte coerenza logica fra gli elementi può corrispondere una denominazione le cui informazioni siano disposte dal generale al particolare, dall'astratto al concreto, con l'intento di articolare, determinare e specificare. Invece a una strategia più attenta a offrire tutte le informazioni utili sembra più consona una denominazione costituita da una sorta di enumerazione, espressa attraverso una sintassi paratattica, di elementi coordinati e quindi di pari grado.

Non essendo stata acquisita dal mondo degli archivi la sensibilità verso i problemi semantici riguardanti la ricerca delle informazioni sviluppata invece dai bibliotecari, gli archivisti si sono proposti di inserire nella denominazione gli elementi che ritengono utili alla ricerca, soprattutto nel caso di strumenti prodotti in formato digitale e sul web. In tal modo il problema delle chiavi di ricerca si è sovrapposto a quello dell'identificazione dell'unità attraverso una denominazione appropriata, autoesplicativa e esauriente. La distinzione fra i due piani dell'identificazione e della reperibilità del documento tuttavia permetterebbe di affrontare il problema dell'elaborazione dei concetti significativi e pertinenti che possano renderla tracciabile e quindi favorirne la ricerca, senza complicare e appesantire la denominazione, cui dovrebbe essere affidato il solo compito di identificare e di rappresentare l'unità archivistica nel contesto della descrizione dell'archivio. D'altro canto anche l'associazione, nell'ambito della produzione dei metadati, di parole chiave a una risorsa digitale da pubblicare sul web costituisce un'operazione che implica metodo e responsabilità e che non può essere affidata al caso o al solo intuito del suo autore.

L'efficacia della ricerca implica l'adozione di regole per la formazione dei dati. In mancanza di queste, accade infatti che non ci sia corrispondenza di vocabolario fra coloro che producono e sedimentano le informazioni e fra questi e i ricercatori. Considerevoli spazi possono aprire al mondo degli archivi lo studio e l'applicazione di metodologie di indicizzazione. L'adozione di vocabolari controllati, di soggettari e di thesauri può fornire efficaci strumenti alla ricerca dei documenti e determinare terreni d'incontro e di condivisione con altri ambiti disciplinari.

L'applicazione di questi strumenti implica un'attenta analisi concettuale del contenuto dei documenti e la traduzione dei concetti così definiti nel linguaggio di indicizzazione prescelto [15].

Note

[1] Mentre nell'ambito del lavoro d'inventariazione, che raramente prende in considerazione entità di livello inferiore al fascicolo, può rendersi utile attribuire un titolo (per la cui definizione da parte delle norme della descrizione archivistica si rinvia alle due note successive) al singolo documento, nell'attività di protocollazione se ne elabora piuttosto l'oggetto, sorta di sintesi degli elementi che si riferiscono al contenuto. Ai problemi legati alla redazione dell'oggetto è dedicato il progetto AURORA, *Amministrazioni unite per la redazione degli oggetti e delle anagrafiche nel protocollo elettronico*, i cui lavori sono attualmente in corso di svolgimento.

[2] Nel glossario dei termini associati alle regole, l'elemento è così definito: *una parola, una locuzione, un carattere alfabetico o un gruppo di caratteri che dà nome ad un'unità di descrizione*.

[3] A proposito della compilazione della denominazione, le ISAD forniscono le seguenti istruzioni: *In un titolo attribuito includere, al livello più elevato, la denominazione del soggetto produttore della documentazione. Ai livelli inferiori può essere incluso, ad esempio, il nome dell'autore del documento e un termine che indichi la tipologia documentaria che costituisce l'unità di descrizione e, se risulta opportuno, una locuzione che faccia riferimento alla funzione, all'attività, all'oggetto, alla localizzazione geografica o all'argomento*.

[4] A. CICALESE, *Titolazione e immagini nella stampa quotidiana*, in *Scrivere per comunicare*, a cura di G. PALLOTTI, Milano, Bompiani, 2001, p. 140.

[5] *Ibidem*.

[6] *Ibidem*.

[7] M.E. PIEMONTESE, *La comunicazione pubblica e istituzionale. Il punto di vista linguistico*, in *Manuale della comunicazione*, a cura di S. Gensini, Roma, Carocci, 1999, pp. 337-338.

[8] Direttiva del Ministero per i beni e le attività culturali, 18 ottobre 2007 su *Carta della qualità dei servizi degli istituti e dei luoghi della cultura.*

[9] Dal 1995 la pubblicazione della Carta della qualità dei servizi costituisce un obbligo.

[10] La norma ISO 9241/1993 identifica l'usabilità con efficacia, efficienza e soddisfazione con le quali determinati utenti raggiungono determinati obiettivi in determinati contesti.

[11] Nell'archivio del Ministero della Presidenza del Consiglio dei ministri del Regno delle Due Sicilie (1806-1869) un fascicolo porta il seguente titolo: *Per l'infausta morte di S.A.R. il principe di Salerno Leopoldo di Borbone*.

[12] P. FRANZESE, *Le Disposizioni di massima. Formazione dell'archivio e analisi del modello descrittivo*, in *L'archivio della Questura di Napoli. Inventario delle Disposizioni di massima*, inventario a cura di Giuliana Buonaurio, Napoli, Luciano editore, 2000, pp. 7-16.

[13] Nell'esempio che segue si può notare come dell'elemento attività siano presenti, nella denominazione di questo fascicolo di polizia, due informazioni, la prima delle quali è cataforica rispetto all'altra: Attori di varietà (soggetto). Divieto di esibirsi in pubblico (attività). Travestimenti femminili (attività).

[14] Le regole raccomandano il rispetto del titolo originale, eventualmente accorciato se troppo esteso, "a condizione che ciò non determini la perdita di qualche informazione essenziale". In ogni caso il titolo originale va distinto da quello attribuito.

[15] A questo problema si riferisce la norma UNI ISO 5963 dell'ottobre 1989, intitolata *Metodi per l'analisi dei documenti, la determinazione del loro soggetto e la selezione dei termini d'indicizzazione*.

Bibliografia

Cicalese A., *Titolazione e immagini nella stampa quotidiana*, in *Scrivere per comunicare*, a cura di G. Pallotti, Milano, Bompiani, 2001, p. 140.

Franzese P., *Le Disposizioni di massima. Formazione dell'archivio e analisi del modello descrittivo*, in *L'archivio della Questura di Napoli. Inventario delle Disposizioni di massima*, inventario a cura di G. Buonaurio, Napoli, Luciano editore, 2000, pp. 7-16.

ISAD(G): General International Standard Archival Description, Adopted by the Committee on Descriptive Standards, Stockholm, Sweden, 19-22 September 1999, Ottawa, 2000.

ISO 9241 - 1993, *Ergonomic requirements for office work with visual display terminals.*

Ministero per i beni e le attività culturali, *Carta della qualità dei servizi degli istituti e dei luoghi della cultura*, 18 ottobre 2007.

Piemontese M.E., *La comunicazione pubblica e istituzionale. Il punto di vista linguistico*, in Manuale della comunicazione, a cura di S. Gensini, Roma, Carocci, 1999, pp. 337-338.

UNI ISO 5963 - 1989, *Metodi per l'analisi dei documenti, la determinazione del loro soggetto e la selezione dei termini d'indicizzazione*.

Terminologia e documenti per la formalizzazione standardizzata della conoscenza

ELENA CARDILLO, ANTONIETTA FOLINO

The aim of this work is to present the results of an ongoing project on tacit knowledge representation and on the creation of a Knowledge Based System in the domain of artistic handicrafts. We have chosen CommonKADS methodology as a formalization standard. We have used the PCPACK software as a support application which allows for the extraction and modelling of the acquired knowledge. The peculiarity of the application domain, represented by the field of goldsmith handicraftsmanship in Calabria, has determined, during the different phases of the project, an interesting activity of personalisation, aimed at solving terminological and documentary criticalities. The ultimate objective is the creation of a system able to support the individual decisional process, to manage and to represent the whole and the complexity of the information related to the professional know-how of the artisans of Calabria, and to keep track of both the terminological and technical specificities of experts.

Keywords: Knowledge Management – CommonKADS – KBS development – Goldsmith Handicraft – Classification System

Introduzione

L'esigenza di creare un Sistema a Base di Conoscenza (KBS) per settori specialistici, quali quello dell'artigianato, in cui il *know-how* degli esperti, consolidatosi quasi esclusivamente con l'esperienza, è quasi privo di formalizzazione, nasce dalla necessità di salvaguardare tale sapere, rendendolo formale, esplicito e condivisibile anche ai fini della continuazione dell'attività.

La tradizione artigiana in Calabria continua a presentare – anche in virtù del perdurare di specifiche condizioni socio-politiche – i tratti tipici di un patrimonio di esperienze e conoscenze non sempre riconducibile alla sola finalità economica e produttiva. Nonostante ciò, la "globalizzazione" del mercato richiede prodotti e materiali che presentino precisi *standard* di qualità e consentano la descrizione e ripetibilità delle procedure utilizzate. Il tramandarsi della tradizione dei maestri artigiani è ulteriormente minacciato da altri fattori, quali: lo stato delle conoscenze disperso e non organizzato; la carenza di conoscenze tecniche sul dominio e sulla specificità locale delle tecniche e,

infine, il limitato accesso ai dati e alle informazioni. Si rende, quindi, sempre più in-dispensabile definire procedure e modalità per una gestione di tale conoscenza che tenga conto della particolarità delle tecniche di lavorazione, dei materiali e degli strumenti utilizzati, della manualità e dell'unicità dei modelli. Punto focale diventa quindi la possibilità di capitalizzare e standardizzare i processi senza per questo omologarli a tecniche più tipicamente industriali ma anzi, utilizzando metodi e tecniche tipiche del *Knowledge Management*, esaltandone la tipicità e rendendoli fruibili per specifici e successivi utilizzi.

Per perseguire tale finalità, si è deciso di sperimentare l'applicabilità, allo specifico settore, di una metodologia di acquisizione e rappresentazione della conoscenza, il CommonKADS, che prevede l'utilizzo di differenti tecniche di elicitazione per estrarre l'*expertise* e la terminologia di base del dominio di applicazione, e la costruzione di modelli formali che descrivono i diversi aspetti del dominio, permettendo – infine – la formalizzazione della conoscenza.

Ci si è proposti, in primo luogo, di riorganizzare e analizzare il *corpus* di conoscenza esplicita esistente sul dominio e, in seguito, di estrarre le conoscenze tacite relative all'àmbito di applicazione sperimentale individuato. A tale scopo sono state utilizzate varie tecniche di elicitazione, applicate durante incontri periodici ai maestri artigiani aderenti al progetto. Esse si sono concretizzate in: interviste non strutturate, semi-strutturate, *self report*, *shadowing*, *card sorting*, *repertory grid*. Interviste e *self report* sono stati poi trascritti e analizzati per mezzo del *toolkit* PCPACK5 (CommonKADS), un *software* adattabile alla metodologia CommonKADS, realizzato dalla Epistemics nel 1994, ma ancora utilizzato nei processi di acquisizione e concettualizzazione della conoscenza. Per la seconda parte del progetto, la formalizzazione della conoscenza estratta, ci si prefigge di creare un modello formale, ontologico, unificato e condiviso con gli attori coinvolti nel processo, e infine di creare un sistema di supporto alle decisioni che utilizzi la base di conoscenza come fonte da cui attingere per rispondere alle *query* degli utenti. Un'ulteriore ipotesi è quella di accedere alla base di conoscenza ottenuta, tramite un sistema di classificazione a faccette, che migliori l'organizzazione delle informazioni, l'accesso e la navigabilità della stessa base di conoscenza. Ci si propone quindi di adottare un FCS (*Faceted Classification System*), per dimostrare come classificare e organizzare le informazioni in gerarchie multidimensionali sia più accessibile rispetto a una singola tassonomia, a un'unica dimensione gerarchica.

2. Obiettivi e finalità

L'obiettivo di questo lavoro è quello di creare una piattaforma sperimentale che permetta di sviluppare, grazie al recupero di modelli tradizionali basati sulla ricerca do-

cumentale, un modello di innovazione locale nel settore dell'artigianato, nel quale, come abbiamo precedentemente accennato, si avverte fortemente il bisogno di formalizzare e rappresentare quella conoscenza che i maestri artigiani hanno acquisito in anni di esperienza e che permette loro di realizzare prodotti esclusivi e di alto valore artistico, attraverso tecniche tradizionali. Ci si propone di creare un Sistema a Base di Conoscenza, che riduca, di fronte ad una problematica operativa, le incertezze del processo decisionale e favorisca la rapidità di scelta strategica. Un monitoraggio quindi della gestione del saper fare operativo dei maestri artigiani calabresi, dei materiali tradizionali e della documentazione esplicita della tradizione artigianale che possa, successivamente, condurre ad un sistema documentale interrogabile in grado di risolvere in tempo reale i problemi posti dagli utenti.

3. Metodologia

Le metodologie di acquisizione e modellazione della conoscenza relativa ad un determinato dominio, prevedono in generale 2 tipi di approcci: *bottom-up*, e *top-down*.

Il primo approccio consiste nel raccogliere la maggior quantità di dati verbali a partire dagli esperti del dominio, dati che successivamente vengono organizzati in un modello. Il secondo approccio si focalizza, da subito, sulla definizione di un modello di *expertise* al fine di filtrare la conoscenza acquisita e di guidare efficacemente i processi di acquisizione della stessa.

Per la costruzione di un KBS consultabile e interrogabile, capace di analizzare e gestire la conoscenza tacita nel campo di applicazione del settore artigianale, si è ritenuto opportuno seguire un approccio misto, partendo dall'estrazione della conoscenza tacita degli esperti del dominio, fino alla costruzione del modello di Conoscenza (*Knowledge Model*) e all'implementazione del modello concettuale, che ha poi supportato il successivo processo di elicitazione. Tutto ciò secondo i dettami della metodologia su cui si basa lo Standard KADS (*Knowledge Acquisition and Documentation Structuring*) (Martin, 1994), sviluppato dall'Università di Amsterdam, come parte del programma ESPRIT, in cooperazione con numerosi partner europei. Lo standard fornisce un *framework* di rappresentazione della conoscenza e suggerisce i processi necessari per la costruzione di un *Knowledge Based System* (KBS), a differenti livelli di astrazione. In particolare è stata scelta, con i dovuti adeguamenti alle esigenze del caso, una versione semplificata dello standard, la metodologia CommonKADS, la quale è stata da sempre utilizzata per problemi di *Knowledge Elicitation*, e più in generale come *baseline* per lo sviluppo di sistemi e progetti di ricerca orientati alla conoscenza. La particolarità di questa metodologia risiede nell'approccio strutturato e basato sull'uso di modelli, e nell'importanza

attribuita alla fase di analisi, completamente indipendente da qualsiasi decisione relativa all'implementazione del sistema.

Il CommonKADS si basa su un modello di ciclo di vita, detto *Spiral Model* proposto da (Boehm, 1988), che sostituisce la sequenzialità del tradizionale modello a cascata (*Waterfall*). Lo *Spiral Model* include le seguenti fasi: *Scoping and feasibility study*, *Impact and improvement study*, *Knowledge Analysis*, *Communication interface analysis*, *System design*, *Knowledge-system implementation*. Per ciascuna di queste fasi sono previste quattro attività cicliche: *Review*, *Risk*, *Plan e Monito*r (Akkermans, Anjewierden et alii., 2000), che consistono, rispettivamente, nell'analisi dello stato del progetto, nell'identificazione e valutazione dei rischi, nella pianificazione delle attività successive alla luce dei rischi precedentemente indicati e, infine, nel monitoraggio del lavoro e riesame del nuovo stato del progetto. Lo scopo di tale approccio è di rendere flessibile la progettazione del sistema per mezzo di cicli consecutivi adattabili sulla base dell'esperienza di quelli precedenti.

Nell'applicazione di tale metodologia al dominio di riferimento, si è deciso di far confluire queste attività consecutive in fasi più generali quali: analisi, concettualizzazione e formalizzazione, design e implementazione.

Per quanto riguarda la fase di analisi, che consiste nell'analisi interna del dominio di riferimento, è previsto il processo di acquisizione della conoscenza, tramite tecniche di elicitazione, e la successiva sua interpretazione, analisi e modellazione. Tale fase prevede la creazione di vari modelli per la scomposizione dei *task* di *Knowledge Engineering*, utilizzati proprio per enfatizzare alcuni aspetti della costruzione del KBS. Tali modelli si distinguono in:

- *Organization Model*, che supporta l'analisi delle principali caratteristiche dell'organizzazione, al fine di evidenziare i problemi e le opportunità legate all'introduzione di un KBS,
- *Task Model*, che analizza i *task* eseguiti all'interno dell'organizzazione, con particolare riferimento anche alle risorse e alle competenze necessarie per la loro realizzazione,
- *Agent Model*, che presenta gli agenti che eseguono i *task* evidenziati nel precedente modello.

In particolare, vengono messe in evidenza le competenze e le responsabilità degli stessi, così come i rapporti comunicativi che si istaurano tra di loro. Questo ultimo aspetto viene meglio esplicato tramite il

- *Communication Model*, nel quale vengono forniti chiarimenti relativi alla tipologia di oggetti informativi o di transazioni scambiate tra i vari agenti nell'esecuzione di un determinato *task*.

Il modello più rilevante in termini di funzione svolta e di informazione veicolata è rappresentato dal *Knowledge Model*, che, per l'appunto, descrive e struttura la cono-

scenza necessaria all'esecuzione di un particolare *task*. È costituito da tre livelli di rappresentazione della conoscenza stessa, ovvero il *domain level*, nel quale viene identificata la conoscenza di dominio, *l'inference level*, che descrive le inferenze che possono essere fatte a partire da questa conoscenza e infine il *task level* (Hickman, Killin et alii., 1989), che specifica l'ordine delle varie inferenze relativamente all'esecuzione di un dato *task*. Preliminare alla costruzione di questo modello è la scelta di un *template* generico (Breuker, Van de Welde et alii., 1994) che possa adattarsi al dominio di riferimento e agli obiettivi da raggiungere con la realizzazione del sistema: diagnosi, classificazione, predizione, valutazione, ecc. Indipendentemente dal *template* scelto, la sua realizzazione, e l'eventuale adattamento allo specifico contesto, supporta il processo di acquisizione della conoscenza e l'analisi dei dati verbali. La metodologia prevede inoltre, per la realizzazione del *Knowledge Model* , l'utilizzo del linguaggio strutturato e semiformale CML, *Conceptual Modelling Language* (Breuker, Schreiber, et alii., 1993), che consente la definizione di ciascuna delle parti previste dal modello e in particolare del *domain schema*, della *knowledge base*, dei concetti, delle relazioni, e di simili elementi di conoscenza. La sintassi alla base del CML fa si che questo linguaggio si collochi a metà strada tra un linguaggio formale, quale l'XML, pensato per essere interpretato dalla macchina e un linguaggio più informale vicino a quello naturale e perciò più ambiguo.

Il *design model* è l'unico modello relativo all'omonima fase di *design* e presenta l'architettura del sistema finale e le sue caratteristiche più propriamente tecniche, cercando di preservare nelle strutture e nei contenuti quanto specificato nei modelli costruiti durante la fase di analisi (Akkermans, Anjewierden, et alii., 2000). Le relazioni tra i modelli previsti sono mostrate nella seguente figura:

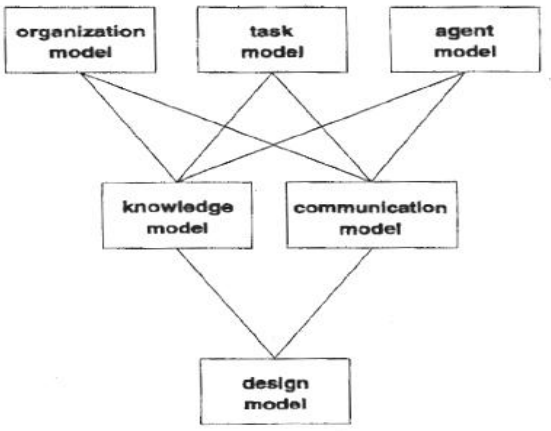


Figura 1 - The CommonKADS model suite

La costruzione dei modelli è supportata dalla fase di acquisizione della conoscenza, a partire dagli esperti del dominio, che rappresenta il punto nodale dell'intero processo di costruzione di un KBS e prevede l'applicazione di una serie di tecniche di elicitazione previste dalla stessa metodologia. Queste tecniche vengono utilizzate in momenti diversi del processo di elicitazione, perchè ciascuna di esse consente di catturare una certa tipologia di conoscenza e di raggiungere determinati scopi. Le tecniche più comuni sono le interviste che, in base al grado di flessibilità e di formalizzazione, si distinguono in non strutturate, semi-strutturate e strutturate.

Vi è poi una serie di tecniche basate sull'osservazione diretta delle *performance* dell'esperto, che consente di estrarre conoscenza procedurale, maggiormente legata alla manualità (*Think Aloud Problem Solving*, *Self-report,* e *Shadowing)*. Se si è invece interessati a capire come l'esperto concettualizzi la conoscenza relativa al proprio dominio di riferimento, è opportuno ricorrere a tecniche quali il *Card Sorting* e il *Repertory Grid*. Oltre a queste tecniche, la metodologia prevede la possibile applicazione di altre tecniche per l'estrazione della conoscenza tacita, tra le quali: *Twenty-Questions technique,* l'utilizzo di *Ladders*, mappe concettuali, matrici e diagrammi (Milton, 2007).

4. Applicazione

4.1. Analisi

Come precedentemente accennato, il settore scelto per valutare l'applicazione di questo tipo di approccio metodologico per la costruzione di un KBS è quello dell'oreficeria, la cui scelta è stata motivata dalle specificità esposte in premessa.

Di notevole importanza, in tal senso, è stata la possibilità della preconoscenza delle macro-caratteristiche del dominio forniteci dal Consorzio delle imprese Artigiane (Coser) della Regione Calabria che ha collaborato anche alla selezione dei sedici orafi che hanno ottenuto il riconoscimento della qualifica di Maestro Artigiano (L.R. n.15, 2002), che viene conferita sulla base di specifici criteri, quali un'anzianità professionale di almeno 10 anni, un adeguato grado di capacità professionale e un'elevata attitudine alla trasmissione delle abilità.

Attività preliminare per la costruzione del KBS è stata quella della ricerca bibliografica e documentale al fine di recuperare ogni informazione disponibile sullo specifico comparto, con particolare attenzione ai disciplinari tecnici, alle tecniche di lavorazione, agli strumenti utilizzati, ed alla terminologia di base. Le descrizioni reperite sono risultate, comunque, non approfondite e quasi esclusivamente limitate alla finalità turistica e propagandistica. Nessuna indicazione è stata, invece, reperita su esperienze specifiche di formalizzazione della conoscenza. Per le specifiche conoscenze sulle tecniche di lavo-

razione dei metalli ci si è avvalsi delle competenze presenti nel Dipartimento di Meccanica dell'Università della Calabria.

Si è passati poi alla fase di elicitazione della conoscenza e quindi al contatto diretto con gli esperti del settore orafo. La prima tecnica di elicitazione utilizzata è stata quella dell'*Interviewing*.

È stato quindi, innanzitutto, creato un questionario semi-strutturato da sottoporre all'esperto, evitando di formulare domande troppo specifiche che avrebbero potuto ingenerare difficoltà o aumentare la già naturale reticenza che è fenomeno comune in situazioni di questo genere, per estrarre una prima conoscenza superficiale del dominio di riferimento. Nel corso dei primi incontri sono state formulate domande circa la specificità della professione, la tipologia dei prodotti, le tecniche di lavorazione dei singoli metalli e la tipologia degli strumenti utilizzati. Del colloquio è stata effettuata una registrazione *audio*, ai fini della sua successiva trascrizione per l'inserimento nel *tool* PC-PACK5.

Nella consapevolezza della non esaustività delle interviste circa l'estrazione di conoscenza tacita dall'esperto, si è convenuto di utilizzare, a seguire, la tecnica della *Protocol Analysis*.

L'analisi dei protocolli, ovvero delle registrazioni *audio* o *video* delle *performance* degli esperti/professionisti, permette di sottolineare gli elementi di base della conoscenza: concetti, attributi, valori, *task*, relazioni. La tecnica da noi scelta per la *Protocol Analysis*, è il *Self-report*, tramite il quale ci siamo proposti di avere un'idea più chiara del processo di lavorazione completo dell'oggetto prezioso, dalla fusione alla commercializzazione. In questo caso è stata effettuata una registrazione *video* di tutto il processo e una successiva identificazione dei *task* e trascrizione ai fini dell'analisi.

Al *Self-report*, in alcuni casi, sono stati aggiunti lo *Shadowing*, il *Repertory Grid* e il *Card-Sorting,* precedentemente citati.

4.2. Concettualizzazione e Formalizzazione

Alla fase di acquisizione della conoscenza segue una fase di analisi e concettualizzazione della stessa. A tal proposito sono stati utilizzati vari *tools* presenti nel *software* citato. In particolare il *Protocol tool*, che ha permesso l'analisi dei testi trascritti, la costruzione delle classi (strumenti, tecniche, materiali, ecc.) e la successiva disaggregazione degli elementi informativi contenuti nelle classi individuate. Così categorizzati, i vari elementi confluiscono automaticamente nella base di conoscenza e in un *ontology browser* che ne consente la visualizzazione e l'utilizzo nei restanti *tool* di modellazione.

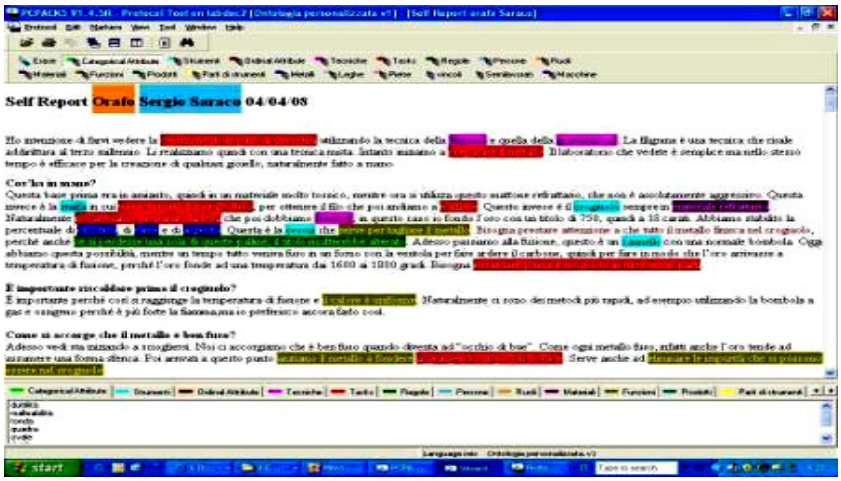


Figura 2 - Protocol Tool: trascrizione intervista

I *tool* utilizzati per modellare la conoscenza di interesse sono stati il *Ladder Tool*, il *Diagram Tool*, il *Matrix Tool* e l'*Annotation Tool*. Con il primo sono state realizzate delle strutture arborescenti che stabiliscono una relazione gerarchica tra gli elementi inseriti, siano essi concetti, *task* o attributi. La categoria di elementi inseriti determina anche la natura del *ladder*, che può essere *Concept Ladder*, *task Ladder* e via dicendo; il *Diagram Tool* consente la costruzione dei diagrammi, che possono assumere diverse connotazioni sulla base delle modalità di rappresentazione degli elementi, della tipologia di elementi stessi e di regole che tra essi si stabiliscono (*process map, activity diagram, concept map*, ecc.). In particolare, nei diagrammi sono state inserite relazioni di sequenzialità tra i vari *task* individuati come facenti parte di un dato processo. Il *Matrix Tool* permette di creare, per l'appunto, matrici che, in quanto tali, consentono di mettere in relazione o diversi tipi di oggetti di conoscenza, nel qual caso si otterrebbe una *relationship matrix*, o i concetti con i relativi attributi, realizzando così una *attribute matrix*. Infine è stato utilizzato l'*Annotation Tool*, che permette di ottenere modelli simili a schede terminologiche i quali forniscono indicazioni più dettagliate riguardo agli elementi cui si è maggiormente interessati. I campi inseriti possono comprendere la descrizione, le relazioni, i sinonimi, il contesto in cui l'elemento compare e via dicendo, in modo tale da mantenere anche il legame tra l'elemento stesso e la fonte, in questo caso la trascrizione, dalla quale lo stesso è stato estratto. Il prodotto finale è costituito da un *knowledge model* XML che preserva le informazioni e le relazioni di volta in volta inserite.

Particolare attenzione è stata rivolta a non uniformare le informazioni fornite dai maestri artigiani in un'unica base di conoscenza indifferenziata, ma anzi a mantenere

l'individualità di ciascuno in considerazione delle peculiarità operative di ogni singolo processo.

Uno dei dati, presente in letteratura e verificato sul campo, è quello relativo alla difficile applicazione dello *standard* alla modellazione di un intero dominio di riferimento, prestandosi meglio ad attività relative a singoli processi di lavorazione che siano *knowledge intensive*, ovvero che richiedano un particolare carico cognitivo, e per la risoluzione dei quali è necessario costruire un KBS. Per questo motivo sono stati implementati solo alcuni modelli, descritti in precedenza: *Organization Model*, *Agent Model*, *task Model*, e il *Knowledge Model*. In particolare quest'ultimo, ottenuto in seguito all'utilizzo di tutti i *tool* di PCPACK, fornisce formalismi per la rappresentazione sia delle strutture concettuali statiche sia delle contingenze basate su regole di decisione specifiche del dominio. Per la sua realizzazione è stato utilizzato il modello della diagnosi in considerazione della finalità ultima di supporto alla decisione. CommonKADS sviluppa solo modelli di conoscenza semi-formali (il *Knowledge Model* infatti viene formalizzato in linguaggio CML o XML), che risiedono cioè ad un livello di formalità e risoluzione semantica intermedia tra quella dei prodotti iniziali della conoscenza acquisita e le rappresentazioni in OWL. PCPACK, a questo proposito prevede un *plug-in* che permette la trasformazione automatica della base di conoscenza da XML ad OWL pur tenendo presente, comunque, che il problema principale, in questo caso, è che la struttura della base di conoscenza non può essere facilmente convertita in formalismi centrati su proprietà, in base a quanto previsto dai linguaggi di *Description Logic*. Per questi motivi, una delle possibilità può essere rappresentata da una trasformazione delle informazioni tassonomiche principali in classi OWL definendo, contestualmente, relazioni, vincoli, processi e regole.

4.3. Design *e Implementazione*

La fase di *design* riguarda la produzione della struttura che supporta direttamente il sistema esperto. La metodologia CommonKADS prevede a questo punto l'utilizzo del *Design Model*, che si avvale della conoscenza ottenuta in seguito alla creazione dei modelli precedenti, per fornire in *output* la specificazione della architettura *software* e il *design* dell'applicazione entro tale architettura. Il processo di *design* prevede le seguenti fasi di esecuzione: architettura del sistema (*design architecture*); specificazione della piattaforma *hardware* e *software* (*specify hw/sw platform*); specificazione dettagliata dell'architettura (*detailed architecture specification*); *design* dettagliato dell'applicazione (*detailed application design*). Pur nella convinzione della complessità delle procedure per la costruzione di un sistema esperto – ove si riscontrassero problematicità nella costruzione di strutture OWL – un'ulteriore ipotesi può essere rappresentata dalla formalizzazione

delle regole riguardanti i processi di lavorazione in CLIPS (Milton, 2008). Scegliendo questa alternativa, il punto di partenza sarebbe costituito dal *Knowledge Model* XML ottenuto per mezzo di PCPACK dal quale, attraverso un'intermedia trasformazione in un foglio di stile XSL, si potrebbe passare quasi automaticamente al codice CLIPS.

Questa trasformazione richiede, però, una strutturazione diversa della base di conoscenza creata con PCPACK. Si rende necessaria, infatti, l'identificazione di ulteriori classi, quali *input*, problemi e soluzioni, che consentano la modellazione delle regole decisionali, non altrimenti supportabili dal *software*. In particolare la classe *input* si riferisce alla richiesta di informazione che il sistema rivolge all'utente per meglio comprendere il problema sorto durante la lavorazione, mentre l'associazione delle istanze delle classi problema e soluzione permette di fornire possibili risposte per ciascuno dei problemi individuati.

5. Conclusioni parziali e prospettive

La sperimentazione applicativa che si intende realizzare ha tre obiettivi: acquisire conoscenza tacita; supportare le imprese artigiane nell'assunzione di decisioni inerenti aspetti specifici del processo produttivo e, cosa ben più importante, contribuire a gestire quella delicata fase di transizione che si ha nel passaggio dell'attività da un maestro artigiano all'altro e che, se non adeguatamente supportata, determina – comunque – una dispersione del patrimonio acquisito nel corso della vita lavorativa.

Tra le prospettive possibili, in corso di definitiva sperimentazione, oltre a quelle citate merita di essere indicata la sostituzione del sistema esperto con un sistema di classificazione a faccette o *thesaurus* a faccette, le cui peculiarità risiedono nella flessibilità, nella multi-dimensionalità e nella possibilità di essere adattata a qualsiasi contesto applicativo, da associare al modello di conoscenza XML generato dal *software* PCPACK e la trasformazione dello stesso in un prodotto consultabile. Adottando quest'ultima soluzione, l'accesso alla base di conoscenza potrebbe avvenire per mezzo di una lista terminologica chiusa, all'interno della quale i termini saranno organizzati in preferiti e non preferiti, inserendo anche sinonimi o varianti, in maniera tale che siano molteplici i punti dai quali poter accedere all'informazione. In tale contesto sarebbe possibile mantenere l'individualità terminologica di ciascun orafo. La specificità del dominio e del relativo lessico, oltre al diverso livello di scolarità, infatti, fa si che spesso siano presenti incongruenze terminologiche, per cui ad uno stesso oggetto, al quale viene riconosciuta una stessa funzione, vengono attribuiti termini diversi, o addirittura anche laddove si riconosca una funzione diversa, una stessa denominazione viene attribuita ai due oggetti distinti.

Bibliografia

Akkermans H., Anjewierden A. *et alii*, (2000). *Knowledge engineering and management: the CommonKADS methodology,* Cambridge, The MIT Press.

Boehm B., (1998). *A spiral model of software development and enhancement*, "EEE Computer", pp. 61-72.

Breuker J, Van de Velde W., (1994). *CommonKADS library for expertise modelling: reusable problem solving components,* Amsterdam, IOS Press.

Breuker J., Schreiber, G., (1993). *KADS: a principled approach to knowledge-based system development*, London, Academic Press.

Esprit Project 12, (1983). *A methodology for the design of knowledge based systems*, Université de Amsterdam et Knowledge Based System Center, Polytechnique de South Bank.

Hickman *et al.*, (1989), *Analysis for knowledge-based systems: a practical introduction to the KADS methodology*, Ellis Horwood, Chichester.

L.R. n.15 del 15 Marzo 2002: *Norme sulla tutela, il recupero e la promozione dell'artigianato artistico e tipico della Calabria*.

Martin P., (1994). *La Méthodologie d'acquisition des connaissances KADS et les explications*, Rapport de recherche N. 2179, Paris, INRIA;

Milton N.R. (2007). *Knowledge acquisition in practice: a step-by-step guide*. London, Springer.

Milton N.R. (2008). *Knowledge technologies*. Milano, Polimetrica.

Sitografia

CommonKADS, PC-PACK5, <www.commonkads.uva.nl/frameset-commonkads.html>.

Coser, <www.artigianatocalabria.it/mailer.php>,

Fra Terminologia e Documentazione: estrazione automatica di voci indice da *corpora* documentali della Pubblica Amministrazione

MARIA TAVERNITI

The purpose of this article is to analyse the problems related to descriptors, the contents of documents, and their role in the process of retrieval and management of information. The starting point for the case study is the construction of a knowledge extraction system from a corpus of documents produced by Italian Ministries and used for request for conformity opinions at the Centro Nazionale per l'Informatica nella Pubblica Amministrazione (CNIPA).

Keywords: Information Science – descriptor – terminology extraction – thesaurus, t2k

Introduzione

Quando ci troviamo nella necessità di effettuare una estrazione terminologica finalizzata o meno alla definizione di termini di indicizzazione, la domanda che ci poniamo è: "quali sono le unità informative e di conoscenza, ovvero i termini che dobbiamo rilevare al fine di meglio rappresentare un àmbito specialistico?". Per rispondere a tale quesito è necessario prima di tutto individuare gli elementi capaci di riconoscere le unità significanti proprie di uno specifico dominio di conoscenza. Ciò diventa tanto più importante nella costruzione e utilizzazione di sistemi automatici di estrazione terminologica, nei quali tali attività costituiscono la necessaria analisi di fattibilità e l'indispensabile banco di prova del sistema. Tutto ciò in linea generale e con l'obbiettivo di definire teoricamente se la finalità dell'azione – documentale o terminologica – sia o meno influente ai fini della determinazione delle azioni da mettere in essere e delle specificità dei termini da estrarre.

Il caso applicativo è la volontà del Centro Nazionale per l'Informatica nella Pubblica Amministrazione di costruire un vocabolario di indicizzazione per la gestione normalizzata e l'estrazione di conoscenza dai documenti relativi ai pareri obbligatori sugli acquisti – economicamente significativi – di beni e servizi informatici proposti dalle pubbliche amministrazioni centrali.

Il *corpus* di riferimento è composto da 46 documenti multipagina – provenienti da varie amministrazioni – per 346.000 *token* (Lenci *et alii,* 2005). Esso, pur se di dimen-

sione quantitativa ridotta – ben rappresenta il dominio o, meglio, il sotto-dominio di riferimento.

Il presupposto dal quale partiamo è: "chi usa termini – oltre ad affermare, negare, domandare – presuppone definizioni condivise entro una comunità"(Gobber, 2007). Lo stesso concetto è espresso da Maria Teresa Cabré nel corso della *I Jornada de Terminologia i Documentació*. In altre parole, la Cabré precisa che non vi è alcun dubbio sul fatto che gli specialisti dispongano tutti di una terminologia – non necessariamente formalizzata – capace di descrivere il loro àmbito di specialità (Cabré, 2000).

In questo lavoro consideriamo come Unità Informativa (Cabré, 2000), il singolo parere composto da una pluralità di tipologie documentali sostanzialmente identificabili nel parere vero e proprio e nella documentazione preparatoria e di corredo, che alla formulazione di quel parere porta. Ognuna, quindi, rappresenta l'insieme dei documenti concorrenti a formare il flusso documentale delle singole azioni contrattuali (parere decisionale, capitolato tecnico, schema di riferimento, gara d'appalto, proposta di progetto, ecc.). Assumiamo pertanto che i termini candidati dell'unità informativa presa in esame, con il valore tf*idf – che illustreremo in seguito – più alto rappresentino gli elementi più significativi dell'unità stessa. Ciò in quanto il numero di volte che essi occorrono nei documenti di tutta la collezione analizzata, equivale al valore della loro frequenza assoluta nell'unità informativa. Nel nostro caso, abbiamo suddiviso il nostro *corpus* iniziale in 11 unità informative. Nelle tabelle che seguono sono riportate le informazioni quantitative relative sia alle tipologie contrattuali che alle unità informative così come le abbiamo intese.

ID_Unità informativa	*Token*	N. Candidati estratti da t2k
1	42536	718
2	39064	888
3	66153	997
4	41310	836
5	37702	899
6	31059	645
7	10636	385
8	21386	555
9	8944	388
10	20525	531
11	26558	522
Totale	**345873**	**7364**

Tabella 1

Tipologia contrattuale:	*Token*
Parere	68371
Contratto	104647
Relazione	68358
Gara d'appalto/Capitolato tecnico	81009
Disciplinare di gara	23488
Totale *Token*	**345873**

Tabella 2

Come illustrato nei due grafici (Figura 1 e 2), la scelta di considerare come "unico" documento le differenti tipologie documentali relative alla stessa attività contrattuale è dettata dall'esigenza d'avere un corpus di riferimento bilanciato (Lenci *et al.,* 2005*)*. Diversamente, una specifica tipologia avrebbe quantitativamente prevalso su un'altra.

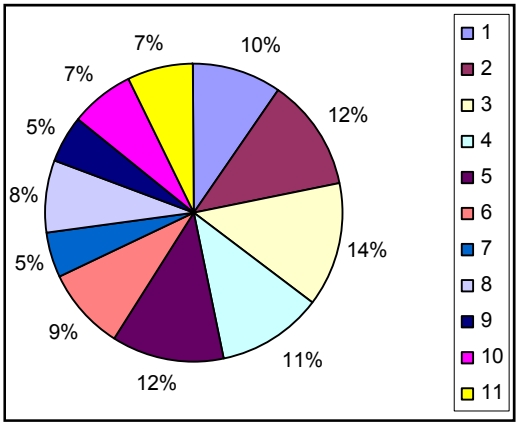


Figura 1 - Rilevanza delle Unità informative

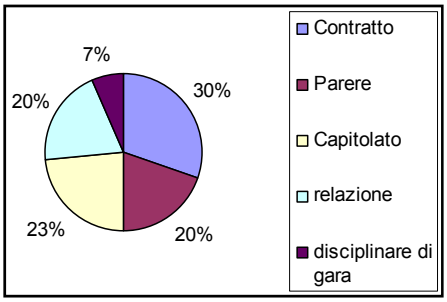


Figura 2 - Rilevanza delle tipologie documentali nel corpus

Terminologia e Documentazione: un rapporto bilaterale

Terminologia e Documentazione sono le due discipline portanti di questo lavoro. Lo scopo di entrambe è rappresentare e trasmettere la conoscenza specialistica appartenente ad un dominio ben definito. In questo caso la comunicazione tra gli attori del settore presuppone l'uso di un linguaggio, convenzionalmente condiviso, per mezzo del quale i termini veicolano concetti aventi un significato specifico, non ambiguo. «La terminologia, in quanto insieme di unità rappresentative della conoscenza specialistica, è necessaria per rappresentare e far comunicare i settori specialistici, e che ogni prassi che abbia relazione con la rappresentazione e/o con il trasferimento della conoscenza specialistica richiede in misura più o meno consistente la terminologia» (Cabré, 2001).

In particolare, la terminologia si occupa della descrizione e dell'analisi delle unità terminologiche, appartenenti al linguaggio naturale, capaci di rappresentare e trasmettere il campo semantico e concettuale del dominio al quale appartengono.

La documentazione, invece, focalizza il proprio interesse sull'informazione presente e organizzata nei documenti e sulle modalità di recupero della stessa attraverso strumenti quali i vocabolari di indicizzazione controllata ove l'astrazione della conoscenza avviene in maniera indiretta sfruttando unità terminologiche che assumono lo *status* di descrittore ovvero di unità di indicizzazione (Adelstein, Feliu, 2000). Queste ultime non appartengono più al linguaggio naturale ma diventano gli elementi sui quali si basa il linguaggio documentale per descrivere, sintetizzare ed estrarre le informazioni dai documenti. Di fatto, la documentazione non ha come oggetto di studio i termini in quanto tali ma solo in quanto costituenti voci indice e/o descrittori i quali, a loro volta, sono un "termine" con un valore aggiunto. Essi comprendono uno o più unità terminologiche che concorrono insieme a rappresentare il contenuto documentale di un più vasto dominio di conoscenza, sia esso specialistico o generale.

Dal nostro punto di vista, ciò che possiede in più un descrittore rispetto ad un "termine" è la possibilità di consentire la navigazione, per mezzo di *link* tra le unità terminologiche, all'interno del campo semantico del dominio specialistico che rappresentano.

La relazione esistente tra la terminologia e la documentazione è, come sottolinea Teresa Cabré (Cabré, 1998), bilaterale poiché mentre da una parte il lavoro del terminologo non può prescindere da quello del documentalista, dall'altra il lavoro di quest'ultimo, relativo alla descrizione del contenuto documentale, non può attuarsi senza il ricorso alla terminologia. Eppure non sempre tra i due àmbiti vi è la contaminazione culturale che sarebbe necessaria.

In sintesi, se lo scopo della documentazione è quello dell'analisi e dell'organizzazione del materiale documentale, l'indicizzazione e la classificazione sono le due tecniche utilizzate per rappresentare il contenuto concettuale di quanto analizzato.

Come risultato di queste attività, avremo un insieme di descrittori la cui funzione è quella di permettere il recupero dell'informazione. Essi devono essere considerati come etichette funzionali alla descrizione dell'informazione. Non devono essere pertanto confusi con i "termini" propriamente detti. Questa distinzione è necessaria in quanto in terminologia i termini sono le unità tipiche impiegate dagli esperti di un settore specialistico per comunicare tra di loro, e sono concepite come unità lessicali che rappresentano e permettono di trasferire la conoscenza specialistica dei settori scientifici tecnici (Cabré, 2001), mentre per i documentalisti i termini sono tanto unità suscettibili di diventare descrittori del contenuto documentale quanto elementi di controllo di un'attività classificatoria. Come già detto, le unità terminologiche perdono la loro funzione lessicale nel momento in cui diventano voci d'indice. Non appartenendo più al linguaggio naturale, in questo nuovo stato esse assumono la qualità tipica dei linguaggi documentali, espressa dalla funzione di meta-rappresentazione della conoscenza (Adelstein, Feliu, 2000).

In àmbito documentale, lo strumento utilizzato per il controllo terminologico è il *thesaurus*. La UNI ISO 5127-6:1988 *Documentazione e informazione - Vocabolario - Linguaggi documentari* lo definisce «vocabolario di indicizzazione controllato usato per tradurre il linguaggio naturale in un linguaggio formalizzato in modo tale da poter riconvertire tale linguaggio formalizzato in linguaggio naturale. Secondo la sua "struttura" un *thesaurus* è un vocabolario controllato e dinamico di termini correlati semanticamente e genericamente che copre un dominio specifico della conoscenza». Per la sua progettazione è fondamentale seguire una metodologia coerente sia relativamente alla selezione dei termini sia nella scelta della rete di relazioni tra i termini/descrittori inseriti nel vocabolario. Tale procedura è codificata, a livello internazionale, dalla norma UNI/ISO 2788: 1993, *Linee guida per la costruzione e lo sviluppo di thesauri monolingue*, la quale stabilisce che «il thesaurus è il vocabolario di un linguaggio di indicizzazione controllato, organizzato formalmente in maniera da rendere esplicite le relazioni "a priori" fra i concetti».

Nel 2005 è stata rilasciata la versione finale dello standard ANSI/NISO Z39.19-2005 *Guidelines for the construction, format, and management of monolingual controlled vocabularies* che del thesaurus formalizza la seguente definizione: «a controlled vocabulary arranged in a known order and structured so that the various relationships among terms are displayed clearly and identified by standardized relationship indicators». Lo standard ANSI fa un notevole passo avanti rispetto agli standard già in vigore poiché, come ben sottolinea il titolo, l'interesse ora è rivolto non solo verso i thesauri ma verso l'intera classe dei vocabolari controllati. Inoltre nelle linee guida dello standard è evidente l'attenzione ad adeguarsi ai cambiamenti avvenuti nel mondo dell'informazione e dei nuovi media documentali (Casson E. 2006).

Storicamente, è a cavallo tra 1950 ed il 1960 e con l'irruzione delle tecnologie digitali nei sistemi documentali che nascono i primi tentativi di indicizzazione controllata di banche dati bibliografiche. Bisognerà, però, aspettare l'inizio degli anni Novanta affinché gli estrattori automatici di termini comincino ad ottenere dei risultati positivi. Come ben sottolinea Rosa Estopà Bagot, nella sua tesi di dottorato (2002): *Extacción de terminología: elementos para la construcción de un SEACUSE: Sistema de extracción automatica de candidatos a unidad de significacción especializada,* l'aggettivo "automatici" riferito a tali strumenti non è del tutto pertinente. La definizione più appropriata dovrebbe essere estrattori semi automatici di termini. Questa specificazione riguarda il fatto che le "parole" estratte automaticamente, per mezzo di strumenti informatici, acquisiscono la dignità d'essere definite "descrittore" oppure "termine" solo in seguito ad un processo di validazione da parte di un esperto di dominio. Nel frattempo il loro *status* è quello di termine candidato.

Nel nostro caso applicativi, quindi, in un primo momento dalle UT verranno estratti i termini candidati i quali, ripetiamo, solo in séguito alla validazione da parte di un esperto di dominio, diventeranno veri e propri termini e/o descrittori (Estopà Bagot, 2002).

Riprendendo quanto già precedentemente affermato, non dobbiamo però confondere i termini specialistici con i descrittori documentali. Il documento AFNOR z47-100 stabilisce che i descrittori sono parole o gruppi di parole contenuti in un *thesaurus* e scelti tra un insieme di altri termini equivalenti al fine di rappresentare senza ambiguità una nozione contenuta in un documento, base di dati o in una domanda di ricerca. Pertanto, dal nostro punto di vista, un descrittore è un elemento appartenente ad un vocabolario d'indicizzazione controllato la cui forza ed il cui valore aggiunto è dato dall'insieme di relazioni semantiche che riesce ad innescare.

Estrazione terminologica: T2k

Il programma d'estrazione terminologica impiegato per la rilevazione dei termini candidati è *Text to Knowledge*, d'ora in avanti T2K. Il *tool* è stato sviluppato dall'Istituto di Linguistica Computazionale del Consiglio Nazionale delle Ricerche in sinergia con il dipartimento di Linguistica dell'Università di Pisa. La validità di questo strumento è confermata dal suo uso in differenti lavori di estrazione e rappresentazione della terminologia di àmbiti specialistici tra i quali citiamo: *Automatic extraction of semantics in law documents, (*Lenci A., *et.alii*), *NLP- based ontology lerning from legal texts. A case study* (Lenci A., Venturi G., et *alii, 2007*).

Una volta effettuata l'analisi linguistica dei testi, T2K acquisisce in maniera semiautomatica delle ontologie come supporto avanzato alla gestione documentale e genera,

inoltre, come *output* finale un vocabolario terminologico il cui valore aggiunto è rappresentato dalle informazioni semantico-concettuali dei termini del glossario stesso, i quali concorreranno a formare le entrate del nostro *thesaurus*. Queste ultime, come evidenziato nell'esempio in Tabella 3 che segue, sono strutturate attraverso relazioni gerarchiche di iponimia/iperonimia ricostruite a partire dalla struttura linguistica interna dei termini, ovvero dalla condivisione della medesima testa lessicale, dei modificatori, ecc., così come di seguito evidenziato:

AREA	AREA APPLICATIVA
AREA	AREA COMUNICAZIONE
AREA	AREA DIREZIONALE
AREA	AREA DI GOVERNO AMMINISTRATIVO
AREA	AREA PATRIMONIALE

Tabella 3

Il sistema d'analisi computazionale adottato da T2k è AnIta. Questa piattaforma è usata per il trattamento automatico della lingua italiana. Attraverso l'analisi linguistica dei testi essa fornisce una rappresentazione avanzata del contenuto informativo dei testi esaminati (Montemagni, 1996).

In particolare, T2k, integra sistemi di analisi linguistica automatica del testo, (di cui in tabella 4 segue un esempio relativo alla frase "Capitolato tecnico per la riorganizzazione dei servizi"), con algoritmi stocastici di identificazione e *clustering* terminologico e concettuale, strumenti di annotazione contenutistica o *knowledge mark-up* del testo e, infine, dati strutturati di supporto all'indicizzazione terminologico-concettuale di documenti.

[[CC: N_C] [AGR: @MS] [POTGOV: CAPITOLATO#S@MS]]	CAPITOLATO
[[CC: NA_C] [AGR: @MS-@MS] [POTGOV: TECNICO#A@MS TECNICO#S@MS]]	TECNICO
[[CC: P_C] [PREP: PER#E]	PER
[DET: LO#RD@FS] [AGR: @FS]	LA
[POTGOV: RIORGANIZZAZIONE#S@FS]]	RIORGANIZZAZIONE
[[CC: di_C] [DET: IL#RD@MP] [AGR: @MP]	DEI
[POTGOV: SERVIZIO#S@MP]]	SERVIZI

Tabella 4 - Esempio di testo chunkato

Quanto appena detto è schematizzato nella figura 3.

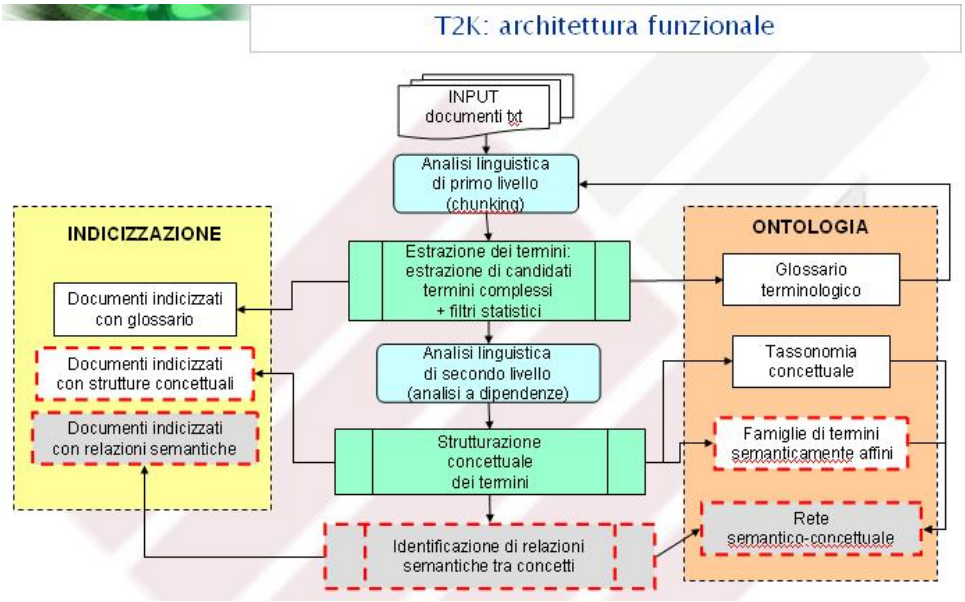


Figura 3 - Architettura funzionale di t2k

In sintesi, lo schema in figura 3, descrive le differenti fasi dell'estrazione terminologica condotte da T2k alternando processi di analisi linguistica e di analisi statistica. Precisiamo soltanto che l'estrazione inizia con l'inserimento dei documenti in formato TXT. L'analisi è effettuata, innanzi tutto, con un sistema di NLP (*natural language processing)*. Segue la suddivisione del testo in *chunk* (CHUG-IT, Federici *et. alii,* 1996) ovvero in unità testuali etichettate a livello morfo-sintattico (vedi esempio di testo chunkato proposto in Tabella 4). I termini candidati rilevati da T2K possono essere tanto unità lessicali monorematiche quanto unità lessicali polirematiche e rappresentano il repertorio terminologico del dominio indagato. In generale, i termini del glossario terminologico acquisito con T2K possono essere a loro volta raggruppati secondo diverse relazioni lessicali. Ad esempio, il termine ATTIVITÀ è relazionato con: ATTIVITÀ CONTRATTUALI, ATTIVITÀ DI DIGITALIZZAZIONE, ATTIVITÀ DI MONITORAGGIO, ATTIVITÀ DI MONITORAGGIO AGGIORNATO, ATTIVITÀ PREVISTE ATTIVITÀ REDAZIONALE, ATTIVITÀ DI VERIFICA. Tutti questi condividono il concetto più generale di "attività" a cui possono essere ricondotti per astrazione iperonimica. La strutturazione concettuale operata da T2K non è, tuttavia, circoscritta alle sole relazioni gerarchiche di iperonimia/iponimia appena esemplificate. Di fatto in *output* avremo un'ontologia di dominio composta, in particolare, da un glossario terminologico delle famiglie di termini semanticamente affini e dalla rete

semantico-concettuale dei termini. Ai fini del nostro lavoro, queste fasi sono ulteriormente interessanti poiché, T2k, una volta effettuata l'identificazione delle relazioni semantiche tra i concetti, costruisce un ponte tra i documenti indicizzati con l'indicazione delle relazioni semantiche e la rete semantico-concettuale dell'ontologia. Questo ci consente, in ogni momento, di risalire dal termine al documento d'origine e, nel medesimo tempo, di identificare i descrittori. Nel nostro caso, assumiamo come descrittori le entrate monolessicali del *thesaurus* aventi il valore tf*idf più alto dell'unità informativa presa in esame. Essi sono capaci di rappresentare il contenuto informativo del nostro *corpus* documentale. Di fatto, in questa fase di estrazione è già possibile condurre un'indicizzazione terminologica dei documenti affinché, come appena detto, ogni termine possa essere ricondotto al suo documento originario. In Figura 4 riportiamo un esempio:

did	termine	funzione	ird	irg
1	AGENZIA DELLE DOGANE	13.02252992938036157986	84	84
1	MONITORE	11.4951218203758358527	312	313
1	CONTRATTO DI MONITORAGGIO	10.7821542257426017388	33	33
1	SERVIZI DI DIREZIONE	10.6322369572833945736	31	31
1	RENDICONTO SULL' INTERVENTO	10.5536103961402503159	30	30
1	PERSONALE DELL' AGENZIA	10.2104695596408188151	26	26

Figura 4

A questo punto, dobbiamo precisare che il valore riportato nella colonna "funzione" indica la misura adottata per calcolare la rilevanza dei termini all'interno della collezione documentale. La funzione utilizzata è la tf*idf (*term frequency * inverse document frequency.* Essa calcola la frequenza di ogni "termine" all'interno del documento di riferimento (*TF = term frequency*), relazionata con la frequenza del termine stesso all'interno del complessivo *corpus* documentale. Maggiore è la frequenza del termine, minore sarà la sua significatività all'interno dell'intero *corpus* e, di conseguenza, minore sarà il valore tf*idf assegnato al candidato termine. Conseguentemente, i candidati a termine estratti da T2k con il valore tf*idf più alto, appartenenti ad ogni singola UT, saranno i nostri candidati a descrittore.

Conclusione

Il progetto è ancora in fase di sviluppo per cui sono stati trattati – nei limiti concessi – solo alcuni aspetti concettuali propedeutici. Di fatto, il *corpus* analizzato appartenendo al settore ICT è composto da lessico sia italiano sia inglese. Attualmente, è stato solo analizzato il linguaggio italiano applicando al *corpus* differenti soglie di occorrenze per

l'estrazione dei termini candidati. Questi ultimi sono stati ritenuti tali se nel *corpus* si ripetevano almeno tre volte. Al fine di valutare la precisione di estrazione automatica dei termini candidati da parte del *software*, verranno effettuati più *run* con differenti valori da applicare alle soglie di estrazione.

In un'ulteriore fase del progetto, la collezione documentale iniziale verrà arricchita da ulteriori UT, fornite dal CNIPA, il quale a sua volta avrà anche il compito di validare la terminologia prodotta automaticamente dai *software* di estrazione. L'operazione di estrazione terminologica verrà effettuata, inoltre, utilizzando anche altri *tools* – i cui risultati verranno confrontati con quelli prodotti da t2*k.* In particolare, TaLTaC2 - *Trattamento Automatico Lessicale e Testuale per l'Analisi del Contenuto di un Corpus* – software sviluppato dall'Università degli Studi di Salerno e di Roma "La Sapienza" e *Terminus*, programma per la gestione terminologica e per la costruzione di *corpora* documentali sviluppato dal gruppo IULATERM dell'Università Pompeu Fabra di Barcellona.

Abbiamo cominciato questa presentazione chiedendoci quali sono le unità informative e di conoscenza, ovvero quali sono i termini che dobbiamo rilevare al fine di meglio rappresentare un àmbito specialistico. In prosieguo abbiamo chiarito che ai fini del presente lavoro consideriamo come Unità Informativa l'insieme dei documenti che concorrono a formare tutto l'*iter* contrattuale seguito dalle differenti tipologie documentali relative al medesimo affare. Abbiamo anche affermato che in questo lavoro i descrittori sono i termini, preferibilmente monolessicali, che compaiono nel *thesaurus* costruito da T2k ordinati secondo il valore di rilevanza nel *corpus*, calcolato con la misura tf*idf . La nostra intenzione, a questo punto, è quella di utilizzare, oltre alla misura appena menzionata, altri algoritmi di estrazione che ci consentano di ottenere un ulteriore livello concettuale del nostro vocabolario di indicizzazione controllato. Attualmente, la nostra scelta è rivolta verso l'algoritmo C-value, precedentemente usato in un lavoro analogo (Vuono, 2007) il cui utilizzo sarà rivolto principalmente all'estrazione dei termini composti.

Il descrittore ed il termine appaiono, quindi, equivalenti dal punto di vista della struttura lessicale ma divergono nella loro strutturazione funzionale e nel valore che viene loro attribuito dalla comunità degli utilizzatori. Terminologi, documentalisti ma anche specialisti dell'ICT, a volte, usano in maniera impropria e indistinta termini e descrittori senza curarsi della loro similarità ma anche della loro sostanziale diversità concettuale. Ciò è causa di non poche criticità nel complesso processo di recupero dell'informazione documentale.

Bibliografia

Adelstein A., Feliu J., *Relations semàntiques entre unitati lèxiques amb valor especializat i descriptores*, 2001, in CABRÈ M.T *et al. Terminologia i Documentacio'. I Jornada de terminologia i documentacio',* Barcellona: Istituto di Linguistica Applicata - IULA, Università Pompeu Fabra, 2000, pp. 121-131.

Biagioli C., Francesconi E., Montemagni S., Passerini A., Soria C., (2005), *Automatic semantics extraction in law documents,* in atti dell'International Conference of Artificial Intelligence and Law, ICAIL 2005 (Bologna, 6-11 giugno 2005).

Cabré M.T, *La terminologia tra lessicologia e documentazione,* Barcellona: Istituto di Linguistica Applicata - IULA, Università Pompeu Fabra, 2001.

Cabré M.T, *et al. Terminologia i Documentacio'. I Jornada de terminologia i documentacio',* Barcellona: Istituto di Linguistica Applicata - IULA, Università Pompeu Fabra, 2000.

Cabré M.T, *Terminologia y Documentacion,* in Gonzalo c., garcia v *Documentacion, Terminologia y Traduccion,* Madrid, Sintesis, Fondazione Duques de Soria, pp. 31-43, 2000.

Casson E., *Dai thesauri ai vocabolari controllati:alcune novità introdotte nell'ultima edizione dello standard ANSI/NISO Z39.19-2005,* in «AIDAinformazioni», Roma, a. 24, n. 1-2, pp. 69-77 2006.

Chaudiron S., *Tecnologies languistques et modes de représentation de l'information textuelle*, in «Documentaliste - Sciences de l'information», ADBS, Parigi, vol. 44/01, pp. 30-39, 2007.

Estopà Bagot R., *Extracción de terminología: elementos para la costrucción de un SEACUSE: Sistema d'extracción automatíca de a unidades de candidatos de significación especializada,* Tesi di dottorato, IULA, Università Pompeu Fabra, Barcellona, 2002.

Federici S., Montemagni S., Pirrelli V., *Shallow Parsing and Text Chunking: a view on underspecification in syntax.* in *Proceedings of the workshop on robust parsing,* Praga, 1996.

Gobber G., *Breve nota sulla doppia natura linguistica e logico-semantica dei termini nelle scienze,* in atti del convegno Terminologie specialistiche e tipologie testuali (Milano, 26-27 maggio 2006), a cura di M. T. Zangola, Milano, ISU, 2007, p. 31.

Lenci A., Montemagni S., Pirrelli V., Venturi G., *NLP-based ontology from legal texts. A case study.* In Atti del LOAIT 2007.

Lenci A., Montemagni S., Pirrelli V., *Testo e computer. Elementi di linguistica computazionale,* Roma, Carocci, p.124.

Menon B., (2007), *Les langages documentaires: un panorama, quelques remarques critiques et un assai de bilan,* in «Documentaliste - Sciences de l'information», ADBS, Parigi, vol. 44/01, pp. 18-28.

Montemagni S., (1996), *Architecture and functioning of a System for the acquisition of Taxonomical information from Dictionary Definitions.* In atti della 4° conferenza su *Computational lexicography and text research* (COMPLEX, 1996), Budapest.

Vuono S., (2007), *Estrazione terminologica e Text Clustering: applicazione di C-Value a un corpus testuale*, tesi di dottorato, Università della Calabria, Cosenza.

Forma e contenuto nella terminologia della gestione documentale: ipotesi per la costruzione di un glossario specialistico

PIERA BELCASTRO

During the creation of a record management glossary, terms and their relations were studied. This led to certain observations on their morphology and significance starting from de Saussure's idea of regarding language as a "complex equilibrium of terms influencing each other".

Keywords: glossary – record management – terminology

1. Introduzione

La gestione documentale, come ogni altro dominio, fa uso e necessita di un linguaggio "speciale", che determini inequivocabilmente l'ambito, la tipologia di utenti e la tipologia di situazioni nelle quali prende vita il processo comunicativo (Cabrè, 1999) [1]. La specificità dei termini e la corrispondenza univoca di un termine ad un dato concetto sono, infatti, garanzia di comprensibilità, all'interno della comunità linguistica che adotta un determinato linguaggio. La definizione e l'uso di una terminologia specifica ben definita, oltre ad essere strumento di comunicazione in grado di evitare confusione ed incomprensioni agli addetti ai lavori diventa anche un importante ed efficace mezzo di conoscenza per coloro che pur non appartenendo ad un specifico ambito disciplinare si confrontano con il linguaggio "speciale" che ad esso appartiene. I linguaggi "speciali" evolvono, infatti, parallelamente al linguaggio naturale, ma con ritmi più accelerati; ne consegue che i loro neologismi tecnici vengono spesso assimilati nel linguaggio comune senza adeguata consapevolezza della loro reale portata semantica.

In tale contesto l'evoluzione, la formalizzazione e la diffusione del linguaggio speciale ha seguito un percorso arduo la cui complicazione è stata determinata dai cambiamenti cui la disciplina archivistica ed, in particolare, quella specifica accezione che è la gestione documentale, è stata ed è tuttora sottoposta, nonché dal ruolo che essa ha assunto nella contemporanea società dell'informazione. In campo strettamente archivistico, il lavoro costante di studiosi ed esperti, che negli anni ne hanno seguito l'evoluzione preoccupandosi di dare una determinazione dei suoi principi attraverso la definizione della sua terminologia, ha determinato una armonizzazione del lessico e, quindi, un'univocità semantica per ognuno dei termini che ne fanno parte. Ampio riscontro del

raggiungimento di tale obiettivo è dato dal vasto utilizzo di tale terminologia all'interno del cospicuo numero di manuali di riferimento nonché dall'esistenza di un dizionario (Walne a cura di, 1988) [2] e un glossario (Nogueira, a cura di, 1988) [3] che, però, oltre ad essere ormai datati [4], presentano in lingua italiana solo una lista di termini a cui non è associata nessuna definizione o riferimento applicativo.

Per quanto attiene la gestione documentale [5] il discorso assume connotati differenti. In questo ambito, infatti, le continue evoluzioni tecnologiche ed il proliferare di nuove normative hanno causato un'assoluta assenza di armonizzazione terminologica e l'introduzione di numerosi neologismi (Adamo e Della Valle, 2005) [6]. I termini, quindi, assumendo spesso una valenza semantica diversa, vengono definiti in maniera differente comportando, così, ambiguità di uso e di interpretazione. Tale situazione assume particolare rilievo nei testi delle norme nei quali definizioni diverse di uno stesso termine o definizioni errate possono, a volte, causare confusione nell'applicazione e nell'interpretazione delle norme stesse. Al riguardo si posso elencare diverse esemplificazioni presenti all'interno del Decreto Legislativo 7 marzo 2005, n. 82 - *Codice dell'Amministrazione Digitale* dove, ad esempio, dall'affermazione «…i documenti degli archivi, le scritture contabili, la corrispondenza ed ogni atto o documento di cui è prescritta la conservazione…» (l'art 43, comma 1) si desume l'errato concetto che le scritture contabili e la corrispondenza non sono documenti di archivio. Oppure quando, all'articolo 47, si parla di «comunicazione di documenti», riferendosi erroneamente alla loro trasmissione o invio, e di «protocollo informatizzato» anziché di protocollo informatico (Giuva, 2005) [7]. Così come la definizione di "originale non unico", ovvero «i documenti per i quali sia possibile risalire al loro contenuto attraverso altre scritture o documenti di cui sia obbligatoria la conservazione, anche se in possesso di altri», risulta inconsistente e contestabile sia dal punto di vista archivistico che da quello giuridico (Giuva, 2005) [8].

Del tutto diverso è lo stato dell'arte a livello internazionale. Si pensi, infatti, che già nel 1996 in Australia è stato pubblicato il primo *standard* per la gestione documentale (AS 4390, 1996) e che, tra il 2001 e il 2004, l'International *standard*isation Organisation (ISO) ha pubblicato in questo ambito altri due importanti *standard*: ISO 15489/2001, ISO TS 23081-1/2006 e ISO TS 23081-2/2007 [9]. Ciò testimonia il maturare di una disciplina al cui interno vanno sviluppandosi nuove professionalità (Hofman, 2005) [10]. In particolare, grande attenzione all'evoluzione ed alla normalizzazione del linguaggio, sia in campo archivistico che in quello di gestione documentale viene rivolta nel mondo anglosassone [11] soprattutto ad opera delle associazioni professionali. È di recente pubblicazione, infatti, a cura della Society of American Archivists, *A Glossary of Archival and Records Terminology* (Pearce-Moses, 2005) [12], il cui scopo è quello di «provide the basic foundation for modern archival practise and theory» (Pearce-Moses ,2005) [13], dal momento che «the archival world has changed

considerably …» e «a rich monographic research literature is developing» (Pearce-Moses, 2005) [14].

2. Costruzione, analisi ed esplorazione del *corpus*: forma e contenuto dei termini

Dal punto di vista metodologico l'attività terminologica, svolta per la costruzione del glossario, è basata sull'approccio semasiologico testuale e quindi sulla *linguistica dei corpora*. Innanzitutto, nella costruzione di un *corpus* è necessario individuare i destinatari del lavoro terminologico per poter effettuare una scelta della tipologia testuale da includere. In questo caso gli utenti del glossario sono rappresentati da coloro che operano in questo settore ma anche da chi, pur non essendo un esperto, si trova nelle condizioni di doversi relazionare con tale terminologia: si pensi ad esempio agli organi preposti a legiferare in materia di gestione documentale, agli organismi di *standard*izzazione o a tutti coloro che necessitano di tale strumento per effettuare traduzioni in e dall'italiano di testi specialistici del dominio in oggetto. In virtù della tipologia di utenti e considerato che le norme e gli *standard* costituiscono il primo livello di approccio ad un dominio, il *corpus* è stato così costituito:

1. UNI ISO 5127:1987 *Documentazione e informazione. Vocabolario*
2. UNI ISO 5963:1989 *Documentazione. Metodi per l'analisi dei documenti, la determinazione del loro soggetto e la selezione dei termini di indicizzazione*
3. Decreto del Presidente della Repubblica 28 dicembre 2000, n. 445 - *Testo unico delle disposizioni legislative e regolamentari in materia di documentazione amministrativa*
4. UNI ISO 15489-1:2001 *Information and documentation - Gestione dei documenti di archivio - parte 1 - Principi generali*
5. UNI ISO 15489-2:2001 *Information and documentation - Gestione dei documenti di archivio - parte 2 - Linee guida*
6. Decreto del Presidente del Consiglio dei Ministri 13 gennaio 2004 - *Regole tecniche per la formazione, la trasmissione, la conservazione, la duplicazione, la riproduzione e la validazione, anche temporale, dei documenti informatici*
7. Deliberazione Cnipa 19 febbraio 2004, n. 11 - *Regole tecniche per la riproduzione e conservazione di documenti su supporto ottico idoneo a garantire la conformità dei documenti agli originali*
8. Deliberazione Cnipa 17 febbraio 2005, n. 4 - *Regole per il riconoscimento e la verifica del documento informatico*
9. Decreto Legislativo 7 marzo 2005, n. 82 - *Codice dell'Amministrazione Digitale*
10. Decreto del Presidente della Repubblica 11 febbraio 2005, n. 68 - *Regolamento recante disposizioni per l'utilizzo della posta elettronica certificata, a norma dell'articolo 27 della legge 16 gennaio 2003, n. 3.*

Il *corpus* è pertanto composto per il 33% da *standard*, per il 22% da deliberazioni, per il 45% da leggi (Figura 1).

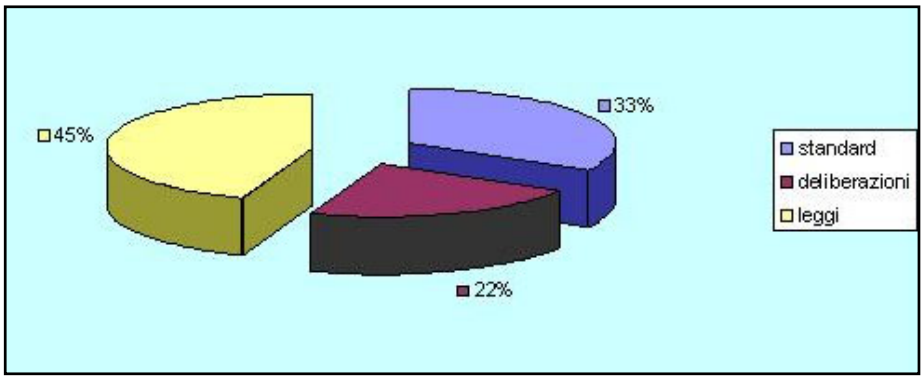


Figura 1 - Composizione percentuale del corpus

Si è quindi proceduto all'analisi ed esplorazione del *corpus* mediante un'estrazione terminologica a *cluster* composti da più lemmi utilizzando il *software* Oxford Wordsmith Tools Versione 5.0 (Scott, 2008) [15]. La scelta della estrazione terminologica a più lemmi nei *cluster* è stata dettata dalla consapevolezza che la tipologia terminologica di tale dominio, ma anche la struttura della tipologia documentaria utilizzata per la costruzione del *corpus*, è costituita da termini semplici (documento), termini complessi (documento informatico) e fraseologia (registrare documenti) (ISO 1087-1:2000) [16]. In particolare nella lista dei candidati a termini, che totalizza 2127 elementi, si riscontrano 656 termini semplici, 1425 termini complessi con numero di lemmi variabile e 46 esempi di fraseologia. La composizione del *corpus* così ottenuta è riportata nella tabella sottostante (Figura 2).

Cluster	1	2	3	4	5	6
Text File	CORPUS	CORPUS	CORPUS	CORPUS	CORPUS	CORPUS
Bytes	885.168	885.168	885.168	885.168	885.168	885.168
Tokens	130.303	130.303	130.303	130.303	130.303	130.303
Types	5.407	12.481	10.636	7.496	5.492	4.197
Type/Token Ratio	4,15	9,58	8,16	5,75	4,21	3,22
Standardised Type/Token	39,41	74,96	86,90	91,61	94,56	95,88

Figura 2 - L'estensione del *corpus*

L'andamento della *type/token ratio* all'aumentare del numero di lemmi presenti nei *cluster*, riportato nella Figura 3, mostra chiaramente come il numero dei *type* sia più elevato nelle *wordlist* di bigrammi, trigrammi e quadrigrammi, con un valore partico-

larmente rilevante per i bigrammi. Con riguardo, invece, a monogrammi e pentagrammi i valori sono simili e non particolarmente elevati, mentre il valore della *wordlist* a sei lemmi è inferiore pur non discostandosi molto da quello dei monogrammi. Tale analisi conferma la tipologia terminologica del dominio dando valore alla metodologia applicata di estrazione terminologica a *cluster* composti da più lemmi.

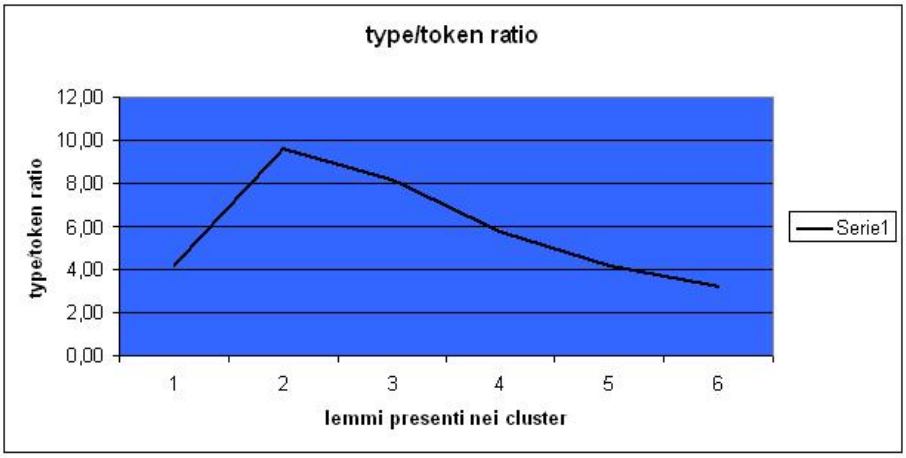


Figura 3 - Type/token ratio in funzione del numero di lemmi presenti nei cluster

Dall'analisi delle liste dei candidati termini, necessaria alla costruzione degli alberi e alla definizione dell'intero sistema concettuale del dominio, sono scaturite delle osservazioni in merito alla relazione tra forma e contenuto dei termini che richiamano la metafora degli scacchi, utilizzata da Saussure per spiegare come la lingua sia «un equilibrio complesso di termini che si condizionano reciprocamente (…), ovvero un sistema nel quale ciascuno dei suoi elementi ha un determinato valore solo in rapporto agli altri elementi che ne fanno parte» (de Saussure, 1965) [17]. Come negli scacchi dove il valore di ogni singolo pezzo non è determinato dalle sostanze di cui essi sono composti, ma esclusivamente dal confronto della funzione del singolo rispetto a quella degli altri, anche nella lingua è la posizione che i termini occupano e le relazioni (gerarchiche, sinonimiche, di opposizione) che contraggono ad attribuire loro un valore. Questa affermazione trova riscontro nella rappresentazione e nell'analisi dei termini del dominio esaminato. In quasi tutti gli alberi concettuali realizzati fino a questo momento, infatti, il termine semplice non ha alcuna rilevanza o valore particolare all'interno del dominio, mentre assume il suo significato specifico ed il suo ruolo all'interno del sistema concettuale se associato ad un altro termine o se letto nel contesto applicativo di riferimento. Inoltre, in alcuni casi, ogni bigramma o trigramma costituisce un

approfondimento o una specifica del termine che gerarchicamente lo precede; altre volte, invece, esso determina il concetto o la definizione a cui il termine fa riferimento. Infine, vi sono casi in cui i bigrammi o trigrammi connotano significati totalmente differenti. Questa relazione forma-contenuto è stata osservata, in particolare, negli alberi concettuali già realizzati di *firma*, *registrazione* e *classificazione* (Figure 4, 5, 6).

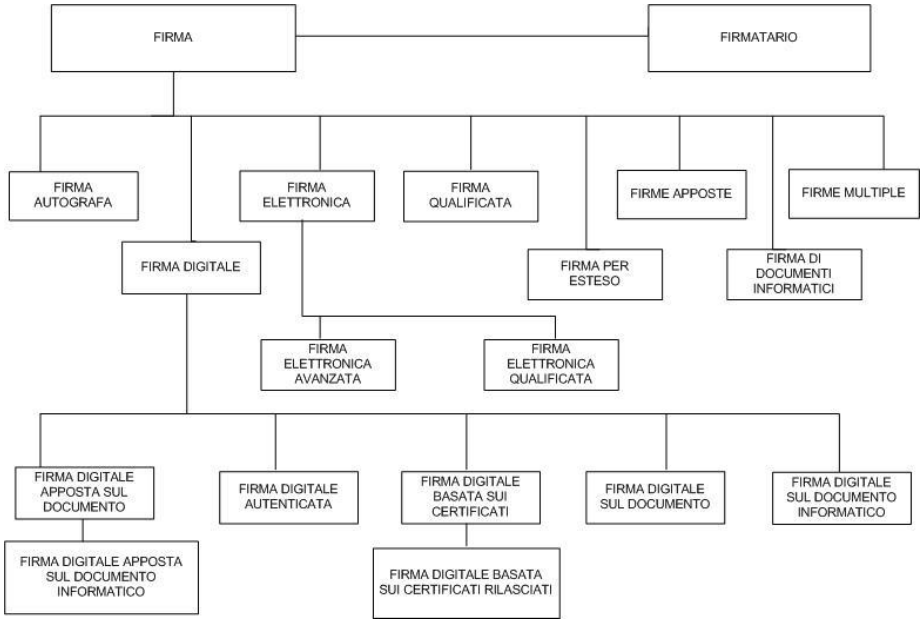


Figura 4 - Albero Concettuale. Firma

Nel caso di *firma*, sebbene il temine occorra ben 247 volte all'interno del *corpus*, esso non assume la medesima rilevanza nel dominio come, invece, accade per *firma elettronica* (39 occorrenze) e *firma digitale* (86 occorrenze). Da una analisi più dettagliata delle concordanze del termine, si evidenzia che esso occorre da solo poche volte mentre si trova associato, nella maggioranza dei casi, a *digitale* ed *elettronico*. È proprio in associazione con questi due elementi che *firma* acquisisce il ruolo di termine all'interno del dominio della gestione documentale. Si consideri che sebbene i due bigrammi possano apparire sinonimi, essi connotano invece, concetti differenti. Con il termine complesso *firma digitale* si indica «un particolare tipo di firma elettronica qualificata basata su un sistema di chiavi crittografiche, una pubblica e una privata, correlate tra loro, che consente al titolare tramite la chiave privata e al destinatario tramite la chiave pubblica,

rispettivamente, di rendere manifesta e di verificare la provenienza e l'integrità di un documento informatico o di un insieme di documenti informatici» (D. Lgs n. 82, 2005) [18], mentre con *firma elettronica* si intende «l'insieme dei dati in forma elettronica, allegati oppure connessi tramite associazione logica ad altri dati elettronici, utilizzati come metodo di autenticazione informatica» (D. Lgs n. 82, 2005) [19]. Esaminando in particolare la suddivisione di quest'ultimo termine troviamo che *firma elettronica qualificata* è «la firma elettronica ottenuta attraverso una procedura informatica che garantisce la connessione univoca al firmatario e la sua univoca autenticazione informatica, creata con mezzi sui quali il firmatario può conservare un controllo esclusivo e collegata ai dati ai quali si riferisce in modo da consentire di rilevare se i dati stessi siano stati successivamente modificati, che sia basata su un certificato qualificato e realizzata mediante un dispositivo sicuro per la creazione della firma, quale l'apparato strumentale usato per la creazione della firma elettronica» (D. Lgs n. 82, 2005) [20]. Inoltre, osservando l'albero concettuale, si può notare come fino al terzo livello i termini costituisco delle specifiche dell'iperonimo. Dal quarto livello in giù i termini, lasciano spazio a concetti che rappresentano, in parte, le definizioni dei loro iperonimi.

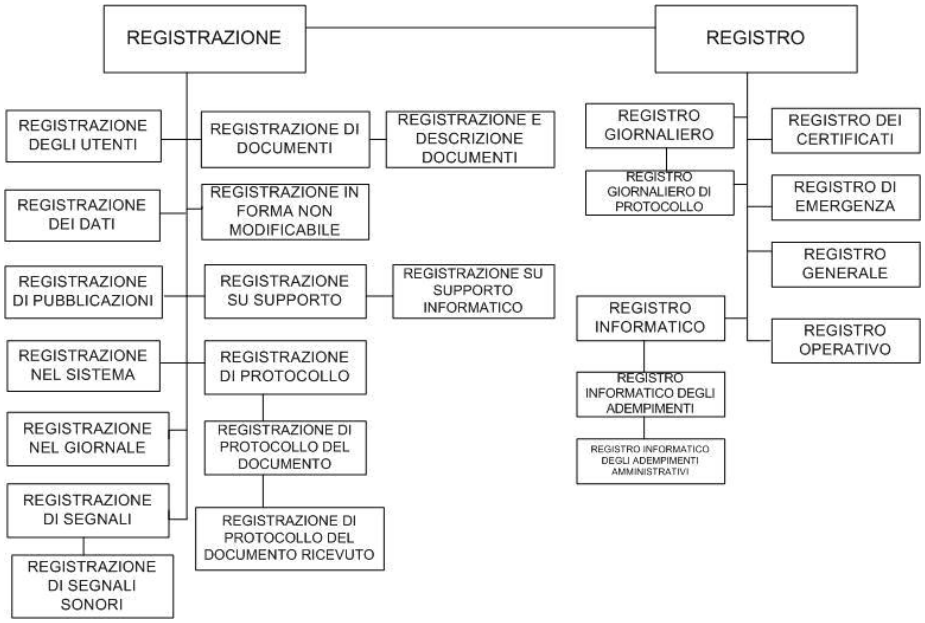


Figura 5 - Albero Concettuale, Registrazione

La Figura 5 evidenza come l'albero concettuale di *registrazione* sia costituito da due iperonimi, rappresentanti uno il processo e l'altro lo strumento utilizzato. L'uso comune dei termini, tuttavia, comporta assai spesso una sovrapposizione di piani identificando, così, il processo con lo strumento.

Inoltre, è possibile notare come, al contrario di quello che accade per *firma*, il termine *registrazione* assume rilevanza da solo nel dominio se viene considerato nel suo significato tanto in senso generale che in quello più strettamente archivistico. Esso infatti può essere definito, nel primo caso, come una «attività compiuta da uffici rientranti nell'organizzazione finanziaria dello Stato (Uff. del Registro) e volta ad uno scopo probatorio e fiscale. Sono soggetti a registrazione tutti gli atti compiuti nel territorio dello Stato sia in forma pubblica che privata. Effetto della registrazione è quello di attestare l'esistenza dell'atto, nonché di stabilirne la data certa» <62.149.227.181/glossarioR.php> [21] e, nella seconda accezione, come «iscrizione nell'inventario di un documento destinato ad essere conservato» (UNI-ISO 5127/05, 1987) [22]. Tale rilevanza all'interno del dominio trova riscontro nell'elevata occorrenza che il termine assume all'interno del *corpus* (116 occorrenze) e nell'analisi delle concordanze che lo presentano quasi sempre non associato ad altri termini.

Dal punto di vista del significato che il termine *registrazione* assume in associazione con altri termini, sarà possibile osservare come i nuovi vocaboli, pur facendo riferimento sempre alla registrazione di documenti, costituiscono una specifica maggiore del termine generico. Ad esempio, *registrazione di protocollo* si riferisce alla registrazione di ogni «documento ricevuto o spedito dalle pubbliche amministrazioni ed effettuata mediante la memorizzazione delle seguenti informazioni: a) numero di protocollo del documento generato automaticamente dal sistema e registrato in forma non modificabile; b) data di registrazione di protocollo assegnata automaticamente dal sistema e registrata in forma non modificabile; c) mittente per i documenti ricevuti o, in alternativa, il destinatario o i destinatari per i documenti spediti, registrati in forma non modificabile; d) oggetto del documento, registrato in forma non modificabile; e) data e protocollo del documento ricevuto, se disponibili; f) l'impronta del documento informatico, se trasmesso per via telematica, costituita dalla sequenza di simboli binari in grado di identificarne univocamente il contenuto, registrata in forma non modificabile» (DPR 445, 2000) [23]. La *registrazione di pubblicazioni in serie* viene intesa come «registrazione regolare dei singoli fascicoli di una pubblicazione in serie nel momento in cui vengono ricevuti, così da rendere possibile, in qualsiasi momento la verifica della consistenza» (UNI-ISO 5127/05, 1987) [24]. Si è notato che in alcuni casi, però, l'associazione con altri termini come, ad esempio, nel caso di *registrazione di segnali sonori*, si fa riferimento ad un'altra tipologia di registrazione, ad una differente tipologia di supporto ed ad un diverso oggetto di tale attività.

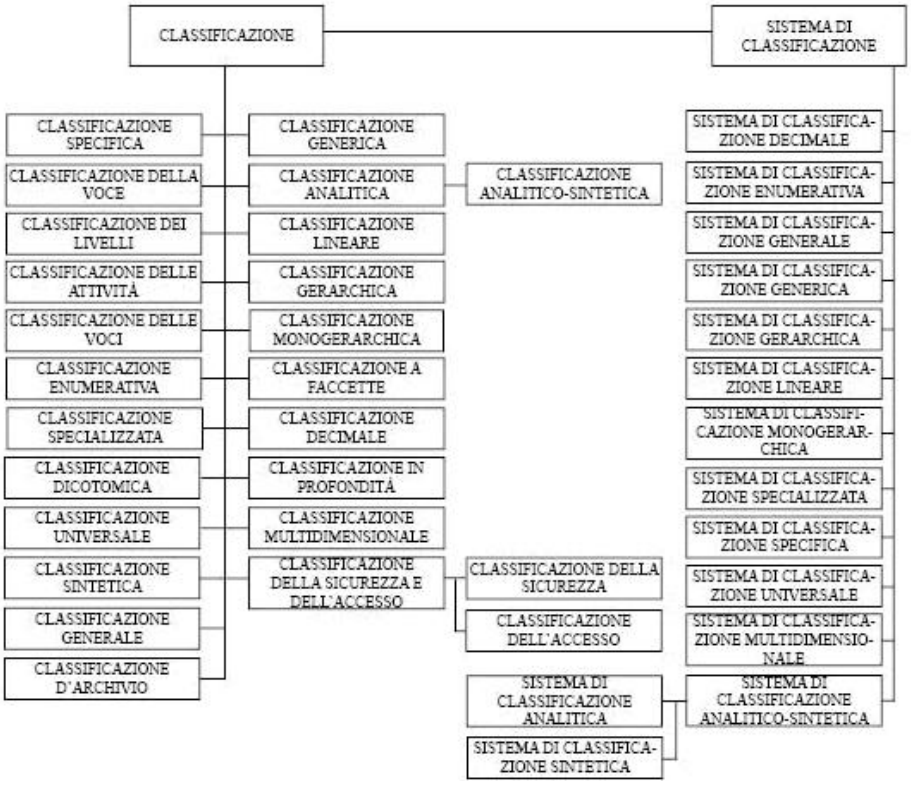


Figura 6 - Albero Concettuale. Classificazione

Anche l'albero concettuale del termine *classificazione* presenta due iperonimi, rappresentanti uno l'attività e l'altro lo strumento e, anche in questo caso, esiste una sovrapposizione dei piani concettuali che porta, spesso, erroneamente all'identificazione dei due termini.

Come per *registrazione* il termine assume, da solo, rilevanza all'interno del *corpus* così come dimostrato dal numero di occorrenze (308) e dalle concordanze effettuate. Esso, nella sua accezione singola, viene inteso come «ordinamento di concetti in classi e loro suddivisioni per esprimere le relazioni semantiche tra loro esistenti; le classi sono rappresentate da una notazione» (UNI-ISO 5127/06, 1987) [25], mentre la sua associazione con altri termini produce un cambiamento non tanto del significato intrinseco del vocabolo ma della metodologia sottesa a tale procedimento. Possiamo pertanto distinguere tra:

- *classificazione analitica* basata sulle relazioni formali fisse tra le classi (UNI-ISO 5127/06, 1987) [26],

- *classificazione sintetica* in cui le relazioni formali sono stabilite durante la classificazione (UNI-ISO 5127/06, 1987) [27],
- *classificazione analitico-sintetica* ovvero un sistema di classificazione la cui applicazione caratteristica consiste nell'analisi di soggetti composti o complessi seguita dalla loro espressione mediante un simbolo di classe costruito per sintesi (UNI-ISO 5127/06, 1987) [28],
- *classificazione dicotomica* in cui ogni classe può essere suddivisa in due classi subordinate (UNI-ISO 5127/06, 1987) [29],
- *classificazione enumerativa* in cui ogni classe è indicata come voce principale (UNI-ISO 5127/06, 1987) [30],
- *classificazione decimale* è quella in cui viene usata una notazione decimale (UNI-ISO 5127/06, 1987) [31].

3. Conclusioni

La metodologia utilizzata e lo studio dei termini singoli e delle loro relazioni, ha portato ad analizzarne forma e contenuto. Da questa analisi si è rilevato come i termini si influenzano generando significati e definizioni non contemplati nell'uso e nelle analisi dei singoli termini e che, inoltre, l'associazione di termini e le relazioni che intercorrono tra di loro concorrono a determinare concetti diversi in relazione anche al contesto d'uso in cui questa relazione va ad agire.

Note

[1] Cabrè M.T. (1999), *Terminology: Theory, methods and applications,* John Benjamins B.V., Amsterdam/Philadelphia, p. 65.

[2] Walne P., (a cura di) (1988), *Dictionary of Archival Terminology: English and French with equivalents in Dutch, German, Italian, Russian and Spanish*, ICA Handbooks Series, Volume 3, München; New York; London; Paris: Saur.

[3] Nogueira C.C. (a cura di) (1988), *Glossary of basic archival and library conservation terms: English with equivalents Spanish, German, Italian, French and Russian*, ICA Handbooks, Volume 4, München; New York; London; Paris: Saur.

[4] Si noti che una precedente edizione fu pubblicata nel 1984, in Dryden, J. (2005), *A Tower of Babel: standardizing Archival Terminology*, "Archival Science", 5, p. 8.

[5] Convenzionalmente intendiamo con il termine l'equivalente italiano del francese Gestion Electonique des documents e quindi tutto quel complesso di attività legate all'applicazione di tecnologie informatiche alla gestione della documentazione corrente.

[6] I neologismi possono «…avere origine da parole già in uso (…) o essere prelevati dal lessico di altre lingue, nella loro forma originaria o in una adattata», o essere considerati come «… significato nuovo attribuito a parole già esistenti …» in Adamo, G., Della Valle, V. (2005), *Duemilasei parole nuove. Un dizionario di neologismi dai giornali*, Sperling & Kupfer editori, Milano, p. V.

[7] Giuva L. (2005), *I sistemi di gestione informatica dei documenti: esperienze e modelli. Un'introduzione*, "Archivi e Computer", a. XV, n. 1, p.11-12.

[8] *Ibidem*.

[9] ISO 15489-1:2001 - *Information and documentation - Record management - part 1 General*; ISO/TR 15489-2:2001 - *Information and documentation - Record management - part 2 Guidelines*; ISO/TS 23081-1:2006 - *Information and documentation - Records management processes - Metadata for records - Part 1 Principles*; ISO/TS 23081-2:2007 - *Information and documentation - Records management processes - Metadata for records - Part 2 - Conceptual and implementation issues*. Si precisa che è in corso di revisione la ISO 15489 attraverso la TC46/SC11/WG5. A tal proposito si veda <www.iso.org/iso/standards_development/technical_committees/list_of_iso_technical_committees/iso_technical_committee.htm?commid=48856> (ultima consultazione 15 aprile 2008).

[10] Hofman H. (2005), *Standardisation in records management,* «Archivi e Computer», a XV, n. 1, p. 83

[11] Si vedano a proposito: Association of Records Managers and Administrators (ARMA), (1989). *Glossary of Records Management Terms.* ARMA; Bellardo L., Bellardo L.L., (1992), *A Glossary for Archivists, Manuscript Curators, and Records Managers,* Society of American Archivists, Chicago; Cox L., (comp.) (1990), *Glossary for Electronic Archives and Records Management*, Annex V, Management of Electronic Records: Issues and Guidelines, Advisory Committee for the Coordination of Information Systems, United Nations; National Archives and Records Administration (NARA) (1993), *A Federal Records Management Glossary,* (2nd ed.), Washington, DC: NARA Office of Records Administration.

[12] Pearce-Moses R. (2005), *A Glossary of Archival and Records Terminology, "*Archival fundamental series II"*,* Society of American Archivists.

[13] *Ibidem*.

[14] *Ibidem*.

[15] Scott M., (2008), *Wordsmith Tools 5.0*, Lexical Analysis Software, Liverpool. Si precisa che la versione 5.0 fornisce le statistiche del *corpus* solo per l'estrazione terminologica di monogrammi pertanto per poter analizzare la *type/token ratio* al crescere dei lemmi nei *cluster* si è dovuto ricorrere alla versione 3.0 che permette si estrarre *wordlist* di bigrammi, trigrammi etc. fornendo per ognuno di loro i relativi dati statistici.

[16] ISO 1087-1:2000(E/F) - *Terminology work - Vocabulary.*

[17] de Saussure F. (1965), *Cours de linguistique générale*, Paris: Payot, Ch. Bally et A. Sechehaye, p. 125
[18] Decreto Legislativo 7 marzo 2005, n. 82, *Codice dell'Amministrazione Digitale*, p. 3.
[19] *Ibidem*.
[20] *Ibidem*.
[21] <62.149.227.181/glossarioR.php> (ultima consultazione 15 aprile 2008).
[22] UNI-ISO 5127/05, 1987, *Documentazione ed informazione, Vocabolario, Linguaggi documentari*, Milano, UNI, p. 25.
[23] Decreto del Presidente della Repubblica 28 dicembre 2000, n. 445, *Testo unico delle disposizioni legislative e regolamentari in materia di documentazione amministrativi*, p. 30.
[24] *Ibidem*.
[25] UNI-ISO 5127/06, 1987, *Documentazione ed informazione, Vocabolario, Linguaggi documentari*, Milano, UNI, p. 6-7, 12.
[26] *Ibidem*.
[27] *Ibidem*.
[28] *Ibidem*.
[29] *Ibidem*.
[30] *Ibidem*.
[31] *Ibidem*.

Bibliografia

Adamo G., Della Valle V. (2005), *Duemilasei parole nuove. Un dizionario di neologismi dai giornali*, Milano, Sperling & Kupfer editori.

AS 4390, 1996.

Association of Records Managers and Administrators (ARMA), (1989), *Glossary of Records Management Terms,* ARMA.

Bellardo L., Bellardo, L.L., (1992), *A Glossary for Archivists, Manuscript Curators, and Records Managers,* Society of American Archivists, Chicago.

Cabrè M.T. (1999), *Terminology: Theory, methods and applications,* Amsterdam/ Philadelphia, John Benjamins B.V.

Cox L., (comp.) (1990), *Glossary for Electronic Archives and Records Management*, Annex V, Management of Electronic Records: Issues and Guidelines, Advisory Committee for the Co-ordination of Information Systems, United Nations.

de Saussure F. (1965), *Cours de linguistique générale*, Paris: Payot, Ch. Bally et A. Sechehaye.

Decreto Legislativo 7 marzo 2005, n. 82, *Codice dell'Amministrazione Digitale.*

Decreto del Presidente della Repubblica 28 dicembre 2000, n. 445, *Testo unico delle disposizioni legislative e regolamentari in materia di documentazione amministrativi.*

Dryden J. (2005), *A Tower of Babel: standardizing Archival Terminology*, «Archival Science».

Giuva L. (2005), *I sistemi di gestione informatica dei documenti: esperienze e modelli. Un'introduzione*, «Archivi e Computer», a XV, n. 1.

Hofman H. (2005), *standardisation in records management*, «Archivi e Computer», a XV, n. 1.

ISO 1087-1:2000(E/F) - Terminology work - Vocabulary.

ISO 15489-1:2001- Information and documentation - Record management - part 1 General .

ISO/TR 15489-2:2001 - Information and documentation - Record management - part 2 Guidelines.

ISO/TS 23081-1:2006 - Information and documentation - Records management processes - Metadata for records - Part 1 - Principles.

ISO/TS 23081-2:2007 - Information and documentation - Records management processes - Metadata for records - Part 2 - Conceptual and implementation issues.

National Archives and Records Administration (NARA) (1993), *A Federal Records Management Glossary,* (2nd ed.), Washington, DC: NARA Office of Records Administration.

Nogueira C.C. (a cura di) (1988), *Glossary of basic archival and library conservation terms: English with equivalents Spanish, German, Italian, French and Russian*, ICA Handbooks, Volume 4, München; New York; London; Paris: Saur.

Pearce-Moses R. (2005), *A Glossary of Archival and Records Terminology,* Archival fundamental series II*,* Society of American Archivists.

Scott M., (2008), *Wordsmith Tools 5.0*, Lexical Analysis *software*, Liverpool.

UNI-ISO 5127/05, 1987, Documentazione ed informazione, Vocabolario, Acquisizione, identificazione e analisi di documenti e dati, Milano, UNI.

UNI-ISO 5127/06, 1987, Documentazione ed informazione, Vocabolario, Linguaggi documentari, Milano, UNI.

Walne P., (a cura di) (1988), *Dictionary of Archival Terminology: English and French with equivalents in Dutch, German, Italian, Russian and Spanish*, 2nd ed., ICA Handbooks Series, Volume 7, München; New York; London; Paris: Saur.

<62.149.227.181/glossarioR.php> (ultima consultazione 15 aprile 2008).

<www.iso.org/iso/standards_development/technical_committees/list_of_iso_technical_committees/iso_technical_committee.htm?commid=48856> (ultima consultazione 15 aprile 2008).

Strutturazione dell'informazione e integrazione della conoscenza

ANNA ROVELLA, GIOVANNI MARRÈ

This article reports the results of a survey and an experiment conducted in collaboration between ItConsult and the Laboratorio di Documentazione at the University of Calabria in order to determine the effects of the integration of a document management system in a knowledge Management Platform. The work aims at the extraction of information from unstructured documents and their content management.

Keywords: Knowledge – Information – BMP – Document Management System

Il rapporto tra informazione e conoscenza è – da sempre – oggetto di studio e di riflessione in un contesto multidisciplinare che spazia dall'epistemologia all'informatica, all'economia e alle scienze cognitive, per citarne solo alcuni. La molteplicità di approcci con cui tale relazione è stata affrontata ha determinato la produzione di differenti definizioni finalizzate a delimitare i confini concettuali tra l'uno e l'altro termine [1] e ha, in maniera diversa, influenzato le realizzazioni pratiche conseguenti. Se nel sentire comune informazione e conoscenza sono, non di rado, utilizzate come sinonimi, la differenza che le separa sotto il profilo semantico è stata, per contro, formalizzata in modo evidente. «Quando si cerca di caratterizzare l'attività intellettuale i termini che vengono in mente sono numerosi. A priori, possono essere presi in considerazione tre livelli di analisi:

1. la conoscenza, potremmo dire la comprensione, è il livello superiore, quello della comprensione dei sistemi;
2. il sapere, che è una conoscenza operativa, un *savoir-faire*, un'attitudine senza dubbio, anche un saper essere un saper vivere, ecc.;
3. l'informazione, che è ciò attraverso cui la conoscenza e il sapere possono scambiarsi, ma anche, e sempre più essere prodotti» [2].

In questa tripartizione gerarchica dei tre livelli di acquisizione cognitiva l'informazione si qualifica come momento di iniziale riduzione dell'incertezza e di *input* del processo mentre la conoscenza diventa anche il momento di ripensamento critico e di acquisizione al patrimonio individuale e collettivo di nuovi descrittori della realtà esperienziale, utili alla definizione di strategie decisionali, organizzative e operative. Si delinea, quindi, un processo concettualmente strutturato all'interno del quale i diversi momenti postulano concretizzazioni pratiche definite e strettamente conseguenti. A

fronte di tutto ciò, aziende e pubbliche amministrazioni evidenziano sostanziali difficoltà nella realizzazione di processi di integrazione tra i diversi elementi della catena della conoscenza, tanto da spingere quasi tutte le realtà imprenditoriali del settore a misurarsi con lo specifico problema coniando anche un apposito acronimo: CMIS (*Content Management Interoperability Services*) [3]. Archivi cartacei ed elettronici, sistemi informatici disomogenei, dati strutturati e documenti tradiscono il risultato di una gestione del capitale cognitivo priva di un disegno organico di sviluppo, tesa a privilegiare, di volta in volta, aspetti particolari e segmentati, se non la singola applicazione tecnologica. L'informazione, chiusa all'interno di circuiti ristretti e limitativi, imbocca percorsi ripetitivi e stagnanti che ne inibiscono fortemente la potenziale azione di diffusione e condivisione. Molto spesso i dati, quali elemento primario e grezzo, vengono immagazzinati in maniera quasi compulsiva, senza un necessario riferimento a processi di sistematizzazione e le informazioni derivanti, spesso eccessivamente frammentate, lungi dall'agevolare il processo decisionale ne generano un sostanziale rallentamento. «Informazione e ignoranza, scelta, previsione e incertezza, sono tutte intimamente correlate (…) Al confine della completa conoscenza e della completa ignoranza, sembra intuitivamente ragionevole parlare di gradi di incertezza. Più vasta è la scelta, più esteso è l'insieme delle alternative che si aprono davanti a noi, più incerti noi siamo circa come procedere e di maggiore informazione abbiamo bisogno per prendere la nostra decisione» [4]. L'accumulo di dati e di informazioni non gestite, e la mancata evoluzione ed integrazione tra informazioni e conoscenza, producono disorientamento e attivano un'azione di rallentamento o addirittura paralizzano le capacità decisionali e strategiche delle organizzazioni in una paradossale confutazione dell'assunto di Shannon: pur in presenza di una bassa entropia positiva non aumenta il supporto alla decisione [5].

La conoscenza rifugge dalla frammentazione: essa consegue e consolida il suo valore mettendo a sistema le informazioni e strutturandone organicamente le unità costitutive mediante un'accurata operazione di definizione dei concetti cui fa seguito l'integrazione di contenuti e di forma nel processo di comunicazione che si occupa di socializzarle. L'informazione, così elaborata, è in grado di offrire una nuova prospettiva di interpretazione di eventi e oggetti, lasciando cogliere significati in precedenza nascosti e evidenziando relazioni inattese. Ed è proprio in questo fluido compenetrarsi e comparteciparsi che essa diventa un fattore di mediazione, elemento necessario a produrre e costruire dinamicamente nuova conoscenza attraverso la ristrutturazione e l'integrazione di nuove valenze. Tuttavia, senza voler qui ulteriormente approfondire la complessa disquisizione terminologica e le problematiche connesse alla formalizzazione ed utilizzazione della conoscenza, vorremmo limitare la nostra riflessione ad un caso determinato e specifico: la possibile integrazione tra gestione dei documenti e gestione dei contenuti all'interno di *workflow* di *Business Process Management* (BPM). Il contesto tecnologico di implementazione e sperimentazione è "josh Protocol!" [6], piattaforma di gestione della co-

noscenza, i cui componenti di protocollo e gestione documentale sono frutto di una collaborazione tra l'azienda produttrice e il Laboratorio di Documentazione dell'Università della Calabria [7].

Per loro stessa natura, i *workflow* di BPM si configurano come strumenti di ricerca attivi non solo sull'intero patrimonio documentale ma anche in ragione delle relazioni che gli stessi intrattengono con i dati e le informazioni provenienti dalle fasi dei *workflow* di processo. I BPM, infatti, nascono e vengono normalmente elaborati nella fiduciosa aspettativa di poter fornire risposte concrete all'esigenza di interoperabilità tra sistemi informativi esistenti e specifici applicativi progettati per l'accesso e la distribuzione dei dati. Tuttavia, malgrado gli sforzi effettuati, le tecnologie messe a punto per la classificazione delle informazioni e la gestione dei contenuti agiscono, di fatto, prevalentemente su dati strutturati, mentre l'integrazione della conoscenza destrutturata presente all'interno dei processi richiede ancora il prevalente ricorso ad attività manuali, vincolate all'utilizzo di personale specificatamente qualificato con conseguente dispendio di tempo e incremento dei costi. In tale prospettiva, la definizione concettuale di regole e la realizzazione di applicativi adattivi capaci di utilizzare le risorse terminologiche per generare descrittori informativi da testi non strutturati mediante procedure di analisi concettuale, assume una sicura rilevanza teorica ed operativa [8]. Particolarmente evidente – dopo un primo periodo d'uso – è l'incremento dei benefici che l'adozione di simili strumenti comporta in termini di contenimento dei costi, riduzione sostanziale dei tempi di ricerca e di recupero delle informazioni. Al più tradizionale approccio della strutturazione dei dati destrutturati si sostituisce la modellizzazione concettuale di *features* capaci di descrivere comunque modelli documentali diversificati contestualizzando la rilevanza dei termini estratti come possibili descrittori [9]. In questa prospettiva, nella quale la virtualizzazione dei supporti affievolisce la significanza dei legami organici tra gli atti di uno stesso complesso, classificazione e indicizzazione si configurano come delle vere e proprie chiavi di accesso alle varie fasi che accompagnano il ciclo vitale del documento, dalla sua produzione, alla selezione, alla conservazione temporanea e permanente, alla consultabilità e fruibilità.

Il modello è un sistema di accesso all'informazione, costituito da un'integrazione tra tecniche di *information retrieval* e *ontology engineering*, combinate mediante un approccio *knowledge-based*, concettuale e automatico-statistico. In particolare, la base di conoscenza del sistema è strutturata su uno schema concettuale elaborato mediante la combinazione di dati estratti dall'organigramma e dal titolario di classificazione associati a tipologie documentali a loro volta relazionate alle norme e ai regolamenti in modo da permettere la strutturazione e la definizione delle *query*.

Il modello progettato consente, nel dettaglio, di effettuare:

- la ricerca di informazione estratta mediante analisi semantica, automatica e semiautomatica, dei testi, condotta su basi di dati testuali non strutturate;

- la ricerca di documenti attraverso tecniche avanzate di interazione con l'utente, (*profiling*, espansione semantica delle *query*, *relevance feedback*, *clustering* dei documenti);
- la valutazione della *performance* del motore di ricerca e sua ottimizzazione attraverso l'analisi dei flussi di documenti e dei processi di *business*;
- la costruzione di un'ontologia dell'organizzazione in grado di guidare il processo di ricerca delle informazioni tramite la definizione di classi, relazioni, e strutture informative rilevanti;
- l'estrazione di informazione e l'identificazione di contenuti informativi specifici all'interno dei documenti e del flusso;
- il trattamento automatico del testo per l'identificazione di occorrenze di informazione strutturata nei documenti;
- la costruzione di strumenti di annotazione automatica dei documenti per la strutturazione di relazioni tassonomiche e l'implementazione di *corpora* terminologici.

Il modello concettuale qui sommariamente delineato, applicato ad amministrazioni con strutture decisionali definite ma con produzione documentale non formalizzata, potrebbe produrre vantaggi operativi anche in tempi estremamente brevi aumentando considerevolmente il supporto informativo e la conseguente capacità di scelta dei decisori.

In ogni caso l'integrazione di metodologie e di strumenti tecnologici per l'automazione delle operazioni di classificazione e di indicizzazione all'interno di piattaforme di gestione delle conoscenza è uno stimolante terreno di confronto per l'elaborazione di esperienze e studio di casi la cui riproducibilità è in grado di apportare benefici, ad ampio raggio, in ogni organizzazione sia sul piano organizzativo sia su quello delle *performance*.

Non sono da sottovalutare, tuttavia, le criticità che il modello ancora presenta relativamente agli aspetti linguistico terminologici determinati dall'analisi dei contenuti applicata a contesti fortemente eterogenei e dalla necessità di usare una pluralità di vocabolari di dominio nell'interazione dei quali emergono significative problematiche di gestione delle sinonimie e di definizione delle relazioni, specie in contesti vocati ad una naturale multilinguismo [10]

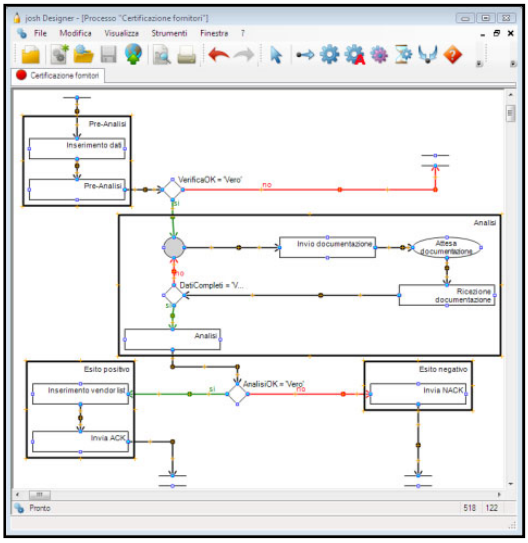
Il caso applicativo.

Quello della gestione dei dati strutturati è stato tradizionalmente considerato un mondo separato da quello dei dati non strutturati. Il primo, fatto essenzialmente di *database*, dispone di tecnologie e metodologie consolidate e stabili, sia pure in evoluzione; il secondo, fatto soprattutto di documenti, non solo testuali, e anche da processi e altri oggetti è, viceversa, sostanzialmente privo di *standard*, presenta problemi meto-

dologici e di interoperabilità. L'interpretazione, l'elaborazione, la semplice estrazione di dati da documenti – intesi in senso lato – è una operazione tuttora particolarmente complessa e difficilmente generalizzabile.

Le tecnologie di *Business Process Management* (BPM) si possono intendere ed utilizzare come un ponte fra i due mondi. Ad esempio, la piattaforma *josh* dispone di un *Workflow Management System* (WFMS) che:

- consente la descrizione semi-formale dei processi con un linguaggio grafico di modellazione [11],
- manda in esecuzione tali diagrammi, chiamando direttamente in causa le persone coinvolte nel processo, attraverso il loro *browser* Internet.

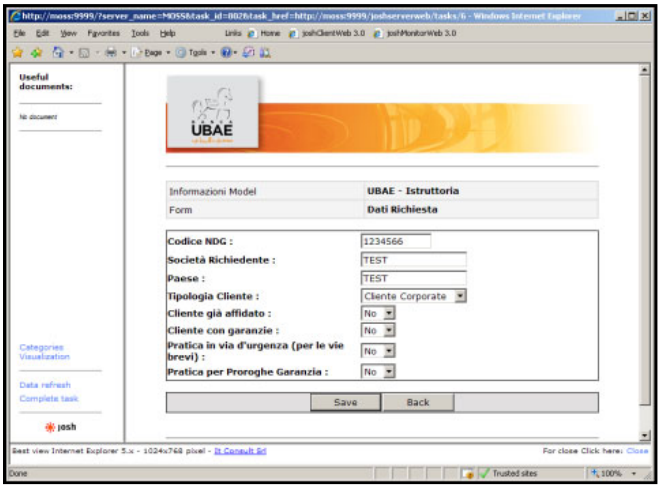


Si tratta di una tecnologia che costituisce una efficace modalità di rappresentazione della conoscenza incorporata nei processi ma anche di un meccanismo di astrazione e disaccoppiamento dei dati dai processi, che però consente di associare gli uni agli altri, ad esempio associando metadati (strutturati) ai documenti (non strutturati).

Il modo tradizionale di sviluppare quelle applicazioni che necessitano, durante l'esecuzione, dell'intervento successivo di diversi operatori, implica l'incorporazione del flusso di azioni all'interno del codice sorgente con, nella migliore delle ipotesi, una parziale configurabilità di opzioni attraverso impostazioni utente. In pratica il progettista/sviluppatore codifica sia la logica dell'applicazione sia le singole *form* che determinano l'interazione dell'applicazione stessa con l'utente.

Ciò che, invece, un sistema *software* di BPM permette, è di disaccoppiare la logica, che viene disegnata graficamente, dalle *form* che possono essere generate automaticamente o continuare ad essere sviluppate con linguaggi tradizionali, ma che comunque beneficiano della parcellizzazione e modularizzazione dello sviluppo di singole componenti di gran lunga più piccole e più semplici di una grande applicazione monolitica che, peraltro, coinvolgendo una elevata quantità utenti, deve spesso possedere elevati requisiti di solidità e scalabilità.

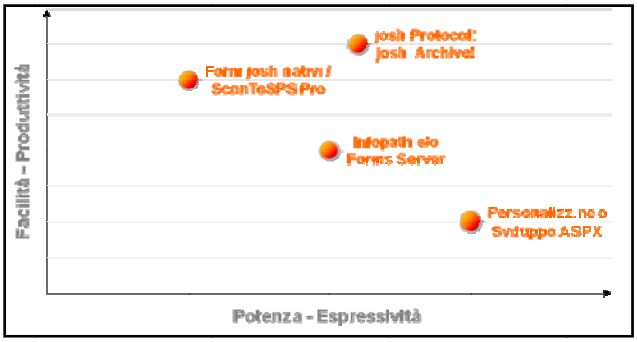
Questa modalità operativa configura un nuovo modo di costruire applicazioni *software*, in cui i processi da analizzare ed automatizzare trovano collocazione nell'ambito di processi generali di un livello di astrazione più elevato, per i quali è l'esperto di dominio applicativo che analizza e definisce quel flusso delle azioni (il *workflow*) che poi, contestualmente o successivamente viene formalizzato in forma grafica, utilizzando lo specifico linguaggio dello strumento software. Tipicamente, una volta descritto un processo in termini di *workflow*, è attraverso successivi passi di raffinamento (*task activity*) che si giunge a definire il dettaglio dei singoli *task*. Il *task* è un'unità elementare di lavoro all'interno di un processo ed è eseguita da un singolo *attore* (tipicamente umano, ma talvolta automatico). Alcune *task activity* consistono nella visualizzazione e/o modifica di dati strutturati collocati su *database* e manipolati attraverso delle *form*.



In josh la costruzione delle *form* di accesso ai *database* si può realizzare in diverse modalità, che hanno un livello decrescente di facilità d'uso (produttività) e parallelamente crescente in termini di espressività (potenza), che sono:

- *form* di josh nativi generati attraverso un *wizard* [12] da cui si scelgono le variabili di processo da editare o visualizzare sul *browser* Internet;
- Microsoft Infopath e/o Forms Server; lo strumento Microsoft per generare e pubblicare *form* che debbono interagire con josh attraverso una modesta attività di sviluppo;
- personalizzazione o sviluppo di *task activity* e di *form*, in ambo i casi con una vera e propria programmazione in ASP.NET;

È opportuno precisare che in josh esistono alcune applicazioni verticali che, oltre a specializzare josh e SharePoint nella risoluzione di specifiche tematiche, gestiscono alcuni dati strutturati *out-of-the-box*. È il caso, ad esempio, in josh Protocol!, dei dati di protocollo dei singoli documenti protocollati registrati e archiviati.



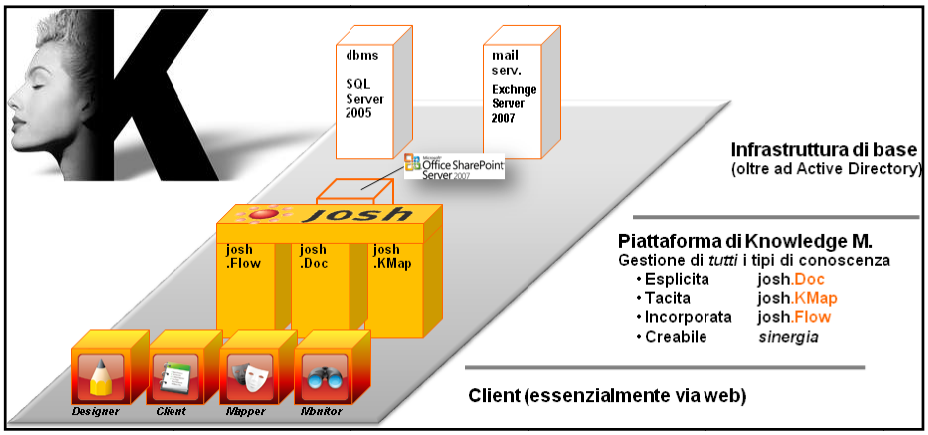
Josh è una piattaforma *software* di *Knowledge Management* (Gestione della Conoscenza Organizzativa - KM) di tipo *enterprise* che consta di tre moduli principali, ciascuno corrispondente ad una tipologia di conoscenza gestita, e, in particolare:

a) Josh.Doc: un sistema di gestione dei documenti formalizzati;
b) Josh.KMap: un sistema di mappatura delle competenze individuali e per la formalizzazione della gestione della conoscenza tacita;
c) Josh.Flow: un *workflow management* per gestire la conoscenza Incorporata nei processi.

Su questa piattaforma si innestano le componenti verticali come joshArchive! (Archiviazione e Conservazione Sostitutiva) e joshProtocol! (Protocollo Informatico) come anche altre applicazioni *ad hoc* che di volta in volta possono essere sviluppate.

Questo approccio comporta una serie di benefici. In primo luogo il fatto che la logica con cui si sviluppano queste "nuove" applicazioni è direttamente presidiata dall'esperto del dominio applicativo e non più dal solo sviluppatore; conseguentemente le

modifiche che consentono la reale riconfigurabilità dei processi aziendali in maniera reattiva divengono rapidissime ed economiche oltre che sicuramente pertinenti.



Infine, non meno importante, il fatto che il sistema BPM consente di migliorare l'efficacia delle ricerche e la produttività individuale, rilevando quali documenti vengono maggiormente consultati durante i *task* di un processo in modo tale di poter successivamente proporre agli utenti i documenti più utili in quel contesto specifico (perché già usati dai colleghi) e/o classificarli automaticamente in base al loro utilizzo, aggiungendo le parole chiave legate ai *task* in cui vengono impiegati, i quali si configurano – di fatto – come dei *training set*. Su quest'area sono possibili numerose migliorie, coerentemente al modello presentato nella prima parte dell'articolo.

In conclusione, l'esperienza di sinergia realizzata è stata sicuramente positiva ed ha portato alla realizzazione di un prodotto commerciale che rappresenta un primo importante tassello concettuale verso l'integrazione dei sistemi e l'estrazione di conoscenza da testi non strutturati.

Note

[1] «Prima di entrare nel vivo della nostra teoria, è bene chiarire con tre osservazioni, somiglianze e differenze fra i concetti di conoscenza e di informazione. La prima osservazione è che la conoscenza, diversamente dall''informazione, concerne le credenze e il coinvolgimento. È cioè funzione del punto di vista, della prospettiva o dell'intenzione del singolo. La seconda osservazione è che la conoscenza, diversamente dall'informazione, riguarda l'azione. È sempre diretta a un fine. La terza osservazione è che la conoscenza, come

l'informazione, concerne significati; è specifica del contesto e relazionale». Ikujiro Nonaka, Hirotaka Takeuchi, *The knowledge creating company*, Oxford University Press, 1995, trad. it. 1997, p. 94.

[2] Laurent Gille, *La protezione della proprietà intellettuale fattore della divisione internazionale della conoscenza*, in: Antonio Pilati, Antonio Perucci (a cura di), *Economia della Conoscenza. Profili teorici ed evidenze empiriche*, Il Mulino, Bologna, 2005, p. 211.

[3] «Sous l'acronyme de CMIS sont regroupées les specifications d'une interface, des techniques, un langage commun et des protocols qui doivent permettre de developer des moyens de consultation et d'échanges d'objets (documents, fichiers) entre les référentiels de plusieurs logiciels de GEIDE ou d'ECM». *CMIS: une interface de services pour l'interopérabilité entre des solutions de gestion de contenu ou de GEIDE*, in: "MOS", n. 251, settembre 2008, p. 5. Cfr. Anche Alain Garnier, *L'information non structurée dans l'entreprise*, Lavoisier, Paris, 2007.

[4] Sigmund Koch, *Information Theory*, in: *Psychology: A Study of a Science*, 1959, pp. 614-615.

[5] Nel 1948 Claude E. Shannon pubblicava *A mathematical theory of communication* ("Bell System Technical Journal", vol. 27, pp. 379-423 and 623-656, July and October, 1948) gettando le basi della teoria dell'informazione. Egli partiva dall'idea che un messaggio inviato attraverso un qualsiasi canale subisce nel corso della trasmissione deformazioni diverse per cui al suo arrivo esso ha perso una parte delle informazioni che conteneva originariamente. Egli stabilì quindi una correlazione tra tale perdita di informazioni e l'entropia, ovvero la funzione matematica che esprime la degradazione dell'energia che si verifica in ogni trasformazione del lavoro meccanico in calore, in quanto la trasformazione inversa dal calore al lavoro meccanico non risulta mai completa. In base a questa analogia la quantità di informazione trasmessa può essere calcolata come entropia negativa giacché nella trasmissione dei messaggi come nella trasformazione dell'energia, l'entropia negativa decresce continuamente in quanto quella positiva (perdita di informazione o degradazione di energia) cresce continuamente.

[6] L'azienda produttrice è ItConsult di Fermignano (Urbino).

[7] Referente di progetto per l'Università della Calabria è stata la prof. Anna Rovella, coautrice del presente testo.

[8] «The link between terminology and cognitive science is created by concepts. Concepts are units constituting the basis of knowledge and concept systems describe the way each field organizes knowledge. In this sense, the theory of terminology and the theory of knowledge are closely related». Maria Teresa Cabré, *Terminology. Theory, methods and applications*, Amsterdam, 1998, p. 52.

[9] Cfr. C. Beghtol, *Semantic Validity: concepts of warrant in bibliographic classification systems*, in: "Library resources and technical services", 30 (1986), n. 2, pp. 109-125, nonché Alberto Cheti, *Le categorie nell'indicizzazione. Indagine su alcuni modelli di analisi e di organizzazione concettuale*, in: "Biblioteche oggi", n. 8 (1990) n. 1, pp. 29-49.

[10] «Lo sforzo di determinazione degli elementi costitutivi di una terminologia, dei valori semantici dei termini delle regole di combinazione accettate è lo sforzo di conferire a parti

del linguaggio verbale le caratteristiche dei codici più semplici e dei calcoli. Costruire un linguaggio formale significa costruire un'area d'uso della lingua in cui ogni discorso sia un testo, ogni comprensione un processo di interpretazione certa e a termine. Un linguaggio speciale, tanto più se le sue regole costitutive sono esplicitate ed esso sia dunque formalizzato, è per così dire un tentativo di servirsi dei materiali della lingua per uscire fuori dalla storia, fuori dalla durata e dalla fluttuante massa parlante, per arrivare a costruire testi valevoli oltre il tempo e la contingenza in cui dapprima si produssero». Tullio De mauro, *Minisemantica*, Laterza, Bari, 1982, pp. 148-149.

[11] In particolare, josh utilizza l'evoluzione di un linguaggio di modellazione denominato WIDE (*Workflow on Intelligent Distributed database Environment*), frutto di un progetto di ricerca ESPRIT, cofinanziato dalla Commissione Europea, che ha avuto come *partner* accademici il Politecnico di Milano e la University of Twente (NL). Il progetto è descritto in Paul Grefen, Barbara Pernici, Gabriel Sanchez, *Database support for Workflow Management - The WIDE Project*, Kluwer Academic Publishers, 1999.

[12] Ossia un'autocomposizione, in cui l'utente è assistito a passo a passo.

Hanno collaborato a questo numero:

CHIARA BASILE, Università di Bologna. Dipartimento di Matematica, Bologna, <basile @dm.unibo.it>

DOMENICO BOGLIOLO, Associazione Italiana per la Documentazione Avanzata (AIDA), Roma, <ingo.bogliolo@gmail.com>

PIERA BELCASTRO, Università della Calabria. Dipartimento di Linguistica, Cosenza, <piera.belcastro@unical.it>

FRANCO BERTACCINI, Università di Bologna. Scuola Superiore di Lingue Moderne per Interpreti e traduttori (SSLMIT), Forlì, <bertaccini@sslmit.unibo.it>

ANDREA BOCCO, Politecnico di Torino. Dipartimento Casa Città (DICAS), Torino, <andrea.bocco@polito.it>

ENRICA BODRATO, Politecnico di Torino. Dipartimento Casa Città (DICAS), Torino, <enrica.bodrato@polito.it>

VALENTINA BONO, Università di Bologna. Scuola Superiore di Lingue Moderne per Interpreti e traduttori (SSLMIT), Forlì

MARIA TERESA CABRÉ, Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada, Barcelona, <teresa.cabre@upf.edu>

ELENA CARDILLO, Università della Calabria. Dipartimento di Linguistica, Cosenza, <elena.cardillo@unical.it>

GIUSEPPE CAVARRETTA, URT-CNR. Dipartimento Sistemi di Produzione, Cosenza, <giuseppealfredo.cavarretta@.cnr.it>

MADEL CRASTA, BAICR, Roma, <segreteria@baicr.it>

FELICE DELL'ORLETTA, Università di Pisa. Dipartimento di Informatica. Istituto di Linguistica Computazionale del CNR, Pisa, <felice.dellorletta@ilc.cnr.it>

ANTONIETTA FOLINO, Università della Calabria. Dipartimento di Linguistica, Cosenza, <antonietta.folino@unical.it>

ÁGOTA FÓRIS, University of West Hungary. TermIK, Szombathely, <aforis@t-online.hu>

PAOLO FRANZESE, Ministero per i beni e le attività culturali, Roma, <paolo.franzese@beniculturali.it>

BERTRAND GAIFFE, Analyse et Traitement Informatique de la Langue Française, UMR 7118 Nancy-Université, CNRS, <bertrand.gaiffe@atilf.fr>

CLAUDIO GIOVANARDI, Università Roma Tre, Roma, <giovanar@ita.uniroma3.it>

ROBERTO GUARASCI, Università della Calabria. Dipartimento di Linguistica, Cosenza, <guarasci@unical.it>

EVELYNE JACQUEY, Analyse et Traitement Informatique de la Langue Française, UMR 7118 Nancy- Université, CNRS, <ejacquey@atilf.fr>

LAURENCE KISTER, Analyse et Traitement Informatique de la Langue Française, UMR 7118 Nancy- Université, CNRS, <laurence.kister@univ-nancy2.fr>

MAURIZIO LANA, Università degli Studi del Piemonte Orientale "A. Avogadro" a Vercelli. Dipartimento di Studi Umanistici, Vercelli, <m.lana@lett.unipmn.it>

CLAUDIA LECCI, Università di Bologna. Scuola Superiore di Lingue Moderne per Interpreti e traduttori (SSLMIT), Forlì, <claudia.lecci2@unibo.it>

ALESSANDRO LENCI, Università di Pisa. Dipartimento di Linguistica, Pisa, <alessandro.lenci@ilc.cnr.it>

SIMONE MARCHI, Istituto di Linguistica Computazionale del CNR, Pisa, <simone.marchi@ilc.cnr.it>

GIOVANNI MARRÈ, ItConsult, Roma, <gmarre@itconsult.it>

SIMONETTA MONTEMAGNI, Istituto di Linguistica Computazionale del CNR, Pisa, <simonetta.montemagni@ilc.cnr.it>

ANTONELLA PERIN, Politecnico di Torino. Dipartimento Casa Città (DICAS), Torino, <antonella.perin@polito.it>

SONIA PIOTTI, Università Cattolica del Sacro Cuore, Brescia, <sonia.piotti@unicatt.it>

VITO PIRRELLI, Istituto di Linguistica Computazionale del CNR, Pisa, <vito.pirrelli@ilc.cnr.it>

DONATELLA PULITANO, Cancelleria di Stato del Cantone di Berna, <donatella.pulitano@sta.be.ch>; Scuola per Traduttori ed Interpreti dell'Università di Ginevra, <donatella.pulitano@eti.unige.ch>

ROGER ROBERTS, Président Titan asbl, RTBF/Titan, <rro@rtbf.be>

LUCIANO ROMITO, Università della Calabria. Laboratorio di Fonetica, Cosenza, <luciano.romito@unical.it>

ANNA ROVELLA, Università della Calabria. Dipartimento di Linguistica, Cosenza, <anna.rovella@unical.it>

MARIA TAVERNITI, Università della Calabria. Dipartimento di Linguistica, Cosenza, <maria.taverniti@unical.it>

MARIA TUCCI, Università della Calabria. Laboratorio di Fonetica, Cosenza, <tucci.maria@libero.it>

GIULIA VENTURI, Istituto di Linguistica Computazionale del CNR, Pisa, <giulia.venturi@ilc.cnr.it>

MARIA TERESA ZANOLA, Università Cattolica del Sacro Cuore, Milano, <mariateresa.zanola@unicatt.it>

Norme per i collaboratori

a) La collaborazione ad "**AIDA**informazioni" è libera e gratuita. I contributi sono sottoposti per la pubblicazione al vaglio preventivo del Comitato scientifico e, successivamente, della Redazione, che si riserva di chiedere agli autori modifiche e adeguamenti. Il contributo, una volta approvato, dovrà essere inviato come file **ODT** o **RTF** allegato a un messaggio di posta elettronica all'indirizzo: <redazione @aidainformazioni.it>.

b) Ordinariamente (salvo cioè il caso di numeri doppi, speciali, etc.) la **data di scadenza** ultima per la presentazione dei testi accettati per la stampa è la seguente:
n. 1 (gennaio-marzo) = 10 febbraio
n. 2 (aprile-giugno) = 10 maggio
n. 3 (luglio-settembre) = 10 settembre
n. 4 (ottobre-dicembre) = 15 novembre.

c) L'autore fornirà **obbligatoriamente** insieme al contributo
- il titolo del contributo in inglese
- due *abstract* (della lunghezza massima di dieci righe ciascuno):
  - uno in lingua inglese
  - uno in lingua italiana
- e due elenchi di parole chiave:
  - uno in lingua inglese
  - uno in lingua italiana.

d) I diritti d'autore appartengono ad **AIDA**, che si riserva di diffondere il contenuto della Rivista anche in formato elettronico e in rete.

e) Le note al testo vanno inserite in fondo al contributo (non "a piè di pagina" e senza usare la funzione NOTE del programma di elaborazione testi, nemmeno per le "note di chiusura") e numerate progressivamente.

f) La forma delle citazioni bibliografiche prevede per le opere a stampa le seguenti forme:

Laura Leonardi, *La dimensione sociale della globalizzazione.* Roma: Carocci, 2001

Paolo Bisogno, *L'informazione e i processi di comunicazione scientifica*, in *Documentazione e utenti: cultura del servizio, marketing, multimedialità. Atti del Convegno Nazionale Aida. Roma 10-12 febbraio 1993*, a cura di M.P. Carosella e P. Fratarcangeli. Padova: Mediagraf, 1994, p. 9-14

Gabriele Gatti, *Portali di piombo.* "AIDAinformazioni", 19 (2001), n. 1, p. 9-16

g) Le opere collettanee o comunque dovute a più di tre autori devono essere intestate al titolo.
La citazione da fonti elettroniche, oltre agli altri dati bibliografici, deve includere l'URL nella forma: <http://www.aidaweb.it>, cui va aggiunta la data dell'ultima consultazione in Internet, nella forma anno-mese-giorno (per es. «consultato in data 2001-01-31»).

h) Eventuali immagini – comprese le tabelle e i grafici – devono essere inviate già correttamente posizionate all'interno del contributo.

i) La segreteria di redazione provvederà, a richiesta e solo nei casi in cui l'articolo le sia pervenuto nei tempi prescritti, ad inviare all'autore, esclusivamente via *e-mail*, l'ultima bozza; se entro 48 ore l'autore non restituisce il messaggio di posta elettronica con le eventuali variazioni, la bozza s'intende confermata.