AGGIORNAMENTO E RECUPERO DELLE INFORMAZIONI RILEVANTI NEL MARS: LO STRUMENTO TERMINOLOGICO E LE SUE POTENZIALITÀ

P. Agnello[1], C. Kirchsteiger[2]

1. ISPESL-DIPIA, Centro Ricerche di Monteporzio, Via di Fontana Candida 1 – 00040 Monteporzio Catone (RM)
2. European Commission, DG-JRC Institute for the Protection and Security of the Citizen (IPSC)- TP 670, 21020 Ispra (VA)

SOMMARIO

Scopo del lavoro è testare ed ampliare una metodologia per il recupero dell'informazione basata sull'indagine terminologica, applicando cioè un algoritmo per l'analisi del testo, ai rapporti di incidente rilevante contenuti nel database Major Accident Reporting System (MARS), curato dalla Commissione Europea secondo le indicazioni della Seveso II.

Tale analisi permette di realizzare il "*Master Thesaurus*" e, per mezzo di uno strumento di interrogazione del database, costruire dei *thesauri* su argomenti specifici *"user-defined"*.

Questo lavoro presenta i risultati sull'efficienza della metodologia e ne porta alcuni esempi, delinea, inoltre, i possibili sviluppi dei *thesauri* sia per individuare nuovi sotto-domini di interesse per i diversi tipi di utente, sia rispetto alle problematiche di ambiguità di interpretazione determinate dalle lingue di compilazione dei rapporti di incidente quando sono diverse dall'inglese, lingua in cui vengono caricati nel database, analizzati e resi disponibili per la consultazione.

1. APPROCCIO TERMINOLOGICO

1.1 Il thesaurus come strumento

Lo sviluppo di un linguaggio "naturale" compatibile con l'uso del computer prevede principalmente tre passi: la creazione di una base di conoscenza cioè una rete semantica di partenza che fornisca la lista organizzata dei concetti elementari e le loro combinazioni fondamentali; seguono gli *statements* ovvero le leggi empiriche che rappresentano i collegamenti possibili tra concetti per espandere dinamicamente la lista precedente generando così concetti di senso compiuto propri del dominio di interesse; ultimo elemento è il motore, ovvero i metodi di deduzione, costituito dal software in grado di utilizzare gli *statements* per dedurre i nuovi concetti complessi a partire dai concetti elementari memorizzati nella base di conoscenza e dalle regole di composizione.

Il sistema descritto utilizza algoritmi di analisi del testo che nella fase di analisi individuano parole e termini altamente significativi nel dominio di studio e le loro combinazioni più ricorrenti, mentre, nella fase di interrogazione della base di dati, permettono di recuperare l'informazione ricostruendola di volta in volta nella maniera più significativa rispetto alla richiesta del singolo utente.

Nella nostra applicazione è stato scelto il *thesaurus* che, tra gli strumenti per la rappresentazione formalizzata della conoscenza, risulta particolarmente indicato per l'indicizzazione ed il recupero dell'informazione espressa in testo libero.

Il *thesaurus* è un dizionario di termini controllati strettamente legato al dominio di applicazione da cui deriva. È stato sviluppato, in origine, per classificare soprattutto la documentazione a carattere bibliografico, ed è stato poi applicato più in generale a tutta quella documentazione la cui parte principale è composta da testo libero.

Le caratteristiche principali sono: presenza di codici posizionali, trattazione estensiva a volte ridondante della conoscenza, presenza di una gerarchia di concetti elementari, presenza di un alto numero di sinonimi e varianti lessicali, relazioni tra concetti di tipo posizionale, concetti complessi rappresentati da combinazioni di codici secondo le definizioni degli utenti che possono necessitare di dettagli differenti a seconda della loro professionalità.

L'organizzazione della conoscenza è in capitoli estesi che contengono una struttura ad albero e questo permette di suddividere il *thesaurus* in diversi *thesauri* orientati ciascuno ad uno specifico sotto-dominio dell'argomento trattato.

1.2 Ambito applicativo: il MARS

Il MARS (Major Accident Reporting System) è la base di dati della Commissione Europea, contenente la notifica degli incidenti rilevanti che coinvolgono sostanze pericolose in accordo con le indicazioni della Direttiva "Seveso II" 96/82/EC, e mira a collezionare in modo coerente i dati forniti su tali incidenti dalle Autorità Competenti degli Stati Membri dell'Unione Europea.

Visto l'enorme contenuto informativo, memorizzato in campi di testo libero molto estesi, è sorta spontanea la necessità di utilizzare lo strumento terminologico per una migliore fruibilità dell'informazione contenuta.

Il MARS rappresenta un esempio di sistema orientato al recupero dell'informazione e non alla sola gestione della stessa, e contiene la documentazione sui maggiori incidenti industriali in particolare nelle due parti chiamate *'short report'* e di *'full report'* nel formato prevalente di testo libero. L'informazione è memorizzata ed è mantenuta in un ambiente di database distribuito accessibile ai responsabili di ciascuno degli Stati Membri.

Ad oggi il contenuto del MARS è di circa 500 incidenti e di questi circa un terzo ha solamente lo *short report*. La struttura del database consta di circa 270 variabili per la descrizione completa del singolo evento e di queste 30 sono variabili di testo libero.

Se valutato su tutti i campi riempiti per i circa 500 incidenti, il contenuto è di circa 60.000 categorie individuali e dati numerici, e circa 15.000 campi di testo libero, per cui l'utente perde facilmente i dettagli di informazione registrati. Solitamente si utilizza una ricerca guidata per eventi simili codificati sulla base delle loro caratteristiche principali insieme ad alcune possibilità di ricerca per singole parole in testo libero o loro combinazioni.

Riassumendo e interpretando gli eventi così ritrovati si possono imparare delle lezioni: quale potrebbe essere, ad esempio, il beneficio ricavato dal designer di un certo progetto industriale potendo identificare specifici punti deboli della strumentazione oppure quali potrebbero essere le reazioni del personale sotto certe condizioni? L'attenzione cioè si sposta dall'informazione trasmessa sul singolo evento incidentale,ovvero il dato, al contenuto informativo che si può estrarre da più eventi nei quali si sono verificate condizioni o scenari simili, alla conoscenza che si può acquisire da una analisi di tali eventi simili.

Ci si può aspettare che per l'utente del database, quindi, non sia sufficiente ritrovare gli eventi in cui compare una certa parola chiave ma che invece sia più interessato a trovare eventi che nel loro complesso siano confrontabili con un certo scenario che include quella parola.

Poiché la ricchezza del linguaggio umano non può essere compressa in schemi gerarchici di categorie (schemi di codifica) senza una significativa perdita di informazione e di oggettività, l'uso di valori appartenenti a categorie (categoriali) in uno strumento di interrogazione (*query)* sul database è inevitabile rispetto a un basso "tasso di riconoscimento". Quest'ultimo è un parametro che serve a valutare due aspetti: la 'completezza' nel recupero dell'informazione (eventi rilevanti recuperati / tutti gli eventi recuperati nel db), e la 'correttezza' (eventi rilevanti recuperati / tutti gli eventi recuperati) degli eventi selezionati rispetto alla *query* fatta.

Quando si cerca per singoli elementi nel testo libero (le parole) il tasso di riconoscimento risulta abbastanza basso a causa della inevitabile ambiguità del linguaggio. Ad esempio l'uso della parola chiave "*fatigue*" originariamente assunta per indicare la presenza di un errore umano tra le possibili cause di un incidente riportato nel MARS, se utilizzata nella *query* basata su singoli elementi di testo libero ritrova solo incidenti dove "*fatigue*" è intesa come difetto del materiale che ha portato all'incidente (quindi tasso di riconoscimento = 0).

La soluzione per un accrescimento significativo del tasso di riconoscimento è tentare di ricostruire lo scenario di interesse partendo da elementi di informazione indicizzati disponibili nel corpo di testo degli eventi usando termini diretti e termini associati per similarità.

Tuttavia, poiché la creazione di un algoritmo che porti al 100% di correttezza il tasso di riconoscimento è praticamente impossibile anche per la presenza di parecchio "rumore di fondo" dovuto all'origine dei dati spesso in lingue diverse dall'inglese, è importante disegnare algoritmi di recupero di testo libero con una idea chiara delle mancanze così che quando gli errori vengono inevitabilmente fatti, questi siano in qualche modo comprensibili e prevedibili.

2. METODOLOGIA

2.1 Approcci terminologici e loro applicazione

Per sfruttare al massimo il contenuto informativo dei dati registrati, che di solito non viene utilizzato perché disperso in campi di testo libero molto estesi, è stato sviluppato uno strumento per l'indicizzazione sulla quale si possono poi effettuare ricerche tramite *query*. A tal fine sono stati presi in considerazione due approcci terminologici:

1. derivare le frasi e costruire gli indici dalle collezione di documenti di testo con tecniche statistiche.
2. utilizzare un *thesaurus* per derivare le associazioni tra termini, che conterrà un certo numero di parole e frasi raggruppate insieme in vari modi.

Nel primo approccio il successo è limitato in quanto l'associazione di parole nei documenti di testo occorre per una tale varietà di ragioni, delle quali le tecniche statistiche non riescono a tenere conto, piuttosto

che non perché si aggregano sempre allo stesso modo per produrre un utile indice di concetti, che alla fine si ottiene un sottoinsieme molto ridotto dell'effettivo contenuto informativo del testo analizzato.

Nel secondo approccio si tiene conto anche dei gruppi di sinonimi e delle classi di parole che sono correlate le une con le altre per la loro associazione ad un concetto più astratto o di livello superiore, e questo comporta una certa ridondanza che però non appesantisce la struttura del *thesaurus* ma ne determina la ricchezza informativa e la flessibilità.

Con lo strumento per il recupero dell'informazione tramite *query* sviluppato per il MARS si ha un misto dei due approcci. Per ciascun sotto-dominio individuato si costruisce una nuova applicazione basata su un *thesaurus* contenente termini e frasi derivati dalla collezione di testi generati in modo semi-automatico da tecniche statistiche e da decisioni dell'utente.

Poiché il *thesaurus* così costruito nel nostro caso porta ad avere diversi livelli di astrazione legati alla complessità delle aggregazioni, e ciò farebbe crescere la difficoltà computazionali al fine di ritrovare il livello di pertinenza di una certa richiesta di informazione da parte di un utente, si è preferito realizzare un *thesaurus* principale o "*Master thesaurus*" e diversi *thesauri* relativi ciascuno ad un sotto-dominio individuato nel dominio di applicazione.

L'altra parte importate della classificazione dell'informazione in un *thesaurus* convenzionale è per "parti di discorso", cioè l'insieme di leggi empiriche (grammatica) che regolano l'associazione di senso compiuto dei concetti elementari presenti nel *thesaurus* e che si manifestano poi nella struttura guidata di composizione di una *query* a seconda dello scenario informativo al quale l'utente vuole pervenire.

Ogni applicazione è limitata dalla base di conoscenza su un singolo sotto-dominio ma il tool terminologico sviluppato fornisce un'accessibilità multi-dominio attraverso un'applicazione, di facile utilizzo da parte dell'utente, che sfrutta la ridondanza e i codici posizionali contenuti in ciascun *thesaurus* dello specifico sotto-dominio per collegare le informazioni in maniera significativa a seconda della interrogazione ricevuta.

In seguito a studi statistici sulla frequenza delle diverse possibili interrogazioni da parte di utenti con diverse professionalità, al fine di determinare una scala di priorità relativa ai sotto-domini di maggior interesse per gli utenti del sistema, sono stati realizzati alcuni *thesauri* su sotto-domini specifici come ad esempio i sistemi di gestione della sicurezza, il fattore umano, etc., che vengono tenuti in continuo aggiornamento.

2.2 Creazione del "*Master thesaurus*"

Assumiamo che l'uso di certe parole in un rapporto d'incidente si correla con la descrizione dello specifico scenario d'incidente. Ci saranno per esempio parole che si correlano a diversi livelli (possibile ambiguità: vedi l'esempio della parola '*fatigue'*) con la descrizione di un errore umano come causa d'incidente. L'obiettivo è di creare liste o vocabolari strutturati (*thesauri*) di tali parole che siano abbastanza indicative, nel nostro esempio, di errori umani e che compaiano nel database. Le parole della lista devono quindi essere organizzate in classi in accordo con i tipi di errore umano che possono indicare. Può essere considerato l'uso di parole estratte da una classe come parole chiave in una *query* per il recupero dal database degli incidenti nei quali è presente.

Laddove le parole richiedano una qualificazione per essere indicative della classe "errore umano", il fatto stesso di questa qualificazione necessita di essere stabilita dal contesto della parola. Per la sua costruzione assumiamo che il contesto di una certa parola sia sufficientemente ben descritto dalla distribuzione delle frequenze significative delle parole prossime che compartecipano nell'ambito di una certa distanza nello stesso campo di testo.

Tutte le stringhe di 'rilevanza generale' sono caricate nel file del MARS denominato "*Master Thesaurus*" tutte quelle non rilevanti vengono accantonate in un file-cestino. Nella attuale versione la situazione si può riassumere in circa 12.500 stringhe di caratteri incluse in circa 500 eventi registrati nel MARS, di cui 9000 stringhe (parole) sono state incluse nel *thesaurus*.

Dovendo decidere sulla rilevanza di una parola il gestore principale del database deve raggruppare le parole (atomi) partendo da un *Master Thesaurus* non strutturato nel quale raccoglie, principalmente in base alla frequenza, le stringhe altamente generative, ovvero quelle presenti nel maggior numero di eventi, per muoversi poi verso una collezione di atomi di 'similarità generale' (*molecole*). Le regole usate in questa costruzione sono:

- ✓ selezionare un atomo come testa della molecola che sia il più adatto ad avere il nome più imparziale per descrivere il concetto sottinteso, e
- ✓ collegare a questa testa di molecola tutti gli atomi che differiscono solo minimamente da essa, cioè solo con riguardo ai sinonimi, alle variazioni grammaticali/sintattiche o agli errori di spelling.

Così dalle 9000 parole di 'rilevanza generale' sono state generate 4500 molecole, in altre parole le molecole mediamente constano di due atomi che sono per lo più variazioni grammaticali/sintattiche ed errori di spelling, fatta eccezione per un numero esiguo di esse.

2.3 Creazione dei *thesauri* orientati allo specifico sotto-dominio

La creazione dei *thesauri* relativi ad uno specifico sotto-dominio avviene per selezione delle molecole presenti nel *Master Thesaurus* per mezzo della seguente procedura semi-automatica. L'utente esperto di quel sotto-dominio, con algoritmi per l'analisi del testo si individuano le parole (*atomi*) più rappresentative nel suo contesto e le indica come parole-chiave. Partendo poi da insiemi esigui estratti dall'insieme degli atomi individuati li completa automaticamente con i sinonimi e le varianti lessicali/grammaticali già presenti nel *Master Thesaurus*.

Da queste stringhe, dette anche "*parole-seme*", parte lo sviluppo dei concetti più complessi associando parole cosiddette "*vicine*" per co-occorrenza o per posizione. Al fine di valutare tale vicinanza si utilizza una misura di 'similarità contestuale' tra parole nota come EMIM (Expected Mutual Information Measure) che riguarda la quantificazione della distanza tra le parole nella rete semantica (Church, 1989). La valutazione tramite il valore di EMIM della significatività delle frasi trovate rispetto al contesto individuato (e cioè con EMIM>7) fa sì che esse possano essere inserite nel *thesaurus*.

Dagli esperimenti fatti si ha che tra la distanza massima perché il valore calcolato di EMIM tra le parole-seme scelte e quelle del loro vicinato sia accettabile è con le prime 20 molecole presenti nel *Master Thesaurus*. Tra queste l'utente seleziona quelle fortemente correlate con il contesto di interesse e le aggiunge alle parole-seme ampliando il t*hesaurus* di quel sotto-dominio.

Possiamo considerare come esempio quello dello sviluppo del *thesaurus* relativo al sotto-dominio "misure d'emergenza in risposta ad un incidente industriale". Partiamo con le seguenti 8 *parole-seme*:

AMBULANCE	EVACUATION	SHELTERING	POLICE
CITY	SCHOOL	LOUD-SPEAKER	TRAFFIC

Su questo ristretto set di parole lo strumento per le *query* genera i seguenti 60 "vicini significativi":

INTERVENTION	CENTER	RIVER	NUMBER	LOCATION	ACTIVATION
MOBILIZATION	HIGHWAY	PHASE	MEASURE	LIVING	INTERRUPTION
RAILWAY	RECEIVING	PUBLIC	RECORDING	EXTENSION	BUS
UNIVERSITY	EXTERNAL	LARGE	INFORMATION	AUTHORITY	KEEP
DECISION	RELATIVES	SITE	HELICOPTER	SPILLAGE	ACCOMODATION
ACCOMPANIED	REQUIRING	STOP	WARNING	FACTORY	EMERGENCY
DIRECTION	INDUSTRY	TOTAL	INDICATION	DEVIATION	NEIGHBOUR
COMMUNITY	SERVICE	DANGER	CONNECTION	VEHICLE	LOCALISATION
INSTALLATION	PERSONNEL	STREET	MEMBER	CALLING	PREMISE
PRECAUTION	TOWARDS	ANNEX	FIREMAN	RAIL	HOSPITALISATION

Poiché la maggior parte di queste parole sono già abbastanza significative, l'utente può interrompere il processo di selezione, individuare in questa lista le parole particolarmente appropriate ed includerle nella precedente selezione delle parole-seme aggiungendole così nel *thesaurus* relativo al sottodominio "misure d'emergenza".

Per l'ulteriore ampliamento del *thesaurus* con le molecole pertinenti tra quelle nel *Master Thesaurus*, vengono generate dello strumento frasi composte da due o più parole appartenenti al "vicinato" delle parole chiave del dominio in analisi e, in base al valore di EMIM e alla loro appropriatezza nel contesto, l'utente seleziona quelle da aggiungere al *thesaurus*.

In questa fase l'attenzione dell'utente esperto è fondamentale al fine di evitare l'inserimento di molecole non significative per le quali però risulta buono il valore di EMIM. Si è verificato comunque che la percentuale di combinazioni non significative è molto bassa.

A questo punto il *thesaurus* viene strutturato in maniera semi-automatica secondo le scelte fatte dall'utente ed è pronto per analizzare gli incidenti e selezionare per ciascun evento quelle porzioni di descrizione che rientrano nel contesto individuato per lo studio.

3. CONCLUSIONI

Abbiamo illustrato l'approccio terminologico basato su *thesaurus* per il recupero di informazioni dal database degli incidenti industriali e ne abbiamo fornito esempi di costruzione. Parte del lavoro è sistematica ma altra richiede particolari conoscenze dell'inglese e del MARS stesso. Si è visto anche che lo sviluppo effettivo del *thesaurus* è stato ben più veloce della fase di progetto, alla fine è però risultato un metodo utile visto il tipo di applicazione e di necessità.

Ciononostante l'approccio illustrato presenta dei problemi, ad esempio l'ambiguità generata dal linguaggio tecnico e l'importanza di inquadrare esattamente le parole nel contesto. La verifica con indicatori statistici non serve a migliorare la qualità ma solo a mantenerla di buon livello mentre lo sforzo maggiore dovrebbe essere fatto in Comunità Europea con le verifiche di qualità dei dati per accrescerne la consistenza all'interno dei vocabolari costruiti all'interno del database.

È evidente però che la sistematizzazione delle informazioni contenute nel MARS, la possibilità di estrarle ed aggiornale in modo diretto ma flessibile, legato cioè al tipo di utente che interroga il database, e la potenza di un simile strumento costituiscono una base per esportare e diffondere questa esperienza negli ambiti di contorno all'evento incidentale.

Costruire cioè un modo univoco di comunicare ai diversi livelli informativi è diventata esigenza primaria in qualunque settore, e più in particolare in quello legato agli incidenti rilevanti poiché coinvolge oltre che esperti di diversa estrazione, anche la popolazione. Lo scambio delle conoscenze avviene attraverso diversi canali ma sempre sullo stesso argomento: dall'informazione alla popolazione ai rapporti tecnici a regime o a seguito dell'incidente, dalla pianificazione dell'emergenza allo studio statistico finalizzato degli scenari incidentali. Per ciascuno di questi aspetti lo strumento terminologico basato su un vocabolario controllato che sia anche riconosciuto come univoco permette di ritrovare, organizzare e presentare l'informazione in maniera sempre coerente e corretta ma con una particolare attenzione alle esigenze di chi la sta ricercando o la deve ricevere.

4. SVILUPPI FUTURI

Presi in considerazione i risultati soddisfacenti e promettenti, il progetto di collaborazione con il DG-JRC di Ispra mira a risolvere almeno in parte i problemi incontrati e studiati nello sviluppo dell'attuale software di gestione del MARS e dello strumento specifico per le *query* basato sull'uso del *thesaurus*.

L'attività ha due obiettivi principali: uno riguarda lo sviluppo dei *thesauri* e l'altro è relativo alla realizzazione di un nuovo modulo multilingua nel software di gestione del MARS.

Sulla prima parte si sta lavorando all'aggiornamento ed al completamento dei *thesauri* già implementati, e alla realizzazione di nuovi *thesauri* su altri sotto-domini individuati con il criterio della scala di interesse basata sulle richieste più frequenti da parte degli esperti del settore e degli utenti che utilizzano lo strumento delle *query*.

Un aspetto di particolare interesse sarà il riferimento incrociato tra le definizioni presenti nella normativa specifica e quelle presenti nei *thesauri* associati al MARS in modo da risolvere l'ambiguità dovuta ad un uso non concordato di espressioni tecniche. In questo modo dal dato memorizzato si potrà estrarre il contenuto informativo implicito e di carattere tecnico-specialistico da attribuirgli a seconda dello scenario che si vuole ricostruire.

Spesso infatti, lo scenario finale si compone di valutazioni provenienti dalle persone coinvolte che hanno però differenti professionalità e quindi modi differenti di esprimersi, e riguardo al MARS questo capita nella compilazione dei rapporti d'incidente che diventano poi oggetto del nostro studio.

Tali espressioni possono generare ambiguità di interpretazione se non è individuato il significato univoco da attribuire ai possibili frasari tipici di un contesto, con particolare riferimento a valutazioni di tipo qualitativo e quantitativo del rischio contenute nelle espressioni usate.

La normativa rappresenta una prima fonte di uniformità in quanto spesso contiene al suo interno delle definizioni sui termini ricorrenti nel testo e che si riferiscono all'oggetto stesso della norma. Il carattere è molto generale e di certo non sono molti i termini presentati con la definizione però il lavoro degli esperti del settore in parallelo al lavoro di tipo terminologico permetterà di raggiungere risultati di sicuro valore anche per l'interesse dimostrato in Comunità Europea nei confronti dei gruppi di lavoro su standard terminologici nei diversi settori (medicina, ambiente, gestione del rischio, etc.).

Il secondo obiettivo è lo sviluppo di un modulo multilingua da inserire nel sistema, al fine di rendere indipendente dalla lingua d'origine l'informazione contenuta nel testo. La lingua di riferimento continuerà ad essere l'inglese e si realizzeranno delle traduzioni 'intelligenti' sia dei descrittori, comprensivi di sinonimi e varianti lessicali, facenti parte dei *thesauri*, sia delle regole di associazione che terranno in conto la

grammatica propria di ciascuna delle lingue scelte per un primo esperimento. La registrazione dei nuovi rapporti così come il recupero dell'informazione su quelli già archiviati sarà quanto più possibile libera dalle ambiguità tipiche delle traduzioni di tipo tecnico.

Anche in questo caso ci viene incontro la normativa che costituisce la base di partenza poiché per il recepimento è riportata in ciascuno stato nella propria lingua. Ricordiamo che i rapporti d'incidente vengono spediti alla Commissione Europea per lo più nella lingua dello stato membro dove l'incidente si è verificato e quindi sono soggetti a traduzione prima di essere caricati nel database. Questo comporta problemi di ambiguità legati alle espressioni tipiche presenti in ciascuna lingua che rischiano di essere fraintese nella traduzione.

Partendo dall'esperienza accumulata in anni di raccolta dei dati relativi agli incidenti rilevanti si può stendere una prima versione del modulo multilingua in modo da sollecitare poi i feedback degli altri stati a valutare la bontà di traduzione. Si possono prendere accordi sul significato delle espressioni nelle diverse lingue rispetto alla versione inglese creando dei traduttori automatici 'intelligenti' che siano in grado di riversare dalla lingua originale all'inglese i rapporti d'incidente risolvendo almeno in parte le ambiguità di traduzione.

5. BIBLIOGRAFIA

[1] Chen H. and Ng T. (1995). An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): Symbolic branch-and-bound Search vs. Connectionist Hopfield Net Activation. *Journal of the American Society for Information Science* 46(5) 348-369.

[2] Chung P.W.H. and Jefferson M. (1998). Accident databases – indexing and retrieval. In *Lessons Learnt from Accidents* 139-146 EUR 17733 European Commission.

[3] Church K.W. and Hanks P. (1989). Word association norms mutual information and lexicography. In *Proceedings of the 27th Annual meeting of the Association for Computational Linguistics* 76-83.

[4] Kirchsteiger, C. (ed.), *Proceedings of a seminar on lessons learned from accidents*, Linz, October 1997, European Commission, DG JRC, 1998.

[5] Galeazzi E, Agnello P, Gangemi A, Niinimäki J, Pakarinen V, Rossi Mori A. What is a medical term? Terms and phrases in controlled vocabularies and continuous discourses. *Proceedings of MIE* 94, Lisbon

[6] Kirchsteiger, C., Rushton, A.G. & N. Kawka, Contribution of human errors to accidents notified to MARS, *presented at Lessons Learned From Accidents*, Linz, 1997 (available from EC, DG JRC, Ispra).

[7] Official Journal (OJ) of the European Communities, Council Directive 96/82/EC of 9 December 1996 on the control of major-accident hazards involving dangerous substances ('Seveso II'), Luxembourg, 1997.

[8] C. Kirchsteiger, Using Modern Database Concepts to Facilitate Exchange of Information on Major Accidents in the European Union, *Proceedings of the ESREL '97 International Conference on Safety and Reliability*, Lisbon, Portugal, June 1997.

[9] C. Kirchsteiger, N. Kawka, Identification of significant recurrent patterns in accident descriptions, *Proceedings of the "60th American Power Conference",* Illinois Institute of Technology, Chicago, Illinois, April 14-16, 1998