A. ZAMPOLLI

(Dipartimento di Linguistica - Università di Pisa
Istituto di Linguistica Computazionale - CNR)

RISORSE LINGUISTICHE RIUTILIZZABILI PER LA RICERCA IN INGEGNERIA LINGUISTICA E PER L'INDUSTRIA DELLE LINGUE

SCHEMA SEMPLIFICATO DI UN TIPICO PRODOTTO PER L'INDUSTRIA DELLE LINGUE

RISORSE LINGUISTICHE (RL)

CONOSCENZE, DATI DESCRIZIONI DI UNA LINGUA A VARI LIVELLI

- CORPORA
- LESSICI
- TERMINOLOGIE

- GRAMMATICHE
- STRUMENTI PER L'USO

ANALIZZATORE (AN)

MECCANISMO PER ANALIZZARE IL TESTO DI INPUT UTILIZZANDO LE CONOSCENZE FORNITE DALLE RL

INPUT: "TESTO"

RF

RAPPRESENTAZIONE FORMALE DEL "TESTO" A DETERMINATI LIVELLI DI DESCRIZ. E INTERPRETAZIONE

SFO

SISTEMA APPLICATIVO OPERA SULLA RF PER ESEGUIRE COMPITI APPLICATIVI SPECIFICI

RISULTATI DELLA APPLICAZIONE

INTERFACCIA UTENTI

UTILIZZATORI

IL RAPPORTO ALPAC (1966)

- BRUSCO ARRESTO DI QUASI TUTTI I PROGETTI DI TA
- RACCOMANDAVA LA COSTRUZIONE DI ESTESE RL

(corpora, lessici, grammatiche) (monolingui e constrastive)

PER LO SVILUPPO DELLA LINGUISTICA COMPUTAZIONALE
COME DISCIPLINA AUTONOMA (PRIMO USO DI QUESTO TERMINE)

LA LC INVECE, IN PARTICOLARE SOTTO L'INFLUENZA
DELLE SCUOLE LINGUISTICHE NORDAMERICANE E DELLA AI,
HA CONCENTRATO GLI SFORZI ESCLUSIVAMENTE
SULLO STUDIO DI MODELLI FORMALI, FOCALIZZANDOSI SU
FENOMENI LINGUISTICI E COGNITIVI ISOLATI,
"INTERESSANTI" PER LO STUDIO DELLE PROPRIETA' FORMALI,
COMPUTAZIONALI, COGNITIVE DEI MODELLI,
MA NON NECESSARIAMENTE I PIU' RILEVANTI PER IL TRATTAMENTO
DI DATI LINGUISTICI "REALI".

DAL 1986

L'EMERGERE DEL PARADIGMA DELL'INDUSTRIA DELLE LINGUE,
CHE RICHIEDE IL TRATTAMENTO DI USI "REALI" DELLE LINGUE, HA COSTRETTO
LE COMUNITA' DELLA ReS A CONSIDERARE LA NECESSITA' DI:

COMPONENTI ROBUSTI
BASATI SULL'EVIDENZA DELL'USO REALE (analisi di corpora)
RINFORZATI DA CONOSCENZE STATISTICHE (analisi di corpora)
AVENTI ACCESSO A LESSICI MULTIFUNZIONALI ESTESI E
A GRAMMATICHE INGEGNERIZZATE MODULARI RIUTILIZZABILI

Di fatto, in passato, per ogni nuova applicazione, anche all'interno della stessa organizzazione e perfino per versioni aggiornate dallo stesso sistema,

un nuovo lessico e una nuova grammatica venivano creati ripartendo da "zero"

Ci si è resi conto che la realizzazione iterativa di RL adeguate

- è troppo costosa e inefficiente: la duplicazione degli sforzi non può essere sopportata neppure da grandi industrie

- rallenta lo sviluppo, l'aggiornamento, il miglioramento dei prodotti

- impedisce la concentrazione degli sforzi su problemi cruciali per l'avanzamento scientifico e tecnologico

Decisivo è stato il workshop da noi organizzato nel 1986 per conto della CEE a Grosseto, dove per la prima volta sono stati riuniti assieme linguisti computazionali, linguisti, AI, case editrici, industrie di vario tipo

che ha introdotto il concetto di

RISORSE LINGUISTICHE MULTIFUNZIONALI RIUTILIZZABILI

e ha affermato che la loro costruzione deve avere la massima priorità

RIUSABILITA' E MULTIFUNZIONALITA' DELLE RL

Due aspetti complementari:

- Uno Relativo al Passato

  riutilizzare RL esistenti, estraendone informazioni esplicitamente o implicitamente presenti da incorporare in nuove risorse linguistiche multifunzionali

  es. estrazione di informazioni semantiche dalle definizioni di MRDs (es. ACQUILEX)

- Uno Relativo al Futuro

  stabilire nuove, estese RL disegnate con le proprietà di essere MULTIFUNZIONALI, cioè tali da poter servire, attraverso interfacce appropriate, una larga varietà di applicazioni presenti e future e i moduli linguistici in esse incorporati, anche se usano frameworks teorici o computazionali diversi.

Il cosiddetto "Paradigma Compilativo"
(International Pisa Group 1987-89; ET7)

Per es. per il lessico:
osservazioni multiple di fenomeni linguistici nell'uso reale di una lingua

Basi di Conoscenza Lessicali
Informazioni di base: differenze minime osservabili

estrazione selettiva, integrazione, riformattamento

Lessici Computazionali per Applicazioni Specifiche

NECESSITA' DI UNA STRATEGIA COMPLESSIVA PER LE RL IN EUROPA

La disponibilità di RL adeguate e riutilizzabili è una condizione "sine qua non"

per il progresso dello stato dell'arte in NLP e Speech

per lo sviluppo di una tecnologia linguistica adeguata

per il decollaggio di una LI reale

Le RL fanno parte della infrastruttura di base:
sono troppo costose per essere sviluppate dal solo settore privato
(in particolare le SME)

Richiedono continuità di azione, oltre i limiti di un singolo programma

Anche se molti sforzi e denari sono stati investiti per le RL in Europa, ci sono ancora molte necessità non soddisfatte

Ci sono 3 gruppi principali di utilizzatori delle RL:

a) ricerca; b) formazione; c) sviluppo di sistemi e prodotti

Nessuno di questi gruppi è stato adeguatamente servito.

Sono i "developers" di sistemi i più severamente handicappati.

Fattori che hanno determinato la insoddisfacente situazione presente:

- **frammentazione e duplicazione di sforzi dovute a insufficiente coordinamento**
- **la maggior parte delle RL costruite fin qui sono state configurate per scopi specifici**
- **mancano standards che assicurino la riusabilità**
- **la dimensione delle RL necessarie, che deriva dalla pluralità delle lingue in Europa**

RL adeguate devono essere sviluppate e mantenute per tutte le lingue europee.

Lo sviluppo di RL, in Europa, deve essere coordinato centralmente per assicurare riutilizzabilità, compatibilità, connessioni multilingui.
Solo la CEE può assicurare questo coordinamento, in collaborazione con gli stati membri.

I due altri maggiori blocchi economici (USA, Giappone) hanno già preso delle misure per rendere disponibili le RL al loro interno, e cominciano a interessarsi alle lingue europee

E' necessario e urgente
definire e implementare una strategia europea globale per

- **rendere disponibili alla ReS Europei le RL infrastrutturali di cui necessitano**
- **assicurare che esse siano riutilizzabili, multifunzionali, condivisibili, disponibili nel "pubblico dominio"**
- **coordinarle centralmente a livello multilingue**

A mio avviso,
solo la CEE può assumere la responsabilità principale per promuovere le azioni necessarie e determinanti per la creazione e distribuzione di tali RL
nel 4° Programma Quadro

Alcune azioni preparatorie sono in corso, in LRE, ESPRIT, MLAP e sono in gran parte coordinate da partners italiani.

E' essenziale costruire sui loro risultati

La fase di definizione

La definizione di una strategia europea globale comprende vari aspetti:

Tecnico-Scientifici. Per es.

- **Metodi per riutilizzare RL esistenti (ESPRIT-BRA, ACQUILEX)**
- **Specifiche comuni: standards di fatto, basati sul consenso.**

sono le condizioni essenziali per la creazione di nuove RL multifunzionali multilingui (LRE-EAGLES)

Organizzative. (LRE-RELATOR)

Quali RL esistono per le varie lingue e valutare la riutilizzabilità
Identificare gli attori potenziali per la costruzione/fornitura di RL
Disegnare e sperimentare modelli di strutture europee coordinate per la produzione e disseminazione di RL
Considerare i costi ed esplorare possibili fonti di finanziamenti
Considerare possibili modelli di relazione con paesi non CEE
Creare un nucleo iniziale di RL per tutte le lingue europee per es:

- **testare le specifiche**
- **assicurare la adozione di specifiche e standards comuni**
- **promuovere lo stesso trattamento per tutte le lingue europee**

Definire il framework giuridico per la distribuzione delle RL
Definire i ruoli e le responsabilità delle Autorità Nazionali e Internazionali, le Istituzioni pubbliche, il settore privato.

EXPERT ADVISORY GROUP FOR LANGUAGE ENGINEERING STANDARDS (EAGLES)

OBIETTIVI

produrre specifiche per le RL con il consenso dei principali progetti europei del settore, per supportare lo sviluppo, a lungo termine, di standards nazionali e internazionali, incoraggiando la ricerca e l'industria europea ad agire in modo concertato.

RUOLI CHIAVE

Supervisione			
Allocazione di fondi	Managing Board		
Interfacce MB/Hosts	Coordinatore		
Produzioni di rapporti e guidelines	Supporto Amministr.	<——————>	CEC
	Redattori	Contratto	

ESECUZIONE DEL LAVORO SCIENTIFICO LOGISTICO

Corpora Testuali	Lessici per NLP	Formalismi per LC	Assessment e valutazione	Linguaggio Parlato
Host	Host	Host	Host	Host
Chairman	Chairman	Chairman	Chairman	Chairman
Esperti	Esperti	Esperti	Esperti	Esperti

Metodologie e specifiche comuni per la creazione gli scambi di RL, quali corpora, lessici, grammatiche

Valutazione e assessment della qualità di sistemi e componenti per il NLP e lo speech

Formalismi di alto livello per la manipolazione della conoscenza per lo speech e il NLP.

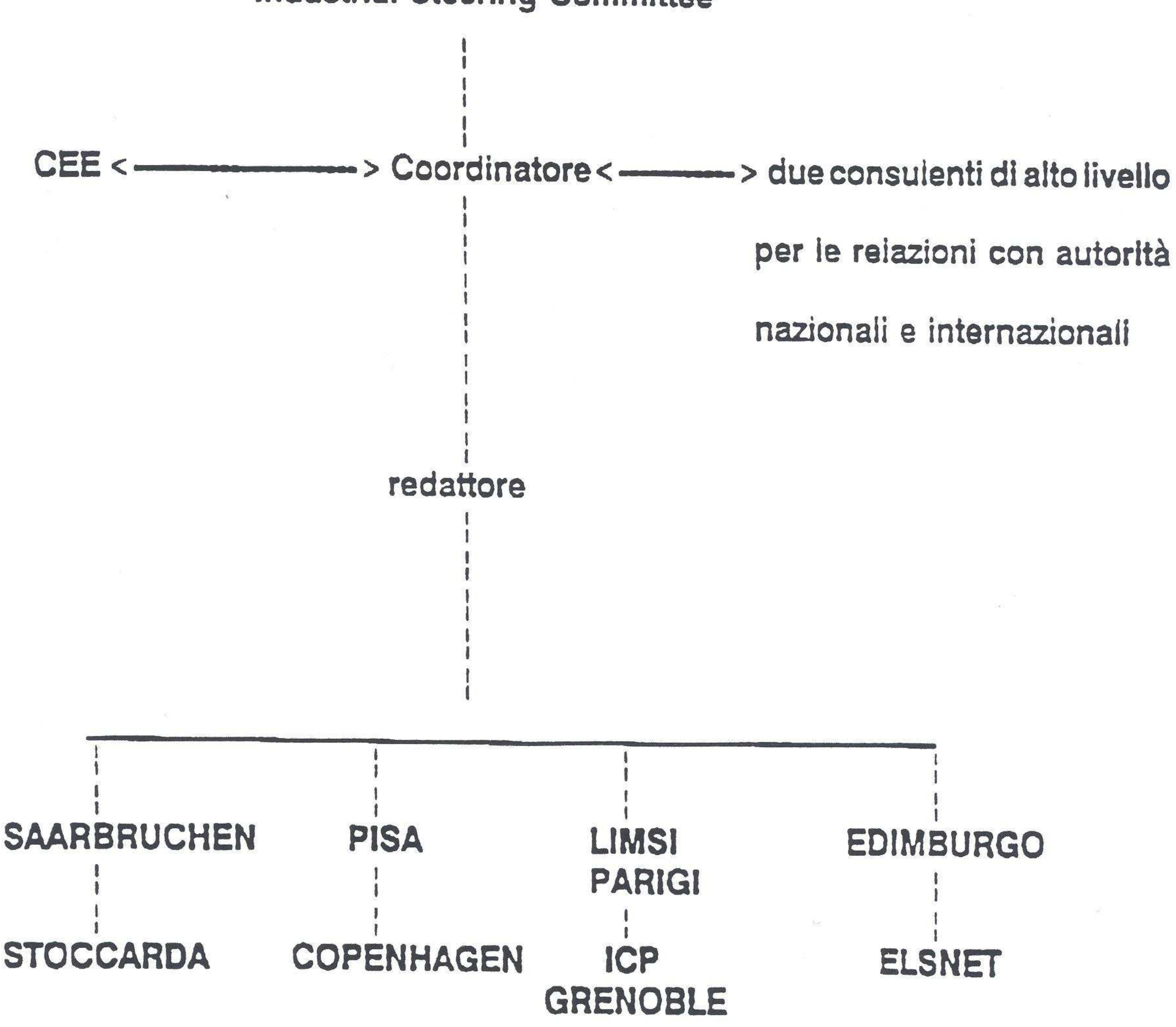
RELATOR

Obiettivi e scopi. Contribuire a:

definire un framework organizzativo integrato per la creazione delle RL che sono necessarie per lo sviluppo di una tecnologia linguistica adeguata e dell'industria delle lingue

creare e sperimentare un network europeo integrato di "depositi" di RL, con la funzione di catalogare, raccogliere, conservare, distribuire tali risorse

Struttura Organizzativa



COLLABORAZIONE CON ISTITUZIONI USA

OBIETTIVI:

Convergenza degli sforzi

Promuovere il consenso internazionale sul lavoro prenormativo europeo

Potenziare le strategie di ricerca

LINEE DI AZIONE

- Survey dello stato dell'arte nel NLP e nello Speech
  (NSF - CEC)
- Cooperazione con la Text Encoding Initiative (ACL - ACH - ALLC)
- Corpus multilingue sperimentale parallelo (analogo a TREC/TIPSTER)
- Definizione di una strategia europea per i rapporti con USA nel settore delle RL
- Coordinazione scientifica Internazionale (LIRIC, COOSDA, ecc.):
  promuovere la collaborazione a livello internazionale tra le comunità dello speech e del NLP
  Stabilire delle strutture congiunte per la collaborazione internazionale