

Trattamento Digitale delle lingue al servizio delle Lingue Meno Usate

Carlo Zoli

Smallcodes - Open Lab

Carlo Zoli, ingegnere elettronico, è nato a Firenze nel 1969, è sposato, ha tre figli e si occupa di internet e di tecnologie della comunicazione dal 1994. Inizialmente si occupa di informatica industriale per la grande impresa; è tra i fondatori di Open Lab, una casa di produzione software piuttosto nota in campo internazionale, vincitrice del prestigioso Jolt Award della Silicon Valley. Ma dal 2003 si dedica unicamente all'informatica linguistica, ed in particolare all'uso delle nuove tecnologie come potente strumento di difesa e promozione della diversità linguistica e culturale. Ha fondato Smallcodes, una rete di condivisione di tecnologie digitali linguistiche di alto livello, che si propone di dotare le lingue di recente normalizzazione di quegli strumenti che sono considerati indispensabili nelle grandi lingue nazionali (dizionari elettronici di riferimento, correttori, traduttori, sistemi di scrittura predittiva, ecc.). Smallcodes collabora con le più importanti cattedre universitarie che si occupano di politica linguistica e linguistica computazionale (Milano Bicocca, Siena, Genova, Piemonte Orientale, ecc.), con centri di ricerca privati (il CELE principalmente), con i principali istituti culturali-linguistici e con molti centri di politica linguistica italiani ed europei. Forse tra molti anni riuscirà a prendere il suo tanto ambito dottorato di ricerca in linguistica storica.

Abstract

Smallcodes è una sistema software web-based per la linguistica computazionale, l'ingegneria linguistica e la lessicografia, con un interesse specifico alle lingue regionali e minoritarie, agli studi dialettologici e alla linguistica storica.

Si compone di quattro moduli integrati tra loro

- Lessicografia e dialettologia computazionale, con un'attenzione particolare alla variazione diatopica, diastratica, diacronica, e alle lingue "meno usate" e "non standardizzate".
- Lessicografia e dialettologia settoriale, per la creazione di dizionari tematici, specialistici, terminologici multi-lingui.
- Strumenti di aiuto alla scrittura (correzione ortografica e grammaticale), con caratteristiche uniche nell'individuare i pattern di errore dovuti al background dialettale degli scriventi e alle interferenze tra il codice maggioritario - lingua tetto e la lingua minoritaria / dialetto.
- Strumenti di analisi e spoglio di corpus per analisi statistiche, stilistiche, per il lavoro di lessicografia, neologia e terminologia tecnica.

La Tecnologia Smallcodes ha come obiettivo quello di colmare il divario che spesso si osserva tra l'alta qualità della ricerca linguistica e socio-linguistica di ambito accademico e le corrispondenti realizzazioni di applicazioni software e di banche-dati le quali, spesso, sono impostate senza rispettare adeguati stan-

dard industriali di qualità, riusabilità, interoperabilità, manutenibilità del software. Si propone, secondo la filosofia del codice aperto, di rappresentare uno standard di livello internazionale nelle applicazioni di dialettologia computazionale e di diventare un formato il più possibile universale di interscambio e confrontabilità dei dati linguistici.

Smallcodes ha in questo momento diversi progetti attivi, per grandi banche dati dei dialetti ladini, cimbri, mocheni, sardi, walser, occitani, trentini, lombardi della Svizzera Italiana etc. Smallcodes è stata finanziata da numerosi enti di ricerca e enti pubblici tra cui l'Istituto di Dialettologia ed Etnografia della Svizzera Italiana.

Affianca Smallcodes Technology una Fondazione Smallcodes (www.smallcodes.org), senza fini di lucro, che si propone di finanziare e sviluppare progetti per colmare il digital divide nel sud del mondo, con la convinzione che le tecnologie digitali siano una grande occasione per la ricerca, il mantenimento e la promozione della diversità linguistica e culturale.

1. INTRODUZIONE

1.1. Manifesto

- Crediamo che la diversità culturale sia un valore da difendere.
- Crediamo che una delle strategie concrete per la difesa delle culture tradizionali, minoritarie, indigene, sia quella di modernizzarle, tentando anche di salvarne alcune specificità.
- Crediamo che il *folklorismo*, e la *laudatio temporis acti*, così come i progetti che mirano **soltanto** al recupero degli aspetti della cultura tradizionale - allo scopo di esporla in qualche forma di museo - senza contemporaneamente dotare queste culture e queste comunità di strumenti propri per affrontare la modernità, siano, in realtà, nemici (magari involontari e in buona fede) di queste culture minori.
- Crediamo che ogni progetto serio di difesa culturale debba partire dalla difesa della lingua; e che la modernizzazione debba passare attraverso una forma scritta, il più possibile coerente e condivisa.
- Crediamo che le tecnologie digitali possano contribuire in modo decisivo al processo di modernizzazione linguistica e a quello della promozione e diffusione presso le generazioni più giovani.
- Crediamo che da un punto di vista tecnologico sia necessario il più alto grado possibile di uniformità negli standard di memorizzazione, salvataggio di testi, formati di banche-dati. Questo per garantire una "lunga vita elettronica" alle risorse linguistiche e per permettere un reale e facile scambio di informazioni, risorse, tecnologie.
- Crediamo che le conoscenze debbano avere la più ampia circolazione possibile, e che in nessun senso si debba limitare l'accesso al patrimonio culturale dell'umanità; e questo in particolare per le risorse linguistiche, essendo forse proprio il linguaggio la caratteristica più distintiva dell'essere umano.

1.2. I fini del progetto *smallcodes*

Visti gli assunti espressi nel "manifesto", questi i fini del nostro progetto

1.2.1. Raggiungere l'eccellenza, arrivando a costituire un punto di riferimento per lo stato dell'arte a livello globale nelle tecnologie digitali per le risorse linguistiche, con

particolare riferimento alle lingue meno usate.
per raggiungere questo scopo, è necessario:

1.2.2. Unire le forze intellettuali, finanziarie, organizzative, di tutte le comunità linguistiche prima italiane, poi europee, poi extra-europee;

Non disperdersi in iniziative isolate, che rischiano di riprodurre errori già commessi e non accolgono spesso le best-practice, perché non le conoscono;

Avvantaggiarsi dell'esperienza degli altri; e partire da dove gli altri sono arrivati, e non ogni volta da zero;

1.2.3. Venendo a ambiti più specificamente tecnici nell'ingegneria del software, lo scopo del progetto è creare una serie di applicazioni open-source, che nel tempo dovrebbero convergere verso un'unica banca dati e un unico macro-servizio on line di risorse linguistiche; queste applicazioni e queste banche dati utilizzeranno, dove esistono, gli STANDARD elettronici e di modellazione dei dati più largamente accettati e consolidati, in armonia con i principi appena espressi. Dove questi standard non esistano, si tenterà di crearli, e, con la forza della qualità e della bontà delle idee, di diffonderli.

1.3. Perché aderire

I motivi per aderire sono molti; citiamone alcuni oltre a quelli ovvi

- spesso nelle istituzioni di politica linguistica o di linguistica applicata i budget destinati alla tecnologia hanno un andamento periodico. Ad anni relativamente "ricchi" seguono periodi in cui il capitolo di bilancio relativi alle tecnologie dell'informazione sono modesti o nulli. Questo modello operativo (vedi dopo, il paragrafo 4.) consente a chi aderisce di vedere il progetto progredire anche durante questi periodi.
- L'esperienza accumulata negli anni e portata dagli aderenti consente a chi aderisce di avvalersi di un'esperienza (e di un investimento pregresso) che da solo non avrebbe potuto probabilmente fare, se non in tempi lunghissimi
- Solo collaborando concretamente tra loro le minoranze d'Italia e d'Europa possono aspirare ad una forza contrattuale reale nelle sedi istituzionali; e la collaborazione è stata finora una buona intenzione, ma spesso non seguita da fatti concreti. Small-codes è e sarà sempre di più un fatto concreto, e per di più in un settore, la società dell'informazione, che gode in questa fase, di "buona stampa", ma in cui ancora le *best practices* sono poco diffuse.

2. I MODULI DEL SISTEMA, IN BREVE

I moduli qui sotto citati sono in stato diverso di maturità e di perfezionamento dello sviluppo: molti sono già funzionanti e pienamente operativi, alcuni sono appena abbozzati. Ci limiteremo qui a citare i moduli per i quali lo sviluppo è in una fase di una certa maturità, tralasciando quelli che sono solo progettati ma non ancora neppure avviati (ad esempio il *trattamento del linguaggio naturale*, il *riconoscimento del parlato*, la *sintesi vocale*, sistemi di vocabolario o di assistenza all'inserimento di testo predittivo per telefonini, ecc.)

Com'è ovvio, tutto il progetto, essendo stato pensato specificamente per lingue "giovani" (in termini di ufficialità o co-ufficialità), mostra particolare attenzione a temi che sono ormai dati per scontati (o, fortunatamente per loro, superati) per le grandi lingue di

cultura standardizzate e normalizzate, ma per che per lingue minoritarie o in corso di normalizzazione sono di importanza cruciale.

2.1. Modello unificato dei dati lessicali

2.1.1. Modello generale del lessico (a oggetti e data base entità-relazioni)

Un'analisi approfondita dei sistemi di lessicografia computazionale, delle esigenze di campo dei lessicografi, e, ancora, degli standard esistenti (pochi, per la verità) nel campo della raccolta di lessico, specie per lingue meno usate e meno standardizzate, ci ha condotto ad una prima linea di sviluppo: il modello unificato di base-di-dati lessicale, applicabile per tutte e quattro le branche della lessicografia classica: lessico generale, lessico tematico/specialistico, lessicografia storica, lessicografia bilingue / multilingue.

Gioverà precisare che si tratta di un modello TEORICO, di una pura struttura di dati; non della sua implementazione in un programma utilizzabile dagli utenti (oggetto del successivo 2.2.). Si tratta solo di uno standard a cui qualunque software può/deve adeguarsi. La comunanza del modello dei dati garantisce l'interoperabilità e l'interscambio dei dati.

Ad oggi sono in corso implementazioni in varie tecnologie del nostro modello teorico: da Microsoft Access, a FileMaker, fino a implementazioni interamente su internet basate su Java.

È interesse di qualunque compilatore di dizionari adeguarsi al modello

- perché è il più completo in circolazione
- perché è abbondantemente provato, testato, raffinato
- perché permette di scambiare esperienze e modalità di lavoro con molte comunità linguistiche, ecc.

Al momento in cui scrivo hanno aderito a questo modello di dati per i loro lavori lessicografici (con l'applicazione illustrata in 2.2., o con altre applicazioni sviluppate in proprio):

- l'Ufitziu pro sa Limba Sarda di Nuoro (Limba Sarda Comuna)
 - la Regione Sardegna (Sardo)
 - l'Istitut Cultural Ladin Micurà de Rü (Ladino Gardenese, Ladino della Val Badia, Ladino Dolomitano std.)
 - l'Istitut Cultural Ladin Majon di Fascègn (Ladino Fassano, Ladino Dolomitano std.)
 - L'Istituto Culturale Mocheno (Bernstoler KulturInstitut, lingua Mòchena)
 - L'Istituto Culturale Cimbri (Institut Lusérn, lingua Cimbria)
- (si noti che tutte le tre minoranze storiche del Trentino sono attivamente partecipanti al progetto. La Provincia Autonoma di Trento, per bocca del suo presidente Dellai ha più volte pubblicamente affermato l'adesione dell'Ente come tale al progetto).
- La Chambrà d'Òc (Occitano Alpino)
 - l'Ufficio per le Minoranze Linguistiche Storiche della Regione Piemonte (Occitani, Francoprovenzali, Walser)
 - la Provincia di Udine (Friulano)
 - l'Espaci Occitan di Dronero (ufficio interprovinciale di promozione e politica linguistica, Occitano)

- la comunità di lingua Arumena di Romania e Bulgaria
- la rete di terminologia LinMiTer (gestita dall'Unione Latina di Parigi, www.unilat.org)
- la comunità di lingua Arumena di Romania e Bulgaria
- il Centro di Dialettologia ed Etnografia della Svizzera italiana (Bellinzona, Canton Ticino)
- la Lia Rumantscha (Coira, lingua Romancia)

2.2. Lessicografia computazionale

Una possibile implementazione del modello di dati sopra descritto, anzi l'implementazione di riferimento, è dato dall'applicazione lessicografica Verba che contiene i seguenti moduli per la lessicografia di campo o sul web.

Il tutto con la tecnologia di Internet, aperta per definizione, e che permette un lavoro di gruppo, con gruppi distribuiti geograficamente, con tempi rapidissimi, e che soprattutto non ha la rigidità (e i costi) degli strumenti "chiusi", come il CD Rom e il volume cartaceo, che una volta andati in stampa non sono modificabili fino all'edizione successiva. È chiaro che lo strumento cartaceo (dizionario normativo generale) non è sostituibile; ma in quest'ottica è un passo successivo che può avvenire solo dopo la piena affermazione e stabilizzazione del materiale del dizionario on-line.

Ecco i moduli in sintesi:

- 2.2.1. Lessicografia generale on line e off-line
- 2.2.2. Lessicografia locale, con varianti locali, dialettali, geosinomimi, variazioni ortografiche sub-standard
- 2.2.2. Lessicografia tematica/settoriale
- 2.2.3. Toponimia (repertorio dei nomi di luogo)
- 2.2.4. Antroponimia (repertorio dei nomi di persona)
- 2.2.5. Lessicografia storica
- 2.2.6. Lessicografia multilingue, lessici per la traduzione
- 2.2.7. Idee affini, sinonimi, coordinate stile e di affinità semantica

2.3. Correzione ortografica e adattamento varianti locali → variante standard

2.3.1. Correzione ortografica on – line e off line (sul web e tramite word-processor OpenOffice e Microsoft Office)

2.3.2. Correzione ortografica / adattamento di varianti locali verso uno standard-tetto.

2.3.3. Correzione ortografica adattativa

Vale a dire, il motore di correzione si comporta diversamente a seconda, ad esempio: o dell'argomento di cui tratta il testo, o del background linguistico-dialettale di chi scrive, o operando un confronto tra vocabolari e indici di frequenza dei termini in un corpus.

2.4. Analizzatori di testi, concordanze e occorrenze, strumenti per la raccolta di corpus

2.4.1. Motori di ricerca testuali

2.4.2. Analizzatore di frequenze e di contesti

2.4.3. Editing on line di corpus

2.4.4. Strumenti di spoglio del corpus per compilazioni lessicografiche

2.5. Strumenti per l'insegnamento delle lingue

- 2.5.1. Creazione rapida dei contenuti multimediale per l'insegnamento
- 2.5.2. Creazione rapida dei test di autovalutazione e degli esami di sbarramento
- 2.5.3. Ingresso libero o con password/pagamento
- 2.5.4. Statistiche d'uso e di profitto

2.6. Strumenti per la pubblicazione di contenuti su internet, periodici on line, ecc.

- 2.6.1. Creazione rapida di portali internet auto-gestiti
- 2.6.2. Creazione rapida periodici elettronici con esportazione verso formati da stampa
- 2.6.3. Integrazione dell'editoria elettronica con gli strumenti di analisi del corpus

2.7. Terminologia/Neologia/Industria della lingua – strumenti per la ricerca terminologica

- 2.7.1. Sistemi di lavoro di gruppo on line per la creazione di neologia e terminologia tecnica
Commissioni di lavoro, flussi di approvazione, revisioni successive, submission di proposte, gestione delle motivazioni nella proposta di un termine, ecc.

2.7.2. Sistemi multilingue di confronto di terminologia per ambiti

Sistemi per la creazione/consultazione rapida di lessici multilingue, per confronti e creazione congiunta di terminologia tecnica in diverse "simili" o "politicamente affini" (esempio: gallese e bretone, corso e occitano, ladino e friulano, e così via)

2.8. Toponomastica – strumenti per la ricerca sul campo

- 2.8.1. Schede di rilevazione sul campo, integrazione con GIS
- 2.8.2. Geo-socio-linguistica [Dell'Aquila/Iannaccaro]

2.9. Sistemi di scrittura facilitata

- 2.9.1. Sistemi di "testo predittivo" per telefonini e computer palmare

Sia in modalità inserimento di testo a 9 tasti, sia in modalità auto-completante per tastiere complete.

3. L'OBIETTIVO FINALE DI INTEGRAZIONE

3.1. Che cosa sono le risorse linguistiche

Per *risorse linguistiche* elettroniche si intende il complesso di dati relativi ad una lingua disponibili in forma elettronica per l'elaborazione con calcolatori. Ad esempio: banche dati lessicali mono- e multi-lingui, banche dati terminologiche, banche dati di testi, file audio per la pronuncia o per il parlato naturale, statistiche di frequenza e concordanza, ecc.

3.2. L'obiettivo finale

L'obiettivo finale del progetto è quello di rendere disponibili (on line, e, forse, – marginalmente- off-line) tutte le risorse linguistiche di una lingua in maniera unificata e coerente, e, soprattutto, pubblica e libera.

Per far un esempio, con la stessa facilità con cui oggi chi usa Microsoft Word, premendo il tasto destro del *mouse* su una parola "sbagliata", ottiene i suggerimenti ortografici, in futuro, per quella parola, ottenere con la stessa facilità e immediatezza:

- i sinonimi
 - la frequenza d'uso (magari diversa a seconda dell'argomento)
 - i contesti d'uso
 - le citazioni d'uso da un corpus (magari limitato per periodo, per autore, per argomento)
 - le varianti locali / dialettali
 - la pronuncia standard (o dialettale) in formato audio
 - le declinazione / coniugazione
 - suggerimenti di stile per scrittura creativa (le cosiddette "idee affini")
 - le traduzioni in altre lingue
 - le accezioni, gli "atomi di significato" (come ipotizzato a Princeton dagli autori del progetto WordNet)
- (e molti altre risorse che oggi è forse difficile anche immaginare)
- vocabolari in formato adeguato per inserimento in sistemi di correzione e testo predittivo.

4. IL MODELLO TECNICO-ECONOMICO DEL PROGETTO

4.1. Software libero e gratuito

Il progetto *smallcodes* vuole coniugare l'eccellenza tecnologica nell'ambito delle applicazioni informatiche con le esigenze e le istanze di difesa delle lingue e culture minacciate, e più in generale della normalizzazione e dell'ingegneria linguistica. Esso si ispira al modello del software aperto e libero; in sintesi

- 1) è, e deve essere, in costante miglioramento, su tre assi principali:
 - a) NUOVE FUNZIONALITÀ
 - b) MAGGIORE USABILITÀ E DOCUMENTAZIONE D'USO
 - c) AGGIORNAMENTO E MIGLIORAMENTO TECNOLOGICO (velocità, sicurezza, affidabilità, portabilità, compatibilità, ecc.)
- 2) Chi accede all'uso del sistema lo utilizza nello stato in cui si trova in quel momento, e può (e moralmente **deve**) contribuire, secondo le sue possibilità, al suo miglioramento (nelle direzioni sopra specificate), in termini
 - a) di contributo finanziario per nuovi sviluppi, che poi ricadranno progressivamente su tutta la comunità di utenti
 - b) di contributo tecnico di programmazione, secondo gli standard del progetto (per la verità piuttosto complessi)
- 3) Il progetto deve essere considerato UNICO, anche se in realtà si compone, e sempre di più si comporrà, di diversi moduli, non necessariamente tutti attivi o tutti di interesse per un certo gruppo di utenti

4.2. Gratuito non significa che non ci siano costi di sviluppo

È solo apparentemente un'ovvietà il fatto che lo sviluppo di software di alta qualità e di alta complessità, anche senza perseguire il fine di lucro, è un'attività costosa, in particolare perché richiede risorse umane altamente qualificate; e per dare lunga vita e stabilità al

progetto non può ragionevolmente basarsi solo sui contributi dei volontari che potrebbero sviluppare solo fuori del loro normale orario di lavoro.

Perciò questo modello economico sposta lo sforzo di reperimento di fondi dal campo puramente commerciale a quello, non meno difficile, del fund raising e della ricerca di partner, finanziatori, sponsor. Naturalmente questo garantisce maggiore imparzialità, maggiore trasparenza, e permette a chi ha le competenze di collaborare o anche di utilizzare il sistema (al punto di sviluppo in cui si trova) senza investimenti rilevanti o, addirittura, senza alcun costo.

4.3. La manutenzione

Un discorso analogo a quello fatto per lo sviluppo può essere fatto per la manutenzione del sistema.

Per manutenzione si intendono quelle attività di correzione di difetti, di assistenza al corretto uso e di mantenimento – aggiornamento di installazioni del sistema presso enti che hanno aderito al progetto.

Una cosa è dire che chi ha aderito al progetto ha diritto alle migliorie successive e finanziate da altre entità; un'altra è sostenere i costi vivi dei tecnici che installano le nuove versioni, convertono i dati tra una versione e l'altra dove necessario, mantengono la compatibilità tra versioni, ecc., valutano quali migliorie successive sono eventualmente utili o in conflitto con quelle esistenti, gestiscono ordinariamente il buon funzionamento dei server, ecc. ecc.