Elisabetta Gola
Universita' degli Studi di Cagliari
elisabetta.gola@docenteonline.it

Stefano Federici
Cleverbytes s.r.l.
stefano.federici@cleverbytes.it

Per un atlante linguistico informatizzato della Sardegna

1.1 Introduzione

Da quando le nuove tecnologie della informazione e della comunicazione sono state rese disponibili in forma massiva nelle industrie e negli istituti di ricerca i campi di indagine e applicazione che ne sono stati permeati hanno cambiato inevitabilmente volto. Da questo punto di vista, la linguistica non fa eccezione[1]. Essa non fa eccezione nemmeno per quel che riguarda una distinzione che vale per tutte le aree investite dall'innovazione tecnologica: quella tra l'automazione che semplicemente accelera, ottimizza e rende realizzabili da macchine, invece che da esseri umani, processi di elaborazione di varia natura (simbolica o meccanica) e l'informatizzazione che enuclea nuove possibilita' e proprieta' relativamente ai medesimi processi consentendo sia di comprenderne aspetti non pensabili prima dell'avvento dell'informatizzazione che di scoprire nuove frontiere di conoscenza[2].
Vedremo nel seguito di questo intervento un esempio di entrambi questi modi d'uso delle tecnologie in linguistica e, in particolare, come essi siano stati applicati alla dialettologia, allo scopo di inquadrare la proposta (proposta che rappresenta il fulcro di questo articolo) per la realizzazione di un atlante linguistico dinamico e innovativo in cui l'applicazione delle tecnologie non sia una semplice veste esteriore, pure importante, che permette impensabili collezioni di dati empirici e la loro 'conta', ma costituisca anche una simulazione e un modello teorico dei fenomeni di variazione linguistica.

1.2 Informatica e linguistica

La natura matematica delle discipline informatiche porta un beneficio immediato in qualunque processo di 'conta' che possa essere necessario in qualunque campo del sapere e della tecnica. Le prime applicazioni della teoria dell'informazione alle lingue

---

[*] Questo articolo è frutto di un lavoro di collaborazione, tuttavia relativamente alle norme che regolano la valutazione dei titoli nell'Accademia italiana, precisiamo che Elisabetta Gola è responsabile delle saezioni 1.2, 1.3 e 1.4 e 1.6 e Stefano Federici delle sezioni 1.1 e 1.5 e rispettive sottosezioni.

[1] Non faremo una distinzione di principio tra linguistica e dialettologia, dal momento che i dialetti da un punto di vista esclusivamente teorico non differiscono dalle lingue. Ci atterremo alla considerazione fatta da P. Meyer secondo il quale "non esistono dialetti ma *caratteri dialettali*" e considereremo pertano i fenomeni dialettali come parte del fenomeno della variazione linguistica che riguarda i sistemi linguistici siano essi riferiti a lingue dominanti o a dialetti.

[2] Questa distinzione per quanto riguarda la dialettologia e' stata ampiamente sottolineata e motivata da A. Pennisi, specialmente nel suo articolo "L'informatica per la dialettologia", *Rivista Italia di Dialettologia*, XV, 1992, pp. 137-164.

enfatizzavano proprio questo aspetto: dai tentativi da parte della teoria dell'informazione di misurare la quantita' di informazione presente nei testi (entropia)[3] alla selezione di liste di parole e frasi ordinate (concordanze, repertori e dizionari macchina) contenute nei testi, al conteggio statistico delle medesime, tutti questi casi sono un esito che possiamo definire 'naturale' dell'ingresso delle tecniche informatiche in linguistica. Uno dei primi esempi, forse il primo, dell'applicazione di metodi di automazione linguistica è la progettazione (che risale ai primi anni '40) e realizzazione (che ha richiesto più di 30 anni di lavoro) dell'*index thomisticus* da parte di padre Roberto Busa, internazionalmente riconosciuto come il pioniere della linguistica computazionale e dell'informatica umanistica più in generale. L'*index thomisticus* consiste nella classificazione di tutte le voci del lessico del corpus tomistico, a livello di lemma e di morfologia e delle relative concordanze.

Accanto a questi aspetti tuttavia, sin dagli stessi anni '50 che hanno visto la nascita e lo sviluppo delle tecniche computazionali applicate alle scienze umane, è nata un'area tuttora viva dell'intelligenza artificiale, il *Natural Language Processing*, con obiettivi molto più ambiziosi relativamente alla possibilità di far convergere abilità computazionali delle macchine e teorie del linguaggio. Si trattava infatti di simulare in un modello computazionale i processi mentali e/o comportamentali attraverso i quali si attuano i processi linguistici, nel quadro del paradigma generale della metafora del calcolatore per cui i programmi del calcolatore (*software*) venivano considerati analoghi a presunti processi rappresentazionali interni alla mente (qualunque entità si voglia intendere con questa parola).

Uno dei primi programmi non banali di questo approccio è il famosissimo SHRDLU, un software creato da Terry Winograd[4] in grado di interpretare frasi dell'inglese relative a un mondo fittizio fatto di blocchi di diversi colori e dimensioni e di agire di conseguenza sul tale mondo-giocattolo (*toy-world*) (figure 1 e 2). SHRDLU conteneva una piccola grammatica e una semantica di natura procedurale (cioè il significato delle frasi veniva interpretato come un'azione da compiere). L'idea forte che stava dietro questo modello era l'ipotesi della sua plausibilità in termini esplicativi.

Anche relativamente ai fenomeni di variazione linguistica si possono trovare esempi interpretati all'interno dell'approccio sviluppato nei primi studi automatici del linguaggio e dei nuovi approcci "alla Winograd", ma va detto subito che la bilancia pende fortemente dalla parte del primo caso.

Quasi tutti i progetti di dialettologia di cui siamo riusciti ad avere notizia sono per lo più orientati alla raccolta massiccia dei dati e alla loro facile e immediata consultazione. Il pregio di tali progetti e' la disponibilita' finale di vaste raccolte di dati che possono poi essere studiati indipendemente dagli obiettivi originali con cui

---

[3] La teoria dell'informazione tentava, attraverso dei conteggi probabilistici, di calcolare una misura chiamata entropia, ossia la quantità di incertezza che la presenza di un segnale (informazione) risolve. Per esempio, se si considera che in una parola di cinque lettere ogni lettera può pervenire da un insieme di circa cinquanta caratteri diversi (tra minuscole, punteggiature e altri caratteri), si ottiene per essa un valore informativo complessivo di 30 bit. Infatti $\log_2 50=5{,}64$ e 6 bit moltiplicato per il numero dei fonemi = 30 bit. Ogni lettera dunque, come sostiene il Volli, «contiene» fra i 5 e i 6 bit di informazione. » U. Volli, *Il libro della comunicazione*, Il Saggiatore, Milano, 1994, p. 20.

[4] T. Winograd scrisse questo programma negli anni 1968-70 nel laboratorio di Intelligenza Artificiale del M.I.T. Artificial Intelligence Laboratory. SHRDLU è stato descritto da Winograd nella sua tesi di dottorato dal titolo *Procedures as a Representation for Data in a Computer Program for Understanding Natural Language*. La sua tesi fu pubblicata dalla rivista *Cognitive Psychology* (Vol. 3 No 1, 1972) che gli dedicò un intero numero e uscì in seguito come libro: *Understanding Natural Language* (Academic Press, 1972).

sono stati raccolti. Senza il contributo di tali progetti sarebbero necessari ogni volta considerevoli sforzi organizzativi che non sempre e' possibile organizzare.

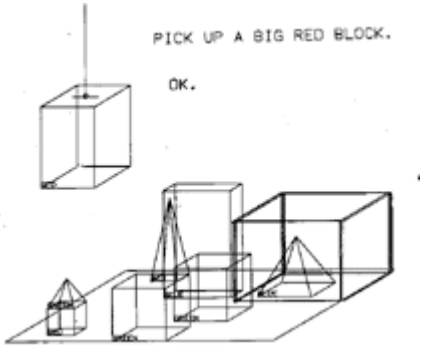

Fig. 1. La schermata di SHRDLU nella sua versione originaria in bianco e nero

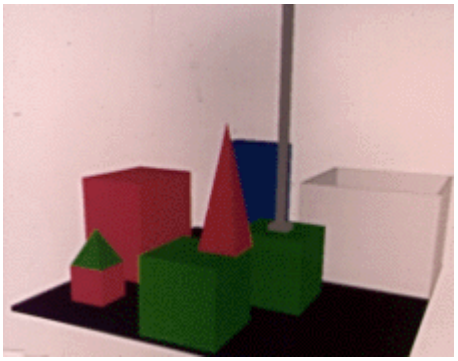

Fig. 2. La schermata di SHRDLU nell'aggiormento del software curato dall'Università dello Utah.

Per quel che abbiamo potuto verificare, benché vi siano diversi progetti di ricerca che nascono già con una metodologia informatizzata, l'unico tentativo che possa essere considerato una consapevole applicazione 'speculativa' dello strumento informatico[5] ci sembra sia il lavoro fatto per l'Atlante Linguistico Siciliano (ALS), ed in particolare alcuni dei suoi aspetti. L'esito più eclatante dovuto all'applicazione di tecniche di intelligenza artificiale all'analisi dei dati raccolti, che svela aspetti del panorama variazionale siciliano inconoscibili se si prescinde dall'applicazione delle tecniche utilizzate, è a nostro avviso la costruzione delle cosiddette carte irradiazionali. Le carte irradiazionali si ispirano ai modelli di rete connessionista[6], e la procedura irradiazionale consiste nella costruzione di una rete che collega i vari punti linguistici[7] di un'area in maniera proporzionale alla presenza/assenza di un determinato fenomeno[8].

---

[5] "[...] la microinformatica applicata a un progetto di ricerca che non ne prevedeva l'utilizzazione già dall'inizio, tende a costituire uno strumento di rafforzamento degli apparati teorici tradizionali. Tende, insomma, ad assumere un ruolo "conservatore", limitandosi semplicemente a rendere più semplice la standardizzazione delle trascrizioni, le procedure di ordinamento, ricerca e analisi dei dati, e, al massimo, a incoraggiare una cartografazione standard più immediata e pubblicabile velocemente. Il che certo non è poco se si tiene conto dei tempi a volte "biblici" che occorrono per portare a compimento un'impresa geolinguistica. Ma è certamente pochissimo se, al contrario, si vuole - come si può - utilizzare al massimo le risorse non solo tecniche ma anche "speculative" del mezzo informatico. In quest'ambito di lavori, comunque, si sono ottenuti risultati interessanti a seconda del tipo di ricerca a cui il piano d'informatizzazione è stato applicato" A. Pennisi, "L'informatica per la dialettologia", *Rivista Italia di Dialettologia*, XV, 1992, pp. 137.

[6] Il *connessionismo* e' una branca dell'intelligenza artificiale che tenta di simulare processi intelligenti tramite reti di unita' che interagiscono tra loro cosi' come fanno i neuroni del cervello. Il risultato di elaborazioni eseguite da una rete neurale e' uno stato di equilibrio tra i segnali elettrici inviati da ciascuna unita' alle unita' limitrofe.

[7] In particolare i punti linguistici nel caso dell'ALS sono punti che includono sia informazioni di natura diatopica che diastratica, per questo sono stati battezzati VUSP (Vettore Sociolinguistico Unitario del Punto).

[8] Per una descrizione più dettagliata si cfr. più avanti il paragrafo 1.4

In questo caso l'applicazione connessionista non rappresenta comunque una simulazione di processi mentali, o meglio cerebrali, ma un'euristica rivolta all'elaborazione di dati linguistici e metalinguistici.
Come vedremo nel capitolo centrale di questo intervento l'ambizione della proposta qui illustrata è invece quella di consentire una navigazione tra i dati che sia assimilabile a quella di un parlante che, per far fronte alle variazioni, confronta e mappa in continuazione i patterns in ingresso (sul piano della *parole*) con i patterns che fanno parte del suo bagaglio di conoscenze linguistiche (*langue*). Il tipo di atlante che verrebbe supportato da questo tipo di teoria e tecnologia sarebbe un atlante variazionale in cui verrebbero presentate contemporaneamente differenze e similarita', che di per se' non emergerebbero in modo immediato dai dati. Questo verra' realizzato utilizzando tecniche di Intelligenza Artificiale che consentono sia di servirsi di procedure di *pattern matching* flessibili sia di ricavare automaticamente regolarita' e omogeneita' presenti nei dati raccolti. Per capire meglio la portata innovativa di questa proposta, tuttavia, prima di illustrarla più in dettaglio ci è parso opportuno da una parte rivolgere uno sguardo, seppur rapido, agli antecedenti storici degli atlanti informatizzati e dall'altra tentare di delineare lo stato dell'arte sulle tecniche più recenti utilizzate nella costruzione di atlanti informatizzati.

1.3 Antecedenti 'remoti' degli atlanti informatizzati

Con la dicitura antecedenti 'remoti' degli atlanti informatizzati, intendiamo indicare quegli atlanti linguistici sviluppati prima dell'avvento delle tecnologie dell'informazione. Li consideriamo antecedenti, infatti, non per presunte anticipazioni delle tecnologie attuali, ma perché le intenzioni e le modalita' nella raccolta dei dati per la compilazione degli atlanti contenevano intrinsecamente aspetti metodologici di natura informazionale. La rappresentazione dei dati secondo determinate coordinate e la loro rappresentazione cartografica, anche quando fatte manualmente, seguivano procedure che si sono prestate poi, appena gli strumenti l'hanno consentito, all'immediara informatizzazione. La costruzione di atlanti linguistici in questo senso risale alla fine dell'ottocento[9].
Un esempio di questo tipo di passaggio è stato il trasferimento dell'index thomisticus dai volumi cartacei, al supporto magnetico (CD-Rom) avvenuto nel 1982.

1.4 Atlanti informatizzati realizzati in altre regioni italiane

In altre regioni italiane gli atlanti informatizzati sono una realtà o come progetti avviati o come lavori 'conclusi' ossia in cui l'obiettivo di costruire un atlante che rispondesse ai criteri del progetto è stato raggiunto. Gli atlanti linguistici che si sono avvalsi dell'ausilio di strumenti informatici e software dedicati hanno portato come beneficio più evidente la possibilita' di esplorazione e rappresentazione non solo delle varietà geografiche, ma anche delle mutazioni di tipo diastratico e diafasico. Cio' ha avuto importanti ritorni nella sociolinguistica dello spazio com'è stato evidenziato tra gli altri da Agostiniani[10], Pennisi e D'Agostino[11].

---

[9] A. Pennisi, "L'informatica per la dialettologia", *Rivista Italia di Dialettologia*, XV, 1992, pp. 137-164.
[10] Agostiniani L., Montemagni S., Paoli M., Picchi E., Poggi Salani T., 1992, "La costruzione di un sistema integrato per il trattamento dei dati dell'Atlante Lessicale Toscano: esperienze, problemi, prospettive". In *Atti del Convegno del Centro di Studi Filologici e Linguistici Siciliani su Atlanti*

Tra gli atlanti linguistici informatizzati di particolare pregio e interesse scientifico meritano senz'altro di essere citati L'Atlante Linguistico della Toscana (ALT)[12] e il gia' citato Atlante Linguistico della Sicilia[13].
Entrambi, per ragioni in parte diverse, sono sicuri esempi dell'importanza del contributo informatico alla realizzazione di una ricerca geo-sociolinguistica complessa, ossia capace di rappresentare tutto l'arco della variabilità negli usi locali e regionali.
Tra gli aspetti interessanti dell'ALT come atlante informatizzato vi è la modalità di ricerca sui dati (*retrieval*). L'ALT è infatti interrogabile attraverso il DBT-ALT[14], una versione specializzata del DBT, un sistema di database testuale per la memorizzazione, gestione ed interrogazione di grandi archivi di testi; il sistema gestisce una varia tipologia di dati linguistici strutturati, che contengono rappresentazioni sia in trascrizione fonetica - forme ottenute in risposta alle domande o a margine di esse, fraseologia o etnotesti - sia in ortografia italiana - descrizioni o note discorsive. Il DBT-ALT permette di consultare i dati sia secondo le canoniche chiavi di accesso al corpus di materiali dialettali sia sulla base delle caratteristiche socio-culturali dell'informatore che li ha attestati. A queste modalità di accesso se ne affiancano altre che permettono esplorazioni personalizzate, a partire ad esempio da un concetto espresso tramite una chiave semantica, o semplicemente da una attestazione dialettale, registrata in trascrizione fonetica.
L'ALS, invece, oltre a permettere la selezione e il recupero articolati dei dati memorizzati, utilizzando anche tecniche mediate dall'I.A., consente anche la produzione di output diversificati, come le carte "parlanti", le carte videografiche e le carte variazionali (sulla base di parametri statistici e sociolinguistici). Come abbiamo anticipato nel paragrafo 1.2, uno degli aspetti relativamente al quale sono state applicate in modo più esplicito e 'visibile' tecniche di intelligenza artificiale è la costruzione delle cosiddette carte irradiazionali. Le carte irradiazionali che ispirandosi ai modelli di rete connessionista[15] permettono la costruzione di reti che collegano i vari punti linguistici di un'area proporzionalmente alla presenza/assenza di un determinato fenomeno.
In sintesi, un collegamento viene creato se in due punti contigui i dati sulla presenza/assenza di un determinato fenomeno (rilevati statisticamente e confermati a mano) superano un valore di soglia prestabilito. La presenza, assenza, debole presenza di interconnessioni consente di diagnosticare la portata omogeneizzatrice del

---

Linguistici Italiani e Romanzi: esperienze a confronto, Centro di Studi Filologici e Linguistici Siciliani, Palermo, pp. 357-393.

[11] M. D'Agostino e A. Pennisi, *Per una sociolinguistica spaziale*, ALS Materiali e ricerche, Palermo, 1995.

[12] L'ALT, avviato nel 1973 e descritto in G. Giacomelli, "Storia, criteri, metodi, prospettive dell'Atlante Lessicale Toscano", *Quaderni della Sezione di Glottologia e Linguistica del Dipartimento di Studi Medievali e Moderni, Università degli Studi "G. D'Annunzio" di Chieti"*, 1991 e Agostiniani, L., Montemagni, S., Poggi Salani, T. "La costruzione di un sistema integrato per il trattamento del dati dell'Atlante Lessicale Toscano: esperienze, problemi, prospettive", in AA.VV., *Atlanti linguistici italiani e romanzi. Esperienze a confronto*, Palermo,1992, pp. 357-393.

[13] Le linee progettuali dell'ALS sono illustrate in Ruffino G. (a cura di), 1990, *Materiali e ricerche dell'Atlante Linguistico della Sicilia*, CSFLS, Palermo, risalgono al 1985,

[14] Cfr. Picchi, E., Montemagni,S., Biagini, L., "DBT-ALT: a System for Storying and Querying the Data of the Atlante Lessicale Toscano (ALT)" , Dialectologia et geolinguistica, DIG 9, 2001.

[15] Una rete connessionista simula il comportamento delle reti neuronali e consiste di un insieme di noti correlati tra loro, ognuno dei quali simula le proprietà di un singolo neurone, rappresentato come unità computazionale in grado di elaborare qualunque funzione Y=F(X). La rete connessionista è un sistema dinamico, in quanto è in grado di esprimere la variabilità dellostato nel tempo.

fenomeno studiato. Per esempio, se in una certa area è presente in maniera diffusa un determinato fenomeno, il programma costruirà una rete abbastanza uniforme che di punto in punto collegherà tra loro tutti i punti di una determinata zona (figura 3). Questa tecnica ha l'obiettivo di far «emergere rapporti di "parentela" (nel senso blando di "analogia") nuovi o non ancora considerati fra diversi esponenti di un medesimo comportamento linguistico»[16].

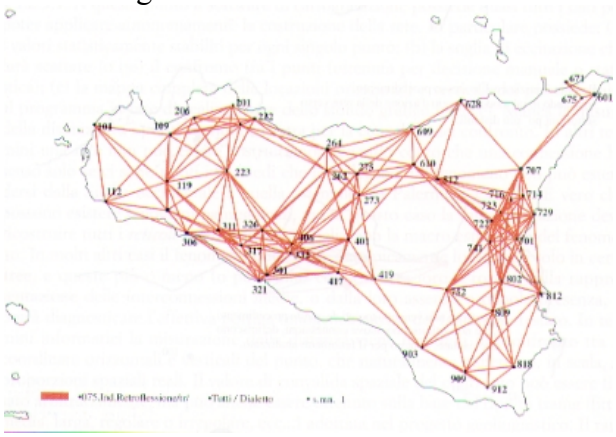

Fig. 3. Carta irradiazionale riferita alla retroflessione dialettale del nesso /tr/.(Tratta da M. D'Agostino e A. Pennisi, *Per una sociolinguistica spaziale*, ALS Materiali e ricerche, Palermo, 1995, p. 42).

Inoltre questo genere di carte associate alle carte polarizzanti consentono «una sorta di 'sintonia fine' nel diagnosticare differenze a prima vista non emergenti»[17]. Un esempio offerto da questa possibilita' può essere esaminato attraverso la carta irradiazionale che analizza la retroflessione del nesso /tr/ rispetto al dato diatratico età: la cartina precedente, se ricostruita tenendo conto della differenza giovani/vecchi rispetto all'uso del tratto considerato non solo in dialetto ma anche in italiano, assume un aspetto molto diverso (figura 4). La prestazione dialettale (retroflessione del /tr/ in dialetto) delle generazioni più anziane copre «tutta la regione (eccezion fatta per alcuni punti centrali, Petralia, Alimena e Mazzarino, che conservano tracce di una storia diversa)»[18]. Nel caso delle generazioni più giovani la percezione del tratto come marcato pone vincoli alla pronuncia che in italiano risulta controllata. Ma se si considera la prestazione in italiano sempre nelle vecchie generazioni si può notare come «anche negli strati più "antichi" della popolazione, sino ai nati nel primo ventennio del secolo, l'autocontrollo di un tratto percepito come fortemente marcato, inizia a porre vincoli alla pronuncia»[19]. A cosa serve in questo caso una proiezione irradiazionale (figura 4.)? Con le parole di D'Agostino e Pennesi

[16] M. D'Agostino e A. Pennisi, *Per una sociolinguistica spaziale*, ALS Materiali e ricerche, Palermo, 1995, p. 42.
[17] *Ibid*.
[18] *Ibid*.
[19] *Ivi*, p. 43.

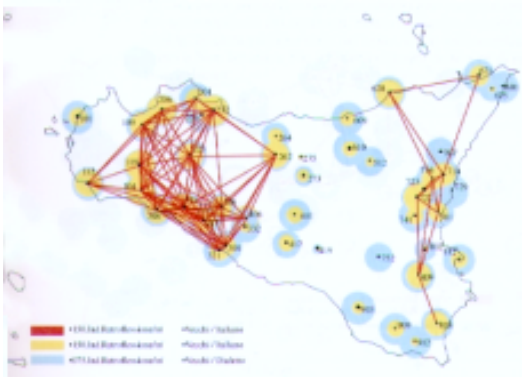

Fig. 4. Carta irradiazionale riferita alla retroflessione dialettale del nesso /tr/ nei Vecchi.(Tratta da M. D'Agostino e A. Pennisi, *Per una sociolinguistica spaziale*, ALS Materiali e ricerche, Palermo, 1995, p. 43).

La proiezione irradiazionale aggiunge [...] una focalizzazione statistico-quantitativa molto precisa: in primo luogo identifica un'area di forte resistenza al cambiamento definita come una fitta trama di linee tra i punti connessi (Sicilia Occidentale, con epicentro nella zona più interna, Corleone); in secondo luogo segnala la presenza di un'origine diatopica forte anche nella zona orientale; infine segnala che i punti dove il sostrato più antico è meno marcato sono anche quelli dove l'intacco del fenomeno in italiano era gia' nettamente compiuto all'inizio del processo[20].

I progetti toscano e siciliano non sono comunque isoltai. Un altro tentativo importante di informatizzazione è rappresentato dall'ALLI (Atlante Linguistico dei Laghi Italiani). Inoltre altre regioni come il Piemonte, la Val d'Aosta, il Salento, la Campania hanno in cantiere la preparazione di atlanti linguistici regionali. Nel prossimo paragrafo analizzeremo lo stato dell'arte per quanto riguarda lo sviluppo di un atlante linguistico della Sardegna nonche' alcune proposte innovative per quanto riguarda l'applicazione delle tecnologie informatiche a questa branca della linguistica computazionale.

1.5 Una proposta informatica innovativa per un atlante linguistico sardo

1.5.1 Stato dell'arte

Illustrare una ricognizione esaustiva dello stato dell'arte dei diversi lavori individuali e di gruppo in cui sono stati sviluppati atlanti linguistici di sottoparti del lessico regionale esula dagli scopi del presente lavoro. Ci limiteremo qui a considerare il fatto che manca in Sardegna, diversamente da altre regioni d'Italia, un atlante linguistico complessivo e informatizzato che sia frutto di un progetto regionale di ampio respiro orientato alla conoscenza e conservazione delle molte e diverse parlate locali presenti nel territorio sardo. Ancor prima dell'Atlante come risultato si sente la mancanza di un progetto costante ed esteso di raccolta e osservazione dei dati (come quello che continua ad essere portato avanti dall'Osservatorio linguistico siciliano, per esempio). Per quel che la nostra breve indagine ha rivelato, il tentativo più simile a quello condotto in altre regioni per quanto riguarda la creazione di un atlante linguistico

[20] *Ibid*.

sardo, fu il lavoro pubblicato nel 1964, ad opera di Benvenuto Terracini e Temistocle Franceschi, intitolato appunto *Saggio di un atlante linguistico della Sardegna*. Dei due volumi del saggio, uno –curato dal Terracini- illustra l'organizzazione dell'opera, mentre l'altro, l'atlante vero e proprio, è costituito dalle carte che illustrano le variazioni delle parole considerate nelle varie aree geografiche utilizzate come campione.
La Sardegna è inoltre inclusa come progetto nell'ambito dell'ALI (Atlante Linguistico Italiano), un ambizioso progetto di raccolta ordinata e sistematica di carte sulle quali sono riprodotte, per ogni località italiana esplorata, le corrispondenti traduzioni dialettali di un concetto o nozione o frase raccolte dalla viva voce del parlanti. Le località individuate in Sardegna nell'ambito di questo progetto sono 109. L'ALI ha una lunghissima storia che ha inizio nel 1924, quando sotto la direzione di M. G. Bartoli (Albona, Istria, 1873 – Torino, 1946) e su iniziativa della Società Filologia Friulana "G.I. Ascoli", il progetto venne avviato. Dopo diverse vicissitudini, difficoltà, ripensamenti e ristrutturazioni del progetto, sul finire degli anni '80, con la soluzione di alcuni gravi problemi di carattere istituzionale e organizzativo, i lavori passarono dalla fase preparatoria a quella vera e propria di redazione e pubblicazione. In quegli anni sono A. Genre prima e L. Massobrio poi a dirigere l'ALI. In collaborazione con l'Istituto Poligrafico e Zecca dello Stato, vengono studiate nuove procedure e sperimentate tecnologie computazionali per la creazione di una banca-dati gestibile elettronicamente. Vengono anche costruiti nuovi set di caratteri fonetici speciali, allestiti software per il trattamento e la cartografazione automatica dei materiali dialettali, messa a punto una nuova base cartografica. In questa fase fu fondamentale l'apporto dell'esperienza maturata dall'Istituto di Linguistica Computazionale del Cnr di Pisa (allora diretto da Antonio Zampolli) in materia di elaborazione di dati linguistici.
L'ALI ha dato luogo alla pubblicazione di alcuni volumi di carte, parole e fotografie e altri sono in fase di pubblicazione e preparazione.
Nonostante questi progetti, non è tutt'ora facilmente disponibile agli studiosi nessun repertorio in formato elettronico sufficientemente ampio da consentire una ricognizione e uno studio delle parlate presenti nella regione Sardegna.
Una revisione e una ricognizione mirata alle specificità della realtà linguistica sarda sarebbe auspicabile e sarebbe altresì desiderabile che qualunque progetto futuro fosse pensato, come abbiamo sostenuto sopra, già nell'ottica di un'informatizzazione che rispecchi e rispetti la dinamicità della lingua.

1.5.2 Analogia e dati linguistici

È nella speranza che si possa progettare e costruire una 'carta parlante' della Sardegna che, facendo tesoro delle esperienze passate, eventualmente utilizzandone i dati già disponibili, rappresenti la ricchezza e la varietà delle parlate regionali. A tale scopo, nel progetto che qui proponiamo vorremmo mettere assieme i vantaggi di un buon metodo di *retrieval* dei dati e quelli derivanti dall'adozione di modelli di natura connessionista. Entrambi gli aspetti verranno raggiunti grazie all'utilizzo di un motore inferenziale basato sui principi dell'analogia in grado di individuare percorsi di similarità tra i dati appresi. Tali percorsi possono essere utilizzati sia in fase di *retrieval* per trovare e raggruppare fenomeni che presentano interessanti similarità ma che, non essendo assolute identità, non potrebbero essere recuperati se non a prezzo di un laborioso (e talvolta impossibile) lavoro manuale; sia per trovare giustificazioni

alle variazioni. Entrambi questi obiettivi, ossia la ricerca articolata delle forme e una sorta di spiegazione della loro presenza in una certa variante dialettale, possono essere raggiunti grazie all'ausilio del sistema proposto in quanto fa uso di un meccanismo che è già intrinsecamente presente nelle lingue e ne guida l'evoluzione: l'analogia[21]. Le inferenze analogiche possono essere tratte grazie alla capacità di apprendimento del motore che in questo modo memorizza in modo dinamico i contesti delle forme che si vogliono studiare e li utilizza come cause che portano a giustificare una determinata forma.
Lo stesso Saussure aveva schematizzato così il processo:

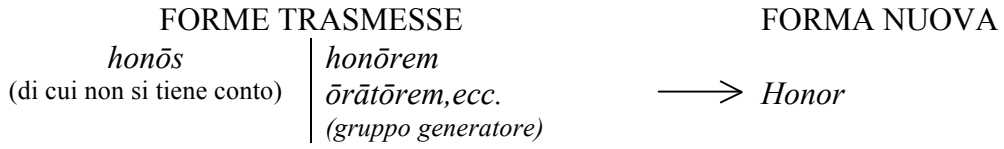

Fig. 5. Tratta da Saussure, *CLG*, p. 197.

Il gruppo generatore, nel nostro sistema, corrisponde all'insieme dei contesti che il motore considera pertinenti relativamente ad una certa conclusione. Ai dialetti, essendo lingue a tutti gli effetti, possono essere applicati tutti gli strumenti di analisi linguistica che sinora sono stati messi a punto utilizzando l'approccio analogico[22]; quello che cambia sono gli interessi che gli studiosi manifestano nei confronti dei caratteri dialettali che spesso sono diversi da quelli delle lingue dominanti e dipendenti dalle peculiarita' di questi sistemi linguistici. Ad esempio la variabilita' geotopologica e' molto maggiore e granulare dato un certo territorio, visto che solitamente i dialetti sono lingue orali per le quali le zone di omogeneita' si estendono sin dove la comunicazione tra gruppi di parlanti riesce a 'raggiungersi' (isoglossa).
Vediamo percio' prima un esempio di come il sistema analogico sia stato utilizzato per l'analisi di fenomeni linguistici relativi all'italiano, per dare un'idea piu' precisa

[21] Nel senso in cui stiamo utilizzando la nozione di analogia in questo caso è esattamente lo stesso che troviamo enunciato da Saussure nel Corso di Linguistica Generale (pp. 197 segg.). Secondo Saussure infatti l'analogia è "il grande fattore di evoluzione delle lingue, il procedimento mediante cui esse passano da uno stato d'organizzazione all'altro" (p. 197). Ogni fatto analogico ha tre componenti in gioco: il tipo trasmesso, il tipo concorrente e tutte le forme che hanno creato il tipo concorrente e come vedremo questi sono esattamente gli elementi che il sistema proposto prende inconsiderazione per giustificare le similarità.

[22] Ad esempio, sistemi analogici sono stati implementati per l'analisi ortofonetica, morfologica, sintattica e semantica di testi in formato elettronico. Si rimanda, per una descrizione di questi diversi aspetti a Federici, S., Pirrelli, V., "The compilation af Large Pronunciation Lexica: the elicitation of letter-to-sound patterns through analogy based networks", in *Proceedings of COMPLEX '94*, Budapest, 1994, per quanto riguarda la trascrizione fonetica; Federici, S. - Pirrelli, V. *MorphEUS: An Analogical Way to Language Modelling*, "Acta Linguistica Hungarica", 41, 1994, pp. 235-264, per una descrizione dell'approccio applicato alla morfologia; Montemagni, S., Federici, S., Pirrelli, V. "Constraints and Preferences at Work: an Analogy-based Approch to Parsing Grammatical Relations", in D. Jones (Ed.), *New Methods in Language Processing*, University College London, London, 1997, per un esperimento sul rilevamento di pattern sintattici; Federici, S., Montemagni, S., Pirrelli, V., "SENSE: an Analogy-based Word Sense Disambiguation System", S*pecial issue on sense disambiguation of Natural Language Engineerin,* volume 7, Cambridge University Press 2001 e Gola, E., Federici, S., "Le regole informali del linguaggio naturale" In *La regola linguistica, Atti del VI Congresso di studi della Società di Filosofia del Linguaggio,* a cura di M. Carapezza e F. Lo Piparo, Palermo, Novecento, 2000, per una discussione su come i sistemi analogici possano risolvere sul piano del significato problemi di ambiguità semantica e metaforicità .

dei suoi principi e del suo funzionamento e poi cerchiamo di indicare alcune possibili direzioni di ricerca perseguibili attraverso l'applicazione del medesimo sistema ai caratteri dialettali in Sardegna.

1.5.3 I sistemi analogici: principi e funzionamento

I sistemi analogici si fondano sull'idea che un sistema automatico intelligente che sia in grado di comprendere un insieme di dati deve poter gestire in qualunque momento l'ingresso di informazioni nuove per il sistema, informazioni cioè per le quali non è possibile, o non è tecnicamente fattibile, dare in anticipo delle regole per il loro trattamento. Il principio generale che guida la costruzione dei sistemi analogici consiste perciò nel progettare programmi che

- da un lato consentano al sistema di acquisire il maggior numero possibile di informazioni utili dai dati che gli saranno forniti come training (*fase di apprendimento*)
- e che in una seconda fase (*fase di test*) consentano al sistema di utilizzare al meglio le conoscenze apprese per dare risposte a dati nuovi e sconosciuti.

Per poter raggiungere questi obiettivi, i sistemi Analogici sono stati sviluppati attorno ai concetti di *apprendimento*, *paradigma analogico*, *estensione di paradigma* e *giustificazione*.

Nella fase di apprendimento i sistemi analogici accumulano conoscenza traendola da un set di dati di riferimento rilevati sul campo, il Set Campione. Un sistema Analogico apprende le risposte fornite dai *casi specifici* presenti nel Set Campione e le struttura internamente in modo da ricavare utili generalizzazioni per analogia (*paradigmi analogici*). Le generalizzazioni rappresentate nei paradigmi analogici rendono esplicite le *similarita'* riscontrate tra casi specifici diversi.
Nella fase di test il sistema cerca la migliore similarità (quella piu' *giustificata*) tra il dato da analizzare e i paradigmi analogici derivati dai casi appresi (*estensione di paradigma*). Questa procedura consente al sistema Analogico di analizzare dati che non corrispondono perfettamente a casi contenuti nel Set Campione: è infatti sufficiente che il nuovo caso sia solo parzialmente simile a casi conosciuti.

Per chiarire il funzionamento generale di un sistema Analogico, si dà un esempio relativo a un compito di comprensione di un testo: supponiamo di voler derivare la giusta interpretazione dell'aggettivo "chiara" (nel senso di 'facilmente comprensibile') contenuto nella frase "questa teoria è chiara". In una frase di questo tipo ci si trova di fronte al problema di un uso metaforico dell'aggettivo "chiara", in quanto non si intende nulla di letterale come "limpida" o "luminosa". Un sistema basato sull'analogia può tuttavia trovare la giusta interpretazione se, ad esempio, conosce le interpretazioni delle frasi

1. "questa idea è chiara"	('*facilmente comprensibile*')
2. "questa idea è oscura"	('*difficilmente comprensibile*')
3. "questa teoria è oscura"	('*difficilmente comprensibile*')

Tabella 1

cioe' se le interpretazioni si queste frasi sono contenute nel Set Campione. La giusta interpretazione è individuata dal sistema Analogico grazie all'inferenza resa possibile dall'estensione di paradigma illustrata in figura 1.

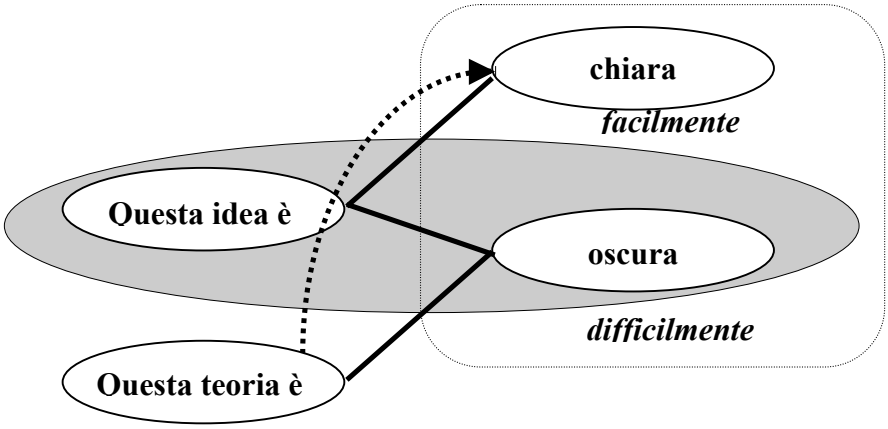

Figura 6: Concetti di base dei sistemi Analogici di *ELN*

Nella figura 6, la conoscenza contenuta nel Set Campione è rappresentata dai cerchi (nodi) connessi dalle linee continue. Ogni coppia di nodi connessa da una linea continua ricostruisce una delle frasi contenute nel Set Campione. In ogni nodo è contenuta una parte del testo della frase. Questa può essere o la parte che è condivisa da testi diversi (ad es. l'aggettivo "oscura", che è condiviso dalle frasi "questa idea è oscura" e "questa teoria è oscura"), oppure la parte che rimane di un testo dopo aver isolato il segmento condiviso (ad es. la sequenza "questa teoria è" che è contenuta in una sola frase). Nell'area racchiusa dalla linea tratteggiata è espresso il *paradigma analogico* relativo alla porzione di testo "questa idea è", rappresentato dagli aggettivi "chiara" e "oscura". Questo tipo di strutturazione della conoscenza contenuta nel Set Campione permette di estendere l'uso dell'aggettivo "chiara" inteso come 'facilmente comprensibile' alla frase "questa teoria è chiara", e cioè di *estendere* il paradigma della porzione di frase "questa teoria è" (rappresentato in questo esempio dal solo aggettivo "oscura") con l'aggettivo "chiara". L'*estensione di paradigma* è rappresentata dalla freccia tratteggiata che termina su "chiara" ed è permessa dal fatto che i paradigmi delle due frasi condividono almeno un elemento in comune, rappresentato in questo caso dall'aggettivo "oscura".
Il processo di estensione di paradigma può essere parafrasato nel seguente modo: se sia le idee che le teorie possono essere oscure nel senso di 'difficilmente comprensibili' e un'idea può essere chiara nel senso di 'facilmente comprensibile', allora è *giustificato* supporre che le teorie, per *analogia* con le idee, possano essere chiare nello stesso senso. La frase "questa teoria è oscura", racchiusa nell'area grigia, costituisce quindi la *giustificazione* necessaria per poter considerare analogicamente motivata questa inferenza.
Come vedremo nelle sezioni seguenti, tutti i sistemi Analogici, pur trattando informazioni linguistiche di natura differente, si basano su questo stesso schema di funzionamento.
Dato l'alto indice di varibilita' dei caratteri dialettali rispetto a quello delle altre lingue, è importante a questo punto affrontare il problema delle eccezioni ed illustrare il loro ruolo nei sistemi Analogici. Finora abbiamo infatti usato i termini similarità, analogia, estensione. Questi termini evocano spontaneamente alla mente i loro

contrari: *differenza*, *caso particolare*, *eccezione*. Nei sistemi inferenziali classici e statistici i casi che non rientrano nella classe più frequente (cioè quelli che non rientrano nella *regola*), e che non hanno quindi una frequenza di apparizione paragonabile a quella dei casi più frequenti, vengono dichiarati eccezioni. In questi sistemi quello che si cerca di derivare e' una base forte, un insieme di regole con la maggiore "pulizia" possibile, che permetta di intepretare e trattare tutti quei fenomeni che vengono detti far parte del *nucleo* principale del set di dati (la parte a bassa deviazione standard), e che lasci al di fuori soltanto quella parte del campione che viene identificata con la *periferia*, cioè le irregolarità e le eccezioni (dati con alta deviazione standard). Le motivazioni con cui tale divisione viene tracciata spesso non possono non essere interpretate come pura ricerca della "pulizia" formale dell'insieme di regole che si intende ricavare: periferia è quindi tutto quello che non si riesce a trattare con un insieme di regole che rispetti particolari canoni di essenzialità e semplicità. Nei sistemi Analogici quelle che vengono viste da altri sistemi come eccezioni sono invece considerate al pari di regolarità (e quindi rappresentate sempre tramite paradigmi analogici) la cui sola peculiarita' e' quella di avere una frequenza di occorrenza più o meno bassa. Il fatto che queste regolarità sottostiano a insiemi di dati grandi, medi o piccoli non fa differenza. Eventualmente, ordinando l'insieme dei paradigmi analogici individuati in un Set Campione, si potra' ottenere una "classifica delle regolarità" ricavate autonomamente dal sistema, dalle più regolari (più frequenti) alle più irregolari (meno frequenti). Ma l'autonomia con la quale i sistemi Analogici scoprono le somiglianze che accomunano casi anche molto diversi ma non del tutto dissimili, consente di evitare la stesura di un complesso sistema di regole, spesso *ad hoc*, per la gestione di quei casi che un sistema inferenziale classico considererebbe eccezioni e relegherebbe nella periferia.
Gli unici casi che si discostano da questo trattamento sono quelli che potremmo chiamare "vere eccezioni", e cioè quei casi che non hanno analogie con nessun altro caso conosciuto, per i quali non c'è una gradazione di irregolarità e non esistono casi simili. In tutti questi casi i sistemi analogici, come qualunque altro sistema artificiale e come gli esseri umani, non possono far altro che imparare il significato dell'eccezione e ricordarlo.
L'alternativa computazionale ai sistemi formali classici proposta dai sistemi Analogici è centrata sull'idea che questo insieme di "eccezioni" siano in realtà indice del reale funzionamento del linguaggio: il modo in cui si commettono gli errori, la scelta delle parole usate metaforicamente, la scelta delle corrette sequenze di parole in una frase suggeriscono un processo che parte da singoli casi conosciuti e li generalizza, seguendo percorsi analogici tra le regolarità individuate in una rete di relazioni linguistiche.

La regola linguistica è quindi distribuita su un insieme di unità interconnesso in rete, un po' come avviene in una rete neurale, tranne che in una rete analogica le unità della rete sono simboliche (e non sub-simboliche come avviene invece nelle reti neurali) ed inoltre la rete si estende senza avere una forma e una dimensione predeterminate a priori.

Nella rete mostrata in figura 6 è evidenziato ciò che il sistema analogico conosce: ogni coppia unita da una linea è stata appresa dal corpus (vedi Tabella 1). In più però la strutturazione in rete analogica permette al sistema di inferire che la coppia unita dalla freccia tratteggiata è altamente probabile. La risposta nel sistema analogico è

prodotta da un bilanciamento tra regolarità supportate da molti fatti, regolarità meno supportate ed eccezioni.
Seppure la regola possa essere "estratta" dalla rete, non è in media nulla di simile ad una semplice espressione del tipo "se…allora", ma piuttosto qualcosa che somiglia a un "se… allora e se invece… allora… purché non…ecc.".
Il sistema è facilmente utilizzabile ed estensibile e finora, in tutti i compiti in cui è stato utilizzato (il suono, la parola, l'accostamento di parole, la semantica), non ha mai dato risultati inferiori a quelli ottenuti da sistemi a regole o statistici.
D'altronde un sistema a regole che fallisce deve essere corretto in maniera estremamente complessa, mentre per un sistema analogico e' sufficiente fornire un insieme sufficiente di conoscenze specifiche relative ai casi in cui fallisce.

1.5.4 I sistemi analogici e i dialetti: progetti per la Sardegna

Dato per assodato che le tecniche per la raccolta e la memorizzazione dei dati e la loro proiezione cartografica possono essere senz'altro applicate ai dati relativi alle lingue presenti sul territorio sardo con le stesse modalita' adottate per gli atlanti delle altre regioni italiane, quello che vorremo qui presentare riguarda quello che si potrebbe fare in aggiunta, di diverso e innovativo, rispetto allo studio delle parlate locali qualora si adotti l'approccio analogico presentato sopra.
A nostro avviso due sono principalmente i campi in cui converrebbe concentrare gli sforzi. Innanzitutto, come abbiamo gia' piu' volte anticipato implicitamente, sarebbe interessante applicare il sistema alla categorizzazione dei dati e al loro recupero. Un sistema come il motore inferenziale basato sull'analogia potrebbe infatti in questo campo contribuire alla costruzione non tanto o non solo di un database relazionale, ma di una base di conoscenza linguistica, storicamente stratificata e dinamica, che verrebbe utilizzata sia per analizzare, giustificare e prevedere l'andamento delle variazioni linguistiche che per tracciare una sorta di isoglosse, ma piu' frastagliate e dinamiche, relative ai fenomeni studiati piuttosto che a un presunto carattere dialettale omogeneo.
Il secondo contributo per il quale l'adozione di un sistema analogico sembra una scelta naturale è relativo alla possibilità di ricavare, a partire dai contesti attestati, le regolarita' in modo automatico e motivato. Il risultato sarebbe una grammatica naturale dei vari caratteri dialettali, che in un momento in cui l'interesse verso la didattica del sardo riscuote particolare attenzione, costituirebbe uno strumento importante e scientificamente più corretto rispetto ai tentativi artificiali e astratti di creare una grammatica di un presunto dialetto sardo.
Nel seguito ci concentreremo solamente sul primo punto, riservando il secondo a lavori futuri.

Analisi variazionali

Ogni obiettivo richiede di formulare quali sono le risposte che si desidera ottenere a partire da un corpus costituito da coppie di informazioni input/output. Nel caso dell'analisi variazionale le coppie saranno costituite da una delle varianti di una stessa parola associata a una qualche informazione semantica che indichi al sistema che si tratta dello stesso item. Tale informazione può essere ad esempio la parola in un'altra lingua (come l'italiano). Se scegliamo questa possibilita' di esprimere l'informazione semantica la struttura dei dati sarà la seguente:

Mabadía/malattia, maladia/malattia, malaria/malattia, maledia/malattia, mobadia/malattia

A partire da questo corpus, ci si aspetta che il meccanismo analogico riesca a formulare ipotesi riguardo a possibili varianti di parole non contenute nel corpus.
Il modo in cui questo viene ottenuto è attraverso un'analisi morfologica realizzata dai paradigmi analogici. La realizzazione di un atlante variazionale dinamico richiedera' la raccolta di una notevole mole di dati che riflettano l'uso reale della lingua. Per dare un'idea concreta del meccanismo di base, daremo ora un esempio di pura analisi morfologica riferita a due verbi "irregolari" sardi: 'beni' (venire) e 'teni'[23] (tenere). Il paradigma analogico risultante e' mostrato in figura 7.

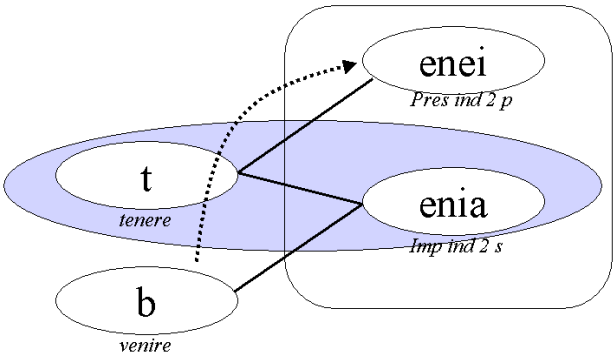

Fig. 6 Paradigma Morfologico in Sardo

Vediamo i dati contenuti nel corpus e quale inferenza analogica ne deriva. Il corpus da cui il paradigma mostrato e' derivato contiene i seguenti fatti:

1. <benia/venire imp ind 2 s>
2. <tenia/tenere imp ind 2 s>
3. <tenei/tenere pres ind 2 p>

La struttura del singolo fatto contiene informazioni morfologiche, oltre al significato della parola espressa con la traduzione italiana: ad esempio il fatto 1 (<benia/venire imp ind 2 s>) esprime la conoscenza che "benia" e' la 2a persona singolare dell'imperfetto indicativo. A partire dai fatti 1-3 il sistema analogico e' in grado di estendere il paradigma del nodo <b/venire> (costituita nel corpus dal solo nodo <enia/imp ind 2 s>) con il nodo <enei/pres ind 2 p>. Questa estensione di paradigma estende la conoscenza contenuta nel corpus con il fatto seguente:

4. <benei/venire pres ind 2 p>

Il caso presentato in questo esempio fornisce anche un chiarimento di quanto discusso in precedenza a proposito del ruolo e dell'interazione di regole ed eccezioni in un sistema analogico e di come entrambi i suddetti concetti convergano nel concetto di

[23] Ci rifacciamo alla trascrizione fonetica della variante dell'infinito coerente con il paradigma proposto. Lo stesso ragionamento vale se si considerano i paradigmi dei medesimi verbi nelle varianti 'benner' e 'tenner' proposte nella norma standard unificata di D. Corraine, consultabile all'indirizzo web: www.linbasarda.it.

regolarita'. I due verbi scelti, venire e tenere, sono considerati come due verbi irregolari della coniugazione in –er (benner, tenner) secondo il documento XXX e –i secondo il dizionario XXX. In realta' e' bene considerare come il concetto di coniugazione sia stato scelto come mezzo di economia descrittiva: la conoscenza degli ultimi caratteri dell'infinito (3 per l'italiano) permette di descrivere tutto il resto della coniugazione per un gran numero di verbi. L'adozione di questa strategia descrittiva forza pero' il dover di conseguenza etichettare come irregolari tutti i verbi per i quali la derivazione di appartenenza ad una determinata coniugazione a partire dalla conoscenza degli ultimi caratteri comporterebbe un'errata assunzione delle rimanenti forme della coniugazione del verbo. La strategia alternativa suggerita da un un sistema di estensione dei paradigmi basato sui principi dell'analogia e' quella di definire due verbi come appartenente alla stessa coniugazione di un secondo verbo se condivide con questo una sequenza finale (suffisso) significativamente piu' lunga di qualunque di quella condivisa con qualunque altro verbo. Questo porta a definire "tenner" e "benner" come due verbi appartenenti ad una stessa coniugazione in quanto condividono ad esempio il suffisso di 8 caratteri "eniausu" (cong. imp. 1 p) mentre non condividono un suffisso piu' lungo con nessun altro verbo. In questo modo si riducono le eccezioni anche se le regole (*regolarita'*) risultano piu' complesse –anche se forse piu' naturali- di quelle che si pretenderebbe di descrivere con l'approccio classico.
Applicando quest'analisi a un corpus sufficientemente ampio di parole, raccolte secondo i metodi standard adottati nelle procedure di costruzione di un atlante linguistico regionale, ci si aspetta di riuscire a stabilire relazioni tra le varie parole che consentano di raggruppare quelle che possono essere considerate le varianti, a posteriori e a partire dalla natura dei dati e non esclusivamente dal giudizio dell'intervistatore o di apposite commissioni di valutazione.

1.6 Conclusioni

In questo articolo abbiamo proposto l'adozione di un metodo informatico innovativo come ausilio alla progettazione e costruzione di una carta parlante dinamica. Abbiamo a tal fine evidenziato la rilevanza teorica generale, e non meramente tecnica, dell'applicazione di tecnologie informatiche alla dialettologia. Da questo punto di vista tecnologie diverse consentono la formulazione di ipotesi differenti. Per questa ragione e' nella fase di progettazione in generale, e di un atlante variazionale in particolare, che devono essere prese attentamente in considerazione le tecnologie che si intendono adottare. Cio' che abbiamo proposto in questo articolo e' una prospettiva teorica di derivazione connessionista, ma con caratteristiche peculiari che la rendono particolarmente adatta alla applicazione a dati linguistici. Essa e' stata gia' utilizzata per l'analisi di altre lingue e sembra un promettente strumento anche per lo svolgimento di compiti propri della dialettologia.