Informatica umanistica

Codifica del testo e intelligenza artificiale

Dino Buzzetti

1. *Il* markup *e il testo*

Qual è lo *status* logico-linguistico del *markup*,[1] dei *marcatori* di cui è cosparso il *file* sorgente di qualsiasi documento reperibile sul *World-WideWeb*? La domanda può sembrare oziosa, ma lo è veramente? Che rapporto c'è tra il documento, così come lo leggiamo, o come lo usiamo per estrarne informazione, e la trama più o meno fitta dei *tag* (marcatori) contenuti nei *file* HTML (*HyperText Markup Language*), quella trama che il programma di visualizzazione (*browser*) filtra e nasconde alla nostra vista? Questa semplice operazione, eseguita ad ogni passo dal nostro familiare strumento di 'navigazione', non è così banale, da un punto di vista concettuale, come a prima vista potrebbe sembrare. Qual è in realtà il *testo* del documento che stiamo leggendo su Netscape o su Explorer? quello che vediamo, o quello che il *browser* nasconde alla nostra vista, quello cioè del *file* sorgente senza il quale il *browser* non potrebbe operare?

Il rapporto tra il *testo* e il *markup* non è un rapporto facile. Anzi, il rapporto dell'informatica stessa col testo, si può quasi dire, non è un rap-

[1] L'uso del termine inglese *markup*, invece dell'italiano *codifica* è deliberato. Il termine italiano è ambiguo, perché si riferisce tanto alla codifica binaria dei caratteri, quanto alla marcatura di stringhe di caratteri con marcatori (*tags*) che ne descrivono la funzione, o l'aspetto fisico, all'interno di un documento. L'uso deliberato del termine *markup* non intende essere lesivo dei diritti della lingua. Per la stessa ragione, tutti i termini inglesi sono qui espressi in corsivo, anche se entrati oramai nell'uso italiano corrente.

porto tranquillo. Che cos'è il 'testo' per l'informatico? Nient'altro che un tipo di dato, o di rappresentazione dell'informazione. "Il *testo*", chiarisce l'autore di un libro dedicato esclusivamente al *text processing*, "può essere descritto come informazione codificata come caratteri, o come sequenze di caratteri". Ma subito si affretta a dichiarare che in questa accezione "'testo' non è usato", per esempio, "nel senso del materiale letterario qual è stato originariamente prodotto da un autore". Infatti, non ha molto senso, per il testo definito come il tipo di dato costituito da "stringhe" di caratteri codificati, porsi domande quali: "Questa versione è diversa dal vero testo"? da quello, cioè, "originariamente prodotto" dall'autore?[2]

Tuttavia non pare che sia solo il filologo a doversi preoccupare. Di perplessità ne sorgono per tutti, perché è la stessa nozione ordinaria di testo che muta in presenza della sua rappresentazione digitale. Per tornare all'esempio precedente, qual è il testo? quello restituito dal *browser*, o il *file* su cui il *browser* lavora? Il *browser* deve operare su un *file* costituito da stringhe di caratteri codificati al cui interno siano stati inseriti dei marcatori, per permetterci di leggerlo come un testo, nel senso ordinario o letterario del termine. Ma allora, il *markup*[3] che dobbiamo aggiungere alla pura e semplice stringa dei caratteri codificati, per ottenere il testo che siamo abituati a leggere, fa parte del testo oppure no? Gli stessi curatori delle *Guidelines* della Text Encoding Initiative (TEI),[4] ossia di quella che può a pieno titolo essere descritta come "la più organica proposta finora avanzata" per la "codifica dei testi" e per l'"interscambio dei dati nel

[2] C. DAY, *Text Processing*, Cambridge, Cambridge University Press, 1984, pp. 1-2.

[3] Storicamente, la parola *markup* è stata usata per indicare le annotazioni che si appongono al testo per dare istruzioni al tipografo. Con l'automazione dell'impaginazione e della stampa, il termine fu esteso a tutti i tipi di codici inseriti nella sequenza dei caratteri per dare istruzioni di impaginazione e di stampa. Con l'introduzione di linguaggi dichiarativi, il *markup* è diventato un mezzo per rendere esplicita ogni caratteristica del testo.

Un linguaggio di *markup* è un insieme di convenzioni usate per la codifica dei testi. Un linguaggio di *markup* deve specificare quali marcatori sono permessi, quali sono necessari, come si distinguono dal testo e che cosa significano. Lo SGML (v. *infra*, nota 19) permette di fare le prime tre cose; la documentazione fornita dalle *Guidelines* della TEI (v. *infra*, nota 6) permette di fare la quarta.

[4] C. M. SPERBERG-MCQUEEN and L. BURNARD (eds.), *Guidelines for Text Encoding and Interchange*, Chicago and Oxford, Text Encoding Initiative, 1994.

campo della ricerca umanistica",[5] paiono esitare di fronte a questo tipo di domanda.[6] Da un lato, infatti, essi definiscono il *markup* come "tutta l'informazione contenuta in un *file* e *diversa* dal testo", e dall'altro affermano che "qualunque *aspetto del testo* che sia di qualche importantanza per il ricercatore" può "essere indicato dal *markup*".[7] Come si vede, proprio mentre affermano che il *markup* rappresenta informazione che "non fa parte del testo"[8] ed è diversa dal testo, i curatori della TEI sostengono anche il contrario, ossia che il *markup* rappresenta informazione che "fa parte del testo e coincide col testo".[9] E se l'incertezza è manifesta proprio in coloro che portano la responsabilità di quelle direttive che sono ormai considerate come un "punto di riferimento ineludibile per chi si occupi di trattamento informatico dei testi", i motivi non possono certo essere dei più banali.[10] Evidentemente dev'esserci qualche problema in radice e su tutta la questione converrà senz'altro ritornare.

Ma che senso ha porsi queste domande sul testo? nascono solo

[5] L. BURNARD, *An Introduction to the Text Encoding Initiative*, in D. GREENSTEIN (ed.), *Modelling Historical Data*, St. Katharinen, Max-Plank-Institut für Geschichte i.K.b. Scripta Mercaturae Verlag, 1991, p. 83.

[6] La TEI è un progetto di cooperazione internazionale per stabilire uno standard interdisciplinare per la rappresentazione digitale del testo. Il progetto, promosso da importanti associazioni scientifiche, ACH (*Association for Computers and the Humanities*), ALLC (*Association for Literary and Linguistic Computing*) e ACH (*Association for Computational Linguistics*) e finanziato dall'UE (DG XIII) e dal National Endowment for the Humanities (USA), si propone tre principali obiettivi: (1) specificare un formato comune per l'interscambio dei dati testuali; (2) proporre raccomandazioni per la codifica di materiali testuali, specificando quali aspetti codificare e in che modo; (3) documentare i più importanti schemi di codifica e introdurre un metalinguaggio per descriverli. <URL: http://www.tei-c.org/>

[7] L. BURNARD and C. M. SPERBERG-MCQUEEN, *Living with the Guidelines: An introduction to TEI tagging*, Text Encoding Initiative, Document Number: TEI EDW18, March 13, 1991, p. 2 (corsivi aggiunti).

[8] J. H. COOMBS, A. H. RENEAR and S. J. DEROSE, *Markup Systems and the Future of Scholarly Text Processing*, "Communications of the ACM", 30, 1987, p. 934.

[9] D. BUZZETTI and M. REHBEIN, *Textual Fluidity and Digital Editions*, in M. DOBREVA (ed.), *Text Variety in the Witnesses of Medieval Texts*, Proceedings of the International Workshop (Sofia, 21-23 September 1997), Sofia, Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences, 1998, p. 31.

[10] G. GIGLIOZZI, *Il testo e il computer: Manuale di informatica per gli studi letterari*, Milano, Bruno Mondadori, 1997, p. 109.

dall'ansia dell'umanista di fronte alla tecnologia? Non pare, se è vero che qualche perplessità si incontra anche sul versante dell'informatica. Da un lato, infatti, "in termini informatici, il testo può essere considerato come la forma più fondamentale di informazione"; dall'altro "l'elaborazione automatica del testo", il testo inteso come tipo di dato, "non è necessariamente priva di complicazioni". È vero che "ogni altro tipo di informazione può essere codificato come testo" specialmente "in fase di *input* e di *output*" e che "quasi tutti i programmi eseguiti su un *computer* hanno bisogno di elaborare testo"; ma è anche vero che "lo *hardware* dà spesso l'impressione di essere stato progettato avendo in mente soprattutto l'elaborazione numerica" e che "la manipolazione di caratteri può sembrare un ripensamento dei progettisti, una cosa possibile, ma certamente non agevole". Così, "se i caratteri creano problemi, le stringhe di caratteri ne creano ancora di più", e se "la loro allocazione in memoria fa nascere delle difficoltà, la loro elaborazione ne solleva in misura ancora maggiore"; per non dire che "i comuni linguaggi di programmazione di solito offrono pochi strumenti per l'elaborazione del testo". Eppure "non c'è forma di rappresentazione dell'informazione così universale e generale come il testo".[11] Da dove nasce, allora, il problema?

2. *La rappresentazione digitale del testo*

Si potrebbe dire, ricorrendo alle categorie dell'umanista, dallo scarto tra la sintassi e la semantica o, se vogliamo, tra la forma logica e la forma grammaticale dell'espressione che costituisce il dato, ossia la rappresentazione dell'informazione. Le espressioni simboliche sono in generale isomorfe alle relazioni logiche tra i contenuti rappresentati, esibiscono immediatamente la forma logica degli enunciati, eliminano lo scarto tra la forma logica e la forma grammaticale dell'espressione. Da queste considerazioni aveva tratto impulso l'orientamento simbolico delle correnti ormai storiche dell'intelligenza artificiale, un punto di vista a cui si sono prevalentemente ispirate le tecniche cognitivistiche di rappresentazione della conoscenza, che ne hanno di conseguenza riproposto le rigidità e le difficoltà di applicazione esaustiva. Ma la rappresentazione digitale del testo ha sollevato di nuovo il problema della divergenza tra la sintassi, ov-

[11] C. DAY, *Text Processing* cit., pp. 2-5.

vero la forma strutturale della *rappresentazione* dell'informazione (la struttura del *dato*), e la semantica, ovvero la forma delle relazioni strutturali tra i contenuti espressi da tale rappresentazione (la struttura del *modello*). In sostanza, con la rappresentazione digitale, era nata una forma di rappresentazione del testo che poteva essere elaborata automaticamente, prescindendo dalla formalizzazione delle relazioni logiche dei contenuti rappresentati. La sintassi si era di nuovo svincolata dalla semantica, senza che se ne avesse immediatamente chiara consapevolezza. Con l'affermarsi del *WorldWideWeb*, la rivalsa del testo sulle forme simboliche di rappresentazione della conoscenza è ora sotto gli occhi di tutti e di nuovo l'applicazione delle tecniche dell'intelligenza artificiale all'analisi e all'elaborazione delle conoscenze si misura con la più tradizionale delle forme di rappresentazione dei contenuti del sapere. Si pensi ai motori di ricerca e si pensi a quello che si sta verificando con la nuova proposta del *Semantic Web*. Ma come si è giunti a questo punto? e soprattutto, ai fini del nostro discorso, non ci troviamo forse di fronte ad una nuova rivalsa del sapere umanistico, il sapere tradizionalmente fondato sullo studio del testo?

2.1 *Rappresentare la struttura*

All'origine di tutto sta, si può dire, la povertà strutturale della stringa di caratteri, il tipo fondamentale di rappresentazione dell'informazione testuale in *machine-readable form*. Finché l'informazione testuale dev'essere sottoposta, come succede "in fase di *input* e di *output*", all'occhio dell'uomo, non sorge nessun problema. È alla mente umana che spetta il compito di elaborarla. L'informazione testuale è così di fatto relegata al ruolo di interfaccia. Ma può l'informazione testuale essere considerata come priva di struttura, per qualunque operazione, anche la più semplice, che si voglia eseguire col *computer*? Il problema si è posto, concretamente, nella progettazione dei "sistemi per la preparazione dei documenti".[12] Nei sistemi di prima generazione, "il testo da stampare" era "inframmezzato da comandi interpretati dal *software* che produceva il docu-

[12] R. FURUTA, *Concepts and Models for Structured Documents*, in J. ANDRÉ, R. FURUTA, and V. QUINT (eds.), *Structured Documents*, Cambridge, Cambridge University Press, 1989, p. 8.

mento fisico". Ma coi sistemi di seconda e terza generazione è emersa la tendenza alla "strutturazione dei documenti".[13] Nei primi sistemi, il *markup*, o l'"informazione aggiuntiva" inserita "a scopo di impaginazione e stampa",[14] era espressa attraverso "linguaggi procedurali"; ma poi, "con l'introduzione di linguaggi dichiarativi", qualcosa è radicalmente cambiato: i "simboli usati per il *markup* non costituiscono più istruzioni di formattazione, bensì marcatori di struttura"[15] e la "struttura logica" del documento, descritta dal *markup*, viene convertita nel suo "modello fisico" da un'operazione successiva di "formattazione".[16] Si è allora pensato di aver realizzato una "separazione tra il *contenuto informativo* del documento e il suo formato [fisico]",[17] e si è sostenuto che il *markup* "non fa parte del testo, o del *contenuto* dell'espressione", ma che "ci dice qualcosa intorno al testo", o al suo contenuto.[18]

Ma è veramente possibile pensare che il testo, inteso come pura sequenza di caratteri, coincida col suo contenuto informativo? o che la struttura logica assegnata al documento da forme di *markup* descrittivo esprima le relazioni logiche tra gli elementi del suo contenuto? Ecco di nuovo un'inavvertita assimilazione della struttura logica (sintattica) della rappresentazione dell'informazione, alla struttura logica (semantica) dell'informazione rappresentata e una pericolosa confusione tra le *strutture di dati* ottenute attraverso il *markup* del testo e il *modello di dati* applicabile all'elaborazione del suo contenuto informativo. E la confusione pare dovuta alla mancata percezione dello scarto tra la rappresentazione digitale, algoritmicamente manipolabile, dell'informazione testuale espressa mediante stringhe di caratteri e il modello delle relazioni formali tra gli elementi del suo contenuto, uno scarto che l'umanista, abituato a misurarsi con l'indeterminatezza dell'interpretazione del testo, difficilmente riesce a trascurare. Eppure su questo equivoco si è fondata anche

[13] V. JOLOBOFF, *Document Representation: Concepts and standards*, in ANDRÉ *et al.*, *Structured Documents* cit., pp. 76 e 86.

[14] D. T. BARNARD, C. A. FRASER, and G. M. LOGAN, *Genealized Markup for Literary Texts*, "Literary and Linguistic Computing", III, 1988, p. 27.

[15] V. JOLOBOFF, *Document Representation* cit., p. 87.

[16] R. FURUTA, *Concepts and Models* cit., p. 11.

[17] J. M. SMITH, *SGML and Related Standards*, New York, Ellis Horwood, 1992, p. 143 (corsivo aggiunto).

[18] J. H. COOMBS *et al.*, *Markup Systems* cit., p. 934 (corsivo aggiunto).

la scelta della TEI di adottare lo SGML (*Standard Generalized Markup Language*), il linguaggio di *markup* dichiarativo approvato come standard ISO nel 1986, quale linguaggio ufficiale per la codifica dei testi letterari.[19] E si è pensato che la struttura del *documento*, definita formalmente dalla DTD (*Document Type Definition*) prevista dallo standard SGML, potesse adeguatamente rappresentare anche la struttura del *testo*, di cui il documento strutturato costituisce soltanto l'espressione materiale. L'*espressione* e il *contenuto*, gli elementi costitutivi del testo, che la tradizione linguistica tiene chiaramente distinti,[20] sono stati erroneamente identificati e alla perentoria domanda *Che cos'è il testo, in realtà?*[21] si è creduto di poter dare un'altrettanto perentoria risposta, considerando il testo "come una 'gerarchia ordinata di oggetti dotati di contenuto', ovvero una 'OHCO'(*Ordered Hierarchy of Content Objects*)". Questa definizione si spiega facilmente, se si assume che un "oggetto dotato di contenuto" sia una porzione di documento, che contiene o è contenuta in altri "oggetti dotati di contenuto", o porzioni di documento, tra i quali si dà una 'gerarchia' di relazioni di inclusione, i cui elementi ultimi sono "ordinati" in successione nella sequenza lineare di caratteri codificati dai quali il documento è materialmente costituito. L'assunzione del modello OHCO come "modello fondamentale" per la rappresentazione del testo si fondava dunque sull'erronea identificazione della struttura testuale

[19] SGML (*Standard Generalized Markup Language*): l'ISO 8879 (1986) è uno standard internazionale per la descrizione e la definizione di linguaggi di codifica per la rappresentazione digitale del testo in modo indipendente da ogni dispositivo e da ogni sistema. Per essere SGML-conforme uno schema di codifica deve: (1) dichiarare un insieme di marcatori mediante una DTD (*Document Type Definition*) che specifica quali elementi possono essere codificati e le loro reciproche relazioni; (2) non deve permettere sovrapposizioni tra gli elementi, che debbono essere inclusi l'uno nell'altro secondo una struttura gerarchica.

HTML (*HyperText Markup Language*): è una DTD o un'applicazione SGML per definire gli elementi dei documenti accessibili sul *Web*.

XML (*eXtensible Markup Language*): è un profilo di applicazione o una forma ristretta – ma non meno potente – di SGML, che definisce un formato universale per documenti e per dati strutturati accessibili sul *Web*.

[20] Cfr. C. SEGRE, *Avviamento all'analisi del testo letterario*, Torino, Einaudi, 1985, pp. 49-50.

[21] Così suona il titolo di un noto e ampiamente discusso intervento di S. J. DEROSE, D. D. DURAND, E. MYLONAS, and A. H. RENEAR, *What is Text, Really?*, "Journal of Computing in Higher Education", 1:2, 1990, pp. 3-26.

con la struttura assegnata dallo standard SGML alla sequenza di caratteri che costituisce la base della sua rappresentazione digitale e non poteva mancare di sollevare obiezioni.[22]

2.2 *Documenti e informazione strutturata*

L'attività concreta di codifica del testo, praticata secondo le norme della TEI, si è in effetti trovata di fronte a seri "problemi pratici", che hanno quindi costretto a "mettere in discussione" il modello OHCO, rivelatosi chiaramente inadeguato al lavoro filologico e critico degli studiosi.[23] Ma anche da parte informatica si è giunti ad una riconsiderazione della natura e del ruolo del *markup* nella rappresentazione digitale del testo. L'avvento dei linguaggi di *markup* dichiarativi è stato salutato dalla *document community*, la comunità che comprende tanto gli studiosi quanto i tipografi interessati all'elaborazione automatica del testo,[24] come un mezzo per dotare meri "flussi di testo digitale contenenti caratteri ortografici", quello che può essere considerato puro e semplice "materiale digitale (*digital stuff*)", di "attributi strutturali o testuali". I linguaggi di *markup* dichiarativi venivano quindi concepiti come lo strumento più idoneo per "rappresentare dati testuali come informazione", ossia per trasformare sequenze di caratteri codificati prive di struttura in informazione strutturata "accessibile all'interno di ambienti di elaborazione intelligenti".[25] E se c'è molto di vero in queste affermazioni, si avvertiva tuttavia l'esigenza di analizzare meglio la natura dei documenti strutturati, distinguendo in modo più preciso tra la struttura dei *dati* testuali e la struttura dell'*informazione* testuale ad essi associata.

Col rappresentare "la struttura di un documento piuttosto che il

[22] Ibid., p. 6.

[23] A. Renear, E. Mylonas, D. Durand, *Refining our Notion of What Text Really Is: The problem of overlapping hierarchies*, in N. Ide and J. Veronis (eds.), *Text Encoding Initiative: Background and contexts*, Dordrecht, Kluwer, 1995, p. 263.

[24] D. Raymond, F. Tompa and D. Wood, *From Data Representation to Data Model: Meta-semantic issues in the evolution of SGML*, "Computer Standards and Interfaces", X, 1995, p. 3 <URL: http://www.csd.uwo.ca/staff/drraymon/.papers/sgml.ps>.

[25] R. Cover, N. Duncan, and D. T. Barnard, *The Progress of SGML (Standard Generalized Markup Language): Extracts from a comprehensive bibliography*, "Literary and Linguistic Computing", VI, 1991, pp. 197-198.

suo aspetto", il *markup* dichiarativo rende possibile un "approccio" che "tratta i documenti come *database*, e non come manufatti la cui unica funzione sia quella di essere visualizzati". Le conseguenze di questo fondamentale mutamento di prospettiva si sono rivelate molto profonde. Infatti, non si è ottenuto solo di poter trattare i documenti come *database*, ma si è presentata anche l'opportunità di fare il contrario, ossia la possibilità di rappresentare e di elaborare ogni tipo di struttura informativa come un documento, vale a dire come un insieme di dati strutturati di natura testuale. All'interesse della *document community* per l'implementazione del *markup* dichiarativo, si è così aggiunto l'interesse della *database community*, ossia di coloro che si occupano professionalmente dell'elaborazione dei dati nella forma che viene ora sempre più spesso descritta come "tradizionale".[26] E questo ha portato anche a riconsiderare e ridefinire in termini più adeguati la natura e la funzione specifica del *markup*.

Ma seguiamo gli orientamenti dei due diversi modi di procedere. L'interesse della *document community* era volto a far sì che "le tecniche disponibili per l'elaborazione di oggetti rigorosamente definiti come i programmi e i *database*" potessero "essere usate anche per l'elaborazione dei documenti".[27] Tuttavia, nell'affrontare i problemi che nascevano dall'assenza di una "netta separazione dell'implementazione dall'applicazione" nei sistemi in uso, prevaleva la preoccupazione di assicurare la "permanenza" dei "documenti" nel trasferimento dei dati da un sistema all'altro. Invece, nella *database community*, per risolvere problemi analoghi derivanti dalla dipendenza delle applicazioni da "informazioni dettagliate sul formato fisico delle registrazioni (*records*) dei dati", prevaleva la preoccupazione di salvaguardare non tanto la "permanenza" della rappresentazione dei dati quanto la stabilità della "semantica delle applicazioni". Sicché mentre la *document community* "sceglieva di standardizzare la rappresentazione dei dati" generalizzando il *markup*, la *database community* "sceglieva di standardizzare la semantica dei dati" sviluppando modelli relazionali e deduttivi dotati di una semantica rigorosamente definita.[28] Mentre in un caso si assicurava l'*invarianza della rappresentazio-*

[26] D. RAYMOND *et al.*, *From Data Representation to Data Model* cit., p. 2 e segg.

[27] CH. F. GOLDFARB, *Introduction to Generalized Markup*, Annex A to ISO standard 8879, 1986, p. 60.

[28] D. RAYMOND *et al.*, *From Data Representation to Data Model* cit., pp. 4-5.

ne dell'informazione, nell'altro si assicurava l'*invarianza del contenuto* informativo, un fatto su cui sarà necessario ritornare.

2.3 *La teoria del* markup

Che cosa se ne traeva a proposito del *markup*? Come si è già ricordato il *markup* consiste, materialmente, nell'"inserimento di speciali ed esclusive combinazioni di caratteri, i cosiddetti *codici*, all'interno della lunga stringa di caratteri" che rappresenta il testo.[29] Da un punto di vista concettuale, invece, il *markup* può essere considerato come "la denotazione di posizioni specifiche" nella stringa dei caratteri mediante l'inserimento di certi "contrassegni (*tokens*)" e quindi come "l'uso di codici, detti marcatori (*tags*), inseriti in un documento, per descriverne la struttura".[30] Ma di quale struttura si tratta? Quella dell'*espressione*, o quella del *contenuto* del testo? La parola *cane* non morde, dicevano già gli Stoici, non solo perché il segno non possiede le proprietà dell'oggetto designato, ma anche perché le proprietà strutturali del segno non sono isomorfe alle proprietà strutturali di ciò che designa. La stringa di caratteri che costituisce la parola *cerchio* non ha forma circolare. Inoltre è difficile "mantenere una chiara e netta distinzione tra il testo stesso e la sua codifica", perché l'operazione che porta a "esplicitare la 'struttura' mediante i codici" è fondamentalmente ambigua: infatti, "non appena qualcosa è stato esplicitato diventa parte del testo, che quindi cambia ed assume una nuova struttura".[31] Corre qui alla mente una proposizione del *Tractatus* di Wittgestein, secondo cui "ciò, che nel linguaggio esprime *sé*, *noi* non possiamo esprimere mediante il linguaggio"; in questo senso, proprio come si dice che "la proposizione non può rappresentare la forma logica", allo stesso modo possiamo dire che il dato, la rappresentazione dell'informazione, non può rappresentare la sua struttura: nell'un caso e

[29] C. HUITFELDT, *Multi-Dimensional Texts in a One-Dimensional Medium*, "Computers and the Humanities", XXVIII, 1995, p. 236.

[30] D. RAYMOND, F. W. TOMPA and D. WOOD, *Markup Reconsidered*, paper presented at the First International Workshop on Principles of Document Processing, Washington DC, October 22-23, 1992 <URL: http://www.csd.uwo.ca/staff/drraymon/.papers/markup.ps>.

[31] C. HUITFELDT, *Multi-Dimensional Texts* cit., p. 237.

nell'altro, il dato, o la proposizione, "mostra" o "esibisce" la sua forma logica, o la sua struttura.[32] Ma ancora una volta, se si parla della struttura del testo, a che cosa si allude? alla sua espressione o al suo contenuto? I due elementi costitutivi del testo non possono essere confusi e occorre specificare chiaramente a quali proprietà strutturali si vuole fare riferimento con l'inserimento del *markup*.

Ed era proprio la specificazione più attenta delle caratteristiche del *markup* che poteva aiutare a rispondere a questo tipo di domande. Rispetto al testo, inteso come successione di caratteri, lo "*status* logico del *markup* è ambivalente":[33] i marcatori, infatti, "fanno parte del testo, eppure possono esserne distinti". Tutti i "tipi di *markup*" sono "forme di struttura", la cui "caratteristica essenziale" è quella di "essere *inserita* (*embedded*) e *separabile*" dal testo.[34] L'"ambiguità del *markup*" è un tratto essenziale della sua natura linguistica.[35] Da un lato, le proprietà del *markup* sono le proprietà stesse della *forma logica* della proposizione: la forma logica non è rappresentata dalla proposizione, ma si mostra o si esibisce con essa. Allo stesso modo, una volta inseriti nel testo, i marcatori ne divengono essi stessi la struttura: non la descrivono, ma la costituiscono; non la rappresentano, ma la esibiscono. Dall'altro lato, i marcatori possono essere distinti e separati dal testo e possono così essere considerati come una forma di notazione metalinguistica per descriverne gli aspetti strutturali. Il *markup* quindi, nello stesso tempo, "è *rappresentazione* della struttura", "ed *è* esso stesso struttura".[36] Sicché, da un punto di vista linguistico, può essere considerato e trattato in modo ambivalente: può es-

[32] L. WITTGENSTEIN, *Tractatus logico-philosophicus*, 4.121, in *Tractatus logico-philosophicus e Quaderni 1914*-1916, trad. it. di Amedeo G. Conte, Torino, Einaudi, 1964, pp. 28-29.

[33] D. BUZZETTI, *Ambiguità diacritica e* Markup*: Note sull'edizione critica digitale*, in S. ALBONICO (ed.), *Soluzioni informatiche e telematiche per la filologia*, Atti del Seminario di studi (Pavia, 30-31 marzo 2000), Pavia, Università degli Studi di Pavia, 2000 (Pubblicazioni telematiche, n.1), <URL: http://dobc.unipv.it/dipslamm/Pubblicazioni/telematiche/dino_buzzetti.htm>.

[34] D. RAYMOND *et al.*, *Markup Reconsidered* cit., pp. 2-3.

[35] D. BUZZETTI, *Ambiguità diacritica* cit.

[36] ID., *Rappresentazione digitale e modello del testo*, in *Il ruolo del modello nella scienza e nel sapere*, Atti del Convegno (Roma, 27-28 ottobre 1998), Roma, Accademia Nazionale dei Lincei, 1999, (Contributi del Centro Linceo Interdisciplinare "Beniamino Segre", N. 100), p. 152.

sere considerato e trattato come una descrizione metalinguistica della struttura del testo, oppure come un'estensione delle risorse espressive del linguaggio naturale, che ne rende esplicita l'intrinseca "metalinguisticità riflessiva".[37]

2.4 *Le forme della testualità e il* markup

Questa ambivalenza del *markup*, chiaramente riscontrabile da un punto di vista linguistico, esprime una delle caratteristiche fondamentali del testo ed è propria di tutte le forme di notazione diacritica, alle quali può essere senz'altro assimilato. E come si vede, le esigenze della rappresentazione digitale del testo costringono a renderne esplicita una delle caratteristiche essenziali. Da un punto di vista informatico, infatti, il *markup* "non è né un automa computazionale, né un modello di dati, né un formalismo matematico"; il *markup* "appartiene" invece, fondamentalmente, "al mondo delle rappresentazioni"[38] e costituisce quindi una proprietà "essenzialmente notazionale" del testo.[39] Da un punto di vista logico, infine, è possibile dimostrare l'equivalenza di asserzioni metalinguistiche sulla struttura del testo e asserzioni ad esse corrispondenti contenenti forme di predicazione di ordine superiore formulate in linguaggio-oggetto. E, di nuovo, l'equivalenza e la convertibilità delle due forme di espressione degli aspetti strutturali del testo, quella che li rappresenta e quella che li mostra, possono essere considerate come una delle proprietà fondamentali che caratterizzano la natura ultima della testualità. Sicché si può dire, in sostanza, che l'inserimento dei marcatori nella sequenza dei caratteri che costituisce la rappresentazione digitale del testo permette di esplicitare l'insieme delle funzioni logico-linguistiche svolte dalla tecnologia alfabetica nel porsi come forma di espressione primaria dell'informazione testuale. La fondamentale ambivalenza diacritica del *markup* può far sembrare particolarmente arduo il compito di assegnare al testo una struttura ben definita, ma a ben vedere, conservare alla rap-

[37] Cfr. T. De Mauro, *Minisemantica dei linguaggi non verbali e delle lingue*, Bari, Laterza, 1982, pp. 93-94 e *Prima lezione sul linguaggio*, Bari, Laterza, 2002, pp. 89 e 91-93.

[38] D. Raymond *et al.*, *Markup Reconsidered* cit., p. 4.

[39] D. Buzzetti, *Rappresentazione digitale e modello del testo* cit., p. 146.

presentazione digitale qualche aspetto dell'elusiva ambiguità del testo letterario, può alla lunga recare, al di là delle difficoltà evidenti, anche qualche vantaggio. Seguiamo ancora, per cercare di metterli in luce, le ulteriori evoluzioni del *markup*.

Ciò significa che per essere veicolo adeguato di informazione testuale la rappresentazione digitale non può assolutamente spogliarsi delle funzioni e delle prerogative più proprie della testualità ordinaria. Ora, l'inserimento di elementi strutturali nel testo attraverso i marcatori può essere di due tipi: può accadere che "la posizione [del marcatore] all'interno del dato esprima informazione", oppure, al contrario, che il marcatore "sia informativo, ma che la sua collocazione all'interno del testo non esprima informazione". Si è così potuto parlare, rispettivamente, di *markup* "inserito in modo vincolato o non vincolato (*strongly or weakly embedded*)" nel testo.[40] Ora, è certo possibile ricorrere al *markup* vincolato (*strongly embedded*) per rappresentare adeguatamente, salvo casi particolari,[41] gli aspetti strutturali dell'espressione del testo, ma non è in generale possibile applicarlo in modo altrettanto soddisfacente alla rappresentazione strutturale del suo contenuto informativo. La struttura espressa dal *markup* vincolato dipende infatti dal formato dei dati, perché il *markup* vincolato "condivide coi dati la forma di rappresentazione" e questa condizione "gli rende difficile esprimere una struttura che non sia un sottoinsieme di posizioni di caratteri all'interno del testo". La struttura espressa dal *markup* non vincolato (*weakly embedded*) non dipende invece dal formato dei dati, perché il *markup* non vincolato può essere collocato "in qualsiasi punto all'interno del testo, o anche al di fuori del testo, senza perdere il suo significato".[42] Per questa ragione viene anche descritto come *markup* "a parte" (*standoff*) o "non in linea" (*out-of-line*)[43] ed è "più propriamente considerato come una forma specifica di struttura esterna".[44] Queste caratteristiche gli per-

[40] D. RAYMOND *et al.*, *Markup Reconsidered* cit., pp. 3-4.

[41] Difficoltà insuperabili sorgono, nella preparazione di un'edizione critica, per la rappresentazione di varianti testuali che si sovrappongono l'una all'altra generando strutture non lineari all'interno del testo (cfr. D. BUZZETTI and M. REHBEIN, *Textual Fluidity and Digital Editions*, cit.).

[42] D. RAYMOND *et al.*, *Markup Reconsidered* cit., pp. 9 e 4.

[43] C. M. SPERBERG-MCQUEEN, C. HUITFELDT, and A. RENEAR, *Meaning and Interpretation of Markup*, "Markup Languages: Theory & Practice", II, 2000, p. 230.

[44] D. RAYMOND *et al.*, *Markup Reconsidered* cit., p. 4.

mettono quindi di rappresentare strutture di tipo non lineare e non gli impongono limiti a priori nella descrizione degli aspetti strutturali del contenuto del testo. In breve, il *markup* vincolato sembra più adatto a rappresentare le proprietà strutturali dipendenti dall'espressione del testo, mentre il *markup* esterno o non vincolato sembra invece essere adatto a rappresentare anche le proprietà strutturali del suo contenuto informativo.

2.5 *Il* markup *e l'indeterminazione strutturale del testo*

Seguire questi dettagli serve a mettere in luce difficoltà di fondo. Infatti, che relazione si può dare tra i due tipi di struttura? Di nuovo può essere utile richiamare considerazioni di filosofia del linguaggio e ricordare che anche a proposito della forma logica degli enunciati di cui si compone un testo si possono distinguere due aspetti affatto corrispondenti. Da un lato possiamo dire, per citare Donald Davidson, che "quando riscriviamo gli enunciati in qualche forma standardizzata, l'inferenza risulta semplificata e meccanizzata", e sappiamo che la validità delle inferenze dipende esclusivamente dalla forma sintattica degli enunciati e non dal loro contenuto semantico. Dall'altro, "dare forma logica ad un enunciato significa", sempre secondo Davidson, "descriverlo in termini tali da condurlo entro l'ambito di una teoria semantica". Ora, l'identificazione dei due aspetti della forma logica o, per analogia, dei due tipi di struttura del testo, presuppone l'uso di un linguaggio formale tale da eliminare lo scarto "tra forma logica e grammatica di superficie" degli enunciati.[45] Sicché la funzione del *markup* potrebbe essere proprio quella di ridurre questo scarto. In questa direzione sembra in effetti orientarsi, da una parte, lo sviluppo dei linguaggi di *markup* e, dall'altra, l'evoluzione di quella che costituisce di fatto la forma più diffusa di testualità digitale, il *WorldWideWeb* – un'evoluzione propiziata proprio dall'introduzione di un nuovo linguaggio di codifica, lo *eXtensible Markup Language* (XML).

Anche negli sviluppi della riflessione teorica sui linguaggi di *markup* – un campo di interesse recentemente emergente e sviluppatosi al punto

[45] D. Davidson, *Action and Reaction*, "Inquiry", XIII, 1970) pp. 203, 210 e 220.

da promuovere la pubblicazione di una nuova rivista[46] interamente dedicata ad ospitarne il dibattito – le difficoltà non sembrano del tutto scomparse. All'obiezione che per la *document community* il termine "modello di dati" non significa, come per la *database community*, "un linguaggio comune per descrivere vincoli sui dati e l'effetto di operazioni su quei dati", ma significa invece "un linguaggio comune per esprimere la struttura dei dati", è stato risposto che benché si dica, di conseguenza, che "lo SGML non è un modello di dati, perché non definisce nessun operatore",[47] in realtà esso "fornisce un modello, per la rappresentazione e l'elaborazione del testo", costituito dalla "struttura ad albero (*tree structure*)" assegnata al documento. La DTD (*Document Type Definition*) collegata al documento specifica infatti una "grammatica libera dal contesto (*context-free grammar*)" che "fornisce un linguaggio di specificazione dei vincoli (*constraint language*)" operanti su tale struttura gerarchica. Ma è chiaro che questo modello di dati applicabile alla struttura dell'espressione del testo (la struttura ad albero del documento) non coincide in generale col modello di dati applicabile alla struttura del suo contenuto informativo. Non sempre infatti è possibile stabilire una "corrispondenza biunivoca tra singole caratteristiche testuali e singoli elementi" della struttura del documento definiti dalla codifica SGML.[48] Un caso evidente è costituito dalla sovrapposizione di diverse strutture gerarchiche all'interno del testo, come ad esempio i versi e i costrutti grammaticali in un componimento poetico, e se esistono metodi praticabili per definire "strutture di dati plausibili per rappresentare documenti con sovrapposizioni",[49] non esiste tuttavia "nessun metodo di specificazione dei vincoli gravanti sull'interconnessione delle distinte DTD" applicate alle diverse strutture gerarchiche sovrapposte.[50]

Ma come può un documento avere strutture diverse? Di nuovo entra in gioco la relazione tra l'espressione e il contenuto e la difficoltà può

[46] Cfr. "Markup Languages: Theory & Practice", ed. by B. T. USDIN and C. M. SPERBERG-MCQUEEN, Cambridge, Mass., MIT Press, 1999.

[47] D. RAYMOND *et al.*, *From Data Representation to Data Model* cit., p. 6.

[48] C. M. SPERBERG-MCQUEEN and C. HUITFELDT, *Concurrent Document Hierarchies in MECS and SGML*, "Literary and Linguistic Computing", XIV, 1999, pp. 30-31.

[49] ID., *GODDAG: A Data Structure for Overlapping Hierarchies*, in *ACH-ALLC'99 Conference Proceedings*, Charlottesville VA, University of Virginia, 1999, p. 198.

[50] ID., *Concurrent Document Hierarchies* cit., p. 41.

essere ricondotta all'intrinseca indeterminazione strutturale del testo. Ancora una volta la consapevolezza dell'umanista sembra essere in grado di orientare la ricerca delle soluzioni informatiche. "Nessun testo poetico", ma a ben vedere nessun testo, "può esistere senza sistemi di 'strutture sovrapposte'", scrive Jerome McGann in un illuminante saggio dedicato all'impatto delle nuove tecnologie dell'informazione sull'idea stessa di testualità. Nel campo testuale "non è possibile assumere nessuna unità come identica a se stessa".[51] L'indeterminazione strutturale del testo si presenta come un carattere costitutivo della sua stessa natura. Un fatto di cui esiste piena consapevolezza anche da parte informatica. Un modello di dati, si riconosce, "è un'interpretazione del mondo", ma "molti testi sono essi stessi mondi da interpretare e quindi la necessità di modelli è potenzialmente infinita" e "ogni modello del testo può riferirsi a diverse strutture". Ma quali sono le conseguenze di quest'indeterminazione? Se ne può solo trarre che "la capacità del *markup* di soddisfare quest'esigenza è seriamente ostacolata"?[52] Ed è davvero impossibile trovare soluzioni?

In realtà, a fronte dell'indeterminazione si dà sempre un'*invarianza*: se è vero che la struttura del testo può essere definita come "l'assieme delle relazioni latenti" sussistenti tra le sue parti[53] e che il numero delle possibili determinazioni strutturali del testo è potenzialmente infinito, è anche vero che è alla *stessa* espressione che possono corrispondere contenuti diversi ed è allo *stesso* contenuto che possono corrispondere espressioni diverse. Sicché "il campo di variabilità del contenuto è vincolato dall'identità dell'espressione e il campo di variabilità dell'espressione è vincolato dall'identità del contenuto".[54] L'identità del contenuto comporta la *sinonimia* delle sue diverse espressioni e l'identità dell'espressione ne comporta la *polisemia* rispetto ai suoi diversi contenuti. Ma "la relazione uno/molti tra identità dell'espressione e varianza del contenuto si può convertire nella relazione uno/molti tra varianza dell'espressione e identità del contenuto" ed è proprio l'ambivalenza linguistica del *markup*

[51] J. McGann, *Radiant Textuality: Literature after the WorldWideWeb*, New York, Palgrave, 2001, pp. 188-189 (cfr., per la traduzione italiana, Id., *La letteratura dopo il World Wide Web*, trad. it. di A. Ferrara e A. Stumpo, Bologna, Bononia University Press, 2002, p. 209).

[52] D. Raymond *et al.*, *Markup Reconsidered* cit., p. 14.

[53] C. Segre, *Avviamento all'analisi del testo letterario* cit., p. 44.

[54] D. Buzzetti, *Ambiguità diacritica*, cit.

che permette di trasferire l'invarianza dall'una all'altra componente del testo. Intesi come espressioni diacritiche del linguaggio oggetto, i marcatori possono essere considerati come varianti testuali distinte o espressioni sinonime di un unico modello. Viceversa, intesi come descrizioni metalinguistiche associate ad una stessa espressione polisemica i marcatori possono essere riferiti a modelli interpretativi diversi. Sicché "il *markup* può trasformare le varianti interpretative in varianti testuali e le varianti testuali in varianti interpretative", diventando così "uno strumento per trasformare la varianza implicita dell'interpretazione di un'espressione identica nella fluidità esplicita dell'espressione di un identico contenuto".[55]

3. *Markup e Intelligenza Artificiale (IA)*

A queste che possono sembrare a prima vista fumose astruserie care al critico letterario, non paiono tuttavia essere così distanti le concrete soluzioni informatiche che si stanno profilando attraverso l'applicazione delle tecniche dell'intelligenza artificiale alla rappresentazione digitale del testo. Ed è proprio il carattere essenzialmente indeterminato della rappresentazione del testo e lo scarto evidente tra la struttura sintattica assegnata alla successione dei caratteri e il modello semantico che le può essere associato a rendere possibile un'applicazione efficace e pervasiva di tecniche confinate in origine ad essere impiegate soltanto per rappresentazioni della conoscenza rigorosamente formalizzate. La chiave della soluzione che si prospetta pare consistere nel livello *pre-linguistico* di formalizzazione della rappresentazione digitale del testo. La duttilità e l'indeterminazione strutturale di tale rappresentazione si rivela del tutto idonea all'implementazione di procedure d'analisi dell'informazione testuale che sono state considerate da sempre patrimonio esclusivo della tradizione ermeneutica umanistica. Ma quali sono le tecniche dell'intelligenza artificiale che, allo stato delle cose, possono essere applicate alla rappresentazione codificata del testo? Due esempi paiono particolarmente significativi: la concezione del *markup* come regola di inferenza e l'implementazione del *Semantic Web*, orientata alla gestione automatica del contenuto informativo dei documenti testuali accessibili in rete.

[55] ID., *Rappresentazione digitale e modello del testo* cit., p. 156.

3.1 *Il* markup *come regola di inferenza*

Una svolta significativa nelle discussioni sul *markup* è ora costituita dalla recente "proposta" di introdurre un metodo "per descrivere il significato del *markup*".[56] La proposta nasce da un'attenta riflessione sulla "funzione del *markup*" (215) nel documento codificato e tende a individuare le caratteristiche essenziali di una forma di "rappresentazione" formale del suo significato. Ma a che cosa mira tutto ciò? Che cosa significa "usare una rappresentazione del *markup* e del significato del *markup*" (217)? e soprattutto in che cosa consiste il significato del *markup* e perché si avverte la necessità di rappresentarlo? Quello che dichiaratamente si vuole ottenere è un metodo per "identificare il significato del *markup* usato in un documento" e si sostiene che a questo proposito "è sufficiente generare l'insieme delle inferenze riguardanti il documento che sono autorizzate (*licensed*) dal *markup*"; anzi, che "in un certo senso, possiamo considerare il significato del *markup* come costituito, e non solo descritto, dall'insieme di quelle inferenze" (231). Anche in questo caso si impongono alcune considerazioni di filosofia del linguaggio. Ma innanzi tutto si deve osservare che il *markup*, in quanto tale, viene introdotto per esibire certe "caratteristiche" del "materiale testuale" o, più precisamente, di certi "passi" del materiale codificato (215). E se già il *markup*, in sé e per sé, esprime queste caratteristiche, perché mai si sente il bisogno di rappresentarlo, o di rappresentarne a sua volta il significato? Una prima risposta viene dalla natura stessa della forma di rappresentazione utilizzata a questo scopo. Si tratta di una "rappresentazione in Prolog" (217) e la scelta indica che lo scopo è quello di tradurre i "costrutti" (225) del linguaggio di *markup* in un linguaggio formale adatto alla rappresentazione della conoscenza. Quindi, da questo punto di vista, il *markup* non fa altro che esplicitare quelle caratteristiche del testo che possono ricevere una rappresentazione formale in una base di conoscenza costituita da asserzioni in Prolog e il metodo proposto non è altro che un semplice processo di formalizzazione di una rappresentazione digitale del testo preventivamente strutturata mediante l'inserimento del *markup*. Dunque è proprio il *markup* ciò che permette di applicare al testo procedimenti di elaborazione tipici dell'intelligenza artificiale. La funzione specifica del

[56] C. M. Sperberg-McQueen *et al.*, *Meaning and Interpretation of Markup* cit., p. 231.

markup diviene così quella di colmare lo scarto tra la struttura sintattica della rappresentazione del testo, costituita da stringhe di caratteri codificati, e la struttura semantica della rappresentazione della conoscenza descritta in modo formale da asserzioni in Prolog, o in ogni altro linguaggio formale impiegato a questo scopo. E già il riscontro di queste caratteristiche è molto istruttivo circa la natura specifica del *markup*.

Ma c'è dell'altro nell'affermazione che il significato del *markup* consiste nelle inferenze che autorizza. Con ciò il *markup* si presenta come una *regola* di inferenza, o un'*inference licence*. L'espressione può essere stata usata senza particolari intenzioni, ma certo non è nuova per un filosofo del linguaggio. Essa rimanda alla cosiddetta teoria dell'*inference-ticket* o dell'*inference warrant* introdotta da filosofi come Gilbert Ryle e Stephen Toulmin.[57] Nella sostanza, e per quello che qui ci interessa, un'*inference licence* "viene usata come un titolo di inferenza (*inference-ticket*), una specie di abbonamento, che autorizza (*licenses*) il suo possessore a passare dall'asserzione di certe proposizioni di fatto all'asserzione di altre proposizioni di fatto"; inoltre, essa appartiene "ad un diverso e più sofisticato livello di discorso" di quello a cui appartengono le affermazioni di fatto alle quali si riferisce, un po' come "le proposizioni dell'algebra si situano su un livello di discorso diverso da quello delle proposizioni aritmetiche che le soddisfano".[58] E se questa è la natura specifica del *markup*, lo si può intendere come l'espressione di regole, o leggi, riguardanti le proprietà del testo, formulate in linguaggio oggetto per mezzo di asserzioni contenenti forme di predicazione autoriflessive di ordine superiore.[59] E sappiamo che il contenuto di tali asserzioni può essere riformulato metalinguisticamente per mezzo di altre e diverse asserzioni che non esibiscono, ma rappresentano o descrivono le proprietà strutturali del testo. La "rappresentazione del significato del *markup*" di cui qui si tratta può così essere intesa come la riformulazione di una proprietà del testo, che il *markup* esibisce in forma diacritica e autoriflessiva, in un'asserzione esterna, o metalingustica, espressa in un linguaggio formale di rappresentazione

[57] Cfr. G. RYLE, *The Concept of Mind*, London, Hutchinson, 1949 e S. TOULMIN, *The Uses of Argument*, Cambridge, Cambridge University Press, 1958.

[58] G. RYLE, *The Concept of Mind*, 2nd ed., Harmondsworth, Penguin Books, 1963, pp. 116-117.

[59] Cfr. D. BUZZETTI, *Ambiguità diacritica*, cit.

della conoscenza. E ciò fa consistere la funzione del *markup* e dei suoi costrutti sintattici nell'esplicitazione delle proprietà strutturali del testo che possono essere rappresentate in modo formale in un sistema di gestione della conoscenza. Tale funzione specifica permette anche, all'inverso, di esprimere in forma diacritica e autoriflessiva le descrizioni formali esterne delle proprietà strutturali del testo.

3.2 Web *semantico e* Markup

Ed è proprio questo, concettualmente, il processo che viene implementato nel cosiddetto *Semantic Web*,[60] il cui scopo è quello di "conferire struttura al contenuto semantico delle pagine *Web*".[61] La struttura assegnata al contenuto informativo di un documento *WWW* è stabilita facendo riferimento a specifiche "ontologie", che ne definiscono il modello astratto. Un'ontologia, intesa nel senso dei "teorici dell'intelligenza artificiale (IA) e del *Web*", è materialmente una risorsa, ossia "un documento o un *file*" univocamente identificato in rete attraverso un URI (*Universal Resource Identifier*), contenente la specificazione "che definisce formalmente le relazioni tra i termini" usati nelle pagine *Web*. Il *Web* semantico è quindi, nella sostanza, il tentativo di assegnare un modello formale al contenuto del *Web*. E il ruolo di mediazione tra il modello e i documenti è svolto proprio dal *markup*. L'introduzione dello XML ha reso possibile l'uso di un linguaggio comune per ridurre a schema tanto la struttura dei documenti, o dell'*espressione* dei testi, quanto la struttura del loro *contenuto* e il problema da risolvere consiste nel mettere in rapporto lo schema che descrive la struttura assegnata

[60] Il *Semantic Web* è un'estensione del *Web* attualmente esistente per potenziarne la natura e trasformarlo da luogo dove l'informazione è semplicemente mostrata, a luogo dove l'informazione possa essere interpretata, scambiata ed elaborata tanto dalle macchine quanto dall'uomo. Associando ai documenti un significato formalmente definito, il *Web* non collegherà solamente i documenti gli uni con gli altri, ma permetterà anche alle macchine di riconoscerne ed elaborarne il contenuto. Programmi che non erano stati progettati per essere tra loro compatibili potranno condividere ed elaborare dati precedentemente non amalgamabili.

[61] T. BERNERS-LEE, J. HENDLER, and O. LASSILA, *The Semantic Web*, in "Scientific American", May 2001, <URL: http://www.sciam.com/2001/0501issue/0501berners-lee.html>.

al formato dei documenti con lo schema che descrive la struttura assegnata direttamente al loro contenuto. Nel *Web* la prima funzione è svolta da uno *Schema XML*, che "è un documento che descrive il formato valido di un insieme di dati (*data-set*) XML",[62] mentre la seconda è svolta dallo *Schema RDF*, che può essere considerato come "un modello logico di asserzione".[63] In entrambi i casi, attraverso lo schema, si tratta di assegnare al testo dei documenti *Web* un modello di dati, ma ancora una volta, nel primo caso si è condizionati dai vincoli imposti dalla struttura dei documenti, mentre nel secondo caso se ne può prescindere. Nel primo caso lo schema espresso dal *markup* XML è riferito alle relazioni tra le diverse porzioni del documento, mentre nel secondo lo schema espresso dal *markup* XML è costituito da una struttura indipendente ed esterna.

Con riferimento al testo ordinario, l'espressione del significato è vincolata alle forme espressive, sintatticamente coerenti, del linguaggio impiegato, mentre sui suoi contenuti astratti si possono compiere operazioni definite da un formalismo che permette di rappresentarli, prescindendo dalla forma superficiale delle asserzioni testuali che li riguardano. Ridurre a schema la struttura dei documenti e, rispettivamente, la struttura dei loro contenuti significa esibirne la forma logica in un linguaggio formale la cui semantica definisce il formalismo, o il modello di dati, che può essere loro applicato. Ed è il testo stesso ciò attraverso cui stanno in relazione tra loro lo schema del documento e lo schema del suo contenuto ed è il testo ciò a cui essi fanno riferimento e assegnano direttamente una struttura. Nel caso dei documenti *Web*, codificati in XML, è invece la struttura sintattica assegnata al testo dal *markup* ciò a cui fanno riferimento gli schemi che ne esibiscono la forma logica astratta. Uno schema XML svolge rispetto al *markup* espresso in questo linguaggio la stessa funzione che "una notazione formale per descrivere la sintassi di un dato linguaggio",[64] la BNF (*Backus Naur Form*) ad esempio, svolge rispetto al

[62] I. STUART, *XML Schema, A Brief Introduction*, <URL: http://lucas.ucs.ed.ac.uk/xml-schema>.

[63] T. BERNERS-LEE, *Design Issues: Architectural and philosophical problems*, <URL: http://www.w3.org/DesignIssues/Architcture.html>.

[64] TH. ESTIER, *What is BNF Notation*, <URL: http://cui.unige.ch/db-research/Enseignement/analyseinfo/AboutBNF.html>.

testo.[65] In questo modo, però, la struttura esibita dal *markup* si rende opaca rispetto alla struttura sottostante del testo e le uniche forme strutturali assegnabili al testo restano quelle espresse dallo schema mediante costrutti XML. Sono i costrutti metalinguistici dello schema, formulabili nello stesso linguaggio di *markup*, che divengono così il luogo della trasformazione dello schema del documento nello schema del suo contenuto. Ed è da questo punto di vista che occorre valutarne l'adeguatezza rispetto alla capacità di espressione dei contenuti testuali.

3.3 *Testualità e Intelligenza Artificiale*

La sintassi XML è una sintassi formale e come tale elimina l'indeterminatezza, quella dell'espressione rispetto al contenuto e quella del contenuto rispetto all'espressione, che è propria del testo e della sua rappresentazione digitale non codificata. Ora, "un documento XML-schema è esso stesso un documento XML".[66] Gli schemi XML sono quindi "documenti XML" con "struttura ad albero", che definiscono "strutture di dati" assegnabili al documento concreto al quale si riferiscono, e che operano "come un meccanismo di scambio per dati strutturati" tra il documento concreto e le "applicazioni orientate ai dati" che ne possono elaborare il contenuto informativo.[67] Ma i diversi schemi applicabili ad un documento restano descrizioni metalinguistiche e non possono essere considerati come espressioni sinonime varianti di un unico contenuto. Allo stesso modo, diverse ontologie riferite al documento concreto attraverso il modello logico assertivo assegnatogli dallo schema RDF, non possono essere considerate contenuti varianti di un'unica espressione polisemica. La serializzazione di dati strutturati di varia natura ottenuta attraverso una rappresentazione XML si approssima solo in modo imperfetto alla natura propria della testualità ordinaria. Se è vero che il *markup* può essere considerato l'espressione in forma diacritica di un aspetto autoriflessivo del testo e quindi assimilato al testo, non è altrettanto vero

[65] Cfr. J. CLARK, *RELAX NG and W3C XML Schma*, <URL: http://www.imc.org/ietf-xml-use/mail-archive/msg00217.html>.

[66] I. STUART, *XML Schema*, cit.

[67] W3C, *The Cambridge Communiqué*, <URL: http://www.w3.org/TR/1999/NOTE-schema-arch-19991007>.

che il testo possa essere fatto coincidere con la sua rappresentazione codificata in XML e quindi assimilato alla struttura formale assegnatagli dal *markup*.

Anche in questo caso, le perplessità non vengono solo da parte dell'umanista preoccupato dell'eclissarsi, nel modello formale proposto, di una delle dimensioni più caratteristiche del testo. Lo stesso Ted Nelson, il conclamato inventore del "termine 'ipertesto'", usato "per esprimere l'idea della scrittura/lettura non lineare in un sistema informatico",[68] denuncia l'incapacità del *Web* semantico di "afferrare le sottigliezze dei problemi testuali". A suo giudizio, lo XML "non migliora le cose" per quanto riguarda le difficoltà poste dai "formati di *markup* vincolato". Inoltre, il riferimento dello schema XML al documento concreto è metalinguistico e la sua introduzione genera una gerarchia discreta di livelli di rappresentazione. Allo stesso modo, le ontologie di riferimento possono essere associate alla descrizione RDF della struttura assertiva del documento concreto attraverso linguaggi, quali il DAML+OIL (ricavato dal *DARPA Agent Markup Language* sviluppato per conto della *Defence Advanced Research Projects Agency* degli USA e dall'*Ontology Interchange Language* sviluppato nel quadro dello *Eropean Union's Information Technologies Program*), che sono estensioni, o nuovi 'vocabolari' XML, ma il loro impiego articola l'architettura del *Web* semantico secondo una struttura a strati definita sarcastimente da Nelson come una sorta di "*hamburger* gerarchico".[69] Lo stesso ideatore del *Web* d'altra parte, Tim Berners-Lee, tiene a dire che "benché i documenti possano essere trasformati, rappresentati, analizzati e indicizzati automaticamente, l'idea che siano *compresi* è un problema che appartiene completamente all'intelligenza artificiale e che non fa parte di quelli che interessano l'architettura del *Web*". Sicché "quando si parla di comprensione automatica dei documenti" si intende semplicemente parlare di "dati predisposti esplicitamente per il ragionamento automatico: ossia appartenenti ad una rete (*web*) semantica".[70] Ma "è tuttavia uno strano fenomeno il fatto che si tenda a lasciar perdere il titolo di Intelligenza Artificiale (IA) proprio quando i problemi affrontati dai suoi teorici possono in una certa misura

[68] P. LÉVY, *Les technologies de l'intelligence: L'avenir de la pensée à l'ère informatique*, Paris, La Découverte, 1990, p. 34.

[69] T. NELSON, *I don't buy in*, <URL: http://ted.hyperland.com/buyin.txt>.

[70] T. BERNERS-LEE, *Design Issues: Architectural and Philosophical Problems*, cit.

essere risolti e producono sistemi concreti". Di fatto, si riscontra "una discreta convergenza tra quello che (alcuni) studiosi di IA amano fare, quello che gli studiosi di biblioteconomia desiderano, e quello che richiede il *Web* semantico", sicché, è stato sostenuto, "il *Web* semantico è", veramente, "un progetto di Intelligenza Artificiale".[71] Piuttosto si tratta di valutare se le procedure applicate nel *Web* semantico risultano pienamente adeguate per l'elaborazione automatica della rappresentazione digitale del testo e il problema pare non consistere tanto nell'accettare o respingere la qualifica di intelligenza artificiale, quanto nel considerare quale forma specifica di intelligenza artificiale possa rivelarsi la più idonea allo scopo.

Il fenomeno dell'indeterminazione strutturale e della "legge di compensazione" tra "la determinazione e l'indeterminazione della struttura dell'espressione e del contenuto del testo"[72] può essere rappresentato come un endomorfismo

$$(A = A \Leftrightarrow A \neq A) \Leftrightarrow A \overset{f}{\rightarrow} A$$ [73]

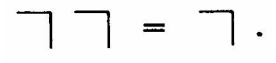
tra elementi strutturali costitutivi del testo. La formulazione della legge come relazione tra identità del tutto e distinzione prodotta dalla partizione primaria del testo in espressione e contenuto, espressa dal primo membro dell'equivalenza, può essere descritta anche attraverso il concetto di *forma della distinzione* introdotto formalmente dal matematico inglese George Spencer-Brown col suo *calcolo delle indicazioni*. L'*indicazione* dell'espressione, o l'*indicazione* del contenuto, presuppongono la loro distinzione, prodotta dalla partizione primaria operata sulla totalità del testo; l'*indicazione* dell'espressione, o del contenuto, li pone come sottounità del testo identiche a se stesse e ne determina la struttura; e la determinazione e l'identità dell'espressione, o del contenuto, con se stessi sono espresse formalmente dalla legge di idempotenza dell'espressione, o del contenuto, rispetto alla loro rappresentazione – primo assioma di Spencer Brown, o *law of calling*, e *forma della condensazione* (1). A sua volta, la relazione tra la forma logica dell'espressione assegnata al testo dal

[71] B. Parsia, *An Introduction to Prolog and RDF*, <URL: http://www.xml.com/pub/a/2001/04/25/prologrdf/index.html>.

[72] D. Buzzetti, *Ambiguità diacritica*, cit.

[73] Id., *Rappresentazione digitale e modello del testo* cit., p. 156.

markup e il modello strutturale associato al suo contenuto può essere considerata un'esemplificazione del secondo assioma di Spencer-Brown, o *law of crossing*, e della *forma della cancellazione* (2). Infatti, il riferimento dell'articolazione strutturale di una delle due sottounità testuali alla totalità strutturale dell'altra ne toglie l'identità con se stessa e ne produce l'indeterminazione.[74]



(1)

(2)

A che pro richiamare questa inconsueta rappresentazione del fenomeno dell'instabilità e dell'indeterminazione strutturale del testo? Al fine di mostrare che il testo può essere considerato la realizzazione materiale di operazioni mentali di natura autopoietica e che proprio la sua materialità e la manipolabilità automatica della sua rappresentazione digitale lo rende idoneo all'applicazione di procedure proprie dell'intelligenza artificiale. Il testo digitale può essere considerato, così, come una macchina algoritmica per la simulazione di operazioni autopoietiche, operazioni proprie di un sistema che "genera e specifica continuamente la propria organizzazione funzionando come un sistema di produzione dei propri componenti".[75] Il concetto di autopoiesi è stato introdotto nel campo delle scienze biologiche da Humberto Maturana e Francisco Varela[76] ed ha influenzato in modo rilevante lo sviluppo delle scienze cognitive e dell'intelligenza artificiale.[77] Non è qui luogo di argomentare diffusamente sulla legittimità di questa prospettiva e sulla sua rilevanza per le questioni affrontate in questa sede. Ci limitiamo soltanto a segnalare che l'estensione del calcolo delle indicazioni di Spencer-Brown pro-

[74] Cfr. G. SPENCER-BROWN, *Laws of Form*, London, Allen and Unwin, 1969.

[75] F. J. VARELA, *Principles of Biologicaal Autonomy*, New York, North-Holland, 1979, p. 13.

[76] Cfr. H. R. MATURANA e F. J. VARELA, *Autopoiesi e cognizione: la realizzazione del vivente* (1980), trad. it. di A. Stragapede, Venezia, Marsilio, 1985.

[77] Cfr. C. DOHERTY, *Reconstructing AI*, in B. McMullin (ed.), *Autopoiesis and Perception*, Dublin, 1992, pp. 137-144 <URL: http://www.univie.ac.at/constructivism/books/mcmullin92/>.

posta da Francisco Varela permette di ottenere rappresentazioni formali dei processi di autoriferimento propri dei sistemi autopoietici che risultano applicabili alla descrizione dei fenomeni espressi dalle notazioni diacritiche autoriflessive del testo.

Varela presenta il calcolo delle indicazioni di Spencer-Brown come "un formalismo per rappresentare l'atto della distinzione", atto di cui l'"indicazione" costituisce il valore.[78] Tale calcolo "può essere considerato come il fondamento della teoria dei sistemi, proprio quanto i matematici possono considerare la teoria degli insiemi come il fondamento del loro campo" di ricerca; e "a questo livello semplice, ma fondamentale di descrizione sistemica, i problemi di circolarità e di autocomputazione divengono più trasparenti e possono essere rappresentati più chiaramente" (120). In questo contesto, un processo autoreferenziale può essere considerato "come un'auto-indicazione" (109), ovvero come "una forma che *rientra* nel suo spazio indicazionale, che informa se stessa". Invocando un'analogia geometrica, possiamo paragonare l'auto-indicazione ad una bottiglia di Klein, "dove l'interno e l'esterno sono irrimediabilmente confusi" (122). Nel calcolo, ogni indicazione è espressa dallo stesso "nome" (110) o "segno"(*token*), che può essere inteso in due modi, come un'operazione o un "atto di distinzione", o come il "valore" dell'indicazione (111). Quest'ambivalenza è "la chiave per risolvere quelli che solitamente sono sembrati preoccupanti paradossi" della circolarità o dell'autoriferimento (113). Così, una forma rientrante può essere assunta come un valore o una forma, oppure come una "prescrizione" (124) o una regola. Il *markup*, nella sua forma diacritica autoriflessiva, può essere considerato come un caso particolare di forma rientrante. In quanto valore, o forma, fa parte del testo; in quanto operatore, può essere inteso come una regola espressa in linguaggio oggetto per fare inferenze riguardanti l'espressione o il contenuto del testo. L'estensione del calcolo delle indicazioni ottenuta con l'introduzione dell'auto-indicazione permette di tener conto di "situazioni autoreferenziali" (137) e offre così una forma di rappresentazione adeguata per il *markup* diacritico.

Può essere proprio questa la forma di rappresentazione associata vantaggiosamente al *markup* per descriverne il nesso con la dinamica della determinazione strutturale dell'espressione e del contenuto del testo e

[78] F. J. VARELA, *Principles of Biological Autonomy* cit., pp. 106-107.

con la forma della loro reciproca compensazione. Si tratta soltanto dell'indicazione di una direzione di ricerca, che pare tuttavia promettente di esiti positivi anche per la costruzione di quel "*database* globale"[79] di documenti testuali che vuol essere il *Web* semantico nell'intenzione dei suoi promotori.

[79] T. BERNERS-LEE, *Design Issues: Architectural and Philosophical Problems*, cit.