UPGRADE
The European Journal for the Informatics Professional
http://www.upgrade-cepis.org

Edizione italiana a cura di ALSI e Tecnoteca
http://upgrade.tecnoteca.it

La lingua nei media digitali: una sfida

di José-Antonio Millàn

Traduzione italiana, a cura di Paola M Hayward, dell'articolo
" Language in the Digital Media: A Political Challenge"
pubblicato sul vol. IV, No. 3, June 2003 della rivista online UPGrade, a cura del CEPIS

"Si tratta di sapere - disse Unto Dunto - chi ha da essere il padrone. Questo è tutto."
(Lewis Carroll, *Attraverso lo specchio*)

Riassunto
Il bene di tutti, che è la lingua, diventa una merce alienabile nell'era dei media digitali. Tuttavia, non vi è ragione per cui chi parla lingue diverse dall'inglese, o coloro che parlano lingue minoritarie o varianti minoritarie di lingue vastamente usate, non possa beneficiare dei numerosi vantaggi che l'uso di una lingua nei media digitali può apportare: interfacce vocali, supporti per la traduzione, ecc. Questo articolo propone misure di politica linguistica che consentirebbero alla ricerca già in atto (molta della quale finanziata da denaro pubblico) di sviluppare programmi che servirebbero ai bisogni della società senza aumentarne la dipendenza tecnologica.

Parole chiave*:* Licenze GPL, politiche linguistiche, lingue minoritarie, interfacce vocali, traduzione.

1. Introduzione

In una società dove i media digitali sono fortemente dominanti, si mostra evidente la tendenza ad incorporare il linguaggio naturale nella comunicazione automatizzata tra sistemi e l'individuo, e tra individui che parlano lingue diverse. Ciò accade per un'ovvia ragione: perché la lingua è un sistema di comunicazione che la società ha dispensato a tutti quanti ed è un sistema di uso quotidiano. Il linguaggio non solo è la più comune interfaccia ma anche la più sofisticata: non esiste un menù con opzioni o una mappa da cliccare capace di fornire tutte le possibilità contenute in una semplice frase. Per quanto riguarda la lingua parlata, persino le persone con problemi di istruzione o coloro che non sanno usare un mouse o una tastiera sarebbero in grado di spiegare a un sistema automatico, uno ben concepito, ciò che vogliono.

Ma quando si arriva al mondo digitale, il linguaggio, un bene di tutti creato dalla collettività, gratuito e dagli usi illimitati, diventa un bene alienabile. Per comprendere e interagire con l'uomo, le macchine devono essere dotate di programmi il cui processo è lungo e costoso e richiede l'esistenza di un sistema strutturato di dati (corpora, dizionari). E anche se tali programmi esistessero o noi fossimo disposti e nella posizione di sostenerne i costi, è molto probabile che non prenderebbero in considerazione tutte le complessità delle nostre società. In questo articolo mi propongo di spiegare i principi che dovrebbero influenzare le future politiche pubbliche a tale proposito.

Inizierò con una descrizione impressionistica dell'uso futuro del linguaggio software, che ho introdotto precedentemente in un altro contesto [1]: *«Che tipo di sistemi il linguaggio userà come interfaccia? Qualsiasi tipo: sistemi di inserimento dati in generale (dai palmari, agende elettroniche, ai sistemi professionali), sistemi di commercio elettronico (che ricerchi prodotti con determinate caratteristiche e che evidenzi descrizioni e confronti), sistemi per attività ricreative (luoghi di spettacoli, ristoranti, punti di informazione turistica), istruzione e formazione (apprendimento automatico e sistemi di valutazione), o ricerca (ricerca di materiali, accesso 'intelligente' a database).»*

Sempre più spesso useremo questo tipo di programmi, talvolta senza rendercene conto. Avranno capacità multilinguistiche e saranno in grado di formulare ipotesi sul grado di interesse di informazioni, di tradurre (con diversi gradi di accuratezza) e di estrapolare riassunti. Saranno i nostri strumenti di lavoro intellettuali e professionali.

2. L'industria linguistica

Tali sistemi contribuiranno alla nascita di un importante settore economico. Ma, come nel caso dello spagnolo, ne deriverà che maggiore sarà la dipendenza tecnologica di quelle nazioni dove si parla spagnolo e contribuirà ad appesantire ulteriormente la loro bilancia dei pagamenti [2]. La verità è che per molte lingue, anche quelle parlate in molte nazioni, si dovrà pagare un costo per utilizzarle sulla rete, e per quelle lingue minoritarie o varianti di queste, non ci sarà possibilità di scelta perché semplicemente non esisteranno programmi per utilizzarle. In un futuro potremo comperare un dizionario computerizzato dei sinonimi di spagnolo dalla Spagna o di francese dalla Francia, quale accessorio di un word processor, ma non potremo ottenerne uno di francese senegalese o di spagnolo boliviano, indipendentemente da quanto possa essere la nostra offerta.

Non riusciamo a spiegarci il perché tale debba essere la situazione, visto che per molte lingue si investono ricerche e risorse, nella maggioranza dei casi finanziate da fondi pubblici, che potrebbero fornire le basi per uno sviluppo di programmi software linguistici a tutti livelli (sia generali che locali). Perché la colonizzazione andrà a limitare quasi completamente questo importante e strategico settore industriale per lingue quali lo spagnolo, il francese e il portoghese? Perché occorrerà così tanto tempo prima che si possa riconoscere l'esistenza di questo settore, se mai accadrà, per le lingue europee meno parlate nel mondo?

Nel caso di alcune lingue la ragione potrebbe essere che non c'è mai stata abbastanza ricerca di base, per ragioni storiche e per mancanza di risorse e fondi erogati a università e altre istituzioni. Ma per lo spagnolo, il portoghese, il francese o l'italiano, ciò si spiega perché i rispettivi governi non hanno mai avuto alcuna politica che regoli il linguaggio digitale. Questo è, in particolare, un argomento spinoso, poiché comprende due aree problematiche che, in termini generali (e qui mi riferisco allo spagnolo), i governi mancano di conoscenze e di volontà politica: politiche linguistiche e politiche digitali. Per quanto riguarda le politiche linguistiche, il potere generalmente non ha coscienza neppure che possa esistere una tale politica (fatta eccezione per quelle comunità autonome con una lingua propria la quale finisce per diventare un'arma politica), e neppure non si comprende appieno l'importanza sociale della questione digitale (da qui la presente inadeguata legislazione in Spagna, il modo poco chiaro di gestione del dominio spagnolo .es, il perpetuamento dei monopoli nel campo delle comunicazioni, e via dicendo).

3. Gli obbiettivi per una politica del linguaggio digitale

Quali dovranno essere gli obbiettivi per una politica del linguaggio digitale?

Garantire che le risorse (corpora e programmi di sviluppo) e sistemi strutturati di dati utlizzati dai sistemi automatici (per esempio dizionari) siano resi disponibili per nuovi progetti.

Aumentare il numero degli agenti impegnati nello sviluppo di software linguistici, in modo da migliorare la qualità e la quantità delle opzioni.

Facilitare l'incorporazione di un software linguistico per lingue minoritarie o varianti locali di lingue largamente diffuse.

In sostanza questi tre punti si possono ridurre a uno solo: consentire l'uso di risorse e dati attraverso una licenza all'utente che assicurerà che i prodotti derivati siano allo stesso modo *accessibili* e *riutilizzabili* a loro volta. Questi sono obbiettivi fondamentali: nel caso dello spagnolo, e forse anche nel caso di altre lingue, esistono risorse in istituzioni pubbliche e storiche che incontrano ostacoli imprevisti quando tali risorse sono richieste per l'utilizzo in progetti. Si dice che certe risorse siano state «rese accessibili» su Internet, ma ciò significa che possono essere solo consultate: un corpus può fornire all'utente un numero di ricorrenze per parola, o analizzare un sistema morfologico. Ma tale uso non è sufficiente per progetti di sviluppo. Quando si parla di *risorse accessibili* si intende che l'intera risorsa sia resa disponibile su DVD o in qualche altro sistema di storaggio per chiunque ne faccia richiesta. Alla fine di questo articolo si considereranno le possibili obiezioni a tale metodologia di lavoro. Per quanto riguarda il secondo punto, il *riutilizzo* si assicura ponendo le risorse all'interno di una GPL (*General Public License* o Licenza pubblica generale) [3] o tipi di licenze cosiddette *Creative Commons* [4].

La situazione attuale (almeno per quanto riguarda lo spagnolo) è che le risorse linguistiche dei centri di ricerca pubblici non giungono in modo trasparente a tutte quelle società che potrebbero utlizzarle, ma solo ad un numero limitato di società, in pratica quelle che sviluppano programmi per gli utenti finali. Una politica efficace sarebbe quella il cui fine è affidare le risorse per lo sviluppo di strumenti linguistici nelle mani di qualsiasi tipo di istituzione, sia pubblica che privata, alla portata di chiunque voglia creare software linguistici. L'opinione comune è che tale compito è riservato solo alle società più importanti (in particolare quelle americane), anche se in realtà, sia in termini di programmi di dati e sviluppo, sia per piccoli programmi ad uso degli utenti, esistono diversi tipi di progetti di sviluppo che potrebbero essere intrapresi, alcuni dei quali con un alto grado di specificità. Per esempio, dizionari specialistici, scritti e orali, per complementare il lessico con convertitori testo/discorso e sistemi di conversazione [5].

4. Quali sono i modi per perseguire una politica del linguaggio digitale?

Diversi sono i modi per ottenerla. Uno potrebbe essere quello di creare una Banca delle risorse linguistiche, gratuita e aperta a qualsiasi ente o individuo che voglia sviluppare un software (regolato dai tipi di licenze menzionati in precedenza) per assicurare che i risultati dell'uso di tali risorse siano ugualmente accessibili e riutilizzabili.

Per costituire una banca delle risorse, l'approccio più realistico consisterebbe nell'ottenere la licenza da quelle istituzioni (università o privati) per l'utilizzazione delle loro risorse linguistiche e dati. Potrebbe sembrare un paradosso, ed infatti lo è, che i risultati di ricerche finanziate con denaro pubblico debbano essere riacquistate a vantaggio della comunità. Tuttavia questa sembra essere la soluzione più pratica rispetto ad altre. Quale azione parallela, dato che una banca delle risorse sarebbe di proprietà pubblica e intesa a realizzare il bene comune, si potrebbe iniziare una campagna per incoraggiare le istituzioni a cedere, escludendo la vendita, le proprie risorse alla banca senza alcuna richiesta di pagamento.

Che scopo avrebbero tali banche delle risorse? Dovrebbero basarsi sulle lingue piuttosto che sulle nazioni. Numerose lingue europee sono diffuse in un grande numero di nazioni, in continenti diversi, come nel caso del francese, portoghese o spagnolo, e sarebbe assurdo limitare solo in base alla variante europea della lingua. Una Banca Linguistica spagnola, per esempio, idealmente dovrebbe conglobare risorse provenienti dal maggior numero di varianti possibili.

In una società interconnessa e multilinguistica come la nostra, potremmo persino cercare di allargare la portata dei benefici di tale azione, estendendola a lingue diverse. Per fare ciò, si potrebbe promuovere le risorse che beneficiano delle similarità esistenti tra molte lingue, per esempio, tra le lingue del ceppo latino: spagnolo, francese, catalano, italiano, per creare nuclei morfologici, sintassi, lessicografia, e così via, comuni a tutte loro.

Il modello funzionale di tali risorse potrebbe basarsi su quello del Linguistic Data Consortium [6] o dell'European Language Resources Association [7].

In precedenza ho citato possibili obiezioni che tale azione potrebbe sollevare: *"Copiare un corpus o un dizionario morfologico in nessun modo li svalorizza. Se tutti gli agenti che desiderano lavorare su progetti linguistici possono ottenere liberamente i risultati di questa indispensabile ricerca basilare, il massimo che può accadere è che in breve tempo potremmo avere una proliferazione di programmi per il riconoscimento di parole, analisi di frasi, e altro. Molti di questi programmi non saranno direttamente fruibili dagli utenti finali, ma potrebbero formare una parte di più elaborati sistemi, e il risultato finale sarà un maggior numero di sistemi, diversificazione di sistemi e sistemi più economici che utilizzino la nostra lingua"* [1].

Di questa semplice proposta che non richiede grandi finanziamenti, chiaramente ne beneficerebbe la società; si tratta di una soluzione che promuove le potenzialità delle società e dei gruppi degli utenti, a solo svantaggio degli interessi delle grandi società. Una proposta che desse il controllo su un settore strategico che coinvolge necessariamente le nostre istituzioni e i cittadini, dovrebbe essere prontamente abbracciata dai rilevanti organi di governo.

Ringraziamenti

Questo articolo ha tratto spunto dalle discussioni tra esperti di lingua portoghese, francese e spagnola, durante le conferenze *Tres Espacios Lingisísticos* (Tre Spazi Linguistici), nel 2001/2002, organizzate dalla *Organisation Internationale de la Francophonie, Organización de Estados Iberoamericanos, Comunidades dos Paises de Lingua Portuguesa, Unión Latina y Secretaria de Cooperación Iberoamericana* <http://www.jamillan.com/tresespa.htm>. Vorrei ringraziare Daniel Pimienta e Isabel Trancoso per i loro contributi e Daniel Prado per il suo costante appoggio.

Note

[1] José Antonio Millan, *"El español en la sociedad digital. una propuesta"* (La lingua spagnola nella società digitale: una proposta), un contributo al *Congreso Internacional de la Lengua Española*, Valladolid, dal 16 al 19 ottobre 2001. <http://cvc.cervantes.es/obref/congresos/valladolid/mesas_redondas/millan_j.htm>

[2] José Antonio Millàn, *"La lengua que era un tesoro "* (La lingua che era un tesoro), 28 marzo 2001,<http://www.jamillan.com/tesoro.htm> e la versione parziale in inglese "How much is a language worth?

A Quantification of the Digital Industry for the Spanish Language". (Quanto vale una lingua? Una quantificazione dell'industria digitale per lingua spagnola). <http://www.jamillan.com/worth.htm>

[3] <http://www.gnu.org/copyleft/gpl.html>

[4] <http://creativecommons.org/.>

[5] Questo tipo di attività in cui, per esempio, il gergo usato in neurobiologia in portoghese o il lessico dell'ingegneria messicana viene incorporato in un sistema preesistente, richiede per prima cosa, che tali sistemi siano accessibili e modificabili e secondariamente che ci sia collaborazione al fine di creare un corpus.

[6] <http://www.ldc.upenn.edu/>

[7] <http://www.icp.grenet.fr/ELRA/home.html>

Notizie sull'autore

José Antonio Millàn, linguista e pioniere *editor* digitale in Spagna, ha diretto l'edizione CD-ROM del Dizionario della Accademia Reale Spagnola (1995), <http://www.rae.es/>, ed è il creatore del Centro Virtual Cervantes (1997), <http://cvc.cervantes.es/portada.htm>. In qualità di consulente si è impegnato in progetti di sviluppo con molte importanti istituzioni, ha lavorato sulla politica linguistica nella cornice di *«Los tres espacios lingúísticos»* (Tre Spazi Linguistici), 2001-2002. È autore dei libri *"Internet y el espanol»* (Internet e la lingua spagnola) (2001) e *"De redes y saberes. Cultura y educación en las nuevas tecnologías"* (1998).

Di particolare interesse è il suo sito personale, <http://jamillan.com>, in lingua spagnola. La sua email è: jam@jamillan. com