ANNOTAZIONE GRAMMATICALE E LEMMATIZZAZIONE DI *CORPORA* IN ITALIANO

FABIO TAMBURINI
Università degli Studi di Bologna
CILTA - Centro Interfacoltà di Linguistica Teorica e Applicata "L. Heilmann"

1. Introduzione

Questo lavoro si inserisce nell'ambito del progetto CORIS[1] per la costruzione di un *corpus* di riferimento per l'italiano scritto. Attualmente la realizzazione di un *corpus* di riferimento comporta numerosi problemi che vanno oltre la costituzione del *corpus*. Dimensioni che superano il centinaio di milioni di parole (o *token*) sono divenute ormai lo standard minimale per *corpus* di riferimento, e richiedono quindi opportune metodologie di supporto al lavoro del ricercatore. Effettuare ricerche in *corpora* di queste dimensioni pone problemi, specialmente con lingue ricche di forme flesse come l'italiano: è possibile infatti ottenere, come risultato di una ricerca, una quantità di concordanze largamente superiore a quella gestibile da un essere umano. Esistono numerose tecniche per campionare efficacemente le concordanze ottenute (estrazione casuale, estrazione di una ogni N, ecc.), ma non risolvono vari problemi di fondo che possono notevolmente complicare l'analisi dei risultati ottenuti. Supponiamo per esempio che il ricercatore voglia fare uno studio sul verbo 'amare'; le forme flesse del verbo 'amo' e 'ami' corrispondono anche al singolare e al plurale del sostantivo 'amo'. È evidente quindi che senza un meccanismo per la selezione, durante l'impostazione della ricerca, di opportuni criteri restrittivi, ogni calcolo di frequenza sulle concordanze ottenute viene invalidato.

Ci è sembrato quindi utile introdurre nel progetto CORIS opportune metodologie che permettano di specificare parametri aggiuntivi nella ricerca dei termini, per esempio rispetto alle parti del discorso (*Part-of-speech tag*) o nei confronti dei lemmi che generano ogni singola parola del *corpus*. Inserire questo tipo di informazioni permette di effettuare ricerche molto più mirate, per esempio rispetto alle forme flesse di un verbo, senza le interferenze generate da eventuali omografi. Sottolineiamo come meccanismi di questo tipo siano ormai richiesti dalla comunità scientifica, per *corpora* di queste dimensioni. È altrettanto importante sottolineare come l'aggiunta di informazioni possa, in linea di principio, corrompere le informazioni stesse contenute nel *corpus*. Si rende quindi necessario mantenere queste informazioni aggiuntive ben separate da

[1] Per la descrizione del progetto si rinvia al contributo di R. Rossini Favretti, in questo volume.

quelli che sono i testi originali che compongono il *corpus* e utilizzarle solo in caso di necessità preservando l'originalità del *corpus* stesso.

Questo lavoro presenta un *work-in-progress* per la relizzazione di un annotatore grammaticale (*tagger*) per la lingua italiana, che permetta di assegnare le parti del discorso e lemmatizzare il CORIS. Annotare grammaticalmente un *corpus* significa, data la sequenza di parole che lo compongono, assegnare ad ognuna di esse la corretta parte del discorso, risolvendo tutte le ambiguità intrinseche nel processo di classificazione grammaticale del linguaggio (Figura 1).

Il	**successo**	**fa**	**male**	**.**
Articolo	*Verbo PP*	*Verbo*	*Aggettivo*	*Punt*
	Nome comune	*Nome comune*	*Nome comune*	
			Avverbio	

FIGURA 1 – *A molte parole, esaminate singolarmente, possono essere assegnate più parti del discorso. Un programma di* tagging *deve risolvere queste ambiguità esaminando il contesto nel quale la parola è inserita.*

Il problema è stato affrontato in molteplici modi negli ultimi decenni, e numerosi sono i lavori bibliografici prodotti. Le tecniche principali coinvolgono modelli stocastici di Markov con apprendimento supervisionato (DeRose 1988; Rabiner 1989, Charniak *et al.* 1993; Charniak 1996; Kempe 1993, Schmid 1994a; Carlberger / Kann 1999, Brants 2000) o autonomo (Rabiner 1989; Cutting *et al.* 1992), annotatori basati su regole (Brill 1992, 1994), sistemi a reti neurali (Benello *et al.* 1989; Schmid 1994b; Marques / Lopes 1996), modelli *Maximum Entropy* (Ratnaparkhi 1996), modelli *Memory-based* (Zavrel / Daelemans 1999a) ecc. Tutti questi lavori presentano annotatori grammaticali con performance piuttosto elevate, che variano dal 95 al 98% di parole correttamente classificate. Questi valori risultano essere lo stato dell'arte nel settore, anche se è molto complesso confrontare annotatori applicati a differenti lingue o che utilizzano differenti insiemi di parti del discorso. Vi sono infatti risultati di confronti tra alcuni di questi metodi (prevalentemente tra i metodi stocastici e quelli basati su regole) che esprimono opinioni contraddittorie (Chanod / Tapainen 1995; Volk / Schneider 1998). Vi sono inoltre studi che riguardano la possibilità di utilizzare uno o più di questi metodi per migliorare il rendimento dei *tagger* grammaticali (Tapanainen / Voutilainen 1994). Alcuni dei lavori precedenti sono stati adattati per annotare la lingua italiana (Cutting *et al*. 1992; Schmid 1994a). Altri lavori, principalmente nel nostro paese, sono stati predisposti e sviluppati specificamente per la lingua italiana: si segnalano De Mauro *et al.* (1993) e il *PiTagger* (Picchi 1994) per quanto riguarda i *tagger* stocastici, mentre l'équipe di Delmonte (1997) ha predisposto un *tagger* basato su regole. Tutti i metodi elencati, pur presentando approcci molto diversi e complessi al problema dell'annotazione grammaticale, forniscono risultati

abbastanza simili, attestati, in media, attorno al 96% di parole correttamente classificate.

Dopo aver sperimentato alcune di queste tecniche e alla luce dei risultati presentati da Zavrel e Daelmans (1999b) e da Brants (2000) appare evidente come le metodologie stocastiche basate su modelli *Hidden-Markov*, unite ad opportune tecniche per la gestione dei problemi connaturati a questi modelli, possano affrontare il problema dell'annotazione grammaticale fornendo ottimi risultati e buone velocità, sia nelle fasi di apprendimento che nelle fasi di annotazione di *corpora*.

2. Struttura di un *tagger* Stocastico

Un *tagger* stocastico, basato su un modello di *Hidden-Markov*, è sostanzialmente formato da tre componenti:

1) un lessico che elenca tutti i termini (sia come lemmi che come forme flesse) e tutte le possibili parti del discorso che un dato termine può assumere (es. "amo" può essere sia sostantivo che prima persona singolare del presente indicativo del verbo amare);
2) un *corpus* già annotato, detto "di apprendimento", utilizzato per ricavare tutte le informazioni, sotto forma di frequenze delle transizioni tra i *tag* (uni/bi/trigrammi) e frequenze delle coppie parola-*tag*, necessarie al *tagger* per risolvere tutti i casi di ambiguità grammaticale che si presentano nella fase di annotazione dei testi;
3) il programma di annotazione vero e proprio, che implementa il modello di Markov e l'algoritmo di Viterbi (1967), con l'ausilio delle opportune tecniche di ottimizzazione.

Il metodo di annotazione, sulla base di queste tre componenti, compie le seguenti operazioni:

- suddivide il *corpus* da annotare in frasi; generalmente nessuna informazione relativa all'annotazione grammaticale viene trasferita da una frase all'altra, quindi il problema dell'annotazione di testi si può ridurre al problema dell'annotazione delle frasi che li compongono. I segni di interpunzione ".", "!" e "?" fungono da separatori;
- la frase viene suddivisa in *token* o parole, che rappresenteranno l'unità alla quale verrà assegnata la parte del discorso.
- utilizzando il lessico associa a ogni *token* tutti i *tag* possibili (es. al *token* "amo" si associano i *tag* "verbo" e "nome").
- applica l'algoritmo di Viterbi, utilizzando le probabilità ricavate dal *corpus* di apprendimento, per risolvere le ambiguità e assegnare a ogni *token* il *tag* che risulta più probabile.

Analizziamo ogni singola componente in dettaglio.

2.1. IL LESSICO DEL *TAGGER*

Il lessico viene utilizzato per associare a ogni singola parola i suoi possibili *tag*. Vi sono almeno due tipologie di approccio alla costruzione di un lessico adatto a essere utilizzato per l'annotazione grammaticale: la prima riguarda la creazione di un vocabolario di forme delle parole con i rispettivi *tag*, mentre l'altra riguarda la creazione di analizzatori morfologici.

LESSICO FINITO

Per costruire questo tipo di lessico è necessario elencare tutte le possibili forme delle parole della lingua presa in esame, in forma tabulare. A ogni forma vanno poi associate tutte le possibili parti del discorso che possono essere assegnate a quella forma (Figura 2).

abaco	Nome	ancora	Nome, Verbo, Avv
ama	Verbo	andare	Verbo, Nome
ami	Nome, Verbo	...	
amo	Nome, Verbo	zuzzurellone	Nome, Aggettivo

FIGURA 2 – *Esempio di lessico in forma tabulare*

Se ci riferiamo all'italiano, ma il problema è simile nelle altre lingue, è facilmente immaginabile come una lista esaustiva possa comprendere milioni di forme. Costruire manualmente elenchi del genere, che abbiano un certo livello di affidabilità, è un compito decisamente arduo, anche utilizzando coniugatori automatici di verbi o programmi simili. Il *TreeTagger* presentato in Schmid (1994a) utilizza un lessico di questo tipo.

Una possibilità alternativa alla costruzione manuale consiste nell'utilizzare il *corpus* di apprendimento per ricavare, con programmi automatici, il lessico in forma tabulare. La dimensione del lessico, e quindi la sua efficacia nella fase di utilizzazione, è però strettamente legata alla dimensione del *corpus* di apprendimento: più grande è il *corpus*, maggiore è la completezza del lessico. I *tagger* presentati in Brants (2000), Brill (1992, 1994) e Ratnaparkhi (1996) sono esempi di annotatori che utilizzano questa tecnica.

ANALIZZATORI MORFOLOGICI

Un approccio radicalmente differente al problema è quello che consiste nel costruire analizzatori morfologici in grado di derivare automaticamente tutte le possibili lemmatizzazioni di una data parola, e quindi tutte le parti del discorso ad esse associate. Opportuni lemmari contenenti tutte le informazioni sulle modalità di generazione delle forme flesse di ogni lemma fungono da base per il procedimento; questi programmi sono in grado di analizzare la parola in

esame confrontandola con tutte le possibili forme flesse generate dal lemmario e fornendo i lemmi che potrebbero generare la parola in esame.

Oltre a questo procedimento, è possibile implementare nell'analizzatore opportune regole euristiche, che permettano di fornire risposte adeguate anche nel caso in cui la parola non sia generabile dal lemmario (ad esempio per prefissazione *ipercubo*, *stracolmo*, ecc.). Questo aumenta considerevolmente la possibilità di riconoscere automaticamente le parti del discorso associate a ogni termine e ricavare quindi le informazioni (lemma e parte del discorso) necessarie.

Questo progetto utilizza un analizzatore morfologico sviluppato da Marco Battista e Vito Pirrelli (1996a, 1996b) che, basandosi su un lemmario appositamente esteso, di circa 100.000 lemmi, è in grado di analizzare i termini fornendo, oltre al lemma e alla parte del discorso, anche ulteriori parametri come il tempo, il modo e la persona di un verbo, il numero, il genere, il caso e il grado. Un lemmario così ampio permette di generare una quantità di forme flesse tale da coprire più del 98% dei termini esaminati. Si veda a tal proposito il paragrafo successivo che si occupa delle parole sconosciute al lessico.

2.2 IL *CORPUS* DI APPRENDIMENTO

Come accennato in precedenza, l'algoritmo di *tagging HMM* utilizza le probabilità di transizione e le probabilità lessicali ricavate da un *corpus* che chiamiamo "di apprendimento". Questo *corpus* deve essere già annotato e corretto al 100%, in modo da poter ricavare tutti i parametri necessari all'algoritmo di *tagging*. È evidente come l'annotazione manuale di un *corpus* sia un lavoro piuttosto lungo e complesso. La dimensione di questo *corpus* deve essere rilevante, al fine di avere un insieme di dati più completo possibile.

Durante lo sviluppo del *tagger* per il progetto CORIS sono stati condotti alcuni studi sulla possibilità di automatizzare in qualche modo questo procedimento.

UTILIZZO DI UN ALTRO TAGGER

Il primo studio riguarda la possibilità di utilizzare un altro *tagger*, per assegnare le parti del discorso al *corpus* di apprendimento.

Esistono numerosi programmi di *tagging* disponibili in rete, ma è apparsa subito chiara la loro inadeguatezza, dato che la maggior parte di essi non è predisposta per la lingua italiana o, pur essendolo, utilizza un insieme di parti del discorso incompatibile con quelle stabilite, almeno nella fase preliminare. Un solo *tagger* fa eccezione, il *TreeTagger* di Schmid (1994a). Dalla bibliografia si ricava che questo programma annota correttamente il 96.36% dei termini; questo dato è tuttavia riferito alla lingua inglese e, anche se è previsto un modulo per l'italiano, non esistono dati certi sulle prestazioni ottenute per la nostra lingua. Anche se si volesse estendere il risultato alla lingua italiana, resterebbe un 3.64% di errori nella classificazione, percentuale che verrebbe assunta

come base per ogni successivo sviluppo dell'annotazione nel progetto CORIS. Una tale diminuzione a priori delle prestazioni non ci è sembrata accettabile. Si potrebbe pensare che il 3.64% di parole da correggere manualmente sia una quantità accettabile rispetto all'intero *corpus*; ma non esiste nessun meccanismo per individuare con certezza quali termini siano errati, pertanto il controllo manuale deve comunque essere esteso all'intero *corpus* di apprendimento.

La fase iniziale del progetto è stata effettuata utilizzando questa modalità, anche se, come vedremo in seguito, la percentuale di errori fornita dal *TreeTagger* è di gran lunga superiore a quella dichiarata per la lingua inglese. In seguito infatti il *TreeTagger* è stato sostituito dalle versioni preliminari del *CORISTagger*, già più preciso di tutti i "colleghi", anche se in fase di sviluppo.

Con questa tecnica si è ottenuto il *corpus* di apprendimento utilizzato in questo lavoro, composto da 84.000 termini. La composizione del *corpus* rispecchia le varietà che compongono il CORIS.

UTILIZZO DI PIÙ TAGGER E DI MECCANISMI DI "VOTAZIONE"

L'idea successiva, sperimentata per velocizzare il processo di annotazione manuale del *corpus* di apprendimento e per aumentarne rapidamente la dimensione, ha riguardato l'utilizzo di più *tagger* e di algoritmi a "votazione". Una volta ottenuto il *corpus* correttamente annotato con le tecniche precedentemente descritte è stato possibile riconfigurare tutti i programmi di *tagging* disponibili in rete e combinarli con la speranza che, fornendo ciascuno un diverso contributo, si potesse arrivare a un risultato globale corretto al 100%. Il primo esperimento è stato fatto utilizzando tre *tagger* di tipo stocastico, in particolare il *TreeTagger*, il *TnT* (Brants 2000) e una versione preliminare del *CORISTagger*.

Ciò che è emerso chiaramente da questo studio è che anche la combinazione delle tre valutazioni non poteva risolvere il problema, in quanto, pur con alcune diversità, i tre *tagger* si trovavano in perfetto accordo nell'annotare erroneamente un largo insieme di termini, mantenendo la prestazione totale ben lontana dal 100% richiesto.

Un successivo esperimento ha coinvolto due *tagger* strutturalmente diversi, il *Brill Tagger* (Brill 1992, 1994) basato su regole, e una versione preliminare del CORIS*Tagger*, di tipo stocastico, fornendo sostanzialmente gli stessi risultati dello studio precedente.

In conclusione, dagli studi effettuati non appare attualmente possibile pensare di creare un *corpus* di apprendimento annotato in maniera automatica o pseudo-automatica, ma è comunque richiesto l'intervento umano per correggere tutte quelle situazioni nelle quali gli annotatori automatici non sono in grado di procedere correttamente.

2.3 IL MODELLO DI MARKOV

Il modello di Markov, in questo contesto, viene utilizzato come modello stocastico del linguaggio. Gli stati del modello rappresentano i *tag*, mentre gli *output* rappresentano le parole. L'obiettivo è determinare la sequenza di stati (*tag*) del modello che meglio interpreta, dal punto di vista probabilistico, la sequenza di parole che si deve annotare. Formalmente, considerando la sequenza di parole $w_{1\ldots n}$ e la sequenza di *tag* ad essa associata $t_{1\ldots n}$ la definizione del problema è

$$\mathrm{T}(w_{1\ldots n}) = \arg\max_{t_{1\ldots n}} P(t_{1\ldots n} \mid w_{1\ldots n})$$

dove l'operatore *argmax* calcola la massima probabilità di ogni sequenza parola-*tag* possibile.

Per l'annotazione grammaticale si utilizza di solito un modello di Markov del second'ordine, introducendo le seguenti restrizioni:

$$P(w_i \mid t_{1..i}, w_{1..i-1}) = P(w_i \mid t_i)$$
$$P(t_i \mid t_{1..i-1}, w_{1..i-1}) = P(t_i \mid t_{i-2}, t_{i-1})$$

La prima indica che ogni parola dipende unicamente dai suoi possibili *tag*, la seconda che ogni *tag* dipende unicamente dai due *tag* precedenti nella sequenza. Il problema del *tagging* può quindi essere completamente formulato come

$$\mathrm{T}(w_{1..n}) = \arg\max_{t_{1..n}} \left[P(w_0 \mid t_0) P(t_0) \;\; P(w_1 \mid t_1) P(t_1 \mid t_0) \cdot \prod_{i=2}^{n} P(w_i \mid t_i) \, P(t_i \mid t_{i-2}, t_{i-1}) \right]. \tag{1}$$

La soluzione del problema di *tagging* consiste quindi nell'esplorare tutte le possibili sequenze di assegnamenti alla sequenza di parole in esame, scegliendo quella che presenta la massima probabilità. Questa ricerca, affrontata direttamente, richiede un numero molto alto di operazioni. Esiste tuttavia un algoritmo introdotto da Viterbi (1967) che consente di risolvere il problema in tempo lineare sul numero di parole da annotare.

Le probabilità coinvolte nella (1) vengono di solito stimate utilizzando il *corpus* di apprendimento descritto nei paragrafi precedenti: in particolare,

$$Unigrammi:\ P(t)=\frac{f(t)}{n}$$

$$Bigrammi:\ P(t_1 \mid t_0)=\frac{f(t_1,t_0)}{f(t_0)}$$

$$Trigrammi:\ P(t_2 \mid t_0,t_1)=\frac{f(t_2,t_1,t_0)}{f(t_1,t_0)}$$

$$P(w \mid t)=\frac{f(w,t)}{f(t)}.$$

Le probabilità dei trigrammi e dei bigrammi tendono ad essere valori piuttosto piccoli; applicando la (1) a questi valori è possibile incorrere in problemi numerici di *underflow*. Per risolvere questo problema, è pratica comune [Kempe, 93] scalare logaritmicamente i valori; trasformando la (1) in

$$\mathrm{T}(w_{1..n})=\arg\max_{t_{1..n}} \left[ln[P(w_0 \mid t_0)P(t_0)]+ln[P(w_1 \mid t_1)P(t_1 \mid t_0)]+\sum_{i=2}^{n} ln[P(w_i \mid t_i)P(t_i \mid t_{i-2},t_{i-1})] \right] \tag{2}$$

si ottiene un problema equivalente che però non soffre degli errori numerici descritti.

PROBLEMA DEI DATI MAL DISTRIBUITI (SPARSE-DATA PROBLEM)

La stima della probabilità dei trigrammi $P(t_i \mid t_{i-2},t_{i-1})$ presenta un problema: se il *corpus* di apprendimento non è sufficientemente grande, vi saranno molti trigrammi e bigrammi con frequenza molto piccola o addirittura nulla. Questo presenta un grosso ostacolo all'applicazione della (1) e della sua versione modificata (2), in quanto valori nulli azzerano il prodotto nella (1) e bloccano ogni processo di valutazione dei possibili *tag*, per tutto il resto della frase.

Vi sono numerosi correttivi applicabili; Kempe (1993) ne ha valutati alcuni, concludendo che tra essi vi è una sostanziale equivalenza, e quindi consiglia di sostituire a una probabilità nulla un valore costante ma molto piccolo da determinarsi sperimentalmente. Un'ulteriore possibilità evidenziata da Carlberger e Kann (1999) e Brants (2000) mostra come, applicando l'interpolazione tra i valori dei trigrammi, bigrammi e unigrammi

$$P(t_i \mid t_{i-2},t_{i-1})=\lambda_1 P(t_i)+\lambda_2 P(t_i \mid t_{i-1})+\lambda_3 P(t_i \mid t_{i-2},t_{i-1}),$$

si possano ottenere miglioramenti e risolvere il problema.

Nella costruzione del *CORISTagger* sono state sperimentate entrambe le soluzioni riscontrando che la seconda fornisce risultati leggermente migliori. Per semplicità è tuttavia possibile applicare anche la prima soluzione, pena un leggerissimo degrado delle prestazioni.

Un problema simile si riscontra anche per le probabilità lessicali $P(w|t)$; anche in questo caso la sostituzione di un piccolo valore costante determinato sperimentalmente permette di migliorare molto le prestazioni del *tagger* (Kempe 1993).

PAROLE NON RICONOSCIUTE DAL LESSICO

Uno dei problemi maggiori da affrontare nella costruzione di un *POS-tagger* è la gestione delle parole non presenti nel lessico di riferimento. L'approccio più comune per la formazione del lessico è quello di ricavare le parole e i loro possibili *tag* dal *corpus* di apprendimento. Questo tipo di approccio crea lessici le cui dimensioni sono funzione del numero di parole nel *corpus* di apprendimento; si va da 10000 forme circa fino a un massimo di 40000/50000. Considerando anche la legge empirica di Zipf (1968) è molto difficile costruire lessici adeguati a soddisfare un'ampia gamma di problemi. Si rende quindi necessario introdurre tecniche di analisi della morfologia della parola al fine di indurre la sua parte del discorso. Le tecniche che sembrano fornire i migliori risultati si basano sull'analisi dei suffissi (Samuelsson 1993).

Una soluzione alternativa a questo grosso problema è l'utilizzazione di analizzatori morfologici potenti e lessici molto ampi. In questo modo è possibile ridurre drasticamente il numero di termini che risultano sconosciuti al *tagger*, permettendo una più efficace previsione dei *tag* caratteristici di un dato termine. Utilizzando il lessico e il *corpus* di apprendimento descritti in precedenza, le parole sconosciute si riducono essenzialmente ai nomi propri.

TAG	**Numero**	**Perc.**
Nomi propri	285	78%
Nomi	37	10%
Aggettivi	25	7%
Altro	19	5%
TOTALE	366	100%

FIGURA 3 – *Le parole sconosciute al lessico, divise per categoria*

La tabella di Figura 3 evidenzia come il 95% delle parole sconosciute al lessico appartengano alle categorie dei nomi (78% nomi propri) e degli aggettivi. Alla luce di questi risultati è possibile introdurre un criterio euristico per la scelta dei possibili *tag* associati a un determinato termine, non appartenente o generabile dal lessico utilizzato. Se si considera che l'iniziale dei nomi propri in italiano è obbligatoriamente scritta in maiuscolo, il criterio euristico po-

trebbe associare a ogni termine sconosciuto ma maiuscolo la classe relativa ai nomi propri, mentre alle altre parole associare entrambi i *tag* relativi ai nomi comuni e agli aggettivi e lasciare decidere al *tagger* quale delle due parti del discorso sia più appropriata nel caso specifico considerato.

SCELTA DELL'INSIEME DEI TAG (TAGSET)

Un punto cruciale nella progettazione di un sistema di annotazione di *corpora* è la definizione dell'insieme delle etichette utilizzato per annotare i termini che lo compongono. La crucialità della definizione coinvolge sia le prestazioni del *tagger*, sia l'effettiva utilità e funzionalità del *corpus* annotato. Questo studio ha seguito due approcci distinti nella definizione del *tagset*.

Come primo approccio al problema si è deciso, anche per uniformarsi alle tipologie assegnate dall'analizzatore morfologico, di utilizzare un insieme di parti del discorso che rispecchiasse la classificazione tradizionale, su base morfosemantica, delle parole; la Figura 4 mostra il *tagset* che chiameremo "tradizionale".

Verbi	*V_AVERE*	**Aggettivi**	*ADJ*	**Pronomi**	*PRON_PER*
	V_ESSERE		*ADJ_DIM*		*PRON_REL*
	V_SRV		*ADJ_IND*		*PRON_DIM*
	V_PP		*ADJ_IES*		*PRON_IND*
	V_GVRB		*ADJ_POS*		*PRON_IES*
Clitici	*CLIT*		*ADJ_NUM*		*PRON_POS*
Nomi	*NN*	**Congiunzioni**	*CONJ_C*	**Simboli**	*NULL*
Nomi propri	*NN_P*		*CONJ_S*	**Punteggiatura**	*P_EOS*
Articoli	*ART*	**Avverbi**	*ADV*		*P_APO*
Preposizioni	*PREP*	**Interiezioni**	*INT*		*P_OTH*
	PREP_A	**Numeri**	*C_NUM*		

FIGURA 4 – Tagset *derivato da una classificazione tradizionale delle parti del discorso*

Questa classificazione fa riferimento alle categorie utilizzate dai dizionari e alla classificazione suggerita da Monachini (1996) nello studio svolto nell'ambito del progetto europeo *EAGLES* per la pianificazione di una struttura comune a tutte le lingue per questo tipo di strumenti per l'analisi dei testi. La classificazione delle parti del discorso suggerita nell'ambito di questo progetto ha vari livelli di dettaglio; il parallelo tra il *tagset* "tradizionale" scelto per questa fase dello studio è col livello EAGLE-L1 di dettaglio descrittivo.

Un'analisi accurata dei risultati ottenuti utilizzando il *tagset* "tradizionale" ha evidenziato però numerosi problemi di classificazione che hanno portato a ripensare la definizione dell'insieme delle parti del discorso, anche alla luce dei suggerimenti contenuti in un altro documento frutto degli studi svolti in seno al progetto *EAGLES* (Teufel *et al*. 1996). Seguendo l'analisi delle parti del di-

scorso fornite da Graffi (1994) e Winograd (1983), è stato definito un *tagset* "sperimentale" basato più su criteri "distribuzionali" dei termini che sulle analisi tradizionali, prevalentemente morfologiche. Il risultato è stata l'introduzione di nuove parti del discorso nel *tagset*, spesso trasversali rispetto alle precedenti, che potessero catturare l'effettivo ruolo nella frase dei termini, come mostrato dalla Figura 5.

Verbi	*V_AVERE*	**Determinanti**	*DET*	**Avverbi**	*ADV*
	V_ESSERE	**Aggettivi**	*ADJ*	**Numerali**	*NUM*
	V_SRV	**Pronomi (pers)**	*PRON*	**Connettivi**	*CONNV*
	V_PP	**Possessivi**	*POS*	**Simboli**	*NULL*
	V_GVRB	**Quantificatori**	*QUANT*	**Punteggiatura**	*P_EOS*
Clitici	*CLIT*	**Intensificatori**	*INTENS*		*P_APO*
Nomi	*NOUN*	**Preposizioni**	*PREP*		*P_OTH*

FIGURA 5 – Tagset *sperimentale*

Questo insieme di parti del discorso è molto simile a quello sviluppato alla *Xerox* per la costruzione di un annotatore grammaticale della lingua italiana.

3. Risultati ottenuti

Per valutare il rendimento complessivo del *tagger* è stato utilizzato un *corpus* di verifica composto da 22.000 termini, anch'esso annotato utilizzando lo stesso metodo applicato al *corpus* di apprendimento. Dopo aver effettuato la fase di apprendimento, utilizzando il *corpus* appositamente predisposto (84.000 termini), si è applicato il *tagger* al *corpus* di verifica. Il risultato ottenuto è stato comparato con quello, supposto corretto, prodotto manualmente, ricavando, oltre alla percentuale di corrette classificazioni, anche altri interessanti risultati. Lo stesso procedimento è stato applicato a tutti i *tagger* reperiti in rete, e citati in bibliografia, forniti degli opportuni moduli di apprendimento. Le percentuali d'errore relative alla classificazione del *corpus* di verifica, per entrambi i *tagset* utilizzati, sono mostrate nella Figura 6.

TAGGER	*TAGSET* TRADIZIONALE	*TAGSET* SPERIMENTALE	METODO UTILIZZATO
CORISTagger	4.81%	3.39%	Stocastico
TnT (Brants)	5.45%	4.39%	Stocastico
MXPOST (Ratnaparkhi)	6.88%	5.27%	*Maximum Entropy*
Brill (Brill)	6.96%	5.66%	Basato su regole
TreeTagger (Schmid)	12.50%	9.24%	Stocastico

FIGURA 6 – *Percentuali d'errore relative alla classificazione del* corpus *di verifica*

La percentuale di errore minore risulta essere quella del *tagger* sviluppato per il progetto *CORIS*, inferiore di almeno un punto percentuale rispetto a quelle ottenute dagli altri sistemi.

Le percentuali di corretta classificazione esibite dagli altri *tagger* considerati è notevolmente inferiore rispetto a quella dichiarata dai corrispondenti autori citati in bibliografia. Ciò è probabilmente dovuto al fatto che il *corpus* di apprendimento considerato in queste prove è molto più piccolo di quelli utilizzati nelle rispettive prove effettuate dagli autori. Utilizzare *corpus* di apprendimento di circa 1.000.000 di *token*, tipica dimensione utilizzata negli altri studi, consente di derivare un lessico molto più ampio e quindi fornire prestazioni molto superiori a quelle mostrate nella figura 6. Il *CORISTagger* invece, utilizzando un analizzatore morfologico basato su un lemmario di circa 100.000 lemmi, consente di fornire prestazioni superiori pur utilizzando un piccolo *corpus* di apprendimento. Questa caratteristica appare molto utile, alla luce delle considerazioni espresse nel paragrafo 2.2 sull'annotazione del *corpus* di apprendimento.

A questo proposito un interessante parametro da valutare è la percentuale di parole che il programma non riconosce, nella fase di analisi morfologica per la determinazione delle parti del discorso assegnabili al termine in esame. Da un esame dei risultati ottenuti si evince che il *CORISTagger* non riconosce solamente l'1.6% delle parole del corpus di verifica e, introducendo su di esse il criterio euristico descritto in un paragrafo precedente solo l'8.5% di queste riceve, in fase di classificazione, un *tag* errato. Ai fini di un confronto si mostrano nella Figura 7 i dati sulla errata classificazione delle parole sconosciute al *tagger* relativi ai programmi per i quali sono disponibili. Purtroppo il confronto deve essere fatto, per mancanza dei dati, su lingue diverse.

TAGGER	PERC. ERRATA CLASSIFICAZIONE	Lingua
CORISTagger	8.5%	Italiano
TnT	11.0%	Tedesco
TnT	14.5%	Inglese
MXPOST	13.0%	Inglese

FIGURA 7 – *Percentuali d'errore relative alla classificazione delle parole sconosciute al* tagger*. Il* tagset *considerato per queste prove è quello sperimentale*

Un altro interessante dato da ricavare riguarda l'analisi degli errori più frequenti commessi dal *tagger*. La figura 8 mostra gli errori più frequenti, in termini di scorretta classificazione dei termini, evidenziando le parti del discorso, riferite al *tagset* sperimentale, che il *CORISTagger* tende a confondere più frequentemente tra loro.

NOUN	ADJ	35.4%
V_PP	ADJ, NOUN	14.3%
CONNV	ADV	7.4%

FIGURA 8 – *Parti del discorso che vengono maggiormente confuse tra loro dal programma di* tagging

4. Conclusioni

È stato presentato lo studio, condotto nell'ambito del progetto CORIS del CILTA, per la costruzione di un programma per la lemmatizzazione e l'annotazione grammaticale di *corpora*, in lingua italiana, rispetto alle parti del discorso associate a ogni termine. Questo studio ha confrontato numerosi annotatori disponibili in rete, basati su modelli teorici differenti; i risultati, in accordo con quanto affermato da Brants (2000), "We have shown that a tagger based on Markov models yields state-of-the-art results, despite contrary claims found in the literature", mostrano come un annotatore stocastico basato su modelli di *Hidden-Markov*, se opportunamente messo a punto, correggendo i difetti strutturali di tale metodo, è in grado di fornire ottime percentuali di corretta classificazione. In particolare il *CORISTagger*, utilizzando come "vocabolario di riferimento" un potente analizzatore morfologico basato su un lemmario composto da 100.000 lemmi, ha mostrato un comportamento apprezzabilmente superiore, soprattutto considerando le dimensioni del *corpus* di apprendimento utilizzato. Da questo studio sembra quindi emergere come non sia strettamente necessario iniziare la costruzione di un annotatore grammaticale basandosi su grandi *corpus* di apprendimento (1.000.000 di parole) annotati manualmente, ma come sia possibile ridurre la dimensione di questo *corpus* di un fattore 10, sostituendo questi dati, almeno per quanto riguarda la parte lessicale, con un adeguato analizzatore morfologico dotato di un ampio lemmario.

Un altro aspetto che sembra rilevante nella costruzione di un *tagger* è la definizione del *tagset*. La composizione dell'insieme dei *tag* risulta essere particolarmente importante sia dal punto di vista delle prestazioni del *tagger* che, in seguito, per l'utilità stessa del processo di annotazione nella fase di interrogazione del *corpus*.

Sottolineiamo ancora una volta come l'obiettivo di questo studio fosse fornire uno strumento aggiuntivo di supporto allo studioso nell'interrogazione del "*data base*" *corpus*, permettendogli una più precisa scelta delle informazioni sulle quali intende lavorare. Tutte le informazioni aggiuntive inserite tramite il processo di annotazione (parti del discorso e lemmi) sono state memorizzate separatamente (su *stream* differenti) rispetto ai testi originali che compongono il *corpus*, conservandone quella caratteristica di autenticità che è alla base di questa metodologia di studio della lingua.

Ringraziamenti

Si ringraziano il prof. A. Moro e la dott.sa C. De Santis, per l'aiuto nella fase di definizione e sperimentazione dell'insieme delle parti del discorso utilizzate.

Bibliografia

Battista M. / Pirrelli V. (1996a), "Monotonic Paradigmatic Schemata in Italian Verb Inflexion", *Proc*. of *COLING-96*, Copenhagen.

Battista M. / Pirrelli V. (1996b), "Descriptive Economy and the Morphology Lexicon", *Proc*. of *EURALEX-96*, Göteborg.

Benello J. / Mackie A.W. / Anderson J.A. (1989), "Syntactic category disambiguation with neural networks", *Computer Speech and Language*, 3: 203-217.

Brants T. (2000), "TnT – A Statistical Part-of-Speech Tagger", *Proc. Conference on Applied Natural Language Processing*, Seattle, WA.

Brill E. (1992), "A Simple rule-based part-of-speech tagger", *Proc*. of *the Third Conference on Applied Natural Language Processing*, ACL, Trento.

Brill E. (1994), "Some advances in rule-based part of speech tagging", *Proc. of the Third International Workshop on Parsing Technologies*, Tilburg, The Nederlands.

Carlberger J. / Kann V. (1999), "Implementing an Efficient Part-of-Speech Tagger", *Software - Practice and Experience*, 29(9): 815-832.

Chanod J.P. / Tapanainen P. (1995), "Tagging French – comparing a statistical and constraint-based method", *Proc of the Seventh Conference of the European Chapter of the Association for Computational Linguistics*, ACL, Dublin, pp. 149-156.

Charniak E. / Hendrickson C. / Jacobson N. / Perkowitz M. (1993), "Equations for Part-of-Speech Tagging", *Proc*. *Eleventh National Conference on Artificial Intelligence*, pp. 784-789.

Charniak E. (1996), *Statistical Language Learning*, MIT Press, Cambridge, MA.

Cutting D. / Kupiec J. / Pedersen J. / Sibun P. (1992), "A Practical Part-of-Speech Tagger", *Proc*. *Third Conf. Applied Natural Language Processing*, ACL, pp. 133-140.

De Mauro T. / Mancini F. / Vedovelli M. / Voghera M. (1993), *Lessico di frequenza dell'italiano parlato*, Etaslibri, Roma.

DeRose S.J. (1988), "Grammatical Category Disambiguation by Statistical Optimization", *Computational Linguistics*, 14(1): 31-39.

Delmonte R. (1997), "Rappresentazioni lessicali e linguistica computazionale", Atti SLI, *Lessico e Grammatica – Teorie Linguistiche e applicazioni lessicografiche*, Roma, Bulzoni, pp. 431-462.

Dermatas E. / Kokkinakis G. (1995), "Automatic Stochastic Tagging of Natural Language Texts", *Computational Linguistics*, 21(2): 137-163.

Ferrari G. (1991), *Introduzione al Natural Language Processing*, Calderini, Bologna.

Graffi G. (1994), *Sintassi*, Il Mulino, Bologna.

Kempe A. (1993), "A probabilistic tagger and an analysis of tagging errors", *Technical Report, Institut für maschinelle Sprachverarbeitung,* Universität Stuttgart.

Marques N.C. / Lopes G.P. (1996), "Using Neural Nets for Portuguese Part-of-Speech Tagging", *Proc. of the Fifth International Conference on Cognitive Science and Natural Language Processing*, Dublin.

Merialdo B. (1994), “Tagging English Text with a Probabilistic Model”, *Computational Linguistics*, 20(2): 155-171.
Monachini M. (1996), “ELM-IT: EAGLES Specification for Italian morphosintax Lexicon Specification and Classification Guidelines”, *EAGLES Document EAG CLWG ELM IT/F*.
Picchi E. (1994). “Statistical Tools for Corpus Analysis: A Tagger and Lemmatizer for Italian”, *Proc. EURALEX ’94*, Amsterdam, pp. 501-510.
Rabiner L.R. (1989), “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”, in Waibel A. / Lee K. (eds.), *Readings in Speech Recognition*, San Mateo, CA, Morgan Kaufmann Publishers.
Ratnaparkhi A. (1996), “A Maximum Entropy Model for Part-of-Speech Tagging”, *Proc. of Empirical Methods in Natural Language Processing Conference*, University of Pennsylvania.
Renzi L. / Salvi G. / Cardinaletti A. (1988-95), *Grande grammatica italiana di consultazione*, Il Mulino, Bologna.
Samuelsson C. (1993), “Morphological tagging based enterely on Bayesian inference”, *Proc. 9th Nordic Conference on Computational Linguistics NODALIDA-93*, Stockholm University, Stockholm, Sweden.
Schmid H. (1994a), “Probabilistic Part-of-Speech Tagging Using Decision Trees”, *Proc. International Conference on New Methods in Language Processing*, Manchester, UK.
Schmid H. (1994b), “Part-of-Speech Tagging with Neural Networks”, *Proc. International Conference on Computational Linguistics*, Kyoto, Japan.
Tapanainen P. / Voutilainen A. (1994), “Tagging accurately – Don’t guess if you know”, *Proc*. *Fourth Conf. Applied Natural Language Processing*, ACL, Stuttgard, pp. 44-52.
Teufel S. / Schmid H. / Heid U. / Schiller A. (1996), “Study of the relation between Tagset and Taggers”, *EAGLES Docuemnt EAG-CLWG-TAGS/V*.
Viterbi A.J. (1967), “Error bounds for convolutional codes and asymptotically optimal decoding algorithm”, *IEEE Transactions on Information Theory*, 13: 260-269.
Volk M. / Schneider G. (1998), “Comparing a statistical and rule-based tagger for German”, *Proc*. *KONVENS-98*, Bonn.
Winograd T. (1983), *Language as a cognitive process, volume 1 – syntax*. Addison-Wesley.
Zavrel J. / Daelemans W. (1999a), “Recent Advances in Memory-Based Part-of-Speech Tagging”, *Actas del VI Simposio Internacional de Comunicacion Social*, Santiago de Cuba, pp. 590-597.
Zavrel J. / Daelemans W. (1999b), “Evaluatie van part-of-speech taggers voor het corpus gesproken nederlands”, *CGN Technical Report*, Katholieke Universiteit Brabant, Tilburg.
Zipf C. (1968), *The Psycho-biology of Language: An Introduction to Dynamic Philology*, MIT Press.