Sulla tipologia dei composti N+N in italiano: principi categoriali ed evidenza distribuzionale a confronto

Marco Baroni, Emiliano Guevara**
*Vito Pirrelli***

(*Università di Bologna, ** ILC, Pisa, e Università di Pavia)

1. INTRODUZIONE

Secondo Downing (1977), i composti morfologici sono la "porta di servizio" del lessico mentale, una chiave di accesso privilegiato ai principi di strutturazione dinamica delle parole e alla rete di relazioni lessicali che informano la memoria a lungo termine dei parlanti. Questa metafora ci appare particolarmente appropriata. I composti sono, a nostro avviso, non soltanto una delle aree di ricerca fondamentali e al tempo stesso più controverse della morfologia teorica degli ultimi anni, ma anche un banco di prova assai impegnativo per qualsiasi modello dell'uso linguistico. In una prospettiva tipologica, la composizione è, infatti, il principale processo di formazione delle parole (per alcune lingue, come il cinese, addirittura l'unico), il cui meccanismo generativo di base, la giustapposizione di unità lessicali piene, riflette una ben nota contiguità filogenetica con la sintassi, ricostruibile diacronicamente attraverso processi di grammaticalizzazione o lessicalizzazione che coinvolgono più di un tipo di struttura sintattica (Bybee 2002, Dressler 2006). Da questo punto di vista, i composti illustrano in modo emblematico il punto di contatto tra l'asse sintagmatico del linguaggio, dominato da relazioni di contiguità temporale a corto raggio (o relazioni *in praesentia*), e l'asse paradigmatico, all'interno del quale si stratificano relazioni a lungo termine tra unità in distribuzione complementare (*in absentia*).

Sarebbe tuttavia fuorviante relegare la composizione allo status di residuo fossile della sintassi nel lessico morfologico. In primo luogo, i composti sono rappresentativi di un'area della morfologia sincronicamente permeabile alla sintassi (Lieber e Scalise 2006), a testimonianza che la relazione tra i due domini linguistici è operativa anche nella competenza grammaticale interiorizzata dai parlanti. In secondo luogo, la composizione è un processo generativo produttivo e flessibile, che risponde a profonde esigenze pragmatiche di comunicazione telegrafica e di compressione dell'informazione e come tale va osservato "sul campo", preferibilmente in quei contesti comunicativi concreti dove queste esigenze risultino prioritarie. Infine, le strategie che i parlanti mettono in atto per produrre e interpretare nuovi composti sono molto diversificate e spesso riconducibili a processi di "composizionalità debole" (Pirrelli, 2001), governati da relazioni lessicali elusive, dipendenti dal contesto e variabili con continuità, quali analogia e salienza (Costello & Keane 2001). Da questo punto di vista, alcune famiglie di composti sembrano definire un ambito di analisi privilegiato per l'osservazione dei processi interpretativi lessicali operativi nell'uso linguistico. Una loro analisi dettagliata è destinata a migliorare la nostra comprensione del modo in cui forme lessicali complesse sono interpretate e rappresentate nel nostro lessico mentale.

Uno degli aspetti più interessanti della composizione come fenomeno linguistico è il fatto che le relazioni sintattico-semantiche fra i costituenti dei composti sono variabili. Da sempre, gli studi sulla composizione (cfr. fra gli altri Tollemache 1945, Marchand 1969, Levi 1978, ecc.) basano la classificazione delle parole composte sull'individuazione di relazioni sistematiche e ricorrenti fra i costituenti. Bisetto e Scalise (2005) hanno recentemente proposto uno schema di classificazione innovativo basato sulla coesistenza di criteri diversi applicati gerarchicamente. Il parametro classificatorio principale è la relazione grammaticale non

esplicita che sussiste tra i costituenti: le possibili relazioni sono quelle di Subordinazione, Coordinazione e Attribuzione, che qualificano i composti corrispondenti come di tipo SUB, CRD e ATT rispettivamente. Lo scopo del presente contributo è quello di definire una rigorosa metodologia d'analisi distribuzionale finalizzata all'induzione e alla classificazione automatiche delle parole composte a partire da corpora testuali di grandi dimensioni, prendendo spunto dalla classificazione proposta da Bisetto e Scalise e, allo stesso tempo, verificandone la validità su dati linguistici realmente attestati. In particolare ci concentreremo sulle problematiche relative alla classificazione dei composti nominali del tipo N+N in italiano.

La struttura dell'articolo è la seguente. Il paragrafo 2 illustra i punti cardine della metodologia di ricerca che abbiamo adottato. Il paragrafo 3 passa in rassegna i più autorevoli tentativi di definire una serie di criteri classificatori dei processi compositivi attestati interlinguisticamente, seguito (paragrafo 4) da un'illustrazione delle principali caratteristiche dei diversi tipi di composto N+N in italiano. Nel paragrafo 5, delineiamo una metodologia semi-automatica volta all'identificazione e alla classificazione dei composti N+N dell'italiano in corpora testuali di vaste dimensioni. La metodologia è indirizzata ad un triplice scopo: (i) la compilazione di un'ampia lista rappresentativa dei principali tipi di composti N+N in italiano, basata su dati linguistici concreti, (ii) la verifica del grado di copertura dei criteri di classificazione proposti in letteratura, a fronte dei composti censiti, (iii) l'induzione di criteri distribuzionali per la classificazione automatica di composti N+N. Infine, il paragrafo 6 riassume le principali indicazioni fin qui fornite dal nostro lavoro e delinea le direzioni future di ricerca.

2. METODOLOGIA

Affrontare un problema così variegato e complesso come la composizione ha richiesto un duplice sforzo metodologico. Da una parte, ci siamo avvalsi degli strumenti categoriali tradizionali della linguistica teorica, al fine di poter disporre di una prima mappatura su vasta scala di un territorio ancora in buona parte inesplorato (§ 3).

All'interno di questa area più ampia, abbiamo deciso di focalizzarci sui composti N+N in italiano, la cui analisi preliminare è l'oggetto del presente contributo. La sotto-famiglia N+N rappresenta infatti una "porta di servizio" ampia e diversificata della composizione in italiano, i cui membri definiscono il tipo di composizione più importante e diffuso interlinguisticamente (Dressler 2006). I composti N+N forniscono dunque una solida base di confronto dei meccanismi di interpretazione e rappresentazione di unità lessicali complesse universalmente attestati. È attraverso questo tipo di confronto, a nostro avviso, che sarà possibile arrivare a comprendere i meccanismi generativi alla base della composizione e, per contrasto, i vincoli interpretativi specifici che ciascuna lingua impone su questi meccanismi. Riteniamo inoltre che una migliore comprensione di questa tipologia di processi compositivi rappresenti un valore aggiunto anche dal punto di vista applicativo, ad esempio in vista della creazione di risorse lessicali di ampia copertura in settori come la didattica linguistica assistita dal calcolatore, lo sviluppo di lessici di dominio e la traduzione semi-automatica. È noto infine che i composti costituiscono uno dei fattori di maggiore criticità nello sviluppo di tecnologie ad ampia copertura per il trattamento automatico del linguaggio naturale, soprattutto per quelle lingue come l'inglese e l'italiano dove i composti in generale (e in particolare i più produttivi) si presentano spesso come sequenze di parole adiacenti separate da spazi bianchi, senza marche morfologiche aggiuntive che ne segnalino lo status sintattico peculiare.

In seconda battuta, ci siamo posti l'obiettivo di dare risposta a una serie di domande relative alle circostanze concrete d'uso dei composti: Come riconosciamo un composto in un testo? Quale tipo di rappresentazione gli assegniamo e attraverso quali processi interpretativi siamo in grado di ricostruire le relazioni concettuali che ne legano i costituenti interni? Per rispondere a queste domande abbiamo adottato una metodologia basata sull'esplorazione

preliminare di attestazioni reali d'uso di composti italiani in repertori testuali di vastissime dimensioni (fig. 1). Questa metodologia ci ha portato a: (i) la definizione di procedure euristiche per l'identificazione automatica dei composti N+N in vasti repertori testuali in formato digitale (§ 3); (ii) la validazione dei risultati preliminari di queste ricerche automatiche attraverso una classificazione manuale dei dati ottenuti in fase esplorativa (§ 4); e infine (iii) l'analisi dei contesti d'uso dei composti validati manualmente (§ 5). In questo modo, abbiamo ottenuto tre risultati preliminari di un certo interesse: (a) lo sviluppo di procedure algoritmiche (in forma di espressioni regolari) per l'individuazione di potenziali composti del tipo N+N in italiano; (b) l'analisi quantitativa dei contesti d'uso ricorrenti dei composti individuati nella prima fase; (c) la messa a punto di misure di correlazione tra classi di composti (definite su base categoriale) e una batteria di variabili quantitative che ne riflettono l'uso tipico in contesti comunicativi attestati.

In prospettiva, i passi (a), (b) e (c) non solo ci consentono di analizzare il comportamento dei composti "sul campo", ma anche di valutare il ruolo potenziale che ciascuna classe di composti (e all'interno di una classe, ciascun composto) riveste nel processo di "schematizzazione" attraverso cui il bambino estrae da un gruppo coerente di unità lessicali complesse schemi relazionali lessicalizzati via via più astratti, per poi generalizzarli a ulteriori contesti d'uso (Tomasello 2006). Il processo di schematizzazione interagisce col modo in cui i composti sono memorizzati nel lessico mentale del parlante e si organizzano reciprocamente, in rapporto alla frequenza d'uso e alla salienza dei loro costituenti interni (Costello e Keane 2001). Ad esempio, è noto che le strategie di riconoscimento di forme morfologicamente complesse sono facilitate dalla dimensione della "famiglia paradigmatica" (*family size effects*, Schreuder e Baayen, 1997): più grande è la famiglia paradigmatica cui è riconducibile la radice lessicale di una forma complessa, più rapido è il riconoscimento di quella forma. Nei composti, l'effetto è complicato dal fatto che una testa lessicale può ricorrere come primo o come secondo costituente di un composto. De Jong et al. (2002) riferiscono che il riconoscimento di un composto olandese A+B del tipo *watermolen* ('mulino ad acqua') è facilitato dalla frequenza della famiglia di composti X+B (che comprende *windmolen*, *koffiemolen* ecc., dove X varia e B *molen* resta costante), ma è indipendente dalla frequenza di B in isolamento. Se da una parte questo effetto illustra l'influenza della struttura interna di un composto sulle strategie di riconoscimento del composto stesso, d'altra parte esso sembra evidenziare che l'organizzazione dei composti nel lessico mentale avverrebbe per strutture lessicali complesse (del tipo X+B) e non sulla base di singole parole decontestualizzate.

In generale, siamo inclini ad ipotizzare che l'incidenza di fattori quali la frequenza e la salienza dei singoli costituenti interni sull'organizzazione dei composti nel lessico mentale possa variare in relazione alla classe specifica di composti. In questa prospettiva, l'analisi della distribuzione e dell'uso dei composti in corpora di grandi dimensioni ci servirà a valutare l'impatto di questi fattori sulla dinamica dei processi di acquisizione, interpretazione e organizzazione dei composti nel lessico mentale. A questo proposito, ci proponiamo di valutare empiricamente la correttezza delle nostre ipotesi di lavoro attraverso la verifica del grado di accettabilità/familiarità di un insieme di composti su un campione di parlanti nativi. In secondo luogo, intendiamo simulare la dinamica dei processi acquisizionali e interpretativi mediante l'uso di modelli artificiali di reti neurali con ben note proprietà auto-organizzative (Pirrelli e Herreros, 2006). Riteniamo infine che la messa a punto di rappresentazioni linguistiche arricchite da indici distribuzionali e di frequenza possa portare a una parziale ridefinizione delle categorie teoriche dalle quali è partita la nostra indagine (fig. 1).

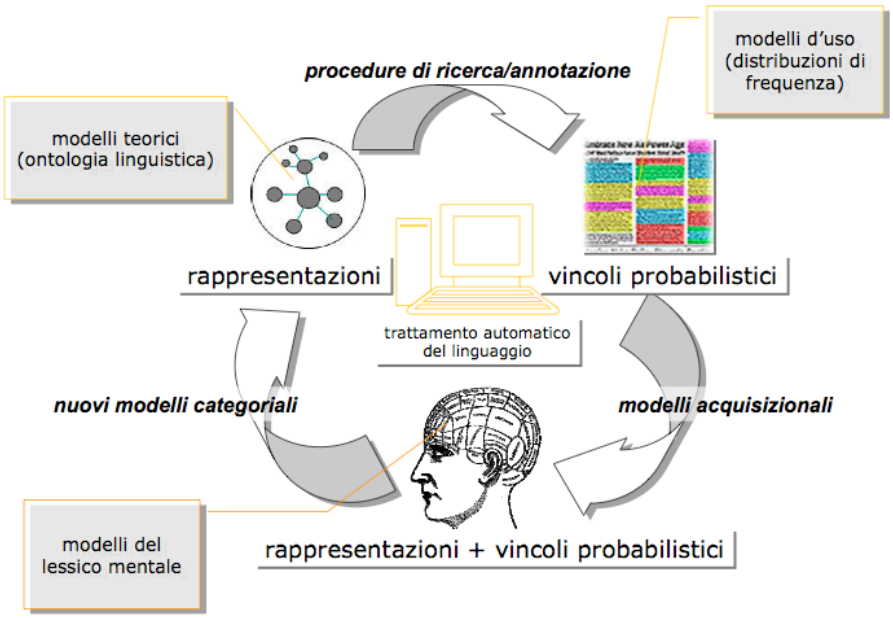


Figura 1: Il ciclo metodologico

3. CLASSIFICAZIONE ED ESTRAZIONE DEI COMPOSTI N+N

Prima di descrivere nei dettagli i passi del nostro ciclo metodologico, riteniamo utile un breve excursus nella tradizione dello studio teorico della composizione. La composizione rappresenta un caso più unico che raro nella storia della linguistica poiché, nonostante sin dall'antichità sia stata oggetto di considerevole attenzione nella letteratura (i composti sono già presenti nella grammatica sanscrita di Pāṇini), la teoria della composizione non ha mai conosciuto uno sviluppo sistematico. Tipicamente sono stati individuati e classificati tipi diversi di composti, dal punto di vista sia strutturale, sia semantico. I principali criteri usati per produrre tali schemi classificatori sono i seguenti:[1]

- Simmetria vs. asimmetria
- Endocentricità vs. esocentricità (testa categoriale e/o formale vs. testa semantica)
- Relazione grammaticale fra i costituenti (e.g. modificazione, modificazione aggettivale, apposizione, coordinazione, ecc.)
- Semantica di tutto il composto vs. semantica dei costituenti
- Caratteristiche strutturali (e.g. categoria dei costituenti, presenza di *linking elements*, ecc.)

3.1 I prototipi compositivi tradizionali e la classificazione dei composti

Tradizionalmente, i composti sono classificati in base ai tre principali tipi individuati nel sanscrito da Pāṇini nei suoi *Aṣṭādhyāyī*: i composti *tatpuruṣa*, i composti *bahuvrīhi* ed infine i composti *dvandva*. Questi tre termini sono essi stessi esempi del tipo compositivo che designano.

I composti *tatpuruṣa* ('(di) questo + uomo, servo = il servo di questa persona') del

1 Per un'esposizione dettagliata della storia delle classificazioni dei composti, cfr. Bisetto e Scalise (2005).

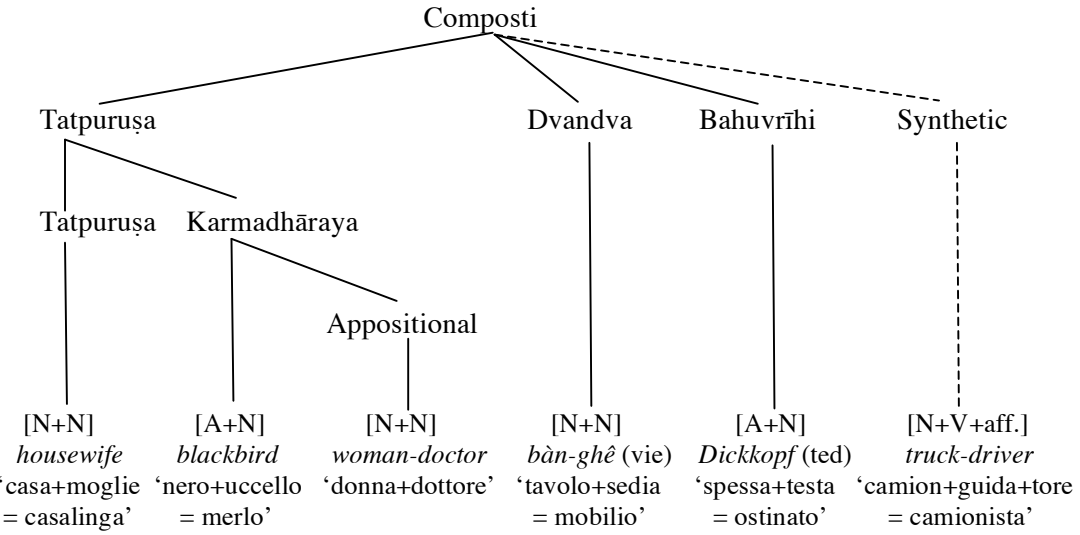
sanscrito sono prototipicamente determinativi: schematicamente, si possono descrivere con la formula "un XY che è un tipo di Y" sia semanticamente, sia formalmente.

I composti *dvandva* ('due + due = coppia, duetto') sono caratterizzati dal fatto che nessuno dei due costituenti è determinato dall'altro, ma fra di loro si stabilisce un rapporto di simmetria: "un XY che è sia X, sia Y".

Infine, i composti *bahuvrīhi* ('molto + riso = persona che possiede molto riso') sono i prototipi del composto esocentrico: "un XY che non è né X, né Y".

Quasi tutti gli schemi classificatori dei composti sin ora prodotti nella letteratura sono fortemente influenzati da questa tripartizione originale. Si veda ad esempio la seguente elaborazione di un recente contributo sulla tipologia della composizione (Bauer 2001, nel grafico in (1) è indicato anche il sottotipo dei composti determinativi *karmadhāraya*, spesso tralasciati nella letteratura):

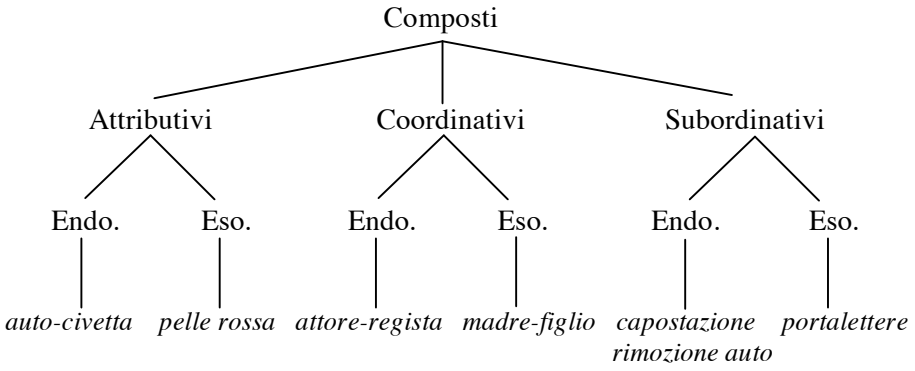
(1)



Si può vedere chiaramente come Bauer riprenda in modo fedele i tipi individuati da Pāṇini, pur aggiungendo qualche novità (e.g. i composti sintetici delle lingue germaniche), e come li applichi meccanicamente all'inglese (ma anche ad altre lingue tedesco e vietnamita). Il problema principale con questo tipo di operazione è che i tre prototipi compositivi tradizionali sono definiti da un insieme di criteri classificatori non omogenei, e la classificazione che ne consegue è uno schema spurio e solo apparentemente "piatto".

Un punto di vista diverso è adottato da Bisetto e Scalise (2005) i quali propongono uno schema classificatorio omogeneo dei composti, basato sull'applicazione gerarchica di due soli criteri (a) *relazione grammaticale fra i costituenti* e (b) *endo-/esocentricità*, che qui illustriamo con esempi dell'italiano:

(2)



Questo schema, rispetto alle proposte precedenti, offre il vantaggio di ridurre l'immensa varietà compositiva nelle lingue del mondo a tre sole classi principali: (a) composti attributivi, (b) composti coordinativi e (c) composti subordinativi. Ognuna di queste classi può successivamente essere suddivisa in base la presenza vs. assenza di una testa lessicale interna.

Nella presente ricerca, adottiamo lo schema classificatorio della composizione proposto da Bisetto e Scalise (2005) come punto di partenza per una serie di indagini esplorative volte all'estrazione ed alla classificazione semiautomatica dei composti in italiano.

Date le vastissime dimensioni e la varietà dei fenomeni compositivi in italiano, in questo studio pilota ci limitiamo al trattamento dei soli composti *N+N endocentrici*, i quali costituiscono uno dei maggiori centri di produttività e regolarità del lessico (peraltro, nel campione di composti N+N non abbiamo riscontrato alcun caso di composto *esocentrico*, un probabile segnale dell'estrema rarità di questo tipo all'interno della classe morfosintattica presa in analisi). A questo fine, inoltre, adotteremo una nozione molto elementare e conservatrice di *testa di un composto*: la testa di un composto è quell'elemento costituente dello stesso che condivide con il tutto sia la categoria lessicale sia i tratti formali di sottocategoria. Inoltre, ipotizzeremo che ogni composto endocentrico possieda soltanto *una testa*, e che nei casi in cui entrambi i costituenti sono candidati equivalenti, lo status di testa sia assegnato al costituente in posizione canonica, vale a dire, in italiano, il costituente situato a sinistra.

3.2 Estrazione

In questa ricerca è stato utilizzato il corpus *itWaC* (Baroni e Ueyama 2006), di circa 2 miliardi di token provenienti da pagine web scritte in italiano, lemmatizzato ed annotato per parti del discorso. La scelta del corpus è ricaduta su *itWaC* per motivi di robustezza delle elaborazioni statistiche descritte nel § 5. In sintesi, per "catturare" i composti e le loro proprietà servono corpora di grandi dimensioni.[2]

Sfruttando le informazioni morfosintattiche e la lemmatizzazione del corpus *itWaC* abbiamo estratto una prima lista di coppie di candidati N+N, puntando a sequenze attestate all'interno di cornici sintattiche plausibili per elementi nominali (identificate con espressioni regolari su sequenze di parti del discorso) in modo da ridurre il numero di "falsi positivi". Al contempo, le cornici sono abbastanza generiche da non (s)favorire particolari tipi di composti (e.g. composti che tendono a ricorrere in contesti target *soggetto di frase*, *oggetto diretto di frase*, o simili). Date le dimensioni del corpus, da questa prima lista di candidati è stato successivamente estratto un campione *random* ridotto, ottenendo i dati da quattro fasce di diversa frequenza di attestazione, per assicurare l'eterogeneità distribuzionale dei candidati scremati:

Rango	Tipi	Campione
1	699.659	300
2 – 5	329.270	300
6 – 3.000	113.147	300
> 3.000	109	109
Totale	1.142.185	1.009

TABELLA 1: Campionamento da quattro fasce di frequenza

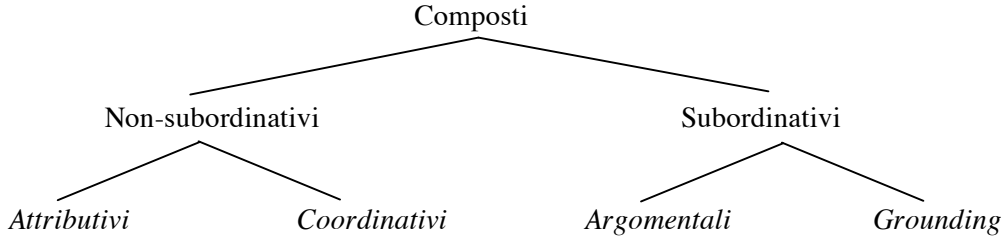
[2] I primi tentativi sperimentali che abbiamo realizzato si basavano su dati estratti dal corpus *La Repubblica* (circa 380 milioni di tokens, cfr. Baroni *et al.* 2004), ma si sono rivelati problematici per la scarsità delle frequenze di co-occorrenza fra i termini dei composti.

Il campione così ottenuto consiste in 1.009 coppie di candidati N+N rappresentative dell'intero spettro di frequenza del corpus (un campionamento *random* unico avrebbe invece favorito fortemente le fasce corrispondenti ai valori di frequenza più bassi).

3.3 Analisi

A partire dal nostro campione ridotto di 1.009 candidati, sono state manualmente identificate come composti 252 sequenze N+N (circa 1/4 del totale). In seguito, questi composti sono stati classificati secondo lo schema classificatorio tripartito di Bisetto e Scalise (2005), al quale, nel corso dell'analisi, sono state apportate delle modifiche. Abbiamo in primo luogo creato una distinzione principale fra i composti subordinativi (o determinativi) ed i composti non-subordinativi, nei quali entrambi i costituenti contribuiscono predicativamente a tutto il composto (cfr. § 4.1). Poi abbiamo suddiviso ulteriormente i composti subordinativi in due categorie: i *composti argomentali* ed i *composti grounding* (cfr. § 4.2).

(3)



Queste modifiche al modello classificatorio hanno poi ricevuto supporto empirico dai dati qualitativi e quantitativi estratti dal corpus. Nelle sezioni seguenti, concentreremo le nostre analisi sui quattro tipi terminali individuati in (3): *attributivi*, *coordinativi*, *argomentali* e *grounding*.

I 252 composti identificati manualmente sono stati quindi classificati secondo il modello quadripartito modificato presentato in (3), ottenendo la seguente distribuzione:

(4) *Distribuzione dei composti campionati in classi*:

a. Coordinativi (COOR):	34	(13,49%)
b. Attributivi (ATTR):	41	(16,27%)
c. Argomentali (ARGU):	51	(20,24%)
d. Grounding (GROU):	118	(46,82%)
e. Altro:	8	(3,17%)

È degno di nota il fatto che circa 2/3 del totale degli esempi identificati corrisponde alla classe dei composti subordinativi (argomentali e grounding insieme), e che il resto del campione (circa 1/3) corrisponde a composti non-subordinativi (attributivi o coordinativi). Un totale di 8 esempi si è rivelato particolarmente difficile da classificare e richiederà analisi ulteriori.

4. LA TIPOLOGIA DEI COMPOSTI N+N IN ITALIANO

Passiamo ora a descrivere a grandi linee le caratteristiche definitorie delle quattro classi di composti N+N italiani identificati manualmente dal corpus *itWaC*, e che serviranno da base empirica per le analisi statistico-distribuzionali dei paragrafi successivi.

4.1 Composti non-subordinativi

La classe dei non-subordinativi raggruppa tutti quei composti i cui costituenti interni non sono legati da una relazione di dipendenza sintattica implicita. Tradizionalmente la classe include il tipo coordinativo e quello attributivo, dettagliati nei sottoparagrafi che seguono. Una menzione a parte merita probabilmente il tipo "appositivo" (o denominativo), che include composti come *cane bassotto*, *effetto notte* e *specie uomo* il cui contenuto può essere in molti casi descritto ricorrendo alla parafrasi "X chiamato/detto Y". Per quanto ci è dato di vedere sulla scorta di alcuni esempi emblematici, la famiglia dei composti appositivi non è riconducibile alle altre famiglie della stessa classe. Tuttavia la penuria di attestazioni nel corpus ci ha consigliato di escluderla per il momento dall'ambito della nostra ricerca esplorativa.

4.1.1 Composti coordinativi (COOR)

I composti coordinativi sono probabilmente il tipo N+N più semplice e regolare dell'italiano: testa e modificatore denotano entità ontologicamente simili o compatibili (appartenenti cioè entrambi allo stesso livello ontologico e a campi semantici affini). Tutto il composto ha interpretazione coordinativa piena: la sequenza *testa* + *modificatore* XY è sia X che Y, e la classe dei suoi *denotata* è definita con buona approssimazione dall'intersezione dei *denotata* di X e di Y.

(5) *viaggio spedizione*, *cantante attore*

Come si può notare dagli esempi apportati in (5), i composti coordinativi corrispondono a varie classi nominali dell'italiano (e.g. nomi evento, nomi d'agente, nomi di luogo, ecc.).

4.1.2 Composti attributivi (ATTR)

I composti attributivi sono più difficili da descrivere, poiché la loro interpretazione non è riconducibile neanche in prima approssimazione a una funzione intersettiva. Nei composti attributivi, l'interpretazione del modificatore si riduce ad una proprietà centrale del suo contenuto semantico pieno, che viene trasferita e *attribuita* alla testa (rapporto che potremmo schematizzare come "un XY che è un X di tipo Y", "un XY che è un X, ma in qualche modo anche un Y", ecc.):

(6) *presidente fantoccio*, *caso limite*, *progetto pilota*

Come si evince dalle interpretazioni dei composti in (6), i modificatori degli attributivi non convogliano la totalità delle proprie informazioni: per esempio un *presidente fantoccio* non ha nessuna delle proprietà "materiali" del *fantoccio*; ad esempio, non è fatto di legno e stracci, non è usato come giocattolo o in rappresentazioni teatrali per bambini, ecc. L'unico attributo di *fantoccio* che è ancora presente in *presidente fantoccio* è quello di essere manovrato da un'altra persona: diremo quindi che *presidente fantoccio* è interpretato grosso modo come 'presidente non-autonomo'. In altre parole, i modificatori dei composti attributivi assumono un senso figurato o metaforico che mette in risalto un loro tratto saliente.

Inoltre, i modificatori dei composti attributivi spesso sviluppano un comportamento distribuzionale di tipo "aggettivale", occorrendo frequentemente in posizione post-nominale:

(7) a. Modificatori attributivi tipici: *base, tipo, simbolo, pilota, modello, guida, chiave, limite, record, fantoccio, fantasma*

b. *pilota* è in posizione post-N in 1/4 delle sue occorrenze nel corpus (13727/52641)

c. *fantoccio* è in posizione post-N in 1/3 delle occorrenze (722/2314)

In sintesi, possiamo descrivere i composti attributivi N+N come caratterizzati da una testa nominale semanticamente piena, e da un modificatore lessicalmente nominale che si comporta semanticamente come una proprietà e che spesso ha la distribuzione di un aggettivo.

4.2 Composti subordinativi

Nei composti subordinativi i costituenti sono chiaramente legati da una relazione di dipendenza sintattica implicita. Uno dei costituenti (la testa) mostra una palese prevalenza formale, semantica e distribuzionale rispetto all'altro costituente. Abbiamo proposto di operare un'ulteriore distinzione all'interno della classe dei subordinativi, quella fra i *composti argomentali* ed i *composti grounding*.

4.2.1 Composti argomentali (ARGU)

I composti argomentali N+N sono costruzioni subordinative in cui la testa è un nome deverbale in grado di proiettare struttura argomentale (tipicamente la testa denota un evento, di solito un nome d'azione). Il modificatore in questi composti esprime l'argomento interno del verbo soggiacente alla testa, e quindi si verificano due casi particolari:

(8) a. Il modificatore è l'oggetto di un verbo transitivo:
protezione persone, *raccolta fondi*, *gestione priorità*

b. Il modificatore è il soggetto di un verbo inaccusativo:
arrivo documenti, *caduta massi*

Questo tipo particolare di composto è stato studiato in dettaglio da Bisetto e Scalise (1999). Si noti che strutturalmente i composti argomentali dell'italiano sono speculari rispetto ai composti argomentali delle lingue germaniche (la letteratura li ha nominati in vari modi diversi, composti secondari, a nesso verbale, sintetici, ecc.):

(9) a. protezione persone $[[V + Suf.N]_N + N]_N$

b. taxi driver 'tassista' $[N + [V + Suf.N]_N]_N$

4.2.2 Composti grounding (GROU)

I composti grounding N+N sono costruzioni chiaramente subordinative in cui la testa è un nome che **non proietta** struttura argomentale di tipo verbale o che in ogni modo non legittima un'interpretazione argomentale del modificatore. La testa ha invece un significato nominale di tipo generico-relazionale che tende ad "attirare" modificatori secondo il proprio contenuto

semantico-enciclopedico.

Il modificatore dei composti grounding contestualizza o specializza la testa (ingl. *the modifier "grounds" the head*), specificandone ulteriormente il contenuto. Le teste dei composti grounding in italiano esprimono concetti intrinsecamente relazionali, che richiedono una specificazione ulteriore per convogliare un contenuto efficacemente informativo; esempi tipici di teste grounding sono i portatori/contenitori di informazione, organizzazioni, luoghi, aggregatori, puntatori nel tempo e nello spazio, proprietà misurabili, ecc. Cfr. (10):

(10) *stanza server*, *fondo pensioni*, *centro città*, *altezza righe*, *posto auto*

4.3 Classificazione manuale del campione: conclusioni preliminari

A questo punto della nostra esposizione, è opportuno trarre alcune conclusioni preliminari. Innanzitutto, è degno di nota il successo nell'applicazione dello schema classificatorio quadripartito basato sulla proposta di Bisetto e Scalise (2005) al campione estratto (successo vicino al 97%). Inoltre, la presenza di generalizzazioni semantiche sulle caratteristiche delle teste e dei modificatori per ogni tipo di composto N+N considerato porta sostegno al modello adottato.

Emerge un quadro in cui la relazione semantica fra testa e modificatore nei composti N+N dell'italiano è legittimata dalle proprietà semantiche della testa e/o del modificatore (per es., dalla natura relazionale delle teste nel tipo subordinativo grounding, dalle caratteristiche salienti dei modificatori negli attributivi, ecc.), fatto che costituisce da solo un punto di differenziazione rispetto alle lingue germaniche, nelle quali una maggiore varietà d'interpretazioni possibili per una stessa sequenza N+N è forse legittimata dalla costruzione N+N medesima. Per esempio, il composto dell'inglese *police protection* può essere interpretato sia come 'protezione della polizia'– lettura argomentale –, sia come 'protezione da parte della polizia, offerta dalla polizia' – con una lettura di tipo grounding. Simili casi non sono attestati in italiano in modo sistematico e regolare: un composto con testa deverbale come *rimozione auto* può solo avere l'interpretazione argomentale.

5. ANALISI DISTRIBUZIONALE

In questa sezione presentiamo vari tipi di indizi distribuzionali che dimostrano come le classi di composti che abbiamo proposto non sono soltanto giustificate dal punto di vista qualitativo, ma trovano anche un riscontro quantitativo nel corpus utilizzato.

5.1 Proprietà di teste e modificatori

Abbiamo innanzitutto contato il numero di teste e modificatori che capitano in più di un composto nel nostro campione, ottenendo i valori riportati nella seguente tabella:

Classe	Teste	Modificatori
ATTR	0 (0%)	4 (12%)
COOR	1 (3%)	1 (3%)
ARGU	6 (14%)	2 (4%)
GROU	15 (18.5%)	10 (9.71%)

TABELLA 2: Teste e modificatori ripetuti nel campione

Per quanto riguarda le teste ripetute, notiamo che né gli attributivi né i coordinati mostrano una

tendenza in questo senso, mentre la tendenza è molto forte per i grounding. Questo è in accordo con la nostra caratterizzazione di questa classe di composti, che sono legittimati dalle proprietà semantiche di teste "forti" che vengono da un numero ristretto di tipi semantici. È dunque plausibile che ci siano delle teste grounding prototipiche che capitino spesso anche in un campione piuttosto limitato come il nostro. Tra le teste più comuni troviamo nomi quali *fine*, *commissione* e *centro* (che capitano in ben 7 composti), che sono in effetti "casi tipo" di teste grounding.

Simmetricamente, la nostra caratterizzazione dei composti attributivi come legittimati dalle proprietà semantiche di modificatori prototipici ci porta a predire che i modificatori di questa classe, e soltanto essi, avranno una tendenza a ripetersi nel campione. Mentre i dati confermano la tendenza degli attributivi ad avere modificatori ripetuti (*base* capita in 5 composti, *quadro* in 3, *limite* e *chiave* in 2), troviamo, inaspettatamente, una simile tendenza anche per i grounding (tra i più frequenti: *sicurezza*, con 4 occorrenze, mentre *stampa*, *lavoro* e *dato*/*dati* capitano 3 volte). Al momento, non abbiamo una giustificazione per questo fatto all'interno del nostro modello, e sono dunque richieste analisi ulteriori.

In ogni caso, la presenza di una notevole proporzione di teste e modificatori ripetute/i anche in un campione relativamente piccolo come il nostro suggerisce che ci sono notevoli forze analogiche all'opera nella creazione dei composti N+N in italiano, un'ipotesi che ci proponiamo di esplorare più approfonditamente in futuro.

Abbiamo in seguito studiato il numero della testa dei composti (e dunque il numero dell'intero composto), ottenendo i dati riportati nella tabella 3:

Classe	Singolare	Plurale
ATTR	64,34%	35,66%
COOR	49,50%	50,50%
ARGU	90,12%	9,88%
GROU	86,20%	13,80%
SIMP	76,47%	23,53%

TABELLA 3: Occorrenze al singolare e al plurale delle teste nel corpus (*macro-averages*)

Nella tabella 3 abbiamo riportato, per ciascuna classe, la proporzione media di occorrenze al singolare e al plurale delle teste dei composti nella classe (per esempio, la prima riga indica che la testa del composto attributivo medio capita al singolare nel 64,34% delle occorrenze, e al plurale nel rimanente 35,66% dei casi). Tali valori sono stati ottenuti con la tecnica del *macro-averaging*, ovvero calcolando prima le proporzioni di singolare e plurale per ciascun composto, e poi la media per ciascuna classe (si veda la discussione in Sebastiani 2002). Riportiamo anche, come metro di paragone, la media di occorrenze al singolare e plurale di un campione di nomi semplici selezionati con criteri simili a quelli usati per i composti. Si noti infine, a proposito di questa tabella (ma anche dei dati discussi in seguito), che abbiamo identificato singolare e plurale con metodi automatici, perciò le proporzioni contengono senza dubbio un certo margine di errore, che dovrebbe però essere simile in tutte le classi.

Il primo dato che risulta evidente nella tabella 3 è la forte tendenza delle teste argomentali ad evitare il plurale. Questo si accorda con la nostra caratterizzazione di tali teste come nominalizzazioni deverbali, o in ogni modo nomi con proprietà argomentali di tipo verbale, che, in particolare quando sono usati all'interno di un composto, hanno una qualità spiccatamente verbale, legittimando il modificatore come argomento interno. Si tratta, insomma, di quelli che Grimshaw (1990) definirebbe *complex event nominals*. Tali nomi eventivi non denotano dunque un'entità concreta o comunque "contabile", ed è perciò naturale che tendano a non essere usati al plurale (si pensi all'innaturalezza di frasi quali: *?Ci sono molte cadute massi in questa zona*). Ulteriore conferma dell'idea che le teste argomentali sono *complex event nominals* viene dal fatto che, in un esperimento non riportato qui, abbiamo trovato che tali teste non sono quasi mai precedute dall'articolo, non avendo, evidentemente,

proprietà di definitezza che sono rilevanti per gli oggetti ma non per gli eventi.

Anche le teste grounding mostrano una forte tendenza ad evitare il plurale, il che sembra conforme alla loro natura di entità astratte (puntatori, aggregatori, ecc.). Invece, come ci aspettavamo data la nostra caratterizzazione delle classi in questione, né le teste attributive né le teste coordinative, che possono tranquillamente riferirsi ad entità concrete e contabili, hanno problemi ad essere usati al plurale.[3]

La seguente tabella presenta invece la distribuzione media di numero del modificatore quando la testa è al singolare e quando la testa è al plurale, mettendo dunque in luce eventuali rapporti d'accordo tra gli elementi di un composto:

Testa al singolare

Classe	Mod. singolare	Mod. plurale
ATTR	95.73%	4.27%
COOR	100.00%	0.00%
ARGU	46.24%	53.76%
GROU	70.75%	29.25%

Testa al plurale

Classe	Mod. singolare	Mod. plurale
ATTR	94.17%	5.83%
COOR	2.71%	97.29%
ARGU	41.05%	58.95%
GROU	62.11%	37.88%

TABELLA 4: Proporzione di occorrenze al singolare e plurale del modificatore nel corpus quando la testa è al singolare e quando la testa è al plurale (*macro-averages*)

Notiamo prima di tutto che i composti coordinativi presentano un accordo pressoché perfetto tra numero della testa e numero del modificatore, come predetto dalla nostra caratterizzazione di questi composti. Nei coordinativi, testa e modificatore si riferiscono ad una stessa entità, che può essere singola/non-contabile o molteplice, e dunque il loro numero, dipendendo da tale entità, deve essere il medesimo.

Per gli attributivi, notiamo che il modificatore ricorre praticamente sempre al singolare, dato conforme anche alla nostra intuizione di parlanti (cfr. *parole chiave* vs. ?*parole chiavi*). Da una parte, il netto contrasto tra i modificatori nei composti coordinativi e attributivi è un indizio forte a favore di una classificazione dei composti in grado di riconoscere questa distinzione. Dall'altra, la motivazione per il comportamento dei modificatori attributivi non ci è del tutto chiara. Un'ipotesi in proposito è che, poiché i nomi usati come modificatori negli attributivi non sono più dei nomi "pieni" con dei referenti veri e propri, ma sono ridotti ad indicare una sola proprietà prototipica del loro referente, proprietà che viene predicata della testa del composto, il tratto nominale di numero non è più pertinente per loro, e dunque essi sono essenzialmente senza numero, e vengono realizzati al singolare semplicemente come valore di default. Alternativamente, si può pensare che costruzioni quali *N di base*, in cui un tipico modificatore attributivo è realizzato in un contesto sintattico che non richiede l'accordo grammaticale tra teste lessicali (**modelli di basi*), abbiano fatto da modello prima ai composti attributivi col medesimo modificatore (**modelli basi*), e poi, per analogia, agli attributivi in genere.

Infine, sia nei composti argomentali che nei composti grounding la testa e il

3 Queste classi mostrano addirittura una maggior propensità a capitare al plurale di quella mostrata dai nomi semplici. Tuttavia, questo può essere dovuto al fatto che tra i nomi semplici a bassa frequenza tendono a capitare più neo-derivazioni (tra cui nominalizzazioni deverbali e altri nomi astratti) che tra i composti con la stessa frequenza (ARGU a parte). Dunque, prima di trarre conclusioni al riguardo, bisogna ripensare il campionamento dei nomi semplici.

modificatore denotano entità distinte, e dunque non ci aspettiamo dipendenze tra numero della testa e numero del modificatore. Questo è confermato dai dati in tabella 4, che mostrano che tutte le combinazioni di numero sono ampiamente attestate sia per i composti argomentali che per i composti grounding.

5.2 Connector patterns

Consideriamo ora un tipo di indizio diverso, osservando la distribuzione di teste e modificatori dei composti quando capitano in costruzioni in cui tra di essi è presente materiale sintattico visibile (per il quale usiamo il termine *connector pattern*). L'idea alla base di questo approccio è che il *connector pattern* possa rendere esplicito il tipo di relazione che è implicitamente presente tra testa e modificatore anche all'interno del composto (per es., che se nel corpus troviamo *stanza per i server* o *stanza con i server*, la relazione tra *stanza* e *server* espressa dalle proposizioni *per*/*con* sia la medesima che è implicita nel composto *stanza server*). L'idea che ci sia una relazione tra composti e sintagmi con gli stessi elementi lessicali è stata esplorata da tempo sia nella letteratura teorica sui composti (si veda in particolare Levi 1978) sia nelle applicazioni computazionali (si veda in particolare Lauer 1995), anche se le dimensioni del nostro corpus ci pongono in una posizione privilegiata, visto che ricerche di questo tipo possono soffrire di gravi problemi di *data sparseness* (in breve, solo in un corpus di grandissime dimensioni possiamo aspettarci di trovare, per ciascun composto studiato, un certo numero di sintagmi con la medesima testa e il medesimo modificatore connessi da uno dei *connector patterns* di interesse – esperimenti preliminari con un corpus di 380 milioni di parole indicano che, anche se queste sono già dimensioni ragguardevoli per un corpus, tale corpus è troppo piccolo a questi fini).

Ci siamo concentrati su 3 tipi di *connector patterns*, ovvero:

- Congiunzioni coordinative (*e*, *o*, *nonché*, *né*, ...)
- La preposizione *di*/*del*
- Altre preposizioni semanticamente "piene": *a*, *in*, *per*, *con*, *su* (e le corrispondenti preposizioni articolate)

Per ciascuna coppia testa-modificatore abbiamo calcolato la significatività statistica della frequenza d'occorrenza dei due termini con ciascun *connector pattern* usando la misura della *log-likelihood ratio*, e calcolando le frequenza osservata e attesa del trigramma testa-connettore-modificatore come segue (Legenda: P = probabilità, H = testa, M = modificatore):

Frequenza osservata: $count(H\ CONN\ M)$

Frequenza attesa: $p_{ind}(HM) \times N_{CONN}$

Dove: $p_{ind}(HM) = p(H) \times p(M)$

$$p(H) = \frac{count(H)}{N_{CORPUS}}$$

$$p(M) = \frac{count(M)}{N_{CORPUS}}$$

$$N_{CONN} = count(NOUN\ \ CONN\ \ NOUN)$$

Riportiamo di seguito, per ciascuna classe di composti e ciascun *connector pattern*, la proporzione di coppie teste-modificatori in quella classe che, in combinazione con il *connector pattern* in questione, hanno una frequenza di co-occorrenza significativa (p<0.01). Altri tipi di analisi statistica hanno dato risultati molto simili a quelli riportati.

La tabella 5 indica la proporzione di teste-connettori-modificatori con co-occorrenza significativa per le quattro classi di composti e scegliendo le congiunzioni coordinative come *connector pattern*:

ATTR	COOR	ARGU	GROU
53.66%	64.71%	52.94%	50.85%

TABELLA 5: Proporzione di coppie testa-modificatore significativamente associate in combinazione con congiunzioni coordinative

Teste e modificatori di composti coordinativi hanno una tendenza molto forte a co-occorrere significativamente con congiunzioni coordinative, come ci aspettavamo avendo ipotizzato che la relazione tra testa e modificatori in questi composti è, appunto, di tipo coordinativo.

Anche se i coordinativi presentano una proporzione di strutture coordinative nettamente più alta che le altre classi, nelle altre classi tale proporzione non è prossima allo zero, come sarebbe stato forse lecito aspettarsi. Attribuiamo tale fatto al "rumore" dovuto all'estrazione automatica delle stringhe usate per contare le co-occorrenze. In particolare, nel caso del *pattern* coordinativo è probabile che le nostre interrogazioni per espressioni regolari abbiano raccolto strutture coordinative spurie che non coinvolgevano sintagmi nominali.

La tabella 6 riporta lo stesso tipo di dati quando *di*/*del* funge da *connector pattern*:

ATTR	COOR	ARGU	GROU
31.71%	20.59%	90.20%	75.42%

TABELLA 6: Proporzione di coppie testa-modificatore significativamente associate in combinazione con *di*/*del*

Poiché la preposizione *di* mette è un tipico connettore sintattico tra due nomi che denotano entità distinte, ci aspettiamo che essa realizzi sintatticamente la relazione tra teste e modificatori dei composti argomentali e grounding. Questa predizione è chiaramente supportata dai dati in tabella 6.

Per quello che riguarda la differenza tra argomentali e grounding, con gli elementi dei primi che mostrano una più spiccata predilezione per *di* dei secondi, questo fatto si spiega probabilmente alla luce dei dati riportati nella tabella 7, che presenta la proporzione di co-occorrenze significative con le altre preposizioni:

ATTR	COOR	ARGU	GROU
46.34%	20.59%	54.90%	68.64%

TABELLA 7: Proporzione di coppie testa-modificatore significativamente associate in combinazione con *a*/*al*, *in*/*nel*, *per*/*per il*, *con*/*col*, *su*/*sul*

Qui la tendenza si ribalta, con teste e modificatori grounding che mostrano una tendenza più forte di teste e modificatori argomentali all'associazione significativa in combinazione con le preposizioni a contenuto semantico pieno. Nella nostra caratterizzazione, le teste argomentali legittimano il modificatore proiettando una struttura argomentale di tipo verbale, e il modificatore assume il ruolo semantico del primo argomento interno (oggetto, soggetto inaccusativo) del verbo corrispondente, perciò non c'è bisogno di specificare ulteriormente la

relazione semantica tra testa e modificatore con una preposizione semanticamente piena (**rimozione alle macchine* non è grammaticale per la stessa ragione per cui non lo è **rimuovere alle macchine*): *di* va bene proprio perché non porta ulteriori informazioni semantiche sulla relazione tra i due nomi.

Invece, nel caso dei composti grounding sarebbe forzato dire che la testa proietta una struttura argomentale di tipo verbale (qual è la struttura argomentale di *angolo*?), e la relazione tra testa e modificatore sembra basarsi più su aspetti di conoscenza enciclopedica riguardanti la funzione delle entità denotate che su relazioni codificate dalla grammatica. Dunque, è più naturale che in questo caso il parlante senta il bisogno di esplicitare la relazione tra testa e modificatore con una preposizione semanticamente piena (*angolo per la cottura*).

Ci sembra che la conferma empirica della predizione non banale secondo la quale i costituenti dei composti argomentali tenderebbero a collocarsi con la preposizione *di*, mentre quelli dei composti grounding con preposizioni semanticamente piene costituisca un argomento forte a favore della nostra proposta di distinguere tra questi due tipi di subordinativi. Notiamo infine che anche questi dati, a parte le tendenze generali che abbiamo appena discusso, appaiono piuttosto "rumorosi": sorprende in particolare il valore relativamente alto di teste e modificatori attributivi in tabella 7, forse dovuto al fatto che espressioni frequenti con tipici modificatori attributivi, quali *in base* o *al limite*, hanno portato all'estrazione di stringhe *testa + preposizione + modificatore* spurie.

6. CONCLUSIONI

Riassumendo, abbiamo proposto e applicato ai composti N+N dell'italiano una classificazione che si basa sulla relazione tra testa e modificatore in composti coordinativi, attributivi, argomentali grounding. Abbiamo mostrato che questa classificazione ci permette di categorizzare gran parte dei composti in un campione estratto da un grande corpus di italiano contemporaneo, e di trovare interessanti generalizzazioni sulle proprietà dei composti in ciascuna classe (in particolare, abbiamo individuato alcune caratteristiche semantiche comuni alle teste grounding e altre caratteristiche tipiche dei modificatori attributivi).

In seguito, abbiamo presentato vari indizi di tipo distribuzionale che, congiuntamente, permettono di distinguere i tipi di composti proposti sulla base di dati quantitativi estratti dal corpus. In breve, i composti coordinativi sono caratterizzati dall'accordo di numero tra testa e modificatore e dalla tendenza di teste e modificatori a co-occorrere anche con congiunzioni coordinative. I composti attributivi sono caratterizzati spesso da modificatori "forti" che tendono a ripetersi in composti diversi (un fenomeno stranamente condiviso dai composti grounding), e dal fatto che il modificatore non viene quasi mai volto al plurale. I composti subordinativi sono caratterizzati dalla presenza di teste prototipiche che tendono a ripetersi (soprattutto nel caso dei grounding), dalla riluttanza a formare il plurale del composto (particolarmente spiccata per gli argomentali), dal fatto che sia la testa che il modificatore possono ricorrere al singolare o al plurale, senza fenomeni d'accordo, e dall'occorrenza di teste e modificatori in sintagmi in cui sono connessi da preposizioni. Oltre che per la tendenza più o meno forte ad aver teste prototipiche ed ad evitare il plurale, gli argomentali e i grounding si distinguono per il tipo di preposizioni che i loro elementi prediligono quando capitano in sintagmi con un connettore esplicito: i costituenti dei composti argomentali sono più facilmente associati a *di*, mentre quelli dei grounding hanno in generale un'associazione significativa anche in combinazione con preposizioni semanticamente piene.

Al momento, stiamo conducendo esperimenti di *clustering* su un campione di composti esteso, nella speranza di mostrare che gli indizi distribuzionali da noi proposti siano sufficienti, se combinati, a far emergere classi simili a quelle proposte senza bisogno di supervisione manuale. Tornando a quanto accennato nel § 2, il passo successivo nel nostro ciclo metodologico sarà poi la ricerca di riscontri piscolinguistici per la nostra proposta, che mostrino

che essa non è soltanto corroborata da evidenza statistica, ma anche plausibile come modello cognitivo.

RINGRAZIAMENTI

Questa ricerca si svolge nell'ambito del progetto nazionale PRIN *COMPONET*, coordinato da S. Scalise (U. di Bologna). Ringraziamo sia il coordinatore del progetto, sia i membri delle diverse unità locali per il loro supporto ed incoraggiamento. Ringraziamo anche gli organizzatori del convegno SLI 2006 ed i presenti al convegno per commenti e suggerimenti.

RIFERIMENTI BIBLIOGRAFICI

Baroni, Marco e Motoko Ueyama, 2006, *Building general- and special-purpose corpora by Web crawling*. In: *Proceedings 13th NIJL International Symposium, Language Corpora: Their Compilation and Application*, Tokyo, 31-40.

Baroni, Marco *et al.,* 2004, *Introducing the La Repubblica corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian*. In: M. T. Lino *et al*. (eds.), *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004, Lisbon, may 26-28)*. Paris, ELRA - European Language Resources Association, pp. 1771-1774.

Bauer, Laurie, 2001b, *Compounding*. In: M. Haspelmath (ed.), *Language Typology and Language Universals*. Mouton de Gruyter. The Hague.

Bisetto, Antonietta e Sergio Scalise, 1999, *Compounding: Morphology and/or Syntax?* In: L. Mereu (ed.), *Boundaries of Morphology and Syntax*. Amsterdam, Benjamins. 31-48.

Bisetto, Antonietta e Sergio Scalise, 2005, *The classification of compounds*. "Lingue e linguaggio", 4(2), 319–332.

Bybee, Joan, 2002, *Sequentiality as the basis of constituent structure*. In: T.Givón e B. Malle (eds.), *The evolution of language out of pre-language*, Amsterdam, Benjamins, 107-132.

Costello, Fintan J. e Mark T. Keane, 2001, *Testing two theories of conceptual combination: Alignment versus diagnosticity in the comprehension and production of combined concepts*. "Journal of Experimental Psychology: Learning, Memory & Cognition", 27(1), 255-271.

Dressler, Wolfgang U., 2006, *Compound Types*. In: G. Libben e G. Jarema (eds.), The Representation and Processing of Compound Words. Oxford, Oxford University Press. 23-44.

Evert, Stefan, 2004, *The statistics of word cooccurrences: Word pairs and collocations*. PhD thesis, University of Stuttgart.

Grimshaw, Jane, 1990, *Argument structure*. Cambridge (Mass), The MIT Press.

De Jong, Nivia H. et al., 2002, *The processing and representation of Dutch and English compounds: Peripheral morphological, and central orthographic effects*. "Brain and Language" 81, 555–567.

Lauer, Mark, 1995, *Designing statistical language learners: Experiments on noun compounds*. PhD thesis, Macquarie University, Sydney.

Levi, Judith, 1978, *The syntax and semantics of complex nominals*. New York, Academic Press.

Lieber, Rochelle e Sergio Scalise, 2006, The Lexical Integrity Hypothesis in a New Theoretical Universe. "Lingue e Linguaggio", 5(1), 7-32.

Marchand, Hans, 1969, *The Categories and Types of Present-Day English Word-Formation: A Synchronic-Diachronic Approach*. 2nd edition. München, C.H. Beck.

Pirrelli, Vito, 2001, *Per un superamento della dicotomia lessico-grammatica. Aspetti di composizionalità "debole" nel linguaggio*. In: E. Ferrario e V. Pulcini (a cura di), *La*

Lessicografia bilingue tra presente e avvenire. Vercelli, Edizioni Mercurio.
Pirrelli, Vito e Ivan Herreros, 2006, *Learning morphology by itself*. To appear in: G. Booij *et al.* (eds.), *On-line proceedings of the Fifth Mediterranean Morphology Meeting*, Bologna.
Schreuder, Robert e R. Harald Baayen, 1997, *How complex simplex words can be*. "Journal of Memory and Language" 37, 118–139.
Sebastiani, Fabrizio, 2002, *Machine learning in automated text categorization*. "ACM Computing Surveys" 34, 1–47.
Tollemache, Federigo, 1945, *Le parole composte nella lingua italiana*. Roma, Rores.
Tomasello, Michael, 2006, *Acquiring linguistic constructions*. In: D. Kuhn e R. Siegler (eds.), *Handbook of Child Psychology*. New York, Wiley.