Dalla *Word Sense Disambiguation* alla Sintassi: il Problema dell'Articolo Partitivo in Italiano

Ignazio Mauro Mirto
Università degli Studi di Palermo
Dipartimento Culture e Società
V.le delle Scienze Ed. 15 - 90128 Palermo
ignazio.mauro.mirto@unipa.it

Emanuele Cipolla
posta@emanuelecipolla.net

Abstract

Italiano. Fuori contesto, un nesso come *dei professori* non dà certezza di dove collocare *dei* in relazione alle parti del discorso. Il nesso può per esempio valere o *alcuni professori* (per es. in *Dei professori intervennero*) o esprimere appartenenza (per es. *i libri dei professori*). Nel primo caso *dei* è l'articolo partitivo di un nesso *nominale*, nel secondo è la preposizione che introduce un complemento di specificazione. Questo caso di omonimia si può far rientrare nell'area del *Word Sense Disambiguation*, ma la sua rilevanza per la sintassi e per il NLP è evidente. Nonostante ciò, in letteratura di esso non abbiamo trovato tracce. Il lavoro distingue diverse funzioni dei membri della serie e propone un algoritmo per disambiguare i due usi riferiti e altri, per esempio i complementi retti (come in *Approfittano dei tuoi fratelli*) che rendono la disambiguazione ancora più complessa.

English. *Out of context, a phrase of Italian such as dei professori 'of.the teachers' is ambiguous: it can either mean some teachers (e.g. Dei professori intervennero 'Some teachers attended') or carry the value of a Saxon genitive (e.g. i libri dei professori 'the teachers' books'). The part of speech to which dei belongs cannot be identified: dei could be a partitive article in a noun phrase or a preposition in a prepositional phrase. This key difference raises a problem in the area of Word Sense Disambiguation. Despite its relevance for NLP, to the best of our knowledge this case of homonymy has so far been disregarded in the literature. The paper distinguishes a number of functions dei carries and proposes an algorithm that can automatically discriminate between the two uses mentioned above, but also identify others that make the picture more complex.*

1 Introduzione

Questo lavoro verte sull'articolo partitivo in italiano, etimologicamente formato da *di* e da un articolo determinativo. L'intera serie, *del*, *dello*, *dell'*, *della*, *dei, degli*, *delle*, si presenta in superficie identica alle omonime preposizioni articolate.

Anche a un primo sguardo, la varietà di esiti che si ottiene collocando una sequenza come *dei professori* in contesti differenti, con *dei* qui preso come elemento rappresentativo dei sette membri della serie, desta stupore per la numerosità degli usi e le conseguenti difficoltà che ciò crea nel NLP.

Obiettivo del nostro lavoro è la disambiguazione automatica. Le difficoltà che un tale compito pone sono numerose. Lo studio è parte di una ricerca più ampia che ha come fine l'individuazione automatica del Soggetto[1] di una frase semplice (Mirto and Cipolla, 2017). In genere, l'articolo partitivo non è elemento frequente nei testi, ma la sua rilevanza al fine di ottenere maggiore precisione nella ricerca del Soggetto è evidente, come si vedrà nel prosieguo.

La sezione 2 è dedicata alle ambiguità semantiche che l'omonimia genera, derivanti da ambiguità strutturali. La sezione 3 presenta alcuni degli àmbiti grammaticali che creano ostacoli per la corretta identificazione degli articoli partitivi. Ognuno di questi àmbiti ha determinato una parte dello script presentato, che è stato messo alla prova su un corpus formato da 463 occorrenze (casualmente scelte tra le complessive 580) degli ele-

[1] Che, a giudicare dal numero di lavori reperibili in letteratura, non sembra argomento che susciti grande interesse, in particolare per l'italiano. Si veda almeno (Dell'Orletta et al., 2005) e i riferimenti ivi contenuti.

menti del paradigma rinvenute nel romanzo *Palomar* di Italo Calvino. La sezione 4 conclude il lavoro presentando i risultati ottenuti.

2 Ambiguità

La frase *Parlarono dei professori* è semanticamente ambigua: il nesso *dei professori* può infatti essere interpretato come complemento di argomento (i professori sono l'argomento di cui qualcuno parla) oppure come Soggetto post-verbale, con *dei* equivalente, in buona sostanza, ad *alcuni* (*Parlarono dei/alcuni professori*).

Anche la frase *Sono dei professori* risulta ambigua, visto che "oscilla" tra un significato di appartenenza (*Questi libri sono dei professori*, se il Soggetto *questi libri* viene omesso) e un significato equativo, cioè con identità referenziale tra nesso preverbale e nesso postverbale (*Loro/Questi sono dei professori*, con *loro/questi* e *professori* che rimandano allo stesso referente). Già da questi casi è possibile intuire alcune delle difficoltà di parsing per l'italiano generate dall'articolo partitivo, che ricorre in ognuno dei due casi di ambiguità presentati.

Caratteristica precipua dell'articolo partitivo dell'italiano è la frequente possibilità di farne a meno, di ometterlo, a parità di significato e mantenendo inalterata l'accettabilità della frase. È possibile farlo, per esempio, in *Parlarono professori*, ovviamente non più ambigua, così come è possibile farne a meno negli usi equativi: *Loro sono professori*. Di contro, l'omissione risulta impossibile nel significato di appartenenza o di possesso: **Questi libri sono professori* e, chiaramente, anche in quello del complemento di argomento (**Loro parlarono professori*), qualora si desideri mantenere identico il significato e l'ineccepibilità della frase.

Ecco succintamente illustrato uno dei frequentissimi casi di ambiguità che si presentano nelle lingue naturali. Chiamata in causa è l'area di ricerca nota come *Word Sense Disambiguation* (Stevenson and Wilks, 2003). È bene riaffermare che l'ambiguità non è di tipo lessicale, essendo *dei* composto da morfemi grammaticali, quindi privi di contenuto descrittivo.

Un paio di tentativi su demo disponibili online[2], che fanno uso di *dependency parsing*, con frasi come *Degli alunni hanno starnutito* o *Dei ragazzi starnutirono*, entrambe con articolo partitivo, hanno dato per *dei* il lemma *di* e la categoria 'preposizione' (si noti che, di fatto, ciò esclude erroneamente il nesso dalla funzione di Soggetto):

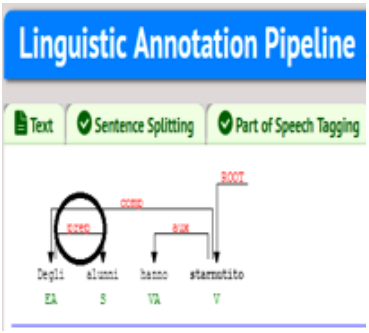

Figure 1: Parsing con LinguA (03.07.2017)

Figure 2: Parsing con TextPro (11.07.2017)

Al di là dei tentativi di soluzione per fini pratici, si può affermare, più in generale, che a questo problema di omonimia in italiano la linguistica teorica e la semantica formale hanno dedicato molte attenzioni. Di contro, nel campo del NLP esso sembra essere passato inosservato.

L'algoritmo che presentiamo è stato implementato nel linguaggio Python 2.7[3]. Per effettuare *part of speech* e *lemma tagging*, al fine di identificare ad esempio nomi, verbi ed aggettivi, è stato utilizzato *TreeTagger* (`http://www.cis.uni-muenchen.de/(~schmid/tools/TreeTagger/)` con il file di parametro per l'italiano realizzati da

[2]Reperibili ai seguenti indirizzi: `http://linguistic-annotation-tool.italianlp.it/syntactic_trees` (figura 1), `http://hlt-services2.fbk.eu/textpro-demo/textpro.php` (figura 2)

[3]A IMM si deve la parte dello script che disambigua i potenziali articoli partitivi. EC si è fatto carico di tutte le indispensabili operazioni di annotazione su *TreeTagger*.

Marco Baroni, richiamato utilizzando il modulo treetagger-python (`https://github.com/miotto/treetagger-python`).[4]

L'algoritmo non si basa sulla nozione di costituente e le strategie adottate non fanno uso di 'alberi' di stampo chomskiano né di *dependency parsing*. Il parsing non è né *bottom up* né *top down*. Riteniamo che ai fini di una maggiore efficacia, cioè per un parsing in grado di identificare e risolvere ambiguità strutturali e semantiche, sarà indispensabile fare ricorso alla struttura argomentale dei predicati, o 'valenza', particolarmente di quelli verbali (Tesnière, 1959).

3 L'Algoritmo di Disambiguazione

Questa sezione mostra la suddivisione dello script di disambiguazione, basata sui diversi contesti di occorrenza dei morfemi della serie indagata. Complessivamente, nel corpus abbiamo identificato sette diversi casi: (I) complementi di specificazione; (II) complementi retti; (III) casi in cui ricorre il verbo *essere* o con funzione di ausiliare perfettivo o come copula; (IV) articoli partitivi con verbi transitivi e intransitivi; (V) comparativi e superlativi; (VI) nessi la cui testa è un pronome indefinito, (VII) locuzioni (*in fin dei conti*, *del resto*, *del tipo* (per es. *un larvato rimprovero del tipo "potresti pensarci un po' tu"*)). Il trattamento degli ultimi tre gruppi (tre occorrenze per (V), cinque per (VI), tre per (VII)) sarà oggetto di un'integrazione successiva.

3.1 *Dei* nel Complemento di Specificazione

Un nesso nominale come *i libri dei professori* esemplifica il complemento di specificazione. La serie che manifesta questo complemento contiene tutti gli elementi già elencati per l'articolo partitivo, ma, significativamente, se ne distingue perché include la forma *di* (*i libri di Leo*). Pur con questa massiccia sovrapposizione di forme, si ottengono distinte parti del discorso: se da un lato il partitivo è una forma di articolo (un determinante), dall'altro ciò che pare lo stesso elemento è invece una preposizione, che può essere articolata o semplice. Con l'unica differenza della preposizione semplice, tuttavia, al parser le forme si presentano identiche, fatto che impone una qualche risorsa che sia in grado di differenziare i due usi. Così, se la frase soggetta al parsing fosse *Abbiamo letto i libri di fisica dei professori*, non si avrebbe difficoltà a collocare *di* tra le preposizioni, mentre per *dei* si rivela necessaria un'operazione di disambiguazione.

Su questo caso di omonimia non siamo stati in grado di trovare in letteratura proposte precedenti. Suggeriamo in questa sede di individuare un complemento di specificazione grazie alla parola che precede la preposizione, che il più delle volte è o un nome o un aggettivo. La parte di codice rilevante, abbreviata e semplificata, è qui di seguito illustrata (`frase[i]` è il pivot):

```
# classificazione: 0=complemento
# di specificazione
for i in range(len(frase)):
    precedenti = frase[0:i]
    successivi = frase[i+2:len(frase)]

    compl_specificazione =
     frase[i] in maybe_partitive \
           and (frase[i-1] in nomi \
           or frase[i-1] in agg)

    if compl_specificazione is True:
            classificazione=0
```

Se tra gli elementi che precedono immediatamente una qualsiasi delle sette forme della serie, inserite nella tupla denominata *maybe_partitive*, si includono (a) i dimostrativi (per es. *il passo delle zampe posteriori [...] quello delle anteriori*), (b) i verbi all'infinito (per es. *l'espandersi della sabbia*), (c) alcune congiunzioni (*l'alfabeto delle onde marine o delle erbe d'un prato*), (d) casi di ricorsività (per es. *del tessuto del fondo*) e, infine, (f) occorrenze multiple con virgola (per es. *la percezione precisa dei contorni, dei colori, delle ombre*), la porzione di script sopra illustrata consente di identificare correttamente 388 complementi di specificazione, pari al $97,7\%$ delle occorrenze. Oltre a questi *true positives* si sono avuti 9 *false negatives*, 3 *false positives* e 63 *true negatives*; ciò dà luogo a una *precision* di 0.99 e ad una *recall* di 0.97; la F_1*-score* è pari a 0.97. Alcuni casi problematici sono: (i) la topicalizzazione del nesso preposizionale (per es. *Della conoscenza mitica degli astri egli capta solo qualche stanco barlume*); (ii) le nominalizzazioni (per es. *tutto il non detto della sua condizione*); oppure (iii) quello di *Ho trovato sul selciato degli uccelli malconci*, in cui *degli* svolge la funzione di articolo partitivo, ma viene erroneamente intercettato come complemento di specificazione a causa del locativo *sul*

[4] Il tratto [±Numerabile] del sostantivo che segue *dei* consentirebbe di escludere che in un nesso come *della penna* ricorra un articolo partitivo (* *Voglio della penna)*. La ricerca ne verrebbe semplificata. Questa risorsa non è stata utilizzata perchè *TreeTagger* non fornisce il tratto.

selciato che ricorre tra il verbo e il nesso nominale post-verbale.

Un paio di osservazioni finali. La prima: dal punto di vista semantico, il complemento di specificazione può esprimere un significato affine a quello di frasi copulative (§ 3.3) come *I libri sono dei professori*, significato cui ci si riferisce comunemente con 'appartenenza' o 'possesso'. La seconda: è bene ribadire che né in *i libri dei professori* né in *I libri sono dei professori* è possibile sottrarre *dei* (**I libri professori*, **I libri sono professori*), proprio perché la sottrazione a parità semantica è caratteristica esclusiva dell'articolo partitivo, anche se tale opzione non è sempre praticabile.

3.2 *Dei* come Complemento Retto

Si tratta del caso esemplificato con il verbo *parlare*. La già discussa ambiguità della frase *Parlarono dei professori* deriva proprio dal fatto che *parlare* è verbo potenzialmente bivalente (o trivalente: *Leo parlò a Luigi di Ada*). Se l'esempio fosse modificato in *Dei professori parlarono*, con Soggetto anteposto, la frase rimarrebbe ancora ambigua, ma in modo diverso: o *dei professori* è un Soggetto canonicamente pre-verbale oppure, se ancora interpretato come complemento di argomento, esso è allora collocato in una posizione marcata e la frase, segmentata, necessita di un particolare profilo intonativo, cioè di una messa in rilievo tramite enfasi, di seguito richiamata con il maiuscoletto: DEI PROFESSORI *parlarono* (*non degli studenti*). L'esplicitazione del Soggetto porrebbe fine a ogni ambiguità: *Loro parlarono dei professori*.

In italiano i predicati che idiosincraticamente legittimano un complemento in *di* non sono necessariamente verbali. Ecco alcuni dei casi rinvenuti nel corpus, con verbi (il nesso non è né Soggetto né Oggetto diretto), aggettivi, avverbi, nomi e polirematiche (si notino le due topicalizzazioni):

- tener conto degli aspetti complessi
- ripaga del sapere che si propaga
- dell'adeguato innaffiamento approfittano le erbacce
- quello che ha pensato del prato
- spera d'essersi appropriato del pianeta
- faccio parte dei soggetti senzienti
- avrebbe più bisogno del nostro interessamento
- è specifico del sesso femminile
- anche del nulla non si può essere sicuri al cento per cento
- prima della sua nascita
- al di là delle abitudini sensoriali
- in balia della sovrapopolazione di questi lumpen-pennuti *[sic]*

Talvolta lo stesso verbo presenta più valenze, con differenze semantiche come *Chiedono dei professori* vs *Chiedono professori*, dunque con un ulteriore caso di ambiguità: *Chiedono a proposito dei professori* vs *Richiedono professori*. Individuare differenze così sottili richiede soluzioni complesse.

Nello script, i complementi di specificazione sono rilevati dopo i complementi retti. Il motivo è semplice: se la frase sottoposta al parsing fosse *Sandro è degno degli onori più grandi*, la funzione rileverebbe nella posizione precedente a *degli* un aggettivo, restituendo quindi un errore, cioè che *degli onori più grandi* è complemento di specificazione. Lo stesso accadrebbe con una polirematica come *tener conto delle proporzioni*, che nella posizione precedente a *delle* presenta un sostantivo.

I complementi retti introdotti da una delle forme omonime a quelle degli articoli partitivi sono complessivamente 33, pari al $7,1\%$ delle 463 occorrenze indagate.

Per l'individuazione dei complementi retti si è creata una lista, denominata *trigger_di*, contenente verbi, aggettivi, avverbi e locuzioni che legittimano un complemento introdotto dalla preposizione *di*. Con la suddivisione della stringa in 'precedenti' e 'successivi' rispetto al pivot l'algoritmo consente di calcolare se il complemento retto è anteposto al predicato che lo regge (ordine marcato) o posposto (ordine canonico):

```
# Classificazione complemento retto:
# 1=posposto, 2=anteposto
for j in range(len(frase)):
    if frase[j] in trigger_di:
        if frase[j] in precedenti:
            classificazione=1
        elif frase[j] in successivi:
            classificazione=2
```

3.3 *Dei* in frasi con *essere* come copula o con *esserci*

È uno dei casi presentati nella sezione 3 con frasi ambigue come *Sono dei professori*. Si noti che la frase *Ci sono dei professori*, in superficie diversa dalla precedente solo per la presenza del clitico *ci*, esemplifica un tipo denominato in letteratura 'esistenziale', che è tutt'altra cosa. Nella frase *Ci sono dei professori* il nesso *dei professori* fornisce un esempio di articolo partitivo. Ne è prova il fatto che *dei* può o essere rimosso senza che la frase collassi (*Ci sono professori*) o essere sostituito con *alcuni* (*Ci sono alcuni professori*). Le due frasi *Sono dei professori* e *Ci sono dei professori* sono dunque diverse dal punto di vista strutturale, al punto che mentre *dei professori* è il Soggetto dell'ultima, nella prima il Soggetto è omesso (*Essi sono dei professori* o *Questi libri sono dei professori*). L'algoritmo deve poter individuare tali differenze strutturali, come si propone nella porzione di codice che segue, che ha individuato due occorrenze di articolo partitivo con *esserci* (*ci sono delle forme e delle sequenze che si ripetono*) senza però essere riuscito ad individuare l'articolo partitivo nel seguente esempio: ([le mani del gorilla] *sono ancora in realtà delle zampe*):

```
# classificazione: 3=nome predicativo,
#                  no soggetto;
# 4=frase esistenziale: partitivo e soggetto

elif is_copulativo is True:
  if is_verbo(tt,frase[i-1],copulativi):
        if frase[i-2] != 'ci':
                classificazione = 3
        else:
                classificazione = 4
elif is_verbo(tt,frase[i-2],copulativi):
    if frase[i-3] != 'ci':
        classificazione = 3
    else:
       classificazione = 4
```

3.4 *Dei* in Soggetti o Oggetti di verbi transitivi e intransitivi

Se, al parsing, un elemento della serie *maybe_partitive* non è riconosciuto come complemento di specificazione, giacché non preceduto né da un nome né da un aggettivo (§ 3.1), oppure se la stringa non contiene né complementi retti (§ 3.2) né un'occorrenza di *essere* copula o di *esserci* (§ 3.3), allora siamo in presenza di un articolo partitivo in un nesso legittimato da un verbo transitivo o intransitivo, come in *Lui per trattenerla le dà dei piccoli morsi a una zampa* e *Esistono delle vie e delle piazze*. In questi casi *essere* può ovviamente ricorrere, ma come ausiliare perfettivo, dunque in combinazione con un participio passato: *Delle ombre silenziose si sono mosse sulla sabbia*. Si tratta in tutto di 10 delle 13 occorrenze complessive di articolo partitivo (2,7% del corpus, tre con *esistere*), così identificate:

```
# classificazione: 5=articolo partitivo
# post-verbale
# 6=articolo partitivo pre-verbale
elif is_forma_verbale is True:
    if frase[j] in precedenti:
      classificazione=5
    elif frase[j] in successivi:
      classificazione=6
```

4 Conclusioni

La procedura di disambiguazione automatica delle sequenze introdotte da *di* + articolo partitivo qui presentata ha dato luogo a risultati promettenti, in particolare per l'identificazione dei complementi di specificazione. Perchè si possa parlare di *information retrieval* è però necessario un campione statistico di una certa rilevanza; l'esiguità del numero di frasi ricadenti nei rimanenti casi di cui alla sezione 3 renderebbe i relativi indicatori privi di utilità, per cui si è scelto di non proporli. Risulta necessario operare ancora sul corpus sia per trattare le rimanenti occorrenze già identificate, sia per arricchirlo di nuove frasi. Inoltre, poichè l'algoritmo lavora per eliminazione, potrebbe essere utile proporre un diverso ordine di valutazione dei casi, in vista di risultati migliori.

References

Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Vito Pirrelli. 2005. Climbing the path to grammar: A maximum entropy model of subject/object learning. In *Proceedings of the Workshop on Psychocomputational Models of Human Language Acquisition*, PMHLA '05, pages 72–81, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ignazio Mauro Mirto and Emanuele Cipolla. 2017. Nooj assisted automatic detection of errors in auxiliaries and past participles in italian. In *Proceedings of the NooJ 2017 International Conference*.

Mark Stevenson and Yorick Wilks. 2003. Word sense disambiguation. *The Oxford Handbook of Comp. Linguistics*, pages 249–265.

Lucien Tesnière. 1959. *Eléments de Syntaxe Structurale*. Paris.