Relazione TLN - Esercizio 3

Davide Giosa[1], Roger Ferrod[1]

Dipartimento di Informatica, Università degli Studi di Torino
davide.giosa@edu.unito.it,
roger.ferrod@edu.unito.it

1 Introduzione

La relazione seguente riguarda la risoluzione dell'Esercizio 3, ossia la traduzione interlingua da inglese ad italiano di alcune frasi date in input. La traduzione interlingua, al contrario delle altre tipologie di traduzione automatica, estrae una rappresentazione astratta e indipendente dal linguaggio (semantica) dalla frase originale e sulla base di essa svolge l'operazione di traduzione, generando una nuova frase. Tale approccio alla traduzione offre numerosi vantaggi, tra cui la possibilità di traduzioni tra lingue molto diverse tra loro (e.g. inglese/arabo), a scapito però di una crescente difficoltà nel definire una rappresentazione semantica su vasti domini. Per questo motivo, in questa esercitazione, ci siamo concentrati su tre frasi d'esempio: su di esse abbiamo costruito una rappresentazione del dominio e, infine, abbiamo aggiunto ulteriori frasi appartenenti allo stesso dominio definito in precedenza.

2 Soluzione proposta

Il progetto sviluppato si compone di diversi moduli eseguibili in cascata che elaborano l'informazione traducendo un insieme di frasi (presenti in un file di testo) da inglese a italiano. Il primo passo consiste nel parsificare ogni frase utilizzando la grammatica sviluppata; inseguito, partendo dalla rappresentazione semantica della frase, viene generato, nella lingua di arrivo, un nuovo albero a dipendenze secondo un preciso template e traducendo i termini in modo univoco attraverso un dizionario; infine viene generata una frase attraverso il software *SimpleNLG*. Il risultato finale viene riportato in un file testuale.

2.1 Risorse utilizzate

Nello sviluppo del progetto sono state impiegate le librerie *nltk* e *SimpleNLG* per elaborare il linguaggio naturale e alcune librerie di supporto (*JSON* e *Jackson*) per manipolare le strutture dati necessarie. In particolare *nltk* fornisce gli strumenti per parsificare il testo, data una grammatica annotata semanticamente, e per navigare l'albero creato secondo la rappresentazione propria di *nltk* (la libreria utilizza il lambda calcolo per ottenere una rappresentazione in logica del primo ordine della semantica della frase).

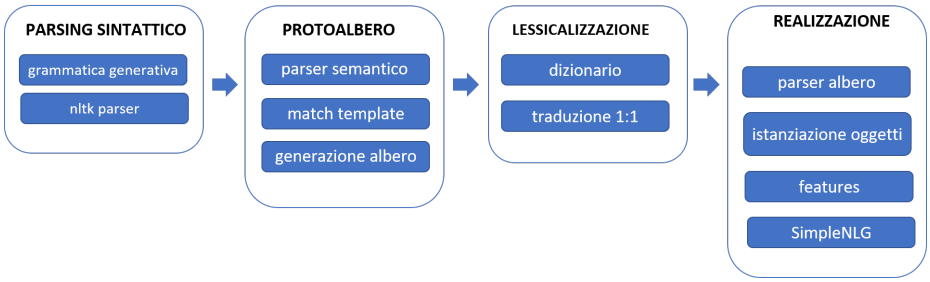

Figura 1. Pipeline del progetto

SimpleNLG-it fornisce invece i metodi per generale la frase e permette di impostare le *features* (tempo verbale, genere eccetera).
Al fine di fare comunicare le due unità software di cui si compone il progetto, si è scelto di rappresentare la struttura dati intermedia (proto-albero[1]) in formato JSON, facilitando in questo modo la creazione e lettura del dato.
Il codice è scritto in Python (3.7.4) e Java (1.8).

3 Il metodo proposto

Di seguito viene analizzato in dettaglio il procedimento seguito.

3.1 Grammatica

La grammatica utilizzata è composta da una serie di simboli non terminali, che rappresentano i costituenti della frase (NP, VP, PP), e simboli terminali dotati di una rappresentazione semantica (lambda espressioni), PoS tag e alcune *features* utili per il processo di traduzione. Le *features* usate sono:

- `LOC` per espressioni locative
- `POSS` per espressioni possessive
- `PERS` per espressioni personali
- `NUM` indica il numero (singolare o plurale)
- `GEN` indica il genere (maschile o femminile)

L'utilizzo di tali *features* si è rivelato essenziale per gestire le parole sconosciute e per differenziare parole che necessitano o meno di articoli.
Data la particolarità di *There is* è stata introdotta una categoria sintattica EX che rappresenta la componente *There* la cui semantica si lega a quella del verbo che segue. È stato introdotto anche un tag CP per rappresentare una copula e AUX per i verbi ausiliari, entrambi hanno una semantica del tipo $\lambda x.x$ in quando privi di significato proprio. Si è rivelato necessario anche l'uso di una produzione NP→N che rappresentasse semanticamente l'articolo implicito che

[1] La struttura dati verrà spiegata in dettaglio nella sezione 3.4

precede il nome: il fenomeno è osservabile, ad esempio, nella frase *"You are imagining (some) things"* o in *"[..] on my head"* traducibile in italiano con *"[..] su la mia testa"*. L'articolo è necessario in quanto, semanticamente, introduce l'esistenziale che quantifica la variabile che rappresenta il nome.

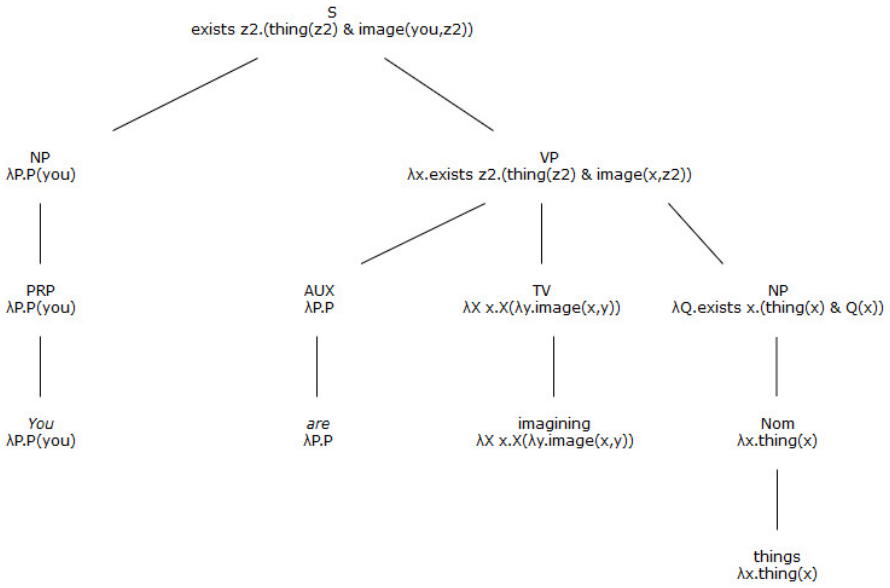

Figura 2. Struttura sintattica e semantica di *You are imagining things*

3.2 Semantica

Nella rappresentazione semantica in logica del primo ordine, ogni variabile è quantificata da un esistenziale. Al fine di rendere disponibile l'uso di avverbi è stato necessario reificare, secondo una rappresentazione Neo-Davidsoniana, i verbi intransitivi ottenendo dunque una rappresentazione del tipo: *verb(e), agent(e,x), adverb(e)*. In modo analogo a quanto avviene per gli avverbi, anche i modificatori di tipo JJ (i.e. aggettivi) vengono trattati come formule del tipo $\lambda x.adj(x)$. Per i verbi transitivi, invece, è stata scelta una rappresentazione del tipo *verb(x,y)* dove x indica il soggetto e y l'oggetto. Infine, attraverso il *type raising* è stato possibile gestire i termini che compaiono con il ruolo di soggetto (ad esempio i nomi proprio o il pronome personale *you*), tali termini sono nelle forma $\lambda P.P(term)$.

3.3 Sentence plan

In generale, partendo dalla rappresentazione semantica delle frasi proposte è possibile categorizzarle in alcune classi:

- ***exists x.(obj(x), verb(subj,x))*** semplice frase dichiarativa con soggetto, verbo (transitivo) e oggetto (diretto)
- ***exists x.(exists e.(VP(e), exists z.(subj(x), complement(x,z))))*** frase con un complemento che mette in relazione il soggetto con qualcosa (un generico NP)
- ***exists x.(subj(x), exists e.(VP(e), exists y.(Pred(e,y))))*** frase con soggetto, verbo e predicato che mette in relazione il verbo con qualcosa (un generico NP)

L'esempio *You are imagining things* appartiene, dunque, alla prima tipologia. Grazie all'utilizzo di espressioni regolari è possibile trovare una corrispondenza tra la frase ed il template, il quale rappresenta solamente il nucleo semplificato della frase; per questo motivo alla prima frase corrisponderanno anche *Irene eat an apple* e *Irene eat a green apple* (l'introduzione di uno o più aggettivi, cosi come avverbi o altri modificatori, sono gestibili tramite espressioni regolari).
Ogni template contiene inoltre un insieme di regole che permettono di costruire un albero a dipendenze in italiano, introducendo tutti gli elementi necessari alla realizzazione della frase (ad esempio in italiano vengono usati articoli che nella frase originale in inglese non compaiono). La ricerca dei componenti della frase (soggetto, verbo, oggetto ecc) avviene navigando l'albero e cercando tutte le occorrenze della variabile presa in esame. Più in dettaglio l'algoritmo parte dal verbo (la cui posizione all'interno della frase è ben nota a seconda della tipologia di frase, in questa analisi è necessario prestare attenzione a tutti i modificatori e componenti opzionali che si aggiungono al kernel della frase), ricava la variabile che rappresenta il soggetto (a seconda della tipologia di verbo: TV o IV) e cerca tutte le occorrenze di tale variabile nell'espressione, il soggetto infine viene facilmente identificato in quanto l'unico di tipo nome (proprio o comune) mentre tutte le altre occorrenze verranno considerate come modificatori del soggetto (il ruolo esatto è determinabile anche in questo caso dal PoS tag). Allo stesso modo si procede per identificare avverbi e complementi. La rappresentazione *exists x.(thing(x) & image(you, x))* fornisce in modo immediato l'informazione del soggetto (*you*) e oggetto (x): ricercando all'interno della frase la variabile x, si conclude che *thing* è l'oggetto. Allo stesso modo in *exists z1.(apple(z1) & green(z1) & eat(irene,z1))*, poiché *apple* e *green* hanno tag diversi, si può stabilire che *apple* rappresenti l'oggetto (tag N) e *green* il modificare dell'oggetto (tag JJ). Un esempio di albero è riportato in Figura 3.

3.4 Lessicalizzazione

Poiché la semantica inglese e italiana coincidono, non vi è la necessità di tradurre il livello semantico; al contrario è necessario tradurre i termini da inglese a italiano. Per semplicità si è scelto di utilizzare un dizionario che associa in modo

univoco (senza tenere conto di ambiguità) una parola in inglese con il corrispettivo termine in italiano. Prima di popolare l'albero a dipendenze è stata dunque applicata una fase di traduzione 1:1.
Si è deciso di creare una struttura dati (di seguito denominata *proto-albero*) contenente:

- nelle foglie, le parole della frase tradotte in italiano
- nei nodi interni, i costituenti della frase. I figli di un nodo interno sono gli attributi che *SimpleNLG* richiede per quel tipo di nodo.

Un esempio è riportato in Figura 4. All'albero risultante sono state aggiunte altre informazioni utili a *SimpleNLG* per la generazione della frase, come i tempi verbali, il genere ed il numero. Infine, l'albero è stato trascritto in formato JSON e passato al modulo che si occupa della realizzazione della frase.

3.5 Realizzazione frase

SimpleNLG è uno strumento, scritto in Java, di realizzazione di frasi nel contesto della *Natural Language Generation* e gestisce: il sistema morfologico-sintattico (possiede un proprio lessico), *realizer* (genera il testo a partire dalla struttura sintattica) e *microplanning*. Il nostro ultimo componente, anch'esso scritto il Java, parsifica il file JSON contenente il proto-albero della frase tradotta e realizza la frase utilizzando le API fornite da *SimpleNLG*. Nello svolgere questa operazione bisogna prestare attenzione alla genericità del problema, si è proceduto quindi nel seguente modo: per ogni nodo del proto-albero (i.e costituente) si crea un oggetto `PhraseElement` (la tipologia esatta viene determinata dal tipo del costituente e può essere `NPPhraseSpec`, `VPPhraseSpec`, `PPPhraseSpec`), ad ogni nodo vengono quindi aggiunti i nodi figli o le foglie a seconda del ruolo che svolgono nell'albero, ad esempio un oggetto di tipo `NPPhraseSpec` è caratterizzato da due figli: *specifier* e *noum* che possono essere impostati attraverso i metodi `.setSpecifier()` e `.setNoum()`. Infine, attraverso il metodo `.setFeature()`, vengono aggiunte le proprietà del nodo (e.g. tempo verbale, genere, numero) che sono state catturate nella fase di parsing. In questo modo è possibile gestire anche i termini che non compaiono nel *lexicon*, come ad esempio *"opportunità"*.

4 Risultati

In questa sezione vengono analizzati i risultati ottenuti dal sistema da noi creato.

4.1 Semantica

Di seguito è riportata la rappresentazione in logica del primo ordine delle tre frasi usate come esempio.

1. *You are imagining things*: exists z2.(thing(z2) & image(you,z2))
   In questa rappresentazione è possibile individuare un predicato binario *image*

(rappresentate il verbo transitivo) che mette in relazione il soggetto (rappresentato dalla costante *you*) e l'oggetto (rappresentato dalla variabile *x*, ossia *thing*).

2. *There is a price on my head*: exists x.(exists e.(presence(e) & agent(e,x)) & exists z8.(my(z8) & head(z8) & price(x) & on(x,z8)))
   Dalla formula logica è possibile estrarre il verbo (*presence*) rappresentato dall'evento *e*. Il predicato *presence* fa riferimento ai termini *"there is"* della frase originale e corrisponde, dunque, al verbo *"essere"* con lo specifico significato di *"esistere"* ovvero *"è presente"*. Il soggetto, indicato dal predicato *agent*, è rappresentato dalla variabile *x* ed è associato al termine *price* e al predicato *on* che lo mette in relazione (*"sopra"*) con l'oggetto identificato da *z8*. Infine, poiché *head* è di tipo N (nome) e *my* di tipo JJ (aggettivo), il sistema è in grado di concludere che la variabile *z8* fa riferimento all'oggetto *head*, il quale è modificato da *my*.

3. *Your big opportunity is flying out of here*: exists x.(your(x) & big(x) & opportunity(x) & exists e.(fly(e) & agent(e,x) & out(e) & exists y.(from(e,y) & here(y))))
   In modo analogo all'esempio precedente, in questa formula *fly* è il verbo ed è modificato dall'avverbio *out* e messo in relazione (*from*) con il predicato *here*. Il soggetto è identificato da *x*, ovvero *opportunity*, ed è modificato da *your* e *big*. Da notare che, poiché *here* può appartenere a diversi tag, in questo caso esso è categorizzato come nome[2] con l'aggiunta di una *feature* LOC.

Allo steso modo, abbiamo provato a tradurre alcune frasi, con struttura simile alle precedenti, di cui riportiamo la formula in FOL:

- *Irene eat an apple* → exists z14.(apple(z14) & eat(irene,z14))
- *Irene eat a green apple* → exists z15.(apple(z15) & green(z15) & eat(irene,z15))
- *Alex explains the lesson* → exists z16.(lesson(z16) & explain(alex,z16))
- *The little mouse escapes quickly from the cat* → exists x.(mouse(x) & little(x) & exists e.(escape(e) & agent(e,x) & quick(e) & exists y.(from(e,y) & cat(y))))

4.2 Albero sintattico

Di seguito sono riporati gli alberi sintattici, annotati semanticamente, ottenuti dal parser:

[2] In realtà, in italiano, l'unica accezione possibile è quella di avverbio e anche in inglese è più corretta la categorizzazione di avverbio ma, dal punto di vista della rappresentazione semantica e della struttura a costituenti, risulta più semplice trattare l'oggetto come un nome facente parte di un NP o PP; si può pensare quindi che *"here"* sostituisca l'espressione "this place".

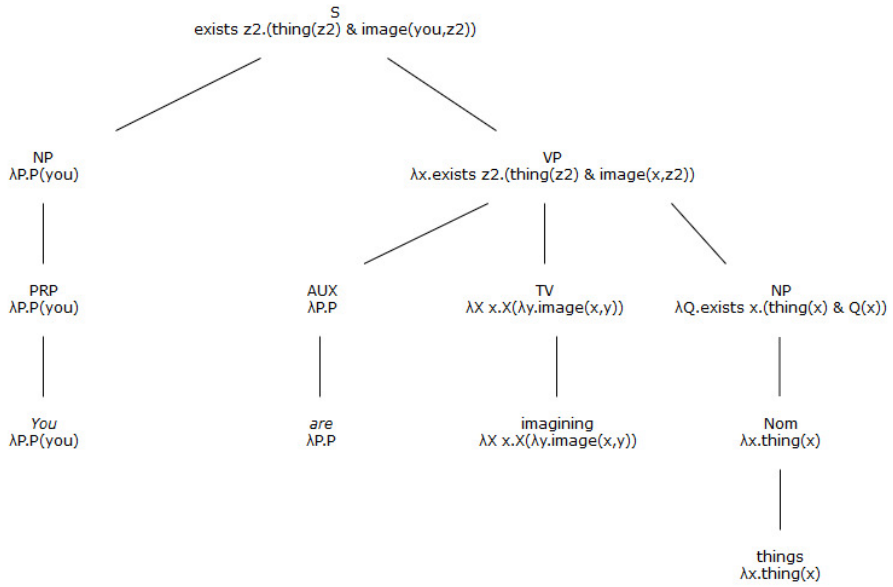

Figura 3. Struttura sintattica e semantica di *You are imagining things*

You are imagining things In Figura 3 è rappresentato l'albero sintattico, dotato di semantica, della prima frase. Bisogna prestare attenzione al nodo NP che include l'informazione implicita di un articolo e, di conseguenza, introduce l'esistenziale e un nuovo predicato Q che, tramite beta-riduzione, verrà unificato con la semantica del costituente *TV*. Alla frase corrisponde il proto-albero di Figura 4 nel quale è presente, in forma esplicita, l'articolo.

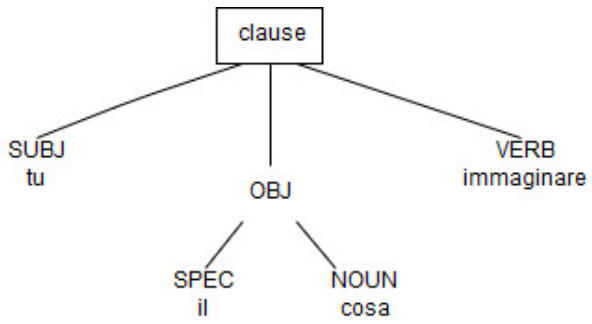

Figura 4. Struttura sintattica e semantica di *You are imagining things*

There is a price on my head In Figura 5 è rappresentato l'albero sintattico corrispondente. La particolarità di questo albero risiede nella ricorsività del costituente *PP*. Alla frase, infatti, corrispondono due *Prepositional Phrase*: il primo introduce la preposizione locativa *on*, il secondo possedimento (*my*) dell'oggetto *head*. Anche in questo caso è presente un articolo implicito che introduce *head*. Il proto-albero, rappresentato in Figura 7, rende esplicito l'articolo e ripropone (in altri termini) la struttura annidata che caratterizza dal struttura originale.

Your big opportunity is flying out of here In Figura 6 è riportato l'albero sintattico della frase. In questo caso è importante ricordare che a *here* è associata una *feature* LOC e, di conseguenza, la formula è priva di esistenziali che quantificano la variabile (ovvero è privo di articolo). Un altro elemento importante è la presenza dell'avverbio *out* che richiede l'utilizzo di una produzione specifica tale da poter legare l'evento *e*, che identifica il verbo, con la rappresentazione dell'avverbio e che, allo stesso tempo, lega *e* con la *Prepositional Phrase*. Il proto-albero corrispondente è riportato in Figura 8.

4.3 Traduzione

Di seguito sono riportate tutte le traduzioni effettuate dal sistema al fine di verificarne l'accuratezza:

1. You are imagining things → Tu stai immaginando le cose.
2. There is a price on my head → Una taglia esiste sopra la mia testa.
3. Your big opportuniy is flying out of here → La tua opportunità grande sta volando via da qui.
4. Irene eat an apple → Irene mangia la mela.
5. Irene eat a green apple → Irene mangia la mela verde.
6. Alex explains the lessos → Alex spiega la lezione.
7. The little mouse escapes quickly from the cat → Il topo piccolo scappa veloce dal gatto.

5 Appendice

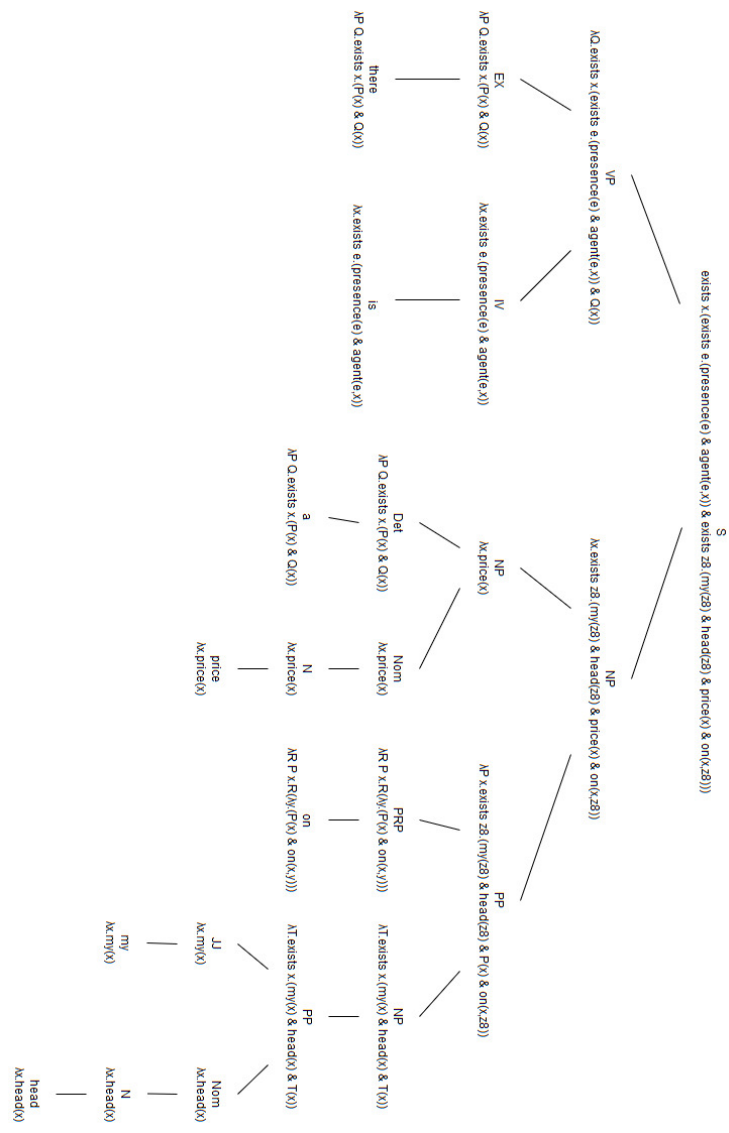

Figura 5. Struttura sintattica e semantica di *There is a price on my head*

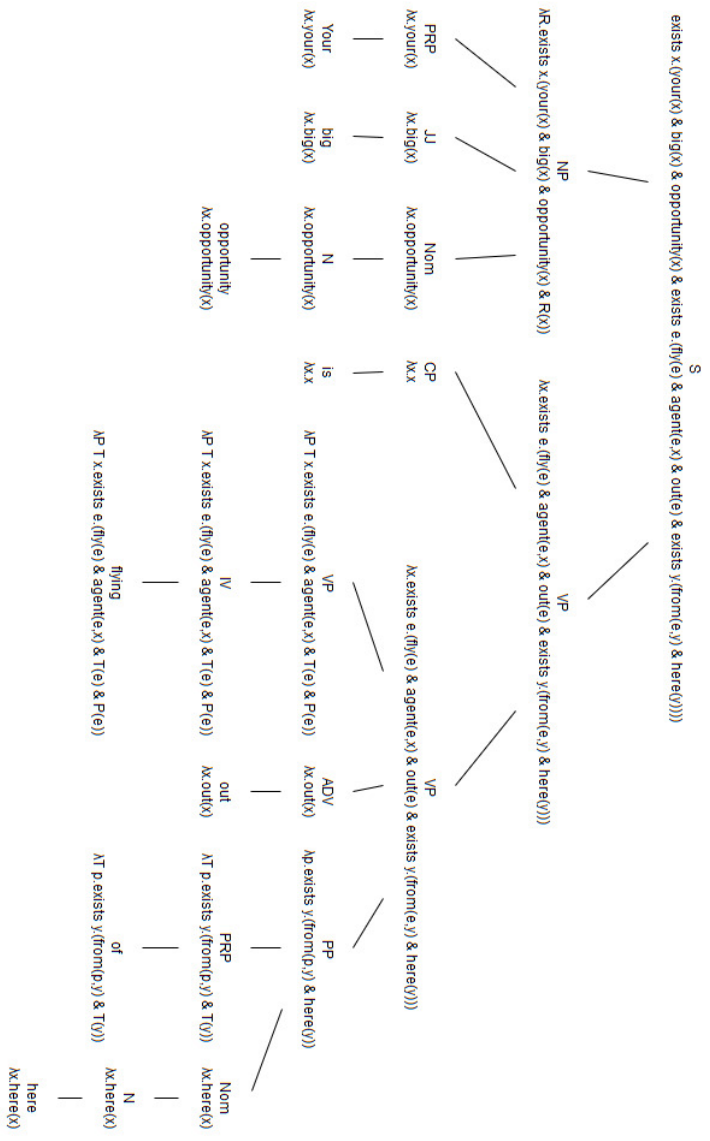

Figura 6. Struttura sintattica e semantica di *Your big opportunity is flying out of here*

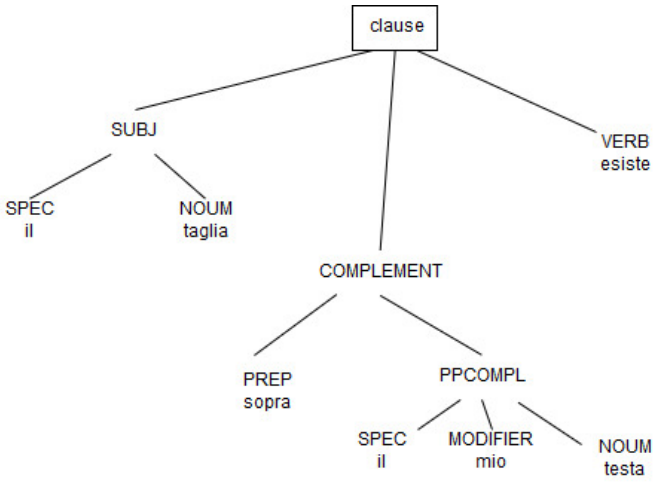

Figura 7. Proto-albero di *There is a price on my head*

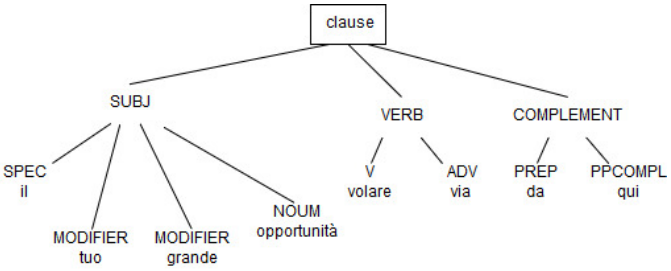

Figura 8. Proto-albero di *Your big opportunity is flying out of here*