

T-PAS: COSTRUIRE UNA RISORSA PER L'ITALIANO BASATA SULL'ANALISI DI UN CORPUS

Anna FELTRACCO

ABSTRACT • This contribution discusses some features of the T-PAS resource, a repository of typed predicate argument structures for Italian, acquired from corpora. We briefly describe T-PAS and the process of acquiring these structures and annotating the corpus. We then highlight the main characteristics of the resource that derive from the methodology used for its creation. We conclude by citing the first works that use T-PAS and proposing further developments of the resource.

KEYWORDS • Predicate Argument Structure, Lexical Resource, Corpus Analysis

1. Introduzione

La risorsa T-PAS (*Typed Predicate Argument Structures*) è una raccolta di strutture argomentali per verbi della lingua italiana (Jezek et al., 2014). In T-PAS le strutture (nel prosieguo *T-PAS*) sono motivate semanticamente e sono acquisite seguendo la procedura lessicografica CPA, *Corpus Pattern Analysis* (Hanks, 2004; Hanks and Pustejovsky, 2005). Le strutture *T-PAS* sono identificate esaminando esempi del verbo in un corpus. Gli esempi sono quindi associati alle strutture così identificate. Un importante riferimento per il progetto T-PAS è PDEV, *Pattern Dictionary of English Verb* (Hanks and Pustejovsky, 2005)¹, la più importante risorsa prodotta seguendo la procedura CPA per la lingua inglese. Per ciò che riguarda la lingua italiana, una risorsa complementare a T-PAS è LexIt (Lenci et al., 2012) che riporta informazioni acquisite in modo totalmente non-supervisionato su verbi, aggettivi e nomi. A differenza di T-PAS, in LexIt le strutture non sono distinte con criteri semantici.²

Il presente contributo si articola in sei sezioni. Nella Sezione 2 descriviamo la procedura di acquisizione delle *T-PAS* e lo stato dell'arte della risorsa alla quale l'autrice ha contribuito per la parte lessicografica. Nella Sezione 3 evidenziamo le principali caratteristiche di T-PAS: in particolare, ne presentiamo le specificità unitamente ad alcune limitazioni. Nella Sezione 4 riportiamo i primi esperimenti in cui la risorsa è stata utilizzata e i risultati ottenuti e nella Sezione

¹ <http://pdev.org.uk/>

²Tra i recenti progetti che si occupano di annotazioni verbali per la lingua italiana citiamo anche il progetto IMAGACT (Moneglia et al. 2012) il cui obiettivo è la costruzione di una ontologia interlinguistica dell'azione le cui entrate sono costituite da scene che rappresentano un prototipo dell'azione. IMAGACT si basa sull'analisi di verbi di azione ad alta frequenza estratti da corpora del parlato di lingua italiana e inglese. Citiamo inoltre il progetto COMBINET (<http://combinet.humnet.unipi.it/>), il cui obiettivo è lo studio delle combinazioni di parole nella lingua italiana e la costruzione di una risorsa lessicale di combinazioni.

5 ne proponiamo alcuni sviluppi. Infine, nella Sezione 6 riportiamo le conclusioni del contributo.

2. La risorsa T-PAS

In questa Sezione viene descritta la risorsa T-PAS e la procedura per l'acquisizione e la definizione delle strutture *T-PAS*. Vengono inoltre presentati alcuni dati relativi alla versione attuale della risorsa.

2.1 La metodologia di acquisizione delle T-PAS

La procedura per l'identificazione delle *T-PAS* prevede varie fasi, che sono di seguito descritte. Come accennato nell'Introduzione, lo sviluppo della risorsa T-PAS avviene solo attraverso l'analisi di un corpus. La prima fase, infatti, prevede l'esame di un campione di 250 concordanze del verbo nel corpus. Successivamente il lessicografo individua le *T-PAS* per il verbo in esame identificando gli elementi rilevanti per la disambiguazione del verbo nel contesto, sia a livello di struttura argomentale sia a livello di elementi lessicali per ciascuna posizione argomentale. Ogni *T-PAS* presenta quindi informazioni lessicali, sintattiche e semantiche.

Attraverso lo strumento di editing³, il lessicografo definisce la struttura della *T-PAS* marcando la presenza di *subject*, *object*, *indirect object*, *complement*, *adverbial* e *clausal*⁴. Queste posizioni argomentali vengono marcate dal lessicografo solo se la loro presenza è ritenuta essenziale per la definizione della struttura della *T-PAS* e quindi per la definizione del senso del verbo. È possibile inoltre precisare che l'argomento, seppure obbligatorio, non è sempre realizzato nei testi esaminati (per esempio nel caso di argomenti impliciti). Inoltre è possibile marcare esplicitamente che la struttura prevede l'assenza di un complemento oggetto diretto ([[NO OBJ]]).

Per ogni *T-PAS* il lessicografo registra anche informazioni di tipo lessicale e semantico, annotando i *Tipi Semantici* (ST) per ogni posizione argomentale. I Tipi Semantici rappresentano delle generalizzazioni degli elementi lessicali più frequentemente osservati nel corpus per un certo verbo⁵. L'esempio 1 riporta la *T-PAS#2* del verbo *divorare* dove [[Document]] rappresenta la generalizzazione di {*documento, rivista, libro, romanzo ...*}:

- (1) *T-PAS#2* del verbo *divorare*:
 [[Human]-subj] *divora* [[Document]-obj]
 Esempio: L'amica divorò il romanzo.

Il lessicografo sceglie i ST più appropriati da una lista di 224 Tipi creata precedentemente operando una generalizzazione a partire dal set di elementi lessicali osservati per le posizioni

³ Lo strumento di editing è in lingua inglese come per la risorsa PDEV (Hanks and Pustejovsky, 2005). Tuttavia, rispetto alla versione inglese sono state apportate alcune modifiche. Ad esempio, mentre nella versione inglese è possibile annotare *that-clause*, nella versione italiana è possibile annotare *che-clause*. Inoltre, è stata aggiunta la possibilità di annotare l'alternanza -che [[CLAUSE]] | -di [[INF[V]]] (ciò permette, ad esempio, per il verbo *dire*, di associare alla stessa *T-PAS* i seguenti esempi: "Per questo dicono che la vita è artificiale.", "La commissione dice di aver inviato..."; *T-PAS#2* del verbo *dire*: [[Human | Human Group] -subj] *dire* che [[CLAUSE]] | di [[INF[V]]]).

⁴ Per un approfondimento su questi ruoli si rimanda a Cinková and Hanks (2010) e El Maarouf (2013).

⁵ Per convenzione i Tipi Semantici vengono indicati tra parentesi quadre e con l'iniziale maiuscola.

argomentali dei verbi in un corpus⁶. In questa lista i ST sono organizzati in una gerarchia i cui elementi sono legati da una relazione “IS-A”; ad esempio, nella Tabella 1 [[Plane]] IS-A [[Vehicle]], [[Vehicle]] IS-A [[Machine]] e così via.

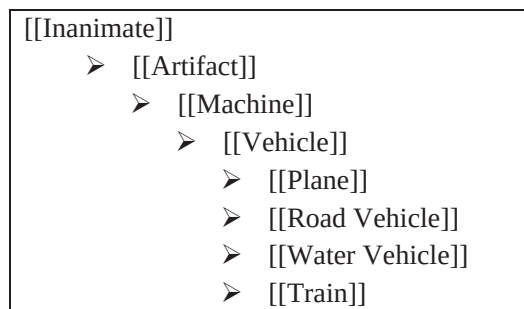


Tabella 1: Sezione della gerarchia dei Tipi Semantici.

Se la generalizzazione possibile è massima, il lessicografo può scegliere il ST [[Anything]] che rappresenta la radice della gerarchia. Viceversa, se nessuna generalizzazione è possibile per la posizione argomentale, il lessicografo annota la lista dei singoli elementi lessicali, come nell'esempio 2 che riporta la *T-PAS#3* del verbo *eleggere*⁷.

- (2) *T-PAS#3* del verbo *eleggere*:
 [[Human]-subj] *elegge* { domicilio | recapito | residenza }

Come detto, la distinzione tra *T-PAS* è principalmente semantica, per cui differenti realizzazioni sintattiche dello stesso argomento (per esempio, un nome e una frase subordinata) afferiscono a una sola *T-PAS*, purché il senso del verbo rimanga invariato, come nell'esempio 3 tratto da Jezek et al. (2014).

- (3) [[Human]-subj] finisce [[Event]-obj | di INF [V]]
 a. Finisce *l'allenamento*.
 b. Non faccio in tempo a finire *di bere la mia birra*.

Anche l'alternanza attivo-passivo viene registrata in un'unica *T-PAS*: nell'esempio che segue, 4a e 4b sono entrambi associati alla *T-PAS#2* di *abbattere*.

- (4) *T-PAS#2* di *abbattere*:
 [[Human | Event]-subj] *abbatte* [[Building]-obj]
 a. Mi parve di sentire muratori occupati ad *abbattere* muraglie o a praticare un'apertura nella roccia.
 b. Perché il muro fu *abbattuto*?

Al contrario, l'alternanza causativo-incoativo viene rappresentata con due distinte *T-PAS* (si veda esempio 5 che riporta *T-PAS#1* e *T-PAS#2* di *cuocere*).

⁶ Questa lista è stata creata analizzando le concordanze per circa 1.500 verbi in Inglese, Italiano, Spagnolo in un lavoro precedente. Si veda Jezek et al. (2014). La stessa lista è attualmente utilizzata per la risorsa inglese PDEV.

⁷ L'alternanza di più elementi è indicata con una barra verticale |.

- (5) T-PAS#1 di cuocere:
 [[Human-subj] *cuoce* [[Food]-obj]
 Esempio: Pulite gli spinaci e *cuoceteli*.

T-PAS#2 di cuocere:
 [[Food]-subj] *cuoce*
 Esempio: Mentre il sugo *cuoce* a fuoco lento [..]

Anche l'alternanza del tipo "spray-load" viene registrata con due *T-PAS* distinte, come nell'esempio in 6 per il verbo *caricare*.

- (6) T-PAS#1 di caricare:
 [[Human]-subj] *carica* [[Animate | Inanimate]-obj] (su | in [[Vehicle]] | su | in {spalle | schiena})
 Esempio: [..] controlliamo che le valige siano state *caricate* sul pullman [..]

T-PAS#2 di caricare:
 [[Human]-subj] *carica* [[Vehicle]-obj] (con | di [[Animate]] | con | di [[Inanimate]])
 Esempio: Stanno *caricando* camion e ambulanze di medicinali destinati a Najaf.

Dopo aver definito le *T-PAS*, il lessicografo associa gli esempi del corpus alla *T-PAS* corrispondente (annotando il numero identificativo della *T-PAS*). L'esempio associato alla *T-PAS* diventa quindi esempio d'uso del verbo per quel determinato senso (Popescu, 2012).

Casi particolari riguardano gli esempi che chiaramente afferiscono a una *T-PAS* ma in cui uno degli elementi lessicali in una determinata posizione argomentale non sembra essere ricompreso nel ST scelto per quella posizione: il lessicografo può allora decidere di aggiungere un ST per la posizione argomentale (qualora molti esempi presentino questa caratteristica) o di associare l'esempio alla *T-PAS* annotandolo come "anomalo". In 7 viene presentato il caso di un esempio associato alla *T-PAS#1* di *divorare* e annotato come "anomalo". La *T-PAS* infatti seleziona il ST [[Food]] come oggetto, dove [[Food]] comprende {carne, provviste, anatre, corpi, vittime, uova, ecc.}: il termine *piatto* è stato annotato come "anomalo".

- (7) Esempio con argomento "anomalo":
 T-PAS#1 di divorare:
 [[Animate]-subj] *divora* [[Food]-obj]
 Ti ho rivisto alla festa dell'Unità, mangiare quel piatto di pasta con i funghi, mi guardavi e *divoravi* quel piatto, stavo andando via quando sei arrivata tu.

Per il lessicografo, è possibile inoltre marcare esempi che contengono errori (come in 8) o esempi che risultino non chiari o non classificabili (ad esempio, perché il contesto è insufficiente per disambiguare il senso del verbo e identificare la *T-PAS* corrispondente, come in 9).

- (8) Esempio di errore per il verbo *abitare*:
 [..] oggi a Milano Iosposa, quattro giorni di sfilate dedicate agli *abiti* nuziali.
- (9) Esempio non classificabile per il verbo *divorare*:
 In fondo se avessero smesso di rapinare i viandanti, avrebbero anche smesso di *divorarli*, pensava fiduciosamente.

Infine, per ogni *T-PAS* viene creata una *implicature*, che consiste in una spiegazione del senso della *T-PAS*. Questa *implicature* è “ancorata” alla *T-PAS* perché riporta, per quanto possibile, gli stessi Tipi Semantici nelle posizioni argomentali (Cinková and Hanks, 2010).⁸

Risulta chiaro dunque che sia la struttura identificata sia i Tipi Semantici scelti dal lessicografo per ogni posizione argomentale sono rilevanti per la disambiguazione del verbo, come si vede nella Tabella 2 che riporta tre *T-PAS* del verbo *abbattere*, le *implicature* per le *T-PAS* e alcuni esempi tratti dal corpus.

Verbo: abbattere	
➤ T-PAS 2:	[[Human Event]] <i>abbattere</i> [[Building]]
IMPLICATURE:	[[Human Event]] demolisce, distrugge, butta giù [[Building]]
CORPUS:	Dovranno essere <i>abbattuti</i> in tutto altri tre immobili. ... muratori occupati ad <i>abbattere</i> muraglie...
➤ T-PAS 3:	[[Human]] <i>abbattere</i> [[Animate]]
IMPLICATURE:	[[Human]] uccide, toglie la vita a [[Animate]]
CORPUS:	...Kenai, il più giovane, <i>abbatte</i> l'orso... ...un bracconiere <i>abbatteva</i> un coniglio...
➤ T-PAS 8:	[[Weather Event]] <i>abbattersi</i> [[NO OBJ]] su [[Location]]
IMPLICATURE:	[[Weather Event]] <i>si rovescia</i> [[NO OBJ]] su [[Location]]
CORPUS:	...un acquazzone <i>si abbatte</i> sull'isola... Una pioggia insistente <i>si abbatte</i> sulla città.

Tabella 2: T-PAS#2, T-PAS #3 e T-PAS#8 del verbo abbattere.

Le tre componenti della risorsa T-PAS (Figura 1) sono dunque: la lista delle *T-PAS*, la lista dei Tipi Semantici, la lista di esempi del corpus associata alle *T-PAS*.

⁸ Per la composizione della *implicature* il lessicografo spesso ricorre all'uso di sinonimi del verbo; tuttavia, può decidere di formulare una *implicature* anche utilizzando delle perifrasi. La *implicature* dunque non rappresenta sistematicamente un elenco di termini sinonimici, ma ha l'obiettivo di chiarire il senso del verbo nella *T-PAS*.

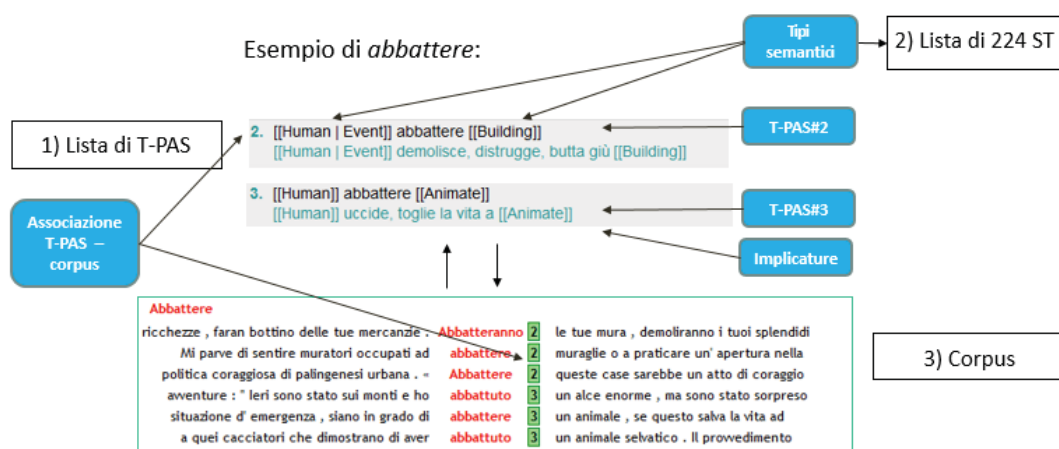


Figura 1: Le componenti di T-PAS.

2.2 La versione corrente di T-PAS

T-PAS è stata sviluppata presso il Dipartimento di Studi Umanistici dell'Università di Pavia, in collaborazione con il gruppo Human Language Technology della Fondazione Bruno Kessler (FBK), Trento e il supporto tecnico della Facoltà di Informatica presso la Masaryk University a Brno (CZ). La risorsa è distribuita sotto una licenza Creative Commons Attribution 3.0⁹.

La prima versione di T-PAS (Jezek et al., 2014) è composta da 1.000 verbi analizzati. Questi verbi sono stati estratti in modo casuale dalla lista totale di verbi nella risorsa Sabatini Coletti 2008 (Sabatini e Coletti, 2007), tenendo conto della loro polisemia nella risorsa e secondo le seguenti proporzioni: il 10% dei verbi presenta 2 sensi nel Sabatini Coletti 2008, il 60% dai 3 ai 5 sensi e il 30% dai 6 agli 11 sensi. Il numero totale delle *T-PAS* identificate nella risorsa è di 4241; il numero minimo di *T-PAS* per verbo è 1 e il numero massimo è 20.

Il corpus di riferimento per la risorsa è un versione ridotta di ItWAC (Baroni and Kilgarriff, 2006)¹⁰.

Numero totale verbi	1000
Numero totale T-PAS	4241
Numero minimo di T-PAS per verbo	1
Numero massimo di T-PAS per verbo	20

Tabella 3: Dati quantitativi di T-PAS.

3. Caratteristiche salienti di T-PAS

In T-PAS vengono fornite informazioni linguistiche per ogni *T-PAS* a livello lessicale (ossia la selezione del Tipo Semantico o del set lessicale per ogni posizione argomentale), a livello sintattico (la struttura della *T-PAS* fornisce informazioni sintattiche relativamente alle relazioni grammaticali degli argomenti) e a livello semantico (le *T-PAS* vengono distinte semanticamente).

⁹ tpas.fbk.eu

¹⁰ Per chiarimenti sulla preparazione del corpus si veda Jezek et al. (2014).

T-PAS è dunque una risorsa importante per studi di linguistica italiana: l'analisi delle *T-PAS* può, ad esempio, agevolare lo studio delle strutture predicato-argomento e lo studio degli argomenti precedentemente definiti “anomali”.

Un'altra caratteristica peculiare di T-PAS è legata alla metodologia di acquisizione delle singole *T-PAS*: diversamente da altre risorse lessicali, le *T-PAS* sono definite partendo dall'analisi di un corpus. Di conseguenza, in T-PAS sono registrati solo sensi del verbo effettivamente utilizzati. Un altro aspetto fondamentale della procedura seguita per lo sviluppo della risorsa riguarda l'associazione degli esempi alle rispettive *T-PAS*, il che permette di eliminare la necessità di creare esempi artificiali o di cercare esempi *ad-hoc* in un corpus. Inoltre il lessicografo, nell'assegnare un esempio alla *T-PAS*, non solo contribuisce ad arricchire la risorsa ma disambigua di fatto il senso del verbo nell'esempio. Infine, visto che gli esempi nel corpus rappresentano evidenze empiriche delle scelte del lessicografo, il processo di assegnazione gli permette di verificare le scelte operate per la definizione della *T-PAS* ed eventualmente di rivedere alcuni elementi della struttura, correggere le *implicature* e rivedere la scelta dei Tipi Semantici.

T-PAS è come si è visto una risorsa ricca di informazioni (utili sia nel campo della ricerca linguistica, sia nel campo del Trattamento Automatico del Linguaggio). L'annotazione manuale garantisce un buon livello di affidabilità, ma vi sono alcune limitazioni insite nel procedimento stesso. Infatti, l'annotazione manuale degli esempi richiede tempistiche elevate e sono possibili errori di disattenzione del lessicografo in tutte le fasi descritte, dalla costruzione delle *T-PAS* all'assegnazione degli esempi. Inoltre, è possibile che si registri un disaccordo tra più lessicografi in merito, ad esempio, al numero delle *T-PAS* identificate (diversa granularità) o alla scelta del ST per una posizione argomentale. Nella Tabella 4 viene riportato un esempio di disaccordo nell'identificazione delle *T-PAS* tra due lessicografi, nominati Annotatore 1 e Annotatore 2, per il verbo *colpire*. Mentre l'Annotatore 1 distingue due *T-PAS* di *colpire* (ossia *T-PAS#1* per “battere, percuotere, picchiare volontariamente”, come in “Condannato ad altri quattro anni di carcere per aver *colpito* un altro detenuto, Ron comincia a ...” e *T-PAS#2* per “toccare, centrare violentemente” come in “una delle parti del chakram *colpisce* in pieno il petto della creatura ...”), l'Annotatore 2 crea una sola *T-PAS* per i due casi.

Verbo: <i>colpire</i>	
Annotatore 1:	
T-PAS 1	[[Human1]] <i>colpire</i> ([[Human2]] [[Human1]] batte, percuote, tocca con violenza, picchia volontariamente ([[Human2]])
T-PAS 2:	[[Animate1 Energy Particle1]] <i>colpire</i> [[Animate2 Inanimate Particle2]] [[Animate1 Energy Particle1]] <i>centra violentemente</i> [[Animate2 Inanimate Particle2]]
Annotatore 2:	
T-PAS 1:	[[Animate1 Physical Object1]] <i>colpire</i> [[Animate2 Physical Object2]] [[Animate1 Physical Object1]] <i>picchia, tocca violentemente</i> [[Animate2 Physical Object2]]

Tabella 4: Esempio di disaccordo nell'identificazione delle *T-PAS* tra due annotatori.

4. Primi esperimenti con T-PAS

Le informazioni raccolte in T-PAS rendono la risorsa estremamente interessante per la sua applicazione nel campo del Trattamento Automatico del Linguaggio, per esempio nei task in cui è necessario disambiguare il senso del verbo. Popescu et al. (2014) propongono un sistema di

Word Sense Disambiguation sviluppando per la disambiguazione del verbo un modello probabilistico che utilizza le *T-PAS* e, in particolare, i Tipi Semantici identificati per ogni posizione argomentale. I risultati mostrano una buona precisione.

In Jezek et al. (2014) viene presentato un esperimento che valuta l'accordo tra due annotatori che hanno contribuito alla creazione della risorsa nell'operazione di raggruppamento degli esempi nel corpus. Viene richiesto ai due annotatori di creare le *T-PAS* per gli stessi verbi (50) analizzando e annotando per ogni verbo gli stessi esempi (250). Ne risultano due distinti set di *T-PAS*, uno per ogni annotatore. L'accordo viene determinato calcolando il grado di similarità tra le due distribuzioni. I risultati mostrano un buon grado di sovrapposizione tra le due distribuzioni, comprovando l'affidabilità della risorsa (Jezek et al. 2014)¹¹.

T-PAS è stata inoltre utilizzata in un esperimento pilota per testare uno schema di annotazione per i tipi di relazioni di opposizione tra frame verbali (Feltracco et al., 2015). L'esperimento coinvolge due annotatori a cui viene richiesto di esaminare coppie di verbi e rispettive *T-PAS*. Per ogni coppia di verbi, gli annotatori devono identificare la presenza o l'assenza di una relazione di opposizione tra le *T-PAS* dei due verbi ed eventualmente marcare il tipo di opposizione secondo lo schema proposto¹². Per svolgere questo compito, entrambi gli annotatori hanno ritenuto essenziale che per ogni posizione argomentale fossero espressi i Tipi Semantici, al fine di disambiguare i vari sensi del verbo e conseguentemente annotare la relazione di opposizione solo tra sensi verbali opposti; era comunque possibile - ed è stata ritenuta utile per svolgere il task - anche l'analisi degli esempi associati alle *T-PAS*. Questo esperimento pilota rappresenta il primo tentativo di annotazione di una relazione semantica tra le *T-PAS* della risorsa e ha permesso di evidenziare alcuni aspetti importanti per l'annotazione delle relazioni semantiche tra *T-PAS*¹³.

T-PAS può inoltre trovare applicazione nei task che richiedono di identificare automaticamente gli argomenti del verbo, come per esempio nel *Semantic Role Labeling* (Gildea and Jurafsky, 2002) in cui i ruoli semantici sono individuati anche sulla base della struttura sintattica della frase.

5. Possibili sviluppi della risorsa

Possibili sviluppi della risorsa *T-PAS* riguardano innanzitutto l'incremento del numero di verbi analizzati. Inoltre, l'annotazione degli argomenti nelle istanze del corpus (per il momento sono taggati nel corpus solo i verbi), inclusi gli elementi considerati "anomali", potrebbe favorire, ad esempio, lo studio delle alternanze più frequenti ed essere un punto di partenza per l'estensione automatica della copertura della risorsa.

¹¹ Gli autori riportano valori medi di BCubed di 0,77 (F1) e di Purity di 0,60 (F1).

¹²La valutazione dell'esperimento prevede il calcolo dell'accordo tra i due annotatori che hanno annotato 25 coppie di verbi. Dai risultati emerge un alto livello di accordo sia nell'identificazione delle relazioni di opposizione (Coefficiente di Dice = 0,98) sia nell'identificazione del tipo di relazione di opposizione (accordo nel riconoscere lo stesso tipo di relazione o nel riconoscere l'assenza di una relazione di opposizione, Coefficiente di Dice = 0,97). Per maggiori dettagli si veda Feltracco et al. (2015).

¹³ Ad esempio, è emersa la necessità di annotare la relazione anche in presenza di Tipi Semantici non condivisi dalle *T-PAS* che instaurano la relazione; ad esempio, la *T-PAS#2* di *abbattere* [[Human | Event]] *abbattere* [[Building]] e la *T-PAS#1* di *costruire* [[Human | Institution]] *costruire* ([[Building | Route]]) sono in relazione di opposizione se [[Human]] è considerato soggetto (la *T-PAS#1* di *costruire* non seleziona [[Event]] come soggetto). Inoltre, è risultato necessario permettere agli annotatori di associare una *T-PAS* di un verbo a più di una *T-PAS* dell'altro verbo.

Un primo tentativo di annotazione automatica degli elementi lessicali in posizione argomentale è stato condotto in via sperimentale in Feltracco et al. (2016) con risultati positivi in termini di precisione. L'esperimento si avvale di una mappatura tra i Tipi Semantici e i lemmi della risorsa MultiWordNet (Pianta et al., 2002) e rappresenta anche il primo lavoro che combina l'utilizzo di T-PAS con un'altra risorsa per l'italiano.

Infine, la risorsa potrebbe essere arricchita con l'annotazione delle relazioni semantiche tra T-PAS; il primo lavoro in questa direzione è descritto alla Sezione 4.

6. Conclusioni

In questo contributo abbiamo presentato le caratteristiche principali di T-PAS, una risorsa consistente in una raccolta di strutture predicato-argomento per i verbi della lingua italiana. In T-PAS, la definizione delle strutture e la selezione dei Tipi Semantici per ogni posizione argomentale è basata sull'osservazione del verbo in esempi tratti da un corpus, esempi che vengono poi associati alle strutture T-PAS di pertinenza. Abbiamo evidenziato, come elementi maggiormente caratterizzanti della risorsa, il fatto che, proprio perché basata sull'analisi di un corpus, T-PAS registra per ogni verbo solo i sensi effettivamente utilizzati e che gli stessi esempi del corpus agevolano la pratica lessicografica di definizione delle T-PAS. Scopo del progetto T-PAS è quindi la costruzione di una risorsa per l'italiano basata su dati empirici. Il risultato è una risorsa ricca di informazioni lessicali, sintattiche e semantiche utili sia per l'indagine linguistica sia per la ricerca in campo computazionale.

BIBLIOGRAFIA

- Baroni, M., Kilgarriff, A. (2006), "Large Linguistically - Processed Web Corpora for Multiple Languages". In Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations, pages 87-90. Association for Computational Linguistics.
- Cinková, S., Hanks, P. (2012), Validation of Corpus Pattern Analysis. - Assigning pattern numbers to random verb samples. http://ufal.mff.cuni.cz/spr/data/publications/annotatio_n_manual.pdf.
- El Maarouf, I. (2013), Methodological aspects of Corpus Pattern Analysis. In *Proceedings of the 34th ICAME conference (ICAME 2013)*, Santiago de Compostela, Spain.
- Feltracco, A., Jezek, E., Magnini, B. (2015), Opposition relations among verb frames. In *Proceedings of The Third Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 16-24, Denver, Colorado, June. Association for Computational Linguistics.
- Feltracco, A., Gatti L., Magnolini S., Magnini, B., Jezek, E. (2016), Using WordNet to Build Lexical Sets for Italian Verbs. In *Proceedings of the Eighth Global WordNet Conference (GWC'16)*, Bucharest, Romania, January.
- Gildea D., Jurafsky D. (2002), Automatic labeling of semantic roles. *Computational linguistics*, 28(3), 245-288.
- Hanks, P., Pustejovsky, J. (2005), A Pattern Dictionary for Natural Language Processing. In *Revue française de linguistique appliquée*, 10(2).
- Hanks, P. (2004), Corpus Pattern Analysis. In *Proceedings of the XI EURALEX International Congress*. Lorient, France, July.
- Jezek, E., Magnini, B., Feltracco, A., Bianchini A., Popescu, O. (2014), T-PAS: A resource of corpus-derived Typed Predicate Argument Structures for linguistic analysis and semantic processing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May.

-
- Lenci A., Lapesa, G., Bonansinga, G. (2012), LexIt: A Computational Resource on Italian Argument Structure. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3712-3718, Istanbul, Turkey, May.
- Moneglia, M., Monachini, M., Calabrese, O., Panunzi, A., Frontini, F., Gagliardi, G. and Russo, I. (2012), The IMAGACT Cross-linguistic Ontology of Action. A new infrastructure for natural language disambiguation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2606-2613, Istanbul, Turkey, May.
- Pianta, E., Bentivogli, L., Girardi, C. (2002), MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet (GWC02)*, Mysore, India, January.
- Popescu, O., Vo, N.P.A., Feltracco, A., Jezek, E., Magnini, B. (2014), Toward Disambiguating Typed Predicate Argument Structures for Italian. In *Proceedings of the First Italian Conference on Computational Linguistics (CLIC-it 2014)*, Pisa, Italy, December.
- Popescu, O. (2012), Building a resource of patterns using semantic types. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2999-3006, Istanbul, Turkey, May.
- Sabatini F., Coletti V. (2007), *Dizionario della lingua italiana 2008*. Milano: Rizzoli Larousse.

ANNA FELTRACCO • Ph.D. student in Linguistic Sciences at the University of Pavia and the University of Bergamo in partnership with Fondazione Bruno Kessler, Trento, Italy where she is pursuing her Ph.D. within the Human Language Technologies Group. She has collaborated in the T-PAS project as annotator, participating in the interannotator agreement evaluation and other experiments using the resource which are cited in this contribution. Her research interests cover corpus linguistics and natural language processing.

E-MAIL • feltracco@fbk.eu