Acquiring typed predicate-argument structures from corpora

Elisabetta Jezek
Università di Pavia

Jornadas internacionales
Los verbos de movimiento
Estudio teórico y aplicación lexicográfica
Madrid, October 4-5, 2012
The goal of the presentation is to illustrate the methodology we are currently using to acquire typed predicate-argument structures from corpora, with the aim of compiling a repository of corpus-based patterns for Italian verbs, and obtaining an empirically sound inventory of argument type shiftings in context.

- Examine the main problems that arise in the identification of the frames.
- Focus on tagging of corpus instances with type mismatches.
- Ongoing evaluation and improvements.
Repository of typed predicate-argument structures for Italian.

Under development at the University of Pavia, with the technical support of the Faculty of Informatics at Masaryk University (Brno) and FBK (Trento).

It currently consists of a nucleus of 350 lexical units (verbs) and about 1300 patterns.

It is a manually annotated resource.

It is conceived for linguistic research and NLP applications.

We plan to make the resource available once the goal of analyzing 1000 “verbi a polisemia media” (average polysemy verbs) is reached.
What counts as a typed predicate-argument structure?

A typed predicate-argument structure is a corpus-derived verb frame with the specification of the expected semantic type for each argument position, populated by lexical sets (Hanks 1986), i.e. the statistically relevant list of collocates that typically fill each position.

[[Human]] partecipa a [[Event]]

- Lexical set [[Event]] = {gara, riunione, selezione, manifestazione, seduta, cerimonia, conferenza, votazione, elezione, celebrazione, esequia, competizione, maratona, discussione, messa, festa, marcia, fiaccolata, trattativa, missione, commemorazione, incontro, concorso, convegno, raduno, iniziativa, stage, evento, seminario, torneo, attività, corso, asta, dibattito, progetto, festival...}
The resource consists in three modules:

- A repository of corpus-derived typed predicate-argument structures.
- An inventory of semantic types conceived as abstractions over the lexical sets found in the argument positions in the corpus.
- A corpus of sentences that represent more prototypical and less prototypical instantiations of the frames, tagged with pattern number and anomalous argument(s) (i.e. arguments that do not satisfy the typing constraints specified in the frame), if there is/are any.

Tools

- A balanced corpus of the language. ITWaC reduced (Baroni and Kilgarriff 2006).
- A methodology of corpus analysis. CPA, Corpus Pattern Analysis (Hanks and Pustejovsky 2005).
- A linguistic model of compositional mechanisms active in V-ARG selection (type coercion etc.). Pustejovsky 2006.
- Corpus tools. Sketch Engine (SkE, Kilgarriff, Rychly, Smrz, Tugwell 2004).
Choose a target verb and create a sample concordance.
Identify the relevant structure (typical syntagmatic patterns).
Assign a type expectation to each argument position in the pattern.
Assign each instance of the sample to one of the frames.
Store the frames (with the associated corpus instances) in the resource.
Associate each frame with at least one sense, expressed in the form of a “primary implicature” linked to the typing constrains specified in the pattern.

[[Human]] essere presente a [[Event]].
Unclassified sample from ITWaC for the verb *arrestare*

Elisabetta Jezek
Università di Pavia

Acquiring typed predicate-argument structures from corpora
Unclassified sample from ITWaC for the verb *arrestare* queried through SkE

### Sketch Engine

User: Jezek  Corpus: ITWaC (reduced)

<table>
<thead>
<tr>
<th>postV_N</th>
<th>preV_N</th>
<th>pp_per</th>
<th>pp_dal</th>
<th>pp_in</th>
</tr>
</thead>
<tbody>
<tr>
<td>pregudicato</td>
<td>polizia</td>
<td>furto</td>
<td>carabiniere</td>
<td>fiagranza</td>
</tr>
<tr>
<td>latitante</td>
<td>carabiniere</td>
<td>rapina</td>
<td>agente</td>
<td>seguito</td>
</tr>
<tr>
<td>scaglia</td>
<td>etiopia</td>
<td>detenzione</td>
<td>nazista</td>
<td>occasione</td>
</tr>
<tr>
<td>spacciatore</td>
<td>manifestante</td>
<td>favoreggiamento</td>
<td>tedesco</td>
<td>relazione</td>
</tr>
<tr>
<td>declino</td>
<td>brigatista</td>
<td>spaccio</td>
<td>cc</td>
<td>appartamento</td>
</tr>
<tr>
<td>emorragia</td>
<td>giornalista</td>
<td>omicidio</td>
<td>doganiere</td>
<td>precedenza</td>
</tr>
<tr>
<td>persona</td>
<td>poliziotta</td>
<td>residenza</td>
<td>soldato</td>
<td>esecuzione</td>
</tr>
<tr>
<td>ladro</td>
<td>terrorist</td>
<td>concussione</td>
<td>fascista</td>
<td>base</td>
</tr>
<tr>
<td>marocchino</td>
<td>romanista</td>
<td>reato</td>
<td>politif</td>
<td>blocco</td>
</tr>
<tr>
<td>dimostrante</td>
<td>marocchina</td>
<td>pedofilia</td>
<td>militare</td>
<td>fondo</td>
</tr>
<tr>
<td>terrorista</td>
<td>spacciatore</td>
<td>rissa</td>
<td>americano</td>
<td>attesa</td>
</tr>
<tr>
<td>rapinatore</td>
<td>no</td>
<td>truffa</td>
<td>regine</td>
<td>massa</td>
</tr>
<tr>
<td>sospetto</td>
<td>tifoso</td>
<td>estorsione</td>
<td>uomo</td>
<td>aula</td>
</tr>
<tr>
<td>traficante</td>
<td>pachistano</td>
<td>violenza</td>
<td>mezzo</td>
<td>operazione</td>
</tr>
<tr>
<td>corsa</td>
<td>dissidente</td>
<td>oltraggio</td>
<td>operazione</td>
<td>momento</td>
</tr>
<tr>
<td>colpevole</td>
<td>latitante</td>
<td>corruzione</td>
<td>modo</td>
<td>tempo</td>
</tr>
<tr>
<td>marcia</td>
<td>persona</td>
<td>vagabondaggio</td>
<td>passato</td>
<td></td>
</tr>
<tr>
<td>caduta</td>
<td>uomo</td>
<td>possesso</td>
<td>condizione</td>
<td></td>
</tr>
<tr>
<td>sospettato</td>
<td>minore</td>
<td>omissione</td>
<td></td>
<td></td>
</tr>
<tr>
<td>manifestante</td>
<td>scafista</td>
<td>danneggiamento</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Patterns 1 and 2 and associated senses for the verb *arrestare*

1. [[Human 1]] arrestare [[Human 2]]
   - [[Human 1]] seize [[Human 2]] by legal authority, [[Human 1]] take [[Human 2]] into custody

2. [[Entity]] arrestare [[Process]]
   - [[Entity]] cause [[Process]] to stop
Tagged contexts Pattern 1 and Pattern 2 *arrestare*

Elisabetta Jezek Università di Pavia

Acquiring typed predicate-argument structures from corpora
Type mismatches

- The paradigmatic sets of words that populate specific argument slots within the same verb sense do not map neatly onto the “expected” type (selected by V).
- Mismatches between “pattern” type (assigned by V) and “instance” type (inherent in N) within the same grammatical relation.
Mismatch classification

Mismatch classification can be classified according to several parameters.

- **Verb class**: (Levin 1993, VerbNet, ...): aspectual verbs, communication verbs, perception verbs, directed motion verbs, verbs of motion using a vehicle ...

- **Targeted grammatical relation**: SUBJ_OF, OBJ_OF, COMPL...

- **Shift type** (domain-preserving vs. domain-shifting mismatch): Artifact as Event, Artifact as Human, Artifact as Sound, Event as Location, Vehicle as Human ....

- **Elasticity/flexibility of noun classes**: (Artifacts vs. Naturals ...)(Lauwers and Willems 2011).
Assuming a qualia-based lexical representation for nouns, as in Generative Lexicon, mismatches may be further classified according to which quale/qualia is/are exploited or introduced in composition.

- **Formal** ($F$): encoding taxonomic information about the lexical item (the is-a relation);
- **Constitutive** ($C$): encoding information on the parts and constitution of an object (part-of or made-of relation);
- **Telic** ($T$): encoding information on purpose and function (the used-for or functions-as relation);
- **Agentive** ($A$): encoding information about the origin of the object (the created-by relation).
Besides the four standard roles, we assume that lexical representations may include values for the following relations (Pustejovsky and Jezek Forth.):

- **Natural Telic (NT)**: property that is necessarily associated with a natural kind (no intentionality).
  - riverNT = flow, heartNT = pump_blood...
- **Conventionalized Attribute (CA)**: property/activity routinely or systematically associated with an object, but not strictly part of the identified Qualia roles (projective manifestation of a specific Quale).
  - dogCA = bark, carCA = park, foodCA = digest ....
We provide a list of examples of mismatches classified according to the parameters introduced above: a) verb class, b) targeted grammatical relation (in color), c) type of shift (instance type as pattern type) and d) targeted Quale of the noun (both relation and value).

In the examples, the instances are being matched to the semantic types derived from a CPA study of these verbs.

This study was used as a base to build the Italian dataset for the SemEval-2010 shared task on coercion (Pustejovsky et al. 2010).
Aspectual verbs

[[Human]-subj] interrompe [[Event]-obj]

- Arriva Mirko e interrompe la conversazione. ‘Mirko arrives and interrupts the conversation’ (matching)
- Il presidente interrompe l’oratore. ‘The president interrupts the speaker’ (Human as Event; T=parlare ‘speak’)

Elisabetta Jezek Università di Pavia
[[Human]-subj] annuncia [[Event]-obj]

- Lo speaker annuncia la partenza. ‘The speaker announces the departure’ (matching)
- Il maggiordomo annuncia gli invitati. ‘The butler announces the guests’ (Human as Event, CA=arrivare ‘arrive’)
- L’altoparlante annunciava l’arrivo del treno. ‘The loudspeaker announces the arrival of the train’ (Artifact as Human; T=usare ‘use’(human, tool))
[[Human1]-subj] avvisa [[Human1]-subj] (di [[Event]-obj])

- I cacciatori hanno avvisato i carabinieri. (matching)
- Una telefonata anonima avvisa la polizia. ‘An anonymous telephone call alerted the police’ (Event as Human; AG=telefonare ‘phone’(human1, human2))
Avoid Verbs

[[Human]-subj] evita [[Event]-obj]

- Abbiamo evitato l’incontro. ‘We avoided the meeting’ (matching)
- Meglio evitare i cibi fritti. ‘It is best to avoid fried food’ (Artifact as Event; T=mangiare ‘eat’)

Elisabetta Jezek Università di Pavia

Acquiring typed predicate-argument structures from corpora
Forbid Verbs

[[Human]-subj] vieta [[Event]-obj]

Nell’Italia di allora la legge vietava l’aborto. ‘At that time in Italy law prohibited abortion’ (matching)

La Francia vieta il velo a scuola. ‘France bans the headscarf in schools’ (Artifact as Event; T=indossare ‘wear’)
Verbs of desire (Bos 2009)

[[Human]-subj] preferire [[Event]-obj]

- Preferisco bere piuttosto che mangiare. ‘I prefer drinking to eating’ (matching)
- Preferisco la birra al vino. ‘I prefer beer to wine’ (Artifact as Event; T=bere ‘drink’)
Perception verbs

[[Human]-subj] ascolta [[Sound]-obj]

- Rilassarsi ascoltando il rumore della pioggia. ‘Relax while listening to the sound of rain’ (matching)
- Ascoltava la radio con la cuffia. ‘He listened to the radio with his earphones’ (Artifact as Sound: T=produrre_suono ‘produce_sound’)
- Rimasi a lungo ad ascoltare il suo respiro. ‘I stayed for a long while listening to his breath’ (Event as Sound; NT=produrre_suono ‘produce_sound’)
- Non ho potuto ascoltare tutti i colleghi ‘I could not listen to all colleagues’ (Human as Sound; CA=parlare ‘speak’)

Elisabetta Jezek Università di Pavia | Acquiring typed predicate-argument structures from corpora
Directed motion verbs 1/3

[[Human]-subj] raggiunge [[Location]-obj]

- Abbiamo raggiunto l’isola alle 5. ‘We reached the island at 5’ (matching)
- Ho raggiunto il semaforo e ho svoltato a destra. ‘I reached the traffic light and turned right’ (Artifact as Location; CA= essere_a ‘be_at’(location))
Alla fine, ormai col buio, sono arrivata a una radura. ‘Finally in the dark I came upon a clearing.’ (matching)

Gli invitati arrivano al concerto in ritardo. ’The guests arrived late at the concert’ (Event as Location; CA=aver luogo_a ‘take place_at’(location))
Gli ospiti possono visitare la città senza usare l’auto. ‘The guests can visit the city without the car.’ (matching)

Per visitare la mostra ci vogliono due ore. ‘To visit the exhibition one needs two hours’. (Artifact as Location; CA=essere_a ‘be_at’(location))
Motion using a vehicle

[[Flying Vehicle]-subj] atterra ([Adv [Location]])

- **Il nostro aereo** atterra alle 21. ‘Our plane lands at 9pm’ (matching)
- **Il pilota** e’ regolarmente atterrato senza problemi. ‘The pilot landed regularly with no problems’ (Human as Vehicle; T=pilotare ‘pilot’(human, vehicle))
- **Tutti i voli civili** sono atterrati. ‘All civilian flights landed’ (Event as Vehicle; ArgStr Exploitation?)
Vehicle Verbs

[[Human]-subj] parcheggiare ([[Vehicle]-obj])

- Luca ha parcheggiato sotto casa. ‘Luca parked near the house’ (matching)
- L’ambulanza ha parcheggiato lontano. ‘The ambulance parked far away’ (Vehicle as Human; T=guidare ‘drive’(human, vehicle))
At present, we treat the entire NP as a markable.

“Regular choices of types within the same argument position are coded as type alternations. Common alternations in subject position are for instance [[Human|Institution]] and [[Human|Body Part]]”.

“Non-canonical lexical items breaking a particular statistical threshold are “coerced” into ‘honorary’ membership of a semantic type in particular contexts”.

Honorary members are tagged as “a” = anomalous arguments.
Contexts for Pattern1 *arrestare*

[[Human1 | Institution]-subj] arresta [[Human2]-obj]
**Contexts with anomalous arguments for Pattern1 atterrare**

[[Flying Vehicle]] atterra (Adv [Location]])

<table>
<thead>
<tr>
<th>Corpus: itWAC (reduced)</th>
<th>Hits: 4892 (5.2 per million) sampled to 250</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotating: atterrare</td>
<td>Info Sort Finish New pattern: Add Number globally: X</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>Context</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>70190621</td>
<td>cercavo tutt’ altro e mi sembra di essere attirato 1.a su uno strano pianeta: gente che raccoglie</td>
<td></td>
</tr>
<tr>
<td>68500741</td>
<td>BALDAZZI e dal Colonnello Roberto ROCCHI, è attirato 1.a con un aereo C - 130 J all’aeroporto di</td>
<td></td>
</tr>
<tr>
<td>103748746</td>
<td>operatori della sala radar di Linate) sono attirato 1.a all’aeroporto di San damiano (piazz</td>
<td></td>
</tr>
<tr>
<td>103749868</td>
<td>sicurezza. Per quasi due anni sono sempre attirato 1.a a Linate … e questo inconveniente non</td>
<td></td>
</tr>
<tr>
<td>103752373</td>
<td>alla minima e fai il circuito a destra e attirere 1.a (anche di notte) Per il decollo per la</td>
<td></td>
</tr>
<tr>
<td>129081889</td>
<td>privato e penso che forse sarebbe stato meglio atterrare 1.e e attendere che passasse tutto questo temporaccio</td>
<td></td>
</tr>
<tr>
<td>145305030</td>
<td>10 settembre 2004 alle 7:25:47 dipende dove attirere 1.a e che genere di vacanza vuoi? se vuoi</td>
<td></td>
</tr>
<tr>
<td>146773219</td>
<td>regolarità durante tutto l’anno. Naturalmente attirete 1.a e scenderete a Manaus, in cui sosterete</td>
<td></td>
</tr>
<tr>
<td>155929206</td>
<td>mongolfiera ad elio , con la quale riusci ad attirare 1.a davanti alla palazzina reale di caccia</td>
<td></td>
</tr>
<tr>
<td>160931401</td>
<td>poi non ho capito , teri pomeriggio sono attirato 1.a a Malpensa e non ho trovato nessuno</td>
<td></td>
</tr>
<tr>
<td>206393108</td>
<td>non si era svegliato nemmeno quand’erano attirati 1.a con qualche scossone , e ciò le feceva</td>
<td></td>
</tr>
<tr>
<td>215314872</td>
<td>costa azzurra … che dall’ aereo quando attirerei 1.a a nizza vedì tutti questi occhi celesti</td>
<td></td>
</tr>
<tr>
<td>219649517</td>
<td>Oriente e l’ Europa: quasi tutti i voli attirano 1.a all’aeroporto Mehrabad di Teheran. La</td>
<td></td>
</tr>
<tr>
<td>232621527</td>
<td>emergenza del genere. Gli aiuti italiani sono attirati 1.a all’aeroporto di Colombo domenica mattina</td>
<td></td>
</tr>
<tr>
<td>243708623</td>
<td>di Zugdidi. Il luogo in Marocco dove è attirare 1.a ‘ il grosso asteroide che ha distrutto</td>
<td></td>
</tr>
<tr>
<td>254133888</td>
<td>a Livorno - Il Messaggero - James Forrest attirerà 1.a in Italia con un giorno di ritardo .</td>
<td></td>
</tr>
</tbody>
</table>
Ongoing work focuses on improving the annotation of corpus instances in regard to two areas:

- annotating instance types,
- annotating the targeted quale/qualia in V-ARG composition.
Based on Pustejovsky et al. 2008, 2010 (SemEval Coercion Task) and previous attempts to annotate metonymic relations in text (Markert and Nissim 2007), in Jezek and Frontini 2010 we finalized a scheme to annotate type mismatches in the resource.
Three layers of semantic annotation:

- the Pattern Type, which records the semantic type that is inherited by the pattern for each argument position;
- the Argument Filler, which contains the lexical material that actually instantiates the semantic position in the instance;
- the Instance Type, which needs to be added when the semantic argument filler instantiates a type that does not match with the Pattern Type, otherwise it is inherited from the Pattern.
I ragazzi hanno bevuto una pinta insieme.
’the boys drank a pint together’

- [[Human]-subj] beve [[Liquid]-obj]
- <instance tid=102> <argument id=a1 pattern_id=p15 instance_sem_type=HUMAN instance_syn_role=subj> I ragazzi </argument> <verb pattern_id=p15> hanno bevuto </verb> <argument id=a2 pattern_id=p15 instance_sem_type=MEASURE_UNIT instance_syn_role=obj> una pinta </argument> insieme. </instance>
In Jezek, Quochi and Calzolari 2009 and Jezek and Quochi 2010 we explored how to integrate qualia specification in the coercion annotation task, in addition to type specification. This may be attained in two ways:

- as “online” specification during the annotation;
- retrieving it from a pre-existing resource (e.g. SIMPLE, QS gold standard, noun-frame repository ....).
Conclusions

- Create a repository of corpus-based typed predicate-argument structure for Italian verbs, intended as *semantically motivated syntagmatic distinctions*.
- Compile a corpus-based inventory of metonymic shifts as a by-product.
- Build an ontology for NOUNS, grouped into clusters according to their argument roles and their ability to predict the interpretation of V in context.
Future work

- Develop a classification of VERBS based on the argument typing (orthogonal to Levin’s and VerbNet classes).
- \([\text{[Human]-subj]}_\text{-}[\text{Location}-\text{obj}]: \text{raggiungere, attraversare, costeggiare}\) (motion verbs), but also \text{invadere, bombardare, conquistare, visitare} \ldots
- \([\text{[Human]-subj]}_\text{-}[\text{Event}-\text{obj}]: \text{finire, iniziare, interrompere}\) (aspectual verbs), but also \text{organizzare, vietare, proibire} \ldots
- Class members are not necessarily synonyms or antonyms.
I thank James Pustejovsky, Alessandro Lenci, Francesca Frontini and Valeria Quochi for coercion-related discussion, and Patrick Hanks for his scientific support for the Italian CPA project.

Thanks also to the participants of the “Semantic and pragmatic annotation of corpora” Workshop, held at Heinrich-Heine-Universität Düsseldorf, July 16, 2012 for their fruitful comments on the annotation scheme.
References

References


References

