Acquiring typed predicate-argument structures from corpora

Elisabetta Jezek
Università di Pavia
Dipartimento di Studi Umanistici
Sezione di Linguistica
jezek@unipv.it

Abstract

In this note, I 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 for linguistic research and NLP applications. The note is organized as follows. I first introduce the resource, then focus on the annotation of type mismatches between argument fillers and verb selectional requirements and their linguistic classification, based on Generative Lexicon Theory (Pustejovsky et al. 2008). Finally, I outline the ongoing attempt to combine typing annotation with standard coarse-grained (high-level) thematic role information (Bonial et al. 2011a, 2011b) carried out in collaboration with Senso Comune (Vetere et al. 2011). A discussion of ongoing evaluation and improvements follows.

1 The resource

Typed predicate-argument structures are corpus-derived verb frames[1] with the specification of the expected semantic type for each argument position (e.g. [[Human]] mangia [[Food]], [Human]] guida [[Vehicle]], [[Human]] partecipa a [[Event]]), populated by lexical sets (Hanks 1986), i.e. the statistically relevant list of collocates that typically fill each position (e.g. [[Event]-iobj of partecipare] = {gara, riunione, selezione, manifestazione, seduta, cerimonia, conferenza, votazione ...}). The repository of corpus-based patterns for Italian verbs is a manually annotated resource under development at the University of Pavia in collaboration with the Faculty of Informatics at Masaryk University (Brno) and FBK (Trento). It currently consists of a nucleus of about 300 lexical units (verbs). In the resource, each lexical unit is linked to a set of frames extracted from the corpus following the Corpus Pattern Analysis technique (CPA, Hanks - Pustejovsky 2005). Each frame is associated with a corpus-derived verb sense (expressed in the form of an implicature linked to the typing constraints) and with a set of corpus instances (a sample of 250 occurrences for each verb), that represent more prototypical and less prototypical instantiations of the frame. Each corpus instance is tagged with information about pattern number and anomalous arguments, i.e. arguments that do not satisfy the typing constrains specified in the frame. At present, in compiling the patterns, we are using a list of shallow semantic types ([[Human]], [[Artifact]] etc.) borrowed from the English project (Pattern Dictionary of English Verbs (PDEV), project page at http://deb.fi.muni.cz/pdev/). The reference corpus for the Italian project is a reduced version of itWaC (Baroni & Kilgarriff 2006). We plan to make the resource available once the goal of analyzing 1000 "verbi a polisemia media" (average polysemous verbs) is reached.

[1] By verb frame we mean the relational semantic structure associated with the verb, specifying information about the linguistically relevant participants in the event encoded by the predicate.

2 Type mismatches

In acquiring the patterns from corpus data, the annotator retrieves the corpus instances, identifies the relevant structure, analyses the lexical set for each grammatical relation and associates a typing assignment to each argument position in the pattern.[2] Once the pattern is identified, each corpus instance is tagged with the associated pattern number. One recurrent problem that arises in this phase is the identification of mismatches between pattern type (assigned by the verb) and instance type (inherent in the argument filler) within the same grammatical relation.

3 Mismatch classification

Mismatches may be classified according to the following 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): Artifact as Event, Artifact as Human, Artifact as Sound, Event as Location, Vehicle as Human ...
- *Elasticity/flexibility of noun class*: 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. Besides the four standard roles, i.e.

- *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).

we may assume that lexical representations include values for the following relations (Pustejovsky and Jezek Forth.):

- *Natural Telic* (NT): property that is necessarily associated with a natural kind (no intentionality). For example: *river*NT=$flow$, *heart*NT=$pump\_blood$.
- *Conventionalized Attribute* (CA): property/activity routinely or systematically associated with an object, but not strictly part of the identified Qualia roles. For example: *dog*CA=$bark$, *car*CA=$park$, *food*CA=$digest$.

3.1 Data

What follows is a list of examples of mismatches classified according to the parameters introduced above: a) verb class, b) targeted grammatical relation (in italics), 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.[3]

(1) *Aspectual Verbs*

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')

(2) *Communication Verbs*

Lo speaker annuncia *la partenza*. 'The speaker announces the departure' (matching)
Il maggiordomo annuncia *gli invitati*. 'The butler announces the guests' (Human as Event,

[2] Types are conceived as abstractions over the lexical sets found in the argument slots in the corpus.

[3] This study was used as a base to build the dataset for the SemEval-2010 shared task on coercion (see below).

CA=arrivare 'arrive')[4]
L'altoparlante annunciava l'arrivo del treno. 'The loudspeaker announces the arrival of the train' (Artifact as Human; T=usare 'use'(human, tool))
Una telefonata anonima avvisa la polizia. 'An anonymous telephone call alerted the police' (Event as Human; AG=telefonare 'phone'(human1, human2))

(3) *Avoid Verbs*

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')

(4) *Forbid Verbs*

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')

(5) *Verbs of desire (Bos 2009)*

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')

(6) *Perception verbs*

Rilassarsi ascoltando *il rumore della pioggia*. 'Relax while listening to the sound of rain' (matching)
Ascoltava *la radiolina* 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')

(7) *Directed motion verbs*

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))
Gli invitati arrivano *al concerto* in ritardo. 'The guests arrive late at the concert' (Event as Location; CA=aver luogo_a 'take place_at'(location))

(8) *Motion using a vehicle*

Il nostro aereo atterra alle 21. 'Our plane lands at 9pm' (matching)
Il pilota è 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?)

(9) *Vehicle Verbs*

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))

4 Mismatch tagging

At present, we treat the entire NP as a markable. Following the CPA procedure, 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]], for example: [[Human|Body Part]] sanguina 'bleeds'. "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.

[4] As noted by one reviewer, this example may be analyzed as an instance of a different sense of *annunciare* with different constraints. We propose instead that the sense is one and the same, and that the interpretation of the specific combination is achieved by exploiting one of the events conventionally associated with the noun.

5 Improving coercion annotation

Ongoing work focuses on improving the annotation of corpus instances in regard to three areas:

- annotating instance types,
- annotating the targeted quale/qualia in V-ARG composition,
- interfacing typing and semantic role annotation.

Each of these points is examined below.

5.1 Annotating instance types

Based on Pustejovsky et al 2008, 2010 (SemEval Coercion Task) and previous attemps 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. The scheme foresees 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 instantiates the semantic position in the instance;
- the Instance Type, which needs to be added when the argument filler instantiates a type that does not match with the Pattern Type, otherwise it is inherited from the pattern.

The following is an example:

(10) 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>

5.2 Annotating the targeted quale in V-ARG composition

In Jezek, Quochi, Calzolari 2009 and Jezek and Quochi 2010 we explored how to integrate qualia specification (relation and/or value) 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 ....).

5.3 Interfacing types with semantic role annotation

In the resource, typing information is sometimes complemented with fine-grained semantic roles. In principle, the semantic type captures the Formal quale of the argument, which is an intrinsic property of nouns normally found in that argument slot (e.g. person, substance, artefact etc.). On the other hand, the semantic role captures an extrinsic property of the nouns in the same slot, namely one that specifies how the referent is involved in the event (e.g. as an intentional agent, an affected entity, a created entity, and so forth). This is illustrated below:

(11) [[Human 1 = Legal Authority]] arresta 'arrest' [[Human 2 = Suspect]]

Ongoing work focuses on improving role annotation with systematic coarse-grained roles annotation. In the context of the Senso Comune initiative (www.sensocomune.it), we designed a set of 27 coarse-grained (high-level) semantic roles based on VerbNet (VN) and LIRICS (Petukhova and Bunt 2008) and the on-going attempt to create a unified standard set for the International Standard Initiative (ISO) (Bonial et al. 2011a, b).[5] We conflated some LIRICS roles (e.g., Medium and Instrument), adopted some suggestions from Bonial et al. 2011a (e.g., the use of co-Agent and co-Patient rather than the unique Partner), and used some classical semantic roles like Experiencer rather than LIRICSs ambiguous Pivot. We adopted the hierarchy in Bonial

[5] Besides the author of this note, the group working at role annotation in Senso Comune includes Fabio Massimo Zanzotto, Laure Vieu, Guido Vetere, and Alessandro Oltramari.

et al. 2011b, but distinguished between *participants* and *context*.

We performed a pilot experiment on 400 usage examples (about 6% of the entire corpus) associated with the sense definitions of 25 fundamental verb lemmas of the Senso Comune resource to release the beta-version of the annotation scheme.

The annotation task involves tagging the usage instances with syntactic and semantic information about the participants in the frame realized by the instances, including argument/adjunct distinction. In semantic annotation, annotators are asked to attach a semantic role and an ontological category to each participant and to annotate the sense definition associated with the filler. We provide them with the hierarchical taxonomy of roles based on Bonial 2011b, together with definitions and examples for each role. The TMEO methodology (cf. Vetere et al. 2011) is used to help them selecting the ontological category in Senso Comune's top-level. For noun sense tagging, the annotator exploits the senses already available in the Senso Comune resource. Drawing on the results of previous experiments on "ontologization" of noun senses (Chiari et al. 2011), we allow multiple classification, that is, we allow the annotators to tag each slot with more than one semantic role, ontological category and sense definition. For example in the context in (12), the subject may be tagged with both Agent and Experiencer, if the annotator assumes that the participant shares entailments which belong to both roles.

(12) [I turisti AG EXP / Human] ammirano i quadri.
'The tourists admire the paintings'

The pilot experiment confirms our expectation that in category assignment, annotators are influenced by the inherent semantic properties of the referents filling the argument positions. For example, in (13) they annotate the referent of the object argument as Human, even though it is metonymically reinterpreted as Document in the context of *leggere* 'read'. Interestingly, the inherent semantic properties of the argument's referents appear to play a role also in semantic role assignment. For example, in the coercive environment in (13), the annotator hesitates whether he/she should annotate the mismatch in object position also at the role level, i.e. assigning Source instead of Theme (the latter is the role chosen for such contexts as *leggere una lettera, il giornale* 'read a letter, the newspaper' and so forth).

(13) leggere [un autore ?SOURCE / Human]
'read an author'

This appears to hold true also when annotations of role and ontological category are performed as separate sub-tasks. That is, if annotators are asked to annotate the semantic role only (besides grammatical relations), semantic role assignment still appears to be performed (also) on the basis of the perceived inherent category of the argument filler. We are currently exploring how to approach this issue (both in theory and in annotation practice), that appears to involve several classes of phenomena, including Instruments (and other kinds of Artifacts) in Subject position, as in (2) and (9) above.

6 Conclusions

In this note, I described the effort of creating a repository of corpus-based patterns for Italian verbs for purposes of linguistic research and NLP application. This involves creating a corpus-based inventory of metonymic shifts as a by-product. Ongoing work focuses on improving mismatch annotation and on examining the interplay between typing and role constraints to argument selection, focusing on coercive environments.

Acknowledgments

I would like to 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. I also acknowledge Guido Vetere, Laure Vieu, Fabio Zanzotto and Alessandro Oltramari for the ongoing collaborative work on semantic role annotation within the Senso Comune initiative. Thanks also to the participants of the "Semantic and pragmatic annotation of corpora" Workshop, held at Heinrich-Heine-Universität Düsseldorf, July 16, 2012, and to four anonymous reviewers for their fruitful comments.

7 References

Baroni, M. and A. Kilgarriff 2006. Large Linguistically-Processed Web Corpora for Multiple

Languages. In *EACL 2006 Proceedings*, 87-90.

Bonial, C., S.W. Brown, W. Corvey, V. Petukhova, Palmer M., Bunt H. 2011a. An Exploratory Comparison of Thematic Roles in VerbNet and LIRICS. In *Proceedings of the Sixth Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*.

Bonial, C., W. Corvey, Palmer M., V. Petukhova, Bunt H. 2011b. A hierarchical unification of LIRICS and VerbNet semantic roles. In *Proceedings of the 2011 IEEE Fifth International Conference on Semantic Computing*, IEEE Computer Society Washington, DC, USA, 483-489.

Bos, J. 2009. Type Coercion in the Contexts of Desire. In P. Bouillon et al. (eds) *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon*, Pisa, ILC-CNR, Sept. 17-19, 2009.

Hanks, P. 1986. Contextual Dependencies and Lexical Sets. In *International Journal of Corpus Linguistics* 1:1, 7598.

Hanks, P. and J. Pustejovsky 2005. A Pattern Dictionary for Natural Language Processing. In *Revue franaise de linguistique appliquée*, 10 (2), 63-82.

Chiari, I. Oltramari, A. Vetere, G. 2011. Di cosa parliamo quando parliamo fondamentale? In S. Ferreri (ed.) *Atti del Convegno della Società di linguistica italiana*, Roma, Bulzoni, 221-236.

Jezek E., V. Quochi and N. Calzolari 2009. Relevance of Qualia Relations in Coercive Contexts. In P. Bouillon et al. (eds) *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon*, Pisa, ILC-CNR, Sept. 17-19, 2009, 128-136.

Jezek E. and F. Frontini 2010. From Pattern Dictionary to PatternBank. In G.M. de Schryver (ed) *A Way with Words: Recent Advances in Lexical Theory and Analysis*, Kampala, Menha Publishers, 215-239.

Jezek E. and V. Quochi 2010. Capturing Coercions in Texts: a First Annotation Exercise. In Nicoletta Calzolari et al. (eds) *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta (May 19-21, 2010), Paris: European Language Resources Association (ELRA), 1464-1471.

Lauwers, P. and D. Willems 2011. Coercion: Definition and Challenges, current approaches and new trends. In *Linguistics* 49:6, 1219-1235.

Markert K. and M. Nissim. 2007. SemEval-2007 task 8: Metonymy resolution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations*, Prague, Czech Republic. Association for Computational Linguistics.

Petukhova,V., Bunt H. 2008. LIRICS semantic role annotation: Design and evaluation of a set of data categories. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, 28–30.

Pustejovsky, J., Rumshisky, A., Moszkowicz, J.L., Batiukova, O. 2008. GLML: A Generative Lexicon Markup Language. ms presented at GL workshop, Pisa, Istituto di Linguistica Computazionale (CNR), Sept. 2008.

Pustejovsky J., Rumshisky A., Plotnick A., Jezek E. Batiukova O., Quochi V. 2010. SemEval Task 7: Argument Selection and Coercion. In Proceedings of the Fifth International Workshop on Semantic Evaluation Uppsala University, Sweden, July 1116 2010.

Pustejovsky, J. and E. Jezek (Forth.). *Generative Lexicon Theory: A Guide*, Oxford, Oxford University Press.

Vetere, G., Oltramari, A., Chiari I., Jezek E. Vieu L., Zanzotto F. 2011. Senso Comune: An Open Knowledge Base for Italian. In *Traitement Automatique des Langues* (TAL), 52:3.