Capturing Coercions in Texts: a First Annotation Exercise

Elisabetta Jezek[1], Valeria Quochi[2]
[1]Department of Theoretical and Applied Linguistics, University of Pavia, Pavia, Italy
[2]Istituto di Linguistica Computazionale, CNR, Pisa, Italy
[1]jezek@unipv.it, [2]valeria.quochi@ilc.cnr.it

Abstract

In this paper we report the first results of an annotation exercise of argument coercion phenomena performed on Italian texts. Our corpus consists of ca 4000 sentences from the PAROLE *sottoinsieme* corpus (Bindi et al. 2000) annotated with Selection and Coercion relations among verb-noun pairs formatted in XML according to the Generative Lexicon Mark-up Language (GLML) format (Pustejovsky et al., 2008). For the purposes of coercion annotation, we selected 26 Italian verbs that impose semantic typing on their arguments in either Subject, Direct Object or Complement position. Every sentence of the corpus contains information about corpus-derived typed selectional preferences for verbs in the targeted argument slots and is annotated with the source type for the noun arguments by two annotators plus a judge. An overall agreement of 0.87 kappa indicates that the annotation methodology is reliable. A qualitative analysis of the results allows us to outline some suggestions for improvement of the task: 1) a different account of inherently polysemous nouns has to be devised and 2) a more comprehensive account of coercion mechanisms requires annotation of the deeper meaning dimensions that are targeted in coercion operations, such as those captured by Qualia relations.

1. Introduction

Automatic recognition and resolution of metonymies in natural language texts (e.g. author for work, organization for members, place for people etc.) has attracted considerable interest within the NLP community and nowadays is recognized as an important complement to and extension of WSD (Markert and Nissim, 2002 and 2009). In this paper we report the first results of an annotation exercise of argument coercion phenomena in Italian text, conducted within the context of a large-scale project aiming at annotating compositional operations based on Generative Lexicon theory (Pustejovsky et al., 2008). In this framework, argument coercion is understood as the operation of type adjustment induced by a predicate over its arguments when they do not match its selectional requirements (Pustejovsky, 1995; Copestake and Briscoe, 1995 inter alia). For example, in the context "he left the concert early", the type expected in object position by the verb *leave* is LOCATION, but the surface type of the argument filler (*concert*) is EVENT. Therefore, it is assumed that in this context, the coercion EVENT→LOCATION has occurred.

The motivation of our work is twofold. First, we intend to provide a reliable and carefully controlled corpus annotated for coercions to be used as training and test set for computational semanticists aiming at developing algorithms for metonymy recognition and/or processing figurative language. Second, we are interested in using the annotation results for improving the GL-based annotation framework, with the overall goal of creating an annotated resource to be used not only for NLP applications but also for linguistic theoretical studies of semantic compositional mechanisms in language.

The paper is organized as follows: in section 2 we introduce the methodology proposed within the Generative Lexicon Markup Language (GLML) project for the annotation of coercion phenomena. In section 3 we describe the methodology we actually followed in building the Italian dataset for our first exercise of coercion annotation. In section 4, we illustrate the preliminary results of the annotation and in section 5 we outline some proposals for the improvement of the methodology of the task, based on the annotation results.

2. A Generative Lexicon annotation of coercion mechanisms

An effort has been made at Brandeis University to "translate" (part of) the theoretical apparatus of the Generative Lexicon theory (Pustejovsky, 1995) into the Generative Lexicon Markup Language annotation framework (Pustejovsky et al., 2008), which tries to establish not only a mark-up language, but also an annotation methodology for compositional operations in natural language text. There are currently four main annotation tasks that are part of the GLML project:

1. Compositional mechanisms of argument selection;
2. Qualia in argument selection;
3. Qualia in modification constructions;
4. Type selection in modification of dot objects.

Here we introduce the methodology foreseen for task 1, which involves identifying whether the compositional operation between a verb and the argument it selects can be characterized as SELECTION or COERCION (Pustejovsky et al., 2009).

Briefly, the GLML methodology proposed for task 1 involves two phases: the construction of the data set to be annotated and the actual human annotation. The data set construction phase consists of four steps: 1) selecting the set of target verbs, 2) compiling a sense inventory for each target, 3) associating a type template with each sense[1] and

[1] A type template is understood as a corpus-derived argument structure with specification of the expected semantic type for the argument fillers (e.g. for *finish*: HUMAN finish EVENT). In GLML-English, type templates are built in the way similar to the context patterns as defined in Corpus Pattern Analysis (CPA)

4) extracting the data containing the selected target verbs from a corpus. The data set construction also assumes a pre-existing shallow type system. In the original proposal, the following list, drawn from the Brandeis Shallow Ontology (BSO) (Pustejovsky et al. 2006), was given:

(1) HUMAN, ANIMATE, PHYSICAL OBJECT, ARTIFACT, ORGANIZATION, EVENT, PROPOSITION, INFORMATION, SENSATION, LOCATION, TIME PERIOD, ABSTRACT ENTITY, ATTITUDE, EMOTION, PROPERTY, OBLIGATION, RULE.

The annotation is organized into three main steps: a) sense disambiguation of the verb in each context/sentence; b) identification of a possible mismatch between the usual type associated with the noun (*source* type) and the type required by the verb (*target* type) c) specification of the source type in case of mismatch. As the task is conceived now, cases where the noun type satisfies the verb selectional requirements are automatically annotated as instances of SELECTION, whereas cases where the noun does not are annotated as COERCIONS. The output of the annotation provides noun source and target types and the compositional operation at play (SELECION or COERCION).

3. The Data Set Construction Phase

3.1 Coercion types and verb selection

In constructing the Italian data set for our annotation exercise we adopted a slightly modified version of the methodology proposed in Pustejovsky et al. 2009 and Pustejovsky and Rumshisky 2009 for English. We first examined previous corpus-informed theoretical studies of argument coercion (Jezek and Lenci, 2007, Pustejovsky and Jezek, 2008 a.o.) and looked in the Pattern Dictionary of Italian Verbs (PDIV, Hanks and Jezek, 2007)[2] with the aim of drawing a preliminary list of verbs that impose semantic typing on their arguments in either Subject (Subj), Direct Object (DObj) or Complement (Comp) position[3].

We started by choosing verbs that instantiate coercions between the types of the original GLML inventory. For example, according to our sources, *arrivare* 'arrive' selects for a LOCATION and instantiates the EVENT→LOCATION shift in its 'reach' sense in Complement position (e.g. "arrivare alla cerimonia" 'arrive at the ceremony'); *finire* 'finish' selects for an EVENT and instantiates the ARTIFACT→EVENT shift in its 'bring to an end' sense in Direct Object position (e.g. "finire il panino" 'finish the sandwich') and so on. Then, in order to obtain a sufficient and varied range of coercion types, we also considered verbs that select for types that were not included in the original type list (i.e. LIQUID, SOUND, DOCUMENT and VEHICLE – the last two classified as subtypes of the superordinate ARTIFACT).

For each target verb, we narrowed our focus on its most coercive sense (generally corresponding to its basic sense) and on the most frequent type shifting(s) that it instantiates within that sense[4]. We drew the relevant sense definition together with the related type template and a few examples of SELECTION and COERCION from PDIV.[5] We mapped sense definitions onto their SIMPLE equivalents, whenever possible (Lenci et al., 2000).

For each coercion type, we first selected a *seed* verb, i.e. a verb that provides good examples for that shift. Then, we looked for more verbs that impose the same type shift on their arguments, in order to ensure that the number of corpus instances aimed at per coercion type would be covered and that a variety of verbs and nouns would be represented in the annotated corpus.[6] The additional verbs may be hyponyms, synonyms, antonyms of the seed verb or they may share with it only the characteristic of selecting the same semantic type for the same argument slot.

Finally, we uploaded the prepared data (verbs, coercive senses, type templates associated with the senses, examples of selection and coercion for each sense) on the GLML-Italian wiki (http://glml-italian.wikidot.com/), to make it available for further steps.

In table 1 below we report the list of target verbs together with the most significant coercion types that they instantiate in the chosen sense according to our

(Pustejovsky et al., 2004).

[2] PDIV is a pre-existing repository of corpus-derived verb patterns (or type templates) for Italian verbs, developed as an extension to the English Pattern Dictionary project reported in Hanks and Pustejovsky 2005. It is being built according to the Corpus Pattern Analysis (CPA) technique and specifies corpus-derived typed selectional preferences for each argument slot associated with a verb in a given sense. Briefly, in CPA each verb is analyzed according to the following procedure: first, a sample concordance for each target verb is created (250 hits); second, the semantic types of the argument fillers are examined and the typical syntagmatic patterns of the verb are identified (e.g. for *leggere* 'read': HUMAN legge DOCUMENT); third, each line of the sample is assigned to one of the drafted patterns; fourth, both the patterns and the associated concordances are stored in the pattern repository. The corpus used for the identification of the verb patterns in PDIV is the Italian Web as Corpus (ItWaC, Baroni and Kilgarriff, 2006).

[3] With respect to the Complement position, we excluded sentential complements for our present purposes.

[4] An exception was made for *leggere* 'read' (analysed in detail in Jezek & Lenci, 2007) for which two senses were included (see table 2 in the Appendix).

[5] The CPA technique used for the identification of verb patterns and senses in PDIV was not originally conceived to encode coercions systematically. For example, regular choices of types within an overall pattern in relation to a target verb (e.g. l'aereo | il pilota | il turista | il volo *è atterrato* 'the plane | the pilot | the turist | the flight landed') are generally specified as type alternations in CPA, although in some cases they could be dealt with in terms of coercions from a basic type. Therefore, a number of adjustments are needed in order to use the patterns and senses stored in PDIV for the purposes of the coercion annotation task.

[6] As byproduct of this annotation project we aimed at producing a dataset for the SemEval-2010 Task 7: Argument Selection and Coercion. Therefore, we wanted to have at least 60 coercions per coercion type, out of 400 overall sentences.

preliminary investigation.[7]

Argument is (source type)	**Verb selects (target type)**	**Verb and targeted Grammatical Relation**
Location	Human	***contattare*** (Subj,DObj) *accusare* (Subj,DObj) *annunciare* (Subj,DObj) *avvisare* (Subj) *informare* (Subj) *organizzare* (Subj)
Organization	Human	***contattare*** (Subj,DObj) *accusare* (Subj,DObj) *annunciare* (Subj,DObj) *avvisare* (Subj) *informare* (Subj) *organizzare* (Subj)
Artifact	Human	***avvisare*** (Subj) *accusare* (Subj) *annunciare* (Subj) *chiamare* (Obj)
Vehicle	Human	***parcheggiare*** (Subj) *guidare* (Subj) *chiamare* (DObj)
Event	Location	***arrivare*** (Comp) *recarsi* (Comp) *raggiungere* (DObj) *visitare* (DObj)
Human	Vehicle	***atterrare*** (Subj) *sbandare* (Subj)
Organization	Vehicle	***guidare*** (DObj) *parcheggiare* (DObj)
Human	Document	***leggere*** (DObj) *divorare* (DObj)
Event	Document	***leggere*** (DObj) *divorare* (DObj)
Artifact	Event	***finire*** (DObj) *cominciare* (DObj) *continuare* (DObj) *interrompere* (DObj)
Document	Event	***finire*** (DObj) *cominciare* (DObj) *continuare* (DObj) *interrompere* (DObj)
Artifact	Sound	***ascoltare*** (DObj) *sentire* (DObj) *udire* (DObj) *echeggiare* (Subj) *rimbombare* (Subj)
Event	Sound	***ascoltare*** (DObj) *sentire* (DObj) *udire* (DObj) *echeggiare* (Subj) *rimbombare* (Subj)
Container	Liquid	**bere** (DObj) versare (DObj) sorseggiare (DObj)

Table 1: Selected verbs and targeted coercion types

3.2 Skimming of the instances to be annotated

In the next step, all sentences containing the selected verbs were extracted from the *PAROLE sottoinsieme corpus* (Bindi et al., 2000) and parsed at the functional level so as to identify all contexts with the relevant argument slots for each target verb. A group of volunteer students in linguistics (either MA or PhD students) (that here we will call selectors) worked on parsed data in order to correct misparses and select the appropriate contexts to be subsequently annotated[8].

Data was provided in the form of tables containing the following information: target verb, argument noun, grammatical relation for the given argument, and the full sentence. For each instance (i.e. sentence) only one grammatical relation was "focused" at a time (= 1 instance 1 noun). The selector needed to do three things: 1) identify which contexts to select (see below); 2) verify if the chosen context instantiated the sense(s) of the verb specified on the wiki; 3) for each chosen context, given the template specified for the verb on the wiki, identify whether it was a case of Selection or Coercion[9].

For each verb we wanted to select the highest possible number of instances with the same coercion type, i.e. same source type for the argument noun (max. 60 sentences) and at least twice the number of instances of selection. This to ensure that the final dataset would contain a sufficient number of coercions and that, even if not representative of the real distribution, the corpus would contain proportionally more selections than coercions. In order to obtain an approximate indication of how frequent a coercion type is, we annotated the first 500 "good" (i.e. correct) sentences. We then went over the file in search of more instances of the targeted/frequent coercion types.

In selecting the instances, we excluded sentences that were either too short or too long (unless the coercion is very interesting or unique); we tended to exclude instances representing a sense that was not present in the given sense inventory; we excluded sentences containing too complex anaphora (e.g. anaphora where the

[7] The chosen senses together with their definition are reported in Table 2 in the Appendix. Some of the verbs and coercion types identified in the first phase were not uploaded in the annotation tool or were excluded from the final data set, for reasons explained below (section 4). *Seed* verbs are marked in bold in table 1.

[8] Since the same group of people helped us in both selecting and annotating the dataset, in distributing the context to annotate we paid attention that the same person would not receive contexts that he/she helped selecting in the first phase.

[9] In order to perform the operation in 3 the selector was asked to identify the source type of the argument-noun using the revised GLML type taxonomy as a reference.

antecedent is very far in the context); we may include cases where the argument filler was a multiword expression (e.g. *ascoltare una colonna sonora* 'listen to a sound track'); if the multiword was exocentric we manually tokenized it by substituting the proposed name (e.g. *colonna* with *colonna_sonora*), otherwise we left the head of the multiword as the token to be later annotated; we included multi-selections (i.e. cases where more than one noun is selected by the same verb in the same context); finally, we included instances with the same verb-noun pair, but preferably only if the target pair was frequent in the corpus.

At this stage of pre-selection of contexts for annotation we were not particularly concerned with high precision (i.e that the coercions were actually the "correct" ones), as actual annotation would provide us with "correct" annotations and statistical significance. Therefore, selectors may include dubious cases for later assessment. For the same reasons, at this stage sentences for each verb were chosen by one single human selector.

4. The Annotation Phase

Annotation was performed by two independent annotators plus a third one acting as a judge. Annotators are volunteer students in linguistics (either Master or PhD students), not necessarily familiar with GL, whereas the two principal investigators plus a PostDoc in Linguistics acted as judges.

The dataset of 3885 sentences selected from the PAROLE *sottinsieme* corpus as described in Section 3 above, was split and reformatted according to the DB requirements and uploaded in the annotation tool.

4.1 The annotation environment

For performing the proper annotation, we adapted a first prototype annotation tool developed for English at Brandeis.

The tool is deployed over the web and has an interface that allows for a user-friendly annotation procedure, possibly usable also by not highly expert annotators. The annotator interface is written in Php and works on a MySQL database[10]. Since the tool is still a prototype, in the current state, the access is password protected and granted only to annotators, since different user access rights have not yet been implemented. However, we plan to improve and optimize the tool and to make it available to the community, if interest arises.

[10] http://wiki.ilc.cnr.it/glml/task1/

Figure 1: Interface for the Italian GLML Annotation Task 1

Reflecting the original annotation methodology as described in Section 2 above, the interface is subdivided into 2 subtasks: one for the disambiguation of senses, the second for the annotation of selections or coercions (see fig. 1). Both subtasks have an area for allowing annotation by annotators and another area for the judges to check the annotator choices and decide in case of disagreement. In both areas, sentences are organised by verb sense and argument grammatical relation.

For the annotation exercise described in this paper, we chose not to perform a proper annotation of verb senses, for the following reasons: 1) the very first experiment we ran with this methodology (Jezek, Quochi and Calzolari, 2009) showed a high and satisfactory interannotator agreement on verb senses[11]; 2) the sentences to be annotated in this exercise have been pre-selected tendentially including only the senses of interests.

Therefore, annotation of verb senses is done only by one of the judges, in order to exclude those sentences that may have been wrongly preselected.

4.2 Source Type Annotation

Two annotators were asked to annotate each sentence for the source type of the argument noun of the given target verb. By source type we mean the type of the noun "outside" the specific context.

The interface displayed the sentence with the target verb and nouns highlighted; the annotator was asked the question "What is the usual semantic type of the noun X? (fig. 2)", and had to choose one from a fixed list of available semantic types.

4.2.1. Revised Type Inventory

As type inventory for this specific annotation effort we revised the original GLML type list (see section 2). Because of the selectional preferences (or type templates) of some coercive verbs selected as described in section 3, some semantic type had to be added to the list, and some

[11] In fact, we had two verbs annotated for senses by two annotators and observed that agreement was 0.98 or above.

types were excluded (Proposition, Obligation and Rule) because not clear in their estensions or because they are not easily mappable to the SIMPLE Ontology (see below).

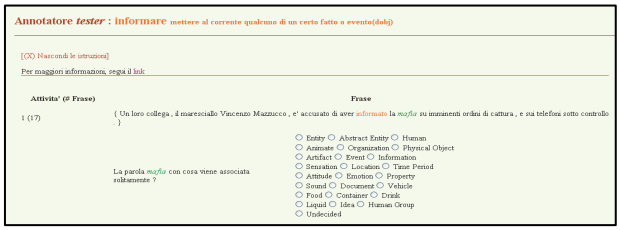

Figure 2: Type annotation interface

The types we added to the list[12] are: SOUND, LIQUID and 3 subtypes of ARTIFACT namely VEHICLE, CONTAINER and DOCUMENT[13]. The type list has also been mapped onto the SIMPLE Ontology so as to make the dataset linked to Italian lexical resources. For most types we were able to establish direct, 1:1, mappings and in some case the label are identical. In few other cases types from the type list can be mapped onto SIMPLE via a type plus a feature (i.e. SOUND :: STIMULI + sound; LIQUID :: SUBSTANCE +liquid).

Type Definitions

Since our previous pilot experiment showed that the simple type labels are not self-explanatory, annotators have been provided with a shallow taxonomy of types and simple guidelines for annotation: they have been instructed to choose the most specific type where appropriate and have been provided with a set of definitions with examples of the extension of each type has been provided.

Definitions and examples have been constructed has follows: first, we mapped the types onto the SIMPLE Ontology; then we compiled the definitions drawing from the SIMPLE specifications (i.e. the definition of the Semantic Classes) and from the type template glosses, when available. Examples are taken from the actual SIMPLE lexical resource (Lenci 2000).

Annotation Adjudication

Finally, a judge adjudicated all cases where the two annotators were in disagreement. Cases for which the judge could not make a choice are left unjudged and therefore automatically excluded from the final corpus[14].

[12] They are also added to the English type list

[13] We are aware that the choice of types may be questionable and this is in fact one of the issues to be further investigated. Of course, the most accurate solution would be to either use all same level concepts or the full range of concepts in a given ontology. This however was not practicable for several reasons, including practical ones related to the current set up of the tool.

[14] In this experiment we do not allow annotators to confront with each other nor force to judge to make a choice, because the cases of disagreement or of impossibility of adjudication are relatively small in number.

```
<instance id="2885">
{ Quel treno non era pero' uno dei diretti che
passano dalla stazione velocemente ma il Milano
Cremona , che aveva gia'
  <SELECTOR sid="sid_2885_22"
  lemma="cominciare" pos="v"
  sense_id="31">cominciato</SELECTOR>
  la
    <TARGET tid="t   _ _  !85_2905_24"
    lemma="FRENATA" pos="n" >frenata</TARGET>
  . }
<CompLink cid="1" selector_id="sid_2885_22"
relatedToTarget="tid_n_2885_2905_24"
gramRel="dobj" compType="SELECTION"
sourceType="Event" targetType="Event"/>
</instance>
```

Figure 3: Example of XML output

Annotations and judgements are stored in the DB and the final dataset can be exported in an XML file compliant with the GLML format (see Pustejovsky et al., 2009 and fig. 3 for an example). The exported dataset includes all and only those sentences on which both annotators agreed or those annotated by the judge.

The resulting dataset consists of a corpus of 3813 sentences with annotation of Selection or Coercion relations among verb-noun pairs, plus the source and target semantic types of the argument in the pair. The final targeted coercion types in the corpus are those reported in Table 2. in the Appendix.

A subset of this corpus was then selected as training and test corpus for the SemEval-2 ASC Task[15] (Pustejovsky and Rumshisky 2009). The SemEval corpus for Italian includes contexts 2893 for 8 coercions types for which we had at least 50 coercive contexts. The choice of the coercion types was also dictated by the need of having variability of target types and of not including too many Named Entities (e.g we excluded the ORGANISATION as HUMAN and the ORGANISATION as VEHICLE coercions, although we had more than 50 coercion examples in the whole annotated corpus). The SemEval corpus has also been expunged of examples of different and sparse coercion types which could constitute noise for classification systems.

4.3 Annotation Results

We measure results calculating the kappa coefficient (Carletta, 2006) on the annotation of coercions (i.e. when the annotated type does not match the type required by the verb) and selections (i.e. when the source type annotated corresponds to the type selected for by the verb) by the two annotators. Results are reported in table 2 in the Appendix. An overall K of 0.87 indicates that the annotation can be considered as reliable[16]. With respect to our first experiment, agreement on type annotation significantly increased. This may be primarily due to the

[15] http://sites.google.com/site/semevalasc/

[16] We adopted this statistics as a current standard in the field. We are aware of discussions on the appropriateness of the kappa statistics as a reliability measure for given uses of datasets (Reidsma & Carletta, 2008). However, it was not our aim, nor it is our expertise, to explore different metrics.

pre-selection of sentences, but also to the clearer annotation guidelines. Annotators reported that the provision of a taxonomic organisation of types (albeit shallow) and of definitions of the estension of classes as well as examples helped them during annotation.
At a qualitative analysis of the results on a verb basis (i.e. on coercion types), we find most disagreement cases with nouns that instantiate regular polysemy (or dot objects, in GL terminology): for example for verbs like *leggere* 'read' disagreement occurs with nouns such as *romanzo* 'novel', *saggio* 'essay', *favola* 'tale', *biografia* 'biography', that have been typed as DOCUMENT by one annotator and INFORMATION by the other.
The highest agreements are observed with HUMAN and LOCATION semantic types, whereas with EVENT agreement seems to vary depending on the verbs.
The qualitative analysis of the results so far allows us to identify some problems with the annotation methodology. We summarise these insights in the next section.

5. Insights for task improvements

Some issues related to noun polysemy remain unsolved in the current methodology: given that no sense disambiguation for nouns is elicited, part of the disagreement between annotators in the selection of the semantic type of the noun depends on whether the noun exhibits regular polysemy or whether it is associated to a complex type. In both cases, annotators may choose different types which are nevertheless both appropriate (see the examples with *leggere* in 4.3 above). Further thinking on such issues seems to be required.
The original GLML annotation framework, in fact, foresees a separate task for the annotation of dot objects (task 4 in section 2 above, cf. Pustejovsky et al., 2008). However, a treatment of complex types needs either to be integrated within the task for the annotation of compositional mechanisms, or to be performed as a first annotation step.
Finally, the current annotation scheme marks the effect but not the licensor of the coercion operation: that is, it allows for the marking of type shifting (e.g. ARTIFACT as SOUND as in "ascoltare la *radio* con le cuffie" 'listen to the radio with the headphones'), but does not foresee annotation of the Qualia role associated with the noun that is acted on by the verb in a coercive context. For example it does not allow for the marking of the constraint to Telic role of the noun *radio* in "ascoltare la radio" (e.g. produce_(SOUND)). However, we can argue that it is precisely the availability of this Quale that licenses the coercion (cf. *ascoltare il tavolo 'listen to the table').
This type of information may also help annotators, as well as automatic systems, to better identify the generative mechanism occurring in a given context.
Again, the original GLML methodology has a separate task for the annotation of Qualia in verb-argument pairs (task 2), but some reflection has to be made on whether to integrate this phase inside task 1 (as proposed in Jezek, Quochi and Calzolari, 2009), or to priorities tasks. We are currently exploring the possibility of integrating Qualia specification in the task as "off line" specification before annotation, e.g. taken from some background lexical resource such as the SIMPLE-PAROLE-CLIPS lexicon, or as "online" specification during the annotation.

6. Conclusion and future work

In the present paper we have described a first effort for annotation of type shifts in verb-argument pairs according to a newly defined Generative Lexicon methodology. The steps and results reported here are the outcomes of a revised approach both w.r.t. the original proposal (Pustejvsky et al. 2009) and to our very first experiment as described in Jezek, Quochi and Calzolari, 2009.
We annotated ca 4000 sentences distributed over 26 verbs with type shifts. The interannotator agreement, calculated with the kappa statistics, appears to be good enough for making the data set interesting for testing computational classification and learning models. However, the dimension of the data set is not very large and, as it is now, it is not representative of the real distribution of the coercions annotated.
In our future work, we plan to reduce the high cost of data set construction by having very little pre-selection of the sentences to be annotated. This will imply to have an overall larger number of sentences uploaded in the annotation tool and therefore higher annotation cost and time. In particular we plan to include a wider range of senses for each target verb (i.e. including not only coercive senses) and to annotate all coercion types that a verb instantiates, regardless of their frequency. This will give a better overview of the distribution of coercion phenomena in natural language and will render the resource more suitable for theoretical investigation.
As manual annotation is cost-intensive, the possibility of having non-expert annotators is interesting given the new potentiality offered by the web. The main idea here is to allow for the possibility of having non experts performing the annotation through the web, by exploiting existing resources like SIMPLE and PDIV as sources of word senses and semantic types (Jezek, Quochi and Calzolari 2009).

7. Acknowledgements

We thank the Istituto di Linguistica Computazionale, the Dipartimento di Linguistica Teorica e Applicata (Università di Pavia) and the BLLC Group (University of Brandeis) for supporting the project. The present work is the outcome of a joint and cooperative effort. However, for the specific concerns of the Italian Academy, Elisabetta Jezek is mainly responsible for section 3 and Valeria Quochi for section 4. A special thanks goes to Riccardo Del Gratta and Roberto Bartolini for helping in customizing and improving the tool and for pre-processing the data. We also acknowledge Francesca Frontini for development and maintenance of the GLML-Italian wiki, which was modeled on the English wiki previously developed at Brandeis, and Anna Rumshisky for providing us with the source code of the annotation tool. Finally, we express our thanks and appreciation to the volunteer students for helping in

selecting and annotating the data set: Vanessa Nardone, Irene De Felice, Filippo Pecorari, Ylenia Proteo, Francesca Frontini, Francesca Strik Lievers, Francesca Mazzariello and Lilian Schrik.

8. References

Baroni, M. & Kilgarriff, A. (2006). Large Linguistically-Processed Web Corpora for Multiple Languages. *Proceedings of European Association for Computational Linguistics* (EACL 2006), pp. 87-90.

Bindi, R. et al. (2000) PAROLE-Sottoinsieme. ILC-CNR Internal Report, Pisa: ILC.

Copestake, A., Briscoe, T. (1995). Semi-productive Polysemy and Sense Extension. *Journal of Semantics* 12(1), pp. 15-67.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistics. *Computational Linguistics* 22(2), pp. 249-254.

Hanks, P. & Jezek, E. (2007). Building Pattern Dictionaries with Corpus Analysis. Oral Presentation at *International Colloquium on Possible Dictionaries*, Roma Tre, Rome, June 6-7, 2007.

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

Jezek E. & Lenci A. (2007). When GL meets the corpus: a data-driven investigation of semantic types and coercion phenomena. *Proceedings of the 4th International Workshop on Generative Approaches to the Lexicon* (GL2007), Paris, France.

Jezek, E., Quochi, V., Calzolari, N. (2009). Relevance of Qualia Relations in Coercive Contexts. *Proceedings of the 5th International Workshop on Generative Approaches to the Lexicon* (GL2009), Pisa, Italy, CD-ROM, pp. 128-136.

Kilgarriff, A., Rychlý, R., Smrž P., Tugwell D. (2004). The Sketch Engine. In G. Williams and S. Vessier (eds.) *Proceedings of the XI Euralex International Congress*, July 6-10, 2004, Lorient, France, pp. 105-116.

Lenci, A., et al. (2000). SIMPLE: A General Framework for the Development of Multilingual Lexicons. *International Journal of Lexicography* XIII (4), pp. 249-263.

Markert, K. & Nissim, M. (2002). Towards a corpus annotated for Metonymies. The case of Location Names. In *Proceedings of the 3rd international Conference on Language Resources and Evaluation* (*LREC 2002*). Las Palmas, Spain.

Markert, K. & Nissim, M. (2009). Data and models for metonymy resolution. *Language Resources and Evaluation* 43, pp. 123-138.

Pustejovsky, J. (1995) *The Generative Lexicon*. Cambridge (MA): MIT Press.

Pustejovsky, J., Hanks, P., Rumshisky, A. (2004). Automated Induction of Sense in Context. In *Proceedings of COLING 2004*. Geneva, Switzerland.

Pustejovsky, J. et al. (2006). Towards a Generative Lexical Resource: The Brandeis Semantic Ontology. *Proceedings of LREC 2006*. Genoa, Italy, pp. 1702-1705.

Pustejovsky J. & Jezek E. (2008). Semantic Coercion in Language: Beyond Distributional Analysis. *Italian Journal of Linguistics/Rivista Italiana di Linguistica* 20(1), pp. 181-214.

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., Moszkowicz, J.L., Batiukova, O. (2009) GLML: Annotating argument selection and coercion. Paper presented at the *IWCS-8 Eighth International Conference on Computational Semantics*, Nijmegen, January 2009.

Pustejovsky, J., Rumshisky, A. (2009). SemEval-2010 Task 7: Argument Selection and Coercion. *SEW-2009: Semantic Evaluations: Recent Achievements and Future Directions (NAACL Workshop)* Boulder, Colorado, USA.

Reidsma, D. & Carletta, J. (2008). Reliability Measurement without Limits. *Computational Linguistics* 34(3), pp. 319-326.

9. Appendix

verb	*sense*	*grel*	*Coercion Types*	*Target argtype*	*#annotated sentences*	*K*
accusare	charge or blame someone of wrongdoing or error	Subj, DObj	Organization→Human Location→ Human	Human	281	0.97subj 1.00dobj
annunciare	make a formal statement to a public audience concerning an event that has recently taken place or a plan that will shortly be put into effect	Subj	Organization→Human Location→ Human Artifact→ Human	Human	227	0.96
arrivare	reach a location	Comp	Event→Location	Location	32	1.00
ascoltare	make conscious effort to hear a sound	DObj	Event→Sound Artifact →Sound	Sound	262	0.87
atterrare	come down to the ground safely and reach the programmed destination	Subj	Human→Vehicle	Vehicle	191	0.93
avvisare	inform or acquaint someone of a certain fact or knowledge	Subj, DObj	Organization→Human Location→ Human Artifact→ Human Vehicle→ Human	Human	249	0.97subj 1.00dobj
chiamare	contact someone by phone	Subj, DObj	Organization→Human Location→ Human Artifact→ Human Vehicle→ Human	Human	238	1.00subj 1.00dobj
cominciare	initiate an undertaking	DObj	Artifact→Event Document → Event	Event	28	0.90
completare	finish an activity	Comp	Artifact→Event Document → Event	Event	104	0.96
concludere	bring an activity to an end	DObj	Artifact→Event Document → Event	Event	56	0.13
contattare	establish communication with someone	Subj, DObj	Organization→Human Location→ Human Vehicle→ Human	Human	406	0.92subj 0.96dobj
divorare	read something eagerly and quickly	Dobj	Human→Document Event→ Document	Document	19	0.10
echeggiare	(of a sound) be repeated or reverberate after the original sound has stopped	Subj	Event→Sound Artifact →Sound	Sound	16	0.75
finire	bring to an end, complete an activity	DObj	Artifact→Event Document → Event	Event	158	0.73
informare	acquaint someone of a certain fact or knowledge	Subj, DObj	Organization→Human Location→ Human Artifact→ Human Vehicle→ Human	Human	109	0.96subj 0.90dobj
interrompere	stop the continuous progress of an activity or process	DObj	Artifact→Event Document → Event	Event	44	0.79
leggere	look at and grasp the meaning of some info contained in written material	DObj	Human→Document Event→ Document	Document	687	0.76
leggere	utter or render out loud some info contained in written material	DObj	Human→Document Event→ Document	Document	108	0.74
organizzare	bring about an event or action by planning and overseeing it	Subj	Organization→Human Location→ Human	Human	31	1.00
parcheggiare	leave a vehicle in a location	Subj	Vehicle→Human Organization→ Human	Human	27	1.00
raggiungere	attain or arrive at a location	DObj	Event→Location	Location	47	1.00
recar(si)	go to a given location	Comp	Event→Location	Location	125	0.98
rimbombare	(of a sound) be loud enough to echo	Subj	Event→Sound Artifact →Sound	Sound	24	0.71
sentire	perceive or pay attention to a sound	DObj	Event→Sound Artifact →Sound	Sound	183	0.68
udire	perceive a sound	DObj	Event→Sound Artifact →Sound	Sound	41	0.83
venire	move toward or into a place near or familiar to the speaker	Comp	Event→Location	Location	64	1.00
visitare	go to and spend some time in a place for turism, business or other purpose	DObj	Event→Location	Location	128	1.00
			All		3885	0.87

Table 2: Annotation: synoptic table