HuRIC: a Human Robot Interaction Corpus

Emanuele Bastianelli(‡), Giuseppe Castellucci(•), Danilo Croce(†)
Luca Iocchi(⋆), Roberto Basili(†), Daniele Nardi(⋆)

(†) DII, (‡) DICII, (•) DIE - University of Roma Tor Vergata - Rome, Italy
{bastianelli,castellucci}@ing.uniroma2.it, {basili,croce}@info.uniroma2.it
(⋆) DIAG - University of Roma La Sapienza - Rome, Italy
{nardi,iocchi}@dis.uniroma1.it

Abstract

Recent years show the development of large scale resources (e.g. FrameNet for the *Frame Semantics*) that supported the definition of several state-of-the-art approaches in Natural Language Processing. However, the reuse of existing resources in heterogeneous domains such as Human Robot Interaction is not straightforward. The generalization offered by many data driven methods is strongly biased by the employed data, whose performance in out-of-domain conditions exhibit large drops. In this paper, we present the *Human Robot Interaction Corpus* (HuRIC). It is made of audio files paired with their transcriptions referring to commands for a robot, e.g. in a home environment. The recorded sentences are annotated with different kinds of linguistic information, ranging from morphological and syntactic information to rich semantic information, according to the *Frame Semantics*, to characterize robot actions, and *Spatial Semantics*, to capture the robot environment. All texts are represented through the *Abstract Meaning Representation*, to adopt a simple but expressive representation of commands, that can be more easily translated into the internal representation of the robot.

Keywords: Human-Robot Interaction, Corpus, Natural Language Processing

1. Human Robot Interaction

Robots are slowly becoming part of everyday life, as they are being marketed for commercial applications (viz. telepresence, cleaning or entertainment). The *Human Robot Interaction* (HRI) research field aims at realizing robotic systems that offer an interaction level as much natural as possible (Scheutz et al., 2011). Several issues arise in translating human generated commands into suitable robotic actions. First, the underlying meaning of an utterance needs to be understood, and then mapped into robot-specific commands, in order to fill the gap between the robot world representation and the linguistic information conveyed in user utterances. In a sense, this is a typical form of *semantic parsing*.

Semantics is the crucial support for grounding linguistic expressions into objects, as they are represented in the robot set of beliefs (i.e. robot knowledge). The complexity of this task largely increases in less restricted scenarios, such as house serving tasks, where people do not follow a-priori known subsets of linguistic expressions. This requires robust command understanding processes, in order to face the flexibility of natural language. In many Natural Language Processing tasks, where robustness and domain adaptation are crucial, e.g. Open Domain Question Answering as discussed in (Ferrucci et al., 2010), methods based on Statistical Learning (SL) theory have been successfully applied.

In this perspective, we are investigating the combination of different state-of-the-art textual inference technologies aimed at modeling and making use of semantic aspects mostly relevant for HRI. We started from the idea that robotic systems are firstly required to exhibit two main features: in order to be useful they are expected to perform actions and these take place in a physical environment. Multiple *semantic theories* can be applied to describe the aspects of the world that should be taken in account for a robust HRI. In this work we point out *Frame Semantics* (Fillmore, 1985) and *Holistic Spatial Semantics* (Zlatev, 2007) as relevant to our goal. Frame Semantics generalizes the notion of action by making reference to a situation, i.e. an experience usually represented by a *Frame*, i.e. a micro-theory about a real world situation, such as `movement actions`. Holistic Spatial Semantics (Zlatev, 2007) defines the basic concepts in the domain of natural language spatial expressions. It helps to make reference to the location or the trajectory of a motion, usually involving one referent in a discourse.

In recent years the adoption of sound linguistic theories brought to the development of large scale resources (e.g. FrameNet (Baker et al., 1998) for the *Frame Semantics* or the CLEF Corpus for *Spatial Semantics*) to support the definition of several state-of-the-art Statistical Learning approaches for NL tasks. However, the reuse of these resources in heterogeneous domain is not straightforward. The generalization offered by ML algorithms is strongly biased by the employed data, whose performance in out-of-domain conditions may exhibit large drops. This is a crucial problem and a specific research topic, i.e. *Domain Adaptation* (Daumé and Marcu, 2006), deals exactly with these classes of problems. For example, as reported in (Pradhan et al., 2008; Johansson and Nugues, 2008), a Semantic Parsing system trained over a specific application-domain corpus shows a significant performance drop, when applied to different domains.

In this paper, we will present the *Human Robot Interaction Corpus* (**HuRIC**) we are collecting[1]. It is made of 570 audio files paired with their transcriptions referring to commands for a robot. The recorded sentences are also annotated with different kinds of linguistic information, ranging from morphological and syntactic information to rich

[1] Available at `http://sag.art.uniroma2.it/huric`

semantic information, according to *Frame Semantics* and *Spatial Semantics* theories. It has to be noticed that the two representations look at independent properties and can cooperate to fully express a model of the sentence meaning, useful for the HRI domain. The integration of these information is not straightforward. In order to accommodate different dimensions in the semantic annotation we are fostering the adoption of the *Abstract Meaning Representation* (AMR) (Banarescu et al., 2013). AMR is a novel representation language that allows to generalize several aspects of the NL semantics. In this representation schema, sentences having different syntactic structures, but basically sharing the same meaning, are seemingly expressed and represented by the same structure. It is very useful to provide a simple but expressive representation of commands, that can be easily translated into the internal representation of the robot.

In the rest of the paper we will first introduce the corpus gathering process (Section 2.) Section 3. then reports in-depth details about the corpus, also discussing some examples and issues raised during the annotation process. A first set of evaluation experiments involving HuRIC are presented and discussed in Section 4.

2. Corpus Collection

During the recent years the effort of providing resources useful for the automatic understanding robot commands has yielded the definition of corpora for Natural Language HRI, as (Kuhlmann et al., 2004; Tellex et al., 2011; Dukes, 2013). However, these corpora are highly domain or system dependent. The basic idea of this work is to build a corpus containing information that are yet oriented to a specific application domain, e.g. the house service robotics, but at the same time inspired by sound linguistic theories, that are by definition decoupled from such a domain. The aim is to offer a level of abstraction that is totally independent from the platform, but yet motivated by supported theories. We exploited three different situations to start gathering user utterances representing possible commands given to a robot in a house environment. Two groups of utterances have been recorded during the Speaky for Robots project[2], while a third one has been collected by interviewing members of the teams participating at the Robocup 2013 competition. At the end of the gathering, three datasets have been collected, each representing a different working condition. Each dataset is mainly characterized by: the complexity of the language used by the user, in terms of degree of variability of syntactic structures and lexicon; background noise conditions; device used for recording. Each utterance is coupled with its correct transcription, directly inserted by the user under an operator's control. Utterances have been pronounced by different users, so that multiple spoken versions of the same sentence are included. In the following, the gathering process is discussed for each of the three datasets.

The **Grammar Generated** (*GG*) dataset contains sentences that have been generated by the speech recognition grammar developed for the Speaky For Robots project. The

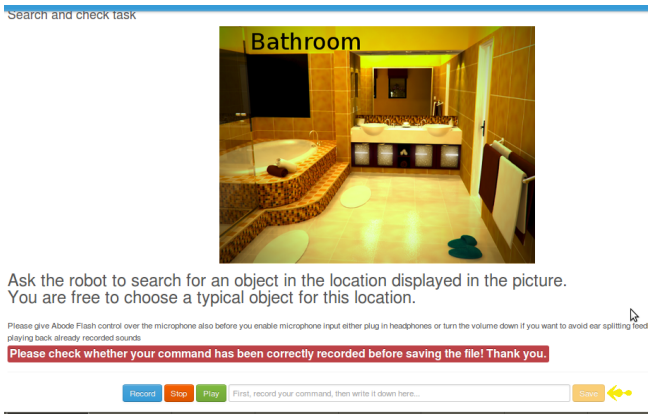

Figure 1: The web portal used for the gathering through crowd-sourcing

generation procedure of this grammar is presented in (Carlucci Aiello et al., 2013). The generated sentences have then been pronounced by three speakers and recorded using a push-to-talk microphone. The acquisition process took place inside a small room, thus with low background noise. Moreover, the push-to-talk mechanism helped in precisely segmenting the audio stream, and further reducing the noise. Due to its constrained nature, the language represented here is free of colloquial forms of interaction.

The **S4R Experiment** (*S4R*) dataset has been gathered in two distinct phases of the Speaky for Robots project experiment. In the first phase, the users were asked to give commands to a real robot operating in rooms set up as a real home, thus capturing all the interferences generated by talking people or sounds of other working devices nearby. The users were aware about the robot capabilities in terms of action it could perform and about the rooms and all the objects the robot was able to recognize. The same device used for the *GG* dataset has been employed here for the interaction. In a second phase, the users could access the Web portal showed in Figure 1 to record other commands. General situations involved in an interaction were described in the portal by displaying text and images. Each user was asked to give a command inherent to the depicted situation. This time the internal microphone of the pc running the portal has been used for recording. Since the users were only partially constrained (they had knowledge only about capabilities of the robot and lexicon handled by the Speech Recognition Engine), the language represented in this dataset is characterized by features that are more similar to free spoken English with respect to the one reported in the *GG* dataset, including richer syntactic structures.

The **Robocup** (*RC*) dataset has been collected during the Robocup@Home (Wisspeintner et al., 2009) competition held in 2013, in the context of the RoCKIn[3] project. The same Web portal used for the *S4R* dataset has been employed, and the recording took place directly in the competition venues or in a cafeteria, thus with different levels of background noise. Again, the internal microphone of the pc running the web portal has been employed. Expressions uttered here exhibit large flexibility in lexical choices

[2] http://labrococo.dis.uniroma1.it/?q=s4r

[3] http://rockinrobotchallenge.eu/

and syntactic structures, since the users did not received any constraint about what to command to the robot, except for the description of the situation. As a consequence, this dataset is much more variable representing a realistic "open" application with respect to the previous two command collections.

All the sentences from each dataset have been then annotated according to *Frame Semantics* and *Holistic Spatial Semantics* by two annotators. POS-tags and syntactic dependency types provided by the CoreNLP[4] (Klein and Manning, 2003) have been also validated during the annotation process, and are provided together with the semantic information. In the last phase of the annotation process, all the tagged information have been validated by a third expert. In order to facilitate the annotation and validation process, a dedicated platform, the *Data Annotation Platform* (*DAP*), has been implemented: its front-end is showed in Figure 2. The tool provides the possibility to tag semantics, syntax in term of dependency types, POS-tag as well as changing the lemma of each word. Moreover, a specific functionality of DAP allows the user to manually assign a quality score to each audio file. Files with a score of 0 are automatically rejected. In a similar way, it is possible to mark syntactically wrong sentences that have been inserted by mistake. The annotations produced are then saved in a database containing also all the information about the three datasets, including speaker's generalities (e.g. age, nationality, background experience in HRI) and the specific device used for the recordings.

3. Corpus Details

In this Section a deep analysis of the corpus characteristics is carried out. General statistics about the composition of the corpus are reported, as well as accurate measurements regarding the annotation process.

3.1. Corpus Statistics

Each of the three datasets composing HuRIC includes a set of audio files representing a robot command, paired with the correct transcription. Each sentence is annotated with the same source of information: lemmas, POS-tags, dependency trees, Frame Semantics and Spatial Semantics.

Table 3.1. reports the number of audio files for each dataset, together with the number of sentences corresponding to their transcriptions. We asked different speakers to pronounce the same command more than one time, in order to provide variable training material to optimize acoustic models for ASR engines. Statistics about the nationality of the different speakers involved are reported in Table 3.1.. As can be noticed, the distribution of the nationalities is different depending on the dataset. For the S4R experiment mainly native-speakers have been selected, or at least very good English speakers, e.g. people that have been living in Anglo-Saxon countries for several years. For the Robocup dataset, instead, speakers from all the teams have been involved, resulting in a more varied distribution. Table 3.1. shows the average number of audio files per speaker for each dataset.

[4] nlp.stanford.edu/software/corenlp.shtml

	#audio files	#sentences	#audio file per sentence
GG	137	48	∼2.85
S4R	141	96	∼1.46
RC	292	177	∼1.64

Table 1: Number of audio files and sentences

Nationality	GG	S4R	RC
Australia	0	0	3
Brazil	0	0	1
UK	0	6	2
Chile	0	0	2
China	0	0	1
Cyprus	0	0	1
Czech Republic	0	0	1
Holland	0	0	5
German	0	0	4
India	1	0	1
Indonesia	0	0	1
Italy	2	2	5
Japan	0	0	1
Mexico	0	0	1
Romania	0	1	0
Spain	0	0	2
Syria	0	0	1
Switzerland	0	1	0
USA	0	2	4
Total	3	12	36

Table 2: Distribution of the nationality of the speakers

The different experiment background also biased the pragmatics of the pronounced sentences in the three datasets. We classified them into three classes: *imperative*, representing the will of a user to have a robot performing an action, e.g. every direct command as "*bring me the bottle of water*"; *descriptive* sentences, used to describe a situation to the robot, e.g. "*there is a bottle on the table*"; within this last set, a specific class is created for definitional sentences, useful to teach the robot about the environment, such as in assigning a category to an entity, e.g. "*this is the kitchen*". The CC and S4R datasets present only imperative sentences, since during the experiments the users were asked only to give direct commands to the robot, as information about the environment was already acquired and known a priori. Instead, in the Robocup dataset in some cases we also asked the users to give descriptions of the involved scene, in order to augment the robot knowledge about the world. The number of sentences belonging to each class are reported in Table 3.1..

With regard to morpho-syntactic information: lemmas, POS-tags and dependency parse trees are uploaded in the Data Annotation Platform and lately validated. Statistics about the fine-grain POS-tags are reported in Table 3.1.. Table 3.1. instead shows the distribution of the general coarse-grain POS-tags, e.g. verbs or nouns.

3.2. Annotating Frame Semantics

The first aim that we had in mind about the realization of this corpus was to provide linguistic information about spoken commands useful to encode knowledge necessary to have the robot to fully understand them. As a first step, we started considering the actions a robot should be able to perform according to a set of general commands. We

	average audio file per speaker
GG	∼45.7
S4R	∼11.8
RC	∼8.1

Table 3: Average number of audio files per speaker

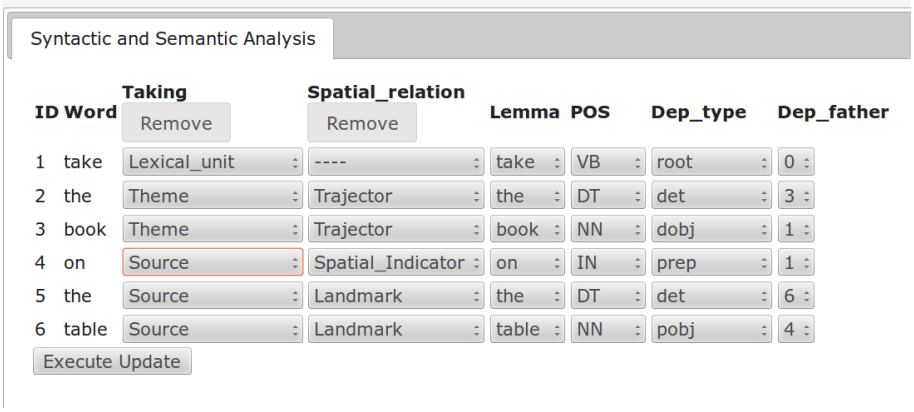

Figure 2: The Data Annotation Platform

	Imperative	Descriptive	Definitional
GG	48	0	0
S4R	96	0	0
RC	150	14	13

Table 4: Distribution of utterance classes

POS	GG	S4R	RC
CC	0	5	31
CD	0	0	16
DT	86	152	285
EX	0	0	11
FW	0	0	1
IN	49	66	134
JJ	0	7	43
JJS	0	0	1
MD	0	2	28
NN	87	188	365
NNS	3	2	22
POS	0	1	0
PRP	2	7	79

POS	GG	S4R	RC
PRP$	3	7	20
RB	5	5	16
RP	0	0	1
TO	12	38	66
UH	0	13	38
VB	47	97	165
VBD	0	4	6
VBG	0	1	2
VBN	0	1	1
VBP	0	1	11
VBZ	0	0	29
WDT	0	0	3
WRB	0	0	2

Table 5: Fine-grain morpho-syntactic information

decided to rely on *frames* as the bridge between the linguistic knowledge contained in the utterances and the robotic actions. *Frame Semantics* generalizes the notion of action by making reference to a situation, i.e. an experience usually represented by a *Semantic Frame* (Fillmore, 1985). A frame is a micro-theory about a real world situation: `movement actions`, such as *moving*, `events`, such as *natural phenomena*, or `properties`, such as *being colored*. A set of semantic roles is associated to each frame, i.e. the descriptors of the different elements involved in the described situation (e.g. the *Agent* of a movement). Linguistic resources providing such information have been produced over the years, as FrameNet (Baker et al., 1998). Frames are interesting primitives in HRI as robot's actions can be linked to semantic frames. This semantic formalism already inspired the RoboFrameNet (Thomas and Jenkins, 2012) framework, where a set of semantic frames contextual to specific robot actions has been defined. The main difference with our work is that in RoboFrameNet the frames have been defined at a finer-grained level with respect to our choices. Higher levels of abstraction of

POS	GG	S4R	RC
CC	0	5	31
CD	0	0	16
DT	86	152	285
EX	0	0	11
FW	0	0	1
IN	49	66	134
J	0	7	44
MD	0	2	28

POS	GG	S4R	RC
N	90	190	387
P	5	15	99
RB	5	5	16
RP	0	0	1
TO	5	38	66
UH	0	13	38
V	47	105	214
W	0	0	5

Table 6: Coarse-grain morpho-syntactic information

FrameNet frames allow to treat a more general set of phenomena and have been settled as a starting point.

In our case, we selected the subset of FrameNet-inspired semantic frames corresponding to the defined robot actions, that are reported in Table 3.2.. Some frames have been slightly adapted, e.g. the frame *Scrutiny* has been called *Searching*. According to our set of frames, in the command "*go near the table of the kitchen*" the *Motion* frame is evoked by the verb *go*. The frame element GOAL, representing the destination of the motion action is assigned to the sequence *near the table of the kitchen*. The frame instantiated in this way finally encodes all the information needed to the robot to understand what action to perform, and which are the arguments involved in the command, i.e. in this case which object to take and where to find it. Besides the number of examples for each semantic frame in the three datasets, Table 3.2. reports also the number of examples of each frame element according to its frame. Again, when speakers had more freedom, as in the RC dataset, the expressivity has grown, resulting in a higher number of different frames and related frame elements. Furthermore, this aspect is reflected also by the higher average number of roles per sentence that goes from $\sim$1.5 of the CC dataset to $\sim$2.0 for the RC dataset.

The automatic annotation process can be decomposed into three subtask: the *Frame Prediction* (*FP*), the *Boundary Detection* (*BD*) and the *Argument Classification* (*AC*). The first is the task of recognizing the type of the intended action, reflected by the events evoked by the targeted sentence, that is recognizing the semantic frames evoked in a sentence; the second is the task of identifying the spans of the arguments of an action, i.e. the *frame elements* related to a given frame; the third aims at assigning a label to each span identified during the *BD* step, e.g. THEME or SOURCE. The *Inter-Annotator-Agreement* (IAA) between the two annotators has been evaluated for each of these subtasks, in terms of Precision, Recall and F-Measure. In turn the tagging of one annotator has been evaluated against the other (thus considering the second as the Gold Standard), and the mean of the two scores has been reported as the IAA for each measure. These scores are showed in Table 3.2.. For the BD and the AC subtasks, both the *exact match* and the *token match* have been reported. The first represents the

	GG	S4R	RC
Attaching	1	0	2
ITEM	0	0	2
GOAL	1	0	0
Being_in_category	0	0	14
CATEGORY	0	0	14
ITEM	0	0	14
Being_located	0	0	20
LOCATION	0	0	11
PLACE	0	0	6
THEME	0	0	20
Bringing	10	22	37
AGENT	0	1	6
BENEFICIARY	2	1	13
GOAL	8	21	24
MANNER	0	0	1
SOURCE	0	0	9
THEME	10	22	37
Change_operational_state	1	3	3
DEVICE	1	3	3
OPERATIONAL_STATE	1	1	2
Closure	6	0	1
CONTAINER_PORTAL	2	0	1
CONTAINING_OBJECT	4	0	0
Entering	4	0	1
GOAL	4	0	1
Following	1	6	30
AREA	0	0	1
COTHEME	1	6	30
GOAL	0	1	5
MANNER	0	3	6
PATH	0	0	1
SPEED	0	0	1
THEME	0	0	6
Giving	0	0	2
RECIPIENT	0	0	2
THEME	0	0	2
Inspecting	0	1	3
DESIRED_STATE	0	1	1
GROUND	0	1	3
INSPECTOR	0	1	1
Motion	9	25	39
AREA	0	0	1
GOAL	9	25	38
MANNER	2	1	1
PATH	0	1	1
THEME	0	0	8
Perception_active	1	0	0
PHENOMENON	1	0	0
Placing	0	7	10
AGENT	0	0	1
GOAL	0	7	10
THEME	0	7	10
Releasing	0	2	0
GOAL	0	2	0
THEME	0	2	0
TIME	0	1	0
Searching	3	27	24
COGNIZER	0	0	5
GROUND	3	16	7
PHENOMENON	3	27	24
PURPOSE	0	0	5
Taking	12	6	12
AGENT	0	0	4
PURPOSE	0	0	2
SOURCE	10	3	5
THEME	12	6	12
Total # of frames	48	99	198
Av. frames per sentence	1.00	1.03	1.12
Total # of roles	74	16	36
Av. roles per sentence	1.54	1.67	2.02

Table 7: Distribution of Frames and related Frame Elements

percentage of roles that have been exactly tagged, meaning that a frame element has been correctly tagged only if its entire span matches the Gold Standard one. The second measure refers to the percentage of token correctly tagged inside the labeled spans. More details and examples about the IAA are reported in Section 3.5.

3.3. Annotating Spatial Semantics

The first consideration that arose while modeling the semantics of commands through *Frame Semantics* has been

	FP			BD			AC		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
	Exact Match								
GG	97.9	97.9	97.9	93.2	93.2	93.2	90.5	90.5	90.5
S4R	95.5	95.5	95.5	93.8	94.4	94.1	93.2	93.8	93.5
RC	95.2	95.2	95.2	84.5	84.5	84.4	82.8	82.8	82.7
	Token Match								
GG	-	-	-	96.3	96.3	96.2	93.0	93.0	92.9
S4R	-	-	-	94.6	95.4	94.9	92.8	93.6	93.2
RC	-	-	-	89.9	89.9	89.8	85.0	85.0	85.0

Table 8: *Frame Semantics* Inter Annotators Agreement

that the related actions take place in an environment. We then focused on the different spatial aspects involved in such interactions, and how the spatial domain is represented in the language. Even though *Frame Semantics* is able to capture some of these aspects, we found out that in some cases the granularity level offered by this theory was not appropriate. Understanding the spatial relations that hold between two or more entities and conveyed through natural language can be crucial for HRI. For example, lets consider the command "*go near the table in the kitchen*". *Frame Semantics*, as defined in FrameNet, is not able to capture the relation holding between *the table* and *the kitchen*, as the whole sequence *near the table in the kitchen* is supposed to be considered as the destination of the motion trajectory, i.e. the GOAL frame element. Identifying such relation would allow a robot to understand which is the table we are referring to, among all the tables the robot is aware of.

We then decided to rely on the *Holistic Spatial Semantics* (Zlatev, 2007) to catch such phenomena. This theory defines the basic concepts in the domain of natural language spatial expressions. It helps to make reference to the location or the trajectory of a motion, usually involving one referent in a discourse. It defines the concept of *spatial relation*, as a relation holding among different *spatial roles* that can be identified in a sentence. This can be a TRAJECTOR, i.e. the entity whose location is of relevance, a LANDMARK, i.e. the reference entity by which the location of the trajectory of the motion is fully specified, or a SPATIAL_INDICATOR, i.e. the part of a sentence holding and characterizing the nature of the whole relation. For example, in the sentence "*go near the table in the kitchen*", the preposition "*in*" is the SPATIAL_INDICATOR of the relation between "*table*" and "*kitchen*", respectively a TRAJECTOR and a LANDMARK.

Spatial Semantics in term of these three roles have been annotated over the whole HuRIC, in line with the annotated CLEF corpus discussed in (Kolomiyets et al., 2013). Table 3.3. reports the number of spatial relations annotated over the three datasets, together with the total number of spatial roles. It is worth noting that the number of LANDMARKS is different from the other two roles because sometimes it can be implicit, e.g. *go near* [*the table*]$_{\text{TRAJECTOR}}$ [*on the right*]$_{\text{SPATIAL\_INDICATOR}}$. The average number of spatial relations and roles per sentence is also reported. The *Inter-Annotator-Agreement* has been evaluated for each spatial role. It has been measured in the same way as for the *Frame Semantics*, and is reported in Table 3.3., considering both the *exact match* and the *token match* measures.

	GG	S4R	RC
Spatial_relation	24	15	47
TRAJECTOR	24	15	47
SPATIAL_INDICATOR	24	15	47
LANDMARK	18	14	41
Av. relation per sentence	0.50	0.16	0.27
Av. role per sentence	1.38	0.46	0.76

Table 9: Distribution of Spatial Relations and Spatial Roles

	TRAJECTOR			SPATIAL INDICATOR			LANDMARK		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
	Exact Match								
GG	94.2	94.2	94.1	98.1	98.1	98.0	92.4	92.4	92.3
S4R	90.6	91.2	90.0	90.6	91.2	90.0	90.0	90.6	89.3
RC	85.8	88.6	85.7	81.4	81.4	81.3	84.7	84.7	84.6
	Token Match								
GG	93.2	93.2	93.2	99.1	99.1	99.1	88.1	88.1	88.1
S4R	90.6	91.2	90.0	93.8	94.0	93.5	90.5	91.0	89.8
RC	81.6	81.6	81.6	86.1	86.1	86.0	83.8	83.8	83.7

Table 10: *Spatial Semantics* Inter Annotators Agreement

3.4. Abstract Meaning Representation

From the considerations made in the previous Sections, it arises that the two selected representations look at independent properties and can cooperate to fully express the meaning of a command. They both represent different forms of annotation of relevant expressions useful for the HRI domain. In order to accommodate different dimensions in the semantic annotation we are fostering the adoption of the *Abstract Meaning Representation* (AMR) (Banarescu et al., 2013). AMR is a novel representation formalism. It allows to generalize several aspects of the NL semantics and it is ideal to abstract some aspects that are currently less relevant (such as quantification) and to focus on spatial (and temporal) concepts, their instances and the individual relations among them. In AMR, sentences that have different syntactic structures, but basically share the same meaning, are seemingly expressed and represented by the same structure. A major advantage in robotic applications is that a fully instantiated AMR annotation can be easily mapped to the corresponding operational commands for the robot, i.e. making a smooth notion of grounding already available.

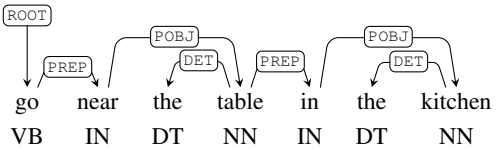

Figure 3: Example of a dependency graph associated to sentence "*go near the table in the kitchen*"

The AMR annotation provided within HuRIC is automatically produced by translating the arguments expressing semantic information. The syntactic tree is used to select the semantic head in each argument in order to instantiate all the variables within the AMR resulting structure. When a spatial relation overlaps a frame element, the semantic head expressing the first is linked to the second, so capturing the spatial specification carried out by the *Spatial Semantics*. For example starting from the two different annotations for the sentence "*go near the table in the kitchen*"

[***go***]$_{Motion}$ [*near the table in the kitchen*]$_{\text{GOAL}}$

go near [*the table*]$_{\text{TRAJECTOR}}$ [*in*]$_{\text{SPATIAL\_INDICATOR}}$ [*the kitchen*]$_{\text{LANDMARK}}$

the spatial relation is used to further specify the GOAL frame element. The syntactic tree shown in Figure 3 suggests that the *table* is the semantic head of the prepositional construction underlying the GOAL role and, at the same time, it reflects a TRAJECTOR. The spatial information can be thus embedded in the GOAL role, recursively generating the following structure.

```
(g1 / go − Motion
    :Goal(i1/ in − Spatial_relation
            :trajector(t1/ table)
            :landmark(k1/ kitchen)))
```

According to the AMR guidelines, the spatial relations can been seen as a surrogate of the information that is represented by the AMR role `be-located-at`.
At the moment of writing, the generation script is able to correctly build the AMR representation for about the 90% of sentences of the whole HuRIC and further work is being done to increase the coverage.

3.5. `HuRIC` Resulting Annotations

In this section a list of annotation examples is showed in order to present some directives we adhered and to discuss about issues we faced during the tagging. The main point is that the level of knowledge the robots are today able to manage is quite far and distant from the huge amount of linguistic information that is present also in simple sentences like the ones contained in HuRIC. To this end, we propose to use the AMR as a container that can be enriched step by step with all the necessary information, according to the robot capabilities. For example, if the robot is not able to deal with quantification, there will be no need of integrating such information inside the AMR. For this reason, we are removing all those aspects that are currently negligible from the final AMRs. For each frame role, we insert in the structure only the semantic head of its textual span. The semantic head is selected by navigating the dependency subtree of the frame element text, and looking for the noun that is closer to the root, as it is possible to see in the tagging of the command "*can you please bring the phone to the bathroom*":

can [*you*]$_{\text{AGENT}}$ please [***bring***]$_{Bringing}$ [*the phone*]$_{\text{THEME}}$ [*to the bathroom*]$_{\text{GOAL}}$

```
(b1 / bring − Bringing
    :Agent(y1/ you)
    :Theme(p1/ phone)
    :Goal(b2/ bathroom))
```

Here the semantic head *phone* and *bathroom* have been respectively selected for the roles THEME and GOAL.
The AMR guidelines define an exhaustive set of non-core roles used to deeply specify most of the semantic phenomena occurring in texts. Since by now we want to capture just the essential information that will allow the robot to understand the commands here reported, we decided to integrate

in the AMR only those features that are strictly necessary for our tasks. Noun modifiers and adjectives are most of the time needed in the discrimination process for a robot, in order to correctly identify the referenced entity. For example, in the command "*carry the mug to the dining room*"

[***carry***]$_{Bringing}$ [*the mug*]$_{\text{THEME}}$ [*to the dining room*]$_{\text{GOAL}}$

```
(c1 / carry − Bringing
    :Theme(m1/ mug)
    :Goal(r1/ room
          :mod(d1/ dining)))
```

the *dining* modifier of the noun *room* is crucial to understand the specific room we are talking about. Without going to deep in the definition of different non-core roles, we used the `:mod` label in those cases.

One main issue, arose during the annotation procedure, concerned the intrinsic ambiguity of some prepositional phrase attachments. This phenomenon has been manifested especially for the *Taking* frame, as showed by the command "*take the bottle on the table*". Here the preposition *on* can be attached directly to the verb or to the noun *bottle*. This gives rise to different interpretations in term of frames and relative frame elements. According to the first case (1), *the bottle* and *on the table* can be tagged respectively as the THEME and the SOURCE for a *Taking* frame; it has to be said that such approach may rise another issue, since *on the table* might be interpreted as the GOAL frame element of a *Bringing* frame, thus changing the whole meaning of the command. In the second case (2), the whole sequence *the bottle on the table* can be labeled as a unique THEME.

1) [***take***]$_{Taking}$ [*the bottle*]$_{\text{THEME}}$ [*on the table*]$_{\text{SOURCE}}$

2) [***take***]$_{Taking}$ [*the bottle on the table*]$_{\text{THEME}}$

take [*the bottle*]$_{\text{TRAJECTOR}}$ [*on*]$_{\text{SPATIAL\_INDICATOR}}$ [*the table*]$_{\text{LANDMARK}}$

```
(t1 / take − Taking
    :Theme(o1/ on − Spatial_relation
           :trajector(b1/ bottle)
           :landmark(t2/ table))
    :Source(t2))
```

As it is possible to notice from the AMR example above, we decided to adhere to the first interpretation (1). This choice has been mainly driven by the CoreNLP syntactic parsing, that uses to attach such prepositions directly to the verb. Moreover, we considered the verb *take* as source from a *Bringing* frame only in those cases in which a frame element introduced by non ambiguous prepositions as *to* or *towards* was present. Another important fact outlined by the *Spatial Semantics* annotation of this example is that the two semantic representations may add information that seem redundant. Since we are still dealing with these kind of issues, we decided to report both the representations in such cases.

Regarding the Inter Annotator Agreement, one of the main source of disagreement in the annotation of the *Frame Semantics* concerned the tagging of the role CATEGORY frame element for the *Being_in_category* frame. Expecially, what differs in the two annotation approaches is the span to which the frame element has been associated. For example, for the command *this is a bedroom with pictures on the wall*, one annotator considered only the portion of text corresponding to the direct entity representing the category, i.e. *a bedroom* in the following example.

[*this*]$_{\text{ITEM}}$ [***is***]$_{Being\_in\_category}$ [*a bedroom*]$_{\text{CATEGORY}}$ *with four pictures on the wall*

The second annotator incorporated the whole span corresponding to the sub-tree starting from the entity representing the category, i.e. *a bedroom with pictures on the wall*.

[*this*]$_{\text{ITEM}}$ [***is***]$_{Being\_in\_category}$ [*a bedroom with four pictures on the wall*]$_{\text{CATEGORY}}$

This generated a significant drop in the Boundary Detection and Argument Classification agreement, as it is possible to see in Table 3.2..

4. A First Empirical Evaluation

As a support to our thesis about the need of liguistic resources for Natural Language HRI and in order to verify their possible impact in the development of HRI systems, we ran some preliminary experiments involving HuRIC. We designed a processing chain able to annotate *Frame Semantics* according to the schema described in Section 3.2. The chain is realized as a cascade of processors exploiting different Statistical Learning algorithms, and it is explained in depth in (Croce et al., 2012). The *Frame Prediction* task is designed as a multi-classification process, using the $SVM^{Multiclass}$ algorithm (Joachims et al., 2009). Both *Boundary Detection* and *Argument Classification* subtasks are realized as a sequential labeling task through the SVM^{Hmm} (Altun et al., 2003), i.e. a Markovian extension of the SVM. In the first subtask, the labeler gives a label to each word indicating whether this is part or not of a frame element. In the second, the labeler assigns a label to each word, corresponding to the associated frame element.

First, we trained the system over the FrameNet corpus, and evaluate its performances over the three dataset composing HuRIC (reported in the FN column of Table 4.). In the second experiment, 66% of annotated examples of each subsets of HuRIC have been added to the FrameNet material for training. In this way a 3-fold evaluation schema has been enabled: cyclically, three different tests have been made possible on the remaining 33% of each test corpus. The macro-average achieved over three measures have been computed as the final performance, and is reported under the $HuRIC$ column in Table 4..

In the labeling chain, we fed each module with gold standard input. The results reported in Table 4. show how adding training from HuRIC results in a significant improvement of the performances, expecially for the *BD* and the *AC* phases. Results show the percentage of correctly assigned frame (the FP column), the percentage of correctly recognized roles (the BD column) and correctly classified with respect to the role labels (the AC column). A constant improvement is shown when adding HuRIC annotated

Gold information at each step						
	FP		BD		AC	
	FN	HuRIC	FN	HuRIC	FN	HuRIC
GG	0.826	0.833	0.684	0.871	0.589	0.822
S4R	0.812	0.817	0.743	0.872	0.736	0.912
RC	0.732	0.758	0.696	0.817	0.701	0.898

Table 11: Semantic Role Labeling analysis with HuRIC

examples, especially in the $S4R$ and RC dataset. It confirms the positive impact of using specialized material, i.e. reflecting the syntactic/semantic peculiarity of robot commands.

5. Conclusion

This paper presents the *Human Robot Interaction Corpus* (HuRIC), a first collection of robot commands to support the HRI research. First, sentences have been recorded as audio files. Then, transcriptions have been annotated with different kinds of linguistic information, ranging from morphological and syntactic information to rich semantic information, according to *Frame Semantics* and *Spatial Semantics*. Finally, the texts annotated in this way have been represented through the *Abstract Meaning Representation* (AMR) formalism, a novel and flexible representation schema. First experimental evaluations underline the positive impact of this resource in the adoption of data driven methods for the Semantic Parsing of natural language commands. Moreover, the adoption of an open representation schema as AMR will enable for extensions of the corpus according to other semantic theories, as explicit temporal annotations. Such information could be useful for HRI in order to better support the planning of sequences of (and not individual) actions.

Acknowledgment: authors want to thank Cristina Giannone for her invaluable support in the development of the DAP system.

6. References

Altun, Y., Tsochantaridis, I., and Hofmann, T. (2003). Hidden Markov support vector machines. In *Proceedings of the ICML*.

Baker, Collin F., Fillmore, Charles J., and Lowe, John B. (1998). The berkeley framenet project. In *Proceedings of the ACL 98*, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Banarescu, Laura, Bonial, Claire, Cai, Shu, Georgescu, Madalina, Griffitt, Kira, Hermjakob, Ulf, Knight, Kevin, Koehn, Philipp, Palmer, Martha, and Schneider, Nathan. (2013). Abstract meaning representation for sembanking. In *Proceedings of the LAW VII & ID*, Sofia, Bulgaria, August.

Carlucci Aiello, L., Bastianelli, E., Iocchi, L., Nardi, D., Perera, V., and Randelli, G. (2013). Knowledgeable talking robots. In *AGI*.

Croce, Danilo, Castellucci, Giuseppe, and Bastianelli, Emanuele. (2012). Structured learning for semantic role labeling. *Intelligenza Artificiale*, 6(2):163–176.

Daumé, III, Hal and Marcu, Daniel. (2006). Domain adaptation for statistical classifiers. *J. Artif. Int. Res.*, 26(1):101–126, May.

Dukes, Kais. (2013). Train robots: A dataset for natural language human-robot spatial interaction through verbal commands. In *ICSR. Embodied Communication of Goals and Intentions Workshop*, October.

Ferrucci, David, Brown, Eric, Chu-Carroll, Jennifer, Fan, James, Gondek, David, Kalyanpur, Aditya A., Lally, Adam, Murdock, J. William, Nyberg, Eric, Prager, John, Schlaefer, Nico, and Welty, Chris. (2010). Building Watson: An Overview of the DeepQA Project. *AI Magazine*, 31(3).

Fillmore, Charles J. (1985). Frames and the semantics of understanding. *Quaderni di Semantica*, 6(2):222–254.

Joachims, Thorsten, Finley, Thomas, and Yu, Chun-Nam. (2009). Cutting-plane training of structural SVMs. *Machine Learning*, 77(1):27–59.

Johansson, Richard and Nugues, Pierre. (2008). The effect of syntactic representation on semantic role labeling. In *Proceedings of COLING*, Manchester, UK, August 18-22.

Klein, Dan and Manning, Christopher D. (2003). Accurate unlexicalized parsing. In *Proceedings of ACL'03*, pages 423–430.

Kolomiyets, Oleksandr, Kordjamshidi, Parisa, Bethard, Steven, and Moens, Marie-Francine. (2013). Semeval-2013 task 3: Spatial role labeling. In *Proceedings of SemEval-2013*.

Kuhlmann, Gregory, Stone, Peter, Mooney, Raymond, and Shavlik, Jude. (2004). Guiding a reinforcement learner with natural language advice: Initial results in RoboCup soccer. In *The AAAI-2004 Workshop on Supervisory Control of Learning and Adaptive Systems*, July.

Pradhan, Sameer S., Ward, Wayne, and Martin, James H. (2008). Towards robust semantic role labeling. *Comput. Linguist.*, 34(2):289–310.

Scheutz, Matthias, Cantrell, Rehj, and Schemerhorn, Paul. (2011). Toward humanlike task-based dialogue processing for human robot interaction. *AI Magazine*, 34(4):64–76.

Tellex, Stefanie, Kollar, Thomas, Dickerson, Steven, Walter, Matthew R., Banerjee, Ashis Gopal, Teller, Seth J., and Roy, Nicholas. (2011). Understanding natural language commands for robotic navigation and mobile manipulation. In Burgard, Wolfram and Roth, Dan, editors, *AAAI*. AAAI Press.

Thomas, Brian J and Jenkins, Odest Chadwicke. (2012). Roboframenet: Verb-centric semantics for actions in robot middleware. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*.

Wisspeintner, T., van der Zant, T., Iocchi, L., and Schiffer, S. (2009). RoboCup@Home: Scientific competition and benchmarking for domestic service robots. *Interaction Studies*, 10(3):393–428.

Zlatev, Jordan. (2007). Spatial semantics. *Handbook of Cognitive Linguistics*, pages 318–350.