Providing Automatic Multilingual Text Generation to Artificial Cognitive Systems

Carles Fernández
Computer Vision Centre
Universitat Autònoma de Barcelona, Spain

Xavier Roca
Computer Vision Centre
Universitat Autònoma de Barcelona, Spain

Jordi Gonzàlez
Institut de Robòtica i Informàtica Ind.
Universitat Politècnica de Catalunya,Spain

{perno | xavir | poal}@cvc.uab.es

Abstract

This contribution addresses the incorporation of a module for advanced user interaction into an artificial cognitive vision system to include the *human-in-the-loop*. Specifically, the document describes a method to automatically generate natural language textual descriptions of meaningful events and behaviors, in a controlled scenario. One of the goals of the system is to be capable of producing these descriptions in multiple languages. We will introduce some relevant stages of the whole system, and concentrate on the linguistic aspects which have been taken into account to derive surface text from conceptual predicates. Some experimental results are provided for the description of simple and complex behaviors of pedestrians in an intercity crosswalk, for Catalan, English, Italian, and Spanish languages.

1 – Introduction

The introduction of Natural Language (NL) interfaces into vision systems has become popular, especially for surveillance systems (Gerber & Nagel, 2008). In this kind of

applications, human behavior is represented by predefined sequences of events. Scenes are evaluated and automatically translated into text by analyzing the contents of the images over time, and deciding on the most suitable predefined event that applies in each case.

Such a process is referred to as Human Sequence Evaluation (HSE) in (Gonzàlez, Rowe, Varona, & Roca, 2008). HSE takes advantage of cognitive capabilities for the semantic understanding of observed situations involving persons. This conception aims to perform an automatic evaluation of generally complex human behavior from image sequences in restricted discourse domains. In our case, the domain of interest has been restricted to urban outdoor surveillance environments.

This automatic analysis and description of temporal events was already tackled by Marburger et al. (Marburger, Neumann, & Novak, 1981), who proposed a NL dialogue system in German to retrieve information about traffic scenes. More recent methods for describing human activities from video images have been reported by Kojima et al. (Kojima, Tamura, & Fukunaga, 2002), and automatic visual surveillance systems for traffic applications have been studied in (Nagel, 2004) and (Buxton & Gong, 1995), among others. These approaches present one or more specific issues such as textual generation in a single language, surveillance for vehicular traffic applications only, restrictions for uncertain data, or very rigid environments.

We aim to build a system which addresses the aforementioned drawbacks by following the proposals of HSE, in order to generate NL descriptions of human behavior appearing in controlled scenarios. There exist several considerations that have been taken into account for the design of such a system fulfilling the aforementioned requirements:

- The resulting system should be *flexible* enough to: (i) enable a multilingual generation of discourse in natural language with average external users, and (ii) enable such a discourse to address the communication of complex events happening in the observed scenario, e.g. interactions among entities, contextualization of actions in a metric-temporal framework, or statements about reasoned interpretations for certain situations.
- This system has also been *restricted* to cover a defined domain of interest, given by the tackled outdoor inner city scenario and the model of possible situations to expect. As a result, we work with particularized linguistic models, which however must still be able to automatically produce natural descriptions of the occurring facts.

Experimental results have been focused to be specialized to a single type of scenario in order to study the problems in-depth, rather attempting to come up with a supposedly generally applicable solution. This agrees with the *situatedness* property of cognitive systems (Wilson & Keil, 2001). Two particular scenes have been considered, which contain complex situations resulting from the interaction of pedestrians and vehicles in an outdoor environment, see **Figure 1** and**Figure 2**. Both consist of crosswalk scenes, in which pedestrians, cars, and objects appear and interact. On the first scene, four pedestrians cross the road in different ways. Several behaviors appear on the second one, e.g. displacements, meetings, crossings, accelerations, object disposals, and more complex situations such as abandoned objects, dangers of running over, and thefts. The recording has been obtained using a distributed system of static cameras, and the scenario has been modeled a priori.

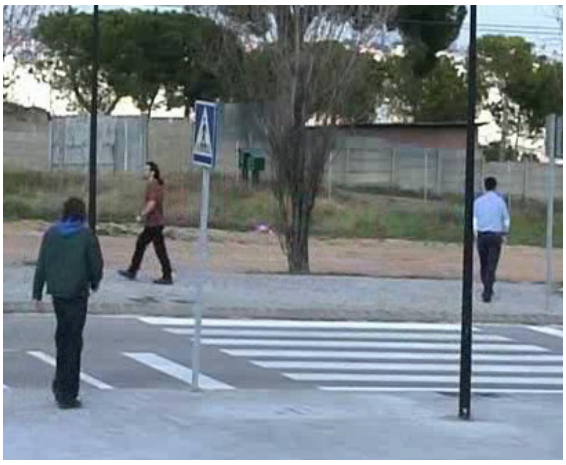


Figure 1: *Crosswalk scene showing simple behaviors*

Figure 2: *Crosswalk scene showing some complex behaviors and interactions*

Next section provides a brief overview about the results obtained at the vision and conceptual levels. After that, we detail the main stages and tasks accomplished specifically at the NL Generation (NLG) module. Finally, some results are shown and evaluated, and last section highlights some general ideas and concludes the work.

2 – Vision and Conceptual Levels

The *Vision level* acquires relevant visual content from the scenes by using a distribution of cameras. The detection and capture of interesting objects within the images is accomplished at this stage, by means of segmentation and tracking procedures which capture the motion information (Huerta et al., 2007; Rowe et al., 2005). As a result, a series of quantitative measurements over time is provided for each detected target, such as positions, velocities, and orientations of the agents.

Quantitative information cannot be naturally evaluated in linguistic terms, and therefore this data must be converted into qualitative facts. The *conceptual level* accomplishes this. First, spatiotemporal data is represented by means of logical predicates created for each frame of the video sequence, in which numerical information is represented by their membership to predefined fuzzy functions. For example, depending of the

instantaneous velocity value (`V`) for an agent, we may assign a `zero`, `small`, `normal`, or `high` tag to it, see **Figure 3.** Apart from categorizing instantaneous facts, a scenario model also enables to situate agents and objects in meaningful regions of the recorded location, e.g. crosswalk, sidewalk, or waiting zones.

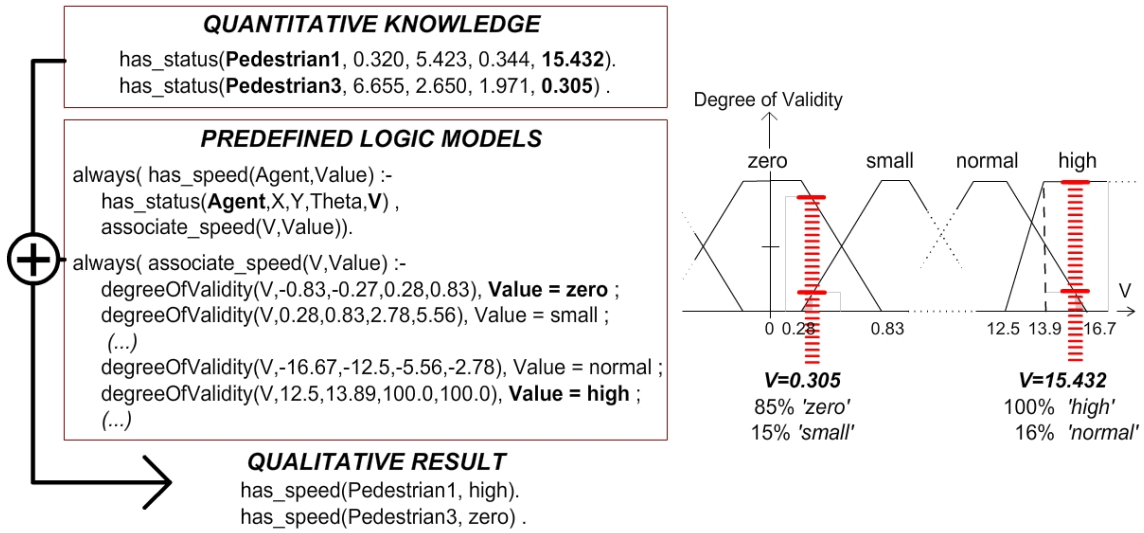


Figure 3 *– Conversion from quantitative to qualitative values. The numerical value of velocity for an agent (last field of* has_status*) at a time step is linked to the most probable membership of the* has_speed *fuzzy function.*

Nevertheless, we obtain a large collection of basic geometric facts that needs to be filtered, so that relevant information and patterns are extracted from it. Concretely, we want to detect admissible sequences of occurrences, which will contextualize geometric and temporal information about the scene, and will let us interpret the situation an agent is in. For instance, a sequence in which an agent walks by a sidewalk and stops in front of a crosswalk probably means that this agent is *waiting to cross*.

Situation Graph Trees are the specific tool used to build these models (Arens & Nagel, 2003; Gonzàlez, Rowe, Varona, & Roca, 2008), see **Figure 4**. They connect a set of defined situations by means of prediction and specialization edges. When a set of conditions is asserted, a high-level predicate is produced as an interpretation of a

situation. An interesting property at this point is that the produced notes are much closer to a linguistic lecture, since they interrelate and put into context different semantic elements such as locations, agents, and objects. Nevertheless, these expressions still keep language independence, and hence are a good starting point for multilingual text generation. More information about this situational analysis can be found in (Fernández, Baiget, Roca, & Gonzàlez, 2007).

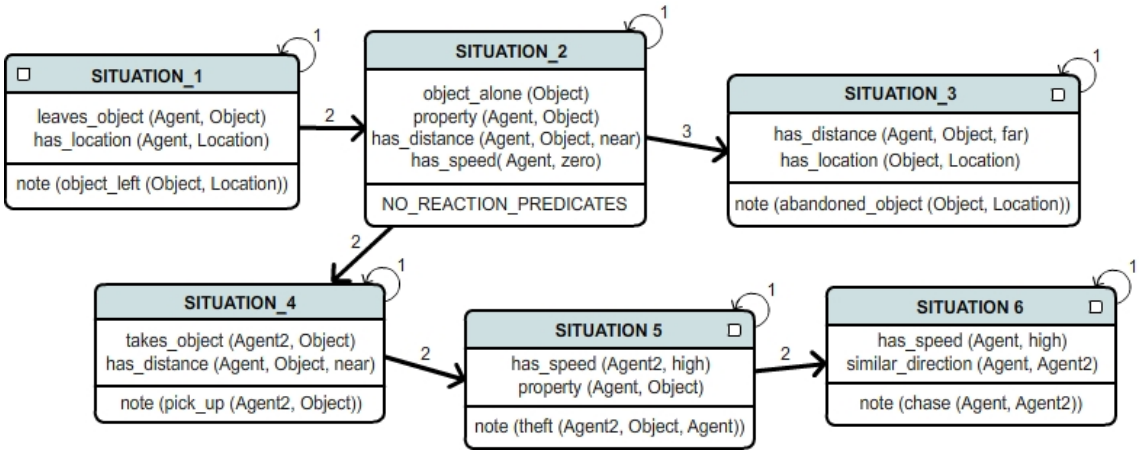


Figure 4 *– Situation Graph Trees are used to model situations and behaviors as predefined sequences of basic events. The example shown allows for complex inferences such as abandoned objects, chasings or thefts, by means of high-level* note *predicates.*

3 – The NLG Module

NLG can be seen as a subfield of both computer science and cognitive science. It focuses on computer systems which can automatically produce understandable texts in a natural human language, so it is concerned with computational models of language and its use. NLG has been often considered as a process of *choice*, in which the most suitable mean has to be selected to achieve some desired end (Reiter & Dale, 2000).

The set of situations that need to be expressed are modeled and made available to the purposed NLG module, so that the main goal for this module consists of selecting one unique form of expressing that information in a clear and natural way, for each of the

languages considered. This module is then built from a deterministic point of view, since it deals with aforeknown situational models.

The overall process of NL text generation is based on a model of architecture similar to the one proposed in (Reiter & Dale, 2000), which includes three modules, see **Figure 5**:

- A *Document Planner*, which produces a specification of the text's content and structure, i.e. what has to be communicated by the NLG, by using both domain knowledge and practical information to be embedded into text.
- A *Microplanner*, in charge of filling the missing details regarding the concrete implementation document structure, i.e. in which way the information has to be communicated: distribution, referring expressions, level of detail, voice, etc.
- A *Surface Realizer*, which converts the abstract specification given by the previous stages into a real text, possibly embedded within some medium. It involves traversing the nodal text specification until the final presentation form.

Visual trackers acquire basic quantitative information about the scene, and the reasoning system decides how this information needs to be structured, gives coherency to the results, and also carries out inferences based on predefined conceptual models. All these tasks are related to the Document Planner, since they provide the structured knowledge to be communicated to the user. Further tasks, such as microplanning and surface realization, are included specifically into the NLG module.

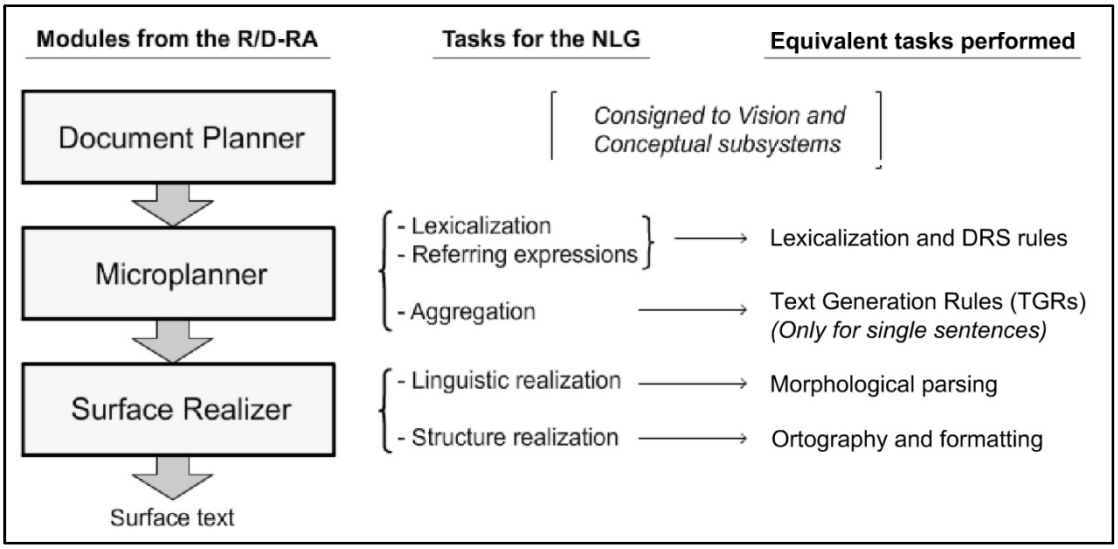


Figure 5 *– Schema of Reiter/Dale Reference Architecture (R/D-RA) [9], including the tasks related to each module that are necessary for a Natural Language Generator.*

The NLG module receives high-level semantic predicates from the reasoning stage, which are eventually converted into surface text. There are several tasks to cover in order to carry out this process; they have been structured into the following stages:

1. Discourse Representation
2. Lexicalization
3. Surface Realization

Besides, the set of lemmata for the domain of interest has to be extracted from a restricted corpus of the specific language. The different corpora have been elaborated based upon the results of several psychophysical experiments on motion description, collected over a significative amount of native speakers of the target language. In our case, ten different people have independently contributed to the corpus with their own descriptions of the sample videos. Four different languages have been implemented for this scenario: Catalan, English, Italian, and Spanish.

3.1 Representation of the Discourse.

The chosen approach towards the implementation of semantics for NL generation is based on Discourse Representation Theory (Kamp & Reyle, 1993). This theory enables to construct semantic structures representing linguistic information contained in NL sentences, in predicate logic formalism. Semantic relationships are stated by means of Discourse Representation Structures (DRSs). Here, the inverse process is implemented, consisting of the retrieval of NL text from logic predicates, by defining a set of DRS construction and transformation rules for each tackled language.

One of the main semantic characteristics to take into account refers to cohesiveness. When a contextual basis is explicitly provided, the maintenance of the meaning for a discourse, including its cross-references, relations and cohesion can be granted. A particularly interesting and comprehensible example of discourse cohesion is the case of anaphoric pronominalization, which allows the generation of some referring expressions; for instance, we typically discard "*The pedestrian waits to cross. The pedestrian crosses*", in favor of "*The pedestrian waits to cross. S/he crosses*".

DRSs are semantic containers which relate referenced conceptual information to linguistic constructions (Kamp & Reyle, 1993). A DRS always consists of a so-called universe of referents and a set of conditions, which can express characteristics of these referents, relations between them, or even more complex conditions including other DRSs in their definition. These structures contain linguistic data from units that may be larger than single sentences, since one of the ubiquitous characteristics of the DRSs is their semantic cohesiveness for an entire discourse.

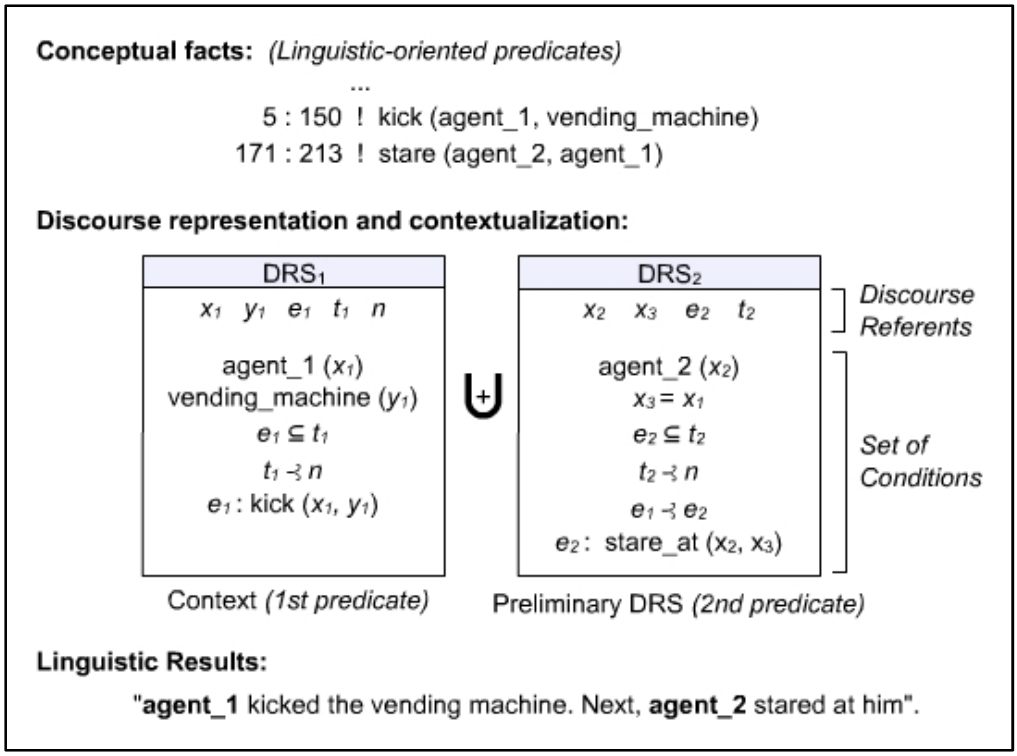


Figure 6 *– A pattern DRS allows to convert a stream of conceptual predicates into a string of textual symbols. Here, two predicates are validated. The first one instantiates a DRS, which serves as context for the following asserted facts. Once a new predicate is validated, it instantiates another DRS which merges with that context, thus providing a new context for subsequent facts. The temporal order of the events is stated by including them within time variables ($e_1 \subseteq t_1$), placing these variables in the past ($t_1 \prec n$), and marking precedence ($e_1 \prec e_2$).*

By using such structures, we will be able to point out the cross-references existing among the semantic constituents of a predicate. The classification of linguistically-perceived reality into thematic roles (e.g. agent, object, location) is commonly used in contemporary linguistic-related applications as a possibility for the representation of semantics, and justifies the use of computational linguistics for describing content extracted by vision processes. In the current implementation, these constituents can be classified as *agents*, *objects*, *locations*, and *events/situations*. Given that a situational analysis is accomplished for each detected agent, we base on previously mentioned information about the focused agent to decide upon referenced expressions or full descriptions. An example which shows how the semantic representation and

contextualization is undertaken by a DRS is illustrated in **Figure 6**. DRSs also facilitate the subsequent tasks for sentence generation. The syntactical features of a sentence are provided by the so-called Text Generation Rules (TGRs), which establish the position for the elements of the discourse within a sentence for a particular language. Due to the specific goals considered for this system, simple sentences are used for effective communication.

The question of how to address temporal references also arises at the semantic level. A natural possibility consists of tensing the statement of recent observations in present perfect (e.g. *He has turned left*), and handle inferences in present time (e.g. *He waits to cross*), although there exists a certain flexibility for the selection of tenses. A discourse referent for the utterance time of discourse (*n*) is required, so that the rest of temporal references t_i can be positioned with respect to it, see **Figure 6**.

3.2 Lexicalization.

As stated in (Reiter & Dale, 2000), *lexicalization* is the process of choosing words and syntactic structures to communicate the information in a document plan, i.e. the interpreted knowledge of logical predicates within a defined domain. Concretely, we will have to map the messages from the predicates, now linked by DRSs, into words and other linguistic resources that explain the semantic contents we want to communicate. It is difficult to bound the lexicalization process to a single module, since the mappings from semantic to linguistic terms are accomplished at several stages of the architecture; in this section we focus lexicalization of prior knowledge, i.e. *agents*, *objects*, and *locations*, which have to be known beforehand.

The lexicalization step can be seen as a mapping process, in which the semantic concepts identifying different entities and events from the selected domain are attached to linguistic terms referring those formal realities. This way, this step works as a real dictionary, providing the required lemmata that will be a basis for describing the results using natural language. Parsing processes will be in charge of traversing the syntactical structures obtained by the Text Generation Rules, and replacing the semantic identifiers by their suitable linguistic patterns. **Figure 7** shows an example of lexicalization for two aforeknown identifiers of semantic regions from the scenario.

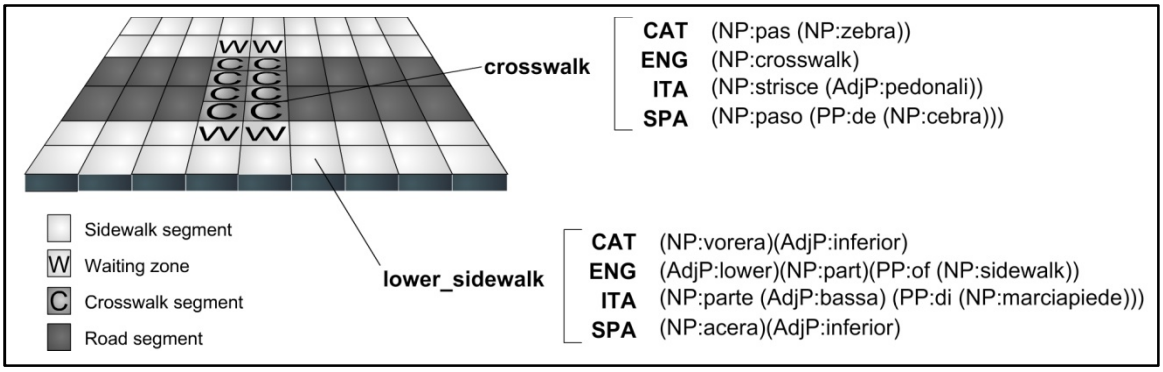


***Figure** 7 – Example depicting lexicalization for locations, in which a linguistic structure is associated with a semantic region of the scenario for each considered language. Only basic structural information is represented here, although morphological characteristics are also provided to the linguistic terms at this step.*

3.3 Surface Realization.

The Surface Realization stage is accomplished in two steps. A first morphological process applies over each single word and partially disambiguates the individual abstraction of that word, by means of morphological attributions such as gender or number. These attributions can be propagated upon the semantic relations previously established by DRSs among the lemmata of a single piece of discourse. After that, a set of post-morphological rules has been conceived to enable interactions among predefined configurations of words, thus affecting the final surface form of the text. This additional step is indispensable for many languages, in which certain phenomena force the surface

form to change, e.g. contractions (*a* + *el* → *al*, in Spanish), or order variation (*es* + *va* + *en* → *se'n va*, in Catalan). **Table 1** shows some examples of morphological rules included in the grammar used for parsing.

VP:<go>¬vL	VP:<gone>¬v	;(ENG)
VP:<meet>¬vL	VP:<met>¬v	;(ENG)
VP:<_>¬vL	VP:<_ed>¬v	;(ENG)
PP:<a>¬p Det:<el>¬dMS	PP:<al>¬pdMS	;(CAT)
PP:<per>¬p Det:<el>¬dMS	PP:<pel>¬pdMS	;(CAT)
[Det:<_>¬dS] ^=vowel	Det:<l'>¬d	;(CAT)
[PP:<de>¬p] ^=vowel	PP:<d'>¬p	;(CAT)
[Det:<quest_>¬dS] ^=vowel	Det:<quest'>¬dS	;(ITA)

Table 1 *- Examples of some simple morphological rules in Catalan, English, and Italian. The upper ones, in English, allow obtaining the participle (tag ¬L) of a verb (¬v). The third rule is general, the two first are examples of exceptions and shall appear first. In the second set of rules, the Catalan and Italian ones, prosodic manipulation is allowed. The two first examples of this second set enable contractions of certain prepositions and determiners; the three last examples show the situation in which certain words appearing in front of a word starting by vowel experiment apostrophication.*

Finally, a general scheme for the entire process of generation is shown in **Figure 8**.

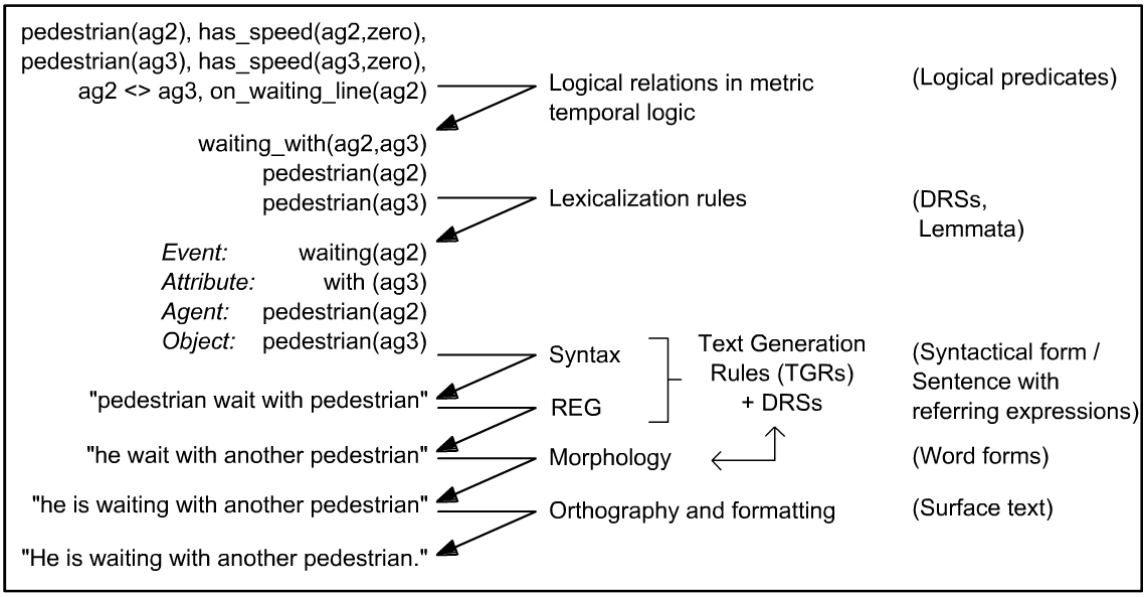
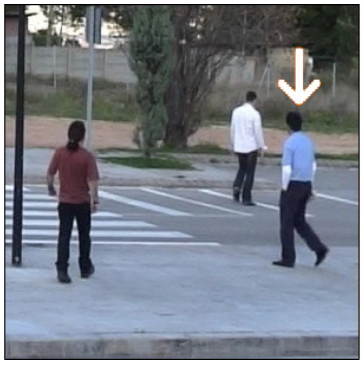
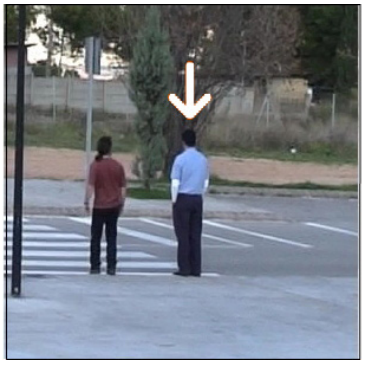
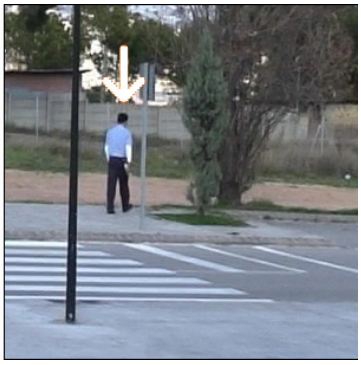


Figure 8 *– Example for the generation of the sentence "He is waiting with another pedestrian" from logical predicates and for the English language. The center column contains the tasks being performed, and the right column indicates the output obtained after each task.*

Experimental results

Next, some results are provided for the two scenes considered. For the first crosswalk scene, textual descriptions in Catalan, English, and Spanish have been selected for Agents 3 and 4, respectively. They include agents appearing or leaving, interactions with locations, and basic interpretations such as waiting with others to cross, or crossing in a dangerous way (i.e. directly by the road and not caring for vehicular traffic).

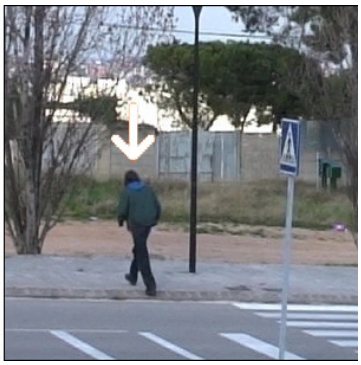
Pedestrian 3 (Catalan)

203 : *Lo vianant surt per la part inferior dreta.*
252 : *Va per la vorera inferior.*
401 : *S'espera per creuar.*
436 : *S'està esperant amb un altre vianant.*
506 : *Creua pel pas zebra.*
616 : *Va per la vorera superior.*
749 : *Se'n va per la part superior dreta.*

Pedestrian 3 (English)

203 : *The pedestrian shows up from the lower right side.*
252 : *S/he walks on the lower sidewalk.*
401 : *S/he waits to cross.*
436 : *S/he is waiting with another pedestrian.*
506 : *S/he enters the crosswalk.*
616 : *S/he walks on the upper sidewalk.*
749 : *S/he leaves by the upper right side.*





Pedestrian 4 (Spanish)

523 : *El peatón aparece por la parte inferior izquierda.*
572 : *Camina por la acera inferior.*
596 : *Cruza sin cuidado por la calzada.*
681 : *Camina por la acera superior.*
711 : *Se va por la parte superior izquierda.*

Pedestrian 4 (English)

523 : *The pedestrian shows up from the lower left side.*
572 : *S/he walks on the lower sidewalk.*
596 : *S/he crosses the road carelessly.*
681 : *S/he walks on the upper sidewalk.*
711 : *S/he leaves by the upper left side.*

Some results for the second scene are presented in Catalan, Italian, and English. In this case there exist more complex interactions and interpretations of events, e.g. abandoned objects, dangers of run over, thefts, or chasings.

470 ! *Un vianant surt per la part superior asquerra.*
470 ! *A pedestrian appears from the upper left side.*
470 ! *Un pedone compare nella parte superiore sinistra.*

492 ! *Lo vianant camina per la vorera superior.*
492 ! *Il pedone cammina sulla parte alta del marciapiede.*
492 ! *The pedestrian walks on the upper part of the sidewalk.*

583 ! *Gira pac a la dreta per la part superior de lo pas zebra.*
583 ! *S/he turns right in the upper part of the crosswalk.*
583 ! *Gira a destra sulla parte alta delle strisce pedonali.*

591 ! *S'ha parat allà mateix.*
591 ! *S/he has stopped in the same place.*
591 ! *Si è fermato in questa posizione.*

615 ! *Ha dixat l'objecte a terra.*
615 ! *S/he has left an object.*
615 ! *Ha lasciato un oggetto in terra.*

630 ! *Un nou vianant surt per la part superior dreta.*
630 ! *A new pedestrian appears from the upper right side.*
630 ! *Un altro pedone compare nella parte superiore destra.*

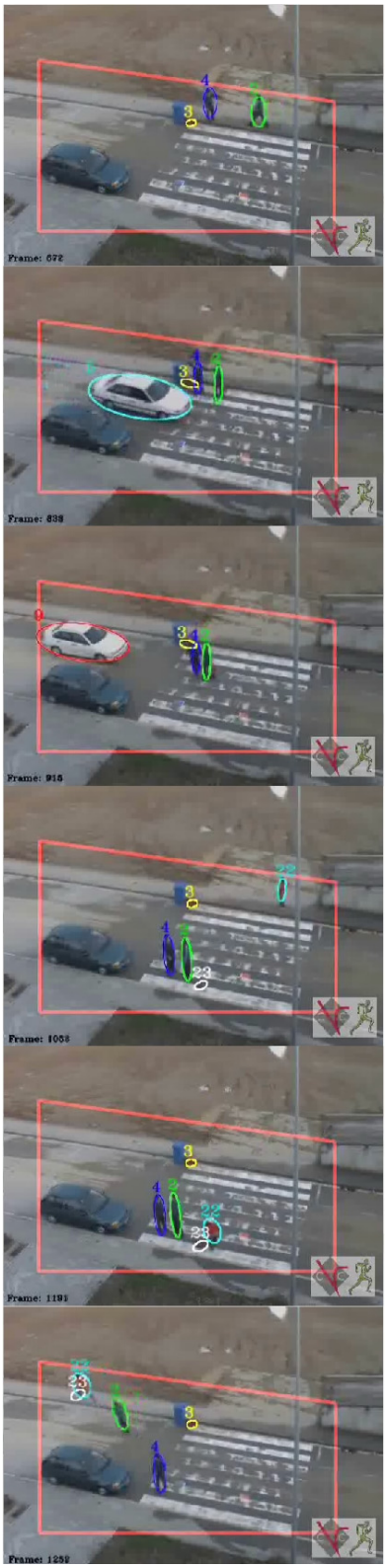
642 ! *Lo vianant camina per la vorera superior.*
642 ! *The pedestrian walks on the upper part of the sidewalk.*
642 ! *Il pedone cammina sulla parte alta del marciapiede.*

656 ! *Lo primer vianant camina per allà mateix.*
656 ! *The first pedestrian walks on the same place.*
656 ! *Il primo pedone cammina in questa zona.*

687 ! *L'objecte pareix haver astat dixat a la part superior de lo pas zebra.*
687 ! *The object seems to have been abandoned in the upper part of the crosswalk.*
687 ! *L'oggetto sembra che sia stato abbandonato nella parte alta delle strisce pedonali.*

692 ! *Lo primer vianant s'ha trobat en lo segon vianant allà mateix.*
692 ! *The first pedestrian has met the second pedestrian in the same place.*
692 ! *Il primo pedone si è incontrato con il secondo pedone in questa posizione.*

822 ! *Un vehicle pareix que astà a punt d'atropellar lo primer vianant.*
822 ! *A danger of runover between the first pedestrian and a vehicle seems to have been detected.*
822 ! *Un veicolo ha rischiato d'investire il primo pedone.*



Discussion

Most of the limitations for the described NLG module come clearly determined by the restrictive domain of work. The linguistic models need to be extended as new situations can be detected by the HSE system, since the content to be communicated is provided entirely by the situational analysis. The deterministic approach that has been chosen limits the variety of produced sentences, but ensures that the output results will be linguistically correct, since they obey the constructions proposed by native speakers and encoded into the models.

The modular architecture proposed for the NLG subsystem apparently allows the common stages to remain unchanged, disregarding the incorporation of new languages or the enlargement of the detection scope. So far, the addition of a new language has only required extending DRS rules and parsing grammars, which allows for a fast and effective implementation of similar languages.

Further steps include an enhancement of the Microplanner to support sentence aggregation. This would allow ordering the information structured in single sentences and mapping it into more complex sentences and paragraphs. Discourse Representation Theory has been proved consistent to accomplish this task (Kamp & Reyle, 1993).

Acknowledgements

This work is supported by EC grants IST-027110 for the HERMES project and IST-045547 for the VIDI-video project, and by the Spanish MEC under projects TIN2006-14606 and CONSOLIDER-INGENIO 2010 (CSD2007-00018). Jordi Gonzàlez also

acknowledges the support of a Juan de la Cierva Postdoctoral fellowship from the Spanish MEC.

References

Arens, M., & Nagel, H.-H. (2003). Behavioral Knowledge Representation for the Understanding and Creation of Video Sequences. *Proc. of the KI 2003. 1*, pp. 149-163. Berlin, Heidelberg, New York: Springer-Verlag.

Buxton, H., & Gong, S. (1995). Visual surveillance in a dynamic and uncertain world. *AI-magazine , I* (78), 431-459.

Fernández, C., Baiget, P., Roca, X., & Gonzàlez, J. (2007). Semantic Annotation of Complex Human Scenes for Multimedia Surveillance. *AI*IA 2007. Tenth Congress of the Italian Association for Artificial Intelligence.* (pp. 698-709). Roma, Italy: Springer LNAI.

Gerber, R., & Nagel, H.-H. (2008). Representation of Occurrences for Road Vehicle Traffic. (Elsevier, Ed.) *Artificial Intelligence , 4-5* (172), 351--391.

Gonzàlez, J., Rowe, D., Varona, J., & Roca, F. (2008). Understanding Dynamic Scenes based on Human Sequence Evaluation. *Image and Vision Computing* .

Huerta, I., Rowe, D., Mozerov, M., & Gonzàlez, J. (2007). Improving Background Subtraction based on a Casuistry of Colour-Motion Segmentation Problems. *3rd IbPRIA* (pp. 475-482). Girona, Spain: Springer LNCS.

Kamp, H., & Reyle, U. (1993). *From Discourse to Logic.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Kojima, A., Tamura, T., & Fukunaga, K. (2002). Natural Language description of Human Activities from Video Images Based on Concept Hierarchy of Actions. *International Journal of Computer Vision , II* (50), 171-184.

Marburger, H., Neumann, B., & Novak, H. (1981). Natural Language Dialogue about Moving Objects in an Automatically Analyzed Traffic Scene. *Proc. IJCAI-81, Vancouver* .

Nagel, H.-H. (2004). Steps toward a Cognitive Vision System. *AI Magazine , II* (25), 31-50.

Reiter, E., & Dale, R. (2000). *Building Natural Language Generation Systems.* Cambridge: Cambridge University Press.

Rowe, D., Rius, I., Gonzàlez, J., & Villanueva, J. (2005). Improving Tracking by Handling Occlusions. *3rd ICAPR* (pp. 384-393). UK: Springer LNCS.

Wilson, R., & Keil, F. (Eds.). (2001). *The MIT Encyclopedia of the Cognitive Sciences.* Bradford Book.