Representation of linguistic and domain knowledge for second language learning in virtual worlds

Alexandre Denis*, Ingrid Falk+, Claire Gardent* and Laura Perez-Beltrachini+

* CNRS/LORIA, +Lorraine University/LORIA
Nancy, France
{alexandre.denis,ingrid.falk,claire.gardent,laura.perez}@loria.fr

Abstract

There has been much debate, both theoretical and practical, on how to link ontologies and lexicons in natural language processing (NLP) applications. In this paper, we focus on an application in which lexicon and ontology are used to generate teaching material. We briefly describe the application (a serious game for language learning). We then zoom in on the representation and interlinking of the lexicon and of the ontology. We show how the use of existing standards and of good practice principles facilitates the design of our resources while satisfying the expressivity requirements set by natural language generation.

Keywords: virtual environments, ontology-lexicon interface, natural language generation

1. Introduction

As shown in (Geroimenko and Chen, 2005), (Ibanez-Martinez and Mata, 2006), semantic annotations of 3D scenes improve 3D content retrieval and managment. More recently, they have also been shown to enable intelligent simulations which support semantics reasoning or semantic-based planning (Kapahnke et al., 2010). In the serious game for learning French called I-FLEG (Interactive French Learning Game, (Amoia et al., 2011)), we exploit semantic annotations of 3D worlds in a novel way namely, as a means to support the generation by a 3D game of natural language sentences. In this game, touching a 3D object triggers the automatic generation of teaching material by a natural language generator. To produce this material, the generator uses (i) the semantic annotations of the 3D world i.e., the ontology describing the 3D objects in the virtual world used for teaching French and (ii) a lexicon which provides the linguistic information necessary to generate text about these objects.

In this paper, we show how to represent and link the ontology and the lexicon used in I-FLEG so as to facilitate maintenance and portability while satisfying the expressivity requirements set by the generation of natural language. In particular, we show how to use the LIR standard (Peters et al., 2009) to represent the lexicon and link it to the ontology; and how the good practice principles proposed in (Bateman, 1990) permits designing an ontology that best supports the requirement set by natural language generation.

We start (Section 2.) by sketching the game scenario and explaining how natural language generation is used to associate 3D objects with language learning exercices. In Section 3., we then zoom in on the semantic annotations of the 3D world and explain the principles underlying the design of the knowledge base. Finally in Section 4., we show how we adapted the LIR standard to develop a lexicon that mediates between the semantic annotations and the linguistic resources used by the generation system.

2. The I-FLEG serious game

To learn French with I-FLEG, the learner moves his avatar inside a virtual house and clicks on objects thereby triggering the display of language learning exercices. These exercises are generated by a natural language generator based on the ontological axioms and facts associated with the objects and on a lexicon linking concepts to words. Here is a toy example illustrating the ontology and the lexicon underlying text generation. Suppose that the world contains a red chair which, in the knowlege base is associated with the following axioms and facts:

Chair(c), SubClassOf(Chair, Noun),
Red(c), SubClassOf(Red, Adjective),
Small(c), SubClassOf(Small,Adjective),
John(j), SubClassOf(John, Noun),
Move(e), Agent(e,j), Theme (e,c)
SubClassOf(Move, Verb),

That is, there are three objects named c, j and e; c is a red chair, j is John, e is a moving event of c by j. Further, $Chair$ is a subconcept of the $Noun$ concept, Red and $Small$ of the $Adjective$ concept and $Move$ of the $Verb$ concept.

Now suppose that the learner wants to work on adjectives. Then the generator will select from this knowledge base a set of facts that can be verbalised as a sentence containing an adjective. In particular, the selection algorithm will search for an $Adjective$ subconcept (to produce an adjective) and a $Verb$ subconcept (to produce a verb). For instance, the following content might be selected:

Chair(c), Red(c), John(j),
Move(e), Agent(e,j), Theme (e,c)

Once a set of facts has been selected, sentence generation is carried out using the GenI (Gardent and Kow, 2007) surface realiser, a generic Feature-Based Lexicalised Tree Adjoining Grammar (FB-LTAG) for French (Gardent, 2008).

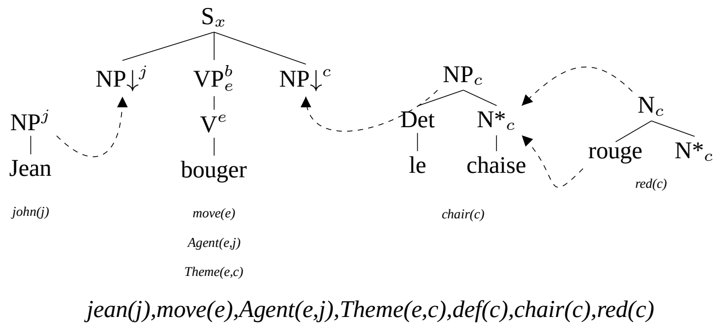


Figure 1: FB-LTAG trees for "Jean bouge la chaise rouge (John moves the red chair)"

Figure 1 shows a toy grammar[1] illustrating the FB-LTAG trees used to generate the sentence *John moved the red chair*. In essence, the grammar consists of a set of trees, each anchored with a lemma and a semantics. To keep the grammar size manageable however, tree schemas are stored separately and a **syntactic lexicon** is used to indicate which lemmas can anchor which tree or family (set) of trees. For instance, the lexical entry for the verb *bouger (to move)* shown below indicates that *bouger* anchors all the trees in the n0Vn1 family that is the set of trees describing the syntactic contexts in which a transitive verb can occur. In addition, the semantics of each lemma is specified using parameterised macros which expands to the appropriate semantics. Thus the parameterised macros Binary_Relation[Move] indicates that the tree anchored by *bouger* is associated with a binary relation with predicate `Move`. In other words, the semantics of *bouger* is {`Move(E),Agent(E,X),Theme(E,Y)`}.

Lemma: bouger
Category: verb
Semantics: Binary_Relation[Move]
Trees: n0Vn1

During generation, trees that are associated with a semantics matching one of the selected facts are retrieved and combined using the grammar tree combination operations. The yields of these trees then give the lists of lemmas making up sentences verbalising the input. For instance, the yield of the tree derived by combining the trees shown in Figure 1 produces the list of lemmas *Jean bouger le chaise rouge*.

To generate a well formed sentence, a **morphological lexicon** is looked up and the appropriate word forms are extracted taking into account the morphosyntactic constraints encoded by the grammar for instance, that the subject must agree in person and number with the verb and similarly, that determiner, noun and adjective must agree in gender and number. The morphological lexicon associates each word form with a lemma and a set of (feature,value) pair describing its morphological properties. For instance, the entries for *chaise* are:

chaise	chaise	{cat:n,num:sg,gender:f,mass:-}
chaises	chaise	{cat:n,num:pl,gender:f,mass:-}

3. Annotating the 3D World with Semantic Annotations

We now describe the semantic annotations we use to describe the virtual world and the general principles underlying these annotations.

3.1. Representing Knowledge using OWL

To annotate the virtual environment with semantic information, we use the Web Ontology Language (OWL, (Horrocks et al., 2003)). As illustrated in the preceding section, we implement both an A-Box and a T-Box. We produce an axiomatic description of concepts in the domain (T-Box). In particular, we place domain concepts in a concept hirarchy where higher-levels are taken from WordNet hypernym structure. For instance, SubClassOf(Chair Piece_Furniture) SubClassOf(Piece_Furniture Artifact). We then describe each object in the virtual world by linking it to an instance in the A-Box and associating this instance with a set of facts describing this object. As a result, each object in the I-FLEG world is associated both with a set of assertions specific to that object and with a set of ontological axioms providing additional information about the class of objects this object belongs to.

3.2. Upper model for content selection

As mentioned in Section 2., the I-FLEG natural language generator needs to check for certain properties in the input. For instance, if the generated output must contain an adjective, then generation must check that the selected content contains a concept that is a subconcept of the $Adjective$ concept. In other words, in our generation task, the communicative goal the generator must realise includes constraints on the syntactic form of the generated sentence. To account for such constraints and support "form driven sentence generation" (that is, generation whose communicative goal includes formal constraints), the I-FLEG ontology includes an *upper model* (Bateman, 1990) i.e., a linguistic ontology capturing how the grammar and/or semantics of a particular natural language carves up the world. The I-FLEG upper model associates domain concepts with linguistic concepts e.g. part-of-speech and predicate argument structure. For instance, MOVE is a

[1] The grammar has been simplified. In reality, the grammar has separate trees for Nouns and for determiners (rather than a single tree including boht the determiner and the noun.

sub-concept of TRANSITIVE. This extension is based on the information contained in the lexicon and permits constraining content selection based on the teaching goal.

3.3. Representing events

OWL does not support ternary relations. However, in natural language, events can be talked about which include three or more participants. For instance, the event described by the sentence “The player puts the glass on the table” involves three participants namely, the player, the glass and the table. To circumvent OWL limitations, we reify events as concepts and linked them to event participants using thematic role relations. More generally, as argued in (Davidson, 1967) and (Franconi, 1994), event reification permits describing modification of eventualities and facilitates the representation of verbs in description logic.

4. Using the LIR standard to link conceptual and linguistic information

As noted in Section 2., to support natural language generation, concepts in the ontology must be linked to words in a lexicon. In addition, this lexicon must provide detailed linguistic information about words and relate these to the grammar the generator makes use of. To represent lexical information and link concepts in the ontology to words in the lexicon, we draw on the LIR model (Peters et al., 2009), (Montiel-Ponsoda et al., 2008) which associates multilingual information with ontologies and is interoperable with several other standards: the Terminological Markup Framework, the Lexical Markup Framework (The LMF Working Group, 2008) and the Multilingual Lexical Information Framework (The MLIF Working Group, 2010). We then derive from this lexicon, the **syntactic** and **morphological** lexicons used by our generator.

4.1. The LIR model

In essence, the LIR model associates a `LexicalEntry` class to the classes `Language` (the language to which the word being described belongs), `Lexicalization` (the word base form) and `Sense` (a definition as in a classic dictionary or a WordNet gloss); lexical semantic equivalences can be established among lexical entries within the same (`hasSynonym`) or different languages (`hasTranslation`); and ontology elements are linked to LIR lexical entries by the `hasLexicalEntry` property. The LIR model is implemented as an OWL ontology so that inconsistencies e.g., between linguistic data and domain knowledge can be detected using OWL reasoning tools. Figure 2 sketches the implementation of the LIR model used in our application.

4.2. Extending the LIR model

To encode the mapping between the ontological and the lexical knowledge used by I-FLEG, we extended the LIR model in two main ways.

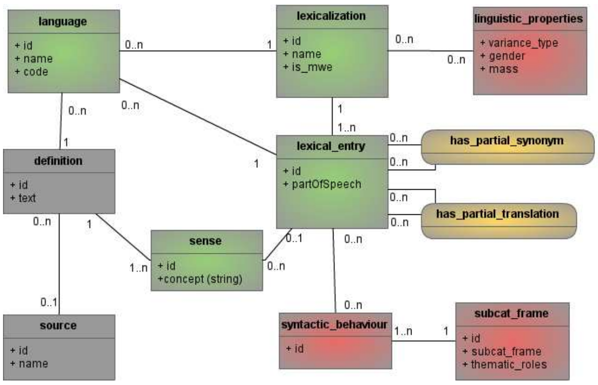


Figure 2: LIR model as used in our application. The green, gray and orange figures are present in the original LIR model and the red boxes represent extensions. The gray boxes are used in the original LIR model, but not in our application and the orange boxes are implemented, but not yet used.

(a) Lexical entry linked to concept E_MOVE

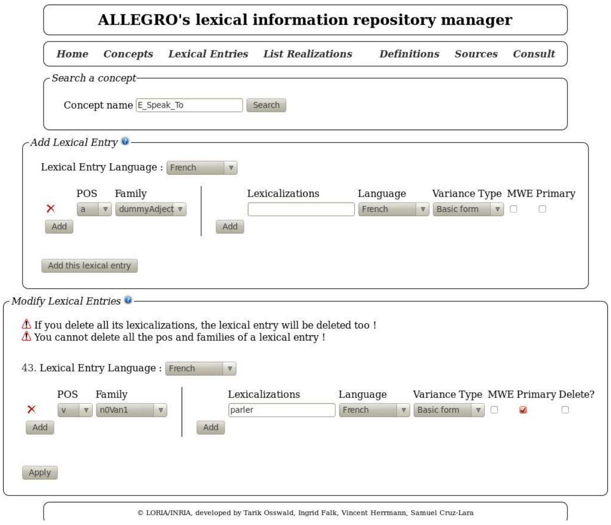
Sense E_MOVE
Language French
POS v
primary yes
Lexicalization bouger
Language French
morph_type 1st group
subcat_frame n0Vn1
semantics Binary_Relation

(b) Lexical entry linked to concept CHAIR

Sense CHAIR
Language French
POS n
primary yes
Lexicalization chaise
Language French
gender f
mass/countable countable
subcat_frame noun
semantics BasicProperty

Figure 3: Sample LIR lexical entries for a verb (a) and a noun (b).

First, we integrated additional syntactic and semantic information about words so as to enable the automatic derivation of the lexicons required by generation. These extensions to the LIR model are represented in Figure 2 by the red boxes *syntactic_behaviour* and *subcat_frame*. In particular, we included in our model information about the semantic type of a word, about its morphology (e.g., whether the inflection of a verb or of a noun follows a regular schema and if so which) and about semantic distinctions such as the mass/count distinction for nouns or the state/event distinction for verbs. In our implementation of the LIR model this information is currently represented as properties of *Lexicalization* instances (red box *linguistic_properties* in Figure 2). Figure 3 sketches sample lexical entries for a verb (Figure 2a) and a noun (Figure 2b). It shows that, for example, the verb *bouger* is a lexicalisation of a lexical entry associated to the concept E_MOVE with syntactic behaviour described by the LTAG family name *n0Vn1* in field *subcat_frame* and its semantic type is shown in field *semantics*: *bouger* is used in a binary relation (Binary_Relation). The verb’s morphology is further specified as a linguistic property (*linguistic_property* class) of the lexicalization: The *morph_type* property shows it is a regular, 1st group verb.



Concept name E_Speak_To

Lexical Entry Language : French

	POS	Family	Lexicalizations	Language	Variance Type	MWE	Primary
✕	a	dummyAdject		French	Basic form	□	□

⚠ If you delete all its lexicalizations, the lexical entry will be deleted too !
⚠ You cannot delete all the pos and families of a lexical entry !

43. Lexical Entry Language : French

	POS	Family	Lexicalizations	Language	Variance Type	MWE	Primary	Delete?
✕	v	n0Van1	parler	French	Basic form	□	☑	□

© LORIA/INRIA, developed by Tarik Osswald, Ingrid Falk, Vincent Herrmann, Samuel Cruz-Lara

Figure 4: Screenshot of web interface for editing lexical entries.

Second, we implemented the lexicon as a database rather than an OWL ontology. In this way, we created a “lightweight” model of the relation between lexicon and ontology which was easier to handle while still keeping domain knowledge and linguistic data separated. Figure 4 shows a screenshot of the web interface for creating and editing lexical entries.

From the ontology and the data-base just described, we then automatically derive the syntactic lexicon used by the sentence realiser (cf. Section 2.). The data-base provides the required (morpho-)syntactic distinctions while the ontology provides the concept names required to inform the semantic field of lexical entries.

4.3. Some features of our extension to LIR

Adopting the LIR model as an intermediate, generic resource mediating between the knowledge base and the specific lexicons used by the generator, allows for a straightforward account of morphology, of synonymy, of multi-word expressions and of multilingualism. We now briefly exemplify each of these points.

Morphology The morphological lexicon can be derived from lexical entries provided that they include morphological features such as *morph_type* or *gender*. The lexicon can be obtained by stemming the lemma and deriving the inflected forms thanks to the morphological features. The irregular cases such as the well known *choux, bijoux, joujoux,...* which lists the few words in French that are inflected with an *-x* in plural form instead of the regular *-s* inflection, can be managed with dedicated morphological features.

Synonymy Synonymy in the original LIR is handled by explicitely stating that two lexical entries are related with a `hasSynonym` relation. However, in our “lightweight” model, this relationship is implicit. Two different lexical entries can be related to the same concept. For instance, the synonymy between couch and sofa can be represented by a single concept COUCH related both to a *couch* lexical entry and to a *sofa* lexical entry.

Multi-Word Expressions Multi-word expressions range from fixed expressions (*in short*, *by and large*), semi-fixed expressions (*spill the beans*, *kick the bucket*), and syntactically flexible expressions (*break up*, *make a mistake*) (Sag et al., 2001). Our framework can handle these three categories. Fixed expressions are considered as words-with-spaces in the lexicalization entry and thus can anchor syntactic trees as whole units. For instance *machine à café* (*coffee machine*) is a single lexicalization unit. Semi-fixed and flexible expressions are dealt with a primary anchor and co-anchors in the syntactic trees, separated by an underscore in the lexicalization entry. For instance the concept of HOUSEMOVING is related to a multi-word expression *déménagement_faire* in which *déménagement* is the primary anchor and *faire* is the verbal co-anchor. By convention, the first word is the primary anchor, the other anchors are ordered as how they appear in the canonical tree of the syntactic family associated to the lexical entry. Hence the entry *déménagement_faire* can be realized as *faire un déménagement*, literally *to do a house moving*. Note that thanks to the synonymy, we can also realize the concept of HOUSEMOVING with the single verb *déménager* (*to move out*).

Translation Multilinguality is supported natively in LIR by means of the `Language` attribute of lexical entries. The same concept, for instance CAT can be linked to an English lexical entry *cat* or to a French lexical entry *chat*. Moreover, it is possible to include compound concepts as senses, for instance, the complex concept CAT ⊓ YOUNG can be linked to the lexical entry *gatito*, the Spanish word for kitten. Vice versa, a primitive concept such as GATITO in a Spanish ontology, can be related to the multi-word lexical entry *cat_young*, where *cat* is the primary anchor and *young* is the co-anchor. This lexical entry can then be associated to syntactic trees realizing it as *young cat*.

5. Conclusion

Semantic annotations of 3D worlds open the door for intelligent simulations and powerful 3D content retrieval and management. In this paper, we present a novel way to exploit these annotations namely to generate text about a virtual world. Using existing standards and best practice principles, we showed how to link semantic annotations not only to the 3D objects but also to words and to the linguistic information required by text generation. As mentioned above, the particular standard used to link ontology and lexicon aims to associate multilingual information with ontologies and is interoperable with several other standards thereby facilitating the integration of other types of information. An obvious interesting extension of the current I-FLEG approach is therefore the portability of the I-FLEG game to English: Does the LIR standard permits an easy adaptation of the I-FLEG game to English learning? Another question we are currently investigating is how to semi-automate the creation of the I-FLEG ontology. We aim to facilitate the authoring of the ontological information required to generate sentence and thereby the automatic creation of situated language learning exercises.

Acknowledgments

The research presented in this paper was partially funded by the European Fund for Regional Development within the framework of the INTERREG IV A Allegro Project.

6. References

M. Amoia, C. Gardent, and L. Perez-Beltrachini. 2011. A serious game for second language acquisition. In *CSEDU 2001*, Noordwijkerout, The Netherlands.

J. A. Bateman. 1990. Upper modeling: A general organization of knowledge for natural language processing. In *1989, USC/Information Sciences Institute*.

D. Davidson, 1967. *The Logic of Decision and Action*, chapter The Logical Form of Action Sentences. Pittsburgh: University of Pittsburgh Press.

E. Franconi. 1994. Description logics for natural language processing.

C. Gardent and E. Kow. 2007. A symbolic approach to near-deterministic surface realisation using tree adjoining grammar. In *ACL07*.

C. Gardent. 2008. Integrating a unification-based semantics in a large scale lexicalised tree adjoining grammar for french. In *COLING'08*, Manchester, UK.

V. Geroimenko and C. Chen. 2005. *Visualizing the Semantic Web: XML-based Internet and Information Visualization*. Springer.

I. Horrocks, PF Patel-Schneider, and F. van Harmelen. 2003. From shiq and rdf to owl: The making of a web ontology language. *Journal of Web Semantics*, pages 7–26.

J. Ibanez-Martinez and D. Mata. 2006. Virtual environments and semantics. *European Journal for the Informatics Professional*, 7(2).

P. Kapahnke, P. Liedtke, S. Nesbigall, S. Warwas, and M. Klusch. 2010. Isreal: An open platform for semantic-based 3d simulations in the 3d internet. In *ISWC 2010*.

E. Montiel-Ponsoda, G. Aguado de Cea, A. Gómez-Pérez, and W. Peters. 2008. Modelling multilinguality in ontologies. In *COLING (Posters)*.

W. Peters, M. Espinoza, E. Montiel-Ponsoda, and M. Sini. 2009. Multilingual and localization support for ontologies. Technical report, D2.4.3 NeOn Project Deliverable.

I. A. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. 2001. Multiword expressions: A pain in the neck for NLP. In *Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15.

The LMF Working Group. 2008. Language Resource Management - Lexical Markup Framework (LMF). Technical report, ISO/TC 37/SC 4 N453 (N330 Rev. 16).

The MLIF Working Group. 2010. MultiLingual Information Framework. Technical report, ISO/DIS 24616.