ULiS: An Expert System on Linguistics to Support Multilingual Management of Interlingual Knowledge bases

Maxime Lefrançois
INRIA Sophia Antipolis – Edelweiss
2004 rt des Lucioles, BP93,
Sophia Antipolis, 06902, France
Maxime.Lefrancois@inria.fr

Fabien Gandon
INRIA Sophia Antipolis – Edelweiss
2004 rt des Lucioles, BP93,
Sophia Antipolis, 06902, France
Fabien.Gandon@inria.fr

Abstract

We are interested in bridging the world of natural language and the world of the semantic web in particular to support multilingual access to the web of data, and multilingual management of interlingual knowledge bases. In this paper we introduce the ULiS project, that aims at designing a pivot-based NLP technique called *Universal Linguistic System*, 100% using the semantic web formalisms, and being compliant with the Meaning-Text theory. Through the ULiS, a user could interact with an *Interlingual Knowledge base* (IKB) in controlled natural language. Linguistic resources themselves are part of a specific IKB: *The Universal Lexical Knowledge base* (ULK), so that actors may enhance their controlled natural language, through requests in controlled natural language. In this paper we propose a basic interaction scenario at the system level, and then we propose and overview the layered architecture of ULiS: meta-ontology, ontology, facts; and ontology, interlingual knowledge, situational knowledge.

1 Introduction

The recently begun ULiS project has first been introduced by Lefrançois and Gandon (2011). It aims at redesigning a pivot-based NLP technique, 100% using the semantic web formalisms, and being compliant with the Meaning-Text theory. The authors envision a *Universal Linguistic System* (ULiS), through which multiple actors could interact with a structured set of knowledge, called an I*nterlingual Knowledge base* (IKB) in multiple controlled (i.e., restricted and formal) natural languages. Each controlled natural language (dictionary, grammar) is to be described in a part of a U*niversal Linguistic Knowledge base* (ULK). Besides this, the ULK consists in one specific interlingual knowledge base. Actors could then enhance their controlled natural language through different actions in controlled natural language (e.g., create, describe, modify, merge, or delete lexical units in the dictionaries and grammar rules; connect situational lexical units to interlingual lexical units; add linguistic attributes with their associated rules, etc.)

The aim of this paper is to overview a proposal for the architecture of ULiS. The rest of this paper is organized as follows: section 2 gives an overview of the related works, and the linguistic theory on which the ULiS relies; section 3 presents the basic interaction scenario at the system level; and section 4 sketches a proposal for a layered architecture for ULiS, with its different modules.

2 Related Work

2.1 The Meaning-Text Theory (MTT)

The MTT is a theoretical linguistic framework for the construction of models of natural language. As such, its goal is to write systems of explicit rules that express the correspondence between meanings and texts (or sounds) in various languages (Kahane, 2003). Meanings correspond to Chomsky's conceptual-intensional level, and texts correspond to Chomsky's articulatory-

perceptual level. Contrary to the Chomskyan approach, five intermediary levels of linguistic representation are supposed for each set of synonymous utterances. The seven levels are namely: a semantic representation that is a network; the deep and surface syntactic representations (DSynR and SSynR) that are trees; the deep and surface morphological representations (DMorphR and SMorphR) that are lists of annotated tokens; and the deep and surface phonological representations (DPhonR and SPhonR) that are also lists of annotated tokens. (Mel'čuk, 1998). Thus, twelve modules containing transformation rules are used to transcribe representations of a level into representations of an adjacent level. The main constituent of the MTT is the dictionary model where lexical units are described, which is called the *Explanatory Combinatorial Dictionary* (ECD), and has been the object of many works on lexical functions, e.g., (Mel'čuk et. al., 1995).

2.2 Lexical ontologies and meaning representation languages

Lexical ontologies, i.e., ontologies of lexicalized concepts, are widely used to model lexical semantics. There exist many of them. Some have broad coverage but shallow treatment (i.e., with no or little axiomatization) such as Princeton WordNet (e.g., Miller et al., 1990), EuroWordNet (Vossen, 1998), and some have small coverage but are highly axiomatized such as SUMO (Niles & Pease, 2001), DOLCE (Gangemi et al., 2002), Mikrokosmos (Nirenburg et al., 1996), HowNet / E-HowNet (Dong & Dong, 2006), FrameNet (Baker et al. 1998). They use different theories of lexical semantics but most of them do not describe phrasemes nor lexical collocations. The French Lexical Network (Lux-Pogodalla & Polguère, 2011) is a growing ECD-compliant lexical resource, but it does not use the semantic web formalisms, and the definitions of the lexical units are not fully formalized.

On the other hand, the Universal Networking Language (UNL) is a meaning representation language, originally designed for pivot techniques Machine Translation. It uses an interlingual lexical ontology based on so-called Universal Words ++, but the lack of argument frames and lexical functions in the UNL dictionary was pointed out in (Boguslavsky, 2002; Boguslavsky, 2005). This is when the idea of an ECD-compliant interlingual lexical ontology was first mentioned. After the semantic web formalisms were introduced at the W3C, an attempt to port the UNL to semantic web formalisms was the topic of a W3C Common Web Language Incubator Group (XGR-CWL, 2008), but no improvement was made to the lexical ontology.

2.3 Collaborative multilingual construction of ontologies

Information systems have been transformed by the integration of web technologies. Beyond the unification of exchange formats and access methods, these web technologies increased tenfold the social dimension of their usage. As numerous communities spring and are assisted by web applications, the interactions of their members generate knowledge bases in which resources are collected and described.

The process of collecting, structuring and maintaining a knowledge base is difficult and costly, particularly as its size, its complexity and the number of actors grows. Consequently, some research works focus on methods for collaborative construction of ontologies or thesaurus (Farquhar et. al., 1996), (Mark et al., 2002), (Fernández, 2006) (Blay-Fornarino et al., 2002). However, new interaction modes offered by the latest web evolutions have opened the way for new scenarios and new usages (Limpens, 2010).

On top of that, emerging technologies open the way for natural interaction with the user (we notably think about natural language interaction), and for the internationalization of contents (representation of contents in an interlingua interpretable by computers and acting as an interface with different natural languages).

2.4 SPARQL Inferencing Notation (SPIN)

Grammar rules are not part of the Common Web Language (CWL) framework, in fact, the construction of grammar modules may be done in any programming language. Knublauch et. al. (2011) introduced SPIN: an RDFS schema to represent SPARQL rules and constraints on Semantic Web models. Using SPIN, one can represent a whole set of SPARQL rules and constraints in the model, and annotate them. A knowledge base in RDF may thus contain vocabularies, facts and SPARQL rules/requests.

2.5 Positioning of the ULiS project

The lexical resource we propose to develop is an interlingual lexical ontology coupled with a situational lexical ontology (the situation of a feature is a generalization of the language in which this feature appears, c.f., section 4.3), both using semantic web formalisms, and that together form an ECD-compliant dictionary. Benefits of using semantic web formalisms are high as it enables us to construct an axiomatized graph-representation of a lexical ontology, with validation and inference rules. Using SPIN, we propose to include transformation rules directly in an RDF format, on top of the ECD-compliant lexical ontologies, thus obtaining an expert system on linguistics.

The ULiS model is somehow similar to the FunGramKB (Periñán-Pascual & Arcas-Túnez, 2010) which is a lexico-conceptual knowledge base for NLP. However, the two projects have different inspiring influence. We choose to comply with the Meaning-Text theory, which gives a thorough understanding of lexical functions that are ubiquitous in every natural language. We also choose to describe the whole ULiS with the semantic web formalisms: we propose to include transformation rules directly in an RDF format, on top of the ECD-compliant lexical ontologies, thus obtaining an expert system on linguistics. This potentially enables the enhancement of the system itself through controlled natural language interactions.

3 Basic Interaction Scenarios with the ULiS

The three basic scenarios of ULiS are illustrated on Figure 1 below.

An actor in a situation *c* inputs some utterance (e.g., in English: "Who killed Mary?") that is first transformed into an RDF situational representation, which undergoes different language-specific process, and which is finally transformed into a CWL-like interlingual representation.

3.1 Machine translation

At this stage, depending on the context, the interlingual representation of the utterance may be translated into another utterance in situation *d* (e.g., in the French situation: "Qui a tué Mary?") through a situational representation (Output1^{TEXT} on Figure 1).

3.2 Management of Interlingual Knowledge Bases

Another possibility is that the interlingual representation of the utterance is transformed in a SPARQL request that is applied on an *Interlingual Knowledge base* (IKB), which eventually produces an RDF output (e.g., ex:John01). This RDF output is then first transformed into an interlingual representation, then into a situational representation and finally into an output utterance: Output2^{TEXT} on Figure 1 (e.g., "John killed Mary").

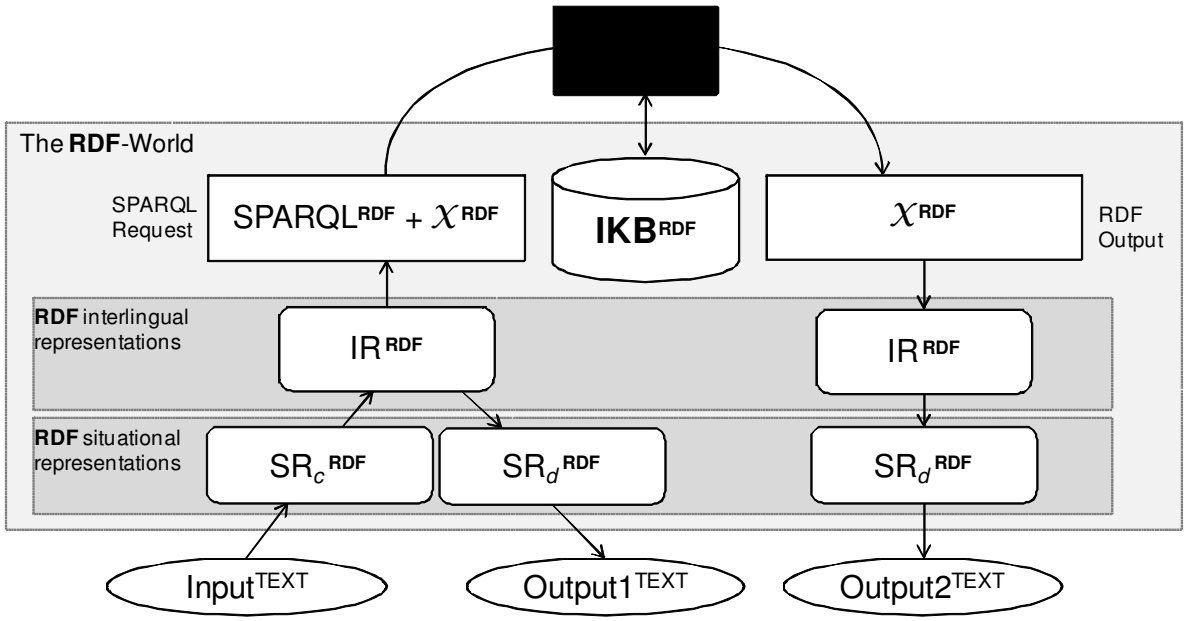

Figure 1. ULiS: The basic interaction scenario with an interlingual knowledge base.

3.3 Management of the Universal Linguistic Knowledge base

Finally, the third scenario is the human-computing scenario: the SPARQL request is applied on the Universal Linguistic Knowledge base, which is the Interlingual Knowledge Base where the whole ULiS is described. Human actors may thus enhance the controlled natural languages through actions stated in controlled natural language.

Thus the interlingual representation format acts as a pivot not only for natural languages, but any interlingual representation may be translated into a SPARQL request, and any RDF graph may be translated to an interlingual representation.

4 The ULiS components

4.1 Overview

Figure 2 illustrates the ULiS, with its three different layers:

The second row represents the interlingual layer (section 4.2), with a meta-ontology that describes the *interlingual lexical ontology* (ILexicOn): the cornerstone of the whole *Universal Linguistic Knowledge base*. The ILexicOn enables inference in *interlingual semantic representations* (ISemRs, on the right).

The first row represents the *interlingual lexical knowledge base* (IKB) layer, with facts (on the right) and an ontology or thesaurus (on the left), augmented with anchors and transformation rules (section 4.4), that enable the transformation of facts into ISemRs, and vice versa. The IKB enables situation-independent inference on utterance representation.

The third row represents the situational layer (section 4.3) with a meta-ontology that describes the *situational lexical ontology* (SLexicOn), that itself enables situation-dependent linguistic inference on utterances' situation-dependent representations (*Situational representations*, SRs, on the right). Situation-annotated links and transformation rules define transformation of utterances among SRs.

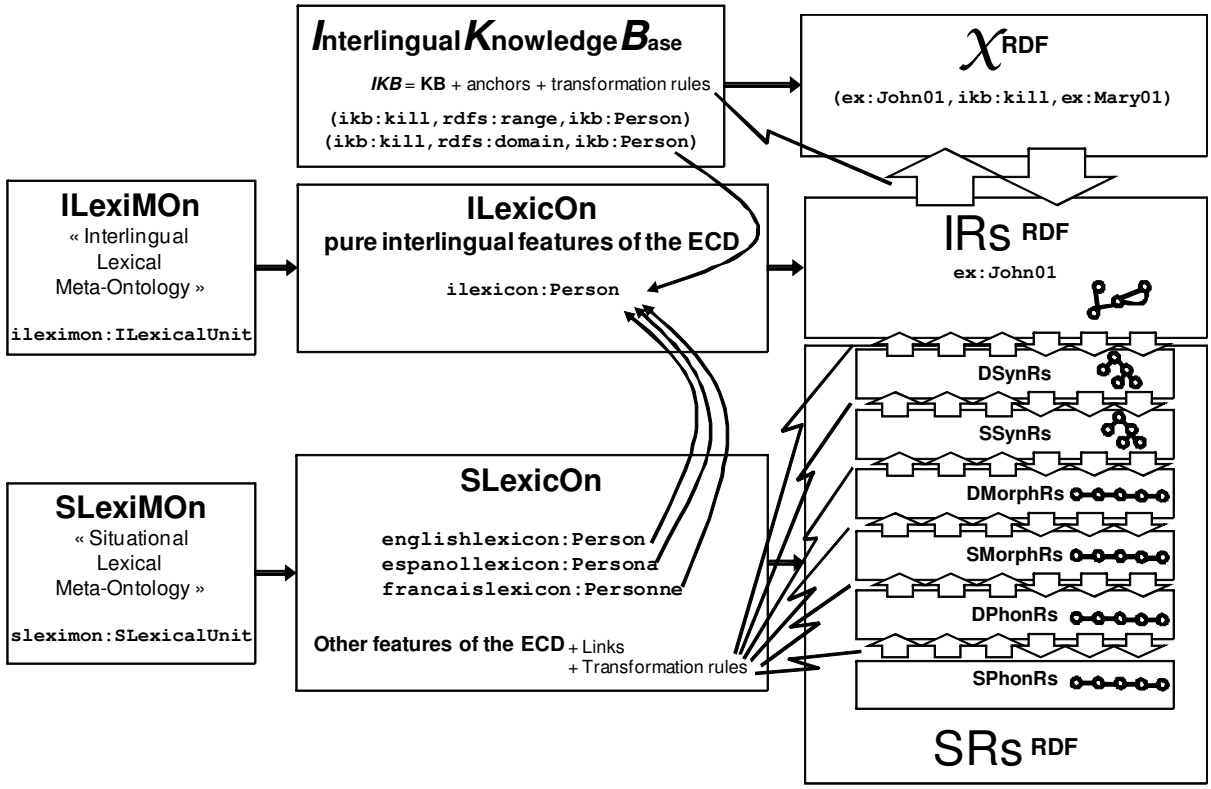

Figure 2. Overview of the architecture of the ULiS.
From top to bottom: the interlingual layer, the interlingual layer, the situational layer.
From left to right: meta-ontologies; ontologies; facts and different representations.

4.2 Architecture in the Interlingual Layer

The pivot module of ULiS is partly described in (Lefrançois & Gandon, 2011). It is divided in three components:

The meta-ontology

The *interlingual lexical meta-ontology* (ILexiMOn) is the schema that the ILexicOn must satisfy to be compliant with the pure semantic features of the Explanatory Combinatorial Dictionary (ECD). It defines meta-classes such as `ileximon:ILexicalUnit`, uses RDFS and some of OWL full's axioms, and contains *ad hoc* SPIN validation and inference rules for the ILexicOn and the *interlingual semantic representations* (ISemRs).

The ontology

The *interlingual lexical ontology* (ILexicOn) is the interlingual dictionary where *interlingual lexical unit classes* (ILU^{c}s) such as `ilexicon:Person` are formally defined as instances of `ileximon:ILexicalUnit`. The ILexicOn contains all the pure semantic features of the *Explanatory Combinatorial Dictionary* (ECD). Any concept expressible in a natural language or a jargon could be defined in the ILexicOn that contains:

- The formal definitions of the ILU^{c}s;
- The definitions of *interlingual attribute classes* (IAtts) (e.g., plural, future, 1st person, indefinite, etc.);
- The definitions of the *interlingual semantic relations* (ISemRels), that are used in the formal definitions of the ILU^{c}s and to construct *interlingual semantic representations* (ISemRs) (e.g., agent, object, manner) ;
- Every purely-semantic lexical links such as synonymy, and purely-semantic generic constructions such as ⁽the center of X⁾, or ⁽stop being X⁾.

The interlingual semantic representations

ISemRs are RDF graphs with nodes being *interlingual lexical unit instances* (ILU^{i}s), and arcs being ISemRels. ILU^{i}s may also be instances of IAtts. Arcs are *interlingual semantic relations* (ISemRels), e.g.,

```
ex:kill01 ilexicon:agent ex:John01.
ex:kill01 ilexicon:object ex:Mary01.
```

4.3 To and from Natural Language facts

Situations

Interlingual-based lexical resources consider connecting language specific dictionaries to some interlingual dictionary. We generalize this by using situations (i.e., the situations of understanding and use of some linguistic element).

The situation of a linguistic element is part of the pragmatics of its use: it represents not only the language used (e.g., EN, FR), but also sociolectal marks (e.g., biologists, architects, official, slang, reverential), topolectal marks (e.g., U.S., Canada), chronolectal marks (e.g., old, neologic), and even individual marks (e.g., a particular group of people). The intersection of situations is also a situation (EN-U.S.-slang), and so is the union of situations (FR-Canada OR FR-France-old).

Architecture of the situational layer

This architecture purposefully mirrors the interlingual layer:

A *situational lexical meta-ontology* (SLexiMOn): describes the SLexicOn, with resources such as `slexicon:SLexicalUnit`;

A *situational lexical ontology* (SLexicOn), contains all non-purely semantic features of the ECD such as:

- Definitions of *situational lexical unit classes*, called SLU^{c}s, (e.g., `enlexicon:Person` in the english lexical ontology, `frlexicon:Person` in the french lexical ontology), by means of a link to an ILU^c, which is annotated by a specific situation.
- A realization unit: either a string, or a semantic representation for idioms.
- Lexical functions such as Instr(X), i.e., the preposition that governs the keyword X and means: ⁽by means of⁾. e.g., $Instr_{EN}$(*hands*)=*by ~* ; $Instr_{FR}$(*hands*)= *at [the ~]* ($Instr_{FR}$(*mains*) = *à [la ~]*)
- Connotations, e.g., C_{EN}(*hot air*) = C_{FR}(*wind*) = C_{RU}(*water*) = *nonsense, void*.
- Situational attribute classes (e.g., invariable English nouns, French 1st verb group, German dative, etc.), their associated situations and rules.
- Situational relations: relations that link two instances of the SLU^{c}s, thus defining the dependency syntax of the utterance, or the order of the words in an utterance.

Situational representations (SRs). The data consist of *situational representations* (SRs): RDF graphs having *situational lexical unit instances* (SLU^is) as nodes and situational relations as arcs. A SR thus represents the different representations of the Meaning-Text theory.

Transformation rules

Contrary to the Common Web Language (CWL), where no grammar rules representation is proposed, we plan to introduce *transformation rules* in the SLexiMOn. Transformation rules form a subclass of the SPIN rules and are attached to a SLU^c to define the correspondence between a generic pattern from a representation level, to another pattern at a deeper or at a higher representation level. Thus, each situation may define its own analysis and production grammar, both made of six sets of transformation rules.

Transformation rules may be sorted according to their level of genericity: transformation rules that are attached to ISemRels, or to IAtts, are less specific than rules that may be triggered only when a complex ISemR patterns is met; also, rules that may be triggered in generic situations are less specific than those that may only be triggered in more specific situations. The important point is that a rule must be triggered if and only if there is not a more specific rule that can be triggered instead. We claim that a reasonably small set of rules will suffice to produce and analyze simple controlled natural languages.

4.4 To and from Interlingual Knowledge Bases facts

Interlingual knowledge bases

The main criterion that an interlingual knowledge base must meet is that any RDF graph inside it must be transformable into an *interlingual semantic representation* (ISemR). We thus propose to form interlingual knowledge bases by augmenting classic knowledge bases with *anchors* and *transformation rules*.

Anchors

An anchor is a triple that links an RDF resource to an ILU^c. For instance, the RDF resource `foaf:Person` will be anchored to a specific ILU^c `ilexicon:Person` that formally defines the concept of a person, and that is itself linked to an English SLU^c that is a pluralizable noun, and that is realized by the string "person".

Transformation rules

The transformation rules are stored in the interlingual knowledge base and form two separated sets of rules: one for producing RDF from an ISemR, the other for producing an ISemR from RDF. Here again, transformation rules may be sorted according to their level of genericity, and the most generic rules must be inhibited when more specific ones can be triggered.

Augmenting classic semantic web formalisms

The output of an ISemR must be a valid SPARQL request, and the output of any RDF graph must be a valid ISemR. This criterion will be satisfied by the introduction of different anchors and generic transformation rules in the classic semantic web vocabularies: RDF, then RDFS, OWL and SPIN, and finally SKOS. Thus an RDF class that has no anchor, e.g., `foaf:Person`, has a correspondence with an ISemR that itself has a correspondence to the textual representation for the EN situation: "The class of persons".

5 Conclusion

We introduced a *universal linguistic system* (ULiS) through which multiple actors could interact with an *interlingual knowledge base* (IKB) in controlled natural language. We explained an interaction scenario with ULiS, which can serve for machine translation and for multilingual management of interlingual knowledge bases. We then gave an overview of the architecture of ULiS: the interlingual module; the situational module; and an interlingual knowledge base.

The main novelty of our proposal is that the characteristics of each controlled natural are stored in a specific interlingual knowledge base. Thus, actors could enhance their controlled natural language through the same actions in controlled natural language they use to interact with the knowledge base (e.g., create, describe, modify, merge, or delete lexical units in the dictionaries and grammar rules; create, describe, modify, merge, or delete linguistic attributes with their associated rules, etc.).

The interlingual module of ULiS has already received much attention, and has been described in (Lefrançois & Gandon, 2011). We plan to validate our results by the design and the experimentation of a web-based prototype with a simple interlingual knowledge base (e.g., the

wine ontology) and the two basic situations English and French.

References

Collin F. Baker, Charles J. Fillmore and John B. Lowe, 1998. The berkeley framenet project. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1, pp.16–90, ACL.

Mireille Blay-Fornarino, Anne-Marie Pinna-Dery, Kjeld Schmidt and Pascale Zaraté, 2002. Cooperative Systems Design: A Challenge of the Mobility Age. In: Proceedings of COOP 2002, Saint-Raphaël, France, 4-7 June 2002, IOS Press, pp.23–37.

Igor Boguslavsky, 2002. Some Lexical Issues of UNL. In Proceedings of the First International Work-shop on UNL, other interlinguas and their applications, Las Palmas, pp.19–22.

Igor Boguslavsky, 2005. Some controversial issues of UNL: Linguistic aspects. Research on Com-puter Science, vol. 12, pp.77–100.

Zhendong Dong and Qiang Dong, 2006. HowNet and the Computation of Meaning. World Scientific, London, UK.

Adam Farquhar, Richard Fikes and James Rice, 1996. The Ontolingua Server: a Tool for Collaborative Ontology Construction. In: International Journal of Human-Computers Studies, vol. 46, num. 6, pp. 707–727.

Miriam Fernández, Iván Cantador, and Pablo Castells, 2006. CORE: A Tool for Collaborative Ontology Reuse and Evaluation. In: Proceedings of the 4th Int. Workshop on Evaluation of Ontologies for the Web (EON'06), at the 15th Int. World Wide Web Conference (WWW'06). Edinburgh, UK.

Aldo Gangemi, Nicola Guarino, Claudio Masolo, Alessandro Oltramari and Luc Schneider, 2002. Sweetening ontologies with DOLCE, Knowledge engineering and knowledge management: Ontologies and the semantic Web, pp.223–233, Springer.

Sylvain Kahane, 2003. The Meaning-Text Theory, Dependency and Valency. In: Handbooks of Linguistics and Communication Sciences 25: 1-2, Berlin/NY: De Gruyter, 32 p.

Holger Knublauch, 2011. SPIN - Overview and Motivation, W3C Member Submission 22 February 2011, http://www.w3.org/Submission/spin-overview/

Freddy Limpens, 2010. Multi-points of view enrichment of folksonomies. Ph.D. INRIA Sophia Antipolis, Edelweiss & Computer Science Dpt of Université Nice Sophia Antipolis.

Maxime Lefrançois and Fabien Gandon, 2011. ILexicOn: toward an ECD-compliant interlingual lexical ontology described with semantic web formalisms. In Proceedings of the 5th Meaning-Text Theory, Barcelona (Spain), pp. 155-164

Veronika Lux-Pogodalla, and Alain Polguère, 2011. Construction of a French Lexical Network: Methodological Issues. International Workshop on Lexical Resources.

Gloria Mark, Victor Gonzalez, Marcello Sarini, and Carla Simone, 2002. Reconciling different perspectives: an experiment on technology support for articulation, in (eds) Blay-Fornarino.

Igor A. Mel'čuk, André Clas, Alain Polguère, 1995. Introduction à la lexicologie explicative et combinatoire. Duculot / Aupelf-UREF, Louvain-la-Neuve, BE.

Igor A. Mel'čuk, 1998. The Meaning-Text Approach to the Study of Natural Language and Linguistic Functional Models. [Invited lecture.] In S. Embleton (ed.): LACUS Forum 24, Chapel Hill: LACUS, pp. 3–20.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Ketherine Miller, 1990. Introduction to wordnet: An on-line lexical database. International Journal of lexicography, 3(4):235-344.

Ian Niles and Adam Pease, 2001. Towards a standard upper ontology. In Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001, ACM.

Sergei Nirenburg, Stephen Beale, Kavi Mahesh, Boyan Onyshkevychy, Victor Raskinyy, Evelyne Viegas, Yorick Wilksx, and Remi Zajac, 1996. Lexicons in the Mikrokosmos project. In Proceedings of the Society for Artificial Intelligence and Simulated Behavior Workshop on Multilinguality in the Lexicon, Brighton, UK.

Carlos Periñán Pascual, and Fernando Arcas Túnez, 2010. The architecture of FunGramKB. In Proceedings of the Seventh International Conference on Language Resources and Evaluation, Valeta (Malta), 2667-2674.

Piek Vossen, 1998. EuroWordNet a multilingual database with lexical semantic networks, Computational Linguistics, 25(4).

XGR-CWL, 2008. Report of W3C Incubator Group on Common Web Language, http://www.w3.org/2005/Incubator/cwl/XGR-cwl/