Hybridizing formal and linguistic semantics for the MSW

Aldo Gangemi[1, 2]

[1]Paris13-CNRS-Sorbonne Cité and [2]STLab, ISTC-CNR, Rome
aldo.gangemi@cnr.it

1. Semantics: a serendipitous chaos

The current uptake of "semantic technologies" requires an effort to design some interoperability for the representation practices among fields as diverse as knowledge representation and reasoning (KR), lexical semantics, information extraction, databases, (semantic) Web standards, Web 2.0 folksonomies, etc.

Multilingual linguistic elements, *ontologies*, and *semantics* are key components that are shared by those fields, but are approached in heterogeneous ways. Due to the enormous amount of legacy data and representation practices, we cannot count only on standardisation efforts to build useful applications. In the forthcoming Multilingual Semantic Web (MSW), we need to live with the "serendipitous chaos" that characterises knowledge (and linguistic) management and engineering.

Too rigorous requirements are not sustainable, as the history of the Semantic Web in the last ten years suggests: logical consistency cannot always be enforced, identity of entities is often questionable, data are not always reliable and usually incomplete, knowledge can take many forms, assignment of predicates to objects can be made for unpredictable reasons, and can change dynamically, the intended meaning of predicates cannot even be studied to a full extent, because any two persons can have various levels of competences, and different needs for their interaction with their environments, often entering a dialectic or even conflicting interaction. Even more importantly, data and content are rarely structured in a cognitively sound way, or in a way that is *relevant* to the humans or applications that use them [1].

For those reasons, we have requirements for an agile semantics that (1) overarches the different representation practices, (2) is able to deal with incompleteness and errors, but also (3) assumes cognitive relevance by default.

In everyday life, any sign that we use or perceive (the perception of a segment of the world, an image, a word, a sentence, a scientific handbook, a novel) is not typically interpreted as it is supposed to be according to an ontology, dictionary, or other quasi-normative resources, but as a function of what we can do with it, i.e. as a relevance function, also known as an *affordance* [2]. For MSW this is a very important assumption, because when we envisage applications that are cross-linguistic, they need to work at the level of cognitive relevance, not at that of single, decontextualized data or term equivalences.

A representation language that integrates ontologies and (multi-, cross-)linguistic data needs then to assume that a sign is interpreted (or produced) with an interaction context in mind. In addition, such representation language should be associated with the practices of accessing, reengineering, or refactoring data when used for a certain

adfa, p. 1, 2011.

purpose, e.g. with natural language processing methods, ontology-based data access, etc., including practices of multilingual corpora matching.

My position, which supports a preliminary sketch of FRASL (FRame ASsignment Language) in later sections, is that we need to define a minimal logical backbone (requirements (1)(2)), and to go back to the (relevance-based) cognitive foundations of KR, which was shared in the seventies (then lost) among AI, linguistics, and cognitive science researchers (requirement (3)), and revisit the way we design ontologies and data accordingly, in the MSW perspective.

2. A minimal model of semantic assignment

Inspired by [3][4][5], I assume folksonomies as used on typical Web2.0 applications as bearing the minimal semantic commitment for our problem. As Figure 1 summarizes, we can imagine a double nature of tagging/annotation on the Web, i.e. that *tags* are *assigned* (and *providing access*) to *resources*, so that the *label* used as face value of that tag expresses a *concept*. Also, a shared assumption on the Semantic Web (and annotation semantics in general) is that those concepts are *instantiated* by the annotated resources.

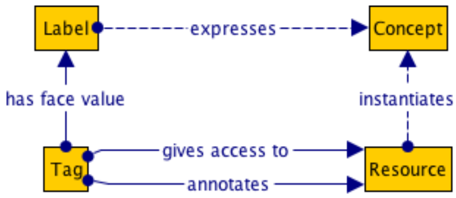

Figure 1: assignment operations and their semantic consequences. Dashed arrows denote the indirect nature of the semantics emerging from assignments.

Of course, there are big differences in labels taken from a folksonomy, extracted from a text, or defined in a formal ontology. The differences are mainly reflected in the way the concept is expected to be interpreted. For example, a label from a Web2.0 tagging action is simply interpreted from the combination of its bare label and the annotated resource(s). A label extracted from text is interpreted also with reference to the text itself, or other text/knowledge known as related to it. Finally, a label from the signature of a formal ontology is interpreted only (or mainly) with respect to its *formal semantics*.

However, despite the differences, the evolution of linked data and semantic applications show that, whatever additional constraints are given in a vocabulary or an ontology, the primary interpretation comes from the *intention of the tagger*, as one can notice from the wild usage of `owl:sameAs`, or the creative reuse of existing vocabularies.

Based on the cognitive semantics hypothesis, the intention of the tagger can be conceived as the relevance function applied in the tagging/annotation action. I call this action *assignment*. Assignments do not require any standpoint on the purely semantic layer: the world of semantics is then accessory, and can be exploited for any added value it can provide besides the basic investigation of assignment actions.

This move frees up the possibility of a KR language that can deal with even purely geometrical accounts of meaning (e.g. from latent semantic analysis, social network analysis, clustering, multi-lingual corpora analysis, etc.), which only work on regularities (patterns) emerging from annotation practices, i.e. devoid of any high-level semantic standpoint.

A notable result is also that *formal* and *linguistic* semantics can be reconciled, provided that they are both grounded in assignments. For example, on one hand the formal interpretation of *hospital* is usually given as the class of 'all' hospitals, but in an assignment-based domain, the class of hospitals is the set of entities that are invariant under certain conditions deriving from compatibility of tagging operations by different agents and with an equivalence class of labels. On the other hand, the linguistic semantics of *hospital* will derive directly from the compatibility of tagging operations, eventually gathering the same grounding as the formal interpretation. An interesting consequence is that within empirically established assignment domains, we can use lexical concepts as formal classes, and vice-versa.

Moreover, my position is that concepts depend on the *relevance function* applied with the assignment. From the hypotheses, relevance functions activate real, fictionary, imagined, or simulated *action (*or more generally *situation) possibilities*. This is what notions like *frame, schema, script*, or *knowledge pattern* typically convey. Frame semantics in this perspective has been reconstructed in [whatsinaschema][towards][cahiers]. The consequence of this position is that whenever we extract or reuse a concept in an assignment scenario, that concept is either a frame (situation type, event type, etc.), or a role of a frame, or a type of a role from a frame. For example, assignment semantics assumes that the label *dog* has only sense in the context of a situation or action where a dog has a role, e.g. *barking* or *chasing*. Any multilingual treatment of *dog* will then need to cope with the contextual binding of that label.

Beaugrande [6] firstly defines "global patterns of knowledge" as a notion encompassing *frames*, *schemas*, *plans*, and *scripts*. Following him, as well as recent work in KR and the Semantic Web [7][1], we call this core notion *knowledge pattern*.

Knowledge patterns seemed appropriate in the seventies to create a positive cross-disciplinary research synergy. KR had a major role in this synergy. Description Logics were among the designs proposed, and for several reasons managed to be a major part of the development of the Semantic Web until nowadays. While DL have been very helpful in understanding the complexity problems behind automated reasoning on frame-like formal languages, they are rather poor when representing sorts like *frame*, *role*, *lexical unit*, *context*, *situation*, etc.

3. FRASL

The proposal that we briefly present here of a FRame ASsignment Language (FRASL), presented fully in [13], derives from previous work (e.g. [8]), but it stands alone in terms of practically covering the wide range of transformations and applications related to the ontology-lexicon interface. FRASL framework has several inspirations, the most evident being Davidson's theory of events [9], Smith's *descriptions* [10], Construction Grammar [11], Discourse Representation Theory [12], etc.

The starting point of FRASL is the *Assignment* relation. An assignment is a semiotic action performed by some *Agent*, during either the production or the interpretation of a discourse fragment, called *Expression*, in an interaction between that agent and its *Environment*, in order to select a *Situation* from the environment.

Frame semantics tries to describe how situations are selected. *Frames* (or knowledge patterns) are situation types featuring roles that are filled by entities of a situation: in this way, situations emerge by filling the role structure of a frame. For example, in the **Cure** frame, a ***healer*** treats an ***affliction*** of a ***patient***, using some ***treatment*** (at some ***time***, ***place***, etc.). If an environment offers entities (e.g. a physician, a medical record, an injured person, and some medicaments) that fill the roles of that frame, we can recognize a *curing* situation within that environment.

In many cases, assignment operations do not provide extensive expressions; i.e. the frame **Cure** can be activated ("evoked") even by the picture of a hospital or a sufferer, the tags *healing* or *emergency* annotating a picture, or a sentence like: *he will undergo radiation treatment*.

A FRASL formula comprises components to represent the elements of assignment operations. For example, this is a template of FRASL components:

(1) **Scope**{"Expression" > (**frame**[***role***:entity(<u>Type</u>), ...]$_{\textit{\textbf{situation}}}$)}

Except scopes, any component can be empty. E.g. this template is almost empty:

(2) **Scope**{(**frame**[])}

For example, sentence (9):

(3) Mustafa said he decided to go alone to Socotra

can be represented as in formula (10):[1]

(4) **Sentence**{"Mustafa said" >
(**say**[***agent***:Mustafa(*x*:<u>Person</u>), ***time***:past(<u>Time</u>), ***sentence***:"he decided" >
(**decide**[***agent***:*x*, ***time***:past, ***sentence***:"to go alone to Socotra" >
(**go**[***agent***:*x*, ***location***:Socotra(<u>Place</u>), ***manner***:**alone** [***agent***:*x*]])])])}

The format of the predicates in (4) reflects that FRASL is a strongly-typed language: besides variables and named entities (individual constants), predicative constants can be sentence types, frames, roles, types, or modal modifiers.

In order to ground FRASL in a formal semantics, we need at least a translation to a many-sorted logic with proposition variables,[2] which gets a formal interpretation from

[1] See [13] for a detailed explanation of the FRASL notation.

[2] The following formula is semantically equivalent to (4): $\exists(x,y,t,z,g,w,p,a)$(say(x,y,t) $\wedge$ agent(*say*,x) $\wedge$ time(*say*,t) $\wedge$ sentence(*say*,y) $\wedge$ Person(x) $\wedge$ x=*Mustafa* $\wedge$ t=past $\wedge$ y="he decided"$\wedge$ expresses(y,z) $\wedge$ g="to go alone to Socotra" $\wedge$ expresses (g,w) $\wedge$ z=(decide(x,w,t)) $\wedge$ w=(go(x,*Socotra*,a)) $\wedge$ agent(*decide*,x) $\wedge$ sentence(*decide*,w) $\wedge$ time(*decide*,t) $\wedge$ agent(*go*,x) $\wedge$ location(*go*,*Socotra*) $\wedge$ Place(*Socotra*) $\wedge$ manner(*go*,a) $\wedge$ a=(alone(x))), with (frames, types, roles): $\mathcal{F}$(say), $\mathcal{F}$(decide), $\mathcal{F}$(go), $\mathcal{T}$(Person), $\mathcal{T}$(Place), $\mathcal{R}$(agent), $\mathcal{R}$(time), $\mathcal{R}$(sentence), $\mathcal{R}$(location), $\mathcal{R}$(manner)s

model theory. Unfortunately, an expressive logic of this kind is not appropriate to the current state-of-art applications of web ontologies. On the other hand, KR for the Semantic Web provides compact and tractable languages with a model-theoretic semantics. The main shortcoming is that the strong typing of FRASL must be reconstructed as "meta-level sugar". As an example, (5-13) encode the first part of (4) as a set of OWL2 axioms:

```
(5)     test:sentence_2 frasl:expression:N "Mustafa"[string]
(6)     test:sentence_2 frasl:expression:VP "said"[string@en]
(7)     test:sentence_2 frasl:evokes say_frame:say
(8)     test:say_1 frasl:occurrenceOf say_frame:say
(9)     test:say_1 say_frame:agent test:Mustafa
(10)    test:say_1 say_frame:sentence test:sentence_2
(11)    test:sentence_2 expression:VP "decided"[string^en]
(12)    test:sentence_2 frasl:evokes decide_frame:decide
(13)    test:decide_1 frasl:occurrenceof decide_frame:decide
```

FRASL can be used to describe very different assignment types, e.g. term extraction:

(14) **TermExtraction**{(**extracts**[***agent***:TermExtractor, ***occurrence*****:**"dog" >
…[...(*x*:<u>Dog</u>)], ***corpus***:BNX, ***relevance***:0.7(<u>float</u>)])}

Term extraction, entity resolution and type induction:

(15) **TermExtraction**{(**extracts**[***agent***:NER+ER+SST, ***occurrence*****:**"Immanuel Kant" >
…[dbpedia:Immanuel_Kant(<u>dbo:Person</u>])}

4. FRED as a FRASL application

FRED[3] [14] is a software tool that makes FRASL concrete and applicable to the rapid extraction of frame structures from text. FRED implements some of the constructs described, in particular it reuses several NLP and KR components in order to produce RDF-OWL triples for either predicative or factual structures. For example, FRED is able to produce the RDF graph depicted in Figure 2, extracted from the sentence:

«The statement by China Foreign Ministry on Friday signaled a possible breakthrough in a diplomatic crisis that has threatened American relations with Beijing.»

For comparison, the complete FRASL representation for that sentence would be:

Sentence{"The statement by China Foreign Ministry on Friday signaled a possible breakthrough in a diplomatic crisis that has threatened American relations with Beijing" >

(**signal**[***agent***(*x*:
statement[***agent***:ChinaForeignMinistry(*y*:<u>Organisation</u>)]), ***time***(*t*:past, *t*=Friday)], ***topic***(*y*:
possibility[***event***(e_1:**breakthrough[*****in***(*z*:<u>**diplomaticCrisis**[</u>***event***(e_2:
threaten[***cause***:*z*, ***experiencer***(*w*:
AmericanRelation[***with***:Beijing(<u>Place</u>)])])])])])])}

Six out of seven frames are detected and represented by FRED (the seventh **possibility** frame requires not yet implemented rules for modality representation).

In addition, FRED provides integration with a named entity recognizer, which resolves one (Beijing) out of two named entities, by linking it to a publicly available multilingual ontology (contextual disambiguation by using inductive classification).

[3] Available at http://wit.istc.cnr.it/stlab-tools/fred

Finally, the conceptual entities extracted can be disambiguated with reference to e.g. WordNet, thus enabling additional conceptual and multilingual interoperability. For example, *statement* can be automatically disambiguated to `wn30:synset-statement-noun-1`, *breakthrough* to `wn30:synset-breakthrough-noun-3`, etc. Disambiguation is also contextual, e.g. with conceptual density or multilingual corpora.

The existence of multilingual ontologies with factual and lexical data (e.g. Wiktionary, DBpedia, WordNet) opens the possibility of rich cross-linguistic queries.

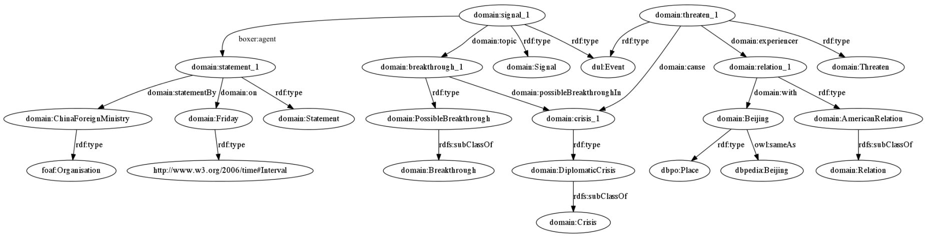

Figure 2: An RDF-OWL graph extracted from the sample sentence by FRED.

References

[1] Gangemi A., Presutti V. A Pattern Science for the Semantic Web, Semantic Web Journal, 1(1), 2010.

[2] Gibson, J. J. The ecological approach to visual perception, Laurence Erlbaum, 1979.

[3] Presutti V., Gangemi A. Identity of Resources and Entities on the Web. International Journal on Semantic Web and Information Systems, 4(2), 2008.

[4] Gruber T. Ontology of Folksonomy: A Mash-up of Apples and Oranges, International Journal on Semantic Web and Information Systems, 3(2), 2007

[5] Abrusci V.M., Romano M. Ontologies, Logic and Interaction: Approaches to Semantic Web ranging from Lexical Semantics to Geometrical Compatibility. Colloque LOCI "Ontologies et Sémantique Lexicale", Roma, 2011.

[6] Beaugrande, R. de. Text, discourse, and process: toward a multidisciplinary science of texts», volume 4 of Advances in discourse processes, ABLEX Pub. Corp., 1980.

[7] Clark P., Thompson J., Porter B. Knowledge Patterns, in Cohn, A.G., et al. (eds.), KR2000, 591-600, Morgan Kaufmann, 2010.

[8] Gangemi A. What's in a Schema?, in Huang, C., Calzolari, N., Gangemi, A., Lenci, A., Oltramari, A., and Prevot, L. (eds.), Ontology and the Lexicon. Cambridge UP, 2010.

[9] Davidson, D. The Logical Form of Action Sentences», in The Logic of Decision and Action (2nd ed.). Pittsburgh: University of Pittsburgh Press, 1968.

[10] Smith, B.C. Levels, layers, and planes. The framework for a theory of knowledge representation semantics. MS Thesis, MIT, 1978.

[11] Fillmore, C., Kay, P., and O'Connor, C. Regularity and Idiomaticity in Grammatical Constructions: The Case of let alone, Language, 64, 501–538 (1988).

[12] Kamp H. A Theory of Truth and Semantic Representation, in J. Groenendijk et al. (eds.). Formal Methods in the Study of Language. Amsterdam: Mathematics Center, 1981

[13] Gangemi A. Back to the Future: Frame Representation and Semantic Technologies. Cahiers de Lexicologie, 2(99), 2011.

[14] Presutti V., Draicchio F., Gangemi A. Knowledge extraction based on Discourse Representation Theory and Linguistic Frames. A. ten Teije and J. Völker (eds.): EKAW2012, LNCS, Springer, 2012.