The Multilingual Procedural Semantic Web
A Position Paper

Sergei Nirenburg and Marjorie McShane
University of Maryland Baltimore County
sergei@umbc.edu, marge@umbc.edu

The stated goal of the Semantic Web community is to turn the Web into a richly annotated resource, making its content more amenable to applications that involve machine reasoning. The most widely discussed language-oriented aspect of this vision involves the creation and use of an inventory of markup tags that indicate select semantic types. So, the "semantics" of the Semantic Web is not the semantics of full texts or even full sentences, but rather of select elements of text and extra-textual information. Moreover, the annotations are expected to be largely carried out manually, so broad coverage is unlikely, as are consistency and universal public-spiritedness on the part of annotators (cf. Doctorow, no date). Compare this to the ideal semantic web, which would be *automatically* generated from the unadorned web by processors that would carry out lexical disambiguation, referential disambiguation, and the interpretation of textual implicatures, such as the recognition of irony and indirect speech acts. Such full semantic interpretations of web content would serve as optimal input for machine reasoners.

It is common practice in the field of AI to assume the availability of such knowledge structures – in fact, practically all work on machine reasoning over the past decades has used hand-crafted, complete, unambiguous knowledge structures as input. *How* that could be achieved automatically was always considered a separate issue, delegated to the NLP community. The NLP community, however, by and large abandoned the task of deep semantic analysis some 20 years ago, opting to pursue either (a) knowledge lean, "low-hanging fruit" tasks that contribute to the configuration of better NLP applications in the near term but do not contribute to the ultimate goal of automatic text understanding or (b) method-oriented work, in which the methods themselves are of first priority and natural language serves primarily as a source of data sets.[1]

The Semantic Web community has largely followed the spirit of the NLP majority by deeming full semantics to be too complex to be pursed. As such, the semantics of the Semantic Web is effectively constrained to selective annotation of text strings in ways that are considered feasible in the short term. The preferences of the Semantic Web community are reflected in

[1] Space does not permit a full motivation for these generalizations. For that see Nirenburg and McShane, forthcoming as well as the historical references cited therein.

the selection of foci of work: the development of formal standards, metadata tag sets, ontologies to be used as the content of tag sets, and so on. While we appreciate the common preferences of the mainstream NLP and Semantic-Web communities, and while the material below describes an attempt to contribute to the near-term gains they seek, our contributions must be framed within the research paradigm that we deem the most promising for the long-term utility of any NLP, be it for the web or any other corpus: computational deep-semantic processing. We will argue that one near-term results can be achieved within a theory and methodology that seek full understanding of texts, along with associated sophisticated behaviors, by intelligent agents.

Our research program is an outgrowth of the theory of Ontological Semantics, which studies the processes of automatically extracting, representing and manipulating meaning in natural language texts. Analysis by the OntoSem text analyzer pursues all of the desiderata listed in the introductory paragraph, seeking fully specified, unambiguous, ontologically grounded meaning representations that are more amenable to machine reasoning than highly ambiguous natural language texts (Nirenburg and Raskin 2004). Of course, the automatically generated structures are not yet perfect, as that would be well beyond the current state of the art. However, we are making direct progress toward this goal, which suggests that the vision of *fully interpreted content* delivered over the internet should not be neglected. A prototype for this vision was demonstrated in the implemented SemNews application (Java et al. 2007), which took web-delivered news feeds as input and generated semantic interpretations of them represented as RTF structures.

Significantly, Ontological Semantics is a language-independent theory, most of whose knowledge bases (e.g., ontology, fact repository, rule sets for agent decision-making) and reasoning engines are language-independent. In fact, in the intentionally provocatively titled "An NLP lexicon as a largely language independent resource" (McShane et al. 2005), we describe how much of the information found even in the lexicons used to support OntoSem language processing can be directly reused across languages (more on this below). Once the input strings from any language have been interpreted using a battery of processors, the resulting text-meaning representations can be reasoned over by a single set of engines. Language-neutrality offers not only great savings in time for the acquisition of knowledge resources and development of processors, it also offers consistency of processing across languages.

The core point of this statement, which follows basic tenets of configuring intelligent agents within the OntoAgent environment, is as follows. *The only realistic way to enhance the Web with useful semantic annotations is automatically*. Semantic analysis is, by its very nature, procedural: a system – hereafter "agent" – receives some input, analyzes it in context, and generates an interpretation. The component functions of this

process, like all functions, are subject to error; as a result, the agent must be able to evaluate its *confidence* in the function's output based on the overall predictive power of the function as well as the confidence in each input parameter value. Depending upon the calculated confidence in output, the agent can decide whether or not to use the output in a given application. Since many of the actual functions used to generate interpretations are identical (or at least very similar) cross-linguistically, they should be reused to support both efficiency and consistency in the treatment of Web content. Since different functions take different types of parameter values as input – and since some parameter values are quite easy to compute with high confidence while others are much more difficult – it is possible to introduce procedural semantic analyses to web content in a progressive manner, over time.

We will now illustrate how automatic annotations, generated using cross-linguistically applicable functions, could be incorporated into the Semantic Web over time. We will use as sample phenomena so-called indexical expressions, which are strings whose absolute meaning can be understood only with reference to a specific context: e.g., *he, themselves, over there, now, in a few minutes, the preceding paragraph*. The reason why one would want all these indexical expressions fully, locally resolved as annotations to Semantic Web content should be self-explanatory: it is more directly useful to an automatic reasoner to have access to the information "John. W. Lacey III of Kansas City, Kansas died on July 5, 1974 in Washington, D.C. from complications of heart disease" rather than an expression that could be synonymous given the right context: "Yesterday, in that same place, that happened to one of our local boys."

There exists an unfortunate, in our opinion, tradition within the NLP community to treat indexicals in a suboptimal way on at least three fronts. (1) **Unrealistic preconditions.** Most work on automatic pronoun resolution, for example, involves supervised learning (i.e., learning from manually annotated corpora), whose resultant engines require that all future inputs be already annotated, to perfection, in the expected way. (2) **All-or-nothing classifications**. Indexicals are regularly (albeit often tacitly) categorized as "easy" (e.g., *he*) or "too hard" (e.g., pronominal *that*), whereas the actual easy/hard distinction is largely based on the contextual usage of the element. (3) **Language specificity**. Most work on indexicals in NLP and descriptive linguistics is language-specific, but many resolution functions are actually cross-linguistically applicable.[2]

Our proposal is to apply to the Semantic Web the same types of cross-linguistically applicable indexical resolution functions that are already used in the OntoSem environment. The key to successful realization of this proposal

[2] Theoretical work, like that grounded in the tradition of theoretical syntax, typically lacks the needed level of descriptive detail to be of practical utility for NLP.

involves *classifying* usage cases for indexicals with respect to *which parameter values* are required for each decision function and *how* and *with what confidence* those parameter values can be obtained and in each context.

Let us begin by considering some *types, sources* and *confidence levels* of input parameters that might contribute to functions for resolving indexicals found on the Web. **The surface string**: always available, maximally high confidence. **Semantic web annotations**: sometimes available for some types of entities; confidence varies depending on the source, type of tag, etc. **Traditional web annotations**: typically available for html documents; some types of tags (as for formatting) are of high confidence but might be noisy and difficult to automatically interpret. **Automatic "preprocessing" of text**: preprocessing (detecting tokens, proper names, dates, etc.) is a cornerstone of NLP, but web content can be error-prone due to the metadata text, embedded media, etc. **Syntactic analysis**: another mainstream NLP task though even the best current parsers achieve far less than perfect results. **Basic semantic analysis (word sense disambiguation and the determination of dependencies)**: carried out by few NLP systems, OntoSem being among them; analyses tend to be extremely useful in supporting high-level tasks like resolving indexicals, but they are error-prone. Procedural semantic routines to resolve indexicals become more complex, and typically of lower confidence, as they incorporate the latter types of features. But, centrally important for this proposal, the difficulty of each usage case and its associated confidence level can typically be automatically calculated, thus suggesting in which types of applications the automatic results might best be used. Let us consider just a few examples of indexical treatment.

Relative time expressions – such as *today*, *now*, *three weeks from tomorrow* and *in a little while* – can readily be resolved to real times (month, day, year, etc.) if (a) the "anchor time" – i.e., the time of the post (article, etc.) – is known, and (b) the time expression is used outside of direct speech. The former is expected to be recorded in Semantic Web tags, and the latter can typically be determined with high confidence using a preprocessor. (If the expression is within direct speech, then the time of speech must be determined, which requires semantic analysis.) Within OntoSem, the actual functions that can calculate, e.g., *today* vs. *three weeks from tomorrow* are recorded in the "meaning procedures" zones of the respective lexicon entries (McShane et al. 2004). As mentioned earlier, OntoSem lexicons are largely language-independent, meaning that their semantic descriptions and procedural semantic routines can be reused across languages (McShane et al. 2005); so the procedure already available for English *today* can be used to derive the full meaning of Czech *dnes* or Hebrew מהיו – assuming, of course, that preprocessors for these languages are available.

A similar example is the pronoun *I*, which can be resolved with high confidence in one of two cases: (1) if it is used outside of direct speech and

the piece has a single author as indicated by a Semantic Web tag or (2) it is used within an instance of direct speech that contains a preceding instance of *I*. In this latter case, although the real-world referent cannot be confidently distinguished, the coreference relationship between instances of *I* can be. Now contrast *I* with its plural counterpart *we*. *We* is substantially more difficult to interpret since a single author often affirms group membership – explicitly or implicitly – then subsequently speaks on behalf of the group. Alternatively, a piece can be written by more than one person, with *we* in a given context referring either to a subset of the authors or to a larger community to which they all, or a subset of them, belong. The extensive analysis required by people to craft a robust function for resolving *we* underscores why we (yes, we!) should take a cross-linguistic approach to developing procedural semantic functions for the web: it will save the community time and foster consistency of interpretations. Our initial work on the resolution of *we* within OntoSem includes subfunctions for resolving *I* and *we* that involve different kinds of heuristic evidence, some of which we can expect to be available for any language in the short term and other aspects of which require full-blown semantic analyses of the type we are working toward.

Let us conclude by stating that there are many more largely cross-linguistically applicable procedural semantic routines beyond indexicals, for example, the procedure for resolving *very* (as applied to different types of expressions) are (McShane et al. 2004).

References Cited

Doctorow, C. (No date) Metacrap: Putting the torch to seven straw-men of the meta-utopia. Available at http://www.well.com/~doctorow/metacrap.htm

Java, Akshay, Sergei Nirenburg, Marjorie McShane, Timothy Finin, Jesse English, Anupam Joshi. 2007. Using a natural language understanding system to generate Semantic Web content. International Journal on Semantic Web & Information Systems, 3(4), 50-74. October-December 2007.

McShane, Marjorie, Sergei Nirenburg and Stephen Beale. 2005. An NLP lexicon as a largely language independent resource. Machine Translation 19(2): 139-173.

McShane, Marjorie, Stephen Beale and Sergei Nirenburg. 2004. Some meaning procedures of Ontological Semantics. Proceedings of LREC-2004.

Nirenburg, S. & McShane, M. Forthcoming. *Natural Language Processing.* To appear in S.Chipman (ed.) The Oxford Handbook of Cognitive Science.

Nirenburg, S. & Raskin, V. 2004. *Ontological Semantics.* The MIT Press.