Senso Comune

A.Oltramari[1] , G.Vetere[2] , M.Lenzerini[3] , A. Gangemi[4] , N.Guarino[1]

[1] Laboratory for Applied Ontology, ISTC-CNR (Trento), [2]IBM Center for Advanced Studies (Roma),
[3]Department of Computer and System Sciences, University "Sapienza", (Roma),
[4]Semantic Technologies Laboratory, ISTC-CNR (Roma)
oltramari@loa-cnr.it, g.vetere@it.ibm.com, lenzerini@dis.uniroma1.it, aldo.gangemi@cnr.it, guarino@loa-cnr.it

Abstract

This paper introduces the general features of Senso Comune, an open knowledge base for the Italian language, focusing on the interplay of lexical and ontological knowledge, and outlining our approach to conceptual knowledge elicitation. Senso Comune consists of a machine-readable lexicon constrained by an ontological infrastructure. The idea at the basis of Senso Comune is that natural languages exist in use, and they belong to their users. In the line of Saussure's linguistics, natural languages are seen as a social product and their main strength relies on the users consensus. At the same time, language has specific goals: i.e. referring to entities that belong to the users world (be it physical or not) and that are made up in social environments where expressions are produced and understood. This usage leverages the creativity of those who produce words and try to understand them. This is the reason why ontology, i.e. a shared conceptualization of the world, can be regarded to as the soil on which the speakers' consensus may be rooted. Some final remarks concerning future work and applications are also given.

1. Introduction

In this work we aim to illustrate *Senso Comune*[1], an open knowledge base for the Italian language. *Senso Comune* consists of a machine-readable lexicon provided with an ontological framework. The knowledge base grounds on a core of basic words excerpted from De Mauro's dictionary[2] (2,075 lemmas with about 16,000 senses) whose occurrence covers more than the 90% of common Italian written sources (De Mauro, 1980). Then, *Senso Comune* has been made available for integration with other resources: a specific web based collaborative platform for linguistic resources has been developed for gathering users contributions. The project, supported by Fondazione IBM Italia, is backed by a vast community of Italian scientists under the supervision of Prof. Tullio De Mauro.

The idea at the basis of *Senso Comune* is that natural languages exist in use, and they belong to their users. In the line of Saussure's linguistics, natural languages are seen as a social product and their main strength is the users consensus. At the same time, language has specific goals: i.e. referring to entities (be them physical or not) and situations of the world that are made up in social environments, where expressions are produced and understood by human agents. This concrete usage leverages the cognitive capabilities of those who produce words and try to understand them. This is the reason why ontology, i.e. a shared theory of the physical and social world, can be regarded to as the framework in which the speakers' consensus may be established. This paper introduces the general features of *Senso Comune* with focus on the interplay of lexical and ontological knowledge, and outlines the approach to onto-linguistic knowledge elicitation we have chosen.

[1]Italian for "Common Sense". For more details refer to: `http://www.sensocomune.it`

[2]See (De Mauro, 1980).

2. The General Framework

Both Knowledge Representation and Human-Language Technologies aim at describing knowledge contents, although from different perspectives, namely the logical and the lexical one. In any case, these disciplines regard ontologies as the formal way of specifying meanings, thus representing an important link between knowledge representation and computational lexical semantics[3]. Ontologies and computational lexicons aim at digging out the basic elements of a given semantic space (domain-dependent or general), characterizing the different relations holding among them. In the simplest case, both ontologies and computational lexicons are hierarchical structures of elements (concepts or terms) concerning the entities of a given domain. Nevertheless, they substantially differ with respect to some general aspects:

1. lexical entries are not barely equivalent to ontological categories;

2. polysemy and synonymy bears upon the lexicon but should do not affect (good) ontologies;

3. the representational features of computational lexicons are not necessarily given a formal semantics;

Computational lexicons are often said to incorporate or even correspond to ontologies, whose purpose is to describe semantic constructs of language (they are bound to grammatical units). However, it is questionable whether the categorial and relational structures of computational lexicons may be acknowledged as bearing ontological commitments or not. For some extent, the problem can be seen as related to the classic "universals debate", where realists (those who think that properties referred by linguistic structures exist in some reality) confront with nominalists (those who

[3]Concerning this topic, see also (Lenci et al., 2002).

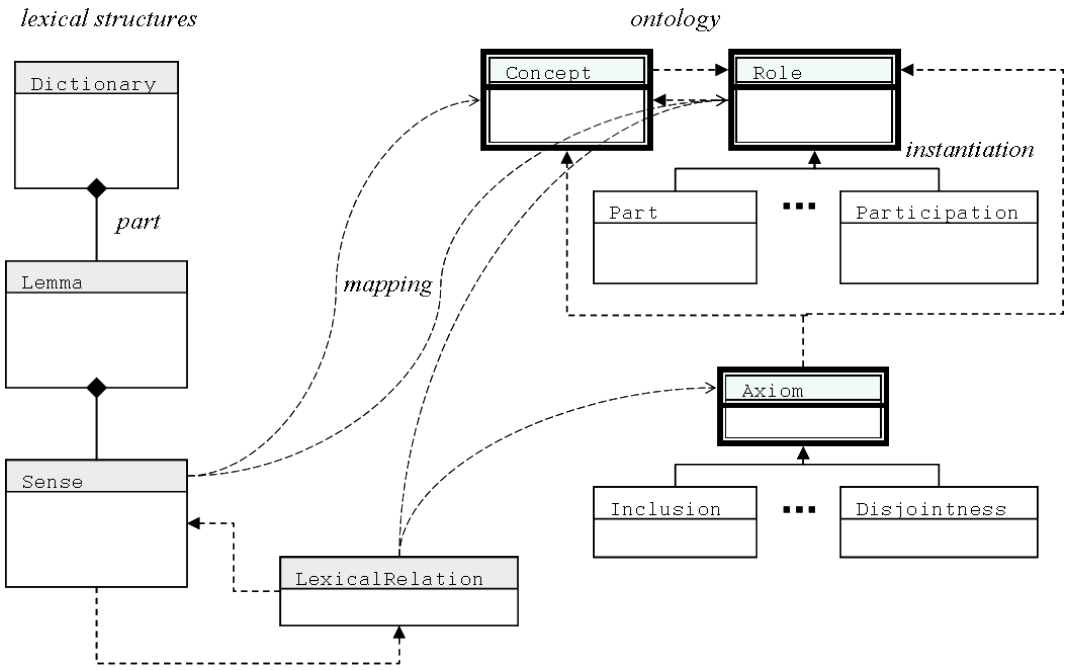

Figure 1: A sketch of *Senso Comune* metamodel

deny it)[4]. In *Senso Comune* we keep language and ontology clearly separated, focusing on how to join the two layers. Linguistic senses (dictionary's definitions) and relationships (e.g. synonymy, hyponymy, and antinomy, etc.) are therefore distinguished from their ontological counterparts (e.g. concept, inclusion, and disjointness, etc.), which are defined with logic descriptive languages (description logics). On one hand, this separation prevents linguistic facts, which are vague by nature, to be directly connected to truth-valued logic constructs; on the other hand, it relieves their entailment towards cogent ontological commitments. Of course, to model the way lexical units refer to real things and situations as described in ontologies, we provide specific mapping constructions. Generally speaking, we think that mapping ontologies with computational lexicons, and developing specific reasoning strategies for each layer along with their mappings, is one of the keygoals of the next generation of knowledge systems.

Basing on this assumption, *Senso Comune* adopts a model in which linguistic meanings (senses) and lexical relations are *mapped to* (rather than *consisting in*) logical concepts and roles. In particular, senses can be associated to either ontology concepts (unary predicates), roles (binary predicates), or individuals, while lexical relations can be associated to either roles or axioms (logical constraints)[5]. Informally, for each sense in the lexicon, a corresponding concept (or role) is introduced in the ontology. This concept is subsumed to the concept to which the sense is mapped to. On the other hand, lexical relations are mapped to two different kinds of logical constructors: roles and axioms. Specifically, relations such as synonymy, hyponymy, and antonymy that aim at capturing formal bindings among senses are mapped to axioms, while relations such as meronymy-holonymy, that express material relationships among the referred entities, are mapped to roles. However, while "substantial" lexical relations can be translated to ontology roles (*object properties*, to use the OWL jargon) quite smoothly, problems arise when mapping "formal" ones to logical axioms. Synonymy and hyponymy, for instance, cannot be directly translated into equivalence and inclusion dependencies, since they don't always exhibit transitiveness, as these axioms require. Researches are underway for understanding how to model these aspects. At the present stage, "formal" lexical relations are registered only in the dictionary, without direct ontological counterparts.

In a nutshell, *Senso Comune* is a lexical resource where each sense is weakly associated to an ontological category. Lexical knowledge and ontological theories can be therefore developed with a certain degree of mutual independence. Still, the development of the whole knowledge base is influenced by how mappings are established. For instance, setting synonymy or hyponymy between two senses whose ontological categories are disjoint, causes a warning to be issued, which can lead a revision of either the lexical information or the ontology mapping. Furthermore, assigning a certain ontological category to a certain sense, may drive the acquisition of lexical knowledge. For instance, by assigning a sense to the category of ARTIFACT, the user can be driven to specify relatedness to senses that describe instances of the category FUNCTION. To make this kind of interplay possible, a complete mapping of linguistic senses to ontological categories is needed[6]. This mapping dif-

[4]See (Lowe, 2002), (Burkhardt and Smith, 1991), etc.

[5]We adopt here the standard Description Logics terminology - see (Baader, 2002)

[6]The mapping process is supported by the TMEO methodology (Tutoring Methodology for Enrichment of Ontologies), which

fers from the individuation of synonym sets and hyponymy chains a posteriori, in that involves the adoption of "foundational ontology" beforehand. This foundation is based on a simplified OWL-lite version of DOLCE (see (Masolo et al., 2002), which we internally refer to as DOLCE-spray: it is characterized only by very general ontological distinctions taken from the original DOLCE and 'spreaded over' *Senso Comune* linguistic surface.

DOLCE[7] consists of about 40 concepts, 100 relations and 80 axioms: it corresponds to a formal model where upper level categories (endurant, perdurant, quality, and abstract) and general relations (part-of, participation, dependence, etc.) are represented in a standard logic language. This ontology has been explicitly developed in order to meet some core cognitive and linguistic features of common sense knowledge.

Although presenting DOLCE-spray in details is out of scope here, we can say that its core structure originates from the most salient cognitive distinctions embedded in DOLCE. Note that we use the meso-level dichotomy 'tangible/intangible' (see figure 2) just to differentiate entities which are capable of being perceived by the senses or the mind from those which are not (or whose shared conceptualization is strongly debatable). Tangible entities are in time (like tables, cars, trees, persons, books, etc.) or happen in time (such as conversations, wars, dives, concerts, weddings, thoughts, et cetera). Intangible entities intuitively regard abstract things (ideas, numbers, facts), time and space. Currently, we are working at "ontologizing" about 8,000 senses of about 1,000 most common Italian nouns. The process of mapping lexical senses to ontological categories tries to preserve the original lexical items, but we expect cases in which a revision of the dictionary content may be needed. For instance, figurative and concrete acceptations could be mixed in the same gloss, which would make it difficult to assign the sense a specific ontological category. In order to better fit our model, this would require the dictionary sense to be split. In our case, since De Mauro's dictionary is an high-quality, fine-grained lexical resource, these cases seem to be quite rare.

3. A collaborative semantic platform

By separating the linguistic layer from the ontology, we allow language users to manifest their knowledge in a free, incremental, natural, collaborative and potentially conflicting way. As Wikipedia demonstrates, collaborative projects produce huge amount of knowledge, which is continuosly updated, amended and extended by wiki-editors. We think that by applying a "crowdsourcing" approach to the collection of human common-sense and linguistic knowledge can also fit the "Semantic Web" paradigm. As a collaborative semantic resource, *Senso Comune* depends on two main levels:

1. top-down: top-level ontological categories and relations are introduced and maintained to constrain lexical semantic space;
2. bottom-up: language users are asked to enrich the semantic resource with linguistic information through a collaborative approach.

As we said above, level (1) is based on DOLCE-spray. In the build-up of *Senso Comune* language users are given access to the lexical level. Ontological commitments are kept "opaque" to ease users' task. These access-restrictions produce an epistemological spread between dimensions (1) and (2), which is a necessary requirement to keep the deep technical aspects of the ontological layer aside from linguistic users. Conversely, to make dimension (2) plainly effective, those lexical concepts and relations which are introduced by users must fit the intended ontological choices. For this reason, we designed TMEO, a tutoring methodology based on ontological distinctions to support enrichment of hybrid semantic resources. The acronym has been chosen in assonance with the title of Plato's fundamental dialogue, *Timaeus*, whose core subjects deal with the general theme of the structural knowledge of the world. TMEO is inspired by Plato's dialectic methodology (the protagonist Socrates drives his disciples to true knowledge, posing questions and arguing on answers): it exploits some suitable ontological properties for posing questions to users in support of domain independent or dependent knowledge modeling. TMEO is an interactive Q/A system based on general distinctions embedded in DOLCE-spray. The *Senso Comune* implementation of TMEO automatically selects the most adequate category of the reference ontology as the super-class of the given lexicalised concept: difference sequences of answers induce different mappings between the lexicon and the (hidden) ontological layer.

Consider the case in which a given user is asked to classifiy the italian fundamental term 'scarpa' (shoe), whose gloss in De Mauro's dictionary is "parte dell'abbligliamento, di cuoio o di altro materiale, che copre e protegge il piede generalmente fino alla caviglia o sopra"[8]. After initializing TMEO wizard, *Senso Comune* interface will put the user through a series of intutive conceptual questions - driven by the underlying DOLCE-spray ontological model - in order to make explicit the intended meaning of the term. The following sequence reflects an experimental trial made with multiple users[9]:

- TMEO-Wizard: Can you touch or see or smell or taste or hear or feel **a shoe**?
  User: Yes
- TMEO-Wizard: Can you count or enumerate **shoe**s?
  User: Yes
- TMEO-Wizard: Can you say that "a **shoe** is happening or occurring"?
  User: No

will be briefly discussed in the sequel.

[7]Descriptive Ontology for Linguistic and Cognitive Engineering (see `http://www.loa-cnr.it/DOLCE.html`).

[8]Basically corresponding to "footwear shaped to fit the foot (below the ankle) with a flexible upper of leather or plastic and a sole and heel of heavier material" in WordNet.

[9]We directly provide here the English translation of the questions originally posited in Italian language.

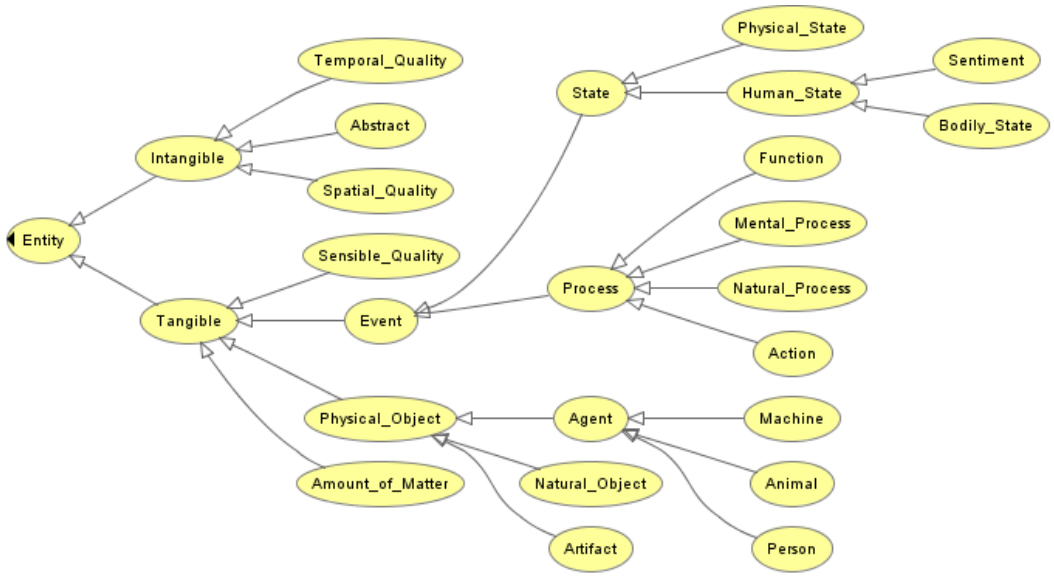

Figure 2: The taxonomy of DOLCE-spray

- TMEO-Wizard: Can you say that a **shoe** is a product of human activity (no matter whether hand-made or with the help of suitable tools and machines)?

  User: Yes

- TMEO-Wizard: **shoe** in the sense of "footwear shaped to fit the foot (below the ankle) with a flexible upper of leather or plastic and a sole and heel of heavier material" has been classified as a kind of ARTIFACT.

- TMEO-Wizard: What are **shoe**s commonly used for?

  User: to walk

- TMEO-Wizard: Select the correct sense of **to walk**: [s_1 - s_2 - s_3 - s_n][10].

  User: s_1

- TMEO-Wizard: **to walk** in the sense s_1 has been classified as the typical FUNCTION of **shoe**s.

Here the algorithm drives the user through tracing the following path of knowledge: *shoes as* ARTIFACT *have the common* FUNCTION *of being used in walking events*. As the above-mentioned scenario suggests, TMEO methodology may therefore be adopted not only in the unilateral classification of a given term ('shoe') but also in making related lexical items explicit. This kind of relatedness between terms actually unwraps the inter-categorial relation(s) holding between the corresponding ontological categories. Indeed, from the ontological viewpoint we can say that there is a relation of *Partecipation* holding between the category ARTIFACT (which is a kind of PHYSICAL OBJECT) and FUNCTION, which is conceptualized in DOLCE-spray as a kind of PROCESS[11].

One advantage of this approach consists in its relative flexibility: it is actually easy to apply and customize to specific domains, given some preliminary conceptual analysis of the entities and relations at play. Towards this direction, TMEO is being adopted, for example, in the TasLab Project[12], concerning the implementation of a semantic portal for fostering territorial ICT innovation, including the use of domain ontologies and thesauri (e.g., Eurovoc[13]), indexing and semantic search techniques[14].

4. Conclusion

Interfacing ontologies with advanced linguistic technologies is the *conditio sine qua non* to allow effective machine-understandability of "human meanings", thus supporting non-trivial applications based on semantic technologies. By implementing a collaborative approach to linguistic knowledge acquisition, *Senso Comune* aims at providing open and high-quality resources to feed semantic technologies applications. Moreover, collaborative construction of hybrid semantic resources may strongly support automatic ontology learning, where resulting ontologies are always incomplete, inconsistent, ambiguous, and machine-centered. Currently, *Senso Comune* provides a core of about 2,000 lemmas with about 16,000 acceptations, that are being classified and interlinked by voluntary contributors. Also, the integration of a suitable partition of MultiWordNet (Pianta et al., 2002) is underway. Future work will include the extension of the dictionary to encompass at least other 7,000 lemmas of common usage, including mul-

[10] For the sake of readability, we don't go through the basic senses of the verb 'to walk', also assuming that s_1 is adequately selected by the user.

[11] Note that we may wish to distinguish descriptions of functions from actual ones, namely those functions which are performed at a certain time by a given object. In the above example we simplify this distinction only focusing on the latter case.

[12] http://www.taslab.eu/

[13] http://europa.eu/eurovoc/

[14] See (Shvaiko et al., 2010).

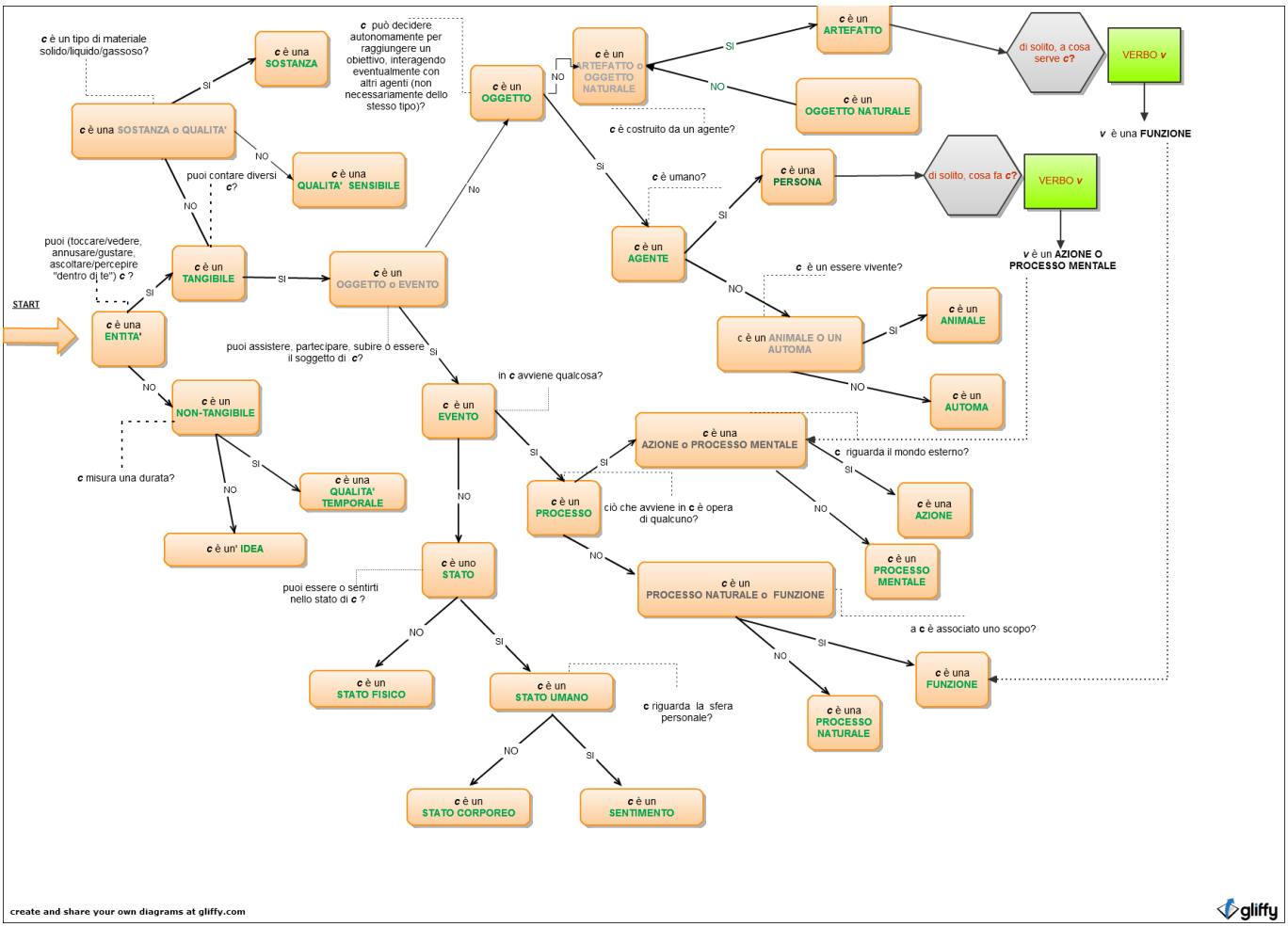

Figure 3: A representation diagram of the TMEO algorithm

tiwords. Then, we plan to proceed to the specification of lexical relationships, which will be possibly extracted from existing resources, consistentently with respect to *Senso Comune* ontological categories (DOLCE-spray will be extended too). Mereology and argumentative structure of verbs will also be part of our next releases. Future work will also be devoted to set up and realize an evaluation plan of the resource for different NLP tasks.

5. References

F. Baader. 2002. *The Description Logics Handbook: Theory, Implementation and Applications*. Cambridge University Press.

Hans Burkhardt and Barry Smith, editors. 1991. *Handbook of Metaphysics and Ontology*. Philosophia Verlag, Munich.

T. De Mauro. 1980. *Guida all'uso delle parole*. Editori Riuniti.

A. Lenci, N. Calzolari, and A. Zampolli. 2002. From text to content: Computational lexicons and the semantic web. In *Eighteenth National Conference on Artificial Intelligence; AAAI Workshop, "Semantic Web Meets Language Resources"*, Edmonton, Alberta, Canada.

E. J. Lowe. 2002. *A Survey of Metaphysics*. Oxford University Press.

C. Masolo, A. Gangemi, N. Guarino, A. Oltramari, and L. Schneider. 2002. WonderWeb Deliverable D17: The WonderWeb Library of Foundational Ontologies. Technical report.

E. Pianta, L. Bentivogli, and C. Girardi. 2002. Multiwordnet: developing an alighend multilingual database. In *First International Conference on Global WordNet*, Mysore, India.

P. Shvaiko, A. Oltramari, R. Cuel, and G. Angelini. 2010. Generating innovation with semantically enabled taslab portal. In *Seventh Extended Semantic Web Conference (ESWC 2010)*, Heraklion, Greece.