Comparing Ontologies — Similarity Measures and a Comparison Study

Alexander Maedche and Steffen Staab

Internal Report No. 408

March 2001

Institute AIFB,
University of Karlsruhe,
76128 Karlsruhe, Germany

http://www.aifb.uni-karlsruhe.de/WBS
{maedche,staab}@aifb.uni-karlsruhe.de

Comparing Ontologies — Similarity Measures and a Comparison Study

Alexander Maedche and Steffen Staab
Institute AIFB, University of Karlsruhe, 76128 Karlsruhe, Germany
{maedche,staab}@aifb.uni-karlsruhe.de
http://www.aifb.uni-karlsruhe.de/WBS

Abstract

Ontologies serve as a means for communication at a semantic level. However, to be able to effectively communicate it is first necessary to determine agreement about the underlying ontological structures. For this purpose, we consider ontologies as semiotic sign systems that are used to communicate meaning. Then we propose a methodology for measuring the extents to which two ontologies overlap and fit with each other at various semiotic levels. The measures that we propose are substantiated by a comparison study, where five human subjects have constructed ontologies about the same domain, but in isolation from each other. Building on this set of ontologies we have performed a multiple-phase cross-evaluation in order to determine adequacy of the measures proposed and inter-subject agreement about domain ontologies.

1 Introduction

A core purpose for the use of ontologies is the exchange of data not only at a common syntactic, but also at a shared semantic level [Wiederhold and Genesereth, 1997]. Especially on the WWW the construction and use of ontologies have begun to replace the old-fashioned ways of exchanging business data via standardized comma-separated formats by standard syntax (like XML/RDF) that adheres to semantic specifications given through ontologies. With the upswing and beginning widespread usage of ontologies, however, new problems of *semantic interoperability* are incurred:

Semantic Diversification. Different (human or machine) agents come with varying conceptualizations of their domains of interest. When they want to communicate they need to find the best way to talk to each other. Very often such an agent, e.g. a trading agent of a business to business marketplace, is staffed with a plethora of semantic views that allow him to adjust to its customers. *Vice versa*, the customers' agents of such a marketplace need multiple terminologies in order to be able to talk to different counterparts, like sellers of various goods or their corporation-internal customers. Therefore, they also need to solve problems regarding their semantic interoperability like *"What are the best matching ontologies or ontology parts of these two agents?"* or *"How should these two ontologies be merged or aligned to each other?"*. Also, when considering automatic trading mechanisms, one may want to consider not only operational (i.e. uptime of service), but also *semantic quality of service* (i.e. agents understand what they tell each other).

Standardization vs. Diversification. Naturally, for some important recurring problems standardization efforts oblige a certain view onto the world. To the larger extent, however, standardization of conceptualizations are infeasible, because they involve the agreement of competing players, because they are too costly, and because the domains of interest they describe are changing too fast (e.g., the product catalogues of electronic products). Where standardization fails for whatever reason, one needs to *compare ontologies* and find out whether and to what degree they overlap and fit to each other in order to bring about satisficing semantic interoperability.

Comparing Ontologies. So, how may we measure the similarity of ontologies or of ontology parts? One could make use of the formal structures of ontologies and try at the unification of ontologies or ontology parts (which is essentially subgraph matching).

The drawback here would be that all real-world ontologies that we know of do not only specify its conceptualization by logical structures, but to a large extent also by reference to terms that are grounded through human natural language use. For instance, modeling that MAN and WOMAN are subordinates of PERSON suffices for many purposes even without any further differentiae. Two ontologies that contain these parts agree on their semantics only to a small extent by formal means, but to a larger extent by reference to common terminology. Furthermore, missing structures need not be problematic. For instance, if one ontology comes with concepts referred to by VEHICLE, CAR, SPORTSWAGON and the other with VEHICLE and SPORTSWAGON only, the semantic exchange of data may still be rather easy, even though the second ontology lacks the two taxonomic links from VEHICLE to CAR and to SPORTSWAGON.

Methodological Inventory. Looking at these requirements, we have found a lack of comprehensive methodological inventory to compare real-world ontologies. as well as practical, reproducable experiences with comparing ontologies.

Firstly, this paper is about introducing the necessary inventory. We break down the ontology comparison task and pro-

pose a set of measures that capture the similarity of ontologies at several different levels. Our similarity measures describe the extent to which one ontology specification is covered by the other — and *vice versa*.

Secondly, this paper is about providing some practical experiences with the proposed measures. Just like early research in information retrieval and library science had to find out about *typical* values for *precision* & *recall* and for *inter-annotator agreement* of different librarians that index the same set of items, we have collected practical experience about the proposed measures in an experimental setting. Five subjects, four novices and one ontology engineering expert, have modelled ontologies in three different phases about a commonly well-known domain given some additional background knowledge in form of domain texts.[1]

In the following, we first prepare the ground for our proposal and our case study by formally specifying the ontology model we refer to subsequently. The two sections thereafter, we propose measures for describing the similarity of different ontology parts. In Section 5, we describe the case study and the results we achieved there, before we relate to other research and conclude the paper.

2 A Semiotics View of Ontologies

Semiotics is the study of signs and the ways in which sign systems convey (and are used to convey) meaning. When comparing ontologies, we look at ontologies as sign systems that are used for communication. Semiotics then differentiates four different levels of sign systems:

Syntax	Which signs do exist?
here	Which lexicalizations convey meaning?
Semantics	What is the meaning of these signs?
here	What relations exist between signs?
Pragmatics	How are signs used for particular purposes?
here	How do the ontologies relate to actual data?
Social	Who uses which signs?
here	Who uses which parts of which ontologies?

The linkage between subsequent levels is introduced by particular connection relations. For instance, syntax and semantics are connected through a reference relation that links a sign with a set of statements.

The measures we propose and use in the following "only" work at the syntactical and semantic levels. The reason is a very pragmatic one: It proved to be very worthwhile this far and it will take a lot of further, though we think seminal, work to approach the third and fourth level.[2]

We now define a frame system as the underlying model for the ontologies that we compare, thereby working on a slightly revised excerpt of the OKBC knowledge model [Chaudhri *et al.*, 1998]:

Definition 1 (Ontology) *An ontology is a sign system* $\mathcal{O} := (\mathcal{L}, \mathcal{F}, \mathcal{G}, \mathcal{C}, \text{ROOT}, \mathcal{H}, \mathcal{S})$*, which consists of*

- *A* ***lexicon****: The lexicon contains a set of signs (lexical entries) for concepts,* $\mathcal{L}^c$*, and a set of signs for template slots,* $\mathcal{L}^s$*. Their union is the lexicon* $\mathcal{L} := \mathcal{L}^c \cup \mathcal{L}^s$*.*
- *Two* ***reference functions*** $\mathcal{F}$*,* $\mathcal{G}$*, with* $\mathcal{F} : 2^{\mathcal{L}^c} \mapsto 2^{\mathcal{C}}$ *and* $\mathcal{G} : 2^{\mathcal{L}^s} \mapsto 2^{\mathcal{S}}$*.* $\mathcal{F}$ *and* $\mathcal{G}$ *link sets of lexical entries* $\{L_i\} \subset \mathcal{L}$ *to the set of concepts and template slots they refer to, respectively, in the given ontology. In general, one lexical entry may refer to several concepts or template slots and one concept or template slot may be refered to by several lexical entries. Their inverses are* $\mathcal{F}^{-1}$ *and* $\mathcal{G}^{-1}$*.*
- *A set of* ***concepts*** $\mathcal{C}$ *(classes in OKBC). About each* $C \in \mathcal{C}$ *exists at least one statement in the ontology, viz. its embedding in the taxonomy.*
- *A particular top concept* ROOT. ROOT *is not in* $\mathcal{C}$*, but in the taxonomy it is above every other concept.*
- *A* ***taxonomy*** $\mathcal{H}$*: Concepts are taxonomically related by the acyclic, transitive relation* $\mathcal{H}$*,* $(\mathcal{H} \subset \mathcal{C} \times (\mathcal{C} \cup \{\text{ROOT}\}))$*.* $\mathcal{H}(C_1, C_2)$ *means that* C_1 *is a subconcept of* C_2*. It holds that* $\forall C \in \mathcal{C} : \mathcal{H}(C, \text{ROOT})$*.*
- *A set of* ***template slots*** $\mathcal{S}$*: A template slot is a relation. It is referred to by its lexical entries and it specifies a pair* (C, R) *with* $C, R \in \mathcal{C}$*. A template slot* S *adds to the description of concept* C *in an object-oriented way by adding a range restriction. An instance* i_1 *of* C *may be related via* S *to another instance* i_2*, only if* $i_2 \in R$*.*

  The functions d and r applied to S *yield the corresponding domain and range concepts* C *and* R*, respectively.*

This model constitutes a core system that captures commonalities of virtually all ontology models. It leaves out standard components that may be ignored here, *e.g.* instances of concepts and (instance) slots (cf. OKBC knowledge model [Chaudhri *et al.*, 1998]), because they would only come into play when we try to determine similarity for the pragmatic level of ontologies. Others like additional axioms in F-Logic or has-value constraints in Description Logics, diverge too much between different ontology systems to be considered here. Thus, the remaining model is quite straightforward and well-agreed upon with one exception: The explicitness of the syntactic level is typically restricted to ontologies for natural language applications[3] — in spite of its general usefulness.

In the following sections we propose and use methods for measuring similarity of ontologies based on syntax and semantics of ontologies. While for the syntactic level we may only refer to similarity of form, i.e. string similarity, for the semantic level we have a richer set of relations that may be exploited, viz. the taxonomy $\mathcal{H}$ and the template slots $\mathcal{S}$.

3 Syntactic Comparison Level

The *edit distance* formulated by Levenshtein [Levenshtein, 1966] is a well-established method for weighting the difference between two strings. It measures the minimum number of token insertions, deletions, and substitutions required to

[1] The ontologies will be made publicly available on the Web when the authors need not remain anonymous anymore.

[2] E.g., for various practical reasons it has been very difficult so far to get data that allow the approaching of the third and fourth levels.

[3] Our distinction of lexical entry and concept is similar to the distinction of word form and synset used in WordNet [Miller, 1995].

transform one string into another using a dynamic programming algorithm. For example, the edit distance, ed, between the two lexical entries "TopHotel" and "Top_Hotel" equals 1, ed("TopHotel", "Top_Hotel") $= 1$, because one insertion operation changes the string "TopHotel" into "Top_Hotel".

Based on Levenshtein's edit distance we propose a *syntactic similarity measure* for strings, the String Matching (SM), which compares two lexical entries L_i, L_j:

$$\text{SM}(L_i, L_j) := \max\left(0, \frac{\min(|L_i|,|L_j|) - \text{ed}(L_i,L_j)}{\min(|L_i|,|L_j|)}\right) \in [0,1].$$

SM returns a degree of similarity between 0 and 1, where 1 stands for perfect match and zero for bad match. It considers the number of changes that must be made to change one string into the other and weighs the number of these changes against the length of the shortest string of these two. In our example from above, we compute SM("TopHotel", "Top_Hotel") $= \frac{7}{8}$. In order to provide a summarizing figure for the syntactic level of two sign systems, e.g. for the lexica referring to concepts $\mathcal{L}_1^c, \mathcal{L}_2^c$ of two ontologies $\mathcal{O}_1, \mathcal{O}_2$, we compare two sets $\mathcal{L}_1, \mathcal{L}_2$ returning the averaged String Matching $\overline{\text{SM}}(\mathcal{L}_1, \mathcal{L}_2)$:

$$\overline{\text{SM}}(\mathcal{L}_1, \mathcal{L}_2) := \frac{1}{|\mathcal{L}_1|} \sum_{L_i \in \mathcal{L}_1} \max_{L_j \in \mathcal{L}_2} \text{SM}(L_i, L_j).$$

$\overline{\text{SM}}(\mathcal{L}_1, \mathcal{L}_2)$ is an asymmetric measure that determines the extent to which the syntactic level of a sign system $\mathcal{L}_1$ (the target) is covered by the one of a second sign system $\mathcal{L}_2$ (the source). Obviously, $\overline{\text{SM}}(\mathcal{L}_1, \mathcal{L}_2)$ may be quite different from $\overline{\text{SM}}(\mathcal{L}_2, \mathcal{L}_1)$. E.g., when $\mathcal{L}_2$ contains all the strings of $\mathcal{L}_1$, but also plenty of others, then $\overline{\text{SM}}(\mathcal{L}_1, \mathcal{L}_2) = 1$, but $\overline{\text{SM}}(\mathcal{L}_2, \mathcal{L}_1)$ may approach zero. Compared to the relative number of hits,

$$\text{RelHit}(\mathcal{L}_1, \mathcal{L}_2) := \frac{|\mathcal{L}_1 \cap \mathcal{L}_2|}{|\mathcal{L}_1|},$$

$\overline{\text{SM}}$ diminishes the influence of string pseudo-differences in different ontologies, such as use vs. not-use of underscores or hyphens, use of singular vs. plural, or use of additional markup characters. Of course, SM may sometimes be deceptive, when two strings resemble each other though they there is no meaningful relationship between them, e.g. "power" and "tower". In our case study, however, we have found that in spite of this added "noise" SM may be very helpful for proposing good matches of strings.

4 Semantic Comparison Level

On the semantic level we may compare semantic structures of ontologies $\mathcal{O}_1, \mathcal{O}_2$, that vary for concepts $\mathcal{C}_1, \mathcal{C}_2$. In our model the semantic structures are solely constituted by $\mathcal{H}_1, \mathcal{H}_2$ and $\mathcal{S}_1, \mathcal{S}_2$.

4.1 Comparing taxonomies $\mathcal{H}_1, \mathcal{H}_2$

Though there has been a long discussion in the literature about comparing the similarity of two concepts in a common taxonomy (cf. Section 6), we have not found any discussion about *comparing two taxonomies*.

We start by determining the extent to which two taxonomies as seen from two particularly identified concepts compare. More precisely, we assume that we have one lexical entry $L \in \mathcal{L}_1^c \cap \mathcal{L}_2^c$ that refers via $\mathcal{F}_1$ and $\mathcal{F}_2$ to two concepts C_1, C_2 from two different taxonomies $\mathcal{H}_1, \mathcal{H}_2$. The semantics of C_1 (C_2) may be seen to be constituted by the *semantic cotopy* (SC) of C_1 (C_2), i.e. all its super- and subconcepts:

$$\text{SC}(C_i, \mathcal{H}) := \{C_j \in \mathcal{C} | \mathcal{H}(C_i, C_j) \vee \mathcal{H}(C_j, C_i) \vee C_i = C_j\}.$$

SC is overloaded to process sets of concepts, too.

$$\text{SC}(\{C_1, \ldots, C_n\}, \mathcal{H}) := \bigcup_{i:=1\ldots n} \text{SC}(C_i, \mathcal{H}).$$

The taxonomic overlap (TO) between $\mathcal{H}_1$ and $\mathcal{H}_2$ as seen from the concepts referred to by L may then be computed by following $\mathcal{F}_1^{-1}$ and $\mathcal{F}_2^{-1}$ back to the common lexicon.

$$\text{TO}'(L, \mathcal{O}_1, \mathcal{O}_2) := \frac{|\mathcal{F}_1^{-1}(\text{SC}(\mathcal{F}(\{L\}), \mathcal{H}_1)) \cap \mathcal{F}_2^{-1}(\text{SC}(\mathcal{F}(\{L\}), \mathcal{H}_2))|}{|\mathcal{F}_1^{-1}(\text{SC}(\mathcal{F}(\{L\}), \mathcal{H}_1)) \cup \mathcal{F}_2^{-1}(\text{SC}(\mathcal{F}(\{L\}), \mathcal{H}_2))|}$$

Averaging over all lexical entries we may thus compute a semantic similarity for the two given hierarchies.

In addition, however, we must consider the case where a lexical entry L is in $\mathcal{L}_1^c$, but not in $\mathcal{L}_2^c$. Then, the simplest assumption is that the L is simply missing from $\mathcal{L}_2^c$, but when comparing the two hierarchies the optimistic taxonomic approximation is the one that searches for the maximum overlap given a fictive membership of L to $\mathcal{L}_2^c$ by

$$\text{TO}''(L, \mathcal{O}_1, \mathcal{O}_2) := \max_{C \in \mathcal{C}_2} \left\{ \frac{|\mathcal{F}_1^{-1}(\text{SC}(\mathcal{F}(\{L\}), \mathcal{H}_1)) \cap \mathcal{F}_2^{-1}(\text{SC}(C), \mathcal{H}_2)|}{|\mathcal{F}_1^{-1}(\text{SC}(\mathcal{F}(\{L\}), \mathcal{H}_1)) \cup \mathcal{F}_2^{-1}(\text{SC}(C), \mathcal{H}_2)|} \right\}$$

Given these premises the averaged similarity $\overline{\text{TO}}$ between two taxonomies $(\mathcal{H}_1, \mathcal{H}_2)$ of ontologies $(\mathcal{O}_1, \mathcal{O}_2)$ may then be defined by:

$$\overline{\text{TO}}(\mathcal{O}_1, \mathcal{O}_2) := \frac{1}{|\mathcal{L}_1^c|} \sum_{L \in \mathcal{L}_1^c} \text{TO}(L, \mathcal{O}_1, \mathcal{O}_2), \text{ with}$$

$$\text{TO}(L, \mathcal{O}_1, \mathcal{O}_2) := \begin{cases} \text{TO}'(L, \mathcal{O}_1, \mathcal{O}_2) & \text{if } L \in \mathcal{L}_2^c \\ \text{TO}''(L, \mathcal{O}_1, \mathcal{O}_2) & \text{if } L \notin \mathcal{L}_2^c \end{cases}$$

Example. A partial example for comparing taxonomies is given in Figure 1:

The taxonomic overlap $\text{TO}'(\text{"hotel"}, \mathcal{H}_1, \mathcal{H}_2)$ is determined by $\mathcal{F}_1^{-1}(\text{SC}(\mathcal{F}(\{\text{"hotel"}\}), \mathcal{H}_1)) = \{\text{"hotel"}, \text{"accomodation"}\}$ and $\mathcal{F}_2^{-1}(\text{SC}(\mathcal{F}(\{\text{"hotel"}\}), \mathcal{H}_2)) = \{\text{"wellness hotel"}, \text{"hotel"}\}$ resulting in $\text{TO}'(\text{"hotel"}, \mathcal{H}_1, \mathcal{H}_2) = \frac{1}{3}$ as input to $\overline{\text{TO}}$.

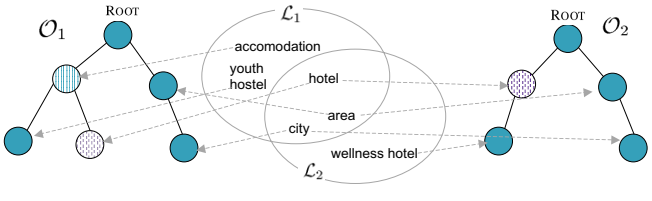

Figure 1: Two Example Ontologies $\mathcal{O}_1, \mathcal{O}_2$

When we consider the lexical entry "accomodation", which is only in $\mathcal{L}_1^c$, we compute the taxonomic overlap as follows: We compute for the lexical entry "accomodation" $\mathcal{F}_1^{-1}(\text{SC}(\mathcal{F}(\{\text{"accomodation"}\}), \mathcal{H}_1)) = \{\text{"youth hostel"}, \text{"accomodation"}, \text{"hotel"}\}$. The concept referred to by "hotel" in $\mathcal{C}_2$ yields the best match resulting in $\mathcal{F}_2^{-1}(\text{SC}(\mathcal{F}(\{\text{"hotel"}\}))) = \{\text{"wellness hotel"}, \text{"hotel"}\}$ and, thus, $\text{TO}''(\text{"accomodation"}, \mathcal{H}_1, \mathcal{H}_2) = \frac{1}{4}$.

The reader may note several properties of $\overline{\text{TO}}$: First, $\overline{\text{TO}}$ is asymmetric. While TO' is a symmetrical measure, TO'' is

asymmetric, because depending on coverage it may be very easy to integrate one taxonomy into another one, but it may be very difficult to do it the other way around. Second, for ease of presentation of the basic principles we have given here a shortened definition. The longer version specially considers the (minority of) cases, where one lexical entry refers to several concepts. The longer version does not consider the semantic cotopies of all referred concepts for computing TO, but only those that eventually optimize TO. Third, obviously $\overline{\text{TO}}$ becomes meaningless when $\mathcal{L}_1^c$ and $\mathcal{L}_2^c$ are disjoint. The more $\mathcal{L}_1^c$ and $\mathcal{L}_2^c$ overlap (or are made to overlap, e.g. through a syntactic merge), the better $\overline{\text{TO}}$ may focus on existing hierarchical structures and not on optimistic estimations of adding a new lexical entry to $\mathcal{L}_2^c$.

4.2 Comparing template slots $\mathcal{S}_1, \mathcal{S}_2$

On the syntactic level a template slot S_1 is referred to by a lexical entry L_1. On the semantic level it specifies a pair $(C_1, R_1), C_1, R_1 \in \mathcal{C}$ describing the concept C_1 that the template slot belongs to and its range restriction R_1.

We determine the accuracy that two template slots match, TSO (template slot overlap), based on the geometric mean value of how similar their domain and range concepts are. The geometric mean reflects the intuition that if either domain or range concepts utterly fail to match, the matching accuracy converges against 0, whereas the arithmetic mean value might still turn out a value of 0.5.

The similarity between two concepts (the concept match CM) may be computed by considering their semantic cotopy. However, the measures derived from complete cotopies underestimate the place of concepts in the taxonomy. For instance, the semantic cotopy of the concept corresponding to "hotel" in $\mathcal{L}_2$ (Figure 1) is identical to the semantic cotopy of the one corresponding to "wellness hotel". Hence, for the purpose of similarity of concepts (rather than taxonomies), we define the upwards cotopy (UC) as follows:

$$\text{UC}(C_i, \mathcal{H}) := \{C_j \in \mathcal{C} | \mathcal{H}(C_i, C_j) \vee C_j = C_i\}.$$

Based on the definition of the upwards cotopy (UC) the concept match (CM) is then defined in analogy to TO':

$$\text{CM}(C_1, \mathcal{O}_1, C_2, \mathcal{O}_2) := \frac{|\mathcal{F}_1^{-1}(\text{UC}(C_1, \mathcal{H}_1)) \cap \mathcal{F}_2^{-1}(\text{UC}(C_2, \mathcal{H}_2))|}{|\mathcal{F}_1^{-1}(\text{UC}(C_1, \mathcal{H}_1)) \cup \mathcal{F}_2^{-1}(\text{UC}(C_2, \mathcal{H}_2))|}.$$

Then TSO' of slots S_1, S_2 may be defined by:

$$\text{TSO}'(S_1, \mathcal{O}_1, S_2, \mathcal{O}_2) := \sqrt{\text{CM}(\text{d}(S_1), \mathcal{O}_1, \text{d}(S_2), \mathcal{O}_2) \cdot \text{CM}(\text{r}(S_1), \mathcal{O}_1, \text{r}(S_2), \mathcal{O}_1)}.$$

In order to take reference by $L \in \mathcal{L}_1^s, L \in \mathcal{L}_2^s$ into account:

$$\text{TSO}''(L, \mathcal{O}_1, \mathcal{O}_2) := \frac{1}{|\mathcal{G}_1(\{L\})|} \sum_{S_1 \in \mathcal{G}_1(\{L\})} \max_{S_2 \in \mathcal{G}_2(\{L\})} \{\text{TSO}'(S_1, \mathcal{O}_1, S_2, \mathcal{O}_2)\}$$

Some lexical entries only refer to slots in $\mathcal{S}_1$:

$$\text{TSO}'''(L, \mathcal{O}_1, \mathcal{O}_2) := \frac{1}{|\mathcal{G}_1(\{L\})|} \sum_{S_1 \in \mathcal{G}_1(\{L\})} \max_{S_2 \in \mathcal{S}_2} \{\text{TSO}'(S_1, \mathcal{O}_1, S_2, \mathcal{O}_2)\}$$

Combined we have for $L \in \mathcal{L}_1^s$:

$$\text{TSO}(L, \mathcal{O}_1, \mathcal{O}_2) := \begin{cases} \text{TSO}''(L, \mathcal{O}_1, \mathcal{O}_2) & \text{if } L \in \mathcal{L}_2^s \\ \text{TSO}'''(L, \mathcal{O}_1, \mathcal{O}_2) & \text{if } L \notin \mathcal{L}_2^s \end{cases}$$

The averaged slot matching accuracy $\overline{\text{TSO}}$ is then defined by:

$$\overline{\text{TSO}}(\mathcal{O}_1, \mathcal{O}_2) := \frac{1}{|\mathcal{L}_1^s|} \sum_{L \in \mathcal{L}_1^s} \text{TSO}(L, \mathcal{O}_1, \mathcal{O}_2).$$

Example. We take Figure 1 as an example setting for computing TSO. We assume one template slot S_1 in $\mathcal{O}_1$, referenced by "located at" and specifying the domain and range corresponding to ("hotel", "area"). In $\mathcal{O}_2$, the same lexical entry may refer to S_2, with domain and range corresponding to ("hotel", "city"). Computing CM for the concepts referred to by "hotel" in $\mathcal{O}_1$ and $\mathcal{O}_2$ results in $\frac{1}{2}$. The CM between the concepts referred to by "area" in $\mathcal{O}_1$ and "city" in $\mathcal{O}_2$ also returns $\frac{1}{2}$. Thus, the TSO' for the lexical entry "located at" boils down to $\sqrt{\frac{1}{2} \cdot \frac{1}{2}} = 0.5$ as input to $\overline{\text{TSO}}$.

The reader may note two major characteristics of $\overline{\text{TSO}}$. First, it depends on the agreement of the lexica and the taxonomies of $\mathcal{O}_1$ and $\mathcal{O}_2$. Without reasonable agreement, $\overline{\text{TSO}}$ may not reach high values of similarity. Second, $\overline{\text{TSO}}$ is also asymmetric reflecting the coverage of template slots of the first by the second ontology.

5 Empirical Evaluation

In this section we present a case study that has been carried out in a seminar on ontology engineering at our institute. We have pursued two main objectives with our evaluation study: *(i)* we wanted to determine the quality of our measures and evaluate them on actual data, and, *(ii)*, we wanted to investigate how similar different ontologies are that have been modeled by different persons.

5.1 Evaluation Study

The experiment was carried out with four subjects, viz. undergraduates in industrial engineering. The modeling expertise of the subjects was sparse. Before actual modeling, they received 3 hours training in ontology engineering in general and 3 hours in using our ontology engineering workbench. Our study required from each of them the building of ontologies in the tourism domain using their background knowledge and using web pages from a WWW site about touristic offers, e.g. hotels with various attractions or cultural events.

Our objective was an overall cross-comparison of ontologies, but we also wanted to test the appropriateness of single measures, To avoid error chaining, we therefore performed the evaluation in three phases (resulting in $4 \cdot 3 = 12$ ontologies). Furthermore, an expert ontology engineer (subject 0) modeled a "gold standard" for the task (a 13th ontology).

Phase I: A small top level structure was given to the subjects.[4] Based on this top level and the available knowledge sources, the subjects had to model a *complete* tourism domain ontology. To keep the ontologies within comparable ranges, the students were required to model around 300 concepts and 80 template slots.

Phase II: The second phase was geared to produce results for $\overline{\text{TO}}$, while avoiding the uncertainties of lexical disagreement. Therefore, the subjects were given 310 lexical entries

[4] It contained four concepts referred to by "thing", "material", "intangible", and "situation".

(for concepts) from the gold standard and the top level structure described before. Then everyone of them had to, first, model the taxonomy for concepts referred to by the 310 lexical entries and, second, model about 80 template slots.

Phase III: The last phase was defined to control $\overline{\text{TSO}}$ in absence of "noise" from different taxonomies and lexica. There the taxonomy (from the gold standard) was given. It consisted of 310 lexical entries, $\mathcal{L}^c$, and a set of 310 corresponding concepts, $\mathcal{C}$, taxonomically related by $\mathcal{H}$. The subjects had to model about 80 template slots.

5.2 Syntactic Comparison Level

The phase I-ontologies described above are used for general cross-comparison, including the syntactic level. The pairwise string matching ($\overline{\text{SM}}$, cf. Section 3) of the five lexica referring to concepts and template slots, respectively, returned the results depicted in Table 1.

Results: The results for computing $\overline{\text{SM}}(\mathcal{L}_1^c, \mathcal{L}_2^c)$ of matching lexical entries referring to concepts vary between 0.38 and 0.65 with an average of 0.45. Comparing lexical entries referring to template slots $\overline{\text{SM}}(\mathcal{L}_1^s, \mathcal{L}_2^s)$ results in values between 0.16 and 0.53 with an average of 0.36. Several typical, though not necessarily good, pairs for which high string match values were computed are shown in Table 2. RelHit$(\mathcal{L}_1^c, \mathcal{L}_2^c)$ ranged between 20 to 25%, *i.e.* this percentage of lexical entries referring to concepts matched exactly. For lexical entries referring to template slots the results were much worse, *viz.* between 10 to 15%.

	Subject				
$i \backslash j$	0	1	2	3	4
0	-	0.51,0.35	0.53,0.21	0.46,0.39	0.5,0.29
1	0.43,0.52	-	0.65,0.43	0.43,0.53	0.39,0.41
2	0.42,0.24	0.54,0.37	-	0.36,0.24	0.4,0.2
3	0.38,0.47	0.43,0.45	0.38,0.28	-	0.38,0.36
4	0.46,0.38	0.41,0.5	0.48,0.16	0.43,0.39	-

Table 1: $\overline{\text{SM}}(\mathcal{L}_i^c, \mathcal{L}_j^c)$, $\overline{\text{SM}}(\mathcal{L}_i^s, \mathcal{L}_j^s)$ for phase I-ontologies.

Interpretation: Analysing the figures we find that human subjects have a considerable higher agreement on lexical entries referring to concepts than on ones referring to slots. Investigating the auxiliary measures we have found that SM values above 0.75 in general retrieve meaningful matches — in spite of few pitfalls (cf. Table 2).

5.3 Semantic Comparison Level

On the semantic level we may compare semantic structures of ontologies $\mathcal{O}_1, \mathcal{O}_2$, that vary for concepts $\mathcal{C}_1, \mathcal{C}_2$. We use the ontologies of phase I, II, and III for evaluating our measures introduced in Section 4.

Results: Table 3 presents the results we have obtained for the phase I-ontologies using the similarity measures taxonomy overlap ($\overline{\text{TO}}$) and template slot overlap ($\overline{\text{TSO}}$). The reader may note that these ontologies have been built without any previous assumptions about the lexica $\mathcal{L}_1$ and $\mathcal{L}_2$, thus their similarity values are well below those of later phases where the lexica for concepts were predefined.

L_1	L_2	SM$(L1, L2)$
Sehenswuerdigkeit [seesight]	Sehenswürdigkeit [seesight]	0.875
Verkehrsmittel [vehicle]	Luftverkehrsmittel [air vehicle]	0.71
Zelt [tent]	Zeit [time]	0.75
Anzahl_Betten [number_beds]	hat_Anzahl_Betten [has_number_beds]	0.77

Table 2: Typical string matches

	Subject				
$i \backslash j$	0	1	2	3	4
0	-	0.33,0.35	0.31,0.25	0.32,0.5	0.29,0.28
1	0.35,0.15	-	0.4,0.41	0.34,0.03	0.28,0.15
2	0.28,0.12	0.36,0.25	-	0.25,0.04	0.24,0.15
3	0.36,0.4	0.31,0.32	0.24,0.04	-	0.26,0.03
4	0.38,0.29	0.31,0.21	0.32,0.2	0.32,0.26	-

Table 3: $\overline{\text{TO}}(\mathcal{O}_i, \mathcal{O}_j)$, $\overline{\text{TSO}}(\mathcal{O}_i, \mathcal{O}_j)$ for phase I-ontologies.

Table 4 depicts the similarity measures computed for phase II-ontologies. Values for $\overline{\text{TO}}$ range between 0.47 and 0.87, the average $\overline{\text{TO}}$ over all 20 cross-comparisons results in 0.56. $\overline{\text{TSO}}$ yields values from 0.34 to 0.82 with an average of 0.47.

	Subject				
$i \backslash j$	0	1	2	3	4
0	-	0.57,0.5	0.54,0.47	0.54,0.48	0.59,0.39
1	0.57,0.44	-	0.86,0.78	0.48,0.45	0.55,0.35
2	0.54,0.46	0.87,0.82	-	0.46,0.46	0.58,0.35
3	0.54,0.44	0.48,0.5	0.46,0.47	-	0.47,0.34
4	0.58,0.4	0.55,0.45	0.57,0.45	0.47,0.35	-

Table 4: $\overline{\text{TO}}(\mathcal{O}_i, \mathcal{O}_j)$, $\overline{\text{TSO}}(\mathcal{O}_i, \mathcal{O}_j)$ for phase II-ontologies.

Interpretation: The figures indicate that subjects tend to agree or disagree on taxonomies irrespective of the amount of material being predefined. In fact, correlation between $\overline{\text{TO}}$ values of phase I- and phase II- ontologies support this indication, because correlation is 0.58 — distinctly positive — for the ontologies with and without predefined lexica. Furthermore, we may conjecture that comparison between $\overline{\text{TO}}$ values (in order to select the best) remains meaningful even with a restricted overlap of lexica.

Results: Table 5 depicts the similarity measures computed for phase III-ontologies, where only $\overline{\text{TSO}}$ has been computed, because the taxonomy was predefined. $\overline{\text{TSO}}$ here ranges between 0.23 and 0.71, the average $\overline{\text{TSO}}$ over all 20 cross-comparisons achieving 0.5.

Interpretation: The correlation of $\overline{\text{TSO}}$ values between phases I and II computes to 0.34, between phases I and III to 0.27, and between phases II and III to 0.16. In general, higher $\overline{\text{TSO}}$ values are reached without a predefined taxonomy — this reflects the observation that subjects found it easy to use a predefined lexicon, but extremely difficult to continue modeling given a predefined taxonomy.

Overall, we may conjecture that the engineers' use of their lexicon correlates rather strongly with their semantic model

	Subject				
$i \backslash j$	0	1	2	3	4
0	-	0.61	0.38	0.51	0.54
1	0.69	-	0.56	0.57	0.55
2	0.4	0.49	-	0.35	0.23
3	0.67	0.71	0.5	-	0.57
4	0.45	0.44	0.3	0.41	-

Table 5: $\overline{\text{TSO}}(\mathcal{O}_i, \mathcal{O}_j)$ for phase III-ontologies.

and *vice versa*: The similarity measures for subject 3 ontologies with subject 4 ontologies result in very low values on the syntactic and on the semantic level. In contrast, subject 1 ontologies reach high similarity values with subject 2 ontologies on all levels.

6 Related Work

Similarity measures for ontological structures have been widely researched, e.g. in cognitive science, databases, software engineering[Spanoudakis and Constantopoulos, 1994], and AI (e.g., [Rada *et al.*, 1989; Agirre and Rigau, 1996; Hovy, 1998]). Though this research covers many wide areas and application possibilities, all of it has restricted its attention to the determination of similarity of lexical entries, concepts, and template slots *within one ontology*. The nearest to our comparison *between two ontologies* come [Bisson, 1992] and [Weinstein and Birmingham, 1999].

[Bisson, 1992] introduces several similarity measures in order to locate a new complex concept into an existing ontology by similarity rather than by logic subsumption. Bisson restricts the attention to the semantic comparison level. In contrast to our work the new concept is described in terms of the existing ontology. Furthermore, he does not distinguish relations into taxonomic relations and template slots, thus ignoring the semantics of inheritance.

[Weinstein and Birmingham, 1999] compute description compatibility in order to answer queries that are formulated with a conceptual structure that is different from the one of the information system. In contrast to our approach their measures depend to a very large extent on a shared ontology that mediates between locally extended ontologies. Their algorithm also seems less suited to evaluate similarities of sets of lexical entries, taxonomies, and template slots.

Applications: Recently, a number of proposals have been made to facilitate ontology merging and aligning [Hovy, 1998; McGuinness *et al.*, 2000; Noy and Musen, 2000]. Though these authors have several similarity measures, their focus has been on the overall system rather than the singular measures. The measures that we propose in this paper may improve suggestions for merging and aligning ontologies. Regarding the evaluation of the tools, our paper offers a baseline of how much agreement human modelers achieve when they model independently of each other.

7 Conclusion

We have investigated ontologies as sign systems, for which different levels exist that contribute to the communication function of ontologies. In particular, we have considered the syntactic and semantic level of a core ontology model describing an original methodological inventory to compare ontologies with each other based on the notions of lexicon $\mathcal{L}$, reference functions $\mathcal{F}, \mathcal{G}$ and semantic cotopy (SC, UC). Then, we have performed a three-phase case study to see how our measures perform in isolation and in combination.

With our investigation we have created a methodological baseline and collected some practical experiences for its applicability rather than a ready-to-use toolset. The reasons are twofold: On the one hand we have not yet fully exploited the potential of the underlying principles presented here. For instance, taxonomy overlap ($\overline{\text{TO}}$) may be used for comparing subtaxonomies of two ontologies that are to be merged/aligned. Thereby, $\overline{\text{TO}}$ allows to take a bird's eyes view onto the ontology merging task rather than a bottom-up view such as prevails in current state-of-the-art merging tools. On the other hand much more experiences are needed about the practical usage of ontology similarity measures. This however is like a hen-and-egg problem. Without techniques for finding the best matching ontology few systems are built that actually deal with a diversity of ontologies. Without such real-life scenarios an improved evaluation with real data is very difficult. We are about to start building a large number of small ontologies for a commercial agent marketplace. There we need methodology for finding best matching ontologies/ontology parts. Later on, we want to further evaluate the inventory described here and replace our lab experiences with real world data.

References

[Agirre and Rigau, 1996] E. Agirre and G. Rigau. Word sense disambiguation using conceptual density. In *Proc. of COLING-96*, 1996.

[Bisson, 1992] G. Bisson. Learning in FOL with a similarity measure. In *Proc. of AAAI-1992*, pages 82–87, 1992.

[Chaudhri *et al.*, 1998] V. Chaudhri, A. Farquhar, R. Fikes, P. Karp, and J. Rice. OKBC: A programmatic foundation for knowledge base interoperability. In *Proceedings of AAAI-98*, pages 600–607, 1998.

[Hovy, 1998] E. Hovy. Combining and standardizing large-scale, practical ontologies for machine translation and other uses. In *Proc. of the First Int. Conf. on Language Resources and Evaluation (LREC)*, 1998.

[Levenshtein, 1966] I. V. Levenshtein. Binary Codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10(8):707–710, 1966.

[McGuinness *et al.*, 2000] D. McGuinness, R. Fikes, J. Rice, and S. Wilder. The Chimaera ontology environment. In *Proc. of AAAI-2000*, pages 1123–1124, 2000.

[Miller, 1995] G. Miller. Wordnet: A lexical database for English. *CACM*, 38(11):39–41, 1995.

[Noy and Musen, 2000] N. F. Noy and M. A. Musen. PROMPT: Algorithm and tool for automated ontology merging and alignment. In *Proc. of AAAI-2000*, pages 450–455, 2000.

[Rada *et al.*, 1989] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1), 1989.

[Spanoudakis and Constantopoulos, 1994] G. Spanoudakis and P. Constantopoulos. Similarity for analogical software reuse: A computational model. In *Proc. of ECAI-1994*, pages 18–22, 1994.

[Weinstein and Birmingham, 1999] P. Weinstein and W. Birmingham. Comparing concepts in differentiated ontologies. In *Proc. of KAW-99*, 1999.

[Wiederhold and Genesereth, 1997] G. Wiederhold and M. Genesereth. The conceptual basis for mediation services. *IEEE Intelligent Systems*, 12(5):38–47, 1997.