

# Measuring Semantic Label Quality Using WordNet

Fabian Friedrich

School of Business and Economics,  
Institute of Information Systems  
Spandauer Strasse 1, Berlin, Germany  
Fabian.Friedrich@informatik.hu-berlin.de

**Abstract:** The automatic determination of defects in business process model and the assurance of a high quality standard are crucial to achieve easy to read and understandable models. Recent research has focused its efforts on the analysis of structural properties of business process models. This paper instead wants to focus on the labels and their impact on the understandability and integratability of process models. Metrics which can help in identifying process model labels that could lead to misunderstandings are discussed and a way to automatically detect labels with a high chance of ambiguity is presented. Therefor the lexical database WordNet is used to obtain information about the specificity and possible synonyms of a word. The derived measures were then applied to the SAP Reference Model and the most interesting findings are presented.

## 1 Introduction

Business Process Modeling has received more and more attention in recent years. Naturally, the interest in the quality of those models grows with their application. Different frameworks have been developed to understand the factors that influence the quality of business process models. A long established framework is Lindland et al. [LSS94] developed in 1994, which served as a basis for many other quality definition approaches [KJ03, KSJ06]. But these Frameworks provide only qualitative statements about process model quality and are rather abstract. So far very little research has been conducted to find appropriate quantitative measurements. Some of them are, for example, the Cross Flow Connectivity or the Density metrics [VCR<sup>+</sup>] which try to analyze the structure of a given process model.

Figure 1 shows an EPC from the SAP Reference Model, which was chosen as a basis for tests for the following analysis. What makes the labels of this model interesting is the fact that the term "wage" and "remuneration" were used within the same model although they are interchangeable. This problem of inconsistent usage of terms arises due to different levels of detail and abstractions used by different modelers [HS06]. Hence, comparing and merging of models or sub-models becomes more complicated because of these conflicts [Pfe08]. To detect and avoid those conflicts this paper will propose solutions by analyzing the labels of process models. The particular approach is to analyze the meaning of these labels, using the well known WordNet semantic database [Mil95] and to define a

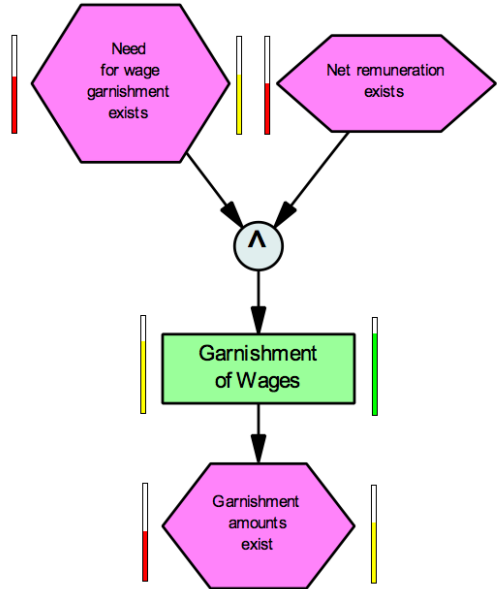


Figure 1: Example of an analyzed EPC. The graphical extension is explained in section 6

quantitative measure which is able to provide clear evidence to whether a label is good or bad.

On the following pages a short introduction of the elements in WordNet is given. Afterwards these elements will be used to derive two quantitative measures for the semantic quality of process model labels. The last chapter will then present the results of the application of those measures to the EPCs of the SAP Reference Model to verify their value. The paper will conclude by critically assessing this application's results and provide an outlook to possible extensions and further research.

## 2 Background

This section wants to introduce the preliminaries for the metrics that were developed. As the focus of this work are the label's semantics, an overview of the lexical database WordNet will be given. Furthermore, section 4 will make use of semantic relatedness measures to determine the meaning of a word. Therefore, the main principles of semantic relatedness will be shortly introduced, too.

## 2.1 A Brief Introduction to WordNet

WordNet is a semantic database which was developed in 1985 at Princeton University, mainly for natural language processing [Mil95]. Since then it has steadily grown and today it contains more than 155,000 words. These words are organized into so called SynSets (Synonym Sets). A SynSet contains several words which share the same meaning. Furthermore, the SynSets in WordNet are linked to each other through pointers with different meanings. Thus a program is able to extract different semantic relations for a given word. For example:

- Synonyms - Words which have the same meaning (to work on - to process)
- Homonyms - Words which are written identically, but have a different meaning
- Hypernyms/Troponyms - Nouns which are superordinate to the given noun. The opposite of a hypernym is a Hyponym. The same principle can also be applied to verbs which is called a troponym then. (sue - challenge, tree - plant)
- Meronyms - Structures nouns in a "part-of" relationship. (car - wheel)
- Antonyms - Mainly used for adjectives and adverbs and describes the opposite (wet - dry, hot - cold)

The quality metrics that will be explained in detail in section 3 will make use of the possibility to extract synonyms and hypernyms/troponyms from WordNet.

## 2.2 Semantic Relatedness

Semantic relatedness is a measure that states in how far two terms are related to each other. Some of the most popular semantic relatedness measures are those of Hirst and St-Onge, Leacock and Chodorow, Resnik, Jiang & Conrath [JC97] and Dekang Lin [Lin98] which were compared in [BH01]. The latter two were used in the conducted experiments. A recent master thesis also investigated on this topic [Scr06]

Both measures leverage the hierarchical structure that is build up within WordNet. But they rely on statistical information on the probability of the occurrence of a word. A possible approach to determine these probability values is to count the number of occurrences within a large corpus of text such as the complete works of Shakespeare, the Brown corpus from the ICAME Collection of English or the British National Corpus<sup>1</sup>, which was used for the experiments in this paper. Once the probability values of a concept  $C_i$  can be determined,  $P(C_i)$ , and the first concept  $C_0$  that subsumes both of those concepts has been extracted from the hierarchical structure, the similarity can be computed as follows:

Lin:

$$sim_{Lin}(C_1, C_2) = (2 * \log P(C_0)) / (\log P(C_1) + \log P(C_2)) \quad (1)$$

---

<sup>1</sup><http://www.natcorp.ox.ac.uk/>

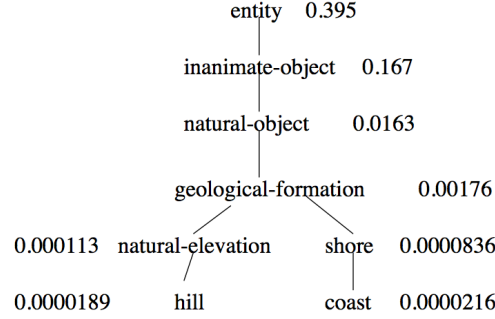


Figure 2: A Fragment of WordNet with its corresponding probability value. [Lin98]

Jiang & Conrath:

$$sim_{JCn}(C_1, C_2) = -\log P(C_0) \quad (2)$$

Taking the example from figure 2 this means that the similarity between the words "coast" and "hill", given that their first common parent is "geological-formation", is 0.59 (Lin) or 9.15 (Jiang & Conrath), respectively. Obviously, the metric defined by Dekang Lin has the advantage that it is always within the bounds of 0.0 and 1.0, but as the word sense disambiguation conducted in section 4 only tries to determine a maximum value it has no influence on the metrics, yet.

### 3 Quality Metrics

The information that can be gathered through the methods previously explained will now be used to determine the semantic quality for a given label. As this quality largely depends on the environment the label is found in, a model cannot be considered independently, but a collection of many models - a model repository - has to be analyzed. To test the metrics presented in this paper, the 604 EPCs of the SAP Reference Model were used. The focus will be the analysis of nouns and verbs, as especially the specificity metrics which will be described below rely on the hypernym/troponym structure within WordNet.

#### 3.1 Consistency

The first problem that is addressed here is the use of several words with the same meaning, as this contradicts the principle of a shared vocabulary and increases the ambiguity of labels and is responsible for misunderstandings. This occurrence of 2 synonyms referring to the same concept or sense respectively, is a conflict that can lead to problems in the process of consolidating a model repository. Resolving those conflicts usually demands

word	usage count	frequency
order	1025	84%
purchase	199	16%
bill	202	60%
invoice	132	40%

Table 1: Example of synonym words used in the SAP Reference Model

the consultation of several technical and domain experts and can thus become very costly. [DHLS08]

An example of such practice would be the usage of both "invoice" and "bill" within the same model repository.

The first step to detect such rarely used synonyms is to count the occurrence of every word within the model repository. To unify different word forms the stemming algorithm of Porter [Por80] was used. Afterwards synonyms of a given word can be acquired through WordNet. The number of occurrences of the given word divided by the total number of occurrences of all synonyms of the word then determines the relative frequency of the given word within the model repository. Another example are the words "order" and "purchase" which both can be found in the SAP Reference Model (see table 3.1).

Now, the consistency quality can be measured. Either an occurrence of 100% can be demanded for each word compared to its synonyms or a continuous quality value based on a minimum frequency  $m$ , which has to be defined externally, can be used. The quality measure for a word  $x$  can then be computed using a quality function like this one:

$$q_{cons}(x) = \min(1.0, (\frac{frequency(x)}{m})^2) \quad (3)$$

This quality function has the advantage that it is scaled between 0.0 and 1.0 and as soon as the frequency value  $x$  of a words falls below  $m$  the quality will rapidly decrease, but a distinction is still possible. Additionally, the quality measure depends on the minimum frequency  $m$ . Taking the example above, this means that at a minimum frequency level of 30% the word "purchase" will be assigned a quality value of  $\sim 0.284$ , whereas at  $m = 50\%$  it is  $\sim 0.098$ . Hence, it can be defined how strict the quality metric is supposed to evaluate the labels of a given model.

### 3.2 Specificity

Another problem, that arises when different people are modeling the same domain, is that each can use a different level of abstraction to describe reality. A good indicator for the used level of abstraction is the depth within the WordNet hypernym tree. The deeper the level of a word, the more specific it is. Thus, to align different sub-models within the repository all the words used should be in the same depth range within the WordNet hypernym tree. As this tree is only available for verbs and nouns the analysis

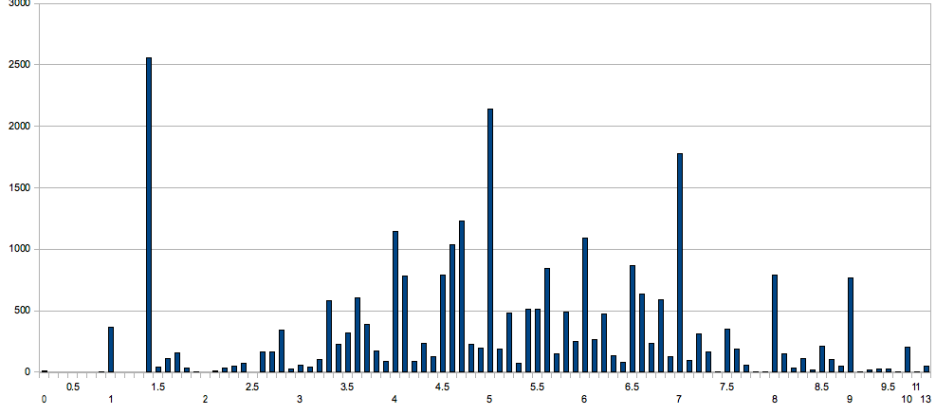


Figure 3: The distribution of average depth values of the words from the SAP Reference Model

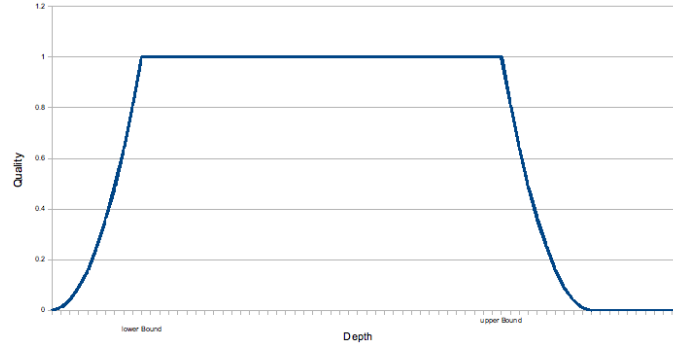


Figure 4: Graphical representation of the  $q_{spec}$  quality function

only uses those and leaves the evaluation of adjectives and adverbs for further research. To compute the depth for a given word a first approach was to determine the average depth of all possible senses which are present in WordNet. The resulting distribution is shown in figure 3. It is visible that most of the words are in a depth level of about 5. Similar to the Consistency Metrics presented before, an arbitrary lower and upper bound can now be defined and used in a quality function  $q_{spec}$ . As an example this could be:

$$q_{spec}(x) = \begin{cases} 0.0, & \text{if } depth(x) > bound_{lower} + bound_{upper} \end{cases} \quad (4)$$

$$\min(1.0, (\frac{depth(x)}{bound_{lower}})^2, \frac{depth(x)}{(bound_{upper} + bound_{lower})^2}), \text{ else} \quad (5)$$

as depicted in figure 4. This function provides the same advantages as mentioned for  $q_{cons}$  before, but punishes deviations into both directions now. Alternatively, a function which is cut off only on one side could be used if it is assumed that only a superficial labeling style causes problems.

The following section will discuss problems of the methods described so far and how they were further optimized.

#### 4 Consistency and Specificity Quality using Semantic Relatedness

The problem with the approach explained above is that words in WordNet can have many different meanings and the usage of averages can lead to unwanted results. On the one hand this can distort the consistency quality as a synonym might just not be appropriate for the word under investigation. On the other hand it could lead to a depth for the specificity metric that deviates strongly from the one the specific meaning intended for this word has. Thus, it is necessary to determine the meaning of a word within a label prior to calculating its quality.

Determining the semantics of a word is a well known task in natural language processing. Some of the problems and ways to solve them can be found e.g. in [Sus93, BGBZ07, Yar92]. The main idea is to use the context of a word and to evaluate which meaning is the most probable. In our case, the context of a word in a label is the rest of the label, of course, but also the labels of its predecessors and successors. If one of them should be a join or split node (XOR, OR, AND) the predecessors/successors of that node will be taken into account, too. Afterwards, the semantic relatedness as described in section 2.2 is computed and averaged for each of the possible meanings a word has. The meaning with the highest relatedness to its context is selected. This procedure is executed independently for each word. Thus the meaning which was selected for one word does not influence the selection of a meaning of another word, contrary to the approach taken in [Sus93].

Applying this technique makes the results much more accurate as only relevant synonyms are taken into account for the consistency metric. The distribution of depth values also becomes clearer when only concrete meanings are evaluated. See figure 5.

The only problem is that the disambiguation itself cannot be guaranteed to be always correct and a false meaning could get selected. Therefore, the quality aspects of both ways (using averages and using the selected meaning) will be regarded. Furthermore, the influence of both metrics on the quality of a word will be scaled by a factor  $\alpha$  which can be externally set. The quality of a single word then becomes:

$$q_{word}(x) = \beta * q_{average}(x) + (1 - \beta) * q_{specific}(x^*) \quad (6)$$

where

$$q_{specific}(x) = q_{average}(x) = \alpha * q_{cons}(x) + (1 - \alpha) * q_{spec}(x) \quad (7)$$

but the specific quality uses the selected meaning (here denoted by  $x^*$ ).

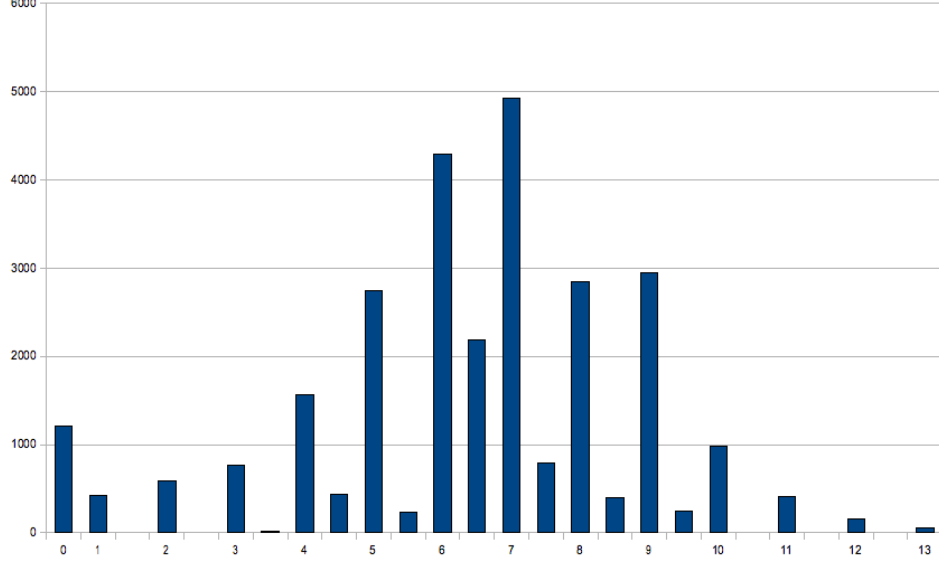


Figure 5: The distribution of depth values after applying the disambiguation technique explained above. Values of .5 exist as a SynSet can have two hypernyms with different levels within the tree (or graph in this case).

## 5 Aggregation on Label and Model Level

To quickly identify and easily visualize these semantic quality metrics it is necessary to aggregate them for a whole label and/or model. This enables the user to quickly identify models with the least quality and to adjust the labeling in a quality assurance process. Our approach was to calculate the arithmetic mean and variance to determine the quality of a given label/model.

$$q_{label}(Label) = \sum_{i=0}^n \frac{q_{word}(Label(i))}{n} \quad (8)$$

where  $n$  is the number of nouns and verbs and  $Label(i)$  denotes the  $i$ -th noun or verb within the label.

$$\sigma_{label}^2(Label) = \sum_{i=0}^n \frac{(q_{word}(Label(i)) - q_{label}(Label))^2}{n} \quad (9)$$

The same applies for the model level:

$$q_{model}(Model) = \sum_{i=0}^m \frac{q_{label}(Model(i))}{m} \quad (10)$$



where  $m$  is the number of labeled elements (in the case of EPCs Functions and Events) and  $Model(i)$  refers to the  $i$ -th label within the model under investigation.

$$\sigma_{model}^2(Model) = \sum_{i=0}^m \frac{(q_{word}(Model(i)) - q_{label}(Model(i)))^2}{m} \quad (11)$$

On the one hand models with a low quality should be subject to further quality checking, but also models with a high variance are interesting. They will have a single label that is evaluated as bad by our metric, surrounded by many good labels. Examples for both cases will be presented in the next section.

## 6 Application to the SAP Reference Model

The quality model described before was prototypically implemented and applied to a part of the SAP Reference Model. In particular the 604 EPCs were examined. The parameter  $\alpha$  was set to 0.5 so the consistency and specificity metrics were regarded equally important. The degree to which the quality using the specific meaning is used  $\beta$  was set to 0.8. A minimum occurrence of 51% was demanded for the consistency quality  $q_{cons}$  and the lower and upper bound for the specificity metric were set to the value of the lowest and highest quantile.

One of the models with the worst total semantic quality was the one shown in figure 1 with a low total semantic quality score of (0.66). The graphical representation was enhanced with two bars which depict the quality using averages (left) and using the specific meaning (right). When the value for the label drops below 0.8 the bar becomes yellow and below 0.6 red. The main issue with this model is the use of the very specific word "garnishment" which lies in the 9th level in the WordNet hypernym tree. On the other hand the very generic verb "exists" is used. Interestingly, even in this small model 2 synonyms "wage" and "remuneration" are used alongside. But while "remuneration" was used 13 times throughout the model repository, wage only appeared twice.

From the information that was acquired, now even recommendations can be given to a user to increase the label quality. Synonym usage can be decreased if "remuneration" is also used in the upper left event and in the function. Furthermore the word "exist" could be replaced with one of its hyponyms. Albeit it is hard to make an automated recommendation as "exist" is very unspecific and lies on the 0-th level within the hypernym tree. Lastly, the word "garnishment" could be replaced by one of its hypernyms e.g. "court order" so the model becomes more abstract and aligns better with the other models in the repository.

Another interesting finding was that a high variance usually arises, when the model contains events which were not labeled at all and were therefore evaluated with a semantic quality of 0.0. But there are also models as the one shown in figure 6 where a single function is responsible for the high variance, although the total quality of the model is not strongly affected. This model is also an example for the problems that arise with our metrics. The quality of the function "RFQ/Quotation" is low because the special abbreviation Request for Quotation (RFQ) is not part of WordNet. A way to solve these kinds of problems would be to create a lexical database similar to WordNet, but specialized for a business

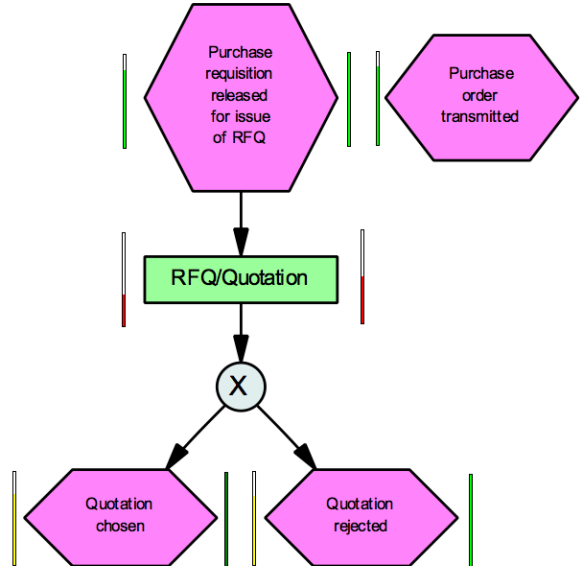


Figure 6: Example of an analyzed EPC with a high variance (0.039) and a semantic quality score in the upper medium range. (0.85)

administration and IT context.

For the following analysis the parameters of the metrics were altered to be a lot stricter with a minimum frequency of 80% and the specificity bounds set to 4.5 and 5.5. The five best and worst labels that resulted from this analysis are depicted in table 2.

By looking into the detailed data for each word within a label it becomes evident why they were evaluated good or bad, although some of the labels seem to be very similar. Table 3 shows in how far the single words of some of the labels shown before do not fulfill the requirements of a consistent labeling style. When interpreting the results it is important to remember that no syntactic features were examined by our quality metric. Thus the label "Distribution" is of high semantic quality regarding its specificity and consistency with the modeling repository, although no clear action can be derived from its syntax. New mechanisms which can be used to tackle those issues are discussed e.g. in [LMS].

The low score of the word "was" could be problematic in the last case as it is used as an auxiliary verb to form the past tense. Although the label could be improved by changing it to "Created Revaluation" the exclusion of some basic vocabulary would prevent such issues.

label	avg. Quality	spec. Quality	total Quality
Material BOM Distribution	0.92	1.00	0.98
Distribution	0.87	1.00	0.97
Electronic Bank Statement	0.98	0.95	0.95
Transaction data	0.95	0.95	0.95
Document Distribution ALE	0.79	0.99	0.95
RFQ/Quotation	0.26	0.06	0.09
Write-up	0.46	0.07	0.15
Post-capitalization	0.38	0.09	0.15
Usage Decision	0.34	0.14	0.18
Revaluation was made	0.20	0.22	0.21

Table 2: The five best/worst labels

word	usage frequency	depth	synonym frequency	possible Hypernym/Troponym
Material	100%	4.5	0% (stuff)	-
BOM	-	-	-	-
Distribution	100%	5.0	no synonyms	-
Transaction	100%	5.5	0% (dealing)	-
data	100%	4.0	0% (information)	raw data; metadata
Usage	51%	6.0	28%(exercise)	activity
Decision	14%	7.0	84%(determination)	choice; selection
Revaluation	100%	8.0	0% (review)	assessment
was	8%	0.0	61%(is) 15%(be)	$\geq 100$
made	10%	0.0	90%(create)	$\geq 50$

Table 3: Closer analysis of the labels

## 7 Related Research

One of the first models used for describing the quality of a process was the one by Lindland et al [LSS94]. As our approach tries to increase the understandability of the model for the audience, it is a part of the pragmatic quality mentioned in this work. Although Lindland also defines the term "semantic quality", meaning the congruence of the model and the domain which has to be modeled, it is not related to what is addressed as semantic label quality in this paper. But, our metrics entirely depend on the WordNet semantic database and try to automatically determine the meaning of a word from its given context. Hence, the term semantic label quality seemed appropriate.

So far quantitative evaluation approaches concentrated on the evaluation of the structure of the process model e.g. [VRM<sup>+</sup>08]. A good overview provides [VCR<sup>+</sup>] and an empirical correlation analysis between understandability and structural features can be found in [MS08]. Some research especially on the impact of labeling and the structure of labels was recently conducted by Jan Mendling [MR08a, MR08b, MRR09]. So far no research that tries to leverage the information of WordNet for the improvement of Business Process Model labels is known to the author.

## 8 Conclusion

In this paper a quantitative approach for determining the quality of process model labels was presented. The metrics developed utilized the information that is available in the WordNet lexical database, with the aim of minimizing misunderstandings between several stakeholders. To achieve this, ambiguity by the usage of synonyms or words from different levels of specificity is penalized. To test the approach defined here it was applied to the EPCs, of the SAP Reference Model with promising results. Although it has been applied to EPCs the model can be seamlessly transferred onto other modeling notations like BPMN, YAWL or UML Activity Diagrams. A different procedure was developed at the University of Muenster [DHLS08], where the agreement upon a restricted set of words and syntactical constructs has to be found before starting to model. While modeling, these constructs are then enforced by a modeling tool. Although this approach guarantees consistent results, it is quite restrictive for the modeler himself. In contrast to that, our approach ensures that a stringent und unified labeling style emerges through agreement between the different parties involved in the modeling process. Thus, the costly process of defining a set of allowed words prior to modeling can be omitted. Furthermore, recommendations on how to improve the alignment of a model's label with the rest of the repository can be automatically derived from the metrics presented. Therefore, it also addresses the main characteristics which are demanded in [Reb09] (dynamic, evolutionary and community-based) for an approach to handle semantic ambiguity.

## 8.1 Limitations

Some general problems arise with the use of our methodology. As briefly discussed some domain specific terms that are completely understandable for a domain expert are not contained in WordNet. Another example is the word "R/3" which is specific to the company of SAP and was mapped to the SynSet containing the physical gas constant "r". This also shows that the simple disambiguation heuristic introduced in section 4 is not perfect as well and could be exchanged for a more sophisticated one. But, due to the fact that both WordNet and the BNC corpus, used for the experiments, were designed for general English such problems will always arise. A considerable reduction can probably only be achieved by extending WordNet with domain specific terms and by changing the corpus that is used to determine the probabilities for the semantic relatedness measure. A good starting point for such a corpus would be e.g. a reference manual or domain specific literature.

Another problem is, that semantic relatedness strongly depends on subjective evaluations. Thus, even if in general the computed similarity is accepted, individual persons could have a different perspective and could object to the identified relations. Another alternative is to give the modeler the possibility to adjust the meaning of a word, if he does not agree with the automatic determination. Finally, as the semantic quality and understandability of a model strongly depend on subjective evaluations, too, these metrics can never give a perfect answer, but they are able to point a user or modeler to peculiar labels and provide hints to increase the general understandability and alignability of the model.

## 8.2 Further Research

Further research could include identifying the possibilities of the information within WordNet which were not used in this paper. Additionally, concepts that are already present, like the determined similarity relation between the labels, could be used to identify labels which do not seem to have a high correlation to their environment. This could lead to the identification of parts that should probably be extracted to a sub-model or integrated with other parts. Another focus will be to conduct an empirical study to verify the usefulness of the metrics.

To conclude, it will also be necessary to combine the measures presented above with techniques that also evaluate the syntactical structure of the label that is investigated, to provide an overall quantitative quality metric. An attempt to that is currently being evaluated at our institute [LMS].

## References

- [BGBZ07] Jordan Boyd-Graber, David Blei, and Xiaojin Zhu. A Topic Model for Word Sense Disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1024–1033, Prague, Czech Republic, June 2007. Association for Computational Linguistics.
- [BH01] A. Budanitsky and Graeme Hirst. Semantic Distance in WordNet: An Experimental, Application-oriented Evaluation of Five Measures, 2001.
- [DHLS08] Patrick Delfmann, Sebastian Herwig, Lukasz Lis, and Armin Stein. Eine Methode zur formalen Spezifikation und Umsetzung von Bezeichnungskonventionen für fachkonzeptionelle Informationsmodelle. pages 23–38, 2008.
- [HS06] I. Hadar and P. Soffer. Variations in conceptual modeling: classification and ontological analysis. In *JAIS*, pages 568–592, 2006.
- [JC97] J. J. Jiang and D. W. Conrath. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *International Conference Research on Computational Linguistics (ROCLING X)*, pages 9008+, September 1997.
- [KJ03] J. Krogstie and H.D. Jorgensen. Quality of Interactive Models. In *Conceptual Modeling - ER 2002, Workshops of the 21st International Conference on Conceptual Modeling, Tampere, Finland*, volume LNCS 2784, pages 351–363. Springer Verlag Berlin Heidelberg, 2003.
- [KSJ06] John Krogstie, Guttorm Sindre, and Håvard Jørgensen. Process models representing knowledge for action: a revised quality framework. *Eur. J. Inf. Syst.*, 15(1):91–102, 2006.
- [Lin98] Dekang Lin. An Information-Theoretic Definition of Similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann, 1998.
- [LMS] Henrik Leopold, Jan Mendling, and Sergey Smirnov. Measuring Label Quality using Part of Speech Tagging. to be published autumn/winter 2009.
- [LSS94] Odd Ivar Lindland, Guttorm Sindre, and Arne Sølvberg. Understanding Quality in Conceptual Modeling. *IEEE Software*, 11(2):42–49, 1994.
- [Mil95] George A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.
- [MR08a] Jan Mendling and Jan C. Recker. Towards Systematic Usage of Labels and Icons in Business Process Models. In Terry Halpin, Erik Proper, John Krogstie, Xavier Franch, Ela Hunt, and Remi Coletta, editors, *CAiSE 2008 Workshop Proceedings - Twelfth International Workshop on Exploring Modeling Methods in Systems Analysis and Design (EMMSAD 2008)*, volume 337, pages 1–13. CEUR-WS.org, June 16-17 2008.
- [MR08b] Jan Mendling and Hajo A. Reijers. The Impact of Activity Labeling Styles on Process Model Quality. In *SIGSAND-EUROPE*, pages 117–128, 2008.
- [MRR09] J. Mendling, H. A. Reijers, and J. Recker. Activity labeling in process modeling: Empirical insights and recommendations. *Information Systems*, April 2009.

- [MS08] Jan Mendling and Mark Strembeck. Influence Factors of Understanding Business Process Models. In Dieter Fensel Witold Abramowicz, editor, *Business Information Systems, 11th International Conference, BIS 2008, Innsbruck, Austria, May 2008*, pages 142–153. Springer-Verlag, 2008.
- [Pfe08] D. Pfeiffer. Semantic Business Process Analysis - Building Block-based Construction of Automatically Analyzable Business Process Models. Münster, 2008.
- [Por80] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
- [Reb09] Michael Rebstock. Technical opinion: Semantic ambiguity: Babylon, Rosetta or beyond? *Commun. ACM*, 52(5):145–146, 2009.
- [Scr06] Aaron D. Scriver. Semantic Distance in WordNet: A Simplified and Improved Measure of Semantic. Master’s thesis, University of Waterloo, Waterloo, Ontario, Canada, 2006.
- [Sus93] Michael Sussna. Word sense disambiguation for free-text indexing using a massive semantic network. In *proceedings of the second international conference on Information and knowledge management (CIKM ’93)*, pages 67–74, New York, NY, USA, 1993. ACM.
- [VCR<sup>+</sup>] Irene Vanderfeesten, Jorge Cardoso, Hajo A. Reijers, Wil Van Der Aalst, and Jan Mendling. Quality Metrics for Business Process Models.
- [VRM<sup>+</sup>08] Irene Vanderfeesten, Hajo Reijers, Jan Mendling, Wil van der Aalst, and Jorge Cardoso. On a Quest for Good Process Models: The Cross-Connectivity Metric. *Advanced Information Systems Engineering*, pages 480–494, 2008.
- [Yar92] David Yarowsky. Word-sense disambiguation using statistical models of Roget’s categories trained on large corpora. In *Proceedings of the 14th conference on Computational linguistics*, pages 454–460, Morristown, NJ, USA, 1992. Association for Computational Linguistics.