Large-scale Semantic Networks: Annotation and Evaluation

Václav Novák
Institute of Formal and Applied Linguistics
Charles University in Prague, Czech Republic
novak@ufal.mff.cuni.cz

Sven Hartrumpf
Computer Science Department
University of Hagen, Germany
Sven.Hartrumpf@FernUni-Hagen.de

Keith Hall*
Google Research
Zürich, Switzerland
kbhall@google.com

Abstract

We introduce a large-scale semantic-network annotation effort based on the MutliNet formalism. Annotation is achieved via a process which incorporates several independent tools including a MultiNet graph editing tool, a semantic concept lexicon, a user-editable knowledge-base for semantic concepts, and a MultiNet parser. We present an evaluation metric for these semantic networks, allowing us to determine the quality of annotations in terms of inter-annotator agreement. We use this metric to report the agreement rates for a pilot annotation effort involving three annotators.

1 Introduction

In this paper we propose an annotation framework which integrates the MultiNet semantic network formalism (Helbig, 2006) and the syntactico-semantic formalism of the Prague Dependency Treebank (Hajič et al., 2006) (PDT). The primary goal of this task is to increase the interoperability of these two frameworks in order to facilitate efforts to annotate at the semantic level while preserving intra-sentential semantic and syntactic annotations as are found in the PDT.

The task of annotating text with global semantic interactions (e.g., semantic interactions within some discourse) presents a cognitively demanding problem. As with many other annotation formalisms, we propose a technique that builds from cognitively simpler tasks such as syntactic and semantic annotations at the sentence level including rich morphological analysis. Rather than constraining the semantic representations to those compatible with the sentential annotations, our procedure provides the syntacitco-semantic tree as a reference; the annotators are free to select nodes from this tree to create nodes in the network. We do not attempt to measure the influence this procedure has on the types of semantic networks generated. We believe that using a soft-constraint such as the syntactico-semantic tree, allows us to better generate human labeled semantic networks with links to the interpretations of the individual sentence analyses.

In this paper, we present a procedure for computing the annotator agreement rate for MultiNet graphs. Note that a MultiNet graph does not represent the same semantics as a syntactico-semantic dependency tree. The nodes of the MultiNet graph are connected based on a corpus-wide interpretation of the entities referred to in the corpus. These global connections are determined by the intra-sentential interpretation but are not restricted to that interpretation. Therefore, the procedure for computing annotator agreement differs from the standard approaches to evaluating syntactic and semantic dependency treebanks (e.g., dependency link agreement, label agreement, predicate-argument structure agreement).

As noted in (Bos, 2008), "*Even though the design of annotation schemes has been initiated for single semantic phenomena, there exists no annotation scheme (as far as I know) that aims to inte-*

*Part of this work was completed while at the Johns Hopkins University Center for Language and Speech Processing in Baltimore, MD USA.

Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions, pages 37–45,
Boulder, Colorado, June 2009.

grate a wide range of semantic phenomena all at once. It would be welcome to have such a resource at ones disposal, and ideally a semantic annotation scheme should be multi-layered, where certain semantic phenomena can be properly analysed or left simply unanalysed."

In Section 1 we introduce the theoretical background of the frameworks on which our annotation tool is based: MultiNet and the Tectogrammatical Representation (TR) of the PDT. Section 2 describes the annotation process in detail, including an introduction to the encyclopedic tools available to the annotators. In Section 3 we present an evaluation metric for MultiNet/TR labeled data. We also present an evaluation of the data we have had annotated using the proposed procedure. Finally, we conclude with a short discussion of the problems observed during the annotation process and suggest improvements as future work.

1.1 MultiNet

The representation of the Multilayered Extended Semantic Networks (MultiNet), which is described in (Helbig, 2006), provides a universal formalism for the treatment of semantic phenomena of natural language. To this end, they offer distinct advantages over the use of the classical predicate calculus and its derivatives. For example, MultiNet provides a rich ontology of *semantic-concept types*. This ontology has been constructed to be language independent. Due to the graphical interpretation of MultiNets, we believe manual annotation and interpretation is simpler and thus more cognitively compatible. Figure 1 shows the MultiNet annotation of a sentence from the WSJ corpus: **"Stephen Akerfeldt, currently vice president finance, will succeed Mr. McAlpine."**
In this example, there are a few relationships that illustrate the representational power of MultiNet. The main predicate *succeed* is a **ANTE** dependent of the node *now*, which indicates that the outcome of the event described by the predicate occurs at some time later than the time of the statement (i.e., the succession is taking place after the current time as captured by the future tense in the sentence). Intra-sentential coreference is indicated by the **EQU** relationship. From the previous context, we know that the *vice president* is related to a particular company, Magna International Inc. The pragmatically defined relationship between *Magna International Inc.* and *vice president finance* is captured by the **ATTCH** (conceptual attachment) relationship. This indicates that there is some relationship between these entities for which one is a member of the other (as indicated by the directed edge). *Stephen Akerfeldt* is the agent of the predicate described by this sub-network.

The semantic representation of natural language expressions by means of MultiNet is generally independent of the considered language. In contrast, the syntactic constructs used in different languages to express the same content are obviously not identical. To bridge the gap between different languages we employ the deep syntactico-semantic representation available in the Functional Generative Description framework (Sgall et al., 1986).

1.2 Prague Dependency Treebank

The Prague Dependency Treebank (PDT) presents a language resource containing a deep manual analysis of texts(Sgall et al., 2004). The PDT contains annotations on three layers:

Morphological A rich morphological annotation is provided when such information is available in the language. This includes lemmatization and detailed morphological tagging.

Analytical The analytical layer is a dependency analysis based purely on the syntactic interpretation.

Tectogrammatical The tectogrammatical annotation provides a deep-syntactic (syntactico-semantic) analysis of the text. The formalism abstracts away from word-order, function words (syn-semantic words), and morphological variation.

The units of each annotation level are linked with corresponding units on the preceding level. The morphological units are linked directly with the original tokenized text. Linking is possible as most of these interpretations are directly tied to the words in the original sentence. In MultiNet graphs, additional nodes are added and nodes are removed.

The PDT 2.0 is based on the long-standing Praguian linguistic tradition, adapted for the current

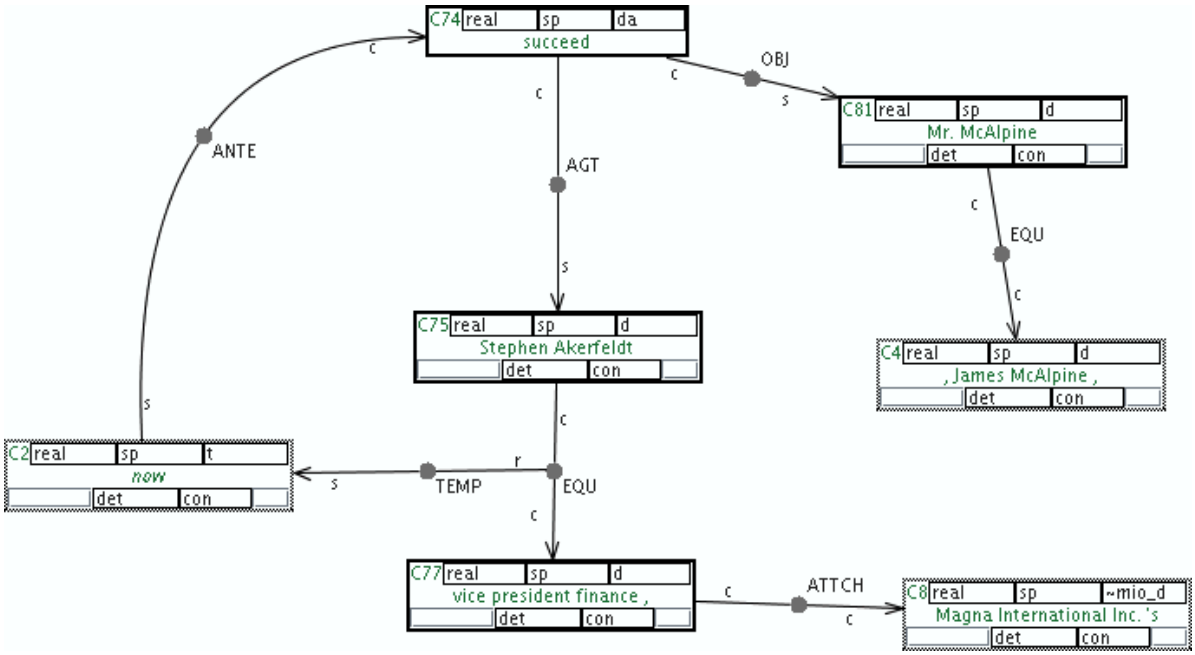

Figure 1: MultiNet annotation of sentence "Stephen Akerfeldt, currently vice president finance, will succeed Mr. McAlpine." Nodes C4 and C8 are re-used from previous sentences. Node C2 is an unexpressed (not explicitly stated in the text) annotator-created node used in previous annotations.

computational-linguistics research needs. The theoretical basis of the tectogrammatical representation lies in the Functional Generative Description of language systems (Sgall et al., 1986). Software tools for corpus search, lexicon retrieval, annotation, and language analysis are included. Extensive documentation in English is provided as well.

2 Integrated Annotation Process

We propose an integrated annotation procedure aimed at acquiring high-quality MultiNet semantic annotations. The procedure is based on a combination of annotation tools and annotation resources. We present these components in the this section.

2.1 Annotation Tool

The core annotation is facilitated by the *cedit* tool[1], which uses PML (Pajas and Štěpánek, 2005), an XML file format, as its internal representation (Novák, 2007). The annotation tool is an application with a graphical user interface implemented in Java (Sun Microsystems, Inc., 2007). The *cedit* tool is platform independent and directly connected to the annotators' wiki (see Section 2.4), where annotators can access the definitions of individual MultiNet semantic relations, functions and attributes; as well as examples, counterexamples, and discussion concerning the entity in question. If the wiki page does not contain the required information, the annotator is encouraged to edit the page with his/her questions and comments.

2.2 Online Lexicon

The annotators in the semantic annotation project have the option to look up examples of MultiNet structures in an online version of the semantically oriented computer lexicon HaGenLex (Hartrumpf et al., 2003). The annotators can use lemmata (instead of reading IDs formed of the lemma and a numerical suffix) for the query, thus increasing the recall of related structures. English and German input is supported with outputs in English and/or German; there are approximately 3,000 and 25,000 semantic networks, respectively, in the lexicon. An example sentence for the German verb "borgen.1.1" ("to borrow") plus its automatically generated and val-

[1]The *cedit* annotation tool can be downloaded from http://ufal.mff.cuni.cz/~novak/files/cedit.zip.

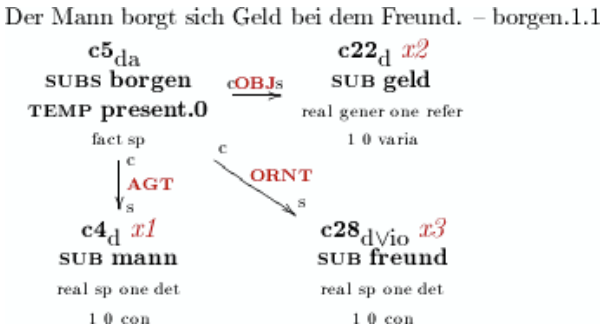

Figure 2: HaGenLex entry showing an example sentence for the German verb "borgen.1.1" ("to borrow"). The sentence is literally "The man borrows himself money from the friend."

idated semantic representation is displayed in Figure 2. The quality of example parses is assured by comparing the marked-up complements in the example to the ones in the semantic network. In the rare case that the parse is not optimal, it will not be visible to annotators.

2.3 Online Parser

Sometimes the annotator needs to look up a phrase or something more general than a particular noun or verb. In this case, the annotator can use the workbench for (MultiNet) knowledge bases (MWR (Gnörlich, 2000)), which provides convenient and quick access to the parser that translates German sentences or phrases into MultiNets.

2.4 Wiki Knowledge Base

A wiki (Leuf and Cunningham, 2001) is used collaboratively to create and maintain the knowledge base used by all the annotators. In this project we use Dokuwiki (Badger, 2007). The entries of individual annotators in the wiki are logged and a feed of changes can be observed using an RSS reader. The *cedit* annotation tool allows users to display appropriate wiki pages of individual relation types, function types and attributes directly from the tool using their preferred web browser.

3 Network Evaluation

We present an evaluation which has been carried out on an initial set of annotations of English articles from *The Wall Street Journal* (covering those annotated at the syntactic level in the Penn Treebank (Marcus et al., 1993)). We use the annotation from the Prague Czech-English Dependency Treebank (Cuřín et al., 2004), which contains a large portion of the WSJ Treebank annotated according to the PDT annotation scheme (including all layers of the FGD formalism).

We reserved a small set of data to be used to train our annotators and have excluded these articles from the evaluation. Three native English-speaking annotators were trained and then asked to annotate sentences from the corpus. We have a sample of 67 sentences (1793 words) annotated by two of the annotators; of those, 46 sentences (1236 words) were annotated by three annotators.[2] Agreement is measured for each individual sentences in two steps.

First, the best match between the two annotators' graphs is found and then the F-measure is computed. In order to determine the optimal graph match between two graphs, we make use of the fact that the annotators have the tectogrammatical tree from which they can select nodes as concepts in the MultiNet graph. Many of the nodes in the annotated graphs remain linked to the tectogrammatical tree, therefore we have a unique identifier for these nodes. When matching the nodes of two different annotations, we assume a node represents an identical concept if both annotators linked the node to the same tectogrammatical node. For the remaining nodes, we consider all possible one-to-one mappings and construct the optimal mapping with respect to the F-measure.

Formally, we start with a set of tectogrammatical trees containing a set of nodes N. The annotation is a tuple $G = (V, E, T, A)$, where V are the vertices, $E \subseteq V \times V \times P$ are the directed edges and their labels (e.g., agent of an action: AGT $\in P$), $T \subseteq V \times N$ is the mapping from vertices to the tectogrammatical nodes, and finally A are attributes of the nodes, which we ignore in this initial evaluation.[3] Analogously, $G' = (V', E', T', A')$ is another annotation

[2]The data associated with this experiment can be downloaded from http://ufal.mff.cuni.cz/~novak/files/data.zip. The data is in *cedit* format and can be viewed using the *cedit* editor at http://ufal.mff.cuni.cz/~novak/files/cedit.zip.

[3]We simplified the problem also by ignoring the mapping from edges to tectogrammatical nodes and the MultiNet edge attribute *knowledge type*.

of the same sentence and our goal is to measure the similarity $s(G, G') \in [0, 1]$ of G and G'.

To measure the similarity we need a set Φ of admissible one-to-one mappings between vertices in the two annotations. A mapping is admissible if it connects vertices which are indicated by the annotators as representing the same tectogrammatical node:

$$\Phi = \Big\{ \phi \subseteq V \times V' \Big| \underset{\substack{n \in N \\ v \in V \\ v' \in V'}}{\forall} \Big(\big((v,n) \in T \wedge (v',n) \in T'\big) \rightarrow (v,v') \in \phi \Big) \wedge \underset{\substack{v \in V \\ v',w' \in V'}}{\forall} \Big(\big((v,v') \in \phi \wedge (v,w') \in \phi\big) \rightarrow (v' = w') \Big) \wedge \underset{\substack{v,w \in V \\ v' \in V'}}{\forall} \Big(\big((v,v') \in \phi \wedge (w,v') \in \phi\big) \rightarrow (v = w) \Big) \Big\} \quad (1)$$

In Equation 1, the first condition ensures that Φ is constrained by the mapping induced by the links to the tectogrammatical layer. The remaining two conditions guarantee that Φ is a one-to-one mapping.

We define the annotation agreement s as:

$$s_F(G, G') = \max_{\phi \in \Phi}(F(G, G', \phi))$$

where F is the F1-measure:

$$F_m(G, G', \phi) = \frac{2 \cdot m(\phi)}{|E| + |E'|}$$

where $m(\phi)$ is the number of edges that match given the mapping ϕ.

We use four versions of m, which gives us four versions of F and consequently four scores s for every sentence:

Directed unlabeled: $m_{du}(\phi) =$

$$\Big| \Big\{ (v,w,\rho) \in E \Big| \exists_{v',w' \in V', \rho' \in P} \Big((v', w', \rho') \in E' \wedge (v, v') \in \phi \wedge (w, w') \in \phi \Big) \Big\} \Big|$$

Undirected unlabeled: $m_{uu}(\phi) =$

$$\Big| \Big\{ (v,w,\rho) \in E \Big| \exists_{v',w' \in V', \rho' \in P} \Big(\big((v', w', \rho') \in E' \vee (w', v', \rho') \in E'\big) \wedge (v, v') \in \phi \wedge (w, w') \in \phi \Big) \Big\} \Big|$$

Directed labeled: $m_{dl}(\phi) =$

$$\Big| \Big\{ (v,w,\rho) \in E \Big| \exists_{v',w' \in V'} \Big((v', w', \rho) \in E' \wedge (v, v') \in \phi \wedge (w, w') \in \phi \Big) \Big\} \Big|$$

Undirected labeled: $m_{ul}(\phi) =$

$$\Big| \Big\{ (v,w,\rho) \in E \Big| \exists_{v',w' \in V'} \Big(\big((v', w', \rho) \in E' \vee (w', v', \rho) \in E'\big) \wedge (v, v') \in \phi \wedge (w, w') \in \phi \Big) \Big\} \Big|$$

These four $m(\phi)$ functions give us four possible F_m measures, which allows us to have four scores for every sentence: s_{du}, s_{uu}, s_{dl} and s_{ul}.

Figure 3 shows that the inter-annotator agreement is not significantly correlated with the position of the sentence in the annotation process. This suggests that the annotations for each annotator had achieved a stable point (primarily due to the annotator training process).

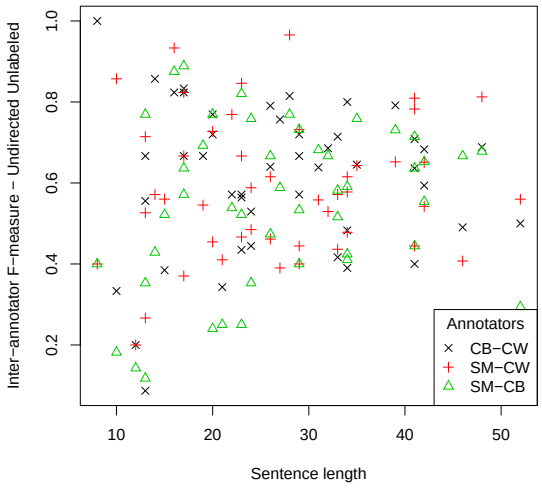

Figure 4: Inter-annotator agreement depending on the sentence length. Each point represents a sentence.

Figure 4 shows that the agreement is not correlated with the sentence length. It means that longer

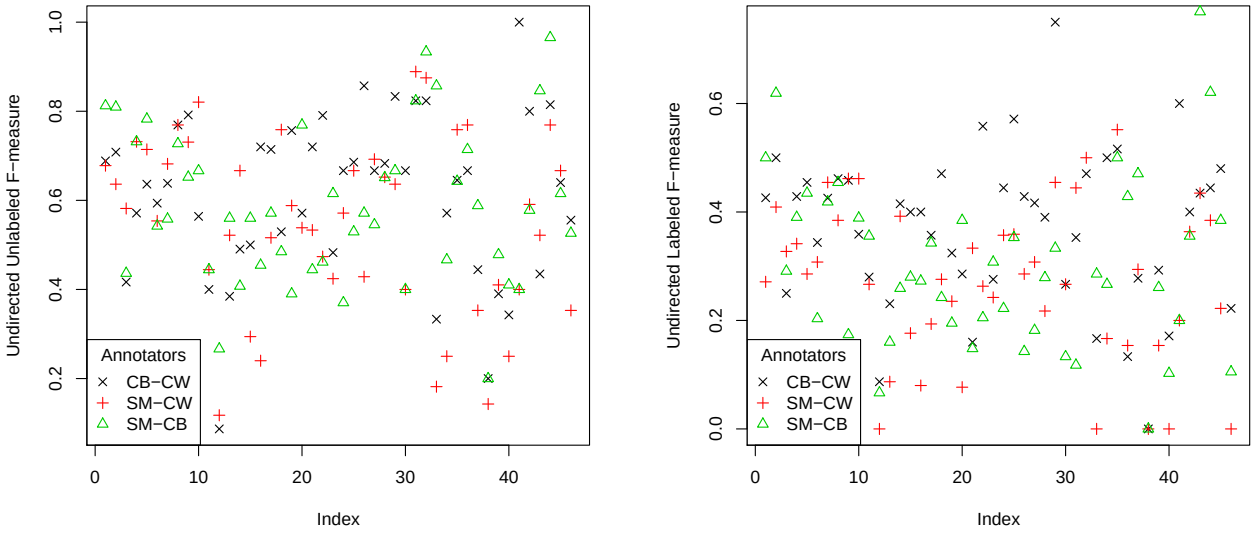

Figure 3: Inter-annotator agreement over time. Left: unlabeled, right: labeled. Each point represents a sentence; CB, CW, and SM are the annotators' IDs.

sentences are not more difficult than short sentences. The variance decreases with the sentence length as expected.

In Figure 5 we show the comparison of directed and labeled evaluations with the undirected unlabeled case. By definition the undirected unlabeled score is the upper bound for all the other scores. The directed score is well correlated and not very different from the undirected score, indicating that the annotators did not have much trouble with determining the correct direction of the edges. This might be, in part, due to support from the formalism and its tool *cedit*: each relation type is specified by a *semantic-concept type* signature; a relation that violates its signature is reported immediately to the annotator. On the other hand, labeled score is significantly lower than the unlabeled score, which suggests that the annotators have difficulties in assigning the correct relation types. The correlation coefficient between s_{uu} and s_{ul} (approx. 0.75) is also much lower than than the correlation coefficient between s_{uu} and s_{du} (approx. 0.95).

Figure 6 compares individual annotator pairs. The scores are similar to each other and also have a similar distribution shape.

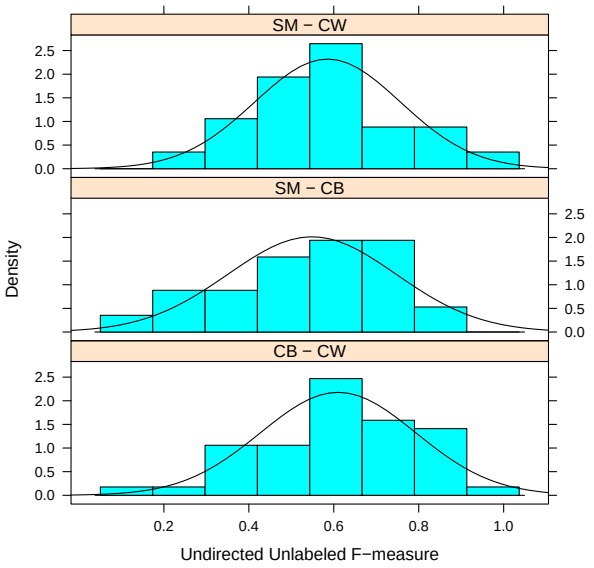

Figure 6: Comparison of individual annotator pairs.

A more detailed comparison of individual annotator pairs is depicted in Figure 7. The graph shows that there is a significant positive correlation between scores, i.e. if two annotators can agree on the

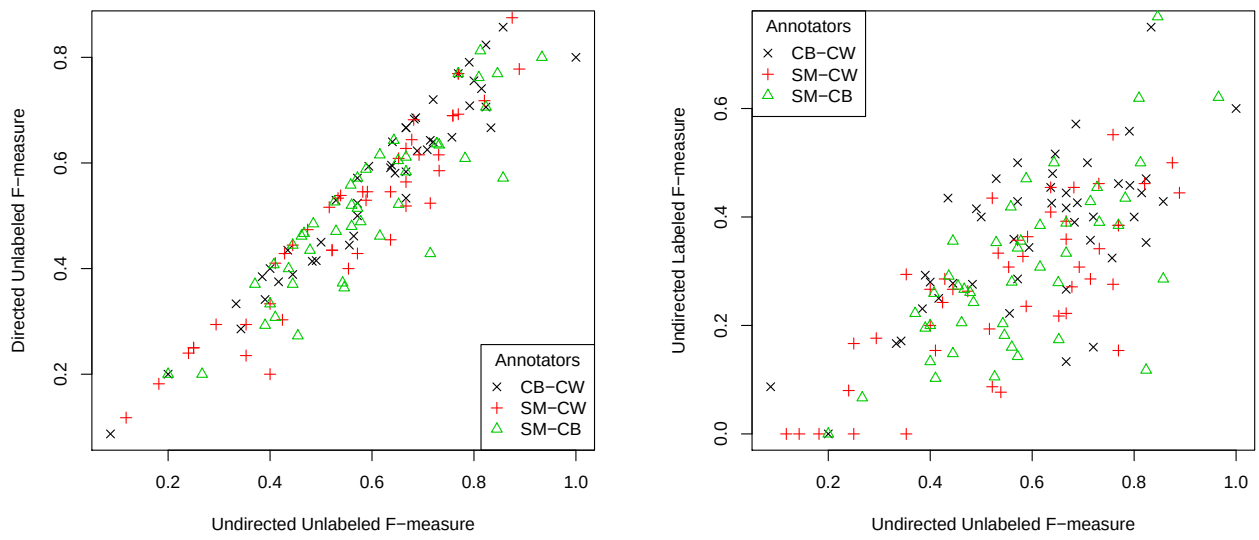

Figure 5: Left: Directed vs. undirected inter-annotator agreement. Right: Labeled vs. unlabeled inter-annotator agreement. Each point represents a sentence.

annotation, the third is likely to also agree, but this correlation is not a very strong one. The actual correlation coefficients are shown under the main diagonal of the matrix.

Sample	Annotators	Agreement F-measure			
		s_{uu}	s_{du}	s_{ul}	s_{dl}
Smaller	CB-CW	61.0	56.3	37.1	35.0
Smaller	SM-CB	54.9	48.5	27.1	25.7
Smaller	SM-CW	58.5	50.7	31.3	30.2
Smaller	average	58.1	51.8	31.8	30.3
Larger	CB-CW	64.6	59.8	40.1	38.5

Table 1: Inter-annotator agreement in percents. The results come from the two samples described in the first paragraph of Section 3.

Finally, we summarize the raw result in Table 1. Note that we report simple annotator agreement here.

4 Conclusion and Future Work

We have presented a novel framework for the annotation of semantic network for natural language discourse. Additionally we present a technique to evaluate the agreement between the semantic networks annotated by different annotators.

Our evaluation of an initial dataset reveals that given the current tools and annotation guidelines, the annotators are able to construct the structure of the semantic network (i.e., they are good at building the directed graph). They are not, however, able to consistently label the semantic relations between the semantic nodes. In our future work, we will investigate the difficulty in labeling semantic annotations. We would like to determine whether this is a product of the annotation guidelines, the tool, or the formalism.

Our ongoing research include the annotation of inter-sentential coreference relationships between the semantic concepts within the sentence-based graphs. These relationships link the local structures, allowing for a complete semantic interpretation of the discourse. Given the current level of consistency in structural annotation, we believe the data will be useful in this analysis.

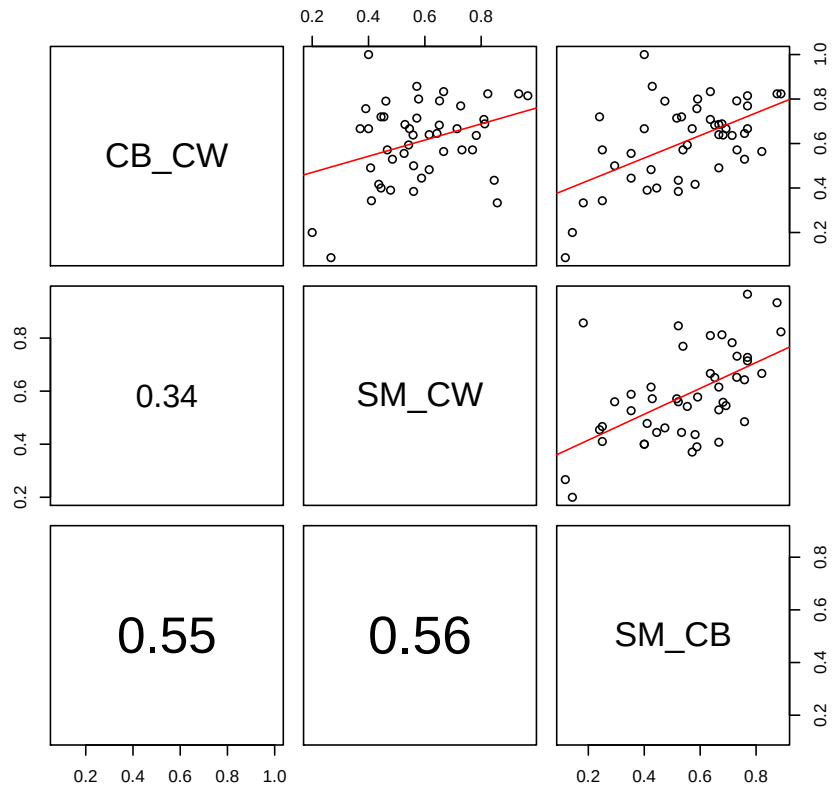

Figure 7: Undirected, unlabeled F-measure correlation of annotator pairs. Each cell represents two different pairs of annotators; cells with graphs show scatter-plots of F-scores for the annotator pairs along with the optimal linear fit; cells with values show the correlation coefficient (each point in the plot corresponds to a sentence). For example, the top row, right-most column, we are comparing the F-score agreement of annotators CB and CW with that of the F-score agreement of annotators SM and CB. This should help identify an outlier in the consistency of the annotations.

Acknowledgment

This work was partially supported by Czech Academy of Science grants 1ET201120505 and 1ET101120503; by Czech Ministry of Education, Youth and Sports projects LC536 and MSM0021620838; and by the US National Science Foundation under grant OISE–0530118. The views expressed are not necessarily endorsed by the sponsors.

References

Mike Badger. 2007. Dokuwiki – A Practical Open Source Knowledge Base Solution. *Enterprise Open Source Magazine*.

Johan Bos. 2008. Let's not Argue about Semantics. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may.

Jan Cuřín, Martin Čmejrek, Jiří Havelka, and Vladislav Kuboň. 2004. Building parallel bilingual syntactically annotated corpus. In *Proceedings of The First International Joint Conference on Natural Language Processing*, pages 141–146, Hainan Island, China.

Carsten Gnörlich. 2000. MultiNet/WR: A Knowledge Engineering Toolkit for Natural Language Information. Technical Report 278, University Hagen, Hagen, Germany.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, and Marie Mikulová. 2006. Prague Dependency Treebank 2.0.

CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia, Pennsylvania.

Sven Hartrumpf, Hermann Helbig, and Rainer Osswald. 2003. The Semantically Based Computer Lexicon HaGenLex – Structure and Technological Environment. *Traitement Automatique des Langues*, 44(2):81–105.

Hermann Helbig. 2006. *Knowledge Representation and the Semantics of Natural Language*. Springer, Berlin, Germany.

Bo Leuf and Ward Cunningham. 2001. *The Wiki Way. Quick Collaboration on the Web*. Addison-Wesley, Reading, Massachusetts.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19(2):313–330.

Václav Novák. 2007. Cedit – semantic networks manual annotation tool. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 11–12, Rochester, New York, April. Association for Computational Linguistics.

Petr Pajas and Jan Štěpánek. 2005. A Generic XML-Based Format for Structured Linguistic Annotation and Its Application to Prague Dependency Treebank 2.0. Technical Report 29, UFAL MFF UK, Praha, Czech Republic.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel, Dordrecht, The Netherlands.

Petr Sgall, Jarmila Panevová, and Eva Hajičová. 2004. Deep syntactic annotation: Tectogrammatical representation and beyond. In Adam Meyers, editor, *Proceedings of the HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 32–38, Boston, Massachusetts, May. Association for Computational Linguistics.

Sun Microsystems, Inc. 2007. *Java Platform, Standard Edition 6*. http://java.sun.com/javase/6/webnotes/README.html.