META-RESEARCH Workshop on Advanced Treebanking

22 May 2012

ABSTRACTS

Editors:

Jan Hajič, Koenraad De Smedt, Marko Tadić, António Branco

Workshop Programme

22 May 2012

09:00 – 10:30 Oral presentations – Session I
co-chaired by Jan Hajič and Koenraad De Smedt

09:00 – 09:10 Jan Hajič *Welcome and Introduction to the Workshop*

09:10 – 09:35 Tom Vanallemeersch *Parser-independent Semantic Tree Alignment*

09:35 – 10:00 Philippe Blache and Stéphane Rauzy
Hybridization and Treebank Enrichment with Constraint-Based Representations

10:00 – 10:25 Bruno Guillaume and Guy Perrier
Semantic Annotation of the French Treebank with Modular Graph Rewriting

10:25 – 10:30 Jan Hajič *Introduction to the Poster Session*

10:30 – 11:30 Coffee break with poster presentations

Victoria Rosén, Koenraad De Smedt, Paul Meurer and Helge Dyvik
An Open Infrastructure for Advanced Treebanking (with demo)

Oleg Kapanadze *A German-Georgian Parallel Treebank Project*

Tomáš Jelínek, Vladimír Petkevič, Alexandr Rosen and Hana Skoumalová
Czech Treebanking Unlimited

João Silva, António Branco, Sérgio Castro and Francisco Costa
Deep, Consistent and also Useful: Extracting Vistas from Deep Corpora for Shallower Tasks

11:30 – 13:00 Oral presentations and invited speech – Session II
co-chaired by Marko Tadić and António Branco

11:30 – 11:45 Hans Uszkoreit, coordinator of META-NET (invited speech)
High-Quality Research, New Language Resources and their Sharing

11:45 – 12:10 Rajesh Bhatt and Fei Xia
Challenges in Converting between Treebanks: a Case Study from the HUTB

12:10 – 12:35 Maytham Alabbas and Allan Ramsay
Arabic Treebank: from Phrase-Structure Trees to Dependency Trees

12:35 – 13:00 Gyri Losnegaard, Gunn Inger Lyse, Martha Thunes, Victoria Rosén, Koenraad De Smedt, Helge Dyvik and Paul Meurer
What We Have Learned from Sofie: Extending Lexical and Grammatical Coverage in an LFG Parsebank

13:00 End of Workshop (start of lunch break)

Workshop Organizers

Jan Hajič	Charles University in Prague, Czech Republic
Koenraad De Smedt	University of Bergen, Norway
Marko Tadić	University of Zagreb, Croatia
António Branco	University of Lisbon, Portugal

Workshop Programme Committee

António Branco	University of Lisbon, Portugal
Sabine Buchholz	Toshiba Research Europe, Cambridge, UK
Khalid Choukri	ELRA/ELDA, Paris, France
Silvie Cinková	Charles University in Prague, Czech Republic
Dan Cristea	University of Iaşi, Romania
Koenraad De Smedt	University of Bergen, Norway
Rebecca Dridan	University of Oslo, Norway
Nancy Ide	Vassar College, New York, USA
Valia Kordoni	DFKI, Berlin, Germany
Sandra Kuebler	Indiana University, Bloomington, USA
Krister Lindén	University of Helsinki, Finland
Paul Meurer	Uni Computing/Uni Research, Bergen, Norway
Adam Meyers	New York University, USA
Joakim Nivre	University of Uppsala, Sweden
Stephan Oepen	University of Oslo, Norway
Marco Passarotti	Catholic University of the Sacred Heart, Milan, It.
Eiríkur Rögnvaldsson	University of Reykjavik, Iceland
Victoria Rosén	University of Bergen, Norway
Mária Šimková	Slovak Academy of Sciences, Bratislava, Slovakia
Barbora Vidová Hladká	Charles University in Prague, Czech Republic
Fei Xia	University of Washington, USA
Daniel Zeman	Charles University in Prague, Czech Republic

Preface

Many R&D projects and research groups are creating, standardizing, converting and/or using treebanks, thereby often tackling the same issues and reinventing methods and tools. While a fair amount of treebanks have been produced in recent years, it is still a challenge for researchers and developers to reuse treebanks in suitable formats for new purposes. Standardization of interchange formats, conversion and adaptation to different purposes, exploration with suitable tools, long term archiving and cataloguing, and other issues still require significant efforts.

In this spirit, the present workshop has been conceived by four projects, namely T4ME, META-NORD, CESAR and META4U, which under the META-NET umbrella project strive to make many treebanks and other language resources and tools available for R&D. It is hoped that the workshop will contribute to innovative insights that will promote development, dissemination, use and reuse of treebanks in the future.

Thirteen papers were submitted to the workshop, of which ten were accepted for presentation at this half-day workshop. Six were selected for oral presentation while four were selected for poster presentation. We thank all our reviewers for their constructive evaluation of the papers.

Jan Hajič
Koenraad De Smedt
Marko Tadić
António Branco

Session I

Tuesday 22 May, 9:00 – 9:10
Chairperson: Koenraad De Smedt

Welcome and Introduction to the META-RESEARCH: Workhop on Advanced Treebanking

Jan Hajič

Session I

Tuesday 22 May, 9:10 – 9:35
Chairperson: Koenraad De Smedt

Parser-independent Semantic Tree Alignment

Tom Vanallemeersch

We describe an approach for training a semantic role labeler through cross-lingual projection between different types of parse trees, with the purpose of enhancing tree alignment on the level of syntactic translation divergences. After applying an existing semantic role labeler to parse trees in a resource-rich language (English), we partially project the semantic information to the parse trees of the corresponding target sentences, based on word alignment. After this precision-oriented projection, we apply a method for training a semantic role labeler which consists in determining a large set of features describing target predicates, roles and predicate-role connections, independently from the type of tree annotation (phrase structure or dependencies). These features describe tree paths starting at or connecting nodes. The semantic role labeling method does not require any knowledge of the parser nor manual intervention. We evaluated the performance of the cross-lingual projection and semantic role labeling using an English parser assigning PropBank labels and Dutch manually annotated parses, and are currently studying ways to use the predicted semantic information for enhancing tree alignment.

Session I

Tuesday 22 May, 9:35 – 10:00
Chairperson: Jan Hajič

Hybridization and Treebank Enrichment with Constraint-Based Representations

Philippe Blache and Stéphane Rauzy

We present in this paper a method for hybridizing constituency treebanks with constraint-based descriptions and enrich them with an evaluation of sentence grammaticality. Such information is calculated thanks to a two-steps technique consisting in: (1) constraint grammar induction from the source treebank and (2) constraint evaluation for all sentences, on top of which a grammaticality index is calculated. This method is theoretically-neutral and language independent. Because of the precision of the encoded information, such enrichment is helpful in different perspectives, for example when designing psycholinguistics experiments such as comprehension or reading difficulty.

Session I

Tuesday 22 May, 10:00 – 10:25
Chairperson: Jan Hajič

Semantic Annotation of the French Treebank with Modular Graph Rewriting

Bruno Guillaume and Guy Perrier

We propose to annotate the French Treebank with semantic dependencies in the framework of DMRS starting from an annotation with surface syntactic dependencies and using modular graph rewriting. This system has been experimented on the whole French Treebank with the prototype which implements the rewriting calculus.

Poster Session Introduction

Tuesday 22 May, 10:25 – 10:30
Chairperson: Jan Hajič

Introduction to the Poster Session

Poster Session (Coffee Break)

Tuesday 22 May, 10:30 – 11.30

An Open Infrastructure for Advanced Treebanking (with demo)

Victoria Rosén, Koenraad De Smedt, Paul Meurer and Helge Dyvik

Increases in the number and size of treebanks, and the complexity of their annotation, present challenges to their exploration by the research community. Adhering to different formalisms, lacking clear standards, and requiring specialized search and visualization and other services, treebanks have not been widely accessible to a broad audience and have remained underexploited. The INESS project is providing the first infrastructure integrating treebank annotation, analysis and distribution, bringing together treebanks for many different languages, spanning different annotation schemes and including parallel treebanks. The infrastructure offers a uniform interface, interactive visualizations, leading edge search capabilities and high performance computing.

Poster Session (Coffee Break)

Tuesday 22 May, 10:30 – 11.30

A German-Georgian Parallel Treebank Project

Oleg Kapanadze

This poster reports about efforts on building a parallel treebank for a typologically dissimilar language pair, namely German and Georgian. The project aims at supporting interdisciplinary collaboration in the field of jurisprudence adding a Natural Language Technology (NLT) angle to the human translation issue. The objective of this project is development of a bilingual Treebank which will be based on the German-Georgian parallel legislative texts.

Poster Session (Coffee Break)

Tuesday 22 May, 10:30 – 11.30

Czech Treebanking Unlimited

Tomáš Jelínek, Vladimír Petkevič, Alexandr Rosen and Hana Skoumalová

We build a large treebank of Czech, avoiding manual effort by using existing parsers, supplemented by a rule-based correction tool. A potentially underspecified morphological and syntactic annotation scheme offers multiple visualisation and export options, customisable in shape and detail according to the preferences of humans or computer applications. The annotation scheme consists of three layers: graphemics, morphology and constituency-based syntax, and is supported by lexicon (with a morphological, multi-word and syntactic part) and grammar. Annotation on any of the interlinked layers can be missing; ambiguous or undecidable phenomena are represented by underspecification and distributive disjunction.

Poster Session (Coffee Break)

Tuesday 22 May, 10:30 – 11.30

Deep, Consistent and also Useful: Extracting Vistas from Deep Corpora for Shallower Tasks

João Silva, António Branco, Sérgio Castro and Francisco Costa

Annotated corpora are fundamental for NLP, and the trend in their development is to move towards datasets with increasingly detailed linguistic annotation. To cope with the complexity of producing such resources, some approaches rely on a supporting deep processing grammar that provides annotation that is rich and consistent over its morphological, syntactic and semantic layers. However, for some purposes, the deep linguistic corpora thus produced are "too deep" and unwieldy. For instance, if one wishes to obtain a probabilistic constituency parser by learning a model over a treebank, the full extent of the annotation created by a deep grammar is not needed and can even be detrimental to training. In this poster, we report on procedures that, starting from a deep dataset produced by a deep processing grammar, extract a variety of vistas---that is, subsets of the information contained in the full dataset. This allows to take a single base dataset as a starting point and deliver a variety of corpora that are more streamlined and focused on particular tasks.

Session II

Tuesday 22 May, 11:30 – 11:45
Chairperson: Marko Tadić

High-Quality Research, New Language Resources and their Sharing

Invited speech

Hans Uszkoreit, main coordinator of META-NET

Session II

Tuesday 22 May, 11:45 – 12:10
Chairperson: Marko Tadić

Challenges in Converting between Treebanks: a Case Study from the HUTB

Rajesh Bhatt and Fei Xia

An important question for treebank development is whether high-quality conversion from one representation (e.g., dependency structure) to another representation (e.g., phrase structure) is possible, assuming that annotation guidelines exist for both representations. In this study, we demonstrate that the conversion is possible only under certain conditions, and even when the conditions are met, the conversion is complex as we need to examine the two sets of guidelines on a phenomenon-by-phenomenon basis and provide an intermediate representation for phenomena with incompatible analysis.

Session II

Tuesday 22 May, 12:10 – 12:35
Chairperson: António Branco

Arabic Treebank: from Phrase-Structure Trees to Dependency Trees

Maytham Alabbas and Allan Ramsay

The aim here is to create a dependency treebank from a phrase-structure treebank for Arabic. Arabic has a number of characteristics, described below, which make it particularly challenging to any natural language processing (NLP) applications. We describe an encouraging semi-automatic technique for converting phrase-structure trees to dependency trees by using a head percolation table. One of the most significant challenges here is the determination of the head of each subtree. We therefore examined different versions of the head percolation table to find the best priority list for each entry in the table. Given that there is no absolute measure of the 'correctness' of a conversion of a phrase structure tree to dependency form, we tested the various transformations by seeing how well a state-of-the-art dependency parser learnt the generalisations that were embodied by the converted trees.

Session II

Tuesday 22 May, 12:35 – 13:00
Chairperson: António Branco

What We Have Learned from Sofie: Extending Lexical and Grammatical Coverage in an LFG Parsebank

Gyri Losnegaard, Gunn Inger Lyse, Martha Thunes, Victoria Rosén, Koenraad De Smedt, Helge Dyvik and Paul Meurer

Constructing a treebank as a dynamically parsed corpus is an iterative process which may effectively lead to improvements of the grammar and lexicon. We show this from our experiences with semiautomatic disambiguation of a Norwegian LFG parsebank. The main types of grammar and lexicon changes necessary for achieving improved coverage are analyzed and discussed. We show that an important contributing factor to missing coverage is missing multiword expressions in the lexicon.