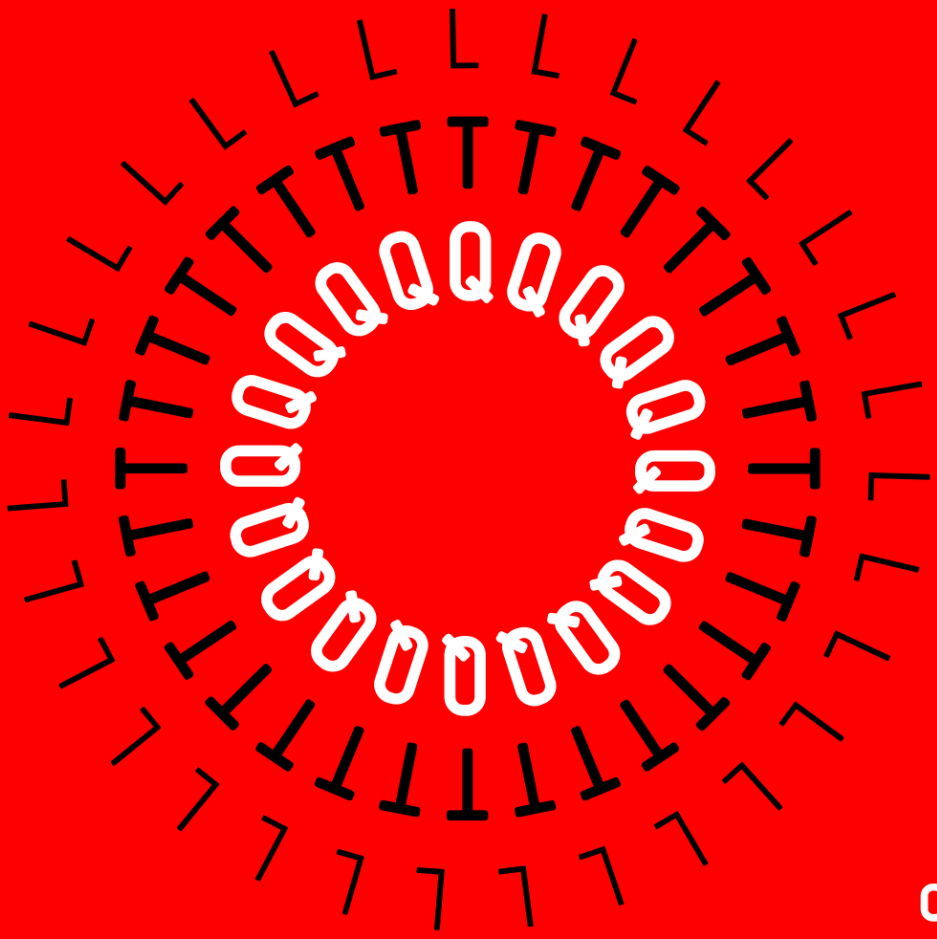


qtleap

quality translation by deep language engineering approaches

Report on the State of the Art concerning Deep Semantic Processing

DELIVERABLE D4.1

VERSION 8.0 | 2015-06-15

QTLeap

Machine translation is a computational procedure that seeks to provide the translation of utterances from one language into another language.

Research and development around this grand challenge is bringing this technology to a level of maturity that already supports useful practical solutions. It permits to get at least the gist of the utterances being translated, and even to get pretty good results for some language pairs in some focused discourse domains, helping to reduce costs and to improve productivity in international businesses.

There is nevertheless still a way to go for this technology to attain a level of maturity that permits the delivery of quality translation across the board.

The goal of the QTLeap project is to research on and deliver an articulated methodology for machine translation that explores deep language engineering approaches in view of breaking the way to translations of higher quality.

The deeper the processing of utterances the less language-specific differences remain between the representation of the meaning of a given utterance and the meaning representation of its translation. Further chances of success can thus be explored by machine translation systems that are based on deeper semantic engineering approaches.

Deep language processing has its stepping-stone in linguistically principled methods and generalizations. It has been evolving towards supporting realistic applications, namely by embedding more data based solutions, and by exploring new types of datasets recently developed, such as parallel DeepBanks.

This progress is further supported by recent advances in terms of lexical processing. These advances have been made possible by enhanced techniques for referential and conceptual ambiguity resolution, and supported also by new types of datasets recently developed as linked open data.

The project QTLeap explores novel ways for attaining machine translation of higher quality that are opened by a new generation of increasingly sophisticated semantic datasets and by recent advances in deep language processing.

www.qtleap.eu

Funded by

QTLeap is funded by the 7th Framework Programme of the European Commission.

Supported by

And supported by the participating institutions:

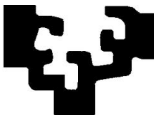
Faculty of Sciences, University of Lisbon

German Research Centre for Artificial Intelligence

Charles University in Prague

Bulgarian Academy of Sciences

Humboldt University of Berlin



University of Basque Country

University of Groningen

Higher Functions, Lda

Revision history

Version	Date	Authors	Organisation	Description
1.0	Jan 01, 2014	Petya Osenova	IICT-BAS	First draft
2.0	Jan 15, 2014	Kiril Simov, Petya Osenova	IICT-BAS	Second draft
3.0	Jan 22, 2014	António Branco	FCUL	Second draft with Integrated feedback
4.0	Jan 31, 2014	Kostadin Cholakov, Valia Kordoni, Markus Egg	UBER	Feedback from internal review integrated
5.0	Feb 7, 2014	Kepa Sarasola, Aljoscha Burchardt, Martin Popel, Gertjan van Noord	UPV-EHU, DFKI, CUNI, UG	Integrated feedback
5.5	Oct 03, 2014	Petya Osenova	IICT-BAS	Addition of Section 6 Expected Benefits for the Real Usage Scenario in response to M6 reviewers' remarks
5.7	Oct 07, 2014	Markus Egg	UBER	Feedback integrated
6.0	Oct 11, 2014	Kepa Sarasola, Eneko Agirre and Nora Aranberri	UPV-EHU	Integrated feedback
6.5	Oct 14, 2014	João Silva	FCUL	Integrated feedback
6.5	Oct 14, 2014	Aljoscha Burchardt	DFKI	Integrated feedback
7.0	Oct 14, 2014	Dieke Oele, Gertjan van Noord	UG	Integrated feedback
8.0	May 27, 2015	João Silva	FCUL	Integrated feedback
8.0	May 27, 2015	Aljoscha Burchardt	DFKI	Integrated feedback
8.0	June 1, 2015	Dieke Oele, Gertjan van Noord	Integrated feedback	
8.0	June 5, 2015	Kepa Sarasola, Eneko Agirre and Nora Aranberri	UPV-EHU	Integrated feedback
8.0	June 8, 2015	Petya Osenova, Kiril Simov	IICT-BAS	Integrated feedback
8.0	June 15, 2015	Markus Egg, Kostadin Cholakov	UBER	Integrated reviewer feedback

Statement of originality
This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Report on the State of the Art concerning Deep Semantic Processing

DOCUMENT QTLEAP-2015-D4.1
EC FP7 PROJECT #610516

DELIVERABLE D4.1

completition
FINAL
status
SUBMITTED
dissemination level
PUBLIC

responsible
Kiril Simov (WP4 Coordinator)
reviewers
Kostadin Cholakov, Valia Kordoni, Markus Egg
contributing partners
FCUL, DFKI, IICT-BAS, UBER, UPV-EHU, UG

authors
Kiril Simov, Petya Osenova

Contents

1 Introduction **8**

2 Deep Dependency Representation and Processing **11**
2.1 Prague Dependency Treebank: Tectogrammatical Layer 11
2.1.1 Tectogrammatical Layer Representation 11
2.1.2 Tectogrammatical Layer Processing 12
2.2 PropBank: Predicate-Argument Layer 13
2.2.1 Propbank Representation 13
2.2.2 Propbank-enhanced Processing 15
2.3 Comparison with the PDT representation 16

3 Underspecified Semantics: Minimal Recursion Semantics **16**
3.1 Minimal Recursion Semantics Representation 16
3.1.1 MRS Definition 16
3.1.2 RMRS Definition 17
3.2 Deepbanks 19
3.3 MRS Processing 19
3.4 MRS for Deep Semantic Transfer in MT 21
3.5 Related Underspecified Semantic Formalisms 22

4 Groningen Meaning Bank (GMB) **22**

5 Beyond the State of the Art **23**

6 Expected Benefits for the Real Usage Scenario **25**
6.1 Problems that Relate to Both Translation Directions. 25

7 Conclusions **31**

List of Abbreviations

AMR	Abstract Meaning Representation
CCG	Combinatory Categorical Grammar
CSLI	Center for the Study of Language and Information, Stanford University, USA
DRT	Discourse Representation Theory
EP	Elementary Predicate
ERG	English Resource Grammar
GMB	Groningen Meaning Bank
HPSG	Head-driven Phrase Structure Grammar
LOD	Linked Open Data
LFG	Lexical Functional Grammar
MRS	Minimal Recursion Semantics
MT	Machine Translation
MWE	Multiword Expressions
NLP	Natural Language Processing
OOV	Out-Of-Vocabulary
PCFG	Probabilistic Context Free Grammar
POS	Part-Of-Speech
PDT	Prague Dependency Treebank
RMRS	Robust Minimal Recursion Semantics
SRL	Semantic Role Labeling
SNoW	Sparse Network of Winnows
WSD	Word Sense Disambiguation

1 Introduction

Deep Semantic Processing plays a central role in many robust NLP applications, such as information extraction, text mining, semantic web exploration, big data understanding, etc. It also plays a crucial role in Machine Translation, especially in the definition of transfer models for high quality translated content. These transfer models are always considered 'deep' in the sense that they would always operate on the top level of the Vauquois triangle (Fig. 1, below). However, they would vary in the different configurations of 'deep' models, thus combining deep syntax with shallow semantics; deep semantics with discourse elements, etc. In this deliverable we present the main achievements in deep semantic processing, considered in the broader context of integrating semantics to language technology. In the last decade this area has been under dynamic development. Following Le and Zuidema [2012], we define semantic processing as *the task of translating natural language sentences to formal meaning representations* (which is not the same as offering a full interpretation). The first question raised by many researchers is the one of which phenomena belong to semantics. As a first (possibly not yet exhaustive) answer to this question, we have combined the lists from Hajič [2011] and Bos [2013]. The lists enumerate phenomena crucial for semantic representation, including some that are pragmatic in nature, but interact with the semantic contribution of the text:

- Semantic Roles (words vs. predicates)
- Lexical Semantics (WSD), MWE
- Metonymy
- Named Entities
- Co-reference (pronominal, bridging anaphora)
- Verb Phrase Ellipsis
- Collective/Distributive NPs
- Scope (Negation, Quantifiers)
- Presuppositions
- Tense and Aspect
- Illocution Force
- Textual Entailment
- Discourse Structure/ Rhetorical Relations
- Neo-Davidsonian Events
- Background Knowledge
- Information Structure ...
- + any combination of the above

However, for the purposes of the QTLeap project, the following entities from the above list will be treated with preference: Lexical Semantics (including WSD), Named Entities, Co-reference, and Background Knowledge. Also, sentence semantics is added to the list. Focusing on these phenomena will advance the state-of-the-art in MT for two reasons: First, there are either suitable linguistic resources available for them — or such resources can be created/detected in manageable ways — and, second, the linguistic knowledge (lexicons, corpora, etc.) often needs common knowledge (LOD datasets) and discourse (co-reference).

Many of these phenomena in isolation cannot constitute a sufficient basis for the semantic transfer step in machine translation. Thus, we impose an additional requirement for the corresponding analysis: it has to reflect the semantic content of all lexical units in the sentence.

In this respect, for example, Named Entities and Co-reference on their own do not suffice for semantic transfer in MT, but they might be part of the semantic processing chain for such a transfer. The notion of *deep semantic processing* is not well-defined and thus leaves room for interpretation. There are at least two dimensions of defining deep and shallow semantic processing: (1) how well the semantic content of a sentence is represented; and (2) what kind of additional sources are necessary for completion of the task. For example, Named Entities are considered shallow analysis in the sense of the first dimension, because many elements of the semantic content of the sentence will not be covered. On the other hand, Named Entities could be considered deep semantic processing by using complex language resources and preprocessing like dependency parsers, Linked Open Data based gazetteers.

Thus, in order to fulfil the demands of MT, deep semantic processing must include a considerable amount of *semantic language resources* such as syntax/semantic treebanks (DeepBank, Prague Dependency Treebank, PropBank, Groningen Meaning Bank, etc.), semantic lexicons (WordNet, Ontology-based Lexicons, Valency Lexicons, etc.), and background knowledge (ontologies, linked open data, etc.), *which complement the semantic content of the text in considerable depth and scope*. This definition allows for many approaches to semantic processing to be considered as deep semantic processing. Here we present mainly approaches to deep semantic processing related to the languages involved in the project. Some other approaches are also given for comparison purposes.

Deep semantic processing might be and in most cases is still language dependent to a great extent. For instance, the predicates involved in the analyses could be based on the lemmas of the word forms in the sentences. The addition of background knowledge provides some language independent elements in the semantic content of the text and we hope that this ensures a better semantic transfer. In terms of the Vauquois triangle, deep semantic processing would facilitate mainly transfer in the upper part of the triangle:

The formalisms used to represent the semantic analyses range from graph representations (e.g. semantic dependency trees, semantic graphs) to lambda calculus over some logical representation (e.g., first order logic). Within the conference STEP 2008 a shared task on comparing semantic representations as output by practical wide-coverage NLP systems was organized [Bos and Delmonte, 2008]. There were seven participant groups in this shared task. The used formalisms included: Logical Form (2 systems), Situation Semantics (1 system), Discourse Representation Theory (1 system), Minimal Recursion Semantics (1 system), Ontological Semantics (1 system), and Semantic Triples (1 system). The number of different formalisms and the analyses described in the papers and devoted to each system corroborate the observation of Bos and Delmonte [2008]: *because there*

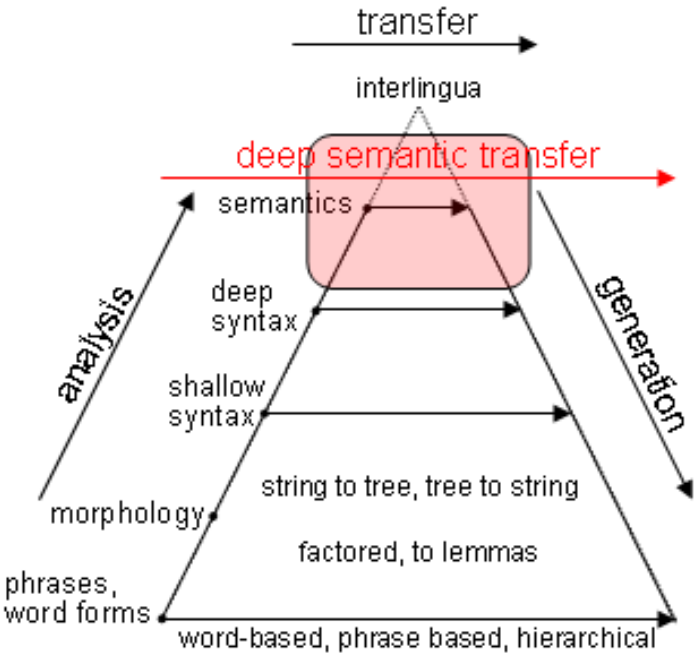


Figure 1: The Vauquois triangle. Deep semantic transfer could be defined by combining several phenomena from the above list including some syntactic features.

are several "competing" semantic formalisms, and the depth of analysis is arbitrary, it is hard to define a universal theory-neutral gold standard for semantic representations.

The two CoNLL shared tasks: Joint Parsing of Syntactic and Semantic Dependencies [Surdeanu et al., 2008] and Syntactic and Semantic Dependencies in Multiple Languages [Hajič et al., 2009] defined semantic processing in terms of predicate identification and semantic dependencies (identification of predicate arguments and their roles). The number of participating systems was impressive (19 in 2008 and 20 in 2009 on closed task restricted to data provided by the organizers). All this demonstrates the diversity of deep semantics frameworks. Also, the approaches to deep semantic processing are highly varied. Many of them consider semantic processing as compositional over the syntactic structure of the sentences, and implement rule-based systems that construct the semantic interpretation for word forms, then applying the rules to construct whole-sentence semantic representation. Other systems construct different elements of the semantic analysis in separate steps and then combine them in one representation. Such systems usually implement the following tasks: semantic role labeling, word sense disambiguation, coreference and anaphora resolution, treatment of quantification, etc. There are also systems that construct the semantic representation directly from semantically annotated corpora on the basis of machine learning algorithms.

In this deliverable we present some of best-practice approaches to deep semantic processing related to the aims of the project. We specially pay attention to approaches developed and/or used by the partners in their previous work on semantic processing. We then outline our steps beyond state-of-the-art and discuss the expected benefits of the project for the real usage scenario.

2 Deep Dependency Representation and Processing

Here we present deep dependency representations of the sentence meaning in terms of predicates and semantic roles. The roles establish connection of the predicate with their arguments in the sentences. Treebanks annotated on the level of deep dependency level exist for Czech, English and Basque as project languages. This kind of analysis is already exploited in systems for machine translation.

2.1 Prague Dependency Treebank: Tectogrammatical Layer

The Prague Dependency Treebank (PDT)[1] is a Czech treebank, annotated in accordance to the linguistic theory of Functional Generative Description [P. Sgall and Panevova, 1986]. The tectogrammatical layer[2] is the third layer of the PDT. The treebank has three layers: morphological, analytical, and tectogrammatical. The morphological layer operates over tokens, assigning to them POS and lemma tags. The analytical layer reflects the surface sentence structure. The tectogrammatical layer represents the syntactic-semantic interface, adding the functional dimension and collapsing the structural information, thus aiming at a more language-independent level of abstraction.

2.1.1 Tectogrammatical Layer Representation

The tectogrammatical annotation builds on the analytical level. It presents the deep semantic structure of the sentence. At the tectogrammatical level, each sentence has at least one representation unambiguously characterizing the meaning of the sentence. The tectogrammatical level representation contains all the information necessary for translating the tectogrammatical representation into the lower levels, as well as for its interpretation in the sense of intentional semantics.

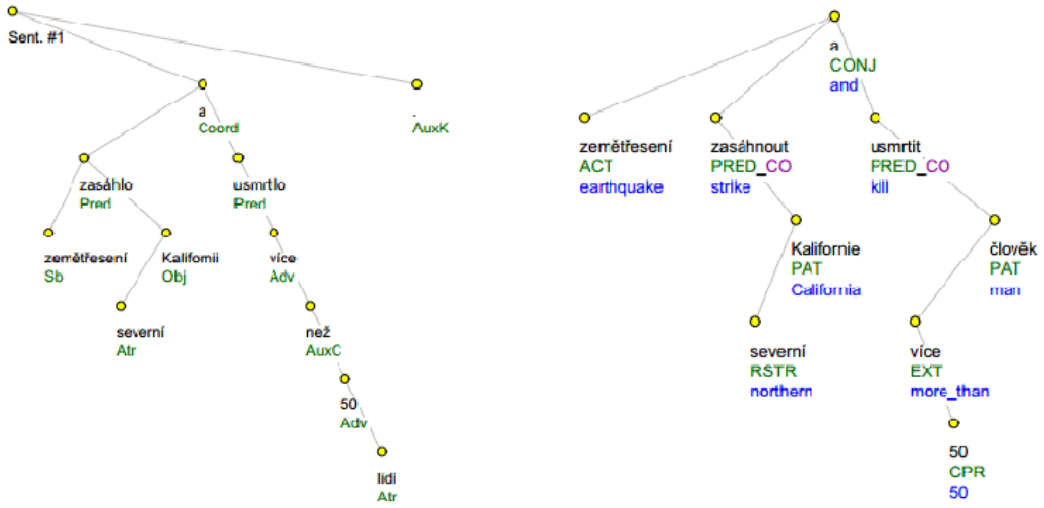


Figure 2: Analytical and Tectogrammatical representation of the sentence *The earthquake hit Northern California and killed more than 50 people*

[1] https://ufal.mff.cuni.cz/pdt2.0/

[2] https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/ch01.html

In contrast to the analytical level, which follows the surface sentence structure and encodes analytical functions (in particular, grammatical relations like *Subject, Object, Predicate, Attribute*, etc.), while preserving the word order, the tectogrammatical level highlights the functional dimension (such as the semantic roles *Actor, Patient, Addressee*, etc.). What is more, it abstracts away from elements from the synsemantic (functional) parts-of-speech (prepositions, conjunctions, etc.) in the dependency trees, thus focussing on the autosemantic (content) words (nouns, verbs, adjectives, etc.). The structural information is not lost, but just "collapsed" into the content words representations. In this way, a more abstract level of language representation is achieved, which then is used for the transfer step within the MT systems.

The tectogrammatical level also encodes phenomena not yet encoded at the analytical level, viz., ellipsis, coreference, and information structure of the sentence (topic-focus). This level uses information from valence lexicons.

The two representations of the same translated sentence are shown in Fig. 2 on page 11. The sentence is: *The earthquake hit Northern California and killed more than 50 people.*

As can be seen, the analytical representation (on the left) displays all the tokens with their grammatical labels, while the tectogrammatical representation (on the right) operates over autosemantic lemmas only assigning to them semantic roles. Also, the function words are compressed.

2.1.2 Tectogrammatical Layer Processing

For tectogrammatical-oriented parsing, dependency parsers are used [McDonald, 2006] and [Nivre et al., 2006]. The result on the tectogrammatical level heavily depends on the results from the processed analytical level.

Recent developments on deep semantic parsing are related to the following system and its services:

Treex [Popel and Žabokrtský, 2010] is a highly modular system for NLP solutions (https://ufal.mff.cuni.cz/treex/). Two of its applications that have been extensively used in recent years with respect to deep processing are as follows:

1. *TectoMT.* (https://ufal.mff.cuni.cz/tectomt/)
2. *Depfix.* (http://ufal.mff.cuni.cz/depfix/)

TectoMT provides an English-Czech translation with deep transfer, based on the tectogrammatical layer. Depfix is a system for rule-based correction of English-Czech statistical MT outputs.

TectoMT is a hybrid statistical/rule-based translation system, which uses a pipeline approach; its pipeline consists of a deep dependency analysis, transfer at the deep layer, and generation of the surface form.

The tectogrammatical layer serves as the transfer step for EN-to-CS translation[3]:

The advantages of using the tectogrammatical layer in the transfer are as follows:

- There are fewer differences between the languages
- The nodes encode only autosemantic words, thus leaving target-language specific issues to the generation step between the analytical and tectogrammatical layers

[3]The picture is from Bojar et al. [2008].

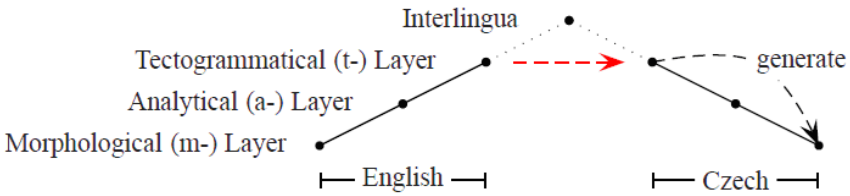


Figure 3: Tectogrammatical level transfer for En-to-CS translation.

- This layer operates on lemmas. Consequently, all the morphological complexity is handled between morphological and analytical layers.

Depfix is a system for automatic rule-based post-processing of English-to-Czech MT outputs designed to increase their fluency. The ones, handled at the tectogrammatical level, are as follows: pro-drops; verb tenses; negation; valence (like argument case correction).

Additionally, consider the strengths of both systems: TectoMT and Depfix have been combined with a factored Moses setting into a new architecture called Chimera[4]. In Bojar et al. [2013] the reported results from human evaluation show that Chimera in its two variant systems (58 % and 57.8 %) outperforms Google translation (56.2 %) in the translation direction from English to Czech.

Within the project this approach will be exploited for several language pairs. For the details on Pilot One, readers are referred to deliverable D2.4.

2.2 PropBank: Predicate-Argument Layer

The PropBank project[5] created a text corpus annotated with information about basic semantic propositions. PropBank results in adding a layer of predicate-argument relations (semantic roles) to the syntactic trees of the Penn Treebank. The syntactic structure of the trees in Penn Treebank is used in the PropBank annotation to assign semantic labels to nodes in the trees. The Penn Treebank does not distinguish the different semantic roles played by a verb's grammatical functions. Since the same verb used with the same syntactic subcategorisation, but with different semantic role sets[6] can assign different semantic roles, roles cannot be deterministically added to the Treebank by an automatic process. Thus every instance of every verb in the treebank is covered.

2.2.1 Propbank Representation

The goal behind Propbank project was the creation of a broad-coverage hand-annotated corpus of semantic roles together with the verb alternations. The annotation relies on the linking between semantic roles and syntactic realization. The syntactic frames are a direct reflection of the underlying semantics [Levin, 1993]. The Propbank approach consists of two steps:

1. Framing: collection of framesets for each lexeme. They:

   (a) examine a sample of corpus sentences for the verb under consideration

[4] http://ufal.mff.cuni.cz/chimera/
[5] http://verbs.colorado.edu/ mpalmer/projects/ace.html
[6] In case of different verb senses.

(b) group the instances into one or more major senses

(c) turn each major sense into a single frameset

2. Annotation. It includes the following steps:

(a) a rule-based argument tagger is run on the corpus; 83% accuracy on pilot data is achieved

(b) the tagger output is corrected by hand, on a verb by verb basis

(c) differences between annotators are resolved via adjudication

An individual verb's semantic arguments are numbered, beginning with 0. For a particular verb, *Arg0* is generally the argument exhibiting features of a prototypical agent while *Arg1* is a prototypical patient or theme. The Roleset is set of roles for a distinct usage of a verb. The Frameset is a role set associated with a set of syntactic frames indicating allowable syntactic variations. The numbered arguments plot a middle course among many different theoretical viewpoints and can be mapped onto any theory of argument structure.

Propbank annotation relies on the traditional thematic roles, such as:

- Agent: animate, volitional; initiates action

  Anna prepared chicken for dinner.

- Patient: animate or inanimate; undergoes (and is affected by) action

  Anna baked a cake for her daughter.

- Experiencer: animate; undergoes perceptual experience

  The storm frightened Anna.

- Theme: animate or inanimate; undergoes motion, or an action that does not affect it significantly

  Anna sent Tim a letter.

- Recipient: generally animate; receives something

  Tim kicked Bob the ball.

Propbank's *ArgM* Modifier Roles include: LOC: location; EXT: extent; DIS: discourse connectives; ADV: general-purpose; NEG: negation marker; MOD: modal verb; CAU: cause; TMP: time; PCN: purpose; MNR: manner; DIR: direction.

EPEC-RolSem [Aldezabal et al., 2009] is a Basque corpus labeled at predicate level following the PropBank-VerbNet model, but being adapted to the specificities of the language. The predicates are aligned to their English counterparts. The EPEC corpus contains 1,457 different verbs, only 270 occurring 30 or more times. The BVI (Basque Verb Index) lexicon [Aldezabal et al., 2013] is the first repository of syntactic-semantic information of Basque verbs, derived from the process of manual annotation at the predicate level for the verbs in the EPEC corpus. So far, it contains syntactic-semantic models of 244 Basque verbs and 364 different senses. This set of verbs covers 71% of the EPEC Corpus, and there is an ongoing process of enlarging the number of verbs by means of a semi-automatic process.

2.2.2 Propbank-enhanced Processing

Initially, there was built a rule-based argument tagger — see [Palmer et al., 2001], with 83 % accuracy. The results were then improved by post-tagging manual corrections. The Semantic Role Labelling (SRL) became an important NLP task in semantic tagging area. For example, the CoNLL shared tasks in 2004 and 2005 were devoted to semantic role labelling — [Carreras and Màrquez, 2004], [Carreras and Màrquez, 2005]. The tasks were defined for English with data from PropBank. Most of the systems exploited a full syntactic parse in order to define argument boundaries and to extract relevant information for training classifiers that disambiguate between role labels — [Carreras and Màrquez, 2004]. With respect to the learning component of the systems, many different approaches were used: pure probabilistic models, Maximum Entropy, Decision Trees, etc. The labelling task was divided mainly in two steps: (i) identification of the arguments (recognition or filtering) and (ii) the labelling itself. Granularity of the processed elements also plays role in the systems. The most successful systems worked at phrasal level. The systems used all the available linguistic information as features - word forms, POS tags, chunk labels, named entities, etc. The second shared task event (2005) also was devoted to Semantic Role Labelling. It aimed at evaluating the contribution of full parsing in SRL. The complete syntactic trees produced by two parsers have been provided as input information for the task. The annotations included: POS tags, chunks, clauses, full syntactic trees and named entities. Nineteen systems participated in the CoNLL 2005 shared task. They used eight different learning algorithms: Maximum Entropy, Support Vector Machines, Sparse Network of Winnows (SNoW), Decision Trees, Memory-Based Learning, Relevant Vector Machine, Tree Conditional Random Fields, Consensus in Pattern Matching. The best system performed with F1 measure: 79.44 %. As it was mentioned in the Introduction section, the two CoNLL shared tasks: *Joint Parsing of Syntactic and Semantic Dependencies* — [Surdeanu et al., 2008] and *Syntactic and Semantic Dependencies in Multiple Languages* — [Hajič et al., 2009], defined the semantic processing in terms of syntactic and semantic dependency parsing and semantic role labeling in English (2008) and in several other languages (2009). The shared task in 2009 included also pure semantic role labelling subtask over dependency trees. The best system for 2008 task performed with 90.13% in-domain test and 82.81% out-of-domain for Labelled Attachment Score (syntactic dependencies) and 81.75% in-domain and 69.06% out-of-domain for Labelled F1 (semantic dependencies). This system used second-order parsing model (using as parsing features siblings, grandparents, etc.), argument identification/classification models separately tuned for different parts of the corpus, reranking inference for the SRL task and joint optimization of the complete task using metalearning. For the dependency parsing both most popular frameworks were used: transition-based and graph-based ones. For different tasks joint approaches were implemented.

A recent tutorial at The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies [Palmer et al., 2013], on Semantic Role Labelling[7] provides a comprehensive overview on the linguistic annotation background and methods for solving the problem.

[7] http://naacl2013.naacl.org/TutorialsAccepted.aspx

2.3 Comparison with the PDT representation

Some comparison and compatibility of semantic labels across frameworks, languages, and resources is provided in Rambow et al. [2003]. The paper considers extracting tectogrammatical labels from other resources, such as Propbank. Local semantics refers to the specific verb meanings, while global one goes beyond verbs and meanings. Thus, the local semantics of verbs in Propbank contrasts with the global semantics of verbs in PDT, because the global semantics reflects a specific grammar framework. For example, the relations in PDT are richer than these in Propbank, which causes difficulties in getting optimal compatibility of data annotation models. The paper proposes a strategy in which resources are annotated with local semantics first, and then a translation lexicon of the type Propbank-to-X language resource is created.

Both treebanks handle propositions via semantic role labelling. However, PDT has implemented its theory from the start, while PropBank is a build-on over Penn Treebank. PropBank relies on bigger generalizations, using notations, such as Arg0, Arg1, Arg2, etc., while PDT introduces a rich set of roles. Despite different levels of underspecification, both resources make use of rich valence and beyond-valence lexicons (PropBank - from VerbNet and FrameNet[8]; PDT - from their in-house constructed lexicon Vallex[9]). It should be noted that nowadays these two types of treebanks have become best practices for creating similar resources for other languages.

3 Underspecified Semantics: Minimal Recursion Semantics

In this section we present Minimal Recursion Semantics (MRS) as an example of underspecified semantic formalisms. MRS underspecifies scope ambiguities for quantifiers and other scope-bearing elements. The selection of MRS is motivated by several facts: (1) it has already been implemented as part of HPSG grammars for several project languages: English [Copestake and Flickinger, 2000], German [Crysmann, 2007], Spanish [Marimon et al., 2007], Portuguese [Branco and Costa, 2008], [Costa and Branco, 2010] and Bulgarian [Osenova, 2010]; (2) it is already used as a basis for semantic transfer in MT systems for several language pairs — [Bond et al., 2005] and [Oepen et al., 2004]; (3) it allows the construction of semantic representation over shallow analyses or dependency syntactic structures — Copestake [2004/2006], [Copestake, 2007], and [Simov and Osenova, 2011]; (4) there exist corpora annotated with MRS structures, including some parallel ones — [Flickinger et al., 2012b] and [Flickinger et al., 2012a].

3.1 Minimal Recursion Semantics Representation

3.1.1 MRS Definition

MRS is introduced as an underspecified semantic formalism [Copestake et al., 2005]. It is used to support semantic analyses in the HPSG English Resource Grammar — [Copestake and Flickinger, 2000], but also in other grammar formalisms like LFG. The main idea is the formalism to rule out spurious analyses resulting from the representation of logical operators and the scope of quantifiers. Spurious analyses of logical form over an

[8] http://verbs.colorado.edu/semlink/
[9] http://ufal.mff.cuni.cz/vallex/2.5/doc/home.html

utterance could be result from different NLP analyses which produce equivalent logical but syntactically different expressions, like the following two formulas: $\lambda x[\text{fierce}(x) \wedge (\text{black}(x) \wedge \text{cat}(x))]$ and $\lambda x[\text{cat}(x) \wedge (\text{black}(x) \wedge \text{fierce}(x))]$. In MRS such spurious analyses are excluded by the flat representation of the body in the formulas. The determination of the scope of quantifiers in a sentence very often requires information which is not available during the sentence processing. Thus, MRS provides a compact representation which allows further specialization of the quantifiers scope when the necessary information becomes available.

Here we will present only basic definitions from Copestake et al. [2005]. For more details the cited publication should be consulted. An MRS structure is a tuple $< GT, R, C >$, where GT is the top handle, R is a bag of EPs (elementary predicates) and C is a bag of handle constraints, such that there is no handle h that outscopes GT. Each elementary predication contains exactly four components: (1) a handle which is the label of the EP; (2) a relation; (3) a list of zero or more ordinary variable arguments of the relation; and (4) a list of zero or more handles corresponding to scopal arguments of the relation (i.e., holes). Each scopal argument could be assigned to a handle. In this way the elementary predicates with this handle become arguments of other elementary predicates. For example, a quantifier requires such kind of arguments as a body or restriction of the quantifier. Thus handles are used to represent different readings of the same set of elementary predicates via different assignments from handles to the corresponding arguments — see the example below, taken from the cited paper. The handle constraints have the following form: $h_i = h_j$ which states that h_i outscopes h_j.

Here is an example of a complex MRS structure for the sentence "Every dog chases some white cat."

$$< h_0, \{h_1 : every(x, h_2, h_3), h_2 : dog(x), h_4 : chase(x, y), h_5 : some(y, h_6, h_7),$$

$$h_6 : white(y), h_6 : cat(y)\}, \{\} >$$

The top handle is h_0. The two quantifiers are represented as relations $every(x, y, z)$ and $some(x, y, z)$ where x is the bound variable, y and z are handles determining the restriction and the body of the quantifier. The conjunction of two or more relations is represented by sharing the same handle (h_6 above). The outscope relation is defined as a transitive closure of the immediate outscope relation between two elementary predications — EP immediately outscopes EP' iff one of the scopal arguments of EP is the label of EP'. In this example the set of handle constraints is empty, which means that the representation is underspecified with respect to the scope of both quantifiers.

The representation of the MRS structure via bags of elementary predicates and handle constraints allows an easy definition of compositional rules via union of bags and variable substitution. This feature of the MRS provides an easy mechanism for implementation of MRS processors in combination with a syntactic parsing.

3.1.2 RMRS Definition

Robust Minimal Recursion Semantics (RMRS) is introduced as a modification of MRS which captures the semantics resulting from a shallow analysis — see [Copestake, 2004/2006] and [Copestake, 2007]. The main motivations for this development are the facts that

currently no single system can do everything: both deep and shallow processing have inherent strengths and weaknesses; on the other hand, the domain-dependent and domain-independent processing must be linked. The ideal level on which this linking can take place is semantics. Therefore, Copestake [2004/2006] and Copestake [2007] propose a semantic representation which allows to build a comparable semantic representation for both deep and shallow processing. The justification behind RMRS is to add more underspecification to MRS. This is done by e.g. the separation of arguments from the predicates. Thus each predicate is represented via its name (constructed on the basis of the lemma of the word form in the text) and its main argument which depends on the part of speech - *referential index* for nouns and some pronouns or *event index* in other cases. In this way it is possible that the predicates and their arguments are added to the structure separately from each other.

Here we present a formal definition of RMRS as defined in Jakob et al. [2010]. An RMRS structure is a quadruple

$$< hook, EPbag, argumentset, handleconstraints >$$

where a hook consists of three elements $l : a : i$, l is a label, a is an anchor and i is an index. Each elementary predication is additionally marked with an anchor — $l : a : r(i)$, where l is a label, a is an anchor and $r(i)$ is a relation with one argument of appropriate kind — referential index or event index. The argument set contains argument statements of the following kind $a : ARG(x)$, where a is anchor which determines for which relation the argument is defined, ARG is the name of the argument, and x is an index or a hole variable or handle (h) for scopal predicates. The handle constraints are of the form $h =_q l$, where h is a handle, l is a label and $=_q$ is the relation expressing the constraint similarly to MRS. $=_q$ sometimes is written as qeq.

Both representations MRS and RMRS could be transferred to each other under certain conditions. For MRS-to-RMRS it will be necessary to have access to the word forms in the text from which the corresponding predicates were inferred. For RMRS-to-MRS it will be necessary to unify the number of arguments of predicates via some kind of a lexicon.

The separation of the predicates from their arguments facilitates the construction of RMRS structures over shallow analyses. Shallow processors usually do not have access to a lexicon. Thus they cannot predict the amount of the arguments that have the corresponding predicate. The forming of the relation names follows such conventions that provide possibilities to construct a correct semantic representation only on the base of information provided by a POS tagger, for example. The arguments are introduced separately by argument relations between the label of a relation and the argument. The names of the argument relations follow some standardized convention like RSTR, BODY, ARG1, ARG2, etc. These argument relations are grouped in a separate set in a given RMRS structure.

RMRS was used in analyses of two dependency treebanks — the TIGER treebank of German and the Prague Dependency Treebank of Czech. The work on Prague Dependency Treebank presented in [Jakob et al., 2010] first assigns elementary predications to each node in the tectogrammatical tree. Then the elementary predications for the nodes are combined on the basis of the dependency annotation in the trees. A similar approach could be taken also in cases when the input trees are surface syntactic trees instead of tectogrammatical trees. The difference is that the surface trees contain nodes for each token in the sentence.

3.2 Deepbanks

DeepBank is an English treebank with annotation on the original Penn Treebank, made with the English Resource Grammar (ERG). The grammar is augmented with a robust approximating PCFG for complete coverage. DeepBank contains rich linguistic annotation on both syntactic and semantic structures of the sentences and is available in a variety of representation formats (http://moin.delph-in.net/DeepBank). Following earlier practice in the development of Redwoods treebanks, manual annotations are done using the discriminant-based treebanking environment provided by `[incr tsdb()]` [Oepen, 1999]. It identifies the correct full analysis among the candidate analyses proposed by the English Resource Grammar. The process of DeepBank annotation of the Wall Street Journal corpus is organized into iterations of the cycle:parsing, treebanking, error analysis and grammar/treebank updates, with the goal of maximizing the accuracy of annotation through successive refinement [Flickinger et al., 2012b].

Similar to DeepBank several other treebanks are in process of construction and within the project QTLeap they will be extended. They are based on translation of the texts annotated in DeepBank. Then the annotation follows the same methodology as in the DeepBank. The current version of parallel deep banks (ParDeepBank) exists for Portuguese and Bulgarian. They are aligned to the English DeepBank on word level from which the alignment is inferred for the next levels - see [Flickinger et al., 2012a]. We plan to construct an MRS annotated treebank for Basque parallel to the Spanish one (and hence also linked to English).

3.3 MRS Processing

The construction of MRS structures for sentences is implemented as part of HPSG deep grammars for several languages. The most developed grammar is the English Resource Grammar (ERG). ERG has been under continuous development at CSLI since 1993. It provides syntactic and semantic analyses for the large majority of common constructions in written English text. The current grammar consists of more than a 35,000-word lexicon instantiating 980 leaf lexical types (without subtypes), as well as 70 derivational and inflection rules, and 220 syntactic rules. Similar grammars exist for Portuguese, Spanish, German, and Bulgarian. These grammars will be extensively used within the project, since they provide high precision of the analyses. However, the main problem with them is their narrow coverage over different types of texts.

In order to solve this problem Copestake [2004/2006] and Copestake [2007] suggested to construct RMRS analyses generated over partial and shallow analyses. The idea is to extract as much as possible semantic information from a partial or shallow processed text. In the worst case — from POS tagged text. Copestake [2007] demonstrates how RMRS structures can be constructed over the output of a robust statistical parser RASP, which do not have access to subcategorisation information [Briscoe and Carroll, 2002].

Similarly, Simov and Osenova [2011] implement a set of rules for transfer of dependency parses into RMRS presentations. The input for the RMRS structures is based on the following linguistic annotation — the lemma (*Lemma*) for the given wordform; the morphosyntactic tag (*MSTag*) of the wordform, and the dependent relations (*Rel*) in the dependency tree. In cases of quantifiers we have access to the lexicon used in the Bulgarian HPSG grammar. The algorithm for producing of RMRS from a dependency parse is implemented via two types of rules:

$$< Lemma, MSTag > \rightarrow EP - RMRS$$

The rules of this type produce an RMRS structure representing an elementary predicate.

$$< DRMRS, Rel, HRMRS > HRMRS'$$

The rules of this type unite the RMRS constructed for a dependent node ($DRMRS$) into the current RMRS for a head node ($HRMRS$). The union ($HRMRS'$) is determined by the dependency relation (Rel) between the two nodes.

First, we start with assigning EPs for each lemma in the dependency tree. These EPs are similar to node EPs of [Jakob et al., 2010]. Each EP for a given lemma consists of a predicate generated on the basis of the lemma string. Additionally, the morphosyntactic features of the wordform are presented. On the basis of the part-of-speech tag the type of ARG0 is determined — referential index or event index. After this initial step the basic RMRS structure for each lemma in the sentence is compiled. Then these structures are incorporated in each other in bottom-up manner. Here are examples of two RMRS structures constructed in this way. They are in Bulgarian: (1) an RMRS for the verb *'cheta'* (to read) [10]:

$< l1 : a1 : e1, \{l1 : a1$:cheta_v_$rel(e1)\}, \{a1 : ARG1(x1)\}, \{\} >$

In this example we also include information for the unexpressed subject (ARG1) which is always incorporated in the verb form. The RMRS structure for a sentence with an explicit subject and an explicit direct object follows. The sentence is *momche mu chete kniga* [Boy him-dative reads book], 'A boy reads a book to him'):

$< l2 : a3 : e1,$
$\{l1 : a1$:momche_n_$rel(x1), l2 : a3$:chete_v_$rel(e1), l3 : a4$:kniga_n_$rel(x2)\},$
$\{a3 : ARG1(x1), a3 : ARG2(x2), a3 : ARG3(x3)\},$
$\{\} >$

In this case the information coming from clitics is represented only in the argument set.

The construction of RMRS for sentences, based on shallow processing, suffers from the pipeline processing effect — error accumulation. Errors in earlier processing stages cause suboptimal performance during the next steps. In order to escape from this effect there are attempts to derive MRS structures from text directly. In order to achieve this goal, first the gold standard treebanks (DeepBank, etc.) are converted into MRS treebanks. The main approach to this task is the construction of semantic dependency structures based on MRS structures, which are linked to the word forms in the sentences. Having such a semantic dependency treebank, one might exploit the existing state-of-the-art dependency statistical parsers to perform MRS-based analyses. Ivanova et al. [2012] and Ivanova et al. [2013] provided an algorithm for conversion of HPSG analyses from the LinGO Redwoods treebank and DeepBank into the Elementary Dependency Structure of [Oepen and Lønning, 2006]. Ivanova et al. [2013] also provide the results from training several available parsers over the transformed treebanks.

As currently there is no HPSG grammar for Basque, we will use the approach of deriving RMRS structures from the output of a Basque dependency analyzer.

[10] Please note that the examples from Bulgarian are presented in their transliterated equivalents.

3.4 MRS for Deep Semantic Transfer in MT

Minimal Recursion Semantics was originally developed for the purposes of Machine Translation. The main idea was that an underspecified semantic representation is appropriate for machine translation because it provides an abstract level to semantic transfer, but at the same time it postpones the difficult decisions for later stages of the processing. These difficulties are assumed to be less important in the area of machine translation. MRS was applied in the past in two ways to support machine translation: (1) rule-based semantic transfer, and (2) factor-based statistical machine translation.

The rules-based semantic transfer uses transfer rules working on the MRS representation of the source language MRS structures and constructing the target language MRS structure. Thus, the transfer rules in this framework are rewriting rules over MRS (Minimal Recursion Semantics) structures. The basic format of the transfer rules is:

$[\mathcal{C} :]\mathcal{I}[!\mathcal{F}] \rightarrow \mathcal{O}$

where $\mathcal{I}$ is the *input* of the rule, $\mathcal{O}$ is the *output*. $\mathcal{C}$ determines the *context* and $\mathcal{F}$ is the *filter* of the rule. $\mathcal{C}$ selects the positive and $\mathcal{F}$ the negative context for the application of a rule. For more details on the transfer rules, see [Oepen, 2008]. This type of rules allows an extremely flexible transfer of factual and linguistic knowledge between the source and the target languages. The rules have access to the MRS structure for the source language, but can also access the (partially) constructed target language MRS structure. Thus elements in each rule could include parts from both MRS structures. The rules could be encoded manually or extracted from existing sources. For the latter we have constructed parallel deep treebanks in the project. Each treebank has to contain parallel sentences, their syntactic and semantic analyses and correspondences on the level of MRS. The aim of constructing parallel deep treebanks within the project is to use them as a source for learning of statistical transfer rules for deep semantic transfer in machine translation along the lines of [Bond et al., 2011].

The factor-based translation model is built on top of the factored SMT model proposed by Koehn and Hoang [2007], as an extension of the traditional phrase-based SMT framework. Instead of using only the word form of the text, it allows the system to take a vector of factors to represent each token, both for the source and target languages. The vector of factors can be used for different levels of linguistic annotations, like lemma, part-of-speech, or other linguistic features, if they can be (somehow) represented as annotations to each token. This approach was explored for Bulgarian-to-English Statistical Machine Translation by Wang et al. [2012a] and Wang et al. [2012b].

In this setup the data was processed by the NLP pipe for Bulgarian [Savkov et al., 2012] including a morphological tagger, a lemmatizer and a dependency parser. On the top of the analyses MRS structure were (partially) constructed. From the whole result the following factors were considered on the source language side (Bulgarian):

- WF – word form is just the original text token.
- Lemma is the lexical invariant of the original word form. We use the lemmatizer, which operates on the output from the POS tagging. Thus, the 3rd person, plural, imperfect tense verb form 'varvyaha' ('walking-were', They were walking) is lemmatized as the 1st person, present tense verb 'varvya'.
- POS – part-of-speech of the word. We use the positional POS tag set of the BulTreeBank, where the first letter of the tag indicates the POS itself, while the next letters refer to semantic and/or morphosyntactic features, such as: Dm - where 'D' stands for 'adverb', and 'm' stand for 'modal'; Ncmsi - where 'N' stand for 'noun', 'c' means 'common', 'm' is 'masculine', 's' is 'singular',and 'i' is 'indefinite'.

- LING – other linguistic features derived from the POS tag in the BulTreeBank tagset.
- DEPREL is the dependency relation between the current word and the parent node.
- HLEMMA is the lemma of the current word's parent node.
- HPOS is the POS tag of the current word's parent node.

For the Pilot one of QTLeap project we exploit this model adding factors for the target language and also training a model for the direction English to Bulgarian. The new results are reported within deliverable D2.4.

3.5 Related Underspecified Semantic Formalisms

One of the MRS-related semantic formalisms is the Abstract Meaning Representation (AMR[11]), which also aims at achieving whole-sentence deep semantics instead of addressing various isolated holders of semantic information (such as, NER, coreferences, temporal anchors, etc.). AMR also builds on the available syntactic trees, thus contributing to the efforts on sembanking. It is English-dependent and it makes an extensive use of PropBank framesets [Kingsbury and Palmer, 2002] and [Palmer et al., 2005]. Its concepts are either English words or special keywords. AMR uses approximately 100 relations. They include: frame arguments, general semantic relations, relations for quantities and date-entities, etc. Below a typical AMR representation is given:

```
(d / describe-01
    :arg0 (m / man)
    :arg1 (m2 / mission}
    :arg2 (d / disaster))

The man described the mission as a disaster.
The man's description of the mission:
        disaster.
The man's description of the mission:
        disaster.
As the man described it, the mission was a
        disaster
```

It can be seen that several syntactic realizations have the same AMR representation.

4 Groningen Meaning Bank (GMB)

This is a large corpus of public domain texts in English with deep (logical) semantics. The annotation is automatically produced, but manually corrected. This resource integrates various phenomena in one formalism. It also has a linguistically motivated, theoretically solid (CCG[12]/DRT[13]) background.

The semantics layer of the treebank includes various types of semantics-related phenomena, such as: scope phenomena (negation, quantification), word senses (Word Net),

[11] http://www.isi.edu/natural-language/amr/a.pdf
[12] Combinatory Categorical Grammar
[13] Discourse Representation Theory

anaphoric pronouns, thematic roles (Verb Net), presuppositions, tense, aspect, events, rhetorical relations, compositional semantics layer, segmented discourse representation structure. Below a chart is given, which lists the combined parts of the semantic annotation — [Basile et al., 2012]:

Level	Source	DRS encoding example
POS tag	Penn (Miltsakaki et al., 2004)	
named entity	ENE (Sekine et al., 2002)	named (X, 'John', 'Person')
word senses	WordNet (Fellbaum, 1998)	pred (X, loon, n, 2)
thematic roles	VerbNet (Kipper et al., 2008)	rel (E, X, 'Agent')
syntax	CCG (Steedman, 2001)	
semantics	DRT (Kamp and Reyle, 1993)	drs (Referents, Conditions)
rhetorical relations	SDRT (Asher, 1993)	rel (DRS1, DRS2, because)

Table 1: Groningen Meaning Bank Representation

Discourse Representation Theory — [Kamp, 1981], is a theoretical framework that incorporates discourse-sensitive linguistic phenomena such as tense and anaphora, within and across sentences. It uses Discourse Representation Structures (DRS) to represent a mental representation of the hearer as the discourse unfolds — [Geurts and Beaver, 2011].

The traditional representation for DRS is a box-like structure which contains two components:

- *discourse referents* (e.g. x, y, z) representing entities in the discourse, and
- discourse conditions (e.g. man(x), love(y, x)) representing the information about the discourse referents which is encoded in the discourse.

Current practices use the algorithm of Zettlemoyer and Collins [2007] and achieve 59% accuracy on semantic parses when the sentence input is of 7-word length[14]. Another approach to this kind of deep semantic processing uses the output of the CCG parser as input to the Boxer tool — [Curran et al., 2007]. This tool derives DRT representations in various formats, including first-order logic formulas.

5 Beyond the State of the Art

The aim of deep semantic processing within the QTLeap project is to construct transfer models for several language pairs in order to provide quality machine translation. In the project we will do experiments with several approaches like tree-to-tree transfer models adapted to support non-isomorphic tree transfer and to exploit document-level context features and Linked Open Data or transfer rules along the lines of Bond et al. [2005]. In order to achieve better translations, we have to develop robust parsing that includes the combination of data-driven models in shallower formalisms with the deep linguistic grammars to deliver full-fledged syntactic and semantic analyses with broader text coverage. The deep semantic processing will implement development beyond the state-of-the-art in several directions:

[14] http://jssp2013.fbk.eu/sites/jssp2013.fbk.eu/files/Johan.pdf

- **Better shallow processing.** Most of the existing deep semantic processing approaches exploit some kind of shallow processing. Thus, we plan to improve the NLP pipelines for syntactic dependency, which produce deep semantic analyses, beyond the state of the art. Usually such a pipeline includes modules like a tokenizer, a POS tagger, a Lemmatizer, a Dependency Parser, and a semantic compiler. Currently, such pipelines exist only for some of the languages in the project. Thus, our goal is to complete the pipelines for each of the languages. The error accumulation effect will be solved by two approaches: (1) improving the performance of the individual components by using the last achievement of machine learning and parsing technology; and (2) solving several tasks as one joint task. For instance, POS tagging, dependency parsing and anaphora resolution within sentences interact with each other and their simultaneous execution could perform better than a pipeline.

- **Direct deep semantic analyzer.** Here we will follow and expand the work of Ivanova et al. [2012] and Ivanova et al. [2013] on learning the derivation of deep semantic structures directly from text without the creation of syntactic analyses. If successful, this approach could completely substitute the pipeline approach. We will seek to further advance the grammar approximation approach as reported in Zhang and Krieger [2011], which utilizes the auto-parsed data by the deep grammars to construct parsing models that robustly analyze out-of-scope sentences with full coverage: this further guides the semantic composition process through unification. Such a direct approach of deep semantic analysis will help in the learning of transfer rules from parallel corpora.

- **Intelligent handling of MWE and OOV.** To further improve the robustness in semantic composition, we will proceed also in the following directions: incorporation of the achievements of tasks 4.1 OOV (out-of-vocabulary items) and 4.2 MWE (multi-word expressions) in parsing tools and grammars; usage of additional lexical information (valency) and linguistic knowledge (partial parses) with statistical parsers.

- **Conceptual Information and Linked Open Data.** We will incorporate ontological information from the LOD framework beyond Ontonotes project — [Pradhan et al., 2007], and the Ontology-to-Text-relation of Simov and Osenova [2008] towards integrating it in deep semantic analyses. This will be done via aligning lexicons for the languages to a common ontology selected within WP5 — Task 5.4 "Using semantic linking and resolving to improve MT". The alignment will use two relations between the senses of lexical units of the lexicons and the concepts in the ontology: *equality* and *subsumption*. Additionally, the elements of the frames for valency lexicons will be aligned to the corresponding participants for the corresponding event concept. We will not be able to provide a complete ontology annotation within the project. Thus, our goal is to add unambiguous concepts and relations from a more general part of the selected ontology. This work will be done by exploiting the results from WP5 on word sense disambiguation and named entity recognition. We consider the addition of this conceptual information to deep semantic analyses as a step to less language dependent analysis as much as the transfer of the conceptual information is practically copied as it is.

- **Text level coreference resolution.** The result of this task will provide interlinks between deep semantic analyses for different sentences in the text. Here the results

from WP5 will be exploited. Via the coreference chains the conceptual information between different sentences in the text will be distributed, leading to concept coverage of the texts.

To sum up, the extensions beyond the state-of-the-art of deep semantic processing aim at the extension of MRS with additional semantic information in order to cover more phenomena from the list given in the introduction, making them less language dependent, and the implementation of modules for quality analysis of several languages. All of these improvements will provide a basis for better MT systems.

6 Expected Benefits for the Real Usage Scenario

The deep semantic processing is envisaged to improve the MT on the IT-helpdesk scenario in separate language pairs in both directions. In such a QA scenario the questions and answers are quite short, but even in such cases deep semantics would be useful, since the failure to analyze and translate the data correctly would lead to total miscommunication. To quote an example, let us consider in more detail translation into and from English. Here comes some brief comparison: Basque, Bulgarian, Dutch, German, Portuguese, and Spanish are highly inflecting languages, while English is not. English has a relatively fixed word order, while other languages have relatively flexible word order.

Despite the fact that the main efforts in the project are directed into translation from English to the other languages, the reverse direction has been also included in this section in order to contrast the situations. It turns out that deep processing would be especially useful also in translating from English to other languages.

Below the translations from Pilot 0 are discussed. The focus is on the phenomena that are problematic and can be fixed with the help of the knowledge rich approach. The examples presented below are considered as a test case towards handling the project usage-based scenario. We present more extended sections on Bulgarian and German, as they adopt their own approaches. The languages that use TectoMT approach are presented together in a distinct section.

6.1 Problems that Relate to Both Translation Directions.

Observations on English-to-Bulgarian and Bulgarian-to-English In both directions the following typical problems have been identified: ungrammatical constructions and wrong lexical choices.

1. Ungrammatical constructions

   Under the notion of ungrammatical constructions various phenomena can be included: wrong POS, wrong postpositions, wrong word order, wrong constructions, etc.

   (a) Wrong case on the pronouns

   EN-BG English sentence: How do I activate it? Pilot 0 translation to Bulgarian: Kak az aktivira **tya**? [How I activate.3Pers she?] (expected: 1st Person of the verb and accusative case of the pronoun - her)

(b) Wrong person on the verb and wrong compound construction
BG-EN Bulgarian sentence: Kak moga da smenya **snimkata v profila** si v Hangouts? Pilot 0 translation: How can I change **the picture in profile** in hangouts ? (expected: In Hangouts, how can I change the **profile picture**?)

(c) Wrong POS (noun), which should be verb
EN-BG. English sentence: When I **scan** my computer, it becomes very slow. Pilot 0 translation: Kogato az **skanirane** moya kompyutyr, stava mnogo bavno. (expected verb, not noun)

(d) Wrong possessive construction and wrongly set constructions
BG-EN. Bulgarian sentence: Prozorecyt **mi za razgovori se otdeli** ot **spisyka** mi **s kontakti**. Pilot 0 translation: The window me for talks split from my list of contacts. (expected: My conversation window got separated from my contact list.)

All the reported problems in both directions have a better chance of being handled appropriately by the inclusion of several deep-oriented components. Elaborate valency lexicons would make reasonable predictions that would discard any partial phrasings, wrong POS, cases, or PP attachments. The addition of MRS or other predicate-layer annotation would further enhance the subject and object control, the correct illocutionary force of the sentence, etc.

2. Wrong lexical choice

The notion of 'wrong lexical choice' comprises several cases: wrong sense in the target language due to: the availability of more corresponding senses or inability to get the correct sense at all; incorrect handling of multiword expressions, etc.

(a) Wrong sense due to the availability of more senses in the target language
EN-BG. English sentence: How can I link my Facebook and YouTube **accounts**? Pilot 0 translation: Kak moga da **svyrzhem moyata** Facebook i YouTube **smetki**? (Human translation: Kak moga da svyrzha **akauntite si** vyv Facebook i YouTube?)
In the Bulgarian translation the word 'account' (being a profile in some Internet place) has been mixed with the meaning of the word 'bill' (being a sum to pay).

(b) Wrong sense due to the inability to get the correct sense in the target language
EN-BG. English sentence: How can I **turn on** parental control? Pilot 0 translation: Kak moga da **palya** roditelski kontrol? (expected verb 'vklyucha'(turn on), not 'palya' (burn))

(c) Incorrect handling of multiword expressions
BG-EN. Bulgarian sentence: Tazi komanda ste **vleze v sila**, kogato restartirate kompyutyra. Pilot 0 translation: This command will **enter into force** when you restart the computer. (expected: take effect)

Here the improvement might be done with the help of the following resources: adding a customized multiword expression component; integration of conceptual (NER, WSD, etc.) and LOD information for better discrimination of senses; incorporation of coreference resolution modules for constraining the propagation of the conceptual information.

For addressing the domain specific issues, domain lexicons are needed to ensure the correct senses in the data.

Specific problems for Bulgarian-to-English translation. In this direction the main problem with the translation is the non-translation of common Bulgarian words into English. These concern mainly two types: prefixed verbs and loan words. The former type is very productive in proliferating various senses from a basic verb sense, while the latter requires processing of foreign roots with Bulgarian affixes and inflection. The Bulgarian examples below (similarly to the Bulgarian examples above) are output of Moses system on aligned BG-EN parallel data. Such cases would require deeper and more focused linguistic treatment due to the high level of generated senses and the hybrid lexical nature of the loan words.

1. **BG-EN.** Bulgarian sentence: **Iztrih** fajl ot Google Disk po pogreshka. Pilot 0 translation: **Iztrih** file by google drive by mistake. (expected: I removed a file from Google Drive by mistake.)

2. **BG-EN.** Bulgarian sentence: Kak moga da go **razkompresiram**? Pilot 0 translation: How can I **razkompresiram** it? (expected: How do I extract it?)

These problems can be handled by: an intelligent OOV (out-of-vocabulary) component for analyzing the missing words; and a deep semantic analyzer for prediction of the missing content data.

Specific problems for English-to-Bulgarian translation. In this direction the main problems with the translation are as follows:

1. Wrong Agreement in NPs and VPs

2. Wrong POS, mainly due to the linguistic phenomenon of conversion in English (for example, 'scan' can be a noun or a verb)

These issues are typical for all settings in which the target language has rich morphology.

Their resolution can be advanced by using the deep semantic analyzer, which would incorporate all the morphosyntactic, valency and predicate information in one module. We also compared the wrong grammatical and lexical choice cases in both directions of translation in a selection of 100 sentences. It turned out that concerning the grammar failures the BG-EN direction has 20 wrong sentences, while the EN-BG direction has 65 wrong sentences. This observation supports the fact that the translation from a morphologically poor to a morphologically rich language requires a proper handling of the target language grammar. Most of the errors have to do with control verbs and wrong selection of part-of-speech (such as noun instead of verb and vice versa). The best performing constructions in both directions are imperatives. Concerning the wrong lexical choice, the difference in both directions is not so big: 10 wrong sentences in BG-EN direction and 20 wrong sentences in EN-BG direction. When propagated in more sentences, however, these errors constitute a serious body of semantically failed utterances. Additionally, the non-translated words in the BG-EN direction affect 16 sentences in the set of 100 ones.

Observations on German-to-English and English-to-German

Problems that Relate to Both Translation Directions. Many of the problems reported for English-Bulgarian and other languages (in both directions) can also be found in the Pilot 0 translations for English-German. In general, terminology is a frequent issue here as well, as it interrelates with, e.g., MWEs. In some cases, the verbal style of the input causes issues as in the following example where the literal German translation of *where it says* renders the sentence below hardly understandable. Omitting the phrase increases the readability of the German sentence significantly.

English sentence: In the documents list, click the arrow next to **where it says** "Last modified", and select "Quota Used".

Pilot 0: In den Dokumenten Liste, klicken Sie auf den Pfeil neben, **wo es heißt** "Last modifiziert," und wählen Sie "Quota Used".[15]

Even if not all the issues found have the same impact on the QTLeap IT-helpdesk scenario, improved accuracy and fluency will help in all scenarios.

1. Ungrammatical constructions

   Several mistreated input structures and in particular structural differences between English and German result in ungrammatical translations, often showing wrong word or phrase order. For example:

   (a) Different word/phrase order
   DE-EN. German sentence: Wie kann ich in Notepad++ eine Datei öffnen?
   Pilot 0: How can I in Notepad++ open a file?
   Expected: How can I open a file in Notepad++?

   (b) Nested sentence
   DE-EN. German sentence: Ich habe in dem Dokument, **das ich in Publisher bearbeite**, sehr viele Zeilen.
   Pilot 0: I have in the document which I in Publisher edit, very many rows.
   Expected: I have many lines in the document I am editing in Publisher.

   (c) Sentence is treated as two imperatives
   EN-DE. English sentence: You can click on the photo [...].
   Pilot 0: Sie können, klicken Sie auf das Foto [...].
   Expected: Sie können auf das Foto klicken [...].

   The examples nicely illustrate that even structurally not very complex sentences can lead to major issues with phrase-based SMT systems. Using deeper technologies, especially parsing technologies with a robust treatment of unknown terms, has a high probability of producing better translations.

2. Wrong lexical realisation

   While many problems on the lexical level are problems due to unknown terminology that result in, e.g., untranslated words and general mistranslations, we also observe some problems where the wrong reading is used in the translation because the domain reading is unknown or dispreferred compared to the general one. Agreement problems are also frequent, but they mostly do not decrease readability too strongly.

[15] In all examples in this section, we focus on the errors to be illustrated. Thus we do not comment, for example, on the untranslated terms in the given sentnece.

(a) Wrong sense

EN-DE. English sentence: MAC Address (Media Access Control) is a **physical** address associated with the communication interface that connects a device to the network.
Pilot 0: Mac Adresse (Media Access Control) ist eine **körperliche** Adresse im Zusammenhang mit dem Kommunikationsbus, dass ein Gerät mit dem Netzwerk verbindet.
Expected: ... **physikalische** ...

(b) Wrong verb form, wrong negation

EN-DE. English sentence: No, Photoshop does not allow vectors.
Pilot 0: Nein, Photoshop **nicht** Vektoren **erlauben**.
Expected: Nein, Photoshop **erlaubt keine** Vektoren.

Again, better modeling of domain knowledge and deeper analysis with generation should be able to generate more consistent translations with less lexical errors.

Specific problems for German-to-English translation. Specific problems when translating from German into English include a proper treatment of verb tense and prepositions where the intended meaning is often still recoverable. Problematic cases include pronouns both on the structural and lexical level:

1. Fake reflexive pronoun in German

   German sentence: Ich sehe **mir** in Media Player Classic ein Video an, aber das Video ist zu schnell.
   Pilot 0: I see **me** in Media Player Classic a video, but the video is too quickly.
   Expected: I am watching a video in Media Player Classic, but the video is too fast.

2. Wrong personal pronoun

   German sentence: Ich möchte in VLC den Ton über Kopfhörer hören, stattdessen kommt **er** durch die Computerlautsprecher.
   Pilot 0: I would like to hear VLC the tone through headphones, instead **he** comes through the computer speakers.
   Expected: In VLC I want to hear the sound through headphones but instead **it** goes through the computer speakers.

Solutions to these problems require a syntactic analysis of the sentence plus semantic knowledge about the antecedent of the pronoun in the second case.

Specific problems for English-to-German translation. When translating from English to German, specific problems concern verbs such as order (modal, auxiliary, and main verb), tense (especially progressive form), or agreement. In the usage domain, a specific problem is the treatment of verbs such as slide or tap that sometimes have German equivalents, sometimes use English loan words and sometimes require different constructions for proper translation. One of the best known problems that also shows frequently are missing verb prefixes. Other problems include MWEs and the right treatment of longer chunks that cannot be handled easily by phrase-based systems.

1. Verb prefix ("eingeben")

   German sentence: In the taskbar there is an icon shaped like binoculars, click and type in what you want to search.
   Pilot 0: In der Taskleiste es ist ein Symbol in Form binoculars, klicken Sie auf und **geben** Sie das, was Sie durchsuchen möchten.
   Expected: In der Aufgabenleiste gibt es ein Symbol, das wie ein Fernglas aussieht. Klicken Sie darauf und **geben** Sie **ein**, wonach was Sie suchen.

2. Modal verbs

   German sentence: You will **have to** install the Google Play Books application that you can find on Google Play.
   Pilot 0: Sie **haben** die Play **zu** installieren, dass Google Books Anwendung finden Sie auf Google Play.
   Expected: Sie **müssen** für die Google Play Books Anwendung installieren, die Sie auf Google Play finden.

3. Noun compound (incl. NE)

   German sentence: In the Kaspersky main panel, click on “Scan” and “Run Scan”.
   Pilot 0: In der **Kaspersky wichtigsten Panel**, klicken Sie auf "Scan" und "Run Scan".
   (Expected: Im **Hauptfenster von Kaspersky**, klicken Sie auf "Scannen" und "Scan beginnen".)

4. Structural difference (fronting)

   German sentence: In Google Drive, select the file you want to share [...].
   Pilot 0: **In Google Drive**, wählen Sie die Datei, die Sie teilen wollen [...].
   Expected: Wählen Sie **in Google Drive** die Datei aus, die Sie teilen wollen [...].

Observations on English to Basque, Dutch, Portuguese and Spanish and the opposite direction

1. Ungrammatical constructions

   (a) Wrong agreement

   EN-ES English text: Tap Import to copy **your Android Browser Favourites**. Pilot 0 translation: Toca Importar para copiar **su navegador de Android favoritos**. (expected: correct noun phrase internal word order: a... los Favoritos del navegador Android)

   EN-PT. English text: It is a term **associated** with... Pilot 0 translation: É um prazo **associados** a... (expected: É um termo **associado** a...)

   The NP “a term” is singular, but the translated verb “associados”, which should agree with the NP, has plural inflection.

   (b) Missing and wrong postpositions

   EN-EU English sentence: Yes it is possible, simply by dragging the profile of the person concerned to the various **circles**. Pilot 0 translation to Basque: Bai posible da, besterik gabe, arrastatu pertsonaren profila hainbat nahia **zirkulu**. (expected: adlative postposition for the bare lemma ’zirkulu’ (circle))

(a) Wrong lexical choice

i. Wrong sense due to the availability of more senses in the target language
EN-PT. English sentence: **Press** the Alt **key** on your keyboard. Pilot 0 translation: Uma **imprensa** alt **chave** em vosso teclado. (Human translation: Terá de **carregar** na **tecla** Alt do seu teclado.)
The English word "Press" has been translated as "imprensa" (in the "news media" sense) instead of "carregar", and the English words "key" has been translated as "chave" (in the "door key" sense) instead of "tecla".
EN-NL. English sentence: **Sign** into the YouTube account. Pilot 0 translation: **Teken** in de Youtube-account. (Human transation: Meldt u aan met uw YouTube-account.)

ii. Wrong sense due to the inability to get the correct sense in the target language
EN-NL. English sentence: In Linux, how can I see all the **ports** that are open on the system? Pilot 0 translation: In Linux, hoe kan ik zie alle de **havens** die openstaan op het systeem? (expected: poorten, not 'harbours')
EN-ES. English sentence: Click **run** where it says vulnerabilities. Pilot 0 translation: Pulse **correr** donde dice vulnerabilidades. (expected 'ejecutar' not 'move quickly'.)
EN-EU. English sentence: ... when you make a gesture of **scrolling** starting from the right edge of the screen, the battery status is presented on the lower left corner of the screen. Pilot 0 translation: ... keinu bat egiten duzunean **korritze** eskuineko ertzetik hasita pantailan , bateriaren egoera agertzen da beheko ezkerreko izkinan pantailan . (expected: 'mugitu' (to move) not 'korritze' (to run))

iii. Incorrect handling of multiword expressions
PT-EN. Portuguese sentence: **Faça o login** no Windows como administrador. Pilot 0 translation: **Of the login** the administrator in the windows. (expected: login)
EN-EU. English sentence: You can use **the app iPP Podcast Player** you find on Google Play. Pilot 0 translation: **Aplikazioa** erabil dezakezu **IPP podcast Player** aurkitu duzu Google erreproduzitu. (expected: iPP Podcast Player aplikazioa)
EN-ES. English sentence: What should I do to change **the Bluetooth name** on my Nexus 7? Pilot 0 translation: ¿Qué debo hacer para cambiar **el nombre** de mi Nexus 7 **Bluetooth**? (expected: el nombre de Bluetooth)

The presented examples from all project languages illustrate a lack of linguistic awareness of Pilot 0, both structural and lexical/conceptual, and it also demosntrates the need of adding more language knowledge into the MT architectures.

7 Conclusions

This deliverable presents a synopsis on the state-of-the-art in deep semantic processing: best practices, such as the Prague Dependency Treebank and PropBank; underspecified

semantics, such as Minimal Recursion Semantics (MRS); deep banks (English Deep Bank) and meaning banks (Groningen Meaning Bank).

Additionally, the survey outlines our aims beyond the state-of-the-art and discusses the envisaged improvements on the IT-helpdesk translation scenario when deep semantic processing has been applied.

The deliverable also outlines some important problems that appeared in Pilot 0, which has not employed any linguistic knowledge in the settings of bidirectional X-to-EN and EN-to-X language translation. Thus, in Pilot 1 (which is reported in D2.4) three deep entry level approaches are introduced: tectogrammatical MT; deep factored MT and Quality system combination MT.

References

I. Aldezabal, Aranzabe M., Arriola J., and Díaz de Ilarraza A. Syntactic annotation in the reference corpus for the processing of basque. In *(EPEC): Theoretical and practical issues Corpus Linguistics and Linguistic Theory 5-2*, pages 241–269. Mouton de Gruyter. Berlin-New York, 2009.

I. Aldezabal, Aranzabe M., Arriola J., and Díaz de Ilarraza A. A methodology for the semiautomatic annotation of epec-rolsem, a basque corpus labeled at predicative level following the propbank-verb net model. Technical report, 2013.

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. Developing a large semantically annotated corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Ondrej Bojar, Silvie Cinková, and Jan Ptácek. Towards english-to-czech MT via tectogrammatical layer. *Prague Bull. Math. Linguistics*, 90:57–68, 2008. URL http://ufal.mff.cuni.cz/pbml/90/art-bojar-et-al.pdf.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W13-2201.

Francis Bond, Stephan Oepen, Melanie Siegel, Ann Copestake, and Dan Flickinger. Open source machine translation with DELPH-IN. In *Proceedings of the Open-Source Machine Translation Workshop at the 10th Machine Translation Summit*, pages 15 – 22, Phuket, Thailand, September 2005.

Francis Bond, Stephan Oepen, Eric Nichols, Dan Flickinger, Erik Velldal, and Petter Haugereid. Deep open source machine translation. *Machine Translation*, 2011. to appear.

Johan Bos. The groningen meaning bank. In *Invited Talk at Joint Symposium on Semantic Processing: Textual Inference and Structures in Corpora*, 2013.

Johan Bos and Rodolfo Delmonte. *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1 of *Research in Computational Semantics*. College Publications, 2008.

Antonia Branco and Francisco Costa. LXGram in the Shared Task "Comparing Semantic Representations" of STEP 2008. In Johan Bos and Rodolfo Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1 of *Research in Computational Semantics*, pages 299–314. College Publications, 2008. URL http://www.aclweb.org/anthology/W08-2224.

Ted Briscoe and John Carroll. Robust accurate statistical annotation of general text. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, Las Palmas, Canary Islands - Spain, May 2002. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2002/pdf/250.pdf. ACL Anthology Identifier: L02-1250.

Xavier Carreras and Lluís Màrquez. Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling. In *Proceedings of the 8th Conference on Natural Language Learning, CoNLL-2004*, Boston, MA USA, 2004.

Xavier Carreras and Lluís Màrquez. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In *Proceedings of the 9th Conference on Natural Language Learning, CoNLL-2005*, Ann Arbor, MI USA, 2005.

Ann Copestake. Robust minimal recursion semantics. unpublished draft., 2004/2006. URL http://www.cl.cam.ac.uk/~aac10/papers/rmrsdraft.pdf.

Ann Copestake. Applying robust semantics. In *Proceedings of the 10th Conference of the Pacific Assocation for Computational Linguistics (PACLING)*, pages 1–12, 2007.

Ann Copestake and Dan Flickinger. An open source grammar development environment and broad-coverage english grammar using hpsg. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athens, Greece, 2000.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan Sag. Minimal recursion semantics: An introduction. *Research on Language & Computation*, 3(4):281–332, 2005.

Francisco Costa and António Branco. Lxgram: A deep linguistic processing grammar for Portuguese. In *Lecture Notes in Artificial Intelligence*, volume 6001, pages 86–89. Springer, Berlin, May 2010. URL http://nlx.di.fc.ul.pt/~fcosta/papers/propor2010.pdf.

Berthold Crysmann. Local ambiguity packing and discontinuity in german. In *ACL 2007 Workshop on Deep Linguistic Processing*, pages 144–151, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W07/W07-1219.

James Curran, Stephen Clark, and Johan Bos. Linguistically motivated large-scale nlp with c&c and boxer. In *Proceedings of the 45th Annual Meeting of the Association*

for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 33–36, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P07-2009.

Daniel Flickinger, Valia Kordoni, Yi Zhang, António Branco, Kiril Simov, Petya Osenova, Catarina Carvalheiro, Francisco Costa, and Sérgio Castro. Pardeepbank: Multiple parallel deep treebanking. In *Proceedings of the Eleventh International Workshop on Treebanks and Linguistic Theories. International Workshop on Treebanks and Linguistic Theories (TLT-11), 11th, November 30 - December 1, Lisbon, Portugal*, pages 97–108. Edições Colibri, Lisbon, 2012a. URL http://www.dfki.de/web/research/publications/renameFileForDownload?filename=ParDeepBank_TLT11.pdf&file_id=uploads_1874.

Daniel Flickinger, Yi Zhang, and Valia Kordoni. Deepbank: A dynamically annotated treebank of the wall street journal. In *Proceedings of the Eleventh International Workshop on Treebanks and Linguistic Theories. International Workshop on Treebanks and Linguistic Theories (TLT-11), 11th, November 30 - December 1, Lisbon, Portugal*, pages 85–96. Edições Colibri, Lisbon, 2012b. URL http://www.dfki.de/web/research/publications/renameFileForDownload?filename=DeepBank_tlt11.pdf&file_id=uploads_1864.

Bart Geurts and David I. Beaver. Discourse representation theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2011 edition, 2011.

Jan Hajič. Machine translation research in META-NET. presentation at META-NET meeting., 2011.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, CoNLL '09, pages 1–18, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-29-9. URL http://dl.acm.org/citation.cfm?id=1596409.1596411.

Angelina Ivanova, Stephan Oepen, Lilja Øvrelid, and Dan Flickinger. Who did what to whom? a contrastive study of syntacto-semantic dependencies. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 2–11, Jeju, Republic of Korea, July 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W12-3602.

Angelina Ivanova, Stephan Oepen, Rebecca Dridan, Dan Flickinger, and Lilja Øvrelid. On Different Approaches to Syntactic Analysis Into Bi-Lexical Dependencies. An Empirical Comparison of Direct, PCFG-Based, and HPSG-Based Parsers. In *Proceedings of The 13th International Conference on Parsing Technologies (IWPT-2013)*, pages 63–72, Nara, Japan, 2013.

Max Jakob, Markéta Lopatková, and Valia Kordoni. Mapping between dependency structures and compositional semantic representations. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 2491–2497, 2010.

Hans Kamp. A theory of truth and semantic representation. In Groenendijk, editor, *Formal Methods in the Study of Language*, pages 189—222. 1981.

Paul Kingsbury and Martha Palmer. From treebank to propbank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, Las Palmas, Canary Islands - Spain, May 2002. European Language Resources Association (ELRA).

Philipp Koehn and Hieu Hoang. Factored translation models. In *Proceedings of EMNLP*, 2007.

Phong Le and Willem Zuidema. Learning compositional semantics for open domain semantic parsing. In *Proceedings of COLING 2012*, pages 1535–1552, Mumbai, India, December 2012. The COLING 2012 Organizing Committee. URL http://www.aclweb.org/anthology/C12-1094.

Beth Levin. *English verb classes and alternations : a preliminary investigation*. 1993. URL http://books.google.com/books?id=Nbh0C3Ac90MC&dq=English+verb+classes+and+alternations+a+preliminary+investigation&printsec=frontcover&source=bl&ots=3faip55J1U&sig=_mJS8VCitUKL1ZpjKoNelYFQMvQ&hl=en&ei=cVp3SrKiNpL0sQP_tpntBA&sa=X&oi=book_result&ct=result&resnum=3#v=onepage&q=&f=false.

Montserrat Marimon, Núria Bel, Sergio Espeja, and Natalia Seghezzi. The spanish resource grammar: Pre-processing strategy and lexical acquisition. In *ACL 2007 Workshop on Deep Linguistic Processing*, pages 105–111, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W07/W07-1214.

Ryan McDonald. *Discriminative Training and Spanning Tree Algorithms for Dependency Parsing*. PhD thesis, 2006.

Joakim Nivre, Johan Hall, and Jens Nilsson. Maltparser: a data-driven parser-generator for dependency parsing. In *Proceedings of LREC-2006*, 2006.

Stephan Oepen. [incr tsdb()] - Competence and Performance Laboratory. Technical report, Saarland University, 1999.

Stephan Oepen. The Transfer Formalism. General Purpose MRS Rewriting. Technical Report LOGON Project. Technical report, University of Oslo, 2008.

Stephan Oepen and Jan Tore Lønning. Discriminant-based mrs banking. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy, May 2006.

Stephan Oepen, Helge Dyvik, Jan Tore Lønning, Erik Velldal, Dorothee Beermann, John Carroll, Dan Flickinger, Lars Hellan, Janne Bondi Johannessen, Paul Meurer, Torbjørn Nordgård, and Victoria Rosén. Som å kapp-ete med trollet? Towards MRS-based Norwegian – English Machine Translation. In *In Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, Baltimore, MD, October 2004.

Petya Osenova. *The Bulgarian Resource Grammar*. VDM, 2010.

E. Hajicova P. Sgall and J. Panevova. *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. Dordrecht: Reidel Publishing Company and Prague: Academia, 1986.

Martha Palmer, Joseph Rosenzweig, and Scott Cotton. Automatic predicate argument analysis of the penn treebank. In *Proceedings of the First International Conference on Human Language Technology Research*, HLT '01, pages 1–5, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics. doi: 10.3115/1072133.1072143. URL http://dx.doi.org/10.3115/1072133.1072143.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist.*, 31(1):71–106, March 2005. ISSN 0891-2017. doi: 10.1162/0891201053630264. URL http://dx.doi.org/10.1162/0891201053630264.

Martha Palmer, Ivan Titov, and Shumin Wu. Semantic role labeling. In *NAACL HLT 2013 Tutorial Abstracts*, pages 10–12, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N13-4004.

Martin Popel and Zdeněk Žabokrtský. Tectomt: Modular nlp framework. In *Proceedings of the 7th International Conference on Advances in Natural Language Processing*, IceTAL'10, pages 293–304, Berlin, Heidelberg, 2010. Springer-Verlag. ISBN 3-642-14769-0, 978-3-642-14769-2. URL http://dl.acm.org/citation.cfm?id=1884371.1884406.

Sameer S. Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. Weischedel, ontonotes: A unified relational semantic representation. *International Journal of Semantic Computing*, pages 405–419, 2007.

Owen Rambow, Bonnie Dorr, Ivona Kučerová, and Martha Palmer. Automatically deriving tectogrammatical labels from other resources: A comparison of semantic labels across frameworks. the prague. *Bulletin of Mathematical Linguistics*, page 3866, 2003.

Aleksandar Savkov, Laska Laskova, Stanislava Kancheva, Petya Osenova, and Kiril Simov. Linguistic processing pipeline for bulgarian. In *Proceedings of LREC*, Istanbul, Turkey, 2012.

Kiril Simov and Petya Osenova. Language resources and tools for ontology-based semantic annotation. In *Proc. of the OntoLex 2008 Workshop at LREC 2008*, pages 9–13, 2008.

Kiril Simov and Petya Osenova. Towards minimal recursion semantics over bulgarian dependency parsing. In *Proceedings of the RANLP 2011*, 2011.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. The conll-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, CoNLL '08, pages 159–177, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. ISBN 978-1-905593-48-4. URL http://dl.acm.org/citation.cfm?id=1596324.1596352.

Rui Wang, Petya Osenova, and Kiril Simov. Linguistically-augmented Bulgarian-to-English Statistical Machine Translation Model. In *Proceedings of the Joint Workshop on Exploiting Synergies Between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, EACL 2012, pages 119–128, Stroudsburg, PA, USA, 2012a. Association for Computational Linguistics. ISBN 978-1-937284-19-0. URL http://dl.acm.org/citation.cfm?id=2387956.2387972.

Rui Wang, Petya Osenova, and Kiril Simov. Linguistically-enriched Models for Bulgarian-to-English Machine Translation. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, SSST-6 '12, pages 10–19, Stroudsburg, PA, USA, 2012b. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2392936.2392939.

Luke Zettlemoyer and Michael Collins. Online learning of relaxed CCG grammars for parsing to logical form. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 678–687, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D/D07/D07-1071.

Yi Zhang and Hans-Ulrich Krieger. Large-scale corpus-driven pcfg approximation of an hpsg. In *Proceedings of the 12th International Conference on Parsing Technologies*, pages 198–208, Dublin, Ireland, October 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W11-2923.