

# The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages

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‡‡ Yi Zhang‡◇

\*: Charles University in Prague, {hajic, stepanek, stranak}@ufal.mff.cuni.cz

†: Google Inc., massi@google.com

\*◇: Uppsala University and Växjö University, joakim.nivre@lingfil.uu.se

‡: University of Trento, johansson@disi.unitn.it

◇: National Institute of Information and Communications Technology, dk@nict.go.jp

\*\*: University of Barcelona, amarti@ub.edu

\*†: Technical University of Catalonia, Barcelona, lluism@lsi.upc.edu

\*‡: New York University, meyers@cs.nyu.edu

\*◇: Uppsala University, joakim.nivre@lingfil.uu.se

‡\*: Stanford University, mihais@stanford.edu

‡‡: Brandeis University, xuen@brandeis.edu

‡◇: Saarland University, yzhang@coli.uni-sb.de

◇\*: Stuttgart University, pado@ims.uni-stuttgart.de

## Abstract

For the 11th straight year, the Conference on Computational Natural Language Learning has been accompanied by a shared task whose purpose is to promote natural language processing applications and evaluate them in a standard setting. In 2009, the shared task was dedicated to the joint parsing of syntactic and semantic dependencies in multiple languages. This shared task combines the shared tasks of the previous five years under a unique dependency-based formalism similar to the 2008 task. In this paper, we define the shared task, describe how the data sets were created, report and analyze the results and summarize the approaches of the participating systems.

## 1 Introduction

Every year since 1999, the Conference on Computational Natural Language Learning (CoNLL) launches a competitive, open “Shared Task”. A common (“shared”) task is defined and datasets are provided for its participants. In 2004 and 2005, the shared tasks were dedicated to semantic role labeling (SRL), in a monolingual setting (English). In 2006 and 2007 the shared tasks were devoted to

the parsing of syntactic dependencies, using corpora from up to 13 languages. In 2008, the shared task used a unified dependency-based formalism, which modelled both syntactic dependencies and semantic roles for English. The CoNLL-2009 shared task, which has built on the CoNLL-2008 shared task (Surdeanu et al., 2008) by extending it to seven languages (Catalan, Chinese, Czech, English, German, Japanese and Spanish), has thus naturally extended the path of the five most recent CoNLL shared tasks.

As in 2008, the CoNLL-2009 shared task combined dependency parsing and the task of identifying and labeling semantic arguments of verbs (and other parts of speech whenever available). Participants had to choose from two tasks:

- Joint task (syntactic dependency parsing *and* semantic role labeling), or
- SRL-only task (syntactic dependency parses have been provided by the organizers, using state-of-the art parsers for the individual languages).

In contrast to the previous year, the evaluation data indicates which words were to be dealt with (for the

SRL task). In other words, (predicate) disambiguation is still part of the task, whereas the *identification* of argument-introducing words was not. This decision was made to compensate for the significant differences between languages and between the annotation schemes used.

The “closed” and “open” challenges have been kept from last year as well; participants could have chosen one or both. In the closed challenge, systems have to be trained strictly with information contained in the given training corpus; in the open challenge, systems can be developed making use of any kind of external tools and resources.

This paper is organized as follows. Section 2 defines the task, including the format of the data, the evaluation metrics, and the two challenges. Section 3 introduces the corpora used and our constituent-to-dependency conversion procedure. Section 4 summarizes the results of the submitted systems. Section 5 discusses the approaches implemented by participants. Section 6 analyzes the results using additional non-official evaluation measures. Section 7 concludes the paper. In all sections, we will concentrate on the differences between last year’s and this year’s tasks while keeping the text self-contained whenever possible. A substantial portion of the paper will be devoted to the conversion and development of the data sets in the additional languages.

## 2 Task Definition

In this section we provide the definition of the shared task; after introducing the two challenges and the two tasks the participants were to choose, we continue with the format of the shared task data, followed by a description of the evaluation metrics used.

For three of the languages (Czech, English and German), out-of-domain data (OOD) have also been prepared for the final evaluation, following the same guidelines and formats.

### 2.1 Closed and Open Challenges

Similarly to the CoNLL-2005 and CoNLL-2008 shared tasks, this shared task evaluation is separated into two challenges:

**Closed Challenge** The aim of this challenge is to compare the performance of the participating systems in a fair environment. Systems have to be built strictly with information contained in the given training corpus, and tuned with the development section. In addition, the lexical frame files (such as the PropBank and NomBank for English, the valency dictionary PDT-Vallex for Czech etc.) can be used. These restrictions mean that outside parsers (not trained by the participants’ systems) cannot be used. However, we do provide the output of a single, state-of-the-art dependency parser for each language so that participants can build a SRL-only system (using the provided parses as inputs) within the closed challenge (as opposed to the 2008 shared task).

**Open Challenge** Systems can be developed making use of any kind of external tools and resources. The only condition is that such tools or resources must not have been developed with the annotations of the test set, both for the input and output annotations of the data. In this challenge, we are interested in learning methods which make use of any tools or resources that might improve the performance. The comparison of different systems in this setting may not be fair, and thus ranking of systems is not necessarily important.

### 2.2 Joint and SRL-only tasks

In 2008, systems participating in the open challenge could have used state-of-the-art parsers for the syntactic dependency part of the task. This year, we have provided the output of these parsers for all the languages in an uniform way, thus allowing orthogonal combination of the two tasks and the two challenges. For the SRL-only task, participants in the closed challenge simply had to use the provided parses only.

Despite the provisions for the SRL-only task, we are more interested in the approaches and results of the Joint task. Therefore, primary system ranking is provided for the Joint task, whereas the comparison of various combinations of parsers and the SRL methods across the tasks and challenges is also interesting.

### 2.3 Data Format

The data format used in this shared task was based on the CoNLL-2008 shared task, with some differ-

ences. The data follows these general rules (same as in 2008):

- The files contain sentences separated by a blank line.
- A sentence consists of one or more tokens and the information for each token is represented on a separate line.
- A token consists of at least 14 fields. The fields are separated by one or more whitespace characters (spaces or tabs). Whitespace characters are not allowed within fields.

The data is thus a large table with whitespace-separated fields (columns). The fields provided in the data are described in Table 1. They are identical for all languages, but they may differ in contents; for example, some fields might not be filled for all the languages provided (such as the FEAT or PFEAT fields).

It was required that participants submit results in all seven languages in the chosen task and in any (or both) the challenges. Submission of out-of-domain data files have been optional.

For the SRL-only task, participants have been provided with all the data but the PRED and APREDs, which they were supposed to fill in with their correct values. However, they did not have to determine which tokens are predicates (or more precisely, which are the argument-bearing tokens), since they were marked by 'Y' in the FILLPRED field.

For the main Joint task, participants could not (in addition to the PRED and APREDs) see the gold-standard nor the predicted syntactic dependencies (HEAD, PHEAD) and their labels (DEPREL, PDEPREL). These syntactic dependencies were also to be filled by participants' systems.

In both tasks, participants have been free to use any other data (columns) provided, except the LEMMA, POS and FEAT columns (to get more 'realistic' results using only their automatically predicted variants PLEMMA, PPOS and PFEAT).

Besides the corpus proper, predicate dictionaries have been provided to participants in order to be able to properly match the predicates to the tokens in the corpus; their contents could have been used

also (e.g.) as features for the PRED/APREDs predictions (or even for the syntactic dependencies, i.e., for filling in the PHEAD and PDEPREL fields).

The system of filling-in the APREDs follows the 2008 pattern: for each argument-bearing token (predicate), a new APRED<sub>n</sub> column is created in the order in which the predicate token is encountered within the sentence (i.e., based on its ID seen as a numerical value). Then, for each token in the sentence, the value in the intersection of the APRED<sub>n</sub> column and the token row is either left unfilled (if the token is not an argument), or a predicate-argument label(s) is(are) filled in.

The differences between the English-only 2008 task and this year's multilingual task can be briefly summarized as follows: only "split"<sup>1</sup> lemmas and forms have been provided in the English datasets (for the other languages, original tokenization from the respective treebanks has been used); added rich morphological features whenever available; syntactic dependencies provided for the SRL-only task; multiple semantic labels are allowed (and properly evaluated) in the APREDs columns; marking of predicates in the test data; and some renaming of the data columns took place.

## 2.4 Evaluation Measures

The main evaluation measure, according to which systems are primarily compared, is the Joint task, closed challenge, Macro F<sub>1</sub> score. However, scores can also be computed for a number of other conditions:

- Task: Joint or SRL-only
- Challenge: open or closed
- Domain: in-domain data (IDD, separated from training corpus) or out-of-domain data (OOD)

Joint task participants are also evaluated separately on the syntactic dependency task (labeled attachment score, LAS). Finally, systems competing in both tasks are compared on semantic role labeling alone, to assess the impact of the the joint parsing/SRL task compared to an SRL-only task on parsed data.

<sup>1</sup>Splitting of forms and lemmas in English has been introduced in the 2008 shared task to match the tokenization convention for the arguments in NomBank.

Field #	Name	Description
1	ID	Token counter, starting at 1 for each new sentence
2	FORM	Form or punctuation symbol (the token; “split” for English)
3	LEMMA	Gold-standard lemma of FORM
4	PLEMMA	Automatically predicted lemma of FORM
5	POS	Gold-standard POS (major POS only)
6	PPOS	Automatically predicted major POS by a language-specific tagger
7	FEAT	Gold-standard morphological features (if applicable)
8	PFEAT	Automatically predicted morphological features (if applicable)
9	HEAD	Gold-standard syntactic head of the current token (ID or 0 if root)
10	PHEAD	Automatically predicted syntactic head
11	DEPREL	Gold-standard syntactic dependency relation (to HEAD)
12	PDEPREL	Autoamtically predicted dependency relation to PHEAD
13	FILLPRED	Contains ‘Y’ for argument-bearing tokens
14	PRED	(sense) identifier of a semantic “predicate” coming from a current token
15...	APRED <sub>n</sub>	Columns with argument labels for each semantic predicate (in the ID order)

Table 1: Description of the fields (columns) in the data provided. The values of columns 9, 11 and 14 and above are not provided in the evaluation data; for the Joint task, columns 9–12 are also empty in the evaluation data.

Finally, as an explanatory measure, precision and recall of the semantic labeling task have been computed and tabulated.

We have decided to omit several evaluation figures that were reported in previous years, such as the percentage of completely correct sentences (“Exact Match”), unlabeled scores, etc.. With seven languages, two tasks (plus two challenges, and the IDD/OOD distinction), there are enough results to get lost even as it is.

#### 2.4.1 Syntactic Dependency Measures

The LAS score is defined similarly as in the previous shared tasks, as the percentage of tokens for which a system has predicted the correct HEAD and DEPREL columns. The unlabeled attachment score (UAS), i.e., the percentage of tokens with correct HEAD regardless if the DEPREL is correct, has not been officially computed this year. No precision and recall measures are applicable, since all systems are supposed to output a single dependency with a single label (see also below the footnote to the description of the combined score).

#### 2.4.2 Semantic Labeling Measures

The semantic propositions are evaluated by converting them to semantic dependencies, i.e., we create  $n$  semantic dependencies from every predicate to its  $n$  arguments. These dependencies are labeled with the labels of the corresponding arguments. Ad-

ditionally, we create a semantic dependency from each predicate to a virtual ROOT node. The latter dependencies are labeled with the predicate senses. This approach guarantees that the semantic dependency structure conceptually forms a single-rooted, connected (but not necessarily acyclic) graph. More importantly, this scoring strategy implies that if a system assigns the incorrect predicate sense, it still receives some points for the arguments correctly assigned. For example, for the correct proposition:

`verb.01: ARG0, ARG1, ARGM-TMP`

the system that generates the following output for the same argument tokens:

`verb.02: ARG0, ARG1, ARGM-LOC`

receives a labeled precision score of 2/4 because two out of four semantic dependencies are incorrect: the dependency to ROOT is labeled 02 instead of 01 and the dependency to the ARGM-TMP is incorrectly labeled ARGM-LOC. Using this strategy we compute precision, recall, and  $F_1$  scores for semantic dependencies (labeled only).

For some languages (Czech, Japanese) there may be more than one label in a given argument position; in general, this happens in special cases of coreference when the same token serves as two or more arguments to the same predicate. The scorer takes this into account and considers such cases to be (as if) multiple predicate-argument relations for the com-

putation of the evaluation measures.

For example, for the correct proposition:

```
v1f1: ACT|EFF, ADDR
```

the system that generates the following output for the same argument tokens:

```
v1f1: ACT, ADDR|PAT
```

receives a labeled precision score of 3/4 because the PAT is incorrect and labeled recall 3/4 because the EFF is missing (should the ACT|EFF and ADDR|PAT be taken as atomic values, the scores would then be zero).

### 2.4.3 Combined Syntactic and Semantic Score

We combine the syntactic and semantic measures into one global measure using macro averaging. We compute macro precision and recall scores by averaging the labeled precision and recall for semantic dependencies with the LAS for syntactic dependencies:<sup>2</sup>

$$LMP = W_{sem} * LP_{sem} + (1 - W_{sem}) * LAS \quad (1)$$

$$LMR = W_{sem} * LR_{sem} + (1 - W_{sem}) * LAS \quad (2)$$

where  $LMP$  is the labeled macro precision and  $LP_{sem}$  is the labeled precision for semantic dependencies. Similarly,  $LMR$  is the labeled macro recall and  $LR_{sem}$  is the labeled recall for semantic dependencies.  $W_{sem}$  is the weight assigned to the semantic task.<sup>3</sup> The macro labeled  $F_1$  score, which was used for the ranking of the participating systems, is computed as the harmonic mean of  $LMP$  and  $LMR$ .

## 3 Data

The unification of the data formats for the various languages appeared to be a challenge in itself. We will briefly describe the processes of the conversion of the existing treebanks in the seven languages of the CoNLL-2009 shared task. In many instances, the original treebanks had to be not only converted format-wise, but also merged with other resources in order to generate useful training and testing data that fit the task description.

<sup>2</sup>We can do this because the LAS for syntactic dependencies is a special case of precision and recall, where the predicted number of dependencies is equal to the number of gold dependencies.

<sup>3</sup>We assign equal weight to the two tasks, i.e.,  $W_{sem} = 0.5$ .

### 3.1 The Input Corpora

The data used as the input for the transformation aimed at arriving at the data contents and format described in Sect. 2.3 are described in (Taulé et al., 2008), (Xue and Palmer, 2009), (Hajič et al., 2006), (Surdeanu et al., 2008), (Burchardt et al., 2006) and (Kawahara et al., 2002).

In the subsequent sections, the procedures for the data conversion for the individual languages are described. The data has been collected by the main organization site and checked for format errors, and repackaged for distribution.

There were three packages of the data distributed to the participants: Trial, Training plus Development, and Evaluation. The Trial data were rather small, just to give the feeling of the format and languages involved.

Due to licensing requirements, every package of the data had to be split into two portions. One portion (Catalan, German, Japanese, and Spanish data) was published on the task’s webpage for download, the other portion (Czech, English, and Chinese data) was invoiced and distributed by the Linguistic Data Consortium under a special agreement free of charge.

Distribution of the Evaluation package was a bit more complicated, because there were two types of the package - one for the Joint task and one for the SRL-only task. Every participant had to subscribe to one of the two tasks; subsequently, he or she obtained the appropriate data (again, from the webpage and LDC).

Prior to release, each data file was checked to eliminate errors. The following test were carried out:

- For every sentence, number of PREDs rows matches the number of APREDs columns.
- The first line of each file is never empty, while the last line always is.
- The first character on a non-empty line is always a digit, the last one is never a whitespace.
- The number of empty lines (i.e. the number of sentences) equals the number of lines beginning with “1”.
- The data contain no spaces nor double tabs.

### 3.2 Catalan and Spanish

The Catalan and Spanish datasets (Taulé et al., 2008) were generated from the AnCora corpora<sup>4</sup> through an automatic conversion process from a constituent-based formalism to dependencies (Civit et al., 2006).

AnCora corpora contain about half million words for Catalan and Spanish annotated with syntactic and semantic information. Text sources for the Catalan corpus are EFE news agency (~75Kw), ACN Catalan news agency (~225Kw), and ‘El Periódico’ newspaper (~200Kw). The Spanish corpus comes from the Lexesp Spanish balanced corpus (~75Kw), the EFE Spanish news agency (~225Kw), and the Spanish version of ‘El Periódico’ (~200Kw). The subset from ‘El Periódico’ corresponds to the same news in Catalan and Spanish, spanning from January to December 2000.

Linguistic annotation is the same in both languages and includes: PoS tags with morphological features (gender, number, person, etc.), lemmatization, syntactic dependencies (syntactic functions), semantic dependencies (arguments and thematic roles), named entities and predicate semantic classes (Lexical Semantic Structure, LSS). Tag sets are shared by the two languages.

If we take into account the complete PoS tags, AnCora has 280 different labels. Considering only the main syntactic categories, the tag set is reduced to 47 tags. The syntactic tag set consists of 50 different syntactic functions. Regarding semantic arguments, we distinguish Arg0, Arg1, Arg2, Arg3, Arg4, ArgM, and ArgL. The first five tags are numbered from less to more obliqueness with respect to the verb, ArgM corresponds to adjuncts. The list of thematic roles consists of 20 different labels: AGT (Agent), AGI (Induced Agent), CAU (Cause), EXP (Experiencer), SCR (Source), PAT (Patient), TEM (Theme), ATR (Attribute), BEN (Beneficiary), EXT (Extension), INS (Instrument), LOC (Locative), TMP (Time), MNR (Manner), ORI (Origin), DES (Goal), FIN (Purpose), EIN (Initial State), EFI (Final State), and ADV (Adverbial). Each argument position can map onto specific thematic roles. By way of example, Arg1 can be PAT, TEM or EXT. For Named Entities, we distinguish six types: Organization, Person, Location, Date, Number, and Others.

An incremental process guided the annotation of AnCora, since semantics depends on morphosyntax, and syntax relies on morphology. This procedure made it possible to check, correct, and complete the previous annotations, thus guaranteeing the final quality of the corpora and minimizing the error rate. The annotation process was carried out sequentially from lower to upper layers of linguistic description. All resulting layers are independent of each other, thus making easier the data management. The initial annotation was performed manually for syntax, semiautomatically in the case of arguments and thematic roles, and fully automatically for PoS (Martí et al., 2007; Màrquez et al., 2007).

The Catalan and Spanish AnCora corpora were straightforwardly translated into the CoNLL-2009 shared task formatting (information about named entities was skipped in this process). The resulting Catalan corpus (including training, development and test partitions) contains 16,786 sentences with an average length of 29.59 lexical tokens per sentence. Long sentences abound in this corpus. For instance, 10.73% of the sentences are longer than 50 tokens, and 4.42% are longer than 60. The corpus contains 47,537 annotated predicates (2.83 predicates per sentence, on average) with 107,171 arguments (2.25 arguments per predicate, on average). From the latter, 73.89% correspond to core arguments and 26.11% to adjuncts. Numbers for the Spanish corpus are comparable in all aspects: 17,709 sentences with 29.84 lexical tokens on average (11.58% of the sentences longer than 50 tokens, 4.07% longer than 60); 54,075 predicates (3.05 per sentence, on average) and 122,478 arguments (2.26 per predicate, on average); 73.34% core arguments and 26.66% adjuncts.

The following are important features of the Catalan and Spanish corpora in the CoNLL-2009 shared task setting: (1) all dependency trees are projective; (2) no word can be the argument of more than one predicate in a sentence; (3) semantic dependencies completely match syntactic dependency structures (i.e., no new edges are introduced by the semantic structure); (4) only verbal predicates are annotated (with exceptional cases referring to words that can be adjectives and past participles); (5) the corpus is segmented so multi-words, named entities, temporal expressions, compounds, etc. are grouped together;

<sup>4</sup><http://clic.ub.edu/ancora/>

and (6) segmentation also accounts for elliptical pronouns (there are marked as empty lexical tokens ‘\_’ with a pronoun POS tag).

Finally, the predicted columns (PLEMMA, PPOS, and PFEAT) have been generated with the FreeLing Open source suite of Language Analyzers<sup>5</sup>. Accuracy in PLEMMA and PPOS columns is above 95% for the two languages. PHEAD and PDEPREL columns have been generated using MaltParser<sup>6</sup>. Parsing accuracy (LAS) is above 86% for the the two languages.

### 3.3 Chinese

The Chinese Corpus for the 2009 CoNLL Shared Task was generated by merging the Chinese Treebank (Xue et al., 2005) and the Chinese Proposition Bank (Xue and Palmer, 2009) and then converting the constituent structure to a dependency formalism as specified in the CoNLL Shared Task. The Chinese data used in the shared task is based on Chinese Treebank 6.0 and the Chinese Proposition Bank 2.0, both of which are publicly available via the Linguistic Data Consortium.

The Chinese Treebank Project originated at Penn and was later moved to University of Colorado at Boulder. Now it is the process of being moved to Brandeis University. The data sources of the Chinese Treebank range from Xinhua newswire (mainland China), Hong Kong news, and Sinorama Magazine (Taiwan). More recently under DARPA GALE funding it has been expanded to include broadcast news, broadcast conversation, news groups and web log data. It currently has over one million words and is fully segmented, POS-tagged and annotated with phrase structure. The version of the Chinese Treebank used in this shared task, CTB 6.0, includes newswire, magazine articles, and transcribed broadcast news<sup>7</sup>. The training set has 609,060 tokens, the development set has 49,620 tokens, and the test set has 73,153 tokens.

The Chinese Proposition Bank adds a layer of semantic annotation to the syntactic parses in the Chinese Treebank. This layer of semantic annotation

mainly deals with the predicate-argument structure of Chinese verbs and their nominalizations. Each major sense (called *frameset*) of a predicate takes a number of *core* arguments annotated with numerical labels *Arg0* through *Arg5* which are defined in a predicate-specific manner. The Chinese Proposition Bank also annotates adjunctive arguments such as locative, temporal and manner modifiers of the predicate. The version of the Chinese Propbank used in this CoNLL Shared Task is CPB 2.0, but nominal predicates are excluded because the annotation is incomplete.

Since the Chinese Treebank is annotated with constituent structures, the conversion and merging procedure converts the constituent structures to dependencies by identifying the head for each constituent in a parse tree and making its sisters its dependents. The Chinese Propbank pointers are then shifted from the entire constituent to the head of that constituent. The conversion procedure identifies the head by first exploiting the structural information in the syntactic parse and detecting six broad categories of syntactic relations that hold between the head and its dependents (*predication, modification, complementation, coordination, auxiliary, and flat*) and then designating the head based on these relations. In particular, the first conjunct of a coordination structure is designated as the head and the heads of the other conjuncts are the conjunctions preceding them. The conjunctions all “modify” the first conjunct.

### 3.4 Czech

For the training, development and evaluation data, Prague Dependency Treebank 2.0 was used (Hajič et al., 2006). For the out-of-domain evaluation data, part of the Czech side of the Prague Czech-English Dependency Treebank (version 2, under construction) was used<sup>8</sup>, see also (Čmejrek et al., 2004). The same conversion procedure has been applied to both sources.

The FORM column was created from the *form* element of the morphological layer, not from the “token” from the word-form layer. Therefore, most typos, errors in word segmentation and tokenization are corrected and numerals are normalized.

<sup>5</sup><http://www.lsi.upc.es/~nlp/freeling/>

<sup>6</sup><http://w3.msi.vxu.se/~jha/maltparser/>

<sup>7</sup>A small number of files are taking out of the CoNLL shared task data due to conversion problems and time constraints to fix them.

<sup>8</sup><http://ufal.mff.cuni.cz/peft>

The LEMMA column was created from the lemma element of the morphological layer. Only the initial string of the element was used, so there is no distinction between homonyms. However, some components of the detailed lemma explanation were incorporated into the FEAT column.

The POS column was created from the morphological tag element, its first character more precisely.

The FEAT column was created from the remaining characters of the tag element. The special feature "Sem" corresponds to a semantical feature of the lemma.

For the HEAD and DEPREL columns, the analytical layer was used. In detail, for every word the DEPREL equals to its analytical function (the afun element). The HEAD is a pointer to its parent, which means the "ord" element of the parent. If a node is a member of a coordination or apposition (is\_member element), its DEPREL obtains the M suffix.

The PRED and APREDS columns were created from the tectogrammatical layer of PDT 2.0 and the valency lexicon PDT-Vallex according to the following rules:

- Every line corresponding to an analytical node referenced by a lexical reference (a/lex.rf) from the tectogrammatical layer has a PRED value filled. If the referring non-generated tectogrammatical node (is\_generated not equal to 1) has a valency frame assigned (val\_frame.rf), the value of PRED is the identifier of the frame. Otherwise, it is set to the same value as the LEMMA column.
- For every tectogrammatical node, a corresponding analytical node is searched for:
  1. If the tectogrammatical node is not generated and has a lexical reference (a/lex.rf), the referenced node is taken.
  2. Otherwise, if the tectogrammatical node has a coreference (coref\_text.rf or coref\_gram.rf) or complement reference (compl.rf) to a node that has an analytical node assigned (by 1. or 2.), the assigned node is taken.

APRED columns are filled with respect to the following correspondence: for a tectogrammatical node P and its effective child C with functor F, the column for P's corresponding analytical node at the row for C's corresponding analytical node is filled with F. (Some nodes can thus have several functors in one APRED column, they are separated by a vertical bar).

PLEMMA, PPOS and PFEAT were generated by the (cross-trained) morphological tagger MORCE (Spoustová et al., 2009).

PHEAD and PDEPREL were generated by the (cross-trained) MST parser for Czech (Chui-Liu/Edmonds algorithm, (McDonald et al., 2005)).

The valency lexicon, converted from (Hajič et al., 2003), has four columns:

1. lemma (can occur several times in the lexicon, with different frames)
2. frame identifier (as found in the PRED column)
3. list of space-separated actants and obligatory members of the frame
4. example(s)

### 3.5 English

The English corpus is almost identical to the corpus used in the closed challenge in the CoNLL-2008 shared task evaluation (Surdeanu et al., 2008). This corpus was generated through a process that merges several input corpora and converts them from the constituent-based formalism to dependencies. The following corpora were used as input to the merging procedure:

- **Penn Treebank 3** - The Penn Treebank 3 corpus (Marcus et al., 1994) consists of hand-coded parses of the Wall Street Journal (test, development and training) and a small subset of the Brown corpus (W. N. Francis and H. Kucera, 1964) (test only).
- **BBN Pronoun Coreference and Entity Type Corpus** - BBN's NE annotation of the Wall Street Journal corpus (Weischedel and Branstetter, 2005) takes the form of SGML inline markup of text, tokenized to be completely compatible with the Penn Treebank annotation.



For the CoNLL-2008 shared task evaluation, this corpus was extended by the task organizers to cover the subset of the Brown corpus used as a secondary testing dataset. From this corpus we only used NE boundaries to derive NAME dependencies between NE tokens, e.g., we create a NAME dependency from *Mary* to *Smith* given the NE mention *Mary Smith*.

- **Proposition Bank I (PropBank)** - The PropBank annotation (Palmer et al., 2005) classifies the arguments of all the main verbs in the Penn Treebank corpus, other than *be*. Arguments are numbered (Arg0, Arg1, ...) based on lexical entries or frame files. Different sets of arguments are assumed for different rolesets. Dependent constituents that fall into categories independent of the lexical entries are classified as various types of adjuncts (ArgM-TMP, -ADV, etc.).
- **NomBank** - NomBank annotation (Meyers et al., 2004) uses essentially the same framework as PropBank to annotate arguments of nouns. Differences between PropBank and NomBank stem from differences between noun and verb argument structure; differences in treatment of nouns and verbs in the Penn Treebank; and differences in the sophistication of previous research about noun and verb argument structure. Only the subset of nouns that take arguments are annotated in NomBank and only a subset of the non-argument siblings of nouns are marked as ArgM.

The complete merging process and the conversion from the constituent representation to dependencies is detailed in (Surdeanu et al., 2008).

The main difference between the 2008 and 2009 version of the corpora is the generation of word lemmas. In the 2008 version the only lemmas provided were predicted using the built-in lemmatizer in WordNet (Fellbaum, 1998) based on the most frequent sense for the form and the predicted part-of-speech tag. These lemmas are listed in the 2009 corpus under the PLEMMA column. The LEMMA column in the 2009 version of the corpus contains lemmas generated using the same algorithm but using the correct Treebank part-of-speech tags. Addi-

tionally, the PHEAD and PDEPREL columns were generated using MaltParser<sup>9</sup>, similarly to the open challenge corpus in the CoNLL 2008 shared task.

### 3.6 German

The German in-domain dataset is based on the annotated verb instances of the SALSA corpus (Burchardt et al., 2006), a total of around 40k sentences<sup>10</sup>. SALSA provides manual semantic role annotation on top of the syntactically annotated TIGER newspaper corpus, one of the standard German treebanks. The original SALSA corpus uses semantic roles in the FrameNet paradigm. We constructed mappings between FrameNet frame elements and PropBank argument positions at the level of frame-predicate pairs semi-automatically. For the frame elements of each frame-predicate pair, we first identified the semantically defined PropBank Arg-0 and Arg-1 positions. To do so, we annotated a small number of very abstract frame elements with these labels (Agent, Actor, Communicator as Arg-0, and Theme, Effect, Message as Arg-1) and percolated these labels through the FrameNet hierarchy, adding further manual labels where necessary. Then, we used frequency and grammatical realization information to map the remaining roles onto higher-numbered Arg roles. We considerably simplified the annotations provided by SALSA, which use a rather complex annotation scheme. In particular, we removed annotation for multi-word expressions (which may be non-contiguous), annotations involving multiple frames for the same predicate (metaphors, underspecification), and inter-sentence roles.

The out-of-domain dataset was taken from a study on the multi-lingual projection of FrameNet annotation (Pado and Lapata, 2005). It is sampled from the EUROPARL corpus and was chosen to maximize the lexical coverage, i.e., it contains of a large number of infrequent predicates. Both syntactic and semantic structure were annotated manually, in the TIGER and SALSA format, respectively. Since it uses a simplified annotation schemes, we have not discarded any annotations.

<sup>9</sup><http://w3.msi.vxu.se/~nivre/research/MaltParser.html>

<sup>10</sup>Note, however, that typically not all predicates in each sentence are annotated.

For both datasets, we converted the syntactic TIGER (Brants et al., 2002) representations into dependencies with a similar set of head-finding rules used for the preparation of the CoNLL-X shared task German dataset. Minor modifications (for the conversion of person names and coordinations) were made to achieve better consistency with datasets of other languages. Since the TIGER annotation allows non-continuous constituents, the resulting dependencies can be non-projective. Secondary edges were discarded in the conversion. As for the automatically constructed features, we used Tree-Tagger (Schmid, 1994) to produce the PLEMMA and PPOS columns, and the Morphisto morphology (Zielinski and Simon, 2008) for PFEAT.

### 3.7 Japanese

For Japanese, we used the Kyoto University Text Corpus (Kawahara et al., 2002), which consists of approximately 40k sentences taken from *Mainichi* Newspapers. Out of them, approximately 5k sentences are annotated with syntactic and semantic dependencies, and are used the training, development and test data of this year’s shared task. The remaining sentences, which are annotated with only syntactic dependencies, are provided for the training corpus of syntactic dependency parsers.

This corpus adopts a dependency structure representation, and thus the conversion to the CoNLL-2009 format was relatively straightforward. However, since the original dependencies are annotated on the basis of phrases (Japanese *bunsetsu*), we needed to automatically convert the original annotations to word-based ones using several criteria. We used the following basic criteria: the words except the last word in a phrase depend on the next (right) word, and the last word in a phrase basically depends on the head word of the governing phrase.

Semantic dependencies are annotated for both verbal predicates and nominal predicates. The semantic roles (APRED columns) consist of 41 surface cases, many of which are case-marking postpositions such as *ga* (nominative), *wo* (accusative) and *ni* (dative). Semantic frame discrimination is not annotated, and so the PRED column is the same as the LEMMA column. The original corpus contains coreference annotations and inter-sentential semantic dependencies, such as inter-sentential zero pro-

nouns and bridging references, but we did not use these annotations, which are not the target of this year’s shared task.

To produce the PLEMMA, PPOS and PFEAT columns, we used the morphological analyzer JUMAN<sup>11</sup> and the dependency and case structure analyzer KNP<sup>12</sup>. To produce the PHEAD and PDEPREL columns, we used the MSTParser<sup>13</sup>.

## 4 Submissions and Results

The official evaluation has been performed centrally. Participants have been uploading the results through the shared task website. Feedback has been provided if any formal problems have been encountered (for a list of checks, see the previous section). One submission had to be rejected because only English results have been provided.

The results have been anonymized and published on the web after the evaluation period had passed.

A total of 21 systems participated in the closed challenge; 14 of them in the Joint task and seven in the SRL-only task. Two systems participated in the Open challenge (Joint task). Moreover, 18 systems provided output in the out-of-domain part of the task (12 in the OOD Joint task and six in the OOD SRL-only task).

The main results for the core task - the Joint task (dependency syntax *and* semantic relations) in the context of the closed challenge - are summarized and ranked in Table 4.

The largest number of systems can be compared in the SRL results table (Table 4), where all the systems have been evaluated solely on the SRL performance regardless whether they participated in the Joint or SRL-only task. However, since the results might have been influenced by the supplied parser, separate ranking is provided for the systems participating in the SRL-only task.

Additional breakdown of the results (open challenge, precision and recall tables for the semantic labeling task, etc.) are available from the CoNLL-

<sup>11</sup> <http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman-e.html>

<sup>12</sup> <http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp-e.html>

<sup>13</sup> <http://sourceforge.net/projects/mstparser/>

Rank	System	Average	Catalan	Chinese	Czech	English	German	Japanese	Spanish
1	Che	82.64	81.84	76.38	<b>83.27</b>	87.00	<b>82.44</b>	<b>85.65</b>	81.90
2	Chen	82.52	<b>83.01</b>	76.23	80.87	<b>87.69</b>	81.22	85.28	<b>83.31</b>
3	Merlo	82.14	82.66	76.15	83.21	86.03	79.59	84.91	82.43
4	Bohnet	80.85	80.44	75.91	79.57	85.14	81.60	82.51	80.75
5	Asahara	78.43	75.91	73.43	81.43	86.40	69.84	84.86	77.12
6	Brown	77.27	77.40	72.12	75.66	83.98	77.86	76.65	77.21
7	Zhang	76.49	75.00	73.42	76.93	82.88	73.76	78.17	75.25
8	Qiu	75.30	70.41	<b>80.66</b>	73.08	80.25	69.87	83.80	69.01
9	Dai	73.98	72.09	72.72	67.14	81.89	75.00	80.89	68.14
10	Lu Li	73.97	71.32	65.53	75.85	81.92	70.93	80.49	71.72
11	Lluís	71.49	56.64	66.18	75.95	81.69	72.31	81.76	65.91
12	Vallejo	70.81	73.75	67.16	60.50	78.19	67.51	77.75	70.78
13	Ren	67.81	59.42	75.90	60.18	77.83	65.77	77.63	57.96
14	Zeman	51.07	49.61	43.50	57.95	50.27	49.57	57.69	48.90

Table 2: Official results of the Joint task, closed challenge. Teams are denoted by the last name (first name added only where needed) of the author who registered for the evaluation data. Results are sorted in descending order of the language-averaged macro  $F_1$  score on the closed challenge Joint task. Bold numbers denote the best result for a given language.

Rank	Rank in task	System	Average	Catalan	Chinese	Czech	English	German	Japanese	Spanish
1	1 (SRLonly)	Zhao	80.47	<b>80.32</b>	77.72	85.19	85.44	75.99	78.15	<b>80.46</b>
2	2 (SRLonly)	Nugues	80.31	80.01	<b>78.60</b>	85.41	85.63	<b>79.71</b>	76.30	76.52
3	1 (Joint)	Chen	79.96	80.10	76.77	82.04	<b>86.15</b>	76.19	78.17	80.29
4	2 (Joint)	Che	79.94	77.10	77.15	<b>86.51</b>	85.51	78.61	<b>78.26</b>	76.47
5	3 (Joint)	Merlo	78.42	77.44	76.05	86.02	83.24	71.78	77.23	77.19
6	3 (SRLonly)	Meza-Ruiz	77.46	78.00	77.73	75.75	83.34	73.52	76.00	77.91
7	4 (Joint)	Bohnet	76.00	74.53	75.29	79.02	80.39	75.72	72.76	74.31
8	5 (Joint)	Asahara	75.65	72.35	74.17	84.69	84.26	63.66	77.93	72.50
9	6 (Joint)	Brown	72.85	72.18	72.43	78.02	80.43	73.40	61.57	71.95
10	7 (Joint)	Qiu	70.87	63.88	77.67	76.66	76.82	61.29	76.62	63.14
11	8 (Joint)	Dai	70.78	66.34	71.57	75.50	78.93	67.43	71.02	64.64
12	9 (Joint)	Zhang	70.31	67.34	73.20	78.28	77.85	62.95	64.71	67.81
13	10 (Joint)	Lu Li	69.72	66.95	67.06	79.08	77.17	61.98	69.58	66.23
14	4 (SRLonly)	Baoli Li	69.26	74.06	70.37	57.46	69.63	67.76	72.03	73.54
15	11 (Joint)	Vallejo	68.95	70.14	66.71	71.49	75.97	61.01	68.82	68.48
16	5 (SRLonly)	Moreau	66.49	65.60	67.37	71.74	72.14	66.50	57.75	64.33
17	12 (Joint)	Lluís	63.06	46.79	59.72	76.90	75.86	62.66	71.60	47.88
18	6 (SRLonly)	Tackstrom	61.27	57.11	63.41	71.05	67.64	53.42	54.74	61.51
19	7 (SRLonly)	Lin	57.18	61.70	70.33	60.43	65.66	59.51	23.78	58.87
20	13 (Joint)	Ren	56.69	41.00	72.58	62.82	67.56	54.31	58.73	39.80
21	14 (Joint)	Zeman	32.14	24.19	34.71	58.13	36.05	16.44	30.13	25.36

Table 3: Official results of the semantic role labeling, closed challenge, all systems. Teams are denoted by the last name (first name added only where needed) of the author who registered for the evaluation data. Results are sorted in descending order of the semantic labeled  $F_1$  score (closed challenge). Bold numbers denote the best result for a given language. Separate ranking is provided for SRL-only systems.

2009 Shared Task website<sup>14</sup>.

## 5 Approaches

Table 5 summarizes the properties of the systems that participated in the closed the open challenges.

The second column of the table highlights the overall architectures. We used + to indicate that the components are sequentially connected. The lack of a + sign indicates that the corresponding tasks are performed jointly.

<sup>14</sup><http://ufal.mff.cuni.cz/con112009-st>

We use the || to indicate that several different

architectures that span multiple subtasks were deployed in parallel.

## 6 Analysis

Section 4 summarized the results in the closed and open challenges using the official evaluation measures. In this section, we analyze the submitted runs using different evaluation measures...

## 7 Conclusion

TBD.

## Acknowledgments

We would like to thank ...

Last but not least, we thank the organizers of the previous four shared tasks: Sabine Buchholz, Xavier Carreras, Ryan McDonald, Amit Dubey, Johan Hall, Yuval Krymolowski, Sandra Kübler, Erwin Marsi, Jens Nilsson, Sebastian Riedel and Deniz Yuret. This shared task would not have been possible without their previous effort.

We also acknowledge the support of the MŠMT of the Czech Republic, projects MSM0021620838 and LC536; the Grant Agency of the Academy of sciences of the Czech Republic 1ET201120505 (for Jan Hajič, Jan Štěpánek and Pavel Straňák).

Lluís Màrquez and M. Antònia Martí participation was supported by the Spanish Ministry of Education and Science, through the OpenMT and TextMess research projects (TIN2006-15307-C03-02, TIN2006-15265-C06-06).

The following individuals directly contributed to the Chinese Treebank (in alphabetic order): Meiyu Chang, Fu-Dong Chiou, Shizhe Huang, Zixin Jiang, Tony Kroch, Martha Palmer, Mitch Marcus, Fei Xia, Nianwen Xue. The contributors to the Chinese Proposition Bank include (in alphabetic order): Meiyu Chang, Gang Chen, Helen Chen, Zixin Jiang, Martha Palmer, Zhiyi Song, Nianwen Xue, Ping Yu, Hua Zhong. The Chinese Treebank and the Chinese Proposition Bank were funded by DOD, NSF and DARPA.

We thank the Mainichi Newspapers for the permission of distributing the sentences of the Kyoto University Text Corpus for this shared task.

## References

- [Brants et al.2002] Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol.
- [Burchardt et al.2006] Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2006. The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy.
- [Civit et al.2006] Montserrat Civit, M. Antònia Martí, and Núria Buñ. 2006. Cat3LB and Cast3LB: from constituents to dependencies. In *Proceedings of the 5th International Conference on Natural Language Processing, FinTAL*, pages 141–153, Turku, Finland. Springer Verlag, LNAI 4139.
- [Fellbaum1998] Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge.
- [Hajič et al.2003] Jan Hajič, Jarmila Panevová, Zdeňka Uřešová, Alevtina Bémová, Veronika Kolářová-Řezníčková, and Petr Pajas. 2003. PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In J. Nivre and E. Hinrichs, editors, *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, pages 57–68, Vaxjo, Sweden. Vaxjo University Press.
- [Hajič et al.2006] Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, and Zdeněk Žabokrtský. 2006. Prague Dependency Treebank 2.0.
- [Kawahara et al.2002] Daisuke Kawahara, Sadao Kurohashi, and Kôiti Hasida. 2002. Construction of a Japanese relevance-tagged corpus. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pages 2008–2013, Las Palmas, Canary Islands.
- [Marcus et al.1994] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1994. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.
- [Màrquez et al.2007] Lluís Màrquez, Luis Villarejo, M. Antònia Martí, and Mariona Taulé. 2007. SemEval-2007 Task 09: Multilevel semantic annotation of catalan and spanish. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 42–47, Prague, Czech Republic.
- [Martí et al.2007] M. Antònia Martí, Mariona Taulé, Lluís Màrquez, and Manu Bertran. 2007. Anotación semiautomática con papeles temáticos de los corpus



- CESS-ECE. *Procesamiento del Lenguaje Natural, SEPLN Journal*, 38:67–76.
- [McDonald et al.2005] Ryan McDonald, Fernando Pereira, Jan Hajič, and Kiril Ribarov. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of NAACL-HLT'05, Vancouver, Canada*, pages 523–530.
- [Meyers et al.2004] A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. The NomBank Project: An Interim Report. In *NAACL/HLT 2004 Workshop Frontiers in Corpus Annotation*, Boston.
- [Pado and Lapata2005] Sebastian Pado and Mirella Lapata. 2005. Cross-lingual projection of role-semantic information. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-2005)*, pages 859–866, Vancouver, BC.
- [Palmer et al.2005] M. Palmer, D. Gildea, and P. Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.
- [Schmid1994] Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.
- [Spoustová et al.2009] Drahomíra "Johanka" Spoustová, Jan Hajič, Jan Raab, and Miroslav Spousta. 2009. Semi-supervised training for the averaged perceptron POS tagger. In *Proceedings of the European ACL Conference EACL'09*, Athens, Greece.
- [Surdeanu et al.2008] Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the 12th Conference on Computational Natural Language Learning (CoNLL-2008)*, pages 159–177.
- [Taulé et al.2008] Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-2008)*, Marrakesh, Morocco.
- [Čmejrek et al.2004] Martin Čmejrek, Jan Cuřín, Jan Hajič, Jiří Havelka, and Vladislav Kuboň. 2004. Prague Czech-English Dependency Treebank: Syntactically Annotated Resources for Machine Translation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*, pages 1597–1600, Lisbon, Portugal.
- [W. N. Francis and H. Kucera1964] W. N. Francis and H. Kucera. 1964. Brown Corpus Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. Revised 1971, Revised and Amplified 1979, available at [www.clarinet/brown](http://www.clarinet/brown).
- [Weischedel and Brunstein2005] R. Weischedel and A. Brunstein. 2005. BBN pronoun coreference and entity type corpus. Technical report, Linguistic Data Consortium.
- [Xue and Palmer2009] Nianwen Xue and Martha Palmer. 2009. Adding semantic roles to the Chinese Treebank. *Natural Language Engineering*, 15(1):143–172.
- [Xue et al.2005] Nianwen Xue, Fei Xia, Fu Dong Chiou, and Martha Palmer. 2005. The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*, 11(2):207–238.
- [Zielinski and Simon2008] Andrea Zielinski and Christian Simon. 2008. Morphisto: An open-source morphological analyzer for german. In *Proceedings of the Conference on Finite State Methods in Natural Language Processing*.