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
‡: 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 and Växjö University, joakim.nivre@lingfil.uu.se
◇⋆: Stuttgart University, pado@ims.uni-stuttgart.de
‡⋆: Stanford University, mihais@stanford.edu
‡‡: Brandeis University, xuen@brandeis.edu
‡◇: Saarland University, yzhang@coli.uni-sb.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 and show their quantitative properties, report 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 (Surdeanu et al., 2008) used a unified dependency-based formalism, which modeled both syntactic dependencies and semantic roles for English. The CoNLL-2009 Shared Task has built on the 2008 results by providing data for six more languages (Catalan, Chinese, Czech, German, Japanese and Spanish) in addition to the original English[1]. It has thus naturally extended the path taken by 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).

[1]There are some format changes and deviations from the 2008 task data specification; see Sect. 2.3

Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL): Shared Task, pages 1–18,
Boulder, Colorado, June 2009.

In contrast to the previous year, the evaluation data indicated which words were to be dealt with (for the SRL task). In other words, (predicate) disambiguation was still part of the task, whereas the *identification* of argument-bearing 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 had to be trained strictly with information contained in the given training corpus; in the open challenge, systems could have been 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. A substantial portion of the paper (Section 3) is devoted to the description of the conversion and development of the data sets in the additional languages. Section 4 shows the main results of the submitted systems in the Joint and SRL-only tasks. Section 5 summarizes the approaches implemented by participants. Section 6 concludes the paper. In all sections, we will mention some of the differences between last year's and this year's tasks while keeping the text self-contained whenever possible; for details and observations on the English data, please refer to the overview paper of the CoNLL-2008 Shared Task (Surdeanu et al., 2008) and to the references mentioned in the sections describing the other 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 was to compare performance of the participating systems in a fair environment. Systems had 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.) were provided and may have been used. These restrictions mean that outside parsers (not trained by the participants' systems) could not be used. However, we did provide the output of a single, state-of-the-art dependency parser for each language so that participants could 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 could have been developed making use of any kind of external tools and resources. The only condition was 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 were 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 an 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 while additional measures

are computed for various combinations of parsers and SRL methods across the tasks and challenges.

2.3 Data Format

The data format used in this shared task has been based on the CoNLL-2008 shared task, with some differences. The data follows these general rules:

- 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).

For the SRL-only task, participants have been provided will 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 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 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 APREDn 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 APREDn 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"[2] lemmas and forms have been provided in the English datasets (for the other languages, original tokenization from the respective treebanks has been used);
- rich morphological features have been added wherever available;
- syntactic dependencies by state-of-the-art parsers have been provided (for the SRL-only task);
- multiple semantic labels for a single token have been allowed (and properly evaluated) in the APREDs columns;
- predicates have been pre-identified and marked in both the training and test data;
- some of the fields (e.g. the `APREDx`) and values (`ARG0` → `A0` etc.) have been renamed.

2.4 Evaluation Measures

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

The main evaluation measure, according to which systems are primarily compared, is the Joint task,

[2] 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	Automatically 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...	APREDn	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.

closed challenge, Macro F_1 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 preparsed 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. Additionally, 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: A0, A1, AM-TMP
```

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

```
verb.02: A0, A1, AM-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 `AM-TMP` is incorrectly labeled `AM-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; for example, this happens in Czech in special cases of reciprocity 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 computation 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:[3]

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

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

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.[4] 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] 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.

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.

3.1 The Input Corpora

The data used as the input for the transformations 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. A visual representation of the Trial data was also created to make understanding of the data easier. Any data in the same format can be transformed and displayed in the Tree Editor TrEd[5] (Pajas and Štěpánek, 2008) with the CoNLL 2009 Shared Task extension that can be installed from within the editor. A sample visualization of an English sentence after its conversion to the shared task format (Sect. 2.3) is in Fig. 1.

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 down-

[4] We assign equal weight to the two tasks, i.e., $W_{sem} = 0.5$.

[5] `http://ufal.mff.cuni.cz/~pajas/tred`

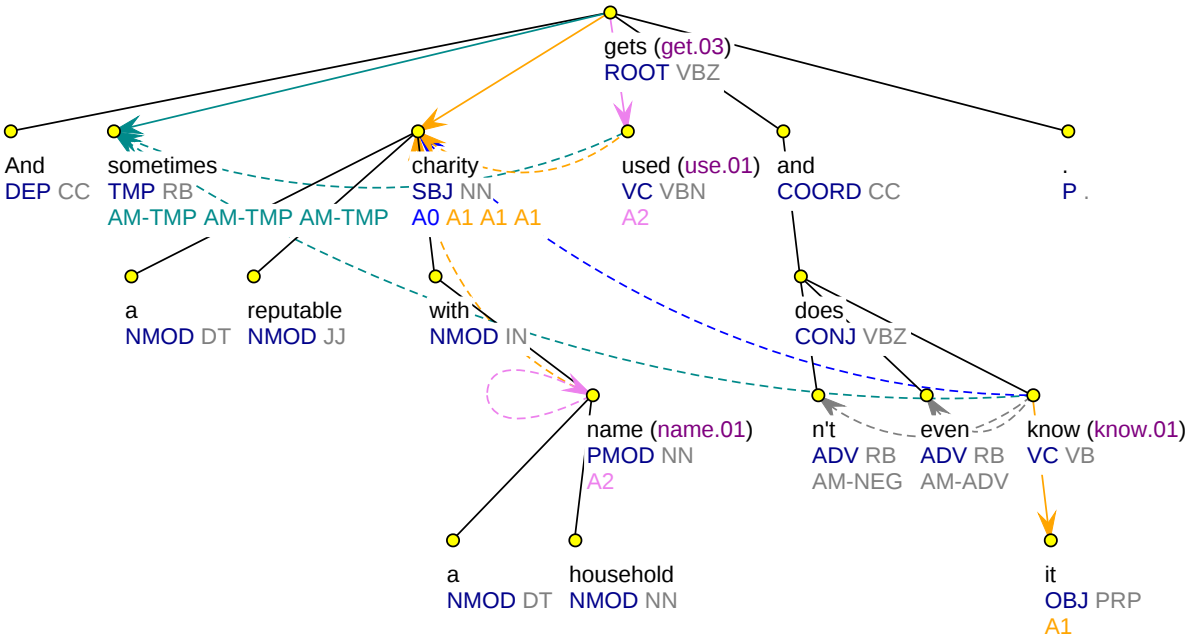

Figure 1: Visualisation of the English sentence *"And sometimes a reputable charity with a houshold name gets used and doesn't even know it."* (Penn Treebank, wsj_0559) showing jointly the labeled syntactic and semantic dependencies. The basic tree shape comes from the syntactic dependencies; syntactic labels and POS tags are on the 2^{nd} line at each node. Semantic dependencies which do not follow the syntactic ones use dotted lines. Predicate senses in parentheses (`use:01`, ...) follow the word label. SRLs (`A0`, `AM-TMP`, ...) are on the last line. Please note that multiple semantic dependencies (e.g., there are four for *charity*: A0 ← *know*, A1 ← *gets*, A1 ← *used*, A1 ← *name*) and self-dependencies (*name*) appear in this sentence.

load, 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 packages - one for the Joint task and one for the SRL-only task. Every participant had to subscribe to one of the two tasks; subsequently, they 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.

Some statistics on the data can be seen in Tables 2, 3 and 4. Whereas the training sizes of the data have not been that different as they were e.g. for the 2007 shared task on multilingual dependency parsing (Nivre et al., 2007)[6], substantial differences existed in the distribution of the predicates and arguments, the input features, the out-of-vocabulary rates, and other statistical characteristics of the data.

Data sizes have been relatively uniform in all the datasets, with Japanese having the smallest dataset

[6] `http://nextens.uvt.nl/depparse-wiki/DataOverview`

containing data for SRL annotation training. To compensate at least for the dependency parsing part, an additional, large Japanese corpus with syntactic dependency annotation has been provided.

The average sentence length, the vocabulary sizes for FORM and LEMMA fields and the OOV rates characterize quite naturally the properties of the respective languages (in the domain of the training and evaluation data). It is no surprise that the FORM OOV rate is the highest for Czech, a highly inflectional language, and that the LEMMA OOV rate is the highest for German (as a consequence of keeping compounds as a single lemma). The other statistics also reflect (to a large extent) the annotation specification and conventions used for the original treebanks and/or the result of the conversion process to the unified CoNLL-2009 Shared Task format.

Starting with the POS and FEAT fields, it can be seen that Catalan, Czech and Spanish use only the 12 major part-of-speech categories as values of the POS field (with richly populated FEAT field); English and Chinese are the opposite extreme, disregarding the use of the FEAT field completely and coding everything as a POS value. While for Chinese this is quite understandable, English follows the PTB tradition in this respect. German and Japanese use relatively rich set of values in both the POS and FEAT fields.

For the dependency relations (DEPREL), all the languages use a similarly-sized set except for Japanese, which only encodes the distinction between a root and a dependent node (and some infrequent special ones).

Evaluation data are over 10% of the size of the training data for Catalan, Chinese, Czech, Japanese and Spanish and roughly 5% for English and German.

Table 3 shows the distribution of the five most frequent dependency relations (determined as part of the subtask of syntactic parsing). With the exception of Japanese, which essentially does not label dependency relations at this level, all the other languages show little difference in this distribution. For example, the unconditioned probability of "subjects" is almost the same for all the six other languages (between 6 and 8 percent). The probability mass covered by the first five most frequent DEPRELs is also almost the same (again, except for Japanese), suggesting that the labeling task might have similar difficulty[7]. The most skewed one is for Czech (after Japanese).

Table 4 shows similar statistics for the argument labels (PRED/APREDs); it also adds the average number of arguments per "predicate" token, since this is part of the SRL task[8]. It is apparent from the comparison of the "Total" rows in this table and Table 3 that the first five argument labels cover more that their syntactic counterparts. For example, the arguments A0-A4 account for all but 3% of all arguments labels, whereas Spanish and Catalan have much more rich set of argument labels, with a high entropy of the most-frequent-label distribution.

3.2 Catalan and Spanish

The Catalan and Spanish datasets (Taulé et al., 2008) were generated from the AnCora corpora[9] 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,

[7] Yes, this is overgeneralization since this distribution does not condition on the features, dependencies etc. But as a rough measure, it often correlates well with the results.

[8] A number below 1 means there are some argument-bearing words (often nouns) which have no arguments in the particular sentence in which they appear.

[9] `http://clic.ub.edu/ancora`

Characteristic	Catalan	Chinese	Czech	English	German	Japanese	Spanish
Training data size (sentences)	13200	22277	38727	39279	36020	4393[a]	14329
Training data size (tokens)	390302	609060	652544	958167	648677	112555[a]	427442
Avg. sentence length (tokens)	29.6	27.3	16.8	24.4	18.0	25.6	29.8
Tokens with arguments[b] (%)	9.6	16.9	63.5	18.7	2.7	22.8	10.3
DEPREL types	50	41	49	69	46	5	49
POS types	12	41	12	48	56	40	12
FEAT types	237	1	1811	1	267	302	264
FORM vocabulary size	33890	40878	86332	39782	72084	36043	40964
LEMMA vocabulary size	24143	40878	37580	28376	51993	30402	26926
Evaluation data size (sent.)	1862	2556	4213	2399	2000	500	1725
Evaluation data size (tokens)	53355	73153	70348	57676	31622	13615	50630
Evaluation FORM OOV[c]	5.40	3.92	7.98/8.62[d]	1.58/3.76[d]	7.93/7.57[d]	6.07	5.63
Evaluation LEMMA OOV[c]	4.14	3.92	3.03/4.29[d]	1.08/2.30[d]	5.83/7.36[d]	5.21	3.69

Table 2: Elementary data statistics for the CoNLL-2009 Shared Task languages. The data themselves, the original treebanks they were derived from and the conversion process are described in more detail in sections 3.2-3.7. All evaluation data statistics are derived from the in-domain evaluation data.

[a]There were additional 33257 sentences (839947 tokens) available for syntactic dependency parsing of Japanese; the type and vocabulary statistics are computed using this larger dataset.

[b]Percentage of tokens with FILLPRED='Y'.

[c]Percentage of FORM/LEMMA tokens not found in the respective vocabularies derived solely from the training data.

[d]OOV percentage for in-domain/out-of-domain data.

DEPREL	Catalan	Chinese	Czech	English	German	Japanese	Spanish
Labels	sn 0.16	COMP 0.21	Atr 0.26	NMOD 0.27	NK 0.31	D 0.93	sn 0.16
	spec 0.15	NMOD 0.14	AuxP 0.10	P 0.11	PUNC 0.14	ROOT 0.04	spec 0.15
	f 0.11	ADV 0.10	Adv 0.10	PMOD 0.10	MO 0.12	P 0.03	f 0.12
	sp 0.09	UNK 0.09	Obj 0.07	SBJ 0.07	SB 0.07	A 0.00	sp 0.08
	suj 0.07	SBJ 0.08	Sb 0.06	OBJ 0.06	ROOT 0.06	I 0.00	suj 0.08
Total	0.58	0.62	0.59	0.61	0.70	1.00	0.59

Table 3: Unigram probability for the five most frequent DEPREL labels in the training data of the CoNLL-2009 Shared Task is shown. Total is the probability mass covered by the five dependency labels shown.

APRED	Catalan	Chinese	Czech	English	German	Japanese	Spanish
Labels	arg1-pat 0.22	A1 0.30	RSTR 0.30	A1 0.37	A0 0.40	GA 0.33	arg1-pat 0.20
	arg0-agt 0.18	A0 0.27	PAT 0.18	A0 0.25	A1 0.39	WO 0.15	arg0-agt 0.19
	arg1-tem 0.15	ADV 0.20	ACT 0.17	A2 0.12	A2 0.12	NO 0.15	arg1-tem 0.15
	argM-tmp 0.08	TMP 0.07	APP 0.06	AM-TMP 0.06	A3 0.06	NI 0.09	arg2-atr 0.08
	arg2-atr 0.08	DIS 0.04	LOC 0.04	AM-MNR 0.03	A4 0.01	DE 0.06	argM-tmp 0.08
Total	0.71	0.91	0.75	0.83	0.97	0.78	0.70
Avg.	2.25	2.26	0.88	2.20	1.97	1.71	2.26

Table 4: Unigram probability for the five most frequent APRED labels in the training data of the CoNLL-2009 Shared Task is shown. Total is the probability mass covered by the five argument labels shown. The "Avg." line shows the average number of arguments per predicate or other argument-bearing token (i.e. for those marked by FILLPRED='Y').

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; 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[10]. Accuracy in PLEMMA and PPOS columns is above 95% for the two languages. PHEAD and PDEPREL columns have been generated using MaltParser[11]. 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

[10] `http://www.lsi.upc.es/~nlp/freeling`
[11] `http://w3.msi.vxu.se/~jha/maltparser`

Boulder. Now it is the process of being to 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 [12]. 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.

[12] A small number of files were taken out of the CoNLL shared task data due to conversion problems and time constraints to fix them.

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[13], see also (Čmejrek et al., 2004). For the OOD data, no manual annotation of LEMMA, POS, and FEAT existed, so the predicted values were used. 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.

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 (see below).

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

The FEAT column was created from the remaining characters of the `tag` element. In addition, the special feature "Sem" corresponds to a semantic feature of the lemma.

For the HEAD and DEPREL columns, the PDT analytical layer was used. The DEPREL was taken from the analytic function (the `afun` node attribtue). There are 27 possible values for `afun` element: `Pred`, `Pnom`, `AuxV`, `Sb`, `Obj`, `Atr`, `Adv`, `Atv`, `AtvV`, `Coord`, `Apos`, `ExD`, and a number of auxiliary and "double-function" labels. The first nine of these are the "most interesting" from the point of view of the shared task, since they relate to semantics more closely than the rest (at least from the linguistic point of view). The HEAD is a pointer to its parent, which means the PDT's `ord` attribute

[13] `http://ufal.mff.cuni.cz/pedt`

(within-sentence ID / word position number) of the parent. If a node is a member of a coordination or apposition (`is_member` element), its DEPREL obtains the `_M` suffix. The parenthesis annotation (`is_parenthesis_root` element) was ignored.

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, separated by a vertical bar (see Sect. 2.4.2).

PLEMMA, PPOS and PFEAT were generated by the (cross-trained) morphological tagger MORCE (Spoustová et al., 2009), which gives full combined accuracy (PLEMMA+PPOS+PFEAT) slightly under 96%.

PHEAD and PDEPREL were generated by the (cross-trained) MST parser for Czech (Chu–Liu/Edmonds algorithm, (McDonald et al., 2005)), which has typical dependency accuracy around 85%.

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)

The source of the out-of-domain data uses an extended valency lexicon (because of out-of-vocabulary entries). For simplicity, the extended lexicon was not provided; instead, such words were not marked as predicates in the OOD data (their FILLPRED was set to '_') and thus not evaluated.

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 Brunstein, 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. Additionally, the PHEAD and PDEPREL columns were generated using MaltParser[14], similarly to the open challenge corpus in the CoNLL 2008 shared task.

[14] `http://w3.msi.vxu.se/~nivre/research/MaltParser.html`

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[15]. 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 did not have to discard any annotation.

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 con-

[15] Note, however, that typically not all predicates in each sentence are annotated (cf. Table 2).

version of person names and coordinations) were made to achieve better consistency with datasets of other languages. Since the TIGER annotation allows non-contiguous constituents, the resulting dependencies can be non-projective. Secondary edges were discarded in the conversion. As for the automatically constructed features, we used TreeTagger (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 pronouns 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 [16] and the dependency and case structure analyzer KNP [17]. To produce the PHEAD and PDEPREL columns, we used the MSTParser [18].

4 Submissions and Results

Participants uploaded the results through the shared task website, and the official evaluation was performed centrally. Feedback was provided if any formal problems were encountered (for a list of checks, see the previous section). One submission had to be rejected because only English results were provided. After the evaluation period had passed, the results were anonymized and published on the web.

A total of 20 systems participated in the closed challenge; 13 of them in the Joint task and seven in the SRL-only task. Two systems participated in the open challenge (Joint task). Moreover, 17 systems provided output in the out-of-domain part of the task (11 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 5.

The largest number of systems can be compared in the SRL results table (Table 6), 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 both types of the systems.

Additional breakdown of the results (open challenge, precision and recall tables for the semantic labeling task, etc.) are available from the CoNLL-2009 Shared Task website[19].

5 Approaches

Table 7 summarizes the properties of the systems that participated in the closed the open challenges.

[16] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman-e.html

[17] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp-e.html

[18] http://sourceforge.net/projects/mstparser

[19] http://ufal.mff.cuni.cz/conll2009-st

Rank	System	Average	Catalan	Chinese	Czech	English	German	Japanese	Spanish
1	Che	82.64	81.84	76.38	**83.27**	87.00	**82.44**	**85.65**	81.90
2	Chen	82.52	**83.01**	76.23	80.87	**87.69**	81.22	85.28	**83.31**
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	Dai	73.98	72.09	72.72	67.14	81.89	75.00	80.89	68.14
9	Lu Li	73.97	71.32	65.53	75.85	81.92	70.93	80.49	71.72
10	Lluís	71.49	56.64	66.18	75.95	81.69	72.31	81.76	65.91
11	Vallejo	70.81	73.75	67.16	60.50	78.19	67.51	77.75	70.78
12	Ren	67.81	59.42	75.90	60.18	77.83	65.77	77.63	57.96
13	Zeman	51.07	49.61	43.50	57.95	50.27	49.57	57.69	48.90

Table 5: 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	**80.32**	77.72	85.19	85.44	75.99	78.15	**80.46**
2	2 (SRLonly)	Nugues	80.31	80.01	**78.60**	85.41	85.63	**79.71**	76.30	76.52
3	1 (Joint)	Chen	79.96	80.10	76.77	82.04	**86.15**	76.19	78.17	80.29
4	2 (Joint)	Che	79.94	77.10	77.15	**86.51**	85.51	78.61	**78.26**	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)	Dai	70.78	66.34	71.57	75.50	78.93	67.43	71.02	64.64
11	8 (Joint)	Zhang	70.31	67.34	73.20	78.28	77.85	62.95	64.71	67.81
12	9 (Joint)	Lu Li	69.72	66.95	67.06	79.08	77.17	61.98	69.58	66.23
13	4 (SRLonly)	Baoli Li	69.26	74.06	70.37	57.46	69.63	67.76	72.03	73.54
14	10 (Joint)	Vallejo	68.95	70.14	66.71	71.49	75.97	61.01	68.82	68.48
15	5 (SRLonly)	Moreau	66.49	65.60	67.37	71.74	72.14	66.50	57.75	64.33
16	11 (Joint)	Lluís	63.06	46.79	59.72	76.90	75.86	62.66	71.60	47.88
17	6 (SRLonly)	Täckström	61.27	57.11	63.41	71.05	67.64	53.42	54.74	61.51
18	7 (SRLonly)	Lin	57.18	61.70	70.33	60.43	65.66	59.51	23.78	58.87
19	12 (Joint)	Ren	56.69	41.00	72.58	62.82	67.56	54.31	58.73	39.80
20	13 (Joint)	Zeman	32.14	24.19	34.71	58.13	36.05	16.44	30.13	25.36

Table 6: Official results of the semantic 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.

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.

It is perhaps not surprising that most of the observations from the 2008 shared task still hold; namely, the best systems overall do not use joint learning or optimization (the best such system was placed third in the Joint task, and there were only four systems where the learning methodology can be considered "joint").

Therefore, most of the observations and conclusions from 2008 shared task hold as well for the current results. For details, we will leave it to the reader to interpret the architectures and methods

System[a]	Overall Arch.[b]	D Arch.	D Comb.	D Inference[c]	PA Arch.	PA Comb.	PA Inference	Joint Learning/Opt.	ML Methods
Zhao	PAIC	(SRL-only)	(SRL-only)	(SRL-only)	class	no	greedy/global search	(SRL-only)	ME
Nugues	(PC+AI+AC) + AIC	(SRL-only)	(SRL-only)	(SRL-only)	class	no	beam search + reranking	(SRL-only)	L2-regularized lin. regression
Chen	P + PC + AI + AC	graph	partially	$MST^{CL/E}$	class	no	greedy (?)	no	ME
Che	D+PC+AIC	graph	no	MST^{HOE}	class	no	ILP	no	SVM, ME
Merlo	DPAIC+D	generative, trans	no	beam search	trans	no	beam search	synchronized derivation	ISBN
Meza-Ruiz	PAIC	(SRL-only)	(SRL-only)	(SRL-only)	Markov LN	no	Cutting Plane	(SRL-only)	MIRA
Bohnet	D + AI + AC + PC	graph	no	MST^{C}+rearrange	class	no	greedy	no	SVM (MIRA)
Asahara	D + PIC + AIC	graph	no	MST^{C}	class	no	n-best relax.	no	perceptron
Dai	D + PC + AC	graph	no	MST^{C}	class	no	prob	iterative	ME
Zhang	D + AI + AC + PC	graph	no	MST^{E}	class	no	classification	no	MIRA, ME
Lu Li	D + (PC || AIC)	graph	for each lang.	$MST^{CL/E}$,MST^{E}	class	no	greedy	no	ME
Baoli Li	PC + AIC	(SRL-only)	(SRL-only)	(SRL-only)	class	no	greedy	(SRL-only)	SVM, kNN, ME
Vallejo[d]	[D+P+A]C + DI	class	no	reranking	class	no	reranking	unified labels	MBL
Moreau	D + PI + Clustering + AI + AC	(SRL-only)	(SRL-only)	(SRL-only)	class	no	CRF	(SRL-only)	CRF
Lluís	D+DAIC+PC	graph	no	MST^{E}	graph	no	MST^{E}	yes, MST^{E}	Avg. Perceptron
Täckström	D + PI + AI + AC + Constraint Satisfaction	(SRL-only)	(SRL-only)	(SRL-only)	class	no	greedy	(SRL-only)	SVM
Ren	D + PC + AIC	trans	no	greedy	class	no	greedy	no	SVM (Malt), ME
Zeman	DI+DC+PC+AI+AC	trans	no	greedy with heuristics	class	no	greedy	no	cooccurrence

Table 7: Summary of system architectures for the CoNLL-2009 shared task; all systems are included. SRL-only systems do not have the D columns and the Joint Learing/Opt. columns filled in. The systems are sorted by the semantic labeled F_1 score averaged over all the languages (same as in Table 6). Only the systems that have a corresponding paper in the proceedings are included. Acronyms used: **D** - syntactic dependencies, **P** - predicate, **A** - argument, I - identification, **C** - classification. **Overall arch.** stands for the complete system architecture; **D Arch.** stands for the architecture of the syntactic parser; **D Comb.** indicates if the final parser output was generated using parser combination; **D Inference** stands for the type of inference used for syntactic parsing; **PA Arch.** stands the type of architecture used for PAIC; **PA Comb.** indicates if the PA output was generated through system combination; **PA Inference** stands for the the type of inference used for PAIC; **Joint Learning/Opt.** indicates if some form of joint learning or optimization was implemented for the syntactic + semantic global task; **ML Methods** lists the ML methods used throughout the complete system.

[a] Authors of two systems: "Brown" and "Lin" didn't submit a paper, so their systems' architectures are unknown.

[b] The symbol + indicates sequential processing (otherwise, parallel/joint). The || means that several different architectures spanning multiple subtasks ran in parallel.

[c] $MST^{CL/E}$ as used by McDonald (2005), MST^{C} by Carreras (2007), MST^{E} by Eisner (2000), MST^{HOE} = MST^{E} with higher-order features (siblings + all grandchildren).

[d] The system unifies the syntactic and semantic labels into one label, and trains classifiers over them. It is thus difficult to split the system characteristic into a "D"/"PA" part.

when comparing Table 7 with the Tables 5 and 6).

6 Conclusion

This year's task has been demanding in several respects, but certainly the most difficulty came from the fact that participants had to tackle all seven languages. It is encouraging that despite this added affort the number of participating systems has been almost the same as last year (20 vs. 22 in 2008).

There are several positive outcomes from this year's enterprise:

- we have prepared a unified format and data for several very different lanaguages, as a basis for possible extensions towards other languages and unified treatment of syntactic depenndecies and semantic role labeling across natural languages;
- 20 participants have produced SRL results for all seven languages, using several different methods, giving hope for a combined system with even substantially better performance;
- initial results have been provided for three languages on out-of-domain data (being in fact quite close to the in-domain results).

Only four systems tried to apply what can be described as joint learning for the syntactic and semantic parts of the task. (Morante et al., 2009) use a true joint learning formulation that phrases syntactico-semantic parsing as a series of classification where the class labels are concatenations of syntactic and semantic edge labels. They predict (a), the set of syntactico-semantic edge labels for each pair of tokens; (b), the set of incoming syntactico-semantic edge labels for each individual token; and (c), the existence of an edge between each pair of tokens. Subsequently, they combine the (possibly conflicting) output of the three classifiers by a ranking approach to determine the most likely structure that meets all well-formedness constraints. (Lluís et al., 2009) present a joint approach based on an extension of Eisner's parser to accommodate also semantic dependency labels. This architecture is similar to the one presented by the same authors in the past edition, with the extension to a second-order syntactic parsing and a particular setting for Catalan and Spanish. (Gesmundo et al., 2009) use an incremental parsing model with synchronous syntactic and semantic derivations and a joint probability model for syntactic and semantic dependency structures. The system uses a single input queue but two separate stacks and synchronizes syntactic and semantic derivations at every word. The synchronous derivations are modeled with an Incremental Sigmoid Belief Network that has latent variables for both syntactic and semantic states and connections from syntax to semantics and vice versa. (Dai et al., 2009) designed an iterative system to exploit the inter-connections between the different subtasks of the CoNLL shared task. The idea is to decompose the joint learning problem into four subtasks – syntactic dependency identification, syntactic dependency labeling, semantic dependency identification and semantic dependency labeling. The initial step is to use a pipeline approach to use the input of one subtask as input to the next, in the order specified. The iterative steps then use additional features that are not available in the initial step to improve the accuracy of the overall system. For example, in the iterative steps, semantic information becomes available as features to syntactic parsing, so on and so forth.

Despite these results, it is still not clear whether joint learning has a significant advantage over other approaches (and if yes, then for what languages). It is thus necessary to carefully plan the next shared tasks; it might be advantageous to bring up a similar task in the future once again, and/or couple it with selected application(s). There, (we hope) the benefits of the dependency representation combined with semantic roles the way we have formulated it in 2008 and 2009 will really show up.

Acknowledgments

We would like to thank the Linguistic Data Consortium, mainly to Denise DiPersio, Tony Casteletto and Christopher Cieri for their help and handling of invoicing and distribution of the data for which LDC has a license. For all of the trial, training and evaluation data they had to act a very short notice. All the data has been at the participants' disposal (again) free of charge. We are grateful to all of them for LDC's continuing support of the CoNLL Shared

Tasks.

We would also like to thank 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.

Adam Meyers' work on the shared task has been supported by the NSF Grant IIS-0534700 "Structure Alignment-based MT."

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

References

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.

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.

Xavier Carreras. 2007. Experiments with a higher-order projective dependency parser. In *Proceedings of EMNLP-CoNLL 2007*, pages 957–961, June. Prague, Czech Republic.

Montserrat Civit, M. Antònia Martí, and Núria Bufí. 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.

Qifeng Dai, Enhong Chen, and Liu Shi. 2009. An iterative approach for joint dependency parsing and semantic role labeling. In *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL-2009), June 4-5*, Boulder, Colorado, USA. June 4-5.

Jason Eisner. 2000. Bilexical grammars and their cubic-time parsing algorithms. In Harry Bunt and Anton Nijholt, editors, *Advances in Probabilistic and Other Parsing Tehcnologies*, pages 29–62. Kluwer Academic Publishers.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge.

Andrea Gesmundo, James Henderson, Paola Merlo, and Ivan Titov. 2009. A latent variable model of synchronous syntactic-semantic parsing for multiple languages. In *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL-2009), June 4-5*, Boulder, Colorado, USA. June 4-5.

Jan Hajič, Jarmila Panevová, Zdeňka Urešová, 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.

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.

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.

Xavier Lluís, Stefan Bott, and Lluís Màrquez. 2009. A second-order joint eisner model for syntactic and semantic dependency parsing. In *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL-2009), June 4-5*, Boulder, Colorado, USA. June 4-5.

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.

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.

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.

Ryan McDonald, Fernando Pereira, Jan Hajič, and Kiril Ribarov. 2005. Non-projective dependency parsing using spanning tree algortihms. In *Proceedings of NAACL-HLT'05, Vancouver, Canada*, pages 523–530.

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.

Roser Morante, Vincent Van Asch, and Antal van den Bosch. 2009. A simple generative pipeline approach to dependency parsing and semantic role labeling. In *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL-2009)*, Boulder, Colorado, USA. June 4-5.

Joakim Nivre, Johann Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The conll 2007 shared task on dependency parsing. In *Proceedings of the EMNLP-CoNLL 2007 Conference*, pages 915–932, Prague, Czech Republic.

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.

Petr Pajas and Jan Štěpánek. 2008. Recent advances in a feature-rich framework for treebank annotation. In *The 22nd International Conference on Computational Linguistics - Proceedings of the Conference (COLING'08)*, pages 673–680, Manchester.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

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 Cenference EACL'09*, Athens, Greece.

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.

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, Morroco.

Martin Čmejrek, Jan Cuřín, Jan Hajič, Jiří Havelka, and Vladislav Kuboň. 2004. Prague Czech-English Dependency Treebank: Syntactically Anntoated 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. 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.

R. Weischedel and A. Brunstein. 2005. BBN pronoun coreference and entity type corpus. Technical report, Lin- guistic Data Consortium.

Nianwen Xue and Martha Palmer. 2009. Adding semantic roles to the Chinese Treebank. *Natural Language Engineering*, 15(1):143–172.

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.

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*.