The PASCAL Challenge on Grammar Induction

Douwe Gelling and **Trevor Cohn**
Department of Computer Science
University of Sheffield, UK
{d.gelling,t.cohn}@sheffield.ac.uk

Phil Blunsom
Department of Computer Science
University of Oxford, UK
Phil.Blunsom@cs.ox.ac.uk

João Graça
L^{2}F Spoken Language Systems Laboratory
INESC ID Lisboa, Portugal
joao.graca@l2f.inesc-id.pt

Abstract

This paper presents the results of the PASCAL Challenge on Grammar Induction, a competition in which competitors sought to predict part-of-speech and dependency syntax from text. Although many previous competitions have featured dependency grammars or parts-of-speech, these were invariably framed as supervised learning and/or domain adaption. This is the first challenge to evaluate unsupervised induction systems, a sub-field of syntax which is rapidly becoming very popular. Our challenge made use of a 10 different treebanks annotated in a range of different linguistic formalisms and covering 9 languages. We provide an overview of the approaches taken by the participants, and evaluate their results on each dataset using a range of different evaluation metrics.

1 Introduction

Inducing grammatical structure from text has long been fundamental problem in Computational Linguistics and Natural Language Processing. In recent years interest has grown, spurred by advances in unsupervised statistical modelling and machine learning. The task has relevance to cognitive scientists and linguists attempting to gauge the learnability of natural language by human children, and also natural language processing researchers who seek syntactic representations for languages with few linguistic resources.

Grammar learning has been popular in previous challenges. For example the CoNLL shared tasks in 2006 and 2007 (Buchholz and Marsi, 2006; Nivre et al., 2007) involved supervised learning of dependency parsers across a wide range of different languages. Our challenge has many similarities to these, in that we focus on dependency grammars, however we seek to evaluate unsupervised algorithms only using syntactically annotated data for evaluation and not for training. Additionally we also consider the related task of part-of-speech (POS) induction, and the next logical challenge: the joint task of POS and dependency induction. Other related challenges can be found in the formal grammar community (e.g., the Omphalos[1] competition) in which competitors seek to learn synthetic languages. In contrast we seek to model natural language text, which entails many different challenges.

Research into unsupervised grammar and POS induction holds considerable promise, although current approaches are still a long way from solving the general problem. For example, the majority of recent research into dependency grammar induction has adopted the evaluation setting of Klein and Manning (2004) who learn grammars on strings of POS tags, rather than on words themselves. One aim of this challenge is to popularise the more difficult and ambitious task of inducing grammars directly from text, which can be viewed as integrating the POS and grammar induction tasks. A second aim is to foster grammar and POS induction research across a wider variety of languages, and improving the standard of evaluation.

We have collated data from existing treebanks in a variety of different languages, domains and linguistic formalisms. This gives a diverse range of

[1] See http://www.irisa.fr/Omphalos

NAACL-HLT Workshop on the Induction of Linguistic Structure, pages 64–80,
Montréal, Canada, June 3-8, 2012.

data upon which to test induction algorithms, yielding a deeper insight into their strengths and shortcomings. One key problem in grammar induction research is how to evaluate the models' predictions given that often many different analyses are linguistically plausible, e.g., the choice of whether determiners or nouns should head noun phrases, or how to represent coordination. Simply comparing against a single gold standard often results in poor reported performance because the model has discovered a different analysis to that used when annotating the treebank. For this reason it has been popular to use lenient measures for comparing predicted trees to the treebank gold standard trees, such as undirected accuracy and the neutral edge distance (Schwartz et al., 2011). As well as evaluating using these popular metrics, we also propose a new method of evaluation which is also lenient in that it rewards different types of linguistically plausible output, but requires consistency in the output, something the previous methods cannot do.

The paper is organised as follows. Section 2 describes the tasks and our data format and section 3 outlines the different treebanks used for the challenge. The baselines, our own benchmark systems and the competitors entries are described in section 5. In section 6 we present and analyse the results for the three different tracks. Finally we conclude in section 7.

2 Task Definition

The three tracks of the WILS challenge are described below. First we describe the data format for the submissions common to the three tracks (POS induction, Dependency induction, and jointly inducing both), and then the three tracks are described along with the respective evaluation metrics.

2.1 Data format

All datasets were presented in a file format similar to that used in the CoNLL tasks, but with slight modifications. In particular the last two columns are removed, as no projective head or projective dependency relations were used, and an extra POS column was inserted at column 6 to accommodate the Universal POS tagset (Petrov et al., 2011). Each line in a file then either consists of 9 columns, separated by a tab character, or is an empty line. Empty lines separate sentences, and all other lines give the annotation for a single token in the sentence as follows:

1. ID: Token counter, gives the index of current word in the sentence. Indexing starts at 1.
2. FORM: Surface form of the token in the sentence.
3. LEMMA: Stemmed form of the word form if available.
4. CPOSTAG: Coarse-grained POS tag.
5. POSTAG: Fine-grained POS tag, or CPOSTAG again if not available.
6. UPOSTAG: Universal POS tag, based on the POSTAG and CPOSTAG.
7. FEATS: List of syntactic / morphological features, separated by a vertical pipe (|).
8. HEAD: Syntactic head of the token, with 0 indicating the root node.
9. DEPREL: The general type of the dependency relation, e.g., subject.

In this setup, the LEMMA, FEATS and DEPREL columns are optional, in which case an underscore (_) will be used as a placeholder. Each treebank was split into training, development and testing partitions. The HEAD and DEPREL entries were only supplied for the development and the final testing sets,[2] but not for the training partition. The competitors were encouraged to develop their unsupervised entries on the union of the three partitions, and make sparse use of the development set, i.e., for sanity checking more than model fitting in order to minimise the extent of supervision.

2.2 POS induction

In the POS induction track, participants developed systems to induce the Part-of-Speech (POS) classes for each word in the testing corpus. In order to train the systems, the same training and development sets were used as for the other tracks. These corpora included manually supplied POS tags for each token,

[2] For the initial test set these fields were omitted.

which were not to be used for training, only evaluation. Participants submitted predicted tags for each token, which were scored against the gold-standard.

For evaluation, we used 4 different metrics. The first is the many-to-one metric (M-1) (also known as cluster purity), which is widely used for cluster evaluation as well as evaluation of POS induction. This metric assigns each word cluster to its most common tag, and then measures the proportion of correctly tagged words. The second metric is the one-to-one mapping (1-1), a constrained version of Many-to-one mapping in which each predicted tag is associated with only one gold-standard tag and vice versa (Haghighi and Klein, 2006). Word clusters are assigned greedily to tags, and in the event of there being more word classes than tags, some word classes will be left unassigned. Another metric that was used is Variation of information (VI) (Meila, 2003), which is based the conditional entropy of between the two different clusterings (Johnson, 2007). Lastly, we use the V-measure (VM) metric (Rosenberg and Hirschberg, 2007), which is another entropy-based measure, but defined in terms of a F score to balance precision and recall terms (we use equal weighting of the two factors). Please see Christodoulopoulos et al. (2010) for further details about these metrics.[3] For these metrics, a higher score is better, with the exception of VI.

For all these metrics, the induced tags are evaluated against the universal pos tags, as this means there are a consistent number of tags across the languages. Using these metrics, the results will vary as a result of predicting a different number of tags (in particular, more tags will mean a higher score for M-1, and the converse is true for 1-1). However, using the universal POS tags, we think will make results less sensitive to large differences in POS inventory between languages (such as for the Dutch dataset).

2.3 Dependency induction

For the Dependency induction track, the training data consisted of the original treebank data, but without dependency annotations. A development set was also provided, which included the dependency annotations, but this was meant mainly as a way to verify systems, as we mean to minimise the amount of supervision in the task. The participants were later supplied with test sets for which the systems could generate predictions. Only after the predictions were submitted were the fully annotated test sets released.

The dependency inductions were evaluated on 3 metrics: directed accuracy, undirected accuracy and Neutral Edge Detection (NED) (Schwartz et al., 2011). Directed accuracy is the ratio of correctly predicted dependencies (including direction) over total amount of predicted dependencies. Undirected accuracy is much the same, but also considers a predicted dependency correct if the direction of the dependency is reversed (e.g. if the predicted dependency is not $A \rightarrow B$, but $B \rightarrow A$). Lastly, the NED metric is a variant of undirected accuracy that also rewards cases where an edge-flip occurs, meaning that the predicted parent of a token is actually the grandparent of that same token in the gold-standard data. Note that before evaluating with these metrics punctuation was removed from all sentences, and any child words under a punctuation node were reattached to their nearest ancestor that wasn't punctuation.

The final 'joint' task consisted of inducing dependency structure from only the tokens in the corpus, without recourse to the gold POS tags. Where POS is predicted (e.g., in a pipeline), we included these in our general POS evaluation. The induced dependency trees were evaluated with the same metrics as in the dependency induction track, but are considered separately. We expect these systems to have lower scores overall due to the lack of gold-standard POS tags.

3 Treebanks

We selected a number of different treebanks for use in the challenge, aiming to represent a wide range of different languages, dialects and genres of text. In total we used ten different treebanked corpora in nine different languages. For the practical reasons of simplifying the administration of the challenge and allowing the data to be reused in future research, we chose corpora with licences allowing either free redistribution, or those held by the Linguis-

[3] Thanks to Christos Christodoulopoulos for sharing his implementation of the POS induction metrics, which we have used in our evaluation.

tic Data Consortium (LDC).[4] Many of these datasets have been used before in dependency grammar or part-of-speech research, particularly the shared tasks at CoNLL 2006 and 2007. For the purpose of the competition, we have updated these datasets to include any annotation updates or additional data, where available. It is important for unsupervised approaches to have sufficient amounts of data, especially given the common sentence length limitations imposed by most dependency grammar models. As described in section 2, we have included an extra field for the universal part-of-speech (UPOS) using Petrov et al. (2011)'s automatic conversion tool.[5]

Below we describe the different treebanks used, and the conversion process into our data format for the purpose of the competition. Please see Table 1 for statistics on each of the treebanks.

Dependency treebanks We used the following dependency treebanks: **Arabic** The Prague Arabic Dependency Treebank V1 (Hajič et al., 2004).[6] **Basque** The Basque 3lb dependency treebank (Aduriz et al., 2003). **Czech** The Prague Dependency Treebank 2.0 (Böhmová et al., 2001).[7] **Danish** The Copenhagen Dependency Treebank version 2 (Buch-Kromann et al., 2007). **English** The CHILDES `US/Brown` subcorpus (Sagae et al., 2007). **Slovene** The jos500k Treebank (Erjavec et al., 2010). [8] **Swedish** The Talbanken treebank (Nivre et al., 2006). The conversion of each of these treebanks was quite straightforward as they were already annotated for dependencies. Moreover, many of these corpora had been used previously in the CoNLL 2006 and 2007 shared tasks, and therefore we were able to reuse this data and/or their conversion scripts. In the case of Arabic and Swedish we used the exact same data, simply converting from CoNLL dependency format into our own format (removing redundant columns and adding a UPOS column). While many of the other corpora had also been used previously, our data is different, making use of subsequent corrections to these treebanks and additional annotated data now available.

First language acquisition provides an important motivation for grammar induction research, consequently we have included data from the CHILDES database of child-directed speech. We use the Brown sub-corpus, a longitudinal study of parent-child interactions for three children aged between 18 months and 5 years old. The corpus has been manually annotated with syntactic dependencies (Sagae et al., 2007) and morphology. From this we take all child-directed utterances, extracting word, morphology, part-of-speech and dependency markup, and developed our own conversion into UPOS. Our testing and development sets were drawn from the first 15 Eve files which were manually annotated for dependency structure. The rest of the corpus, which had not been manually annotated for syntax, was merged to form the training set.

Phrase-structure treebanks As well as dependency treebanks, we used three different phrase-structure treebanks: The **Dutch** Alpino treebank (Bouma et al., 2000), the **English** Penn Treebank V3 (Marcus et al., 1993),[9] and the **Portuguese** Floresta Sintá(c)tica treebank (Afonso et al., 2002). As these treebanks do not explicitly mark dependencies, we automatically extracted these using head finding heuristics. Thankfully the difficult work of creating such scripts has already been done as part of the CoNLL shared tasks. We have reused their scripts to create dependency representations of these treebanks, before converting into our file format and augmenting with UPOS annotation. In the case of Dutch, we have reused the same CoNLL 2006 data; note that this dataset includes predicted part-of-speech rather than gold standard annotation (Buchholz and Marsi, 2006). For the Portuguese, we used the same Bosque 7.3 sub-corpus[10] from CoNLL 2006, additionally including in our training set the recently-annotated Selva 1.0 subcorpus.

The Penn Treebank is the most common data set in parsing and grammar induction. We have patched

[4] In the following corpus descriptions, when not otherwise specified the corpus is freely available for research purposes.

[5] `http://code.google.com/p/universal-pos-tags`

[6] LDC catalogue number LDC2004T23.

[7] LDC catalogue number LDC2006T01.

[8] For the shared task, the annotation was converted to english using the tables found at the JOS website: `http://nl.ijs.si/jos/msd/html-en/index.html`

[9] LDC catalogue number LDC99T42.

[10] An updated version of this corpus is available, however the file format had changed significantly and we were unable to adapt the conversion scripts in time for the competition.

	ar	cs	da	en-childes	en-ptb	eu	nl	pt	sl	sv
annotation	d	d	d	d	p	d	p	p	d	d
Training data										
Tokens	106.6k	1.2M	68.5k	312.8k	1.1M	124.7k	192.2k	196.4k	193k	184.6k
Sentences	2.8k	68.5k	3.6k	57.4k	45.4k	9.1k	13k	8.7k	9.4k	10.7k
Tokens/sent	38.4	17.1	18.8	5.5	23.9	13.7	14.8	22.6	20.5	17.3
CPOSTAG	15	12	25	31	31	16	13	16	13	41
POSTAG	21	61	141	76	45	50	300	22	31	41
FEATS	22	75	338	29	0	269	310	146	46	0
Development data										
Tokens	5.1k	159k	17k	25.3k	32.9k	12.6k	2.9k	10.3k	20.2k	6.9k
Sentences	139	9.3k	1k	5k	1.3k	1k	386	400	1k	389
Tokens/sent	36.8	17.1	17	5.1	24.4	12.5	7.4	25.8	20.2	17.6
% New words	27.5	26	49.8	9.8	11.4	46.1	18.8	27.5	38.7	13.8
Test data										
Tokens	5.1k	173.6k	14.7k	28.4k	56.7k	14.3k	5.6k	5.9k	22.6k	5.7k
Sentences	131	10.1k	1k	5.2k	2.4k	1.1k	386	288	1k	389
Tokens/sent	39.1	17.1	14.7	5.4	23.5	12.7	14.5	20.4	22.6	14.5
% New words	24.3	25.3	43.7	9	12.1	51.5	40.5	25.2	37.1	34.6

Table 1: Properties of the treebanks. We report the linguistic annotation method (dependency vs. phrase-structure), the size of each treebank, the number of types for the different granularities of part-of-speech tags and morphological features (note that UPOS has a fixed set of 12 tags), and the proportion of word types that were not present in training.

the treebank to include NP-internal structure using Vadas and Curran's annotations (Vadas and Curran, 2007), which was then converted to dependency structures using the `penn-converter`[11] script (Johansson and Nugues, 2007). This tool has a number of options controlling the linguistic decisions in converting from phrase-structure to dependency trees, e.g., the treatment of coordination. We extracted five versions of the treebank, each encoding each different sets of linguistic assumptions (Tsarfaty et al., 2011).[12] These are denoted default, oldLTH, CoNLL-2007, functional and lexical; for the main results we used the standard options, we also report separately evaluations using each of the five variants. The treebank was partitioned into training (sections 0-22), development (sec. 24) and testing sets (sec. 23).

4 Baselines and Benchmarks

A number of standard baselines and previously published benchmark systems were implemented for each task in order to place the submitted systems in context.

The standard baseline for grammar induction models is to assume either left branching or right branching analyses (`LB`, `RB`). These capture the tendency for languages to favour one attachment direction over another. The most frequently cited and extended model for dependency induction is DMV (Klein and Manning, 2004). We provide results for this model trained on each of the coarse (`DMV`^{c}), fine (`DMV`^{p}), and universal (`DMV`^{u}) POS tag sets, all initialised with the original harmonic initialiser. As a further baseline we also evaluated the dependency trees resulting from directly using the harmonic initialiser without any training (`H`).

As a strong benchmark we include the results of the non-parametric Bayesian model previously published in Blunsom and Cohn (2010) (`BC`). The stated results are for the unlexicalised model described in that paper where the final analysis is formed by choosing the maximum marginal probability dependency links estimated from forty independent Gibbs sampler runs.

For part-of-speech tagging we include results from an implementation of the Brown word clustering algorithm (Brown et al., 1992) (`B`^{c,p,u}), and the `mkcls` tool written by Franz Och (Och, 1999) (`MK`^{c,p,u}). Both of these benchmarks were trained with the number of classes matching the number in the gold standard of each of the tagsets in turn: coarse (`c`), fine (`p`), and universal (`u`). A notable

[11] `http://nlp.cs.lth.se/software/treebank_converter`

[12] Note that Tsarfaty et al. (2011) also propose an evaluation metric for comparing dependency trees, which we have not used. Note however that it could, in principle, be used for similar evaluations.

property of both of these word class models is that they enforce a one-tag-per-type restriction that ensures there is a one-to-one mapping between word types and classes.

For POS tagging we also provide benchmark results from two previously published models. The first of these is the Pitman-Yor HMM model described in (Blunsom and Cohn, 2011), which incorporates ta one-tag-per-type restriction (BC). This model was trained with the same number of tags as in the gold standard fine tag set for each corpus. The second benchmark is the HMM with Sparsity Constraints trained using Posterior Regularization (PR) described in (Graça et al., 2011). In this model the HMM emission probabilitiy distribution are estimated using small Maximum Entropy models (features set described in the original paper). The models were trained for 200 iterations of PR using both the same number of hidden states as the coarse $\mathtt{G}^c$ and universal $\mathtt{G}^u$ gold standard. All parameters were set to the values described in the original paper.

5 Submissions

The shared task received submissions covering a diverse range of approaches to the dependency and part-of-speech induction challenges. Encouragingly all of these submissions made significant departures from the benchmark HMM and DMV approaches which have dominated the published literature on these tasks in recent years. The submissions were characterised by varied choices of model structure, parameterisation, regularisation, and the degree to which light supervision was provided through constraints or the use of labelled tuning data. In the following sections we summarise the approaches taken by the systems submitted for each task.

5.1 Part-of-Speech Induction

The part-of-speech induction challenge received two submission, (Chrupała, 2012; Christodoulopoulos et al., 2012). Both of these submissions based their induction systems on LDA inspired models for clustering word types by the contexts in which they appear. Notably, the strongest of the provided benchmarks and the two submissions modelled part-of-speech tags at the type level, thus restricting all tokens of a given word type to share the same tag. Though clearly out of step with the gold standard tagging, this one-tag-per-type restriction has previously been shown to be a crude but effective way of regularising models towards a good solution. Below we summarise the approach of each submission, identified by the surname of the first author on the submitted system description.

Chrupała (2012) employed a two stage approach to inducing part-of-speech tags. The first stage used an LDA style probabilistic model to induce a distribution over possible tags for a given word type. These distributions were then hierarchically clustered and the final tags selected using the prefix of the path from the root node to the word type in the cluster tree. The length of the prefixes, and thus the number of tags, was tuned on the labelled development data.

The system of Christodoulopoulos et al. (2012) was based upon an LDA type model which included both contexts and other conditionally independent features (Christodoulopoulos et al., 2011). This base system was then iterated with a DMV system and with the resultant dependencies being repeatedly fed back into the POS model as features. This submission is notable for being one of the first to attempt joint POS and dependency induction rather than taking a pipeline approach.

5.2 Dependency Induction

The dependency parsing task saw a variety of approaches with only a couple based on the previously dominant DMV system. Two forms of light supervision were popular, the first being the inclusion of pre-specified constraints or rules for allowable dependency links, and the second being the tuning of model parameters or selecting between competing models on the labelled development data. Obviously the merits of such supervision would depend on the desired application for the induced parser. The direct comparison of models which include a form of universal prior syntactic information with those that don't does permit interesting development linguistic questions to be explored in future.

Bisk and Hockenmaier (2012) chose to induce a restricted form of Combinatory Categorial Grammar (CCG), the parses of which were then mapped to dependency structures. Restrictions on head-child dependencies were encoded in the allowable cate-

gories for each POS tag and the heads of sentences. Key features of their approach were a maximum likelihood objective function and an iterative procedure for generating composite categories from simple ones. Such composite categories allow the parameterisation of larger units than just head-child dependencies, improving over the more limited conditioning of DMV.

Maraček and Žabokrtský (2012) introduced a number of novel features in their dependency induction submission. Wikipedia articles were used to quantify the reducibility of word types, the degree to which the word could be removed from a sentence and grammaticality maintained. This metric was then used, along with a model of child fertility and dependency distance, within a probabilistic model. Inference was performed by using a local Gibbs sampler to approximate the marginal distribution over head-child links.

Søgaard (2012) presented two model-free heuristic algorithms. The first was based on heuristically adding dependency edges based on rules such as adjacency, function words, and morphology. The resulting structure is then run through a PageRank algorithm and another heuristic is used to select a tree from the resulting ranked dependency edges. The second approach takes the universal rules of Naseem et al. (2010) but rather than estimating a probabilistic model with these rules, a rule based heuristic is used to select a parse rather. This second model-free approach in particular provides a strong baseline for probabilistic models built upon hand-specified dependency rules.

Tu (2012) described a system based on an extended DMV model. Their work focussed on the exploration of multiple forms of regularisation, including Dirichlet priors and posterior regularisation, to favour both sparse conditional distributions and low ambiguity in the induced parse charts. While many previous works have included sparse priors on the conditional head-child distributions the additional regularisation of the ambiguity over parse trees is a novel and interesting addition. The labelled development sets were employed to both select between models employing different regularisation, and to tune model parameters.

5.3 POS and Dependency Induction

There was only a single submission for the task of inducing dependencies without gold standard part-of-speech tags supplied. Christodoulopoulos et al. (2012) submitted the same joint tagging and DMV system used for the POS induction task to the dependency induction task. Results on the development data indicated that this iterated joint training had a significant benefit for the induced tags and a smaller benefit for the dependency structures induced.

6 Results

The main results for the three tasks are shown in Tables 2, 3, and 4, for the POS induction, dependency induction and joint tasks, respectively.[13] We now present a detailed analysis of each of the three tasks.

6.1 POS induction

The main evaluation results for the POS induction task are shown in Table 2, which compares the induced clusters against the gold universal tags (UPOS).[14] Given the diversity of scenarions used by each system (e.g. number of hidden states, tuning on development data) a direct comparison between the systems can only be illustrative. A first observation is that depending on the particular evaluation metric employed the ranking of the systems changes substantially, for instance the G^u system is the best using the 1-1 and VI metric but is the worst of the entries (excepting the baselines) when using the other two metrics. Focusing on the VM metric, which was shown empirically not to have low bias with respect to the word classes (Christodoulopoulos et al., 2010), the best entry is the BC system which has the best performance in 9 out of 10 entries followed by the CGS and the C system. Note that this ranking holds also for the comparison against fine POS tags, shown in Table 7.

An interesting aspect is that almost all systems beat the strong Brown (B) and mkcls (MK) baseline across the different metrics when we restrict our attention to the cases where the same number

[13] Additional tables of results are in the appendix, and further results are online at `http://wiki.cs.ox.ac.uk/InducingLinguisticStructure`.

[14] See also Table 7 for the comparison against the fine POS tags; we base our analysis on UPOS instead as this tag set has a fixed size irrespective of the treebank.

M-1												
testset	BC	CGS	C^c	C^p	G^c	G^u	B^c	B^p	B^u	MK^c	MK^p	MK^u
arabic	**83.80**	N/A	83.33	83.33	65.05	66.61	66.20	69.89	66.22	71.08	72.72	68.03
basque	80.85	79.54	**86.67**	**86.67**	77.37	73.88	74.73	77.49	73.32	74.80	78.63	71.40
czech	**83.10**	66.78	72.27	77.97	N/A	N/A	60.85	75.57	60.42	65.43	79.35	57.16
danish	81.44	77.76	84.13	**84.92**	68.16	53.78	72.12	79.77	47.09	72.26	82.59	53.07
dutch	80.75	70.13	74.04	76.11	63.37	57.64	57.99	84.17	57.31	68.18	**84.78**	63.04
en-childes	90.36	85.42	**91.50**	**91.50**	N/A	N/A	82.65	89.70	70.12	86.27	91.44	75.63
en-ptb	**86.73**	81.93	78.11	84.35	77.14	71.10	77.29	80.88	63.74	79.99	83.88	63.34
portuguese	81.69	77.38	80.38	**81.90**	75.54	74.35	70.07	74.25	67.60	70.79	72.90	68.08
slovene	70.81	65.31	75.53	**75.92**	67.94	59.96	61.58	68.93	58.32	58.43	65.69	50.36
swedish	78.61	**80.45**	79.60	79.60	69.91	58.79	71.69	71.69	57.55	76.45	76.45	57.30
averages	81.82	76.08	80.56	**82.23**	70.56	64.51	69.52	77.23	62.17	72.37	78.84	62.74

1-1												
testset	BC	CGS	C^c	C^p	G^c	G^u	B^c	B^p	B^u	MK^c	MK^p	MK^u
arabic	53.67	N/A	39.44	39.44	39.83	**55.52**	40.55	33.57	43.31	51.54	40.24	51.58
basque	36.10	36.03	47.15	47.15	47.09	**54.70**	32.61	20.53	40.62	34.80	27.28	37.65
czech	31.82	**49.30**	30.49	27.20	N/A	N/A	46.19	26.66	45.10	43.70	24.48	39.25
danish	42.54	42.77	31.67	31.04	39.95	**45.58**	36.04	17.74	39.19	43.89	22.18	44.23
dutch	42.79	56.15	43.10	39.62	**56.45**	45.37	48.18	21.36	43.12	55.99	21.32	54.09
en-childes	38.79	42.57	43.76	43.76	N/A	N/A	40.78	35.54	57.71	43.45	32.00	**59.18**
en-ptb	41.55	39.57	43.86	31.56	42.07	**51.70**	39.79	33.90	46.50	40.55	36.22	51.17
portuguese	**59.66**	47.45	35.90	35.50	46.50	56.08	51.15	42.68	51.58	44.28	35.38	46.31
slovene	39.02	**53.04**	33.18	32.50	50.90	48.50	46.83	40.16	42.28	40.34	39.32	40.58
swedish	42.38	32.44	26.45	26.45	34.99	**54.92**	27.56	27.56	51.34	35.82	35.82	43.60
averages	42.83	44.37	37.50	35.42	44.72	**51.55**	40.97	29.97	46.07	43.44	31.42	46.76

VM												
testset	BC	CGS	C^c	C^p	G^c	G^u	B^c	B^p	B^u	MK^c	MK^p	MK^u
arabic	**61.75**	N/A	51.27	51.27	44.81	47.07	39.93	42.43	39.92	47.47	43.91	44.49
basque	42.17	41.52	**43.04**	**43.04**	40.86	40.05	34.85	33.33	36.08	36.32	34.35	33.42
czech	**52.26**	45.31	40.22	39.20	N/A	N/A	38.56	42.90	37.46	41.70	46.03	37.34
danish	**56.57**	54.63	52.46	52.32	47.26	41.96	47.89	44.37	35.13	50.52	48.17	39.96
dutch	**56.96**	53.35	54.87	52.90	48.57	45.80	43.34	49.33	43.67	51.37	50.11	47.20
en-childes	**64.53**	62.32	62.76	62.76	N/A	N/A	58.87	60.31	57.06	62.76	60.92	60.51
en-ptb	**60.73**	57.99	53.14	52.09	55.10	52.54	54.76	55.08	48.04	56.81	57.29	48.46
portuguese	**64.17**	58.41	52.54	52.32	55.96	58.14	52.09	53.18	50.32	52.48	50.87	50.18
slovene	51.15	**51.29**	46.60	46.50	50.98	45.98	44.49	45.80	38.61	36.79	43.43	36.43
swedish	**57.05**	54.21	47.08	47.08	48.89	45.73	45.87	45.87	40.84	49.77	49.77	42.83
averages	**56.73**	53.23	50.40	49.95	49.05	47.16	46.06	47.26	42.71	48.60	48.48	44.08

VI												
testset	BC	CGS	C^c	C^p	G^c	G^u	B^c	B^p	B^u	MK^c	MK^p	MK^u
arabic	2.48	N/A	3.70	3.70	3.39	2.98	3.78	3.94	3.53	3.31	3.82	3.30
basque	3.82	3.44	3.98	3.98	3.25	**2.82**	3.92	4.98	3.45	3.79	4.76	3.58
czech	3.83	3.41	4.92	5.77	N/A	N/A	3.70	4.76	3.69	3.63	4.53	3.83
danish	3.36	**3.34**	4.31	4.38	3.78	3.46	3.86	5.43	3.79	3.64	4.90	3.59
dutch	3.56	**3.13**	3.28	3.71	3.30	3.44	3.66	5.22	3.60	3.26	5.15	3.39
en-childes	2.81	2.86	3.06	3.06	N/A	N/A	3.13	3.34	2.59	2.84	3.33	2.50
en-ptb	3.18	3.28	3.67	4.36	3.34	**3.03**	3.46	3.62	3.36	3.36	3.52	3.28
portuguese	**2.47**	2.83	3.96	4.09	2.96	2.62	3.19	3.36	3.10	3.21	3.52	3.15
slovene	3.62	**3.14**	4.80	4.86	3.16	3.30	3.61	4.09	3.73	4.15	4.33	3.99
swedish	**3.31**	3.68	4.98	4.98	3.90	3.32	4.46	4.46	3.70	4.07	4.07	3.62
averages	3.24	3.23	4.07	4.29	3.39	**3.12**	3.68	4.32	3.45	3.53	4.19	3.42

Table 2: Results for the POS induction task, showing one-to-one, many-to-one, VM and VI scores, measured against the gold UPOS tags. Each system is shown in a column, where the title is an acronym of the authors' last names, or else the name of a benchmark system (B is the Brown clusterer and MK is mkcls). The superscripts c, p and u denote different applications of the same method with a number of word classes set to equal the true number of coarse tags, full tags or universal tags, respectively, for each treebank.

of hidden states are used (the exception being the G system which occasionally under-performed against MK). Interestingly the assumption of one-tag-per-word, made by all but the G system, works very well in practice leading to consistently strong results. This suggests that dealing with word ambiguity is still an unresolved issue in unsupervised POS induction.

Comparing the performance of the systems for different languages, as expected the languages for which we have a larger corpora (English CHILDES and PTB and Czech) tend to result in systems with better accuracies. An interesting future question is how do the propose methods scale when training on really large corpora (e.g., wikipedia) both in terms of performance (accuracy) but also in the resources they required.

Finally, the wild divergences in the system rankings when considering the different evaluation metrics calls for some sort of external evaluation using the induced clusters as features to other end systems, for instance semi-supervised tagging. The main question is if there will be a definitive ranking between systems for a diverse set of tasks, or if on the contrary the effectiveness of the output of each system will vary according to the task at hand.

6.2 Dependency induction

The main evaluation results for the dependency task are shown in Table 3. From this we make several observations.[15] Firstly, for almost all the corpora the participants systems have outperformed the simple baselines, and by a significant margin. There are three exceptions to this: for Arabic, Basque and Danish the left or right-branching baselines outperforms most or all of the competitors. This may indicate that these languages are inherently difficult, or may simply be a consequence of these three languages having the least data of all of our corpora. Basque and Dutch proved to be the hardest of the treebanks, with the lowest overall scores, and the CHILDES (English) and Portuguese were the easiest. The reasons for this are not immediately clear, although we speculate that Basque is difficult due to its dissimilarity from other European languages, and therefore may not match the assumptions underlying models developed primarily on English. Dutch is difficult as its annotation was non-projective, and it has a very large set of POS tags, while CHILDES is made easier due to its extremely short and simple sentences.

In terms of declaring a 'winner', it is clear that Tu's system ranks best under directed accuracy and NED, and a very narrow second (to the organisers' submission, BC) for undirected accuracy. Moreover Tu's system was a consistent performed across all corpora, with no single result well below the results of the other participants. Note that the three different metrics often predict the same winner across the different treebanks, however there are some large discrepancies, such as Portuguese and Dutch where the directed and undirected accuracy metrics concur, but NED produces a very different ranking. It is unclear which metric should be trusted more than another; this could only be assessed by correlating these metrics with some form of secondary evaluation, such as in a task based setting or obtaining human grammaticality judgements.[16]

The benchmark systems include DMV (Klein and Manning, 2004), which has historical importance in terms of being the first research systems to outperform simple baselines for dependency induction, and also the model upon which most recent dependency induction research is based, including many of the competitors in the competition. We observe that in most cases the competitors have outperformed the DMV models, in many cases by a large margin. In all cases DMV improved over its initialisation condition (the harmonic initialiser), although often this improvement was only slight, underscoring the importance of good initialisation. The effect of inducing DMV grammars from various different granularity of POS tags made little difference in most cases, although for Dutch[17] and the English PTB there change was more dramatic.

[15] Table 3 evaluates against the full test sets, however it is traditional to present results for short sentences mirroring the common training setup. See Tables 8 and 9 for results over sentences with 10 words or fewer, excluding punctuation. Note that our analysis is based on the results for the full test set.

[16] It was our intention to include a task-based evaluation for machine translation, but this proved impractical for the competition due to the volumes of data that we would require each participant to process.

[17] Note that for Dutch the full POS tags were not gold standard, but were system predictions.

Directed														
testset	BC	BH	MZ^c	MZ^p	MZ^u	S^1	S^2	Tu	DMV^c	DMV^p	DMV^u	H	LB	RB
arabic	61.1	47.2	15.7	64.8	64.8	47.2	54.6	**66.7**	46.3	45.4	50.0	42.6	9.3	64.8
basque	56.3	50.3	28.0	30.3	27.2	33.3	22.3	**58.6**	46.3	43.2	31.3	21.8	34.3	24.4
czech	50.0	48.5	**61.3**	57.5	57.7	45.5	51.2	59.0	30.1	31.2	31.8	24.7	28.9	34.3
danish	46.2	49.3	60.2	51.3	**61.4**	56.9	60.5	60.8	47.2	50.2	35.3	36.4	18.7	49.2
dutch	50.5	50.8	37.0	49.5	38.4	38.9	50.0	**51.7**	48.7	39.7	49.1	35.1	34.0	39.5
en-childes	48.1	**62.2**	56.8	47.2	51.8	50.5	53.5	56.0	51.7	51.9	39.0	31.7	36.0	23.3
en-ptb	72.1	73.7	58.9	67.4	52.2	44.8	61.0	**74.7**	31.7	44.7	30.6	35.2	40.4	19.9
portuguese	54.3	**76.3**	63.6	59.9	44.3	47.7	71.1	55.7	27.1	37.2	26.9	31.1	28.1	37.7
slovene	65.8	53.9	42.1	51.4	39.2	39.7	50.3	**67.7**	35.7	37.2	35.6	25.7	35.9	14.7
swedish	65.8	66.7	61.4	63.7	70.8	48.2	72.0	**76.5**	44.2	44.2	45.8	39.1	33.2	31.3
averages	57.0	57.9	48.5	54.3	50.8	45.3	54.7	**62.7**	40.9	42.5	37.5	32.3	29.9	33.9
Undirected														
testset	BC	BH	MZ^c	MZ^p	MZ^u	S^1	S^2	Tu	DMV^c	DMV^p	DMV^u	H	LB	RB
arabic	57.3	29.7	57.6	62.0	58.7	48.0	58.4	59.3	41.8	42.0	43.7	41.2	61.7	**63.9**
basque	**58.0**	47.2	43.3	45.0	43.2	47.5	24.3	53.3	48.1	47.7	40.3	37.6	53.9	53.1
czech	59.0	45.0	57.8	54.3	55.5	49.3	55.8	**61.4**	46.2	46.7	45.3	38.5	51.5	52.3
danish	60.8	50.7	60.7	56.1	60.3	56.6	60.5	**61.6**	55.1	54.1	51.6	46.0	58.7	59.9
dutch	**61.0**	45.0	47.5	51.5	48.9	46.8	51.4	54.6	52.2	45.0	52.2	37.2	50.1	50.8
en-childes	63.5	**68.4**	67.2	59.9	61.4	62.0	62.4	66.9	63.8	64.0	57.5	49.0	50.0	49.9
en-ptb	**66.2**	58.1	49.7	57.6	48.2	49.5	58.8	62.1	43.1	53.1	43.0	36.2	51.7	51.5
portuguese	56.6	**72.4**	61.4	61.9	49.8	52.6	66.9	61.4	44.3	48.1	43.6	41.2	55.7	56.8
slovene	58.1	47.9	45.2	49.1	44.5	42.4	53.5	**61.8**	42.1	40.6	42.1	32.5	40.8	41.1
swedish	**70.0**	58.5	58.8	59.3	60.4	53.5	65.2	66.9	51.1	51.1	53.3	44.5	53.0	53.2
averages	**61.0**	52.3	54.9	55.7	53.1	50.8	55.7	60.9	48.8	49.2	47.3	40.4	52.7	53.2
NED														
testset	BC	BH	MZ^c	MZ^p	MZ^u	S^1	S^2	Tu	DMV^c	DMV^p	DMV^u	H	LB	RB
arabic	63.6	37.3	59.2	67.1	63.2	56.5	65.8	64.1	48.9	48.0	48.8	47.5	62.7	**69.0**
basque	**69.6**	55.8	51.5	55.6	53.4	58.8	38.0	65.8	57.6	57.1	51.5	49.3	67.2	59.1
czech	71.0	55.7	70.2	65.2	67.3	63.2	69.7	**71.6**	53.2	52.9	54.1	47.6	56.3	68.6
danish	72.0	63.1	72.9	69.5	73.5	65.9	71.8	**76.4**	64.8	63.5	58.9	53.5	61.6	71.5
dutch	71.6	58.6	68.6	**72.0**	69.7	60.6	63.8	66.9	63.5	54.5	63.5	46.9	55.1	67.0
en-childes	80.9	79.6	82.8	74.1	72.7	77.1	**83.2**	80.4	78.1	78.3	77.5	67.2	61.0	75.2
en-ptb	**75.2**	69.8	69.4	73.8	67.2	64.1	71.6	69.8	49.8	67.0	49.6	44.8	53.9	68.1
portuguese	67.5	79.8	75.6	75.7	71.7	66.9	78.2	**80.4**	62.1	66.6	61.3	51.8	57.3	75.4
slovene	64.4	60.7	56.9	58.9	57.1	55.9	66.7	**68.7**	49.2	47.3	49.2	38.9	43.8	56.6
swedish	**80.1**	70.9	73.2	73.8	72.7	66.7	77.0	77.1	64.0	64.0	62.0	56.0	56.5	71.0
averages	71.6	63.1	68.0	68.6	66.9	63.6	68.6	**72.1**	59.1	59.9	57.6	50.3	57.6	68.1

Table 3: Directed accuracy, undirected accuracy and NED results for the dependency task (using supplied POS). The first column (BC) is our benchmark system, the next seven are participants systems, and the remaining columns consist of the DMV benchmark and various simple baselines. The superscripts c, p and u denote which type of POS was used, and S^1 and S^2 denote two different submissions for Søgaard (2012).

Overall the full POS tagset lead to the best performance over the coarse and universal tags (considering undirected accuracy or NED), which is to be expected as there is considerably more syntactic information contained in the full POS. This must be balanced against the additional model complexity from expanding its parameter space, which may explain why the difference in performance differences are so small. The same pattern can also be seen in Maraček and Žabokrtský (2012)'s submission, whose system using full POS (M^p) outperformed their other variants.

6.3 Joint task

As we had only one submission for the joint problem of POS and dependency induction, there are few conclusions we can draw for this joint task (see Table 4 for the results, and Table 9 for the short sentence evaluation). Compared to the dependency induction task using gold standard POS, as shown in Table 3, the accuracy for the joint models are lower. Interestingly, the DMV model performs best when using the same number of word clusters as there are POS tags, mirroring the findings reported

directed				
testset	CGS	DMV^c	DMV^p	DMV^u
arabic	N/A	35.3	**44.4**	34.2
basque	24.5	27.5	25.1	**28.7**
czech	24.7	19.9	**33.2**	20.0
danish	21.4	23.3	**31.9**	10.0
dutch	15.1	20.6	**33.7**	20.5
en-childes	29.9	38.6	**42.2**	40.3
en-ptb	21.5	22.5	**23.3**	17.2
portuguese	19.7	**28.5**	28.0	17.1
slovene	**19.2**	13.9	11.5	14.4
swedish	23.6	**26.4**	**26.4**	20.5
averages	22.2	25.7	**30.0**	22.3

undirected				
testset	CGS	DMV^c	DMV^p	DMV^u
arabic	N/A	45.5	**52.5**	45.0
basque	43.5	46.4	**47.3**	47.0
czech	38.9	37.5	**50.9**	38.5
danish	51.4	**52.2**	48.8	37.3
dutch	40.3	41.9	**48.6**	40.8
en-childes	54.9	59.2	**60.8**	58.1
en-ptb	43.4	45.4	**48.8**	39.4
portuguese	45.5	51.8	**52.7**	39.8
slovene	32.8	33.3	**36.7**	32.8
swedish	45.6	**48.9**	**48.9**	40.3
averages	44.0	46.2	**49.6**	41.9

NED				
testset	CGS	DMV^c	DMV^p	DMV^u
arabic	N/A	53.4	**57.6**	53.3
basque	**55.9**	55.6	54.4	54.7
czech	51.2	49.3	**63.4**	51.5
danish	**61.7**	60.3	60.4	46.3
dutch	47.2	**57.5**	56.8	55.2
en-childes	**78.2**	77.7	78.1	76.5
en-ptb	53.9	60.2	**63.5**	47.5
portuguese	50.0	69.4	**70.8**	57.9
slovene	40.7	38.7	**47.5**	40.3
swedish	54.5	**65.4**	**65.4**	54.3
averages	54.8	58.8	**61.8**	53.8

Table 4: Directed, undirected and NED accuracy results for evaluating the predicted dependency structures in the joint task (i.e., not using supplied POS tags). The first column is the participant's system and the next three are DMV models trained on the Brown word clusters (see section 6.1).

above with gold standard tags. The best joint system was the DMV^p model, which only marginally under-performed the equivalent DMV model trained on gold POS. This is an encouraging finding, suggesting that word clusters are able to represent important POS distinctions to inform deeper syntactic processing.

6.4 Analysis

Until now we have adopted the standard metrics in dependency evaluation: namely directed head attachment accuracy, and its more lenient counterparts, undirected accuracy and NED. The latter metrics reward structures that almost match the gold standard tree, by way of rewarding child-parent edges that are predicted in the reverse direction, i.e., attaching the child as the parent (NED takes this further, by also rewarding the grandparent-child edge when this occurs). This allows some degree of flexibility when considering various contentious linguistic decisions such as whether a preposition should head a preposition phrase, or the head of the child noun-phrase. This added leniency comes at a price, as shown in Table 3 where the undirected accuracy and NED results are considerably higher than directed accuracy, and display less spread of values (look in particular at the random trees, Ra). Is is unclear that the predicted trees are truly predicting linguistically plausible structures, but instead that the differences are due largely to chance. Moreover, systems that predict linguistic phenomena inconsistently between sentences or across types of related phenomena are rewarded under these lenient metrics.

For these reasons we also consider a different, less permissive, evaluation method, using multiple references of the treebank where each is annotated with different styles of dependency. As described in section 2, we processed the Penn treebank five times with different options to the LTH conversion tool. This affected the treatment of coordination, preposition phrases, subordinate clauses, infinitival clauses etc. Next we compare the directed accuracy of the systems against these five different 'gold standard' references, which are displayed in Table 5, alongside the maximum score for each system. Note that most systems performed well against the standard, conll2007 and functional references but poorly against the lexical and oldLTH references.[18] Considering the latter two references, a different system would be selected as the highest performing, namely Bisk and Hockenmaier (2012) (BH) over Blunsom and Cohn (2010) (BC) which wins in the other cases.

[18]The common difference here is that the latter two references do not treat prepositions as heads of PPs.

This evaluation method rewards many different linguistically plausible structures, but in such a way that the predictions must be consistent between different sentences in the testing set, and in their treatment of related linguistic phenomena. One caveat is that this method can only be used when there are many references, although in many cases different outputs can be generated automatically, e.g., by adjusting head-finding heuristics in converting between phrase-structure to dependency trees.

The previous analysis has rated each system in terms of overall performance against treebank trees, however this doesn't necessarily mean that the predictions of the best ranked system will be the most useful ones in a task-based setting. Take the example of information extraction, in which a central problem is to identify the arguments (subject, object *etc*) of a given verb. This setting gives rise to some types of dependency edges being more valuable than others. We present comparative results for the Penn treebank in Table 6 showing the directed accuracy for different types of dependency relations. Observe that there is a wide spread of accuracies for predicting the head word of the sentence (ROOT), and similarly for verbs' subject and object arguments. These scores are similar to the scores for the local modifiers shown, such as NMOD which describe the arguments of a noun. This is surprising as noun edges tend to be much shorter than for the arguments to a verb, and thus should be easier to predict. Also interesting are the spread of results for the CC edges (these link a coordinating conjunction to its head), suggesting that the systems learn to represent coordination in very different ways to the method used in the reference.

Figure 1 illustrates the directed accuracy over different lengths of dependency edge. For all systems the accuracy diminishes with edge length, however some fall at a much faster rate. The two best systems (Tu, BC) have similar overall accuracy results, but it is clear that Tu does better on short edges while BC does better on longer ones. The same pattern was also observed when considering the average accuracy over all treebanks (not shown), although the systems' results were closer together.

system	ROOT	SBJ	OBJ	PRD	NMOD	COORD	CC
Tu	71.0	64.8	**53.7**	49.4	56.9	36.8	11.4
LB	17.8	40.1	15.3	18.0	41.9	27.7	9.7
BC	**74.9**	**65.7**	53.0	50.2	56.8	36.3	**71.4**
DMV^c	17.0	11.7	16.0	31.3	27.8	25.7	9.2
DMV^u	17.6	9.3	16.4	25.0	27.8	25.7	8.6
BH	67.5	55.3	44.9	45.6	**58.6**	27.6	62.7
M^u	29.3	42.4	38.8	51.8	34.5	30.5	33.0
R	12.9	9.4	16.1	21.1	12.1	15.7	2.7
M^c	60.7	47.4	39.9	45.8	36.5	33.9	44.3
RB	17.9	12.4	26.2	36.5	15.3	25.4	1.1
H	19.4	29.3	12.2	22.2	17.3	20.9	10.3
DMV^p	54.7	42.0	30.7	30.1	28.9	25.4	24.3
S^2	45.2	41.9	44.2	49.8	39.7	25.4	63.8
M^p	67.8	54.3	49.6	**59.4**	47.7	**37.7**	49.7
S^1	43.1	47.9	36.3	46.7	27.9	23.5	7.6

Table 6: Directed accuracy results on the Penn treebank, stratified by dependency relation. For clarity, only 9 important relation types are shown. The vertical bars separate different groups of relations, from left to right, relating to the main verb, general modifiers and coordination.

7 Conclusion

This challenge set out to evaluate the state-of-the-art in part-of-speech and dependency grammar induction, promoting research in this field and, importantly, providing a fair means of evaluation. The participants submissions used a wide variety of different approaches, many of which we shown to have improved over competitive benchmark systems. While the results were overall very positive, it is fair to say that the tasks of part-of-speech and grammar induction are still very much open challenges, and that there is still considerable room for improvement. The data submitted to this evaluation campaign will provide a great resource for devising new methods of evaluation, and we plan to pursue this avenue in future work, in particular task-based evaluation such as in an information extraction or machine translation setting.

8 Acknowledgements

This challenge was funded by the PASCAL 2 (Pattern Analysis, Statistical Modelling and Computational Intelligence) European Network of Excellence. We would also like to thank the treebank providers for allowing us to use their resources, assisting us in converting these into our desired format, and helping to resolve various questions. In particular, special thanks to Zdenek Zabokrtsky and Jan (Czech and Arabic), Tomaz Erjavec (Slovene), and Eckhard Bick and Diana Santos (Portuguese). We are also indebted to the organisers of the previ-

testset	BC	BH	MZ^c	MZ^p	MZ^u	S^1	S^2	Tu	DMV^c	DMV^p	DMV^u	H	LB	RB
conll2007	**54.9**	51.7	40.4	49.2	36.8	32.2	41.7	54.2	20.9	33.2	20.4	18.0	30.1	20.3
functional	**59.6**	52.4	41.5	47.4	36.2	30.6	40.0	58.5	20.9	37.2	20.6	19.3	29.2	23.7
lexical	40.6	**41.9**	28.5	37.3	24.8	27.7	35.5	39.5	23.5	23.1	23.0	14.4	33.1	10.1
oldLTH	41.4	**43.6**	28.8	37.8	24.6	28.6	36.1	39.5	22.3	23.7	21.8	14.3	32.0	10.7
standard	**56.0**	50.4	41.0	50.3	37.5	32.8	42.5	55.5	22.3	33.5	21.8	18.4	31.4	20.4
best	**59.6**	52.4	41.5	50.3	37.5	32.8	42.5	58.5	23.5	37.2	23.0	19.3	33.1	23.7

Table 5: Directed accuracy results measured against different conversions of the Penn Treebank into dependency trees.

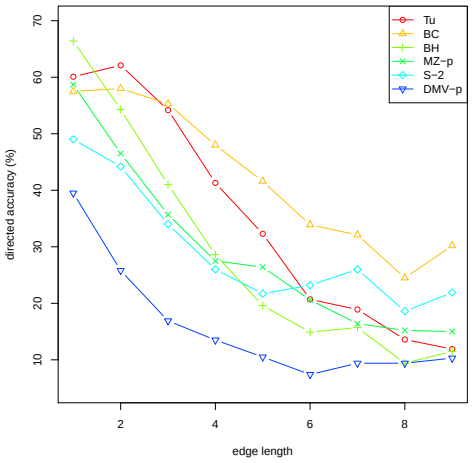

Figure 1: Directed accuracy on the Penn treebank stratified by dependency length. For clarity only a subset of the systems are shown, and edges of length 10 or more were omitted.

ous CoNLL 2006 and 2007 competitions, who contributed significant efforts into collating so many treebanks and developing treebank conversion tools, making our job much easier than it would otherwise have been. Thanks to Sebastian Reidel, Joakim Nivre and Sabine Buchholz for promptly answering our questions. We would like to thank the LDC, who allowed their licenced data to be used free of charge by the competitors, and Ilya Ahtaridis who administered the licencing and corpus distribution. Thanks also to Valentin Spitkovski and Christos Christodoulopoulos who kindly provided us with their evaluation scripts, and finally, the participants themselves for taking part.

References

I. Aduriz, M. J. Aranzabe, J. M. Arriola, A. Atutxa, A. Diaz de Ilarraza, A. Garmendia, and M. Oronoz. 2003. Construction of a Basque dependency treebank. In *Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories (TLT)*.

Susana Afonso, Eckhard Bick, Renato Haber, and Diana Santos. 2002. Floresta sintá(c)tica: a treebank for Portuguese. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1698–1703, May.

Yonatan Bisk and Julia Hockenmaier. 2012. Induction of linguistic structure with combinatory categorial grammars. In *Proceedings of the NAACL-HLT 2012 Workshop on Inducing Linguistic Structure Shared Task*, June.

Phil Blunsom and Trevor Cohn. 2010. Unsupervised induction of tree substitution grammars for dependency parsing. In *Proceedings of the 2010 Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pages 1204–1213, Cambridge, MA, USA.

Phil Blunsom and Trevor Cohn. 2011. A hierarchical Pitman-Yor process HMM for unsupervised part of speech induction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 865–874, Portland, Oregon, USA, June.

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2001. The Prague Dependency Treebank: A Three-Level Annotation Scenario. In Anne Abeillé, editor, *Treebanks: Building and Using Syntactically Annotated Corpora*, pages 103–127. Kluwer Academic Publishers.

Gosse Bouma, Gertjan van Noord, and Robert Malouf. 2000. Alpino: Wide coverage computational analysis of Dutch. In *Proceedings of Computational Linguistics in the Netherlands (CLIN 2000)*, pages 45–59.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Comput. Linguist.*, 18:467–479, December.

Matthias Buch-Kromann, Jürgen Wedekind, and Jakob Elming. 2007. The Copenhagen Danish-English dependency treebank. `http://code.google.com/p/copenhagen-dependency-treebank`.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In

Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X), pages 149–164.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2010. Two decades of unsupervised POS induction: how far have we come? In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 575–584.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2011. A Bayesian mixture model for PoS induction using multiple features. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 638–647, Edinburgh, Scotland, UK., July.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2012. Turning the pipeline into a loop: Iterated unsupervised dependency parsing and pos induction. In *Proceedings of the NAACL-HLT 2012 Workshop on Inducing Linguistic Structure Shared Task*, June.

Grzegorz Chrupała. 2012. Hierarchical clustering of word class distributions. In *Proceedings of the NAACL-HLT 2012 Workshop on Inducing Linguistic Structure Shared Task*, June.

Tomaž Erjavec, Darja Fišer, Simon Krek, and Nina Ledinek. 2010. The JOS linguistically tagged corpus of Slovene. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

João Graça, Kuzman Ganchev, Luísa Coheur, Fernando Pereira, and Benjamin Taskar. 2011. Controlling complexity in part-of-speech induction. *J. Artif. Intell. Res. (JAIR)*, 41:527–551.

Aria Haghighi and Dan Klein. 2006. Prototype-driven learning for sequence models. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL '06)*, pages 320–327.

Jan Hajič, Otakar Smrž, Petr Zemánek, Jan Šnaidauf, and Emanuel Beška. 2004. Prague Arabic dependency treebank: Development in data and tools. In *Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools*, pages 110–117.

Richard Johansson and Pierre Nugues. 2007. Extended constituent-to-dependency conversion for English. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*.

Mark Johnson. 2007. Why doesn't EM find good hmm pos-taggers. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '07, pages 296–305.

Dan Klein and Christopher D. Manning. 2004. Corpus-based induction of syntactic structure: models of dependency and constituency. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 478.

David Maraček and Zdeněk Žabokrtský. 2012. Unsupervised dependency parsing using reducibility and fertility features. In *Proceedings of the NAACL-HLT 2012 Workshop on Inducing Linguistic Structure Shared Task*, June.

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

Marina Meila. 2003. Comparing Clusterings by the Variation of Information. *Learning Theory and Kernel Machines*, pages 173–187.

Tahira Naseem, Harr Chen, Regina Barzilay, and Mark Johnson. 2010. Using universal linguistic knowledge to guide grammar induction. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1234–1244, October.

Joakim Nivre, Jens Nilsson, and Johan Hall. 2006. Talbanken05: A Swedish treebank with phrase structure and dependency annotation. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)*.

Joakim Nivre, Johan 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 CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932, June.

Franz Josef Och. 1999. An efficient method for determining bilingual word classes. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 71–76.

Slav Petrov, Dipanjan Das, and Ryan T. McDonald. 2011. A universal part-of-speech tagset. *CoRR*, abs/1104.2086.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420.

K. Sagae, E. Davis, A. Lavie, B. MacWhinney, and S. Wintner. 2007. High-accuracy annotation and parsing of CHILDES transcripts. In *Proceedings of the ACL-2007 Workshop on Cognitive Aspects of Computational Language Acquisition.*, June.

Roy Schwartz, Omri Abend, Roi Reichart, and Ari Rappoport. 2011. Neutralizing linguistically problematic

annotations in unsupervised dependency parsing evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 663–672.

Anders Søgaard. 2012. Two baselines for unsupervised dependency parsing. In *Proceedings of the NAACL-HLT 2012 Workshop on Inducing Linguistic Structure Shared Task*, June.

Reut Tsarfaty, Joakim Nivre, and Evelina Andersson. 2011. Evaluating dependency parsing: Robust and heuristics-free cross-annotation evaluation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 385–396, Edinburgh, UK, July.

Kewei Tu. 2012. Combining the sparsity and unambiguity biases for grammar induction. In *Proceedings of the NAACL-HLT 2012 Workshop on Inducing Linguistic Structure Shared Task*, June.

David Vadas and James R. Curran. 2007. Adding noun phrase structure to the Penn Treebank. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, pages 240–247, June.

Appendix

Directed				
testset	CGS	DMV^c	DMV^p	DMV^u
arabic	36.1	42.6	**51.9**	49.1
basque	28.4	28.9	27.1	**30.2**
czech	33.1	28.6	**38.2**	28.3
danish	27.9	36.4	**38.4**	18.2
dutch	31.0	39.0	**41.1**	40.3
en-childes	31.2	40.8	**44.3**	42.1
en-ptb	22.7	**25.1**	23.1	23.1
portuguese	26.7	**38.4**	34.5	31.1
slovene	**26.3**	20.6	19.2	22.6
swedish	29.0	**30.9**	**30.9**	26.5
averages	29.3	33.1	**34.9**	31.1
Undirected				
testset	CGS	DMV^c	DMV^p	DMV^u
arabic	58.3	52.8	58.3	**61.1**
basque	49.3	49.2	**50.5**	50.0
czech	48.7	45.9	**57.2**	47.9
danish	56.3	**60.9**	57.0	43.7
dutch	47.0	53.6	**57.2**	53.8
en-childes	56.3	61.0	**62.7**	59.9
en-ptb	50.7	52.9	**54.1**	46.9
portuguese	51.8	**61.1**	59.4	51.6
slovene	40.5	41.9	**45.3**	41.2
swedish	52.5	**57.1**	**57.1**	48.6
averages	51.1	53.6	**55.9**	50.5
NED				
testset	CGS	DMV^c	DMV^p	DMV^u
arabic	62.0	63.0	66.7	**67.6**
basque	**67.4**	62.8	62.5	62.3
czech	65.1	60.7	**72.0**	64.0
danish	72.0	72.4	**73.2**	60.3
dutch	57.7	64.9	**65.0**	64.7
en-childes	**79.9**	79.6	79.9	78.4
en-ptb	67.9	73.4	**74.3**	63.7
portuguese	58.2	**80.7**	79.5	72.6
slovene	55.7	52.3	**59.4**	51.9
swedish	64.8	**78.4**	**78.4**	65.7
averages	65.1	68.8	**71.1**	65.1

Table 9: Evaluation of the joint task on the dependency output using a maximum sentence length of 10.

M-1												
testset	BC	CGS	C^c	C^p	G^c	G^u	B^c	B^p	B^u	MK^c	MK^p	MK^u
arabic	75.06	N/A	**79.00**	**79.00**	62.20	62.24	61.81	64.60	61.55	65.69	67.82	63.23
basque	71.58	69.20	**75.23**	**75.23**	65.97	56.37	64.37	68.58	62.17	63.59	68.22	60.49
czech	74.84	61.53	66.97	**76.00**	N/A	N/A	55.60	72.51	55.01	60.38	73.38	51.96
danish	56.48	55.41	70.28	**71.62**	49.24	35.32	50.21	66.50	33.57	49.40	61.60	35.02
dutch	80.72	70.13	74.04	76.08	63.37	57.64	57.99	83.76	57.31	68.18	**84.64**	63.04
en-childes	84.23	77.57	85.35	85.35	N/A	N/A	76.34	85.11	59.75	77.55	**86.39**	59.55
en-ptb	**78.26**	72.26	62.86	73.46	63.15	56.32	65.10	68.10	48.76	70.31	73.96	47.91
portuguese	76.00	72.05	75.47	**77.13**	68.40	65.86	65.52	69.61	61.84	64.63	66.81	62.95
slovene	67.29	59.78	72.18	**72.71**	63.95	55.23	54.85	63.68	52.15	54.08	59.31	45.40
swedish	66.20	67.86	**73.55**	**73.55**	60.10	48.43	61.21	61.21	47.51	64.39	64.39	46.04
averages	73.07	67.31	73.49	**76.01**	62.05	54.68	61.30	70.37	53.96	63.82	70.65	53.56

1-1												
testset	BC	CGS	C^c	C^p	G^c	G^u	B^c	B^p	B^u	MK^c	MK^p	MK^u
arabic	50.90	N/A	39.15	39.15	41.49	**53.92**	39.89	34.23	40.63	50.53	41.98	50.27
basque	42.45	46.25	**52.38**	**52.38**	48.91	45.55	41.29	33.49	50.80	43.27	35.73	43.43
czech	31.45	**48.24**	32.18	31.74	N/A	N/A	43.12	33.55	41.94	38.93	28.33	35.31
danish	43.08	**43.64**	32.17	31.77	40.56	34.83	33.48	30.87	26.33	38.92	30.59	32.95
dutch	43.22	55.85	43.26	39.98	**56.45**	45.37	48.13	21.88	43.10	55.86	22.42	54.04
en-childes	64.10	63.62	**64.50**	**64.50**	N/A	N/A	59.96	56.87	59.75	63.43	53.40	57.68
en-ptb	**57.63**	56.02	45.52	41.35	48.95	53.15	55.43	49.60	47.57	54.10	51.80	45.43
portuguese	59.71	50.18	36.13	35.38	54.42	**60.08**	49.57	45.00	48.25	46.57	38.37	45.10
slovene	42.62	50.66	33.23	32.59	**56.55**	50.30	44.97	44.34	40.62	41.24	40.01	38.93
swedish	**48.76**	40.54	34.07	34.07	38.21	46.32	36.12	36.12	44.57	41.90	41.90	38.05
averages	48.39	**50.55**	41.26	40.29	48.19	48.69	45.20	38.59	44.36	47.48	38.45	44.12

VM												
testset	BC	CGS	C^c	C^p	G^c	G^u	B^c	B^p	B^u	MK^c	MK^p	MK^u
arabic	**61.59**	N/A	52.95	52.95	46.99	47.18	40.75	43.16	40.40	47.95	45.09	45.14
basque	53.83	51.34	**54.45**	**54.45**	49.34	44.26	44.18	45.46	45.37	45.02	44.90	42.76
czech	**56.80**	50.22	45.06	46.76	N/A	N/A	42.50	49.93	41.51	45.15	51.38	39.62
danish	61.57	59.00	63.39	**63.62**	53.35	43.20	51.83	58.38	33.46	52.52	58.44	39.46
dutch	**57.82**	53.94	55.01	53.40	48.99	46.26	44.08	50.49	44.37	52.02	51.33	47.99
en-childes	**80.17**	76.59	78.18	78.18	N/A	N/A	73.67	76.47	65.44	76.14	76.87	68.25
en-ptb	**71.44**	68.12	59.90	61.31	63.90	60.04	63.79	63.64	52.96	66.40	66.50	54.33
portuguese	**67.49**	60.37	54.61	54.74	58.91	59.58	53.30	54.99	50.26	53.15	52.67	50.76
slovene	**54.80**	52.13	51.85	51.88	52.99	48.55	45.33	48.33	40.13	39.25	45.73	38.68
swedish	**61.52**	58.23	56.09	56.09	55.02	48.69	51.76	51.76	43.39	54.28	54.28	44.51
averages	**62.70**	58.88	57.15	57.34	53.69	49.72	51.12	54.26	45.73	53.19	54.72	47.15

VI												
testset	BC	CGS	C^c	C^p	G^c	G^u	B^c	B^p	B^u	MK^c	MK^p	MK^u
arabic	2.65	N/A	3.76	3.76	3.47	3.19	3.96	4.12	3.73	3.49	3.96	3.48
basque	3.65	3.50	3.78	3.78	3.45	**3.36**	4.09	4.79	3.67	4.00	4.72	3.83
czech	3.80	3.49	4.96	5.48	N/A	N/A	3.92	4.57	3.91	3.85	4.46	4.17
danish	**3.76**	3.85	4.07	4.08	4.29	4.53	4.54	4.91	5.24	4.45	4.78	4.85
dutch	3.53	**3.14**	3.31	3.72	3.33	3.47	3.68	5.16	3.61	3.26	5.07	3.39
en-childes	1.86	2.12	2.11	2.11	N/A	N/A	2.39	2.32	2.59	2.17	2.31	2.47
en-ptb	**2.69**	2.90	3.67	4.03	3.16	3.08	3.24	3.41	3.66	3.05	3.19	3.50
portuguese	**2.40**	2.90	4.01	4.11	2.97	2.74	3.35	3.46	3.35	3.40	3.63	3.37
slovene	3.65	3.40	4.65	4.68	**3.34**	3.48	3.92	4.23	4.03	4.38	4.51	4.25
swedish	**3.36**	3.78	4.57	4.57	3.89	3.65	4.45	4.45	4.11	4.17	4.17	4.07
averages	**3.13**	3.23	3.89	4.03	3.49	3.44	3.75	4.14	3.79	3.62	4.08	3.74

Table 7: One to one, Many to one, VM and VI scores of POS induction results evaluated against fine POS tags (c.f., Table 2 which used UPOS).

Directed														
testset	BC	BH	MZ^c	MZ^p	MZ^u	S^1	S^2	Tu	DMV^c	DMV^p	DMV^u	H	LB	RB
arabic	61.1	47.2	15.7	64.8	64.8	47.2	54.6	**66.7**	46.3	45.4	50.0	42.6	9.3	64.8
basque	56.3	50.3	28.0	30.3	27.2	33.3	22.3	**58.6**	46.3	43.2	31.3	21.8	34.3	24.4
czech	50.0	48.5	**61.3**	57.5	57.7	45.5	51.2	59.0	30.1	31.2	31.8	24.7	28.9	34.3
danish	46.2	49.3	60.2	51.3	**61.4**	56.9	60.5	60.8	47.2	50.2	35.3	36.4	18.7	49.2
dutch	50.5	50.8	37.0	49.5	38.4	38.9	50.0	**51.7**	48.7	39.7	49.1	35.1	34.0	39.5
en-childes	48.1	**62.2**	56.8	47.2	51.8	50.5	53.5	56.0	51.7	51.9	39.0	31.7	36.0	23.3
en-ptb	72.1	73.7	58.9	67.4	52.2	44.8	61.0	**74.7**	31.7	44.7	30.6	35.2	40.4	19.9
portuguese	54.3	**76.3**	63.6	59.9	44.3	47.7	71.1	55.7	27.1	37.2	26.9	31.1	28.1	37.7
slovene	65.8	53.9	42.1	51.4	39.2	39.7	50.3	**67.7**	35.7	37.2	35.6	25.7	35.9	14.7
swedish	65.8	66.7	61.4	63.7	70.8	48.2	72.0	**76.5**	44.2	44.2	45.8	39.1	33.2	31.3
averages	57.0	57.9	48.5	54.3	50.8	45.3	54.7	**62.7**	40.9	42.5	37.5	32.3	29.9	33.9

Undirected														
testset	BC	BH	MZ^c	MZ^p	MZ^u	S^1	S^2	Tu	DMV^c	DMV^p	DMV^u	H	LB	RB
arabic	**69.4**	59.3	59.3	**69.4**	**69.4**	59.3	65.7	67.6	52.8	53.7	55.6	54.6	61.1	**69.4**
basque	65.6	59.5	49.8	50.1	48.4	54.3	32.8	**66.4**	60.1	58.1	48.5	42.2	56.4	53.9
czech	65.9	59.9	69.2	66.6	67.6	62.3	63.5	**70.1**	51.6	51.2	50.5	47.2	54.2	56.9
danish	67.9	63.5	70.6	64.4	**71.1**	67.9	70.7	70.2	65.0	64.3	60.0	56.4	59.7	63.9
dutch	63.2	59.9	57.5	**63.3**	58.0	58.5	58.5	60.5	62.7	56.7	62.9	51.1	56.3	60.7
en-childes	65.3	**70.6**	69.4	62.4	63.7	64.3	63.6	69.1	65.7	66.1	59.5	51.2	51.2	51.0
en-ptb	**79.4**	79.2	65.9	72.5	62.4	62.5	75.1	78.8	53.8	65.3	53.2	52.0	58.8	54.9
portuguese	66.3	**81.9**	71.6	70.2	62.3	65.8	78.5	72.1	54.0	60.9	54.3	53.3	56.7	63.8
slovene	70.6	63.7	56.3	59.1	55.1	54.8	63.7	**72.0**	46.4	53.1	46.3	44.9	45.5	46.0
swedish	82.3	73.5	70.1	71.1	75.4	66.5	77.3	**83.7**	64.5	64.5	66.1	59.2	59.2	59.5
averages	69.6	67.1	64.0	64.9	63.3	61.6	64.9	**71.0**	57.7	59.4	55.7	51.2	55.9	58.0

NED														
testset	BC	BH	MZ^c	MZ^p	MZ^u	S^1	S^2	Tu	DMV^c	DMV^p	DMV^u	H	LB	RB
arabic	**78.7**	68.5	66.7	75.9	75.9	68.5	72.2	71.3	61.1	63.0	63.0	63.0	64.8	75.9
basque	77.9	68.6	62.9	65.2	64.2	70.1	55.0	**79.5**	69.3	69.0	63.8	58.2	73.2	64.0
czech	79.9	72.6	**81.8**	78.6	79.9	76.0	78.1	81.2	62.9	61.7	63.6	60.3	63.2	73.8
danish	81.7	76.3	83.2	78.4	84.3	77.3	84.4	**85.0**	77.9	75.8	69.5	67.8	66.2	77.5
dutch	71.0	71.8	77.1	**78.4**	76.8	72.6	68.5	71.0	73.2	64.6	73.0	60.5	64.4	71.0
en-childes	82.4	81.6	84.5	76.7	74.8	79.0	**84.9**	82.5	79.9	80.4	79.9	70.1	63.2	76.9
en-ptb	86.7	**89.4**	87.5	88.4	84.0	78.7	87.1	84.3	64.0	80.7	64.2	64.3	65.2	75.0
portuguese	78.2	90.7	87.8	87.8	87.5	82.9	**91.9**	90.2	75.6	81.7	76.5	67.7	60.9	82.4
slovene	79.3	76.8	70.8	72.8	70.4	68.1	78.9	**79.8**	59.4	62.2	59.4	55.8	54.3	62.2
swedish	91.7	85.8	83.1	85.6	87.1	80.8	87.6	**92.1**	76.1	76.1	76.4	75.3	67.2	79.1
averages	80.8	78.2	78.5	78.8	78.5	75.4	78.9	**81.7**	69.9	71.5	68.9	64.3	64.3	73.8

Table 8: Evaluation of the dependency task using a maximum sentence length of 10. See also Table 3 which presents the same results with no length restriction.