Multi-lingual Dependency Parsing Evaluation: a Large-scale Analysis of Word Order Properties using Artificial Data

Kristina Gulordava and **Paola Merlo**
Department of Linguistics
University of Geneva
5 Rue de Candolle, CH-1211 Genève 4
kristina.gulordava@unige.ch, paola.merlo@unige.ch

Abstract

The growing work in multi-lingual parsing faces the challenge of fair comparative evaluation and performance analysis across languages and their treebanks. The difficulty lies in teasing apart the properties of treebanks, such as their size or average sentence length, from those of the annotation scheme, and from the linguistic properties of languages. We propose a method to evaluate the effects of word order of a language on dependency parsing performance, while controlling for confounding treebank properties. The method uses artificially-generated treebanks that are minimal permutations of actual treebanks with respect to two word order properties: word order variation and dependency lengths. Based on these artificial data on twelve languages, we show that longer dependencies and higher word order variability degrade parsing performance. Our method also extends to minimal pairs of individual sentences, leading to a finer-grained understanding of parsing errors.

1 Introduction

Fair comparative performance evaluation across languages and their treebanks is one of the difficulties for work on multi-lingual parsing (Buchholz and Marsi, 2006; Nivre et al., 2007; Seddah et al., 2011). The differences in parsing performance can be the result of disparate properties of treebanks (such as their size or average sentence length), choices in annotation schemes, and the linguistic properties of languages. Despite recent attempts to create and apply cross-linguistic and cross-framework evaluation procedures (Tsarfaty et al., 2011; Seddah et al., 2013), there is no commonly used method of analysis of parsing performance which accounts for different linguistic and extra-linguistic factors of treebanks and teases them apart.

When investigating possible causal factors for observed phenomena, one powerful method, if available, consists in intervening on the postulated causes to observe possible changes in the observed effects. In other words, if A causes B, then changing A or properties of A should result in an observable change in B. This interventionist approach to the study of causality creates counterfactual data and a type of controlled modification that is wide-spread in experimental methodology, but that is not widely used in fields that rely on observational data, such as corpus-driven natural language processing.

In analyses of parsing performance, it is customary to manipulate and control word-level features, such as part-of-speech tags or morphological features. These types of features can be easily omitted or modified to assess their contribution to parsing performance. However, higher-order features, such as linear word order precedence properties, are much harder to define and to manipulate. A parsing performance analysis based on controlled modification of word order, in fact, has not been reported previously. We propose such a method based on word order permutations which allows us to manipulate word order properties analogously to familiar word-level properties and study their effect on parsing performance.

Specifically, given a dependency treebank, we obtain new synthetic data by permuting the original order of words in the sentences, keeping the unordered

Transactions of the Association for Computational Linguistics, vol. 4, pp. 343–356, 2016. Action Editor: Joakim Nivre.
Submission batch: 1/2016; Revision batch: 4/2016; Published 7/2016.

dependency tree constant. These permuted sentences are not necessarily grammatical in the original language. They constitute an alternative "language" which forms a minimal pair with the original one, where the only changed property is the order of words, but all the other properties of the unordered tree and the confounding variables between the two datasets are kept constant, such as the size of the training data, the average sentence length, the number of PoS tags and the dependency labels.

We perform two types of word order permutations to the treebanks in our sample: a permutation which minimises the lengths of the dependencies in a dependency tree and a permutation which minimises the variability of word order. We then compare how the parsing performances on the original and the permuted trees vary in relation to the quantified measures of the dependency length and word order variation properties of the treebanks. To quantify dependency length, we use the ratio of minimisation of the length of dependencies between words in the tree (dependency length minimisation, DLM (Gildea and Temperley, 2010)). To quantify the property intuitively referred to as variability of word order, we use the entropy of the linear precedence ordering between a head and a child in dependency arcs (Liu, 2010).

The reason to concentrate on these two word order properties comes from previous parsing results. Morphologically-rich languages are known to be hard for parsing, as rich morphology increases the percentage of new words in the test set (Nivre et al., 2007; Tsarfaty et al., 2010). These languages however also often exhibit very flexible word order. It has not so far been investigated how much rich morphology contributes to parsing difficulty compared to the difficulty introduced by word order variation in such languages. The length of the dependencies in the tree has also been shown to affect performance: almost all types of dependency parsers, in different measure, show degraded performance for longer sentences and longer dependencies (McDonald and Nivre, 2011).[1] We use arc direction entropy and DLM ratio, respectively, as the measures of these two word order properties because they are formally defined in the previous literature and can be quantified on a dependency treebank in any language.

To preview our results, in a set of pairwise comparisons between original and permuted treebanks, we confirm the influence of word order variability and dependency length on parsing performance, at the large scale provided by fourteen different treebanks across twelve different languages.[2] Our results suggest, in addition, that word order entropy applies a stronger negative pressure on parsing performance than longer dependencies. Finally, on an example of one treebank, we show how our method can be extended to provide finer-grained analyses at the sentence level and relate the parsing errors to properties of the parsing architecture.

2 Parsing analysis using synthetic data

In this section, we introduce our new approach to using synthetic data for cross-linguistic analysis of parsing performance.

2.1 Methodology

Our experiments with artificial data consist in modifying a treebank T to create its minimal pair T' and evaluating parsing performance by comparing these pairs of treebanks. We create several kinds of artificial treebanks in the same manner: each sentence in T' is a permutation of the words of the original sentence in T. We permute words in various ways according to the word order property whose effect on parsing we want to analyse. Importantly, we only change the order of words in a sentence. In contrast, the dependency tree structure of a permuted sentence in T' is the same as in the original sentence in T.

For each treebank in our sample of languages and a type of permutation, we conduct two parsing evaluations: $T_{Train} \rightarrow T_{Test}$ and $T'_{Train} \rightarrow T'_{Test}$. The training-test data split for T and T' is always the same, that is $T'_{Train} = Permuted(T_{Train})$ and $T'_{Test} = Permuted(T_{Test})$. The parsing performance is measured as Unlabeled and Labeled Attachment Scores (UAS and LAS), the proportion of correctly attached arcs in the unlabelled or labelled tree, respectively.

[1] But see Titov and Henderson (2007) for an exception and comparison to Malt.

[2] Polish, Italian, Finnish, Spanish, French, English, Bulgarian, Latin (Vulgate, Cicero), Dutch, Ancient Greek (New Testament, Herodotus), German and Persian.

Given the training-testing setup, the differences in unlabelled attachment scores $UAS(T_{Test}) - UAS(T'_{Test})$ can be directly attributed to the differences in word order properties o between T and T', abstracting away from other treebank properties h. More formally, we assume that $UAS(T) = f(o^T, h^T)$ and $UAS(T') = f(o^{T'}, h^T)$. Except for word order properties o^T and $o^{T'}$, the two equations share all other treebank properties h^T — such as size, average dependency length, size of PoS tagset — and f is a function that applies to all languages, here embodied by a given parser.

Our method can be further extended to analyse parsing performance at the sentence level. Consider the pair consisting of a sentence in an original treebank and its correspondence in a permuted treebank. The two sentences share all lexical items and underlying dependencies between them: the explanation for different parsing accuracies must be sought therefore in their different word orders. In standard treebank evaluation settings, instead, exact sentence-level comparisons are not possible, as two sentences very rarely constitute a truly minimal pair with respect to any specific syntactic property. Our approach opens up the possibility of deeper understanding of parsing behaviour at the sentence-level and even of individual dependencies based on large sets of minimal pairs.

2.2 Word order properties

To be able to compare parsing performance across the actual and the synthetic data, we must manipulate the causal property we want to study. In this work, we concentrate on variability of word order and length of dependencies. We define and discuss these two properties and their measures below.

Arc direction entropy One dimension that can greatly affect parsing performance across languages is word order freedom, the ability languages have to express the same or similar meaning in the same context with a free choice of different word orders. The extent of word order freedom in a sentence is reflected in the entropy of word order, given the words and the syntactic structure of the sentence, $H(order|words, tree)$.

One approximation of word order entropy is the entropy of the direction of dependencies in a treebank. This measure has been proposed in several recent works to quantitatively describe the typology of word order freedom in many languages (Liu, 2010; Futrell et al., 2015b).

Arc direction entropy can be used, for instance, to capture the difference between adjective-noun word order properties in Germanic and Romance languages. In English, this word order is fixed, as adjectives appear almost exclusively prenominally; the adjective-noun arc direction entropy will therefore be close to 0. In Italian, by contrast, the same adjective can both precede and follow nouns; the adjective-noun arc direction entropy will be greater than 0.

We calculate the overall entropy of arc directions in a treebank conditioned on the relation type defined by the dependency label Rel and the PoS tags of the head H and the child C:

$$H(Dir|Rel, H, C) = \sum_{rel,h,c} p(rel, h, c) H(Dir|rel, h, c) \quad (1)$$

Dir in (1) is the order between the child and the head in the dependency arc (Left or Right). In other words, we compute the entropy of arc direction $H(Dir) = -p(L) \cdot \log p(L) - p(R) \cdot \log p(R)$ for each observed tuple (rel, h, c) independently and weigh them according to the tuple frequency in the corpora.

DLM ratio Another property that has been shown to affect parsing performance across languages and across parsers is the length of the dependencies in the tree.[3] A global measure of average dependency length of a whole treebank has been proposed in the literature on dependency length minimisation (DLM). This measure allows comparisons across treebanks with sentences of different size and across dependency trees of different topology.

Experimental and theoretical language research has yielded a large and diverse body of evidence showing that languages, synchronically and diachronically, tend to minimise the length of their dependencies (Hawkins, 1994; Gibson, 1998; Demberg and Keller, 2008; Tily, 2010; Gulordava and

[3]The length of a dependency, $DL(arc)$ below, is the number of words in the span covered by the dependency arc.

Merlo, 2015; Gulordava et al., 2015). Languages differ, however, in the degree to which they minimise dependencies. A low degree of DLM is associated with flexibility of word order and in particular with high non-projectivity, i.e., the presence of crossing arcs in a tree, a feature that has been treated in dependency parsing using local word order permutations (Hajičová et al., 2004; Nivre, 2009; Titov et al., 2009; Henderson et al., 2013). To estimate the degree of DLM in a language, we follow previous work which analysed the dependency lengths in a treebank with respect to their random and minimal potential alternatives (Temperley, 2007; Gildea and Temperley, 2010; Futrell et al., 2015a; Gulordava and Merlo, 2015).

We calculate the overall ratio of DLM in a treebank as shown in equation 2.

$$DLMRatio = \Sigma_s \frac{DL_s}{|s|^2} / \Sigma_s \frac{OptDL_s}{|s|^2} \quad (2)$$

For each sentence s and its dependency tree t, we compute the overall dependency length of the original sentence $DL(s) = \sum_{arc \in t} DL(arc)$ and its minimal projective dependency length $OptDL(s) = DL(s')$, where s' is obtained by reordering the words in the sentence s using the algorithm described in the next section (following Gildea and Temperley (2010)). To average these values across all sentences, we normalise them by $|s|^2$, since it has been observed empirically that the relation between the dependency lengths DL and $OptDL$ and the length $|s|$ of a sentence is not linear, but rather quadratic (Ferrer-i-Cancho and Liu, 2014; Futrell et al., 2015a).[4]

In the next section, we illustrate how we create two pairs of (T, T') treebanks, manipulating the two word order properties just discussed.

3 Word order permutations

We create two types of permuted treebanks to optimise for the two word order parameters considered in the previous section.

3.1 Creating trees with optimal DL

Given a sentence s and its dependency tree t in a natural language, we employ the algorithm proposed by Gildea and Temperley (2010) to create a new artificial sentence s' with a permuted order of words. The algorithm reorders the words in a sentence s to yield the projective dependency tree with the minimal overall dependency length $DL(s')$.[5] To do so, it recursively places the children on the left and on the right of the head in alternation, so that the children on the same side of the head are ordered based on their sizes — shortest phrases closer to the head. Children of the same size are ordered between each other as found in the original sentence.

This algorithm is deterministic and the dependency length of each sentence is optimised independently. We exclude from our analysis sentences with any non-final punctuation tokens and sentences with multiple roots. By definition, the DLM ratio for sentences permuted in such a way is equal to 1.

3.2 Creating trees with optimal Entropy

To obtain treebanks with a minimal arc direction entropy equal to zero, we can fix the order of each type of dependency, defined by a tuple (rel, h, c). There exist therefore many possible permutations resulting in zero arc direction entropy. We choose to assign the same direction (either Left or Right) to all the dependencies. This results in two permutations yielding fully right-branching (RB) and fully left-branching (LB) treebanks. We order the children on the same side of a head in the same way as in the OptDL permutation: the shortest children are closest to the head. For RB permutation, children of the same size are kept in the order of the original sentence; for LB permutation, this order is reversed, so that the RB and LB orders are symmetrical. These two permutations are particularly interesting, as they give us the two extremes in the space of possible tree-branching structures. Moreover, since the LB/RB word orders for each sentence are completely symmetrical, the two treebanks

[4] We follow previous work in using $DL(s)$ as the measure for DLM ratio calculation. Equivalently, we could use the average length of a single dependency $\langle DL(arc) \rangle$. Given that $\langle DL(s) \rangle = |s| \cdot \langle DL(arc) \rangle$, the fact that $\langle DL(s) \rangle \sim |s|^2$ can be more naturally stated as $\langle DL(arc) \rangle \sim |s|$: the average length of a single dependency is linear with respect to the sentence length.

[5] In principle, an order with minimal DL can be non-projective. However, such cases are rare in natural language trees, which have limited topology. In particular, natural language trees have small average branching factors, while a non-projective order with minimal DL occurs only if at least one node of out-degree 3 is present in the tree (Chung, 1984).

constitute a minimal pair with respect to the tree-branching parameter.

Importantly, there exist both predominantly right-branching (e.g. English) and left-branching natural languages (Japanese, Persian) and the comparison between LB/RB-permuted treebanks will show how much of the difference in parsing of typologically different natural languages can be attributed to their different branching directions. Of course, the parsing sensitivity to the parameter depends on the parsing architecture. As discussed in detail below, we investigate both graph-based and transition-based architectures. For a graph-based parser, we do not expect to observe much difference in parsing performance due to directionality, given its global optimisation strategy. On the other hand, a transition-based parser relies on left-to-right processing of words and the fully right-branching or fully left-branching orders can yield different results.

3.3 Dependency Treebanks

We use a sample of fourteen dependency treebanks for twelve languages. The treebanks for Bulgarian, English, Finnish, French, German, Italian and Spanish come from the Universal Dependency Project and are annotated with the same annotation scheme (Agić et al., 2015). We use the treebank for Dutch from the CONLL 2006 shared task (Buchholz and Marsi, 2006). The Polish treebank is described in Woliński et al. (2011) and the Persian treebank in Rasooli et al. (2013). In addition, we use two Latin and two Ancient Greek dependency annotated texts (Haug and Jøhndal, 2008) because these languages are well-known for having very free word order.[6] The quantitative properties of these treebanks are presented in Table 1 (second and third column). This set of treebanks includes those treebanks which had at least 3,000 sentences in their training set after eliminating sentences not fit for permutation (with punctuation tokens or multiple roots). This excluded from our analysis some otherwise typologically interesting languages such as Basque and Arabic. Where available, we used the training-test split of a treebank provided by its distributors; in other cases we split the treebank randomly with a 9-to-1 training-test set proportion.

[6]The Latin corpora comprise works of Cicero (circa 40 BC) and Vulgate (Bible translation, 4th century AD). The Ancient Greek corpora are works of Herodotus (4th century BC) and New Testament (4th century AD). Despite the fact that they belong to the same language, these pairs of texts of different time periods show quite different word order properties (Gulordava and Merlo, 2015).

3.4 Word order properties of original and permuted treebanks

Table 1 presents the treebanks in our sample and the values of DLM ratio and Entropy measures calculated on the training set of the original non-permuted treebanks. From these data, we confirm that the DLM ratio and Entropy measures capture different word order properties as they are not correlated (Spearman correlation $r = 0.32$, $p > 0.1$). For example, we can find languages with both low DLM ratio and high Entropy (Finnish) and high DLM ratio and low Entropy (Persian). Furthermore, these two measures are not a simple reflex of genetic similarity between languages of the same family: for example, Polish (Indo-European family) and Finnish (Finno-Ugric family) are clustered together according to their word order properties.

Table 1 also shows how the DLM ratio and Entropy values change, when we apply the two permutations to the treebanks. For the treebanks permuted to obtain minimal dependency length (DLM ratio = 1), we present Entropy values in the column 'OptDL Entropy'. For the treebanks permuted to obtain minimal entropy (Entropy = 0), we present DLM ratio values in the column 'LB/RB DLM ratio'. With respect to the values of the original treebanks, the DLM ratio and Entropy values of the artificial treebanks are much more narrowly distributed: 1.17 ± 0.02 (mean $\pm$ SD) compared to 1.19 ± 0.07 for DLM ratio and 0.59 ± 0.03 compared to 0.27 ± 0.17 for Entropy.

Notice also that, on average, the treebanks in the LB/RB permuted set have both lower entropy and lower DLM ratio than the original treebanks. The treebanks in the OptDL set have lower DLM ratio, but also higher entropy than the original treebanks.

3.5 Parsing setup

To evaluate the impact of word order properties on parsing performance, we use MSTParser (McDonald et al., 2006) and MaltParser (Nivre et al., 2006) — two widely used representatives of two main de-

Language	Abbr.	Size	Av. sentence length	Original treebanks DLM ratio	Original treebanks Entropy	LB/RB DLM ratio	OptDL Entropy
Polish	pl	29k	6.8	1.13	0.34	1.18	0.55
Italian	it	57k	12.1	1.13	0.18	1.18	0.60
Finnish	fi	46k	5.7	1.13	0.34	1.19	0.53
Spanish	es	63k	15.1	1.15	0.15	1.17	0.62
French	fr	72k	14.5	1.15	0.11	1.20	0.62
English	en	62k	9.5	1.17	0.09	1.16	0.58
Bulgarian	bg	30k	8.5	1.17	0.20	1.17	0.58
Vulgate (La)	la.V	63k	8.8	1.17	0.43	1.18	0.59
Dutch	nl	38k	8.4	1.17	0.26	1.12	0.52
NewTest (AG)	grc.NT	69k	10.5	1.19	0.38	1.17	0.62
German	de	65k	11.5	1.24	0.21	1.21	0.62
Cicero (La)	la.C	35k	11.6	1.26	0.42	1.15	0.61
Persian	fa	35k	9.4	1.33	0.13	1.15	0.61
Herodotus (AG)	grc.H	59k	14.4	1.33	0.46	1.20	0.64
		Mean ($\pm$ st. deviation)		$1.19_{\pm 0.07}$	$0.27_{\pm 0.17}$	$1.17_{\pm 0.02}$	$0.59_{\pm 0.03}$

Table 1: Training size (in number of words), average sentence length, DLM ratio and arc direction entropy (Entropy) measures for the treebanks in our sample. The column 'LB/RB DLM ratio' presents the DLM ratio for LB/RB-permuted treebanks optimised for zero entropy; the column 'OptDL Entropy' presents the arc direction entropy for OptDL-permuted treebanks optimised for minimal DLM ratio.

Language	Original		OptDL		LB		RB	
	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS
Polish	92	88	94	88	94	89	94	89
Italian	94	91	91	85	94	90	95	91
Finnish	83	80	85	81	90	85	91	87
Spanish	86	81	80	72	85	76	88	80
French	84	80	81	74	90	82	91	85
English	90	88	85	79	89	83	89	83
Bulgarian	93	89	92	85	92	85	93	87
Vulgate (La)	86	81	88	81	93	86	93	86
Dutch	88	84	93	87	95	90	95	90
NewTestament (AG)	85	79	88	81	93	85	91	73
German	86	80	84	75	89	78	89	81
Cicero (La)	67	59	79	67	88	76	87	76
Persian	83	74	84	73	90	80	90	80
Herodotus (AG)	72	65	83	74	89	79	88	67
Average	85	80	86	79	91	83	91	83

Table 2: Percent labelled and unlabelled accuracies of MaltParser on the original treebanks, on the treebanks permuted for optimal dependency length (OptDL), and on the left-branching (LB) and right-branching (RB) permuted data that minimise entropy.

pendency parsing architectures: a graph-based parsing architecture and a transition-based architecture. The graph-based architecture is known to be less dependent on word order and dependency length than transition-based dependency parsers, as it searches the whole space of possible parse trees and solves a global optimisation problem (McDonald and Nivre, 2011).

To achieve competitive performance, the transition-based MaltParser must be provided with a list of features tailored for each treebank and each language. We used the MaltOptimizer package (Ballesteros and Nivre, 2012), to find the best features based on the training set. By contrast, MSTParser is trained on all the treebanks in our sample using the default configuration (first-order projective).

4 Experiments and results

In this section, we illustrate the power of the technique and the fine-grained analyses supported by it with a range of planned, pairwise quantitative and qualitative analyses of the parsing results.

4.1 Comparison of parsing performance between original and permuted treebanks

Table 2 presents the parsing results for MaltParser for the original treebanks and the three sets of permuted treebanks (OptDL, LB, RB). Table 3 presents the results on the same data for MSTParser. For MST, the parsing performances on the fully left-branching and right-branching treebanks are identical, as expected, when percentages are rounded at the two-digit level, which is what we report here.

As discussed in the introduction, a comparison between parsers in a multilingual setting is not straightforward. Instead, we attempt to understand their common behaviour with respect to the word order properties of languages. The first observation is that, overall, all three sets of permuted data are easier to parse than the original data, for both parsers. We observe an increase of +1% and +6% UAS for OptDL and LB/RB data, respectively, for Malt, and an increase of +4% and +8% UAS for OptDL and LB/RB data, respectively, for MST. The better results on the LB/RB permuted data must be due to the observation above: the LB/RB data have both lower Entropy and DLM ratio than the original data.

Lang.	Original		OptDL		LB/RB	
	UAS	LAS	UAS	LAS	UAS	LAS
pl	93	85	95	84	95	85
it	93	88	91	85	94	87
fi	84	79	87	80	91	84
es	85	67	84	64	91	68
fr	84	70	86	68	92	71
en	88	85	87	79	91	82
bg	93	87	91	83	92	84
la.V	84	75	90	79	94	82
nl	85	79	93	85	95	88
grc.NT	84	74	89	78	93	83
de	87	67	88	66	91	69
la.C	68	54	84	67	89	72
fa	84	73	86	74	91	79
grc.H	69	57	86	71	90	76
Av.	84	74	88	76	92	79

Table 3: Percent labelled and unlabelled accuracies of MSTParser on the original treebanks, on the treebanks permuted for optimal dependency length (OptDL), and on the left/right-branching (LB/RB) permuted data.

Overall, the performance of the parsers on our artificial treebanks confirms that the lengths of the dependencies and the word order variability are two factors that negatively affect parsing accuracy. Two illustrative examples are Latin, a language well-known for its variable word order (as confirmed by the high entropy values of 0.42 and 0.43 for our two treebanks), and German, a language known for its long dependencies (as confirmed by its high DML ratio of 1.24). For the Cicero text, for example, we can conclude that indeed its variable word order is the primary reason for the very low parsing performances (67%–68% UAS). These numbers improve significantly when the treebanks are rearranged in a fixed word order (87%–89% UAS). This permutation reduces DLM by 0.11 and reduces entropy by 0.42, yielding the very considerable increase in UAS of 21%. The other permutation, which optimises DL, reduces DLM by 0.26, but increases entropy by 0.19. This increase in entropy dampens the beneficial effect of DL reduction and performance increases 12%, less than in the fixed-order permutation. For German, our analysis gives the same overall results. The DLM ratio in the RB/LB sce-

nario decreases slightly (from 1.24 to 1.21) and its entropy also decreases (-0.21). The performance of the parsers on RB/LB-permuted data is better than on the original data (89%–91% against 86%–87% UAS). Moreover, when DLM is reduced (-0.24, in the OptDL permutation), but entropy is increased (from 0.21 to 0.62), we find a reduction in performance for Malt (from 86% to 84% for UAS). These data weakly suggest that the word order variability of German, minimised in the RB/LB case, has higher impact on parsing difficulty than its well-known long dependencies.

A more detailed picture emerges when we compare pairwise the original treebanks to the permuted treebanks for each of the languages. For this analysis, we use the measure of unlabelled accuracy, since attachment decisions are more directly dependent on word order than labelling decisions, which are mediated by correct attachments. Hence, we limit our analysis to the space of three parameters: DLM ratio, Entropy and UAS.

Figures 1 (OptDL) and 2 (RB) plot the differences in UAS of MaltParser between pairs of the permuted and the original treebanks for each language to the differences in DLM ratio and Entropy between these treebanks. Our dependent variable is $\Delta UAS = UAS(T') - UAS(T)$ computed from Table 2. The x-axis and the y-axis values $\Delta DLM = DLMRatio(T) - DLMRatio(T')$ and $\Delta Entropy = Entropy(T) - Entropy(T')$ compute the differences of the measures between the original treebank and the permuted treebank based on the numbers in Table 1. We have chosen to calculate these differences reversing the two factors, compared to the ΔUAS value, for better readability of the figures: an increase in the axes values (entropy or dependency lengths) should correspond to the decrease in difficulty of parsing and therefore to the increase of the dependent variable ΔUAS. The same relative values of the measures and the parsing accuracy for MSTParser result in very similar plots, which we do not include here for reasons of space.

For the OptDL data (Figure 1), the overall picture is very coherent: the more DLs are minimised and the less entropy is added to the artificial treebank, the larger the gain in parsing performance (blue circles in the lower left corner and red circles in the upper right corner). Again, we observe an interaction

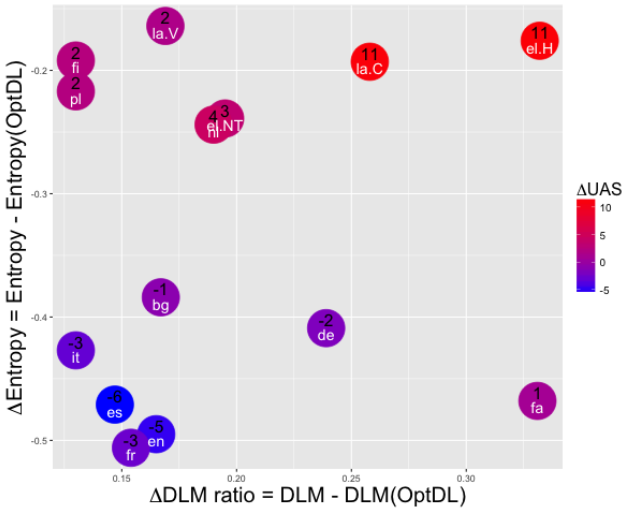

Figure 1: Differences in UAS of MaltParser between OptDL-permuted and original pairs of treebanks for the corpora in our sample.

between DLM ratio and Entropy parameters: for the languages with originally relatively low DLM ratio and low Entropy, such as English or Spanish, the performance on the permuted data decreases. This is because while DLM decreases, Entropy increases and, for this group of languages, the particular trade-off between these two properties leads to lower parsing accuracy.

RB-permuted data show similar trends (Figure 2). An interesting regularity is shown by four languages (Latin Vulgate, Ancient Greek New Testament, Dutch and Persian) on the off-diagonal. Although they have different relative Entropy and DLM ratio values, which span from near minimal to maximal values, the improvement in parsing performance on these languages is very similar (as indicated by the same purple colour). This again strongly points to the fact that both DLM ratio and Entropy contribute to the observed parsing performance values.

We can further confirm the effect of dependency length by comparing the parsing accuracy across sentences.[7] Consider the Dutch treebank and its RB-permuted pair. For each sentence and its permuted counterpart, we can compute the difference in their dependency lengths ($\Delta DLM = DLM - DLM_{RB}$)

[7]The Entropy measure is computed on a whole treebank and cannot be meaningfully compared across sentences.

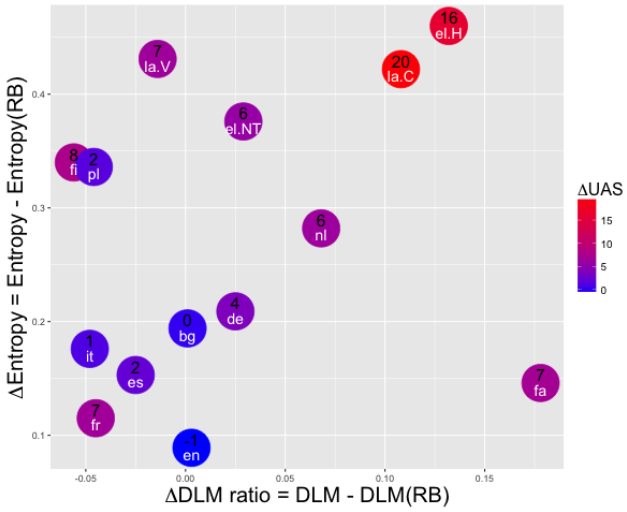

Figure 2: Differences in UAS of MaltParser between RB-permuted and original pairs of treebanks for the corpora in our sample.

and compare it to the difference in parsing performance ($\Delta UAS = UAS_{RB} - UAS$). We expect to observe that ΔUAS increases when ΔDLM increases. Indeed, the parsing results on Dutch show a positive correlation between these two values ($r = 0.40, p < 0.001$ for Malt and $r = 0.55, p < 0.001$ for MST).

All these analyses confirm and quantify that dependency length and, more significantly, word order variability affect parsing performance.

4.2 Sentence-level analysis of parsing performance

Referring back to the results in Table 2, we observe that MaltParser shows the same average accuracy for RB and LB-permuted data. However, some languages show significantly different results between their LB and RB-permuted data, especially in their labelled accuracy scores. The New Testament corpus, for example, is much easier to parse when it is rearranged in left-branching order (91% RB vs 93% LB UAS, 73% RB vs 85% LB LAS). Our artificial data allows us to investigate this difference in the scores by looking at parsing accuracy at the sentence level.

The differences in Malt accuracies on RB-permuted and LB-permuted data are striking, because these data have the same head-direction entropy and dependency lengths properties. The only word order difference is in the branching parameter resulting in two completely symmetrical word orders for each sentence of the original treebank. To understand the behaviour of MaltParser, and of transition-based parsers in general, we looked at the out-degree, or branching factor, of the syntactic trees. The intuition is that when many children appear on one side of a head, the parser behaviour on head-final and head-initial orders can diverge due to sequences of different operations, such as *shift* versus *attach*, that must be chosen in the two cases.[8]

The data for the New Testament shows that the branching factor plays a role in the LB/RB differences found in this treebank. For each pair of sentences with LB/RB orders, we computed the parsing accuracies (UAS and LAS) and the branching factor as the average out-degree of the dependency tree. We then tested whether the better performance on the LB data is correlated with the branching factor across the sentences ($UAS_{LB} - UAS_{RB} \sim BF$). The Pearson correlation for UAS values was 0.08 ($p = 0.02$), but for LAS values the correlation was 0.30 and highly significant ($p < 0.001$). On sentences with larger branching factors, the labelled accuracy scores on the LB data were better compared to the RB data.

We combine our result for the branching factor with an observation based on the confusion matrix of the labels, to provide a more accurate explanation of the comparatively low LAS in the RB-permuted treebank of the New Testament corpus. We found that when a verb or a noun has several one-word children, such as 'aux' (auxiliaries), 'atr' (attributes), 'obl' (obliques), 'adv' (adverbs) etc, these are frequently confused and receive the wrong label, if they appear after the head (RB data), but the labels are assigned correctly if these elements appear before the head (LB data). It appears that the leftward placement of children is advantageous for the transition-based MaltParser, as at the moment of first attachment decision for the child closest to the head it has access to a larger left context. When children appear after the head, the first one is attached before any other children are seen by the parser and the la-

[8]The MaltParser configurations for LB and RB data had the same parsing algorithm (Covington projective).

belling decision is less informed, leading to more labelling errors.

It should be noted that it is not always possible to identify a single source of difficulty in the error analysis. Contrary to New Testament, Spanish is easier to parse when it is rearranged into the right-branching order (88% RB vs 85% LB UAS, 80% RB vs 76% LB LAS). However, the types of difficult dependencies emerging from the different branching of the LB/RB data were not similar or symmetric to that of New Testament. In the case of Spanish, we did not observe a distinct dimension of errors which would explain the 4% difference in UAS scores.[9]

5 General discussion

Our results highlight both the contributions and the challenges of the proposed method. On the one hand, the results show that we can identify and manipulate word order properties of treebanks to analyse the impact of these properties on parsing performance and suggest avenues to improve it. In this respect, our framework is similar to standard analyses of parsing performance based on separate manipulations of individual word-level features (such as omitting morphological annotation or changing coarse PoS tags to fine PoS tags). Similarly to these evaluation procedures, our approach can lead to improved parsing models or better choice of parsing model by finding out their strengths and weaknesses. The performance of Malt and MST (Tables 2 and 3) — while not directly comparable to each other due to differences in the training set-up (Malt features are optimised for each language and permutation) — show that MST performs better on average on permuted datasets than Malt. This can suggest that MST handles the high entropy of the OptDL permuted set as well as the long dependencies of LB/RB permuted sets better, or, conversely, that the MaltParser does not perform well on treebanks with high word order variability between the children attached to the same head (see Section 4.2). When two parsing systems are known to have different strengths and weaknesses they can be successfully combined in an ensemble model for more robust performance (Surdeanu and Manning, 2010).

A contribution of the parsing performance analyses in a multilingual setting is the identification of difficult properties of treebanks. For Cicero and Herodotus texts, for example, our method reveals that their word order properties are important reasons for the very low parsing performances. This result confirms intuition, but it could not be firmly concluded without factoring out confounds such as the size of the training set or the dissimilarity between the training and test sets, which could also be reasons for low parsing performance. For German, our analysis gives more unexpected results and allows us to conclude that the variability of word order is a more negative factor on parsing performance than long dependencies. Together, the knowledge of word order properties of a language and the knowledge of parsing performance related to these properties give us an a priori estimation of what parsing system could be better suited for a particular language.

On the other hand, our method also raises some complexities. Compared to commonly used parsing performance analyses related to word-level features, the main challenges to a systematic analysis of word order lie in its multifactorial nature and in the large choice of quantifiable properties correlated with parsing performance. First, the multifactorial nature of word order precludes one from considering word order properties separately. The two properties we have looked at — DLM ratio and arc direction entropy — cannot be teased apart completely since minimising one property leads to the increase of the other.

Another challenge is due to the fact that formal quantitative approaches to studying word order variation cross-linguistically are just beginning to appear and not all word order features relevant for parsing performance have been identified. In particular, our results suggest that the relative order between the children (and not only the order between heads and their children) should be taken into account (Section 4.2). However, we are not aware of previous work which proposes a measure for this property and describes it typologically on a large scale.

Finally, our method, which consists in creating ar-

[9] Overall, the variance in the LB/RB performances on Spanish is relatively high and the mean difference (computed across UAS scores for sentences) is not statistically significant (t-test: $p > 0.5$) – a result we would expect if errors cannot be imputed to clear structural factors.

tificial treebanks, can prove useful beyond parsing evaluation. For instance, our data could enrich the training data for tasks such as de-lexicalized parser transfer (McDonald et al., 2011). Word order properties play an important role in computing similarity between languages and finding the source language leading to the best parser performance in the target language (Naseem et al., 2012; Rosa and Zabokrtsky, 2015). A possibly large artificially permuted treebank with word order properties similar to the target language could then be a better training match than a small treebank of an existing target natural language.

6 Related work

Much previous work has been dedicated to the evaluation of parsing performance, also in a multilingual setting. The shared tasks in multilingual dependency parsing (Buchholz and Marsi, 2006; Nivre et al., 2007) and parsing of morphologically-rich languages (Tsarfaty et al., 2010; Seddah et al., 2013) collected a large set of parsing performance results. Some steps towards comparability of the annotations of multilingual treebanks and the parsing evaluation measures were proposed and undertaken in Tsarfaty et al. (2011), Seddah et al. (2013) and, most recently, in the collaborative Universal Dependencies effort (de Marneffe et al., 2014; Nivre et al., 2016). However, little work has suggested an analysis of the differences in parsing performance across languages in connection with the word order properties of treebanks.

Some papers have analysed the impact of dependency lengths on parsing performance in English. McDonald and Nivre (2011) demonstrated that parsers make more mistakes in longer sentences and on longer dependencies. Rimell et al. (2009) and Bender et al. (2011) created benchmark test sets of constructions containing long dependencies, such as subject and object relative clauses, and analysed parsing behaviour on these selected constructions. Other analyses on long-distance dependencies can be found in Nivre et al. (2010) and Merlo (2015). We are not familiar with any similar analysis of parsing performance in English addressing other word order variation properties (e.g. head-direction entropy).

In Gulordava and Merlo (2015), the parsing performance on several Latin and Ancient Greek texts is analysed with respect to the dependency length and, indirectly, the head-direction entropy. The authors compare the parsing performance across texts of the same language (Latin or Ancient Greek) from separated historical periods which differ slightly in their word order properties.[10] Gulordava and Merlo (2015) show that texts with longer dependencies and more varied word order are harder to parse. Assuming the same lexical material of the texts, their particular setting allows a more direct comparison of parsing performance than a standard multilingual setting where languages differ in many aspects other than word order.

The calculation of the minimal dependency length through the permutation of a dependency treebank was proposed in the work of Temperley and Gildea (Temperley, 2007; Gildea and Temperley, 2010). In this work and the following work of Futrell et al. (2015a), several types of permutations were employed to compute different lower bounds on dependency length minimisation in English and across dozens of languages.

Artificially permuted treebanks were previously used in Fong and Berwick (2008) as stress-test diagnostics for cognitive plausibility of parsing systems. In particular, Fong and Berwick (2008) permuted the order of words in the English Penn Treebank to obtain 'unnatural' languages. Their permutations included transformations to head-final and head-initial orders (applied with 50%-50% proportion to sentences in the treebank) and reversing the respective order of complements and adjuncts. The parsing performances on these permuted treebanks were 0.5–1 point lower than on the original treebank, which the authors interpreted as too accurate to be a cognitively plausible behaviour for a model of the human parser. From the perspective of our paper, the permuted treebanks of Fong and Berwick (2008) were constructed to have longer dependencies and higher word order variation; the lower performances are therefore in agreement with our own results.

[10] The Latin and Ancient Greek data we used in this work is a subset of the data that Gulordava and Merlo (2015) have analysed, all coming from the PROIEL treebanks (Haug and Jøhndal, 2008).

7 Conclusions

We have proposed a method to analyse parsing performance cross-linguistically. The method is based on the generation and the evaluation of artificial data obtained by permuting the sentences in a natural language treebank. The main advantage of this approach is that it teases apart the linguistic factors from the extra-linguistic factors in parsing evaluation.

First, we have shown how this method can be used to estimate the impact of two word order properties — dependency length and head-direction entropy — on parsing performance. Previous observations that longer dependencies are harder to parse are confirmed on a much larger scale than before, while controlling for confounding treebank properties. It has also been found that variability of word order is an even more prominent factor affecting performance.

Second, we have shown that the construction of artificial data opens a new way to analyse the behavior of parsers using sentence-level observations. Sentence-level evaluations could be a very powerful tool for detailed investigations of how syntactic properties of languages affect parsing performance and could help creating more cross-linguistically valid parsing techniques.

Two avenues are open for future work. First we will investigate more properties related to word order. Specifically, we will apply the method to the non-projectivity property. On the one hand, dependency lengths and non-projectivity are correlated properties, as predicted theoretically (Ferrer-i-Cancho, 2006). Our data confirm this relation empirically: the Pearson correlation between DLM ratio and the percentage of non-projective dependencies across treebanks is 0.66 ($p < 0.02$). On the other hand, this correlation is not perfect and both dependency length and non-projectivity should be taken into account to fully explain the variation in parsing performance.

Second, we have not attempted in the current work to estimate the function f (see section 2.1). This task is equivalent to automatic prediction of parsing accuracy of a treebank based on its properties. Ravi et al. (2008) have proposed an accuracy prediction method for one language (English) based on simple lexical and syntactic properties. Combining their insights with our analysis of word order could lead to a first language-independent approximation of f.

Acknowledgements

We gratefully acknowledge the partial funding of this work by the Swiss National Science Foundation, under grant 144362.

References

Željko Agić, Maria Jesus Aranzabe, Aitziber Atutxa, Cristina Bosco, Jinho Choi, Marie-Catherine de Marneffe, Timothy Dozat, Richárd Farkas, Jennifer Foster, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Jan Hajič, Anders Trærup Johannsen, Jenna Kanerva, Juha Kuokkala, Veronika Laippala, Alessandro Lenci, Krister Lindén, Nikola Ljubešić, Teresa Lynn, Christopher Manning, Héctor Alonso Martínez, Ryan McDonald, Anna Missilä, Simonetta Montemagni, Joakim Nivre, Hanna Nurmi, Petya Osenova, Slav Petrov, Jussi Piitulainen, Barbara Plank, Prokopis Prokopidis, Sampo Pyysalo, Wolfgang Seeker, Mojgan Seraji, Natalia Silveira, Maria Simi, Kiril Simov, Aaron Smith, Reut Tsarfaty, Veronika Vincze, and Daniel Zeman. 2015. Universal dependencies 1.1.

Miguel Ballesteros and Joakim Nivre. 2012. MaltOptimizer: An optimization tool for MaltParser. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL '12)*, pages 58–62, Avignon, France, April.

Emily M. Bender, Dan Flickinger, Stephan Oepen, and Yi Zhang. 2011. Parser evaluation over local and non-local deep dependencies in a large corpus. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*, pages 397–408, Edinburgh, United Kingdom, July.

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 '06)*, pages 149–164, New York City, NY, USA, June.

F. R. K. Chung. 1984. On optimal linear arrangements of trees. *Computers & Mathematics with Applications*, 10(1):43–60.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference*

on Language Resources and Evaluation (LREC '14), pages 4585–4592, Reykjavik, Iceland, May.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Ramon Ferrer-i-Cancho and Haitao Liu. 2014. The risks of mixing dependency lengths from sequences of different length. *Glottotheory*, 5(2):143–155.

Ramon Ferrer-i-Cancho. 2006. Why do syntactic links not cross? *EPL (Europhysics Letters)*, 76(6):12–28.

Sandiway Fong and Robert Berwick. 2008. Treebank parsing and knowledge of language: A cognitive perspective. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 539–544, Washington, DC, USA, July.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015a. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015b. Quantifying word order freedom in dependency corpora. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 91–100, Uppsala, Sweden, August.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Science*, 34(2):286–310.

Kristina Gulordava and Paola Merlo. 2015. Diachronic trends in word order freedom and dependency length in dependency-annotated corpora of Latin and Ancient Greek. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 121–130, Uppsala, Sweden, August.

Kristina Gulordava, Paola Merlo, and Benoit Crabbé. 2015. Dependency length minimisation effects in short spans: a large-scale analysis of adjective placement in complex noun phrases. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 477–482, Beijing, China, July.

Eva Hajičová, Jiří Havelka, Petr Sgall, Kateřina Veselá, and Daniel Zeman. 2004. Issues of projectivity in the Prague Dependency Treebank. *Prague Bulletin of Mathematical Linguistics*, (81).

Dag T. T. Haug and Marius L. Jøhndal. 2008. Creating a parallel treebank of the Old Indo-European Bible translations. In *Proceedings of the 2nd Workshop on Language Technology for Cultural Heritage Data*, pages 27–34, Marrakech, Morocco, June.

John A. Hawkins. 1994. *A performance theory of order and constituency*. Cambridge University Press, Cambridge, UK.

James Henderson, Paola Merlo, Ivan Titov, and Gabriele Musillo. 2013. Multilingual joint parsing of syntactic and semantic dependencies with a latent variable model. *Computational Linguistics*, 39(4):949–998.

Haitao Liu. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 120(6):1567–1578, June.

Ryan McDonald and Joakim Nivre. 2011. Analyzing and integrating dependency parsers. *Computational Linguistics*, 37(1):197–230.

Ryan McDonald, Kevin Lerman, and Fernando Pereira. 2006. Multilingual dependency analysis with a two-stage discriminative parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL'06)*, pages 216–220, New York, NY, USA, June.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Edinburgh, Scotland, UK, July.

Paola Merlo. 2015. Evaluation of two-level dependency representations of argument structure in long-distance dependencies. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 221–230, Uppsala, Sweden, August.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–637, Jeju Island, Korea, July.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, pages 2216–2219, Genova, Italy, May.

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, Prague, Czech Republic, June.

Joakim Nivre, Laura Rimell, Ryan McDonald, and Carlos Gómez Rodríguez. 2010. Evaluation of dependency parsers on unbounded dependencies. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 833–841, Beijing, China, August.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC '16)*, Portoroz, Slovenia, May.

Joakim Nivre. 2009. Non-projective dependency parsing in expected linear time. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Long papers*, pages 351–359, Suntec, Singapore, August.

Mohammad Sadegh Rasooli, Manouchehr Kouhestani, and Amirsaeid Moloodi. 2013. Development of a Persian syntactic dependency treebank. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 306–314, Atlanta, Georgia, June.

Sujith Ravi, Kevin Knight, and Radu Soricut. 2008. Automatic prediction of parser accuracy. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 887–896, Honolulu, Hawaii, October.

Laura Rimell, Stephen Clark, and Mark Steedman. 2009. Unbounded dependency recovery for parser evaluation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP '09)*, pages 813–821, Suntec, Singapore, August.

Rudolf Rosa and Zdenek Zabokrtsky. 2015. Klcpos3 - a language similarity measure for delexicalized parser transfer. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 243–249, Beijing, China, July.

Djamé Seddah, Reut Tsarfaty, and Jennifer Foster, editors. 2011. *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL '11)*. Dublin, Ireland, October.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska, and Eric Villemonte de la Clergerie. 2013. Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 146–182, Seattle, Washington, USA, October.

Mihai Surdeanu and Christopher D. Manning. 2010. Ensemble models for dependency parsing: Cheap and good? In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 649–652, Los Angeles, California, June.

David Temperley. 2007. Minimization of dependency length in written English. *Cognition*, 105(2):300–333.

Harry Joel Tily. 2010. *The Role of Processing Complexity in Word Order Variation and Change*. Ph.D. Thesis, Stanford University.

Ivan Titov and James Henderson. 2007. A latent variable model for generative dependency parsing. In *Proceedings of the 10th International Conference on Parsing Technologies*, IWPT '07, pages 144–155, Prague, Czech Republic, June.

Ivan Titov, James Henderson, Paola Merlo, and Gabriele Musillo. 2009. Online graph planarisation for synchronous parsing of semantic and syntactic dependencies. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI '09)*, pages 1562–1567, Pasadena, California, USA, July.

Reut Tsarfaty, Djamé Seddah, Yoav Goldberg, Sandra Kübler, Marie Candito, Jennifer Foster, Yannick Versley, Ines Rehbein, and Lamia Tounsi. 2010. Statistical parsing of morphologically rich languages (SPMRL): What, how and whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, SPMRL '10, pages 1–12, Los Angeles, California.

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 (EMNLP '11)*, pages 385–396, Edinburgh, Scotland, UK, July.

Marcin Woliński, Katarzyna Głowińska, and Marek Świdziński. 2011. A preliminary version of Skladnicaa treebank of Polish. In *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 299–303, Poznan, Poland, November.