Universal Dependency Evaluation

Joakim Nivre
Dept. of Linguistics and Philology
Uppsala University
joakim.nivre@lingfil.uu.se

Chiao-Ting Fang
Dept. of Linguistics and Philology
Uppsala University
chfa4190@student.uu.se

Abstract

Multilingual parser evaluation has for a long time been hampered by the lack of cross-linguistically consistent annotation. While initiatives like Universal Dependencies have greatly improved the situation, they have also raised questions about the adequacy of existing parser evaluation metrics when applied across typologically different languages. This paper argues that the usual attachment score metrics used to evaluate dependency parsers are biased in favor of analytic languages, where grammatical structure tends to be encoded in free morphemes (function words) rather than in bound morphemes (inflection). We therefore propose an alternative evaluation metric that excludes functional relations from the attachment score. We explore the effect of this change in experiments using a subset of treebanks from release v2.0 of Universal Dependencies.

1 Introduction

The last decade has seen a steadily growing interest in multilingual parsing research, inspired by such events as the CoNLL shared tasks on multilingual dependency parsing (Buchholz and Marsi, 2006; Nivre et al., 2007) and the SPMRL shared tasks on parsing morphologically rich languages (Seddah et al., 2013; Seddah et al., 2014). This has led to a number of conjectures about the suitability of different parsing models for languages with different structural characteristics, but it has been surprisingly hard to study the interplay of parsing technology and language typology in a systematic way. To some extent, this is due to data-related factors such as text genre and training set size, which are hard to control for, but even more important has been the fact that syntactic annotation is not standardized across languages. This has made it almost impossible to isolate the influence of typological variables, such as word order or morphosyntactic alignment, from the effect of more or less arbitrary choices in linguistic representations. The absence of cross-linguistically consistent annotation has also been a constant source of noise in the evaluation of cross-lingual learning of syntax (Hwa et al., 2002; Zeman and Resnik, 2008; McDonald et al., 2011).

Fortunately, there is now also a growing interest in developing cross-linguistically consistent syntactic annotation, which has led to a number of initiatives and proposals (Zeman et al., 2012; McDonald et al., 2013; Tsarfaty, 2013; de Marneffe et al., 2014). Many of these initiatives have now converged into Universal Dependencies (UD), an open community effort that aims to develop cross-linguistically consistent treebank annotation for many languages and that has so far released 70 treebanks representing 50 languages (Nivre, 2015; Nivre et al., 2016). The basic idea behind the UD scheme is to maximize parallelism across languages by focusing on dependency relations between content words, which are more likely to be similar across languages, and to use cross-linguistically valid categories for morphological and syntactic analysis. The UD scheme is illustrated in Figure 1 for two translationally equivalent sentences in English and Finnish. For readability, we display only a subset of the full annotation, in particular suppressing all morphological features except case.

The example shows that English and Finnish have rather different structural characteristics. What is expressed by eight words in English is expressed by four words in Finnish, and whereas word order and function words like *from* are crucial in English for understanding who does what to whom, the same information is encoded in Finnish mainly by nominal case inflection (nominative for the subject, accusative for the object, and ela-

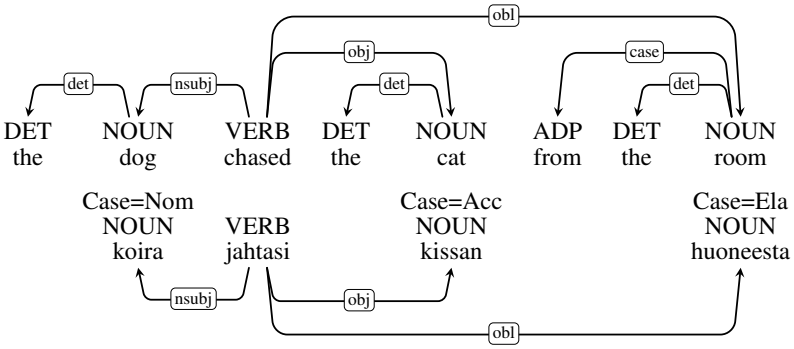

Figure 1: Simplified UD annotation for equivalent sentences from English (top) and Finnish (bottom).

tive for the locative modifier). Moreover, Finnish has no explicit encoding of the information expressed by the definite article *the* in English. Nevertheless, the main grammatical relations are exactly parallel in the two sentences, with the main verb *chased*/*jahtasi* having three direct nominal dependents, which can be categorized in both languages as (nominal) subject (*nsubj*), object (*obj*), and oblique modifier (*obl*). This illustrates how UD maximizes parallelism by giving priority to dependency relations between content words.

It is tempting to assume that cross-linguistically consistent annotation automatically guarantees cross-linguistically valid parser evaluation. Unfortunately, this is not the case, because our old established evaluation metrics may not be adequate for the new harmonized representations. The most commonly used metric in dependency parsing is the (labeled or unlabeled) attachment score, which measures the percentage of words that have been assigned the correct head (with or without taking the dependency label into account). Suppose now that a parser makes a single mistake on each of the sentences in Figure 1, say, by attaching the locative modifier to the object instead of to the verb. It seems intuitively correct to say that the parser has done an equally good job in both cases. However, for simple arithmetical reasons, the English parser will be credited with an attachment score of 87.5%, while the Finnish parser only gets 75%. In other words, the impact of a single error is doubled in Finnish because of the smaller denominator. Using the attachment score for cross-linguistic comparisons can therefore be quite misleading even if the annotation has been harmonized across languages.

What should we do about this? A drastic proposal would be to give up intrinsic evaluation altogether, on the grounds that it will always be biased one way or the other, and instead put all our hope on extrinsic evaluation. In doing so, however, we would run the risk of just moving the problem elsewhere. For example, if we decide to evaluate parsers through their impact on machine translation quality, how do we guarantee that the latter evaluation is comparable across languages? Furthermore, intrinsic evaluation metrics will always be useful for internal testing purposes, so we might as well do our best to develop new metrics that are better suited for cross-linguistic comparisons. This is the purpose of this paper.

More precisely, we want to find an alternative evaluation metric for parsing with UD representations, a metric that puts more emphasis on dependency relations between content words in order to maximize comparability across languages, following the same principle as in the design of the annotation itself. We will begin by dividing the syntactic relations used in UD representations into a number of different groups and study their impact on evaluation scores. We will then propose a new metric called CLAS, for Content-Word Labeled Attachment Score, and analyze in more depth how different languages are affected by excluding different functional relations from the evaluation.

2 Syntactic Relations in UD

Annotation in UD consists of a morphological and a syntactic layer. The morphological layer assigns to each word a lemma, a part-of-speech tag and a set of morphological features. The part-of-speech tag comes from a fixed inventory of 17 tags, which is a revised and extended version of the Google universal tagset (Petrov et al., 2012), and the features come from a standardized but extendable inventory based on Interset (Zeman, 2008). The syntactic layer is essentially a dependency tree with labels taken from a set of 37 syntactic relations,

which is a revised version of the universal Stanford dependencies (de Marneffe et al., 2014).

As explained in the introduction, the syntactic tree gives priority to grammatical relations between content words, while function words are attached to the content word they specify using special relations such as *case* (for adpositions), *mark* (for subordinating conjunctions) and *aux* (for auxiliary verbs). Although these functional relations are formally indistinguishable from other relations in the tree, they can be seen as encoding features of the content word rather than representing real dependency relations.

When applying the standard attachment score metrics to UD, functional relations are scored just like any other relation encoded in the dependency tree (except the special *punct* relation for punctuation, which is often excluded from evaluation). If a language makes frequent use of function words to encode grammatical information, these relations will therefore make a large contribution to the overall score. Since these relations tend to be local and involve highly frequent words, they also tend to have higher than average accuracy, which means that the overall score comes out higher if they are included. For a language that instead uses morphology to encode grammatical information of a similar kind, there will be no corresponding boost to the evaluation score, because morphological features are not included in the parsing score. Moreover, as illustrated earlier, errors on content word dependencies will be more severely penalized in such a language, because the error rate is normalized by the number of words. In this way, languages with a lower ratio of function words are in effect doubly penalized.

One strategy for dealing with this problem could be to come up with a more comprehensive metric that considers the full grammatical representation and abstracts over different realization patterns and puts morphological features and function words on a more equal footing. Such a metric has been proposed in the context of grammar-based parsing by Dridan and Oepen (2011). In the context of UD, however, this would require a substantial research effort in order to establish correspondences between many languages. And while this is precisely the type of research that UD is meant to enable, it would be premature to assume that we already have the required knowledge. For the time being, we will therefore propose a new metric for syntactic dependencies that is limited to those dependencies that we can expect to find in all or most languages. Besides being less biased from a cross-linguistic perspective, such a metric may also be more relevant for downstream language understanding tasks, where errors on functional relations often matter less than errors on argument and modifier relations. And by comparing results for this metric to those obtained with standard attachment scores, we can estimate the degree of bias inherent in the older metric.

As a preliminary to defining the new metric, we first divide the 37 syntactic UD relations into five disjoint subsets, listed in Table 1. FUN is the subset of relations that relate a function word to a content word, including determiners (*det*), classifiers (*clf*), adpositions (*case*), auxiliaries (*aux*, *cop*), and conjunctions (*cc*, *mark*). The first three can be grouped together as *nominal* functional relations, because they are associated with noun phrases (in the extended sense that includes adpositional phrases), while *aux* and *cop* are connected to clausal predicates, and *mark* and *cc* link clauses (or other phrases) in relations of subordination or coordination.

The set MWE contains relations used to analyze (restricted classes) of multiword expressions. The *fixed* relation is used for completely fixed, grammaticized expressions like *in spite of* and *by and large*; the *flat* relation is used for semi-fixed expressions without a clear syntactic head, and the *compound* relation is used for all kinds of compounding. These relations are clearly different from the functional relations, but their distribution can also be expected to vary across languages, sometimes because of typological factors and sometimes simply because of orthographical conventions. For example, noun-noun compounds like *orange juice* are most commonly written as two space-separated tokens in English, which according to the UD guidelines require that they are analyzed as a syntactic combination using the *compound* relation. Exactly parallel expressions in other Germanic languages like German and Swedish are normally written as a single token (for example, *apelsinjuice* in Swedish), which has as a consequence that the compounding relation is not included in the syntactic evaluation for the latter languages. The relation *goeswith*, finally, is different from the (other) MWE relations in that it is primarily intended for annotation of orthographic

FUN	MWE	CORE	NON-CORE			PUNCT
aux	compound	ccomp	acl	discourse	orphan	punct
case	fixed	csubj	advcl	dislocated	parataxis	
cc	flat	iobj	advmod	expl	reparandum	
clf	goeswith	nsubj	amod	list	root	
cop		obj	appos	nmod	vocative	
det		xcomp	conj	nummod		
mark			dep	obl		

Table 1: Subsets of UD relations: core, non-core, functional, multiword and punctuation.

errors, where a single word has accidentally been split into two, but it is similar in that it does not denote a proper syntactic relation.

The remaining UD relations are divided into CORE, NON-CORE and PUNCT. CORE includes relations for core arguments of predicates, which play a central role in the UD taxonomy and arguably in all syntactic representations. NON-CORE includes all other syntactic relations, including modifier relations at various syntactic levels as well as relations for analyzing coordination and special phenomena like ellipsis and disfluencies. PUNCT, finally, contains the single relation *punct*, which has an unclear status as a syntactic relation and is often excluded in evaluation metrics.

3 Labeled Attachment Score

The labeled attachment score (LAS) evaluates the output of a parser by considering how many words have been assigned both the correct syntactic head and the correct label. If parse trees and gold standard trees can be assumed to have the same yield, and if no syntactic relations are excluded, then it reduces to a simple accuracy score, but in general it can be defined as the labeled F_1-score of syntactic relations.

To get a better view of the impact of different relation types on the overall LAS, we performed a simple experiment where we trained and evaluated MaltParser (Nivre et al., 2006) on treebanks from the latest UD release (v2.0). The parser used an arc-standard transition system with online reordering and a lazy oracle (Nivre et al., 2009) and an extended feature model that takes all morphological features into account. We selected one treebank per language[1] but only included treebanks containing morphological features and at least 30,000 words. We used the dedicated training sets for training and the development sets for evaluation. To make evaluation scores comparable across languages, we replaced all language-specific subtypes of syntactic relations by their universal supertypes.

Table 2 shows the results for the 42 treebanks included in the experiment. The LAS column reports the standard LAS score over all relations (including punctuation). The next four columns report the LAS for CORE, NON-CORE, FUN and MWE separately. The last three columns report the difference in LAS score when excluding relations in PUNCT, FUN and MWE, respectively.

The first thing to note is that there is a very large variation in LAS scores, ranging from a high of 88.29 for Slovenian to a low of 56.35 for Lithuanian. Some of this variation can be explained by data set specific properties like text genre and training set size, and it is undeniable that the parsing model used works better for some languages than others. However, the results in Table 2 also show that the exact difference between two languages is sensitive to which syntactic relations are included.

Examining the LAS scores for different subsets of relations, we find that FUN relations on average are parsed with almost 90% accuracy, to be compared with CORE and NON-CORE relations at about 75% and MWE at about 80%. This means that including FUN relations in the LAS score generally leads to higher scores and that languages with a high share of function words receive a boost. It is also worth noting that, even if CORE and NON-CORE relations are on average parsed with the same accuracy, there is considerable variation across languages. Most languages have a higher LAS score for CORE than

[1] For languages with more than one treebank, we selected the treebank without a suffix except in the case of Ancient Greek and Latin, where we selected the PROIEL treebanks to avoid including poetry.

		LAS				LAS Diff		
Language	LAS	CORE	NON-CORE	FUN	MWE	PUNCT	FUN	MWE
Ancient Greek	73.21	67.20	65.76	86.71	66.67	0.00	−6.98	0.01
Arabic	77.00	70.30	73.56	89.49	79.70	0.39	−4.01	−0.02
Basque	73.82	66.61	73.12	86.00	84.95	1.33	−2.79	−0.37
Bulgarian	85.85	76.80	81.68	97.02	83.40	−0.39	−3.94	0.04
Catalan	85.22	79.61	75.43	96.71	89.66	0.87	−7.23	−0.27
Chinese	73.66	63.03	71.09	87.61	91.30	0.86	−5.01	−0.03
Croatian	78.42	78.61	75.16	87.82	54.14	0.14	−2.87	0.49
Czech	84.67	83.35	81.26	94.14	89.15	0.10	−2.37	−0.07
Danish	79.82	81.96	72.82	89.74	82.14	0.49	−3.99	−0.05
Dutch	79.22	68.18	73.21	92.15	89.14	0.43	−5.47	−0.53
English	83.92	86.40	78.04	94.59	79.20	0.96	−4.08	0.28
Estonian	75.58	77.20	71.14	84.86	77.74	−0.31	−1.73	−0.06
Finnish	79.71	80.50	76.22	86.59	78.95	−0.72	−1.28	0.02
French	86.37	87.67	80.35	97.57	81.62	1.99	−6.57	0.15
German	82.58	85.94	75.79	93.52	81.13	1.38	−4.88	0.04
Gothic	76.00	73.27	71.79	86.19	78.79	0.00	−3.70	−0.00
Greek	83.06	82.12	76.82	94.41	86.67	1.28	−6.08	−0.01
Hebrew	80.79	68.04	71.42	96.48	92.78	1.54	−7.93	−1.01
Hindi	85.94	66.94	79.13	96.33	91.46	−0.52	−5.49	−0.66
Hungarian	77.34	77.20	74.74	87.61	91.85	1.67	−2.76	−0.60
Indonesian	75.89	78.67	67.80	90.04	83.65	1.71	−3.14	−1.62
Italian	86.65	80.46	80.49	98.10	91.86	1.70	−7.03	−0.11
Japanese	87.21	47.71	75.23	99.43	96.15	−1.30	−8.98	−0.96
Korean	58.94	49.95	58.19	51.93	56.62	−3.20	0.35	0.54
Latin	70.82	67.58	66.36	82.68	76.43	0.00	−3.97	−0.04
Latvian	73.37	73.18	67.91	84.09	80.44	−1.28	−1.69	−0.10
Lithuanian	56.35	52.84	53.09	72.86	61.54	1.32	−3.70	−0.09
Norwegian (bokmaal)	86.95	86.60	81.93	95.11	85.71	0.33	−3.37	0.03
Norwegian (nynorsk)	86.04	85.79	80.83	94.14	87.72	0.45	−3.51	−0.05
Old Church Slavonic	80.77	78.81	77.99	88.50	80.00	0.00	−2.50	0.00
Persian	81.25	66.63	79.29	90.96	79.16	−0.01	−3.55	0.17
Polish	87.94	85.69	84.72	95.05	80.00	−0.70	−1.63	0.01
Portuguese	87.46	87.10	80.81	97.34	95.00	1.61	−5.72	−0.23
Romanian	79.76	74.24	73.64	92.93	76.78	0.06	−4.83	0.12
Russian	79.79	81.72	76.03	93.56	89.01	1.04	−2.64	−0.38
Slovak	84.61	80.94	82.55	92.82	53.33	−0.34	−1.85	0.19
Slovenian	88.29	85.88	84.75	95.98	83.68	−0.14	−2.59	0.04
Spanish	84.51	79.88	76.74	96.02	81.87	1.04	−7.21	0.07
Swedish	80.08	82.98	75.34	90.54	69.11	1.00	−3.95	0.38
Turkish	60.02	51.84	58.76	75.14	45.43	−1.20	−1.86	0.87
Urdu	78.35	52.88	66.28	93.67	87.24	−0.78	−8.18	−1.27
Vietnamese	64.51	61.93	62.87	74.09	69.33	0.63	−1.77	−0.38
Average	79.09	74.15	74.05	89.77	80.01	0.32	−4.11	−0.13

Table 2: Evaluation scores for 42 UD treebanks (development sets). LAS = Labeled Attachment Score (overall and subsets). LAS Diff = Difference in LAS when excluding a subset of relations. Language families/branches with at least 2 members: Slavonic, Germanic, Romance, Finno-Ugric, Baltic, Greek, Indian, Semitic.

for NON-CORE, including most Germanic and Romance languages. But there are also languages that have a considerably lower score for CORE than for NON-CORE. The most extreme example is Japanese, where the difference is almost 30 percentage points, but large discrepancies can also be found for Basque, Chinese, Korean, Hindi and Persian. It seems that the parsing model used in the experiment fails to learn how core arguments are encoded in these languages, which is an interesting observation but not directly related to the topic of this paper.

Next we examine how the LAS score is affected when different subsets are excluded. Starting with PUNCT, we see that LAS sometimes increases and sometimes decreases. This may be due to inconsistent annotation of punctuation across treebanks, but it could also be due to differences in syntactic complexity, as short and simple sentences increase the frequency of easily predictable punctuation relations while long and complex sentences have the opposite effect. For most languages, the difference is less than a percentage point, but in a few cases it is quite substantial. For Korean, for example, excluding PUNCT from the LAS score decreases the score by over 3 percentage points. We also see that some of the classical languages (Ancient Greek, Gothic, Latin and Old Church Slavonic) lack punctuation completely, and the same would have been true if we had included treebanks of spoken language. This casts additional doubt on the inclusion of PUNCT in an evaluation score for syntactic analysis, and we will propose to exclude it in the new score.

As expected, the relations in FUN have a more significant and differential impact on the score. On average, LAS scores decrease by 4.11 points when these relations are not included, which is consistent with their being parsed more accurately than other relations, but the cross-linguistic variation is considerable. The largest drop is almost 9 points, for Japanese, and 12 languages have a drop of over 5 points. In this group, Romance languages like Catalan, Italian and Spanish and Greek (both ancient and modern) are prominent. At the opposite of the scale, Korean in fact sees a small improvement (0.35) when excluding FUN, and 7 languages have a drop smaller than 2 points. Finno-Ugric language like Estonian and Finnish are in this group, together with Turkish, Vietnamese and a few Baltic and Slavonic languages. This is mostly in line with our expectations based on linguistic typology (although the result for Japanese is unexpected) and consistent with the view that focusing on relations between content words will give a more balanced picture of parsing accuracy. Our new metric will therefore exclude all relations in FUN.

Omitting MWE has a more marginal effect on the evaluation scores. On average, LAS scores decrease by 0.13 points, and for most languages the difference (whether positive or negative) is less than 0.5 points, with a small number of outliers like Indonesian (-1.62), Urdu (-1.27), Hebrew (-1.01) and Turkish ($+0.87$). Based on these results, it is hard to draw any clear conclusions about the status of these relations in a cross-linguistically valid evaluation metric. For the time being, we will therefore simply leave them intact.

4 Content Labeled Attachment Score

Based on the theoretical discussion in the introduction and with further support from the empirical results in the previous section, we propose an alternative evaluation metric for UD parsing called Content-Word Labeled Attachment Score, abbreviated CLAS. CLAS is defined as the labeled F_1-score over all relations except relations in FUN and PUNCT. To make this precise, let S and G be the set of labeled dependencies in the system output and in the gold standard, respectively, and let $C(X)$ denote the subset of labeled dependencies in the set X that are not in FUN or PUNCT. Then we define precision (P), recall (R) and CLAS in the obvious way:

$$P(S,G) = \frac{|C(S) \cap C(G)|}{|C(S)|}$$

$$R(S,G) = \frac{|C(S) \cap C(G)|}{|C(G)|}$$

$$\text{CLAS}(S,G) = \frac{2 \cdot P(S,G) \cdot R(S,G)}{P(S,G) + R(S,G)}$$

The main idea behind this metric is that, by excluding function words, we are left with a set of relations that can be expected to occur with similar frequency across languages, although their structural realization may vary considerably. In this way, we can at least avoid the simple arithmetic biasing effects observed in the introduction and obtain scores that make more sense to compare across languages.

Language	LAS	CLAS	Diff	LAS Diff						
				DET	CLF	CASE	AUX	COP	MARK	CC
Ancient Greek	73.21	66.23	−6.98	−4.05	0.00	−1.67	0.01	0.07	−0.28	0.56
Arabic	77.00	72.95	−4.05	−0.09	0.00	−3.25	−0.08	0.05	−0.24	0.12
Basque	73.82	72.04	−1.79	−0.26	0.00	−0.35	−2.20	0.10	0.03	0.22
Bulgarian	85.85	80.42	−5.43	−0.35	0.00	−2.03	−0.52	0.02	−0.10	−0.26
Catalan	85.22	78.14	−7.08	−2.27	0.00	−2.11	−0.43	−0.02	−0.15	−0.11
Chinese	73.66	68.71	−4.95	−0.51	−0.18	−2.07	−0.06	−0.24	−0.85	−0.10
Croatian	78.42	75.25	−3.17	−0.22	0.00	−1.75	−0.44	0.18	−0.13	−0.05
Czech	84.67	81.91	−2.75	−0.25	0.00	−1.42	−0.14	−0.02	−0.17	−0.05
Danish	79.82	75.64	−4.18	−0.86	0.00	−1.22	−0.39	−0.11	−0.49	0.02
Dutch	79.22	73.28	−5.94	−2.09	0.00	−1.85	−0.16	0.05	−0.17	−0.06
English	83.92	80.42	−3.50	−1.09	0.00	−0.79	−0.52	−0.23	−0.36	−0.15
Estonian	75.58	73.02	−2.55	−0.17	0.00	−0.44	−0.48	0.17	−0.21	−0.36
Finnish	79.71	77.28	−2.42	−0.04	0.00	−0.20	−0.45	−0.05	−0.22	−0.14
French	86.37	81.86	−4.51	−2.18	0.00	−1.95	−0.27	−0.10	−0.08	−0.19
German	82.58	78.69	−3.89	−1.99	0.00	−1.10	−0.46	0.00	−0.07	−0.10
Gothic	76.00	72.31	−3.70	−0.73	0.00	−2.05	−0.04	0.00	−0.54	0.45
Greek	83.06	77.93	−5.13	−2.95	0.00	−1.29	−0.39	0.03	−0.01	0.00
Hebrew	80.79	73.67	−7.11	−2.12	0.00	−3.37	0.02	−0.03	−0.34	−0.21
Hindi	85.94	78.99	−6.95	−0.23	0.00	−2.72	−0.93	−0.08	−0.32	0.10
Hungarian	77.34	76.26	−1.08	−2.21	0.00	−0.32	−0.00	0.01	−0.31	0.38
Indonesian	75.89	74.19	−1.70	−0.20	0.00	−2.10	0.00	−0.18	−0.02	−0.30
Italian	86.65	80.94	−5.71	−2.45	0.00	−2.00	−0.31	−0.04	−0.17	−0.08
Japanese	87.21	74.03	−13.18	−0.03	0.00	−3.53	−1.73	−0.17	−0.71	−0.04
Korean	58.94	55.95	−2.98	0.62	0.00	−0.12	0.00	0.00	−0.10	−0.04
Latin	70.82	66.85	−3.97	−0.48	0.00	−2.11	−0.25	0.14	−0.48	0.03
Latvian	73.37	69.54	−3.84	−0.41	0.00	−0.77	−0.16	0.07	−0.13	−0.11
Lithuanian	56.35	53.26	−3.10	−1.32	0.00	−0.93	−0.12	−0.07	−0.37	−0.34
Norwegian (bokmaal)	86.95	83.40	−3.56	−0.50	0.00	−0.99	−0.38	−0.07	−0.37	−0.23
Norwegian (nynorsk)	86.04	82.56	−3.48	−0.63	0.00	−0.96	−0.33	−0.07	−0.48	−0.16
Old Church Slavonic	80.77	78.28	−2.50	−0.19	0.00	−1.56	−0.43	0.00	−0.33	0.52
Persian	81.25	77.22	−4.03	−0.33	0.00	−2.01	−0.21	0.16	−0.13	−0.37
Polish	87.94	85.01	−2.93	−0.15	0.00	−0.91	−0.13	−0.00	−0.07	−0.17
Portuguese	87.46	83.04	−4.42	−1.98	0.00	−1.95	−0.17	−0.10	−0.05	0.04
Romanian	79.76	73.96	−5.80	−1.00	0.00	−1.86	−0.49	−0.03	−0.26	−0.20
Russian	79.79	77.70	−2.09	−0.19	0.00	−1.80	−0.09	−0.01	−0.08	−0.23
Slovak	84.61	81.94	−2.67	−0.33	0.00	−1.33	0.07	0.05	−0.06	−0.04
Slovenian	88.29	84.96	−3.33	−0.19	0.00	−1.13	−0.38	0.05	−0.22	−0.20
Spanish	84.51	77.58	−6.93	−2.38	0.00	−2.31	−0.16	−0.02	−0.16	−0.13
Swedish	80.08	76.95	−3.13	−0.89	0.00	−1.10	−0.42	−0.04	−0.28	−0.32
Turkish	60.02	56.32	−3.70	−0.70	0.00	−0.90	−0.03	−0.40	0.05	0.26
Urdu	78.35	68.28	−10.07	−0.36	0.00	−3.78	−1.53	−0.13	−0.42	0.03
Vietnamese	64.51	63.15	−1.36	−0.63	0.00	−0.90	−0.07	−0.20	0.26	−0.03
Average	79.09	74.76	−4.32	−0.94	−0.00	−1.59	−0.36	−0.03	−0.23	−0.05

Table 3: Evaluation scores for 42 UD treebanks (development sets). LAS = Labeled Attachment Score. CLAS = Content-Word Labeled Attachment Score. LAS Diff = Difference in LAS when excluding a relation. Language families/branches with at least 2 members: Slavonic, Germanic, Romance, Finno-Ugric, Baltic, Greek, Indian, Semitic.

In order to explore the properties of the new metric, we present additional evaluation scores for the same parsing experiment in Table 3. The first three columns show LAS, CLAS and difference CLAS − LAS. The final seven columns show the difference in LAS when excluding the relations in FUN, one at a time.[2] Comparing CLAS to LAS, we see essentially the same picture as when excluding FUN from LAS in Table 2, although there is sometimes a combined effect when also excluding PUNCT. The average difference is −4.32 points, and the language-specific differences range from −1.08 for Hungarian to −13.18 for Japanese.

Among the languages that exhibit the smallest decrease, we find the Finno-Ugric languages (Estonian, Finnish, Hungarian) together with Basque and Indonesian, which are all agglutinating languages. More surprisingly, the low-decrease group also includes Vietnamese, which is usually described as an analytic language. The explanation seems to be a low LAS for FUN relations in Vietnamese, only 74.09 as shown in Table 2.

Slavonic languages, which are morphologically rich but not agglutinating, mostly have a relatively low decrease in the 2–3 point range, with the exception of Bulgarian (−5.43), which has developed in a typologically different direction from the other Slavonic languages in the sample. The closely related Baltic languages (Latvian, Lithuanian) behave similarly to Slavonic languages, but with a slightly higher decrease, and Germanic languages, which in general are less morphologically rich, are a little higher still with an average decrease of 3–4 points, although Dutch deviates from the general pattern by having an unexpectedly large decrease (−5.94).

Among the languages with the highest decrease we find Japanese, which is again somewhat unexpected and may have to do with particular annotation choices when applying the UD guidelines to Japanese. A high decrease is also observed for Indian languages (Hindi and Urdu), most of the Romance languages (especially Catalan, Italian and Spanish), Greek (both ancient and modern) and the Semitic languages (especially Hebrew).

Zooming in on the individual relations in FUN, we see that most of the difference can be attributed to functional relations in noun phrases, in particular *det* and *case*. (The third relation *clf* is currently only used in Chinese.) The *det* relation has the largest impact on Ancient Greek, followed by modern Greek, a group of Romance languages (French, Italian, Spanish), Hungarian and Hebrew. The *case* relation instead shows the largest effect for Urdu, Japanese, Hebrew and Arabic.

The remaining four FUN relations *aux*, *cop*, *mark* and *cc* have a less significant effect than the nominal relations. The *cc* relation is different from the rest in that scores sometimes go up when it is excluded. This effect is noticeable for two of the classical languages, Ancient Greek and Gothic, and for Hungarian. More research will be needed to find out why this is the case.

5 Conclusion

Proving that one evaluation metric is superior to another is very difficult in general. Ideally, we should show that it correlates better with independent quality criteria, but such criteria are often not available. This is especially tricky for a component task like syntactic parsing, where there are no real end users and where human assessments are notoriously unreliable. In this paper, we have instead relied primarily on rational argumentation. Since UD has been explicitly designed to capture cross-linguistic similarities in grammatical relations between content words, we argue that multilingual evaluation should focus primarily on these relations. The main metric should therefore exclude both function words and morphological features, to prevent bias towards either analytical or synthetic languages. (In addition, the main metric should exclude punctuation, which is completely absent in some of the treebanks and arguably not part of the syntactic structure.) To back up these rational arguments, we have presented empirical results from a parsing experiment, studying the effect of excluding different relations and showing that the new metric behaves, by and large, as we can expect based on typological considerations.

A common objection against only scoring dependencies between content words is that we will lose important information about parsing quality for languages where functional relations are important. If, in addition, we start tuning parsers on the new metric, we risk favoring systems that score well on a subset of relations at the expense of much lower accuracy on functional relations. We think these risks are exaggerated. First of all, the old LAS score is still available and might even

[2] All scores are F_1-scores, which explains why the differences under LAS Diff do not add up to the difference CLAS − LAS. In addition, CLAS also excludes PUNCT.

be the metric of choice for monolingual evaluation, where structural differences across languages are not relevant. Secondly, we are convinced that, in order to achieve high accuracy on argument and modifier relations, a parser must be able to recover other structures that provide information about these relations. For analytical languages, parsers will therefore be forced to pay attention to functional relations, even if they are not scored in the evaluation metric. For more synthetic languages, parsers will instead have to focus more on morphological information. Hence, by *directly* favoring accuracy on major grammatical relations, we are *indirectly* encouraging parsers to pay attention to grammatically relevant information, be it encoded in morphology, function words, or word order patterns. Therefore, we think the risk for an unwanted bias is in fact less of a problem than with the traditional LAS metric, where parsers can score well (for some languages) by being accurate mainly on functional relations, which are highly frequent and easy to parse.

At this point, it is still an empirical question which metric will give the right or wrong kind of bias, although some of the results reviewed in previous sections at least illustrate that the traditional LAS score can be severely inflated for some languages. More research will definitely be needed to better understand the effects of using different metrics, in particular experiments with different parsers and perhaps also different variants of the new metric, but we hope that the proposal made in this paper can be a first step towards more sound metrics for multilingual parser evaluation.

Acknowledgments

The research presented in this paper was supported by grant 2016-01817 of the Swedish Research Council.

References

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL)*, pages 149–164.

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 9th International Conference on Language Resources and Evaluation (LREC)*, pages 4585–4592.

Rebecca Dridan and Stephan Oepen. 2011. Parser evaluation using elementary dependency matching. In *Proceedings of the 12th International Conference on Parsing Technologies (IWPT)*, pages 225–230.

Rebecca Hwa, Philip Resnik, Amy Weinberg, and Okan Kolak. 2002. Evaluating translational correspondence using annotation projection. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 392–399.

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

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

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)*, pages 2216–2219.

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 of EMNLP-CoNLL 2007*, pages 915–932.

Joakim Nivre, Marco Kuhlmann, and Johan Hall. 2009. An improved oracle for dependency parsing with online reordering. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, pages 73–76.

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

Joakim Nivre. 2015. Towards a universal grammar for natural language processing. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 3–16. Springer.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*.

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.

Djamé Seddah, Sandra Kübler, and Reut Tsarfaty. 2014. Introducing the SPMRL 2014 shared task on parsing morphologically-rich languages. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, pages 103–109.

Reut Tsarfaty. 2013. A unified morpho-syntactic scheme of Stanford dependencies. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 578–584.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, pages 35–42.

Daniel Zeman, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič. 2012. HamleDT: To parse or not to parse? In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 2735–2741.

Daniel Zeman. 2008. Reusable tagset conversion using tagset drivers. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, pages 213–218.