Getting More from Morphology in Multilingual Dependency Parsing

Matt Hohensee and Emily M. Bender
University of Washington
Department of Linguistics
Box 354340
Seattle WA 98195-4340, USA
{hohensee, ebender}@uw.edu

Abstract

We propose a linguistically motivated set of features to capture morphological agreement and add them to the MSTParser dependency parser. Compared to the built-in morphological feature set, ours is both much smaller and more accurate across a sample of 20 morphologically annotated treebanks. We find increases in accuracy of up to 5.3% absolute. While some of this results from the feature set capturing information unrelated to morphology, there is still significant improvement, up to 4.6% absolute, due to the agreement model.

1 Introduction

Most data-driven dependency parsers are meant to be language-independent. They do not use any information that is specific to the language being parsed, and they often rely heavily on n-grams, or sequences of words and POS tags, to make parsing decisions. However, designing a parser without incorporating any specific linguistic details does not guarantee its language-independence; even linguistically naïve systems can involve design decisions which in fact bias the system towards languages with certain properties (Bender, 2011).

It is often taken for granted that using linguistic information necessarily makes a system language-dependent. But it is possible to design a linguistically intelligent parser without tuning it to a specific language, by modeling at a high level phenomena which appear cross-linguistically. Such a system is still language-independent; it does not require any knowledge or modeling of specific languages, but it does use linguistic knowledge to make the most of the available data. We present modifications to an existing system, MSTParser (McDonald et al., 2006), to incorporate a very simple model of morphological agreement. These modifications improve parsing performance across a variety of languages by making better use of morphological annotations.

2 Background and related work

2.1 Morphological marking of agreement

Most languages show some morphological agreement via inflected noun, adjective, verb, and determiner forms, although the degree to which this happens varies. At one end of the spectrum are analytic, or "morphologically impoverished", languages. An extreme example is Chinese, which shows no inflection at all; words do not take different forms depending on features such as person or gender. English has some inflection, but is relatively morphologically poor.

At the other end are synthetic or "morphologically rich" languages such as Czech, which has, inter alia, four genders and seven cases. In synthetic languages, words which are syntactically related in certain ways must agree: e.g., subject-verb agreement for gender or determiner-noun agreement for case (Corbett, 2006). Words participating in agreement may be marked explicitly for the property in question (via affixing or other morphological changes), or may possess it inherently (with no specific affix encoding the property). Treebanks are often annotated to reflect some or all of these properties; the level of detail depends on the annotation guidelines.

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 315–326,
Montréal, Canada, June 3-8, 2012. ©2012 Association for Computational Linguistics

zahraniční	investice	rostou
foreign	investment	grow
.F.PL.NOM	.F.3RD.PL.NOM	.3RD.PL.PRES
foreign	investments	grow
foreign	investment	grow
	.3RD.PL	.PL

Table 1: Sentence in Czech (Hajič, 1998) and English

A sample sentence in English and Czech (Table 1) demonstrates this contrast. In Czech, the adjective and noun agree for gender, number, and case, and the noun and verb agree for person and number. In the English version, only the noun and verb agree.

Agreement can be very useful for data-driven dependency parsing. A statistical parser can learn from training data that, for example, a third-person singular noun is a likely dependent of a verb marked as third-person singular. Similarly, it can learn that a determiner showing genitive case and a noun showing dative case are often not syntactically related.

It is often assumed that morphological complexity correlates with degree of variation in word order. This is because synthetic languages use inflection to mark the roles of constituents, while analytic languages generally assign these roles to specific phrase structural locations. Siewierska (1998) investigated this empirically and found that it holds to a certain extent: the absence of agreement and/or case marking predicts rigid word order, though their presence is not particularly predictive of flexible word order.

Many parsers rely on word order to establish dependencies, so they often perform best on languages with more rigid word order. Making use of morphological agreement could compensate for greater variation in word order and help to bring parsing performance on flexible-word-order languages up to par with that on rigid-word-order languages.

2.2 MSTParser

The CoNLL-X (Buchholz and Marsi, 2006) and CoNLL 2007 (Nivre et al., 2007) shared tasks focused on multilingual dependency parsing. Each system was trained on treebanks in a variety of languages and predicted dependency arcs and labels for POS-tagged data. The best performers in 2006 were MSTParser (McDonald et al., 2006), which we use here, and MaltParser (Nivre et al., 2006a).

MSTParser is a data-driven, graph-based parser which creates a model from training data by learning weights for arc-level features. The feature set includes combinations of the word and POS tag of the parent and child of each dependency arc; POS tags of words between the parent and child; and POS tags of the parent and child along with those of the preceding and following words. A similar feature set is conjoined with arc labels in order to perform labeling, and an optional set of "second-order" features includes analogous information about siblings.

Morphological features for an arc are generated by iterating over each pair in the cross product of the parent and child tokens' lists of attributes. For every such pair, thirteen groups of four features each are generated. The thirteen groups represent combinations of the head and child word forms/lemmas and attributes. Each group contains subgroups distinguished by whether they use word forms or lemmas and by whether or not they encode the direction and distance of the dependency. These features are summarized in Table 2. At run time, MSTParser finds the highest-scoring parse for each sentence according to the learned feature weights.

Decoding can be performed in projective or non-projective mode, depending on the type of trees desired. Projective trees are those in which every constituent (head plus all dependents) forms a complete subtree; non-projective parsing lacks this limitation.

2.3 Related work

The organizers of the CoNLL 2007 shared task noted that languages with free word order and high morphological complexity are the most difficult for dependency parsing (Nivre et al., 2007). Most of the participants took language-independent approaches toward leveraging this complexity into better performance: generating machine learning features based on each item in a token's list of morphological attributes (Nivre et al., 2006b; Carreras et al., 2006); using the entire list as an atomic feature (Chang et al., 2006; Titov and Henderson, 2007); or generating features based on each pair of attributes in the cross-product of the lists of a potential head and dependent (McDonald et al., 2006; Nakagawa, 2007).

Language-specific uses of morphological information have included using it to disambiguate function words (Bick, 2006) or to pick out finite verbs

```
<hdIdx>⋆<dpIdx>=<{hdForm|hdLemma}>(<dir+dist>)
<hdIdx>⋆<dpIdx>=<{dpForm|dpLemma}>(<dir+dist>)
<hdIdx>⋆<dpIdx>=<hdAtt>(<dir+dist>)
<hdIdx>⋆<dpIdx>=<dpAtt>(<dir+dist>)
<hdIdx>⋆<dpIdx>=<{hdForm|hdLemma}><{dpForm|dpLemma}>(<dir+dist>)
<hdIdx>⋆<dpIdx>=<{hdForm|hdLemma}><hdAtt>(<dir+dist>)
<hdIdx>⋆<dpIdx>=<{hdForm|hdLemma}><dpAtt>(<dir+dist>)
<hdIdx>⋆<dpIdx>=<{dpForm|dpLemma}><dpAtt>(<dir+dist>)
<hdIdx>⋆<dpIdx>=<{dpForm|dpLemma}><hdAtt>(<dir+dist>)
<hdIdx>⋆<dpIdx>=<hdAtt><dpAtt>(<dir+dist>)
<hdIdx>⋆<dpIdx>=<{hdForm|hdLemma}><hdAtt><dpAtt>(<dir+dist>)
<hdIdx>⋆<dpIdx>=<{dpForm|dpLemma}><hdAtt><dpAtt>(<dir+dist>)
<hdIdx>⋆<dpIdx>=<{hdForm|hdLemma}><{dpForm|dpLemma}><hdAtt><dpAtt>(<dir+dist>)
```

Table 2: Original MSTParser feature templates. `hdForm` and `dpForm` are the head and dependent word forms; `hdLemma` and `dpLemma` are the lemmas. `hdAtt` and `dpAtt` are the morphological attributes; `hdIdx` and `dpIdx` are their indices. `dir+dist` is a string encoding the direction and length of the arc. Each line represents one feature.

Unlabeled	`<attr>_agrees,head=<headPOS>,dep=<depPOS>`
	`<attr>_disagrees,head=<headPOS>,dep=<depPOS>`
	`head_<attr=value>,head=<headPOS>,dep=<depPOS>`
	`dep_<attr=value>,head=<headPOS>,dep=<depPOS>`

Labeled	`<attr>_agrees&label=<label>,head=<headPOS>,dep=<depPOS>`
	`<attr>_disagrees&label=<label>,head=<headPOS>,dep=<depPOS>`
	`head_<attr=value>&label=<label>,head=<headPOS>,dep=<depPOS>`
	`dep_<attr=value>&label=<label>,head=<headPOS>,dep=<depPOS>`

Table 3: Agreement feature templates. `headPOS` and `depPOS` are the head and dependent coarse POS tags.

(Carreras et al., 2006). Schiehlen and Spranger (2007) used language-specific rules to add detail to other features, such as fine-grained POS tags or lemmas. Attardi et al. (2007) modeled agreement explicitly, generating a morphological agreement feature whenever two tokens possess the same value for the same linguistic attribute. The authors note accuracy improvements of up to 0.5% for Italian and 0.8% for Catalan using a transition-based parser. A similar approach was used by Goldberg and Elhadad (2010), who improved the accuracy of their transition-based Hebrew parser by adding features for gender and number agreement in noun phrases.

The potential of morphological information to improve parsing performance has been documented in numerous experiments using MaltParser and with various morphological attributes as machine learning features, on several morphologically rich languages, including: Russian (Nivre et al., 2008); Swedish (Øvrelid and Nivre, 2007); Bangla, Telugu, and Hindi (Nivre, 2009); Turkish (Eryiğit et al., 2008); and Basque (Bengoetxea and Gojenola, 2010). These experiments, however, did not include any higher-level features such as agreement.

Goldberg and Elhadad (2009) found that using morphological features increased the accuracy of MSTParser on Hebrew only when the morphological annotations were gold-standard; automatic annotations decreased accuracy, although MaltParser showed improvement with both gold and automatic annotations. The accuracy of MaltParser on Arabic was improved by different types of morphological features depending on whether gold or automatic annotations were used (Marton et al., 2010).

As far as we can tell, no language-independent approaches to utilizing morphological data thus far have taken advantage of agreement specifically. We take a linguistically informed approach, maintaining language-independence, by explicitly modeling agreement between head and dependent morphology.

3 Methodology

3.1 Modifications to parser

Our approach builds on the observation that there are two kinds of information marked in morphology: symmetric, recorded on both head and depen-

ID	TOKEN	CPOS	MORPH	HEAD	REL	Gloss
1	Vznikají	VERB	num=PL|per=3	0	ROOT	arise.3RD.PL
2	zbytečné	ADJ	num=PL|gen=I|case=NOM	3	ATR	unnecessary.PL.INAN.NOM
3	konflikty	NOUN	num=PL|gen=I|case=NOM	1	SBJ	conflicts.PL.INAN.NOM

num_agrees,head=NOUN,dep=ADJ	num_agrees,head=VERB,dep=NOUN
num_agrees&label=ATR,head=NOUN,dep=ADJ	num_agrees&label=SBJ,head=VERB,dep=NOUN
gen_agrees,head=NOUN,dep=ADJ	head_per=3,head=VERB,dep=NOUN
gen_agrees&label=ATR,head=NOUN,dep=ADJ	head_per=3&label=SBJ,head=VERB,dep=NOUN
case_agrees,head=NOUN,dep=ADJ	dep_gen=I,head=VERB,dep=NOUN
case_agrees&label=ATR,head=NOUN,dep=ADJ	dep_gen=I&label=SBJ,head=VERB,dep=NOUN
	dep_case=NOM,head=VERB,dep=NOUN
	dep_case=NOM&label=SBJ,head=VERB,dep=NOUN

Table 4: Sample sentence (Hajič, 1998) and agreement features generated

dent, and asymmetric, marked on only one or the other. Symmetric information provides a natural, effectively non-lossy type of back-off that parsers can take advantage of; all that matters is whether the information on the head and dependent match.[1] Furthermore, we don't need to know ahead of time which types of morphological information are symmetric. This is extracted from the annotations.

In order to take advantage of this property of natural language, we devised a set of features which model agreement. These allow the learner to operate at a higher level, using agreement itself as a feature rather than having to discover agreement and forming generalizations about whether tokens which agree (or disagree) in various ways are related. Since agreement appears cross-linguistically, such features are applicable to a diverse set of languages.

Since MSTParser breaks down every parse into a set of arcs, our features are defined at the arc level. Each arc is a head and dependent pair, and each of those tokens has a list of morphological features in the normalized form `attribute=value`. We compare these lists and add, for every attribute which is present in both, either an agreement or a disagreement feature, depending on whether the head and dependent have the same value for that attribute. This feature encapsulates the attribute, but not the value, as well as the coarse POS tags of the head and the dependent. If an attribute is present in only one of the lists, we add a feature encapsulating whether the token is the head or the dependent, the single morphological feature (attribute and value), and the two coarse POS tags. We also generate both types of features conjoined with the arc label. Like the original feature set, we include only first-order morphological features. See Table 3 for a summary. A sample sentence in a simplified CoNLL format and the features it would trigger are shown in Table 4.[2]

We hypothesize that these agreement features will function as a type of back-off, allowing the parser to extract more information from the morphological marking. For instance, they can capture case agreement between a determiner and noun. We expect that this would lead to higher parsing accuracy, especially when training on smaller datasets, where morphological data might be sparse.

We made a slight modification to the parser so that underscores used in the treebanks to indicate the absence of morphological annotation for a token were not themselves treated as morphological information. This was necessary to ensure that all feature configurations performed identically on treebanks with no morphological information. Depending on the treebank, this increased or decreased the performance of the system slightly (by less than 0.5%).

3.2 Data collection and preparation

We gathered a range of dependency treebanks, representing as many language families as possible (Table 5). Many of these used the CoNLL shared task treebank format, so we adopted it as well, and con-

[1] If an attribute is marked on both head and dependent and the value matches, the specific value should not affect the probability or possibility of the dependency relationship. If the same attribute is marked on both elements but is independent (not a matter of agreement) we risk losing information, but we hypothesize that such information is unlikely to be very predictive.

[2] A more complete description of the system, as well as source code, can be found in (Hohensee, 2012).

Language	ISO	Treebank	Num. sents.	Ref. size	Avg. atts.	Reference
Hindi-Urdu	hin	HUTB	3,855	2,800	3.6	(Bhatt et al., 2009)
Hungarian	hun	Szeged DTB	92,176	9,000	3.3	(Vincze et al., 2010)
Czech	ces	PDT 1.0	73,068	9,000	2.8	(Hajič, 1998)
Tamil	tam	TamilTB v0.1	600	600	2.8	(Ramasamy and Žabokrtský, 2011)
Slovene	slv	SDT	1,998	1,500	2.6	(Džeroski et al., 2006)
Danish	dan	DDT	5,512	5,500	2.4	(Kromann, 2003)
Basque	eus	3LB*	3,175	2,800	2.4	(Aduriz et al., 2003)
Dutch	nld	Alpino	13,735	9,000	2.4	(Van der Beek et al., 2002)
Latin	lat	LDT	3,423	2,800	2.4	(Bamman and Crane, 2006)
Bulgarian	bul	BulTreeBank	13,221	9,000	2.1	(Simov et al., 2004)
Greek (ancient)	grc	AGDT	21,104	9,000	2.1	(Bamman et al., 2009)
Finnish	fin	Turku	4,307	2,800	2.0	(Haverinen et al., 2010)
German	deu	NEGRA	3,427	2,800	2.0	(Brants et al., 1999)
Turkish	tur	METU-Sabanci	5,620	5,500	1.6	(Oflazer et al., 2003)
Catalan	cat	CESS-ECE*	3,512	2,800	1.5	(Martı et al., 2007)
Arabic	ara	PADT 1.0	2,367	2,300	1.2	(Hajic et al., 2004)
Italian	ita	TUT	2,858	2,800	1.1	(Bosco et al., 2000)
Portuguese	por	Floresta	9,359	9,000	1.0	(Afonso et al., 2002)
Hebrew (modern)	heb	DepTB	6,214	5,500	0.9	(Goldberg, 2011)
English	eng	Penn*	49,208	9,000	0.4	(Marcus et al., 1993)
Chinese	cmn	Penn Chinese	28,035	9,000	0.0	(Xue et al., 2005)

*Acquired as part of NLTK (Bird et al., 2009)

Table 5: Language, ISO 639-2 code, treebank name, total number of sentences, reference size, average number of morphological attributes per token, and reference for each treebank used, ordered by average number of attributes.

verted the other treebanks to the same. It includes for each token: position in the sentence; the token itself; a lemma (not present in all datasets); a coarse POS tag; a fine POS tag; a list of morphological features; the token's head; and the label for the dependency relation to that head.[3] We retained all punctuation and other tokens in the treebanks.

The POS tagsets used in the treebanks varied widely. We normalized the coarse tags to the universal twelve-tag set suggested by Petrov et al. (2011), in order to ensure that every treebank had coarse tags for use in the agreement features, and to make the features easier to interpret. It is unlikely that information was lost in this process: for treebanks with one set of tags, information was added, and for those with two, the universal tags aligned closely with the coarse tags already in the data.

Two of the treebanks we used included no morphological information. We included the Penn Chinese Treebank as a representative of analytic languages.[4] We also included part of the (English) Penn Treebank, converted to dependency trees. For this data we generated morphological annotations based on fine POS tags, consisting of person and number information for nouns and verbs, and person, number, and case information for pronouns. The German NEGRA corpus includes detailed morphological annotations for about 3,400 sentences (of 20,600), and we used only that portion.

Note that the amount of morphological information present in any given treebank is a function of the morphological properties of the language as well as the annotation guidelines: annotations do not necessarily encode all of the morphological information which is actually marked in a language. Furthermore, the presence of a morphological feature does not imply that it participates in an agreement relationship; it merely encodes some piece of morphological information about the token. Finally, annotation guidelines vary as to whether they provide for the explicit marking of morphological properties which are inherent to a lemma (e.g., gender on nouns) and not marked by separate affixes.

[3] The original format also included two more fields, projective head and label; neither is used by MSTParser.

[4] Dependency trees were generated from the Penn Chinese Treebank using the Penn2Malt converter: `http://w3.msi.vxu.se/~nivre/research/Penn2Malt.html`.

We normalized all morphological annotations to the form attribute=value (e.g., case=NOM). For treebanks that provided values only, this involved adding attribute names, obtained from the annotation guidelines. The attributes person, number, gender, and case appeared often; also included in some data were verb tense, adjective degree, and pronoun type (e.g., personal, possessive, or reflexive). We normalized all features in the data, regardless of whether they participate in any agreement relations.

Many of the treebanks include data from multiple domains; to minimize the effects of this, we randomized the order of sentences in each treebank.

3.3 Experimental setup

All experiments were performed using 5-fold cross-validation. Reported accuracies, run times, and feature counts are averages over all five folds. We ran experiments on multiple cross-validation dataset sizes in order to assess the performance of our model when trained on different amounts of data. For each treebank, we report results on a "reference size": 9,000 sentences or the largest size available (for treebanks of less than 9,000 sentences).

For evaluation, we used the module built into MSTParser. We focused on the unlabeled accuracy score (percentage of tokens with correctly assigned heads, ignoring labels). We also looked at labeled accuracies, but found they displayed trends very similar, if not identical, to the unlabeled scores.

4 Results

We ran the system on each treebank at all dataset sizes in projective and non-projective modes, using no morphological features. For each language, subsequent tests used the algorithm which performed better (or non-projective in the case of a tie).

4.1 Overall results

We ran the parser on each treebank with each of four feature configurations: one with no morphological features (no-morph); one with the original morphological features (orig; Table 2); one using the agreement features (agr; Table 3); and one using both feature sets (agr+orig).

Table 6 displays the unlabeled accuracy, run time, and feature counts when parsing each treebank using each feature configuration at the reference size, with the highest accuracy highlighted. Excluding Chinese, agr generated the best performance in all but two cases, outperforming orig by margins ranging from 0.8% (Arabic) to 5.3% (Latin) absolute. In the other cases, agr+orig outperformed agr slightly. In all cases, the total number of machine learning features was approximately the same for no-morph and agr, and for orig and agr+orig, because the number of morphological features generated by orig is very large compared to the number generated by agr. Performance was noticeably faster for the two smaller feature configurations.

Figure 1 shows the error reduction of orig, agr, and agr+orig relative to no-morph, at the reference size. Despite its relative lack of morphological inflection, English shows a fairly high error reduction, because parsing performance on English was already high. Similarly, error reduction on some of the morphologically rich languages is lower because baseline performance was low. Calculating the correlation coefficient (Pearson's r) between average morphological attributes per token and error reduction gives $r = 0.608$ for orig, $r = 0.560$ for agr, and $r = 0.428$ for agr+orig, with $p < 0.01$ for the first two and $p < 0.10$ for the last, indicating moderate correlations for all feature sets.

The strength of these correlations depends on several factors. Languages differ in what information is marked morphologically, and in number of agreement relationships. Annotation schemes vary in what morphological information they encode, and in how relevant that information is to agreement. Some morphologically complex languages have rigid word order, leading to better performance with no morphological features at all, and limiting the amount of improvement that is possible. Finally, it is possible that a stronger correlation is obscured by other effects due to feature set design, as we will find later.

4.2 Performance vs. dataset size

Figures 2 presents unlabeled accuracy when parsing Czech with the orig and agr configurations. Improvement with agr is roughly uniform across all dataset sizes; this was the general trend for all treebanks. This is somewhat unexpected; we had predicted that the agreement features would be more helpful at smaller dataset sizes.

	`no-morph`			`orig`			`agr`			`agr+orig`		
Lang.	UAC	time	feats	UAC	Δtime	Δfeats	UAC	Δtime	Δfeats	UAC	Δtime	Δfeats
hin	90.0	1.4k	1.6m	92.0	116%	893%	**93.8**	50%	1%	93.0	144%	893%
hun	87.9	4.6k	5.3m	88.7	201%	687%	**90.3**	10%	0%	89.9	159%	687%
ces	80.9	3.3k	4.8m	81.6	71%	454%	**85.5**	27%	0%	84.5	114%	454%
tam	79.0	0.1k	0.5m	79.7	237%	329%	**82.1**	64%	1%	81.1	279%	330%
slv	80.8	0.8k	1.0m	80.4	103%	352%	**81.9**	21%	1%	80.8	129%	353%
dan	87.7	2.0k	1.6m	88.4	71%	256%	**89.3**	24%	0%	**89.3**	86%	256%
lat	61.7	1.8k	1.6m	65.0	54%	306%	**70.3**	91%	0%	68.6	119%	306%
nld	88.2	2.0k	3.6m	89.0	83%	270%	**90.5**	16%	0%	90.3	98%	270%
eus	78.7	0.7k	1.7m	80.2	80%	229%	**82.3**	10%	0%	82.3	78%	230%
bul	89.9	1.7k	2.6m	90.1	60%	221%	**93.0**	14%	0%	92.5	54%	222%
grc	74.9	8.6k	3.8m	76.9	36%	314%	**80.7**	45%	0%	79.5	70%	314%
deu	90.0	0.9k	1.3m	90.8	33%	189%	**92.0**	1%	0%	91.7	50%	186%
fin	73.3	0.7k	2.4m	76.3	74%	244%	**79.1**	23%	1%	78.7	84%	245%
tur	80.2	1.2k	2.1m	81.5	13%	178%	81.6	−2%	0%	**81.7**	29%	178%
cat	81.8	3.0k	2.5m	81.9	2%	142%	**84.9**	-9%	0%	84.0	−2%	143%
ara	78.0	3.2k	2.0m	78.1	65%	94%	**78.9**	23%	0%	78.7	20%	94%
ita	88.3	4.2k	1.8m	88.9	−3%	59%	90.2	9%	0%	**90.3**	6%	59%
por	88.1	6.4k	5.0m	88.1	18%	46%	**89.0**	−3%	0%	88.9	27%	46%
heb	87.4	4.3k	3.1m	87.4	−18%	31%	**89.2**	−16%	0%	89.1	−5%	31%
eng	88.1	5.2k	3.1m	88.0	5%	7%	**90.6**	3%	0%	**90.6**	−9%	8%
cmn	**82.4**	7.5k	6.0m	**82.4**	37%	0%	**82.4**	16%	0%	**82.4**	23%	0%

Table 6: Unlabeled accuracy, run time in seconds, and number of features for all treebanks and feature configurations. Run time and number of features for `orig`, `agr`, and `agr+orig` are given as percent change relative to `no-morph`

4.3 Gold vs. automatic tags

The Hebrew treebank includes both automatically generated and gold standard POS and morphological annotations. In order to test how sensitive the agreement features are to automatically predicted morphological information, tests were run on both versions at the reference size. These results are not directly comparable to those of Goldberg and Elhadad (2009), because of the parser modifications, POS tag normalization, and cross-validation described earlier. Comparing results qualitatively, we find less sensitivity to the automatic tags overall, and that the `orig` features improve accuracy even when using automatic tags.

Results appear in Table 7. Using the automatic data affects all feature sets negatively by 2.1% to 2.9%. Since the `no-morph` parser was affected the most, it appears that this decrease is due largely to errors in the POS tags, rather than the morphological annotations. The `orig` features compensate for this slightly (0.2%), and the `agr` features more (0.8%); this indicates that including even automatic morphological information can compensate for incorrect POS tags, and that the `agr` feature configuration is the most robust when given predicted tags.

Feature configuration	**Acc. on gold data**	**Acc. on auto data**	**Difference**
`no-morph`	87.4	84.5	−2.9
`orig`	87.4	84.7	−2.7
`agr`	89.3	87.2	−2.1
`agr+orig`	89.1	86.9	−2.2

Table 7: Unlabeled accuracy on Hebrew dataset, with gold and automatic POS and morphological annotations

4.4 PPL feature

Examining the feature weights from the first cross-validation fold when running the `agr` feature configuration on the Czech dataset indicated that 323 of the 1,000 highest-weighted features are agreement features. Of these, 79 are symmetric ("agrees" or "disagrees") `agr` features, and 244 asymmetric. This was unexpected, as the symmetric features would seem to be more useful, and it suggested that the labeled asymmetric `agr` features might be important for reasons other than their modeling of morphological information. Careful analysis of the MSTParser feature set revealed that it does not include a feature which incorporates head POS, dependent POS, and dependency label. We hypothesized that the labeled asymmetric `agr` features were highly ranked

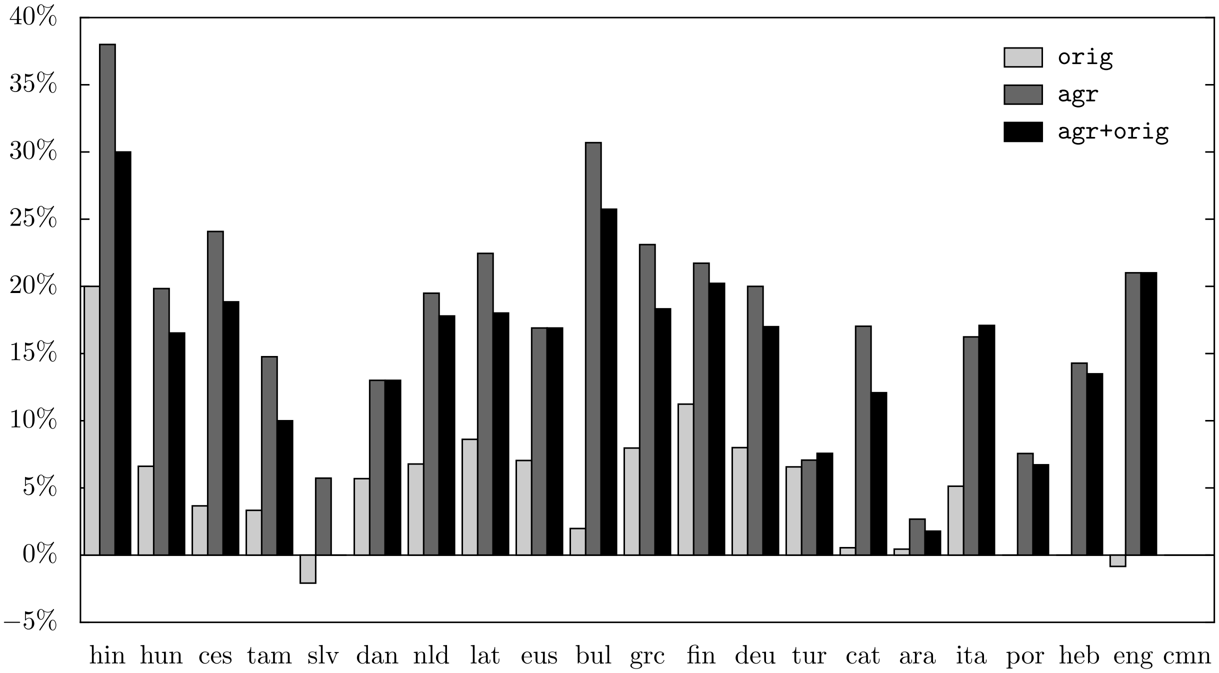

Figure 1: Error reduction relative to `no-morph` vs. language

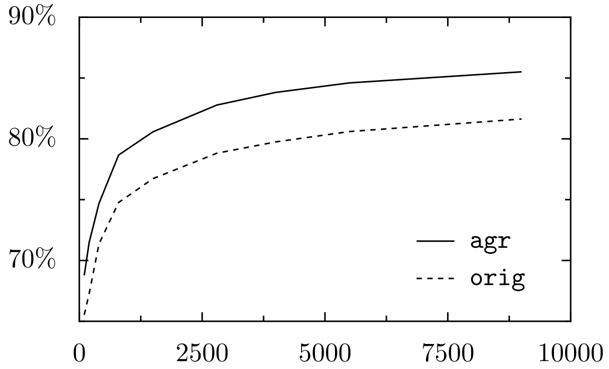

Figure 2: Unlabeled accuracy vs. num. sentences, Czech

because they capture these three arc features, not because they include with morphological information.

To test this, we added a single feature template to MSTParser which encapsulates head POS, dependent POS, and dependency label (the POS-POS-label, or PPL, feature). Running a subsequent experiment on the Czech data and looking at feature weights from the same cross-validation fold, 278 of the 1,000 highest-weighted features were PPL features, and 187 were asymmetric `agr` features. This indicated that the improvement seen with `agr` features was indeed due partly to their inclusion of features combining label and head and dependent POS.

All feature configurations were run on all treebanks with the PPL feature included; results appear in Table 8. Performance increases from `orig` to `agr` are generally smaller, with a maximum of 4.6% absolute. This is seen especially on languages with less morphological information, such as English and Hebrew; this indicates that for those languages, most of the previous improvement was due not to agreement modeling, but to the PPL effect.

Calculating Pearson's r between morphological features per token and the new error reduction data gives a stronger correlation coefficient of 0.748 for `agr`, with $p < 0.01$, demonstrating that improvement due solely to agreement modeling correlates strongly with quantity of morphological information. The earlier error reduction data were likely polluted by improvement due to capturing the PPL information. Correlation for the other feature configurations is still moderate (0.506 with $p < 0.02$ for `orig` and 0.621 with $p < 0.01$ for `agr+orig`).

5 Future work

In future work, we plan to experiment with more careful normalization of treebanks. For instance, if an adjective can agree with either a masculine or a feminine noun, annotating it with both `gen=M`

	`no-morph`			`orig`			`agr`			`agr+orig`		
Lang.	UAC	time	feats	UAC	Δtime	Δfeats	UAC	Δtime	Δfeats	UAC	Δtime	Δfeats
hin	90.0	1.4k	1.6	92.0	116%	893%	**93.8**	50%	1%	93.0	144%	893%
hun	87.9	4.6k	5.3	88.7	201%	687%	**90.3**	10%	0%	89.9	159%	687%
ces	80.9	3.3k	4.8	81.6	71%	454%	**85.5**	27%	0%	84.5	114%	454%
tam	79.0	0.1k	0.5	79.7	237%	329%	**82.1**	64%	1%	81.1	279%	330%
slv	80.8	0.8k	1.0	80.4	102%	352%	**81.8**	21%	0%	80.8	129%	353%
dan	87.8	2.0k	1.6	88.4	71%	256%	**89.3**	24%	0%	**89.3**	86%	256%
lat	61.7	1.8k	1.6	65.0	54%	306%	**70.3**	91%	0%	68.6	119%	306%
nld	88.2	2.0k	3.6	89.0	83%	270%	**90.5**	16%	0%	90.3	98%	270%
eus	78.7	0.7k	1.7	80.2	80%	229%	**82.3**	10%	0%	**82.3**	78%	230%
bul	89.9	1.7k	2.6	90.2	60%	221%	**93.0**	14%	0%	92.5	54%	222%
grc	74.9	8.6k	3.8	77.0	36%	314%	**80.7**	45%	0%	79.5	70%	314%
deu	90.0	0.9k	1.3	90.8	33%	189%	**92.0**	1%	0%	91.7	50%	186%
fin	73.3	0.7k	2.4	76.3	74%	244%	**79.1**	23%	0%	78.7	84%	245%
tur	80.2	1.2k	2.1	81.5	13%	178%	81.6	-2%	0%	**81.7**	29%	178%
cat	81.8	3.0k	2.5	81.9	1%	142%	**84.9**	-9%	0%	84.0	-2%	143%
ara	77.6	5.4k	1.8	77.7	20%	100%	**78.2**	-8%	0%	78.0	4%	100%
ita	88.4	4.2k	1.8	88.9	-2%	59%	90.2	9%	0%	**90.3**	6%	59%
por	88.1	6.4k	5.0	88.2	18%	46%	**89.0**	-3%	0%	88.9	27%	46%
heb	87.4	4.3k	3.1	87.4	-18%	31%	**89.2**	-16%	0%	89.1	-5%	31%
eng	88.1	5.2k	3.1	88.0	5%	7%	**90.6**	3%	0%	**90.6**	-9%	7%
cmn	**82.4**	7.5k	6.0	**82.4**	37%	0%	**82.4**	16%	0%	**82.4**	23%	0%

Table 8: Unlabeled accuracy, run time in seconds, and number of features with PPL feature included. Run time and number of features for `orig`, `agr`, and `agr+orig` are given as percent change relative to `no-morph`.

and `gen=F` (rather than `gen=X`) would ensure that agreement with a noun of either gender would be captured by our features. Furthermore, we may experiment with filtering morphological information based on part-of-speech, on attribute, or on whether the attribute participates in any agreement relationships. We also intend to perform feature selection on the original feature set, and investigate the importance of labeled morphological features, which are included in `agr` but not in `orig`. Finally, we plan to develop metrics to measure the degree of word order flexibility in a treebank, in order to explore the extent to which it correlates with the degree of improvement achieved by our system.

6 Conclusions

We developed a simple, language-independent model of agreement to better leverage morphological data in dependency parsing. Testing on treebanks containing varying amounts of morphological information resulted in substantial improvements in parsing accuracy while reducing feature counts and run times significantly. Although originally intended to compensate for lower accuracy on morphologically rich languages, the model improved performance on all treebanks with any morphological information.

We acknowledge that because our model was tested on treebanks which differ widely in annotation guidelines, variables such as the amount of morphological information included and the treatment of non-projective parses and coordination could affect parsing performance. We did not delve into these factors. However, we believe this is part of the strength of the approach: we were able to achieve performance gains without any detailed knowledge of the languages and treebanks used.

We hope these results will encourage similarly linguistically motivated design in future systems. This case study provides strong evidence that incorporating linguistic knowledge into NLP systems does not preclude language independence, and indeed may enhance it, by leveling performance across typologically differing languages.

Acknowledgments

We would like to thank everyone who assisted us in gathering treebanks, particularly Maite Oronoz and her colleagues at the University of the Basque Country and Yoav Goldberg, as well as three anonymous reviewers for their comments.

References

I. Aduriz, M.J. Aranzabe, J.M. Arriola, A. Atutxa, A.D. de Ilarraza, A. Garmendia, and M. Oronoz. 2003. Construction of a Basque dependency treebank. In *Proc. of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, pages 201–204.

S. Afonso, E. Bick, R. Haber, and D. Santos. 2002. Floresta Sintá(c)tica: A treebank for Portuguese. In *Proc. of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, page 1698.

G. Attardi, F. DellOrletta, M. Simi, A. Chanev, and M. Ciaramita. 2007. Multilingual dependency parsing and domain adaptation using DeSR. In *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 1112–1118.

D. Bamman and G. Crane. 2006. The design and use of a Latin dependency treebank. In *Proc. of the Fifth International Workshop on Treebanks and Linguistic Theories (TLT 2006)*, pages 67–78.

D. Bamman, F. Mambrini, and G. Crane. 2009. An ownership model of annotation: The Ancient Greek Dependency Treebank. In *Proc. of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT8)*, pages 5–15.

E.M. Bender. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology: Special Issue on Interaction of Linguistics and Computational Linguistics*, 6(3):1–26.

K. Bengoetxea and K. Gojenola. 2010. Application of different techniques to dependency parsing of Basque. In *Proc. of the First Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, pages 31–39. Association for Computational Linguistics.

R. Bhatt, B. Narasimhan, M. Palmer, O. Rambow, D.M. Sharma, and F. Xia. 2009. A multi-representational and multi-layered treebank for Hindi/Urdu. In *Proc. of the Third Linguistic Annotation Workshop (LAW III)*, pages 186–189. Association for Computational Linguistics.

E. Bick. 2006. LingPars, a linguistically inspired, language-independent machine learner for dependency treebanks. In *Proc. of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 171–175. Association for Computational Linguistics.

S. Bird, E. Klein, and E. Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

C. Bosco, V. Lombardo, D. Vassallo, and L. Lesmo. 2000. Building a treebank for Italian: a data-driven annotation schema. In *Proc. of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, pages 99–106.

T. Brants, W. Skut, and H. Uszkoreit. 1999. Syntactic annotation of a German newspaper corpus. *Treebanks: Building and using parsed corpora*, 20:73.

S. Buchholz and E. Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proc. of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164. Association for Computational Linguistics.

X. Carreras, M. Surdeanu, and L. Marquez. 2006. Projective dependency parsing with perceptron. In *Proc. of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 181–185. Association for Computational Linguistics.

M.W. Chang, Q. Do, and D. Roth. 2006. A pipeline model for bottom-up dependency parsing. In *Proc. of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 186–190. Association for Computational Linguistics.

G.G. Corbett. 2006. *Agreement*. Cambridge University Press.

S. Džeroski, T. Erjavec, N. Ledinek, P. Pajas, Z. Žabokrtsky, and A. Žele. 2006. Towards a Slovene dependency treebank. In *Proc. of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*.

G. Eryiğit, J. Nivre, and K. Oflazer. 2008. Dependency parsing of Turkish. *Computational Linguistics*, 34(3):357–389.

Y. Goldberg and M. Elhadad. 2009. Hebrew dependency parsing: Initial results. In *Proc. of the 11th International Conference on Parsing Technologies (IWPT'09)*, pages 129–133. Association for Computational Linguistics.

Y. Goldberg and M. Elhadad. 2010. Easy-first dependency parsing of Modern Hebrew. In *Proc. of the First Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, pages 103–107. Association for Computational Linguistics.

Yoav Goldberg. 2011. *Automatic Syntactic Processing of Modern Hebrew*. Ph.D. thesis, Ben Gurion University.

J. Hajic, O. Smrz, P. Zemánek, J. Šnaidauf, and E. Beška. 2004. Prague Arabic dependency treebank: Development in data and tools. In *Proc. of the NEMLAR International Conference on Arabic Language Resources and Tools*, pages 110–117.

Jan Hajič. 1998. Building a syntactically annotated corpus: The Prague Dependency Treebank. In Eva Hajičová, editor, *Issues of Valency and Meaning: Studies in Honor of Jarmila Panevová*, pages 12–19. Prague Karolinum, Charles University Press.

Katri Haverinen, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Filip Ginter, and Tapio Salakoski. 2010. Treebanking Finnish. In *Proc. of the Ninth International Workshop on Treebanks and Linguistic Theories (TLT9*, volume 9, pages 79–90.

M. Hohensee. 2012. It's only morpho-logical: Modeling agreement in cross-linguistic dependency parsing. Master's thesis, University of Washington.

M.T. Kromann. 2003. The Danish Dependency Treebank and the DTAG treebank tool. In *Proc. of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, pages 217–220.

M.P. Marcus, M.A. Marcinkiewicz, and B. Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

M.A. Martı, M. Taulé, L. Márquez, and M. Bertran. 2007. *CESS-ECE: A multilingual and multilevel annotated corpus*.

Y. Marton, N. Habash, and O. Rambow. 2010. Improving Arabic dependency parsing with lexical and inflectional morphological features. In *Proc. of the First Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, pages 13–21. Association for Computational Linguistics.

R. McDonald, K. Lerman, and F. Pereira. 2006. Multilingual dependency analysis with a two-stage discriminative parser. In *Proc. of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 216–220. Association for Computational Linguistics.

T. Nakagawa. 2007. Multilingual dependency parsing using global features. In *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 952–956.

J. Nivre, J. Hall, and J. Nilsson. 2006a. Maltparser: A data-driven parser-generator for dependency parsing. In *Proc. of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, volume 6, pages 2216–2219.

J. Nivre, J. Hall, J. Nilsson, G. Eryiğit, and S. Marinov. 2006b. Labeled pseudo-projective dependency parsing with support vector machines. In *Proc. of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 221–225. Association for Computational Linguistics.

J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. 2007. CoNLL 2007 shared task on dependency parsing. In *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*. Association for Computational Linguistics.

J. Nivre, I.M. Boguslavsky, and L.L. Iomdin. 2008. Parsing the SynTagRus treebank of Russian. In *Proc. of the 22nd International Conference on Computational Linguistics (COLING 2008)*, volume 1, pages 641–648. Association for Computational Linguistics.

J. Nivre. 2009. Parsing Indian languages with MaltParser. In *Proc. of the Seventh International Conference on Natural Language Processing (ICON 2009) NLP Tools Contest*, pages 12–18.

K. Oflazer, B. Say, D.Z. Hakkani-Tür, and G. Tür. 2003. Building a Turkish treebank. *Text, Speech, and Language Technology*, pages 261–277.

L. Øvrelid and J. Nivre. 2007. When word order and part-of-speech tags are not enough–Swedish dependency parsing with rich linguistic features. In *Proc. of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 447–451.

S. Petrov, D. Das, and R. McDonald. 2011. A universal part-of-speech tagset. *Arxiv preprint ArXiv:1104.2086*.

Loganathan Ramasamy and Zdeněk Žabokrtský. 2011. Tamil dependency parsing: Results using rule based and corpus based approaches. In *Proc. of the 12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2011)*, volume 1, pages 82–95, Berlin, Heidelberg. Springer-Verlag.

M. Schiehlen and K. Spranger. 2007. Global learning of labelled dependency trees. In *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 1156–1160.

Anna Siewierska. 1998. Variation in major constituent order: A global and a European perspective. In Anna Siewierska, editor, *Constituent Order in the Languages of Europe*, pages 475–551. Mouton De Gruyter.

K. Simov, P. Osenova, A. Simov, and M. Kouylekov. 2004. Design and implementation of the Bulgarian HPSG-based treebank. *Research on Language & Computation*, 2(4):495–522.

I. Titov and J. Henderson. 2007. Fast and robust multilingual dependency parsing with a generative latent variable model. In *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 947–951.

L. Van der Beek, G. Bouma, R. Malouf, and G. Van Noord. 2002. The Alpino dependency treebank. *Language and Computers*, 45(1):8–22.

V. Vincze, D. Szauter, A. Almási, G. Móra, Z. Alexin, and J. Csirik. 2010. Hungarian dependency treebank. In *Proc. of the Seventh Conference on Language Resources and Evaluation (LREC 2010)*.

N. Xue, F. Xia, F.D. Chiou, and M. Palmer. 2005. The Penn Chinese Treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207–238.