Brown clustering for unlexicalized parsing

Daniel Dakota
Indiana University
Ballantine Hall 844
Bloomington, IN 47405-7005
ddakota@indiana.edu

Abstract

Brown clustering has been used to help increase parsing performance for morphologically rich languages. However, much of the work has focused on using clustering techniques to replace terminal nodes or as a feature for parsing. Instead, we choose to examine how effectively Brown clustering is for unlexicalized parsing by creating data-driven POS tagsets which are then used with the Berkeley parser. We investigate cluster sizes as well as on what information (e.g. words vs. lemmas) clustering will yield the best parser performance. Our results approach the current state of the art results for the German TüBa-D/Z treebank when using parser internal tagging.

1 Introduction

Part of Speech (POS) tags are an essential aspect of any annotated corpus, in particular for treebanks. However, the development of optimal tagsets for a given language is still problematic. The granularity of the linguistic information has both practical and theoretical aspects, but the chosen tagset has direct consequences on performance of a given task, especially to parsing.

The argument can be made that regardless of the morphological complexity of a language, there still only exists a set of primary POS tags. This has resulted in the creation of simplified, coarse-grained tagsets, most notably the Universal Tagset (Petrov et al., 2012) consisting of only 12 primary POS tags. However, this oversimplifies the linguistic complexity of a language. Subsequently, too fine-grained of a tagset also results in a decrease in parser performance (Maier et al., 2014). Although statistical methods for parsing have improved over the past decade, the issue of complex morphology and its direct impact on parsing performance still remains. This is most evident in morphologically rich languages (MRLs) where a single form of a word may have dozens of surface forms. This has resulted in expanded tagsets for many languages that possess more morphology than English, as well as the addition of morphological information directly attached to the tags, which increases both the tagset size and the level of granularity.

With the creation of any tagset, how much linguistic information is relevant becomes a matter of debate. This has traditionally required a discussion about how best to incorporate the relevant linguistic information in order to categorize and sub-categorize various POS into a tagset. We choose to approach this problem by examining whether we can empirically and automatically create POS tags utilizing Brown clustering (Brown et al., 1992), and how effectively these tagsets can be used for parsing. By doing so, we group words together contextually and are able to add additional linguistic information into the process, which reduces the need to manually group morphologically complex words into various linguistic categories. We experiment with the granularity of these tags by clustering words, lemmas, and lemmas with morphological information and subsequently examine to what extent these tagsets still mimic linguistic categories. We utilize the unlexicalized Berkeley parser (Petrov and Klein, 2007) to examine the impacts of these tagsets on parsing performance of the German TüBa-D/Z treebank (Telljohann et al., 2015). Results fall in line with previous research on tagset granularity and show empirically created tagsets can come close to matching our established baseline using pre-defined tags as well as state of the art results when using parser predicted tags for parsing.

The remainder of the article is structured as follows. In section 2, we review previous work on clustering and POS tagset granularity. Section 3 presents the task while section

4 describes our experimental setup. Parsing results and discussion are presented in sections 5 and 6 before section 7 concludes the article.

2 Related Work

2.1 POS Tag Set Granularity

The granularity of a POS tagset is an important aspect of parsing since it directly impacts the parsing performance. English POS tags have continued to be based on the 36 tagset of the Penn Treebank (Marcus et al., 1993), but this has not confined other languages to such tag limits. For German, there is the 54 STTS tagset (Schiller et al., 1995) which can have morphological information attached to the tags increasing the maximum tagset size into the hundreds. This is a common strategy for many tagsets for MRLs which demonstrate much higher degrees of morphology. However, what morphology is optimal for improved parser performance for any given language has not been definitively determined, as the increase in the tagset subsequently increases sparsity of tags which influences the parser.

Although less granular POS tagsets can achieve a high rate of tagging accuracy, this does not necessarily mean they convey enough information for parsing. This was demonstrated by Maier et al. (2014) who utilized the Berkeley parser to tag and parse two German treebanks with three tagset variants, the UTS (Petrov et al., 2012) consisting of 12 tags, the STTS tagset (Schiller et al., 1995) consisting of 54 tags, and the STTS with morphological information resulting in hundreds of tags. Although the use of the UTS tagset resulted in the highest POS accuracy, it did not obtain the highest parsing performance which was obtained by the use of the STTS tagset.

Additionally, Marton et al. (2013) found that particular linguistic information (e.g. person, number, gender) for finer-grained tagsets can be useful when utilized as a gold POS tag for dependency parsing of Arabic, but detrimental when predicted by the parser internally, which benefits from coarser grained tagsets.

Seddah et al. (2009) investigated two tagsets with different granularity on French treebanks and concluded that the granularity of the tagsets can improve results, but with each improving either dependency or constituency parsing results respectively over the other.

2.2 Clustering

Clustering has been used in document classification, but there has also been an increase in its utilization to other areas of NLP such as to help improve POS tagging for Twitter (Owoputi et al., 2010). More recently it has been utilized in parsing to help reduce data sparsity, as statistical parsing suffers from data sparseness, particularly when parsing MRLs which have a higher ratio of word forms to lemmas (Tsarfaty et al., 2010).

Most, if not all, work has focused on replacing terminal nodes with clusters IDs or by using clusters as a feature for dependency parsing. Clustering has been shown to reduce sparsity issues, resulting in increased parser performance. Koo et al. (2008) showed that using Brown clustering to create cluster-based feature sets outperformed the baseline models in both English and Czech dependency parsing.

However, for MRLs how best to use Brown clustering to improve parsing performance is still unclear as clustering on words, lemmas, or lemmas with additional morphological information has yielded various results. Candito and Crabbé (2009) clustered what they termed *desinflected* French words. They removed unnecessary inflection markers using an external lexicon and then combined the *desinflected* form with additional features and replaced terminal nodes with the cluster ID. Although this increased French parsing performance with the Berkeley parser and improved results for both medium and higher frequency words (Candito and Seddah, 2010), the results were comparable to clustering the lemma with the predicted POS tag of the word.

Candito et al. (2010) found that replacing terminals with clustering-based features improved results for the Berkeley parser but not substantially for dependency parsers. Related work by Ghayoomi (2012) and Ghayoomi et al. (2014) used Brown clustering with POS information to resolve homograph issues in Persian and Bulgarian respectively to significantly improve class-based lexicalized parsing results over word-based parsing. Goenaga et al. (2014) created word clusters using both words (for Swedish) and lemmas with morphological information (for Basque) to create features of varying granularities for use in dependency parsing with noticeable improvements. Such findings are supported by Versley (2014) who noted that cluster-based features improved discontinuous constituent

parsing results for German considerably, but were also influenced by the granularities of the feature (i.e. a sequence of 0s and 1s to which every word is assigned indicating the cluster ID with shorten bit-strings representing more general, larger subsets of clusters).

3 Task

The question of how best to determine the granularity for POS tags for optimal parsing continues to persist for many languages, which is made more problematic by language-specific linguistic phenomena. We choose to investigate whether we can create empirically optimal tagsets using Brown clustering and obtain results similar to pre-defined tagsets. As has been shown, clustering has yielded positive results in parsing. However, much of the work has replaced terminal nodes with class-based representations. This has been demonstrated to be useful for lexicalized parsing, but for unlexicalized parsing, although improvements have been shown, the extent to which clusters can be utilized has been minimized. Terminal nodes (i.e. words) are only utilized for unlexicalized parsing when the parser needs more information than just using the tags, thus how often the terminals influence the parser is minimized when the POS tagging accuracy is high. For this reason, we choose to replace POS tags. During the clustering process, we examine the impact of word frequencies, clustering sizes, and granularity of information at the word level on parsing performance. German possesses a richer morphology than English, allowing for different linguistic phenomena that effect parser performance such as case. In particular, German morphology allows for a much freer word order than English, but not as free as other MRLs. For example, articles are inflected for case and gender allowing subjects, direct objects, and indirect objects to freely move in the sentence. One inherent complexity of German morphology is case syncretism. This is seen with articles where the case and gender for one object can mimic another (e.g. *die* is both the definite nominative feminine and definitive accusative plural). This means that grammatical functions improve the usefulness of a parse (Rafferty and Manning, 2008) but that they cannot be determined strictly by their position in the tree (Kübler, 2008).

4 Experimental Setup

4.1 Treebank

We use the German treebank TüBa-D/Z version 10.0 (Telljohann et al., 2015), taking the first 90% for training and performing a 3-fold cross-validation. Each fold consists of 57357 training sentences and 28678 for testing. The final 10% percent was left out for testing after further experiments have been run. The treebank was pre-processed by replacing all grammatical function (GF) dash separators with a "#" and collapsing all occurrences of label-internal dash separators (e.g. R-SIMPX $\rightarrow$ RSIMPX). This was done as the Berkeley parser treats anything after a "-" as a grammatical function and cuts it off.

4.2 Parser

For parsing, we use the Berkeley parser (Petrov and Klein, 2007). The parser is ideal to examine the impact of POS tags as it is unlexicalized. The Berkeley parser uses a system of split/merge cycles that should help to smooth over the variation in the tagset sizes. We evaluate using standard EVALB (Sekine and Collins, 1997) including grammatical functions, using a parameter file to delete VROOT. Non-parsed sentences are not calculated in the evaluation metrics, but we provide their number in the results.

4.3 Word Clustering

We use Brown Clustering (Brown et al., 1992) using the implementation from Liang (2005). Brown clustering is an unsupervised clustering method that obtains a pre-specified number of clusters (C). It assigns the C most frequent word tokens to their own cluster. Every subsequent word is assigned to one of the clusters by creating a new cluster and merging the C+1 cluster with an already defined cluster that minimizes the loss in likelihood of the corpus based on a bigram model determined from the clusters. Brown clustering is a hard clustering algorithm, thus the previous step is repeated for each subsequent word until every word is assigned a cluster, resulting in words having been clustered based on their contextual similarity to one another. The final product is a binary hierarchical structure with each cluster being represented by a bit-string of varying lengths. We cluster using a German wikipedia dump consisting of approximately 175 million words (Versley and Panchenko, 2012), which was also tagged with both POS infor-

mation and morphological information using Mate Tools (Björkelund et al., 2010).

By using Brown clustering, we are empirically creating tagsets that allow for words to be grouped together based on contextual similarity. This also allows for words normally assigned to the same linguistic category (e.g. nouns) to be possibly assigned to different clusters because their contextual similarity differs enough as defined by the clustering algorithm. This subsequently allows for a finer distinction of categories of words than would naturally be assumed. We replace POS tags in the treebank by looking up whether the word has a cluster ID and replacing it with the full bit-string. Any word in the treebank without a cluster ID was given a tag of '0' symbolic of an unknown tag. All punctuation was replaced with a single '-PUNCT-' in order to reduce the overall number of tags. This means for every cluster size C, the true number of tags in the set is C+2. We performed an initial experiment between words and lemmas in order to determine which of the two are a better basis for clustering tags. Since Brown clustering has different thresholds, we examined different minimum frequency of lemmas in the clustering corpus to examine a) what impact decreasing the minimum frequencies has on coverage and performance and b) whether there is a minimum frequency after which there are no longer improvements in results. Finally we performed two additional experiments by adding morphological information to the lemmas. The first experiment added both selected POS tag information and morphological information from Mate Tools (Björkelund et al., 2010) to the lemma. This was done to examine whether the use of some pre-defined STTS tag information with additional morphological information can be utilized in the clustering processing, as it adds additional German-specific linguistic information. The second experiment attached only morphological information to the lemmas. The list of selected tags are presented in Table 1. These tags were selected based on morphological information and not every possible STTS tag was selected. In particular, we focused on tags that tend to represent words that are inflected for case and gender (i.e. articles, adjectives, and personal pronouns). We also chose to simply assign all verbs a single VERB tag. This was done as verbs are particularly challenging to label for granularity in any given language. A summary of the selected tags with morphological information are presented in Table 2.

Name	Description
ART	article
ADJA	adjectives
PRELS	substituting relative pronoun
PIS	substituting indefinite pronoun
PPOSAT	attributive possessive pronoun
PPER	irreflexive personal pronoun
VERB	all verbs given simply VERB

Table 1: The selected POS tags for experiment 1

Name	Description
art+case	attach case to articles
art+gend	attach genders to articles
art+case+gend	attach both case and gender to articles
infl+case+gend	attach case and gender to all lemmas if applicable
verb+person	attach person to verbs
verb+num	attach number to verbs
verb+person+num	attach person and number to verbs
all	all features

Table 2: Description of Lemmas+Features used for clustering

Recall	Precision	F-score	POS Acc.	Unparsed Sent.
83.12	82.93	83.02	97.5	4

Table 3: Average results for 3-fold baseline with STTS tags

N1	N2	N3	F-score Average
83.53	83.25	82.29	83.02

Table 4: Individual F-scores for 3-fold baseline with STTS

4.4 Baseline

We establish a baseline by using the STTS tagset for the TüBa-D/Z treebank and report the average recall, precision, F-score for parsing, and POS accuracy which is calculated by comparing every tag in the gold and test files (Table 3). This was done as a basis of comparison for our experimental setup since there exists no previous findings which we can directly compare our results against. Table 4 provides the F-scores for each fold of the baseline. The varying results on each fold is consistent with other findings (see Levy and Manning (2003)) that have noted that any given section of a treebank may be more or less difficult to parse relative to another section. Here later portions of the treebank are inherently harder to parse.

Cluster Size	Recall	Precision	F-score	POS Acc.	Unparsed Sent.
25	78.16	78.89	78.52	95.10	8
50	79.24	79.85	79.54	95.25	6
75	79.05	79.56	79.30	95.39	2
100	79.41	79.73	79.57	95.34	4
125	79.50	79.75	**79.62**	95.52	3
150	79.35	79.67	79.51	95.59	4
175	79.38	79.64	79.51	95.67	17
200	79.29	79.47	79.38	**95.74**	16

Table 5: Words used as tag with a min. frequency of 100

Cluster Size	Recall	Precision	F-score	POS Acc.	Unparsed Sent.
25	79.17	79.67	79.42	93.32	3
50	79.75	80.15	**79.95**	93.26	4
75	79.57	79.95	79.76	**93.41**	1
100	79.60	79.90	79.74	93.17	3
125	79.77	80.03	79.90	93.03	10
150	79.49	79.56	79.53	93.10	5
175	79.84	79.87	79.85	93.18	6
200	79.34	79.32	79.33	93.16	5

Table 6: Lemmas used as tags with a min. frequency of 100

Min Occurrence	% of Words	% of Word Types
100	88.5	30.2
50	90.9	39.3
20	93.3	51.2
3	96.4	70.5
1	97.4	78.5

Table 7: The percentage of words and word types found in TüBa-D/Z from clustering corpus

Min Occurrence	% of Lemmas	% of Lemma Types
100	89.9	31.3
50	91.8	40.3
20	93.6	51.3
3	96.1	69.1
1	96.9	76.8

Table 8: The percentage of lemmas and lemma types found in TüBa-D/Z from clustering corpus

5 Results

5.1 Word vs. Lemma

When comparing POS tags created strictly on the words (Table 5) versus tags created on lemmas (Table 6) in all cases, except for a cluster size of 200, lemmas outperform words. However, for tags created on words, the highest F-score is obtained using a cluster size of 125, whereas for lemmas, the highest F-score is obtained with a cluster size of 50. Interestingly, the POS accuracy for word clusters increases with the cluster size which stands in contrast to the POS accuracy for lemmas, which tends to decrease in accuracy as the cluster size increases. A cluster size of 200 trained on just words obtained the highest POS accuracy of any of our experiments at 95.74%. However, this is consistent with the findings from Maier et al. (2014) that a higher POS accuracy does not necessarily result in the best parsing performance. This is further supported by the lower POS accuracies of the equivalent lemma POS cluster sizes which although lower, demonstrate a consistently higher F-score. None of the results reach our baseline; the closest, a cluster size of 50 using lemmas, is still more than 2.5% absolute below the baseline.

In order to investigate the coverage of clusters on words and lemmas in the treebank, we extracted the percentage of words and lemma tokens covered by the clusters, as well as extracting type coverage, the results of which are presented in Tables 7 and 8. We do not include punctuation, since stand-alone punctuation is not utilized during Brown clustering. Using a minimum frequency of 100 in the Wikipedia data, the resulting clusters cover 88.5% of total word tokens in the treebank, but represents merely 30.2% of all word types. Decreasing the minimum frequency to 1 increases coverage of the overall corpus to about 97% for both the raw words and lemmas but still about 25% of types are not covered. Using lemmas instead of word forms does not alter coverage percentages substantially. This is surprising given that reducing words to their lemmas should help decrease sparsity but the overall coverage between unlemmatized forms and their lemmas is comparable. However, by lemmatizing we increase the frequency of a given token type in the data which should help the parser, as given cluster tags will occur more frequently.

5.2 Lemmas

Noting the slightly better performance of lemmas over words, experiments were conducted clustering on lemmas but reducing the minimum frequency of a lemma for clustering to 50 times and 3 times, as presented in Tables 9 and 10 respectively. We choose not to utilize a minimum frequency of 1 to help reduce the number of possible typos or erroneous words for clustering given the nature of web data. We can see a general rise in F-scores as the minimum frequency of a lemma's occurrence for clustering decreases. However, it is not absolute, as there are several instances in which a higher minimum frequency outperforms a lower minimum frequency. This can be seen in Table 10 where the F-score for minimum frequency lemma of 3 with a cluster size of 50 is lower than the F-score in Table 6 for minimum lemma frequency of 100 for a cluster size of 100, which was the highest performing

Cluster Size	Recall	Precision	F-score	POS Acc.	Unparsed Sent.
25	79.48	79.95	79.72	**93.15**	0
50	80.03	80.39	80.21	92.79	1
75	80.13	80.46	**80.30**	92.96	1
100	79.69	79.98	79.83	92.67	3
125	79.71	79.94	79.82	92.47	0
150	79.79	79.79	79.79	92.53	5
175	79.36	79.35	79.36	92.54	11
200	79.81	79.75	79.78	92.50	6

Table 9: Lemmas used as tags with a min. frequency of 50

Cluster Size	Recall	Precision	F-score	POS Acc.	Unparsed Sent.
25	79.85	80.29	80.07	**93.63**	5
50	79.25	79.61	79.43	92.86	570
75	80.20	80.49	80.34	92.93	5
100	80.38	80.58	**80.48**	92.51	2
125	79.82	80.02	79.92	91.96	7
150	79.87	79.84	79.86	91.51	5
175	79.96	79.94	79.94	91.37	7
200	79.74	79.65	79.70	91.38	4

Table 10: Lemmas used as tags with a min. frequency of 3

cluster size for a minimum lemma frequency of 100. The overall trend of increased performance is supported with the percentages presented in Table 8 that showed slight increases in token coverage, but larger increases in type coverage as the minimum frequency decreases for lemmas to be clustered. On average only a few sentences are not parsed, but an anomaly occurs in Table 10 where a cluster size of 50 resulted in 570 sentences not being parsed. A reason for this has not been identified.

5.3 Lemma + Morphology

To examine the effect of adding morphological information to lemmas, we select the highest obtained F-score of 80.48% , which was with lemmas with a minimum frequency of 3 and a cluster size of 100. The results for adding selected POS tags plus morphology are presented in Table 11. Adding lemma and morphological information alters results, in some cases significantly. By simply adding the STTS article tag and case information, there is a decrease of almost 4% absolute. However, when adding person information to the verb, there is an increase in performance in both experiments. Interestingly, when using the VERB tag, the F-score is further increased when combining person and number, even though VERB tag and number information alone decreases performance from just the lemma. In contrast, when not using a VERB tag, adding number information decreases performance. Combining all the features reduces overall performance in both experiments.

Cluster Size	Recall	Precision	F-score	POS Acc.	Unparsed Sent.
art+case	80.07	73.53	76.67	92.14	3
art+gend	79.93	80.14	80.03	**92.42**	4
art+case+gend	80.1	80.21	80.15	92.16	3
infl+case+gend	79.37	79.55	79.46	92.13	6
verb+person	80.64	80.60	80.62	**92.42**	9
verb+num	80.04	80.08	80.06	92.37	4
verb+person+num	80.80	80.76	**80.78**	92.20	16
all	80.08	80.01	80.04	89.83	11
best word performance	79.50	79.75	79.62	95.52	3
best lemma performance	80.38	80.58	80.48	92.51	2

Table 11: Results for lemmas and selected POS tags with morphology

Cluster Size	Recall	Precision	F-score	POS Acc.	Unparsed Sent.
art+case	79.86	73.49	76.54	92.16	7
art+gend	79.93	80.14	80.03	92.42	4
art+case+gend	80.1	80.21	80.15	92.16	3
infl+case+gend	79.69	79.94	79.81	**92.87**	6
verb+person	81.11	81.04	**81.08**	92.42	5
verb+num	79.98	80.56	80.27	92.37	2
verb+person+num	80.80	80.76	80.78	92.20	16
all	79.06	79.06	79.06	90.81	8
best word performance	79.50	79.75	79.62	95.52	3
best lemma performance	80.38	80.58	80.48	92.51	2

Table 12: Results for only lemmas and morphology

6 Discussion

Currently state of the art results for German constituency parsing for the Berkeley parser on TüBa-D/Z is an F-score of 83.97 (Petrov and Klein, 2008), however this was done using Gold POS tags. We compare results to our own baseline using the STTS tagset as well as noting consistencies found by Maier et al. (2014).

Examining Tables 11 and 12 we see that in some cases we are able to increase the F-score by adding morphological information into the clustering process, but in other cases there is a decrease in performance. Simply adding case information to articles significantly decreases performance for both experiment. This can partially be attributed to the case syncreticsm seen in German. Our decision to treat all verbs with a single coarse-grained POS tag while selecting finer grained STTS tags for tags containing case and information most likely influenced the results between the two linguistic categories. This suggests that coarse tags may be slightly more beneficial when combined with morphological information. Overall, our results are consistent with issues regarding tag granularity and parsing performance.

Interestingly, there are three identical sets of results in the experiments. This could indicate that these particular morphological features are more important than the granularity of the tag itself (i.e. detailed information of the verb is not as important as the person and number information of the verb).

We are not able to match our baseline F-score of 83.02 using the original STTS tagset. However,

Tag	Recall	Precision	F-score	% in Gold	Majority Tags
0	75.58	86.39	80.62	11.09	unknown
00010010	99.43	99.89	99.66	6.11	nouns/adjectives
0010	92.55	88.36	90.41	5.00	proper nouns
0001001110	98.20	98.59	98.39	4.88	3rd person verbs
11000	90.82	78.76	84.36	4.08	nouns
000100110	91.45	85.66	88.46	3.29	3rd person verbs
01011	88.03	85.14	86.56	2.91	nouns
110111	89.33	81.04	84.98	2.89	nouns
11010	88.69	86.06	87.35	2.58	nouns
00011110	99.69	99.79	**99.74**	2.49	mixed

Table 13: POS Tag Analysis of fold 3 for lemmas and selected POS tags with morphology

Tag	Recall	Precision	F-score	% in Gold	Majority Tags
0	87.08	92.97	89.93	19.32	unknown
10010100	99.61	99.95	99.78	6.10	mixed
1010	92.59	88.39	90.44	5.05	proper nouns
1001010110	99.92	99.96	**99.94**	4.14	3rd person verbs
0111	90.62	79.09	84.47	3.96	nouns
0101	88.46	84.39	86.38	2.89	nouns
0010	90.23	79.39	84.47	2.89	nouns
000	88.11	85.40	86.73	2.56	nouns
10011110	99.60	99.71	99.65	2.49	mixed
10000010	89.36	88.10	88.73	1.85	adjectives

Table 14: POS Tag Analysis of fold 3 for only lemmas with morphology

our results do show that is possible to create empirically driven POS tags that are created using Brown clustering that can approach results using a predefined tagset as our best results perform only less than 2% absolute lower than our baseline. Furthermore we can individually demonstrate the effects of a single piece of morphological information has on parsing performance. This provides further evidence that there is a balance between granularity and optimal performance. Given the selective nature of what morphology we chose to add to the lemmas, it is possible that a different combination of morphological information may further improve results. Additionally, our results further reinforce that a high POS accuracy does not necessarily correlate to a higher parsing performance. In both experiments, the experiments achieving the highest POS accuracy did not obtain the highest parsing results.

In an attempt to ascertain what sort of clusters are more accurate in terms of tagging than others, an analysis was performed on individual tags. However, given the nature of clustering, it is difficult to provide too much detailed information on the clusters themselves, but rather one can extrapolate general patterns within the clusters by examining them manually.

In Tables 13 and 14 we present the top 10 most frequent POS tags from the 3rd fold from the "all" experiments of the results in Tables 11 and 12 by using the EVALB implementation in Disco-dop (van Cranenburgh et al., 2016) which provides more detailed POS tag information. We also provide what we manually identified as the majority tag (i.e. a manually assigned POS tag based on the majority of words in the cluster).

The 'unknown' tag of '0' indicating that the word did not have a cluster constitutes more than 11% of the overall tags in the fold in Table 13. As seen in Table 8, this is a higher than expected percentage given that only about 4% of the lemma tags in the entire treebank are not found in the clusters. However, this can be attributed to the addition of morphological information to the lemmas. Certain tags are tagged with a very high degree of accuracy at over 99%, while other tags are more difficult for the parser. We can assume however, that if 10% of the tags in the entire treebank are only accurately tagged 80% of the time (e.g. the '0' tag), this will introduce problems for the parser leading to a decrease in parser performance. Worth noting is that although the '0' tag constitutes almost 20% of the treebank when not using POS tag information, the F-score is 9% absolute higher. This may suggest that it is easier for the parser to correctly tag unknown words using morphological information over POS information. To help further reduce the number of unknown tags in future experiments, it may be beneficial to add the treebank corpus into the clustering corpus, as well as additional domain specific texts to help increase domain specific type coverage. By simply adding the lemmatized TüBa-D/Z corpus into the clustering data alone, and using a minimum frequency of 3, we can increase the lemma token coverage of the clusters on TüBa-D/Z corpus to 97.4% and the type coverage to 75.1%. This should also help increase parser performance, as out of domain parsing impacts parsing results (Gildea, 2001).

In order to further examine the size and frequency counts of individual clusters, Tables 15 and 16 contain the number of types in each cluster, and the percentage of types with less than or equal to 10 total counts in the clustering corpus.

At first glance, it appears that if the frequency of rare words are relatively high in the cluster, then the accuracy of the tags is higher. Although the two clusters with high rates of less frequent words obtain higher POS tagging rates, this does not mean there is direct association, although it most likely attributes to the higher accuracy. A counter example can be seen however with tag

Tag	Types	POS F-Score	% ≤ 10 Freq
00010010	16842	99.66	81%
0010	145065	90.41	58%
0001001110	4143	98.39	64%
11000	98360	84.36	62%
000100110	4143	88.46	64%
01011	84061	86.56	64%
110111	75296	84.98	63%
11010	90846	87.35	64%
00011110	7888	99.74	87%

Table 15: Cluster analysis of fold 3 for lemmas and selected POS tags with morphology

Tag	Types	POS F-Score	% ≤ 10 Freq
10010100	16428	99.78	75%
1010	147868	90.44	58%
1001010110	4352	99.94	66%
0111	96025	84.47	63%
0101	82531	86.38	64%
0010	76019	84.47	62%
000	89149	86.73	63%
10011110	7719	99.65	87%
10000010	23119	88.73	57%

Table 16: cluster analysis of fold 3 for lemmas with morphology

0001001110 in Table 15. Interestingly, this cluster consists predominantly of 3rd person plural verbs (e.g. gehen.VERB.3p "go") of high frequencies. This is also seen with cluster 10000010 in Table 16. This cluster has a relatively low percentage of rare words compared to the other clusters, but still has a relatively high F-score for tag accuracy. Manually inspecting the cluster reveals that it predominantly consists of adjectives without morphological information.

When further manually examining other tags that demonstrate lower F-scores, it appears that tags that represent clusters consisting of words with a high frequency of common words that have not been tagged with additional morphological information (particularly nouns) are tagged with lower accuracy. When examining the cluster for the least accurate tag 11000 in Table 15, it consists predominantly of common nouns (e.g. Raum "room"). This low accuracy may be due to the decision not to add additional morphological information to nouns (e.g. singular vs. plural) which, if provided, may have increased tagging performance for these clusters. It also confirms that the most frequent words in the clusters have the largest influence on the tagging accuracy regardless of size and proportion of rare words.

7 Conlusion and Future Work

We have shown that we can use Brown clustering to empirically create POS tags for parsing that yield results only slightly below than that of our baseline using the pre-defined STTS tagset, as well as similar results for Berkeley internal tagging and parsing on the German TüBa-D/Z treebank.

We can increase performance by simply clustering on lemmas instead of words to create tags, which can be further increased by adding additional morphological information. However, simply adding even a single piece of morphological information can either reduce or improve results, in some cases drastically. This aligns with previous research indicating that granularity of tags effects parsing performance (Maier et al., 2014; Marton et al., 2013), but further experimentation is still needed in order to better determine how best to incorporate additional morphological information into the POS tagset for clustering to improve parsing performance, and what introduces additional parser errors. However we have demonstrated a possible mechanism for creating empirically driven tagsets possessing different granularities using readily available tools. This allows both the incorporation of linguistic information into the tagsets, but bypasses the need to manually assign words to various finer grained tags and testing how different tagset sizes and granularities affect parsing. In order to improve performance using clustering, we must better understand how language specific clustering techniques need to be utilized. This is compounded by the fact that languages possess starkly different linguistic principles, so optimal settings for German may not work for other MRLs. Similar techniques need to be performed on a set of starkly different languages in order to see if a general pattern emerges, or whether for clustering to be effective, very specific language parameters must be fine-tuned.

Acknowledgments

We would like the thank Wolfgang Seeker and Bernd Bohnet for tagging the clustering corpus with morphological information as well as Djamé Seddah and Yannick Versley for providing the data for clustering and additional pertinent information.

References

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 33–36.

Peter Brown, Vincent Della, Peter Desouza, Jennifer Lai, and Robert Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 19(4):467–479.

Marie Candito and Benoît Crabbé. 2009. Improving generative statistical parsing with semi-supervised word clustering. In *Proceedings of the 11th International Conference on Parsing Technologies*, IWPT '09, pages 138–141, Paris, France.

Marie Candito and Djamé Seddah. 2010. Parsing word clusters. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, SPMRL '10, pages 76–84, Los Angeles, California.

Marie Candito, Joakim Nivre, Pascal Denis, and Enrique Henestroza Anguiano. 2010. Benchmarking of statistical dependency parsers for French. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 108–116, Beijing, China.

Masood Ghayoomi, Kiril Simov, and Petya Osenova. 2014. Constituency parsing of bulgarian: Word- vs class-based parsing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4056–4060, Reykjavik, Iceland.

Masood Ghayoomi. 2012. Word clustering for Persian statistical parsing. In Histohsi Isahara and Kyoko Kanzaki, editors, *Advances in Natural Language Processing*, volume 7614 of Lecture Notes in Computer Science: JapTal 12: Proccedings of the 8th International Conference on Advances in Natural Language Processing, pages 126–137, Kanazawa, Japan.

Daniel Gildea. 2001. Corpus variation and parser performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 167–202, Pittsburgh, PA.

Iakes Goenaga, Koldo Gojenola, and Nerea Ezeiza. 2014. Combining clustering approaches for semi-supervised parsing: the BASQUE TEAM system in the SPRML2014 shared task. In *In First Joined Workshop of Statistical Parsing of Morphologically Rich Language and Syntactic Analysis of Non-Canonical Languages*, Dublin, Ireland.

Terry Koo, Xavier Carreras, and Michael Collins. 2008. Simple semi-supervised dependency parsing. In *Proceedings of ACL-08: HLT*, pages 595–603, Columbus, Ohio.

Sandra Kübler. 2008. The PaGe 2008 shared task on parsing German. In *Proceedings of the Workshop on Parsing German*, PaGe '08, pages 55–63, Columbus, OH USA.

Roger Levy and Christopher Manning. 2003. Is it harder to parse Chinese, or the Chinese Treebank? In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 439–446, Sapporo, Japan.

Percy Liang. 2005. Supervised learning for natural language. Master's thesis, Massachusetts Institute of Technology.

Wolfgang Maier, Sandra Kübler, Daniel Dakota, and Daniel Whyatt. 2014. Parsing German: How much morphology do we need? In *Proceedings of the First Joined Workshop of Statistical Parsing of Morphologically Rich Language and Syntactic Analysis of Non-Canonical Languages (SPMRL-SANCL 2014)*, pages 1–14, Dublin, Ireland, August.

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

Yuval Marton, Nizar Habash, and Owen Rambow. 2013. Dependency parsing of Modern Standard Arabic with lexical and inflectional features. *Computational Linguistics*, 39(1):161–194, March.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, and Nathan Schnieder. 2010. Part-of-speech tagging for twitter: Word clusters and other advances. Technical report, Carnegie Mellon University.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 404–411, Rochester, NY.

Slav Petrov and Dan Klein. 2008. Parsing German with latent variable grammars. In *Proceedings of the Workshop on Parsing German at ACL '08*, pages 33–39, Columbus, Ohio.

Slav Petrov, Das Dipanjan, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey, May.

Anna N. Rafferty and Christopher D. Manning. 2008. Parsing three German treebanks: Lexicalized and unlexicalized baselines. In *Proceedings of the Workshop on Parsing German*, pages 40–46, Columbus, Ohio.

Anne Schiller, Simone Teufel, and Christine Thielen. 1995. Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, Universität Stuttgart and Universität Tübingen.

Djamé Seddah, Marie Candito, and Benoît Crabbé. 2009. Cross parser evaluation: A French treebanks study. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT)*, pages 150–161, Paris, France.

Satoshi Sekine and Michael Collins. 1997. Evalb bracket scoring program. http://nlp.cs.nyu.edu/evalb/.

Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. 2015. Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Seminar für Sprachwissenschaft, Universität Tübingen, Germany.

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.

Andreas van Cranenburgh, Remko Scha, and Rens Bod. 2016. Data-oriented parsing with discontinuous constituents and function tags. *Journal of Language Modelling*, 4(1):57–111.

Yannick Versley and Yana Panchenko. 2012. Not just bigger: Towards better- quality web corpora. In *Seventh Web as Corpus Workshop (WAC7)*, pages 44–52, Lyon, France.

Yannick Versley. 2014. Incorporating semi-supervised features into discontinuous easy-first constituent parsing. In *In First Joined Workshop of Statistical Parsing of Morphologically Rich Language and Syntactic Analysis of Non-Canonical Languages*, Dublin, Ireland.