Quantifying the morphosyntactic content of Brown Clusters

Manuel R. Ciosici
UNSILO A/S and
Aarhus University
Aarhus, Denmark
manuel@cs.au.dk

Leon Derczynski
IT University of
Copenhagen
Copenhagen, Denmark
leod@itu.dk

Ira Assent
Department of Computer Science
Aarhus University
Aarhus, Denmark
ira@cs.au.dk

Abstract

Brown and Exchange word clusters have long been successfully used as word representations in Natural Language Processing (NLP) systems. Their success has been attributed to their seeming ability to represent both semantic and syntactic information. Using corpora representing several language families, we test the hypothesis that Brown and Exchange word clusters are highly effective at encoding morphosyntactic information. Our experiments show that word clusters are highly capable of distinguishing Parts of Speech. We show that increases in Average Mutual Information, the clustering algorithms' optimization goal, are highly correlated with improvements in encoding of morphosyntactic information. Our results provide empirical evidence that downstream NLP systems addressing tasks dependent on morphosyntactic information can benefit from word cluster features.

1 Introduction

Distributionally generated word classes (often referred to as *word clusters*) are hard clusters, containing all word types observed in a corpus, allocated to clusters based on contextual information observed in the corpus. They have found wide use in Natural Language Processing (NLP) systems as an alternative to word embeddings such as word2vec (Mikolov et al., 2013). Word clusters differentiate themselves from word embeddings by requiring estimation of many fewer parameters, and by their ability to derive qualitative representations from smaller corpora (Qu et al., 2015; Bansal et al., 2014).

Brown Clusters (Brown et al., 1992) are a well-known approach based on hard, hierarchical, distributionally derived groups of word types observed in a corpus of unstructured text, with Average Mutual Information (AMI) as the optimization goal. Exchange Clusters are an alternative approach obtained by applying the Exchange Algorithm (Kneser and Ney, 1993) to the same optimization goal. Unlike Brown, Exchange outputs a flat clustering, with no hierarchy (Martin et al., 1998). When only the bottom of the hierarchy is used, like in this paper, Exchange and Brown clusters are interchangeable.

Both Brown and Exchange clusters have been used as word representations for various Natural Language Processing tasks such as Part of Speech tagging in clean and noisy text (Swain and Cole, 2016; Owoputi et al., 2013; Derczynski et al., 2015), dependency parsing (Koo et al., 2008; Bansal et al., 2014), Chinese Word Segmentation (Liang, 2005), and Named Entity Recognition (Swain and Cole, 2016; Derczynski et al., 2015; Liang, 2005). Word clusters distinguish themselves from word embedding models by their ability to learn from little data (Bansal et al., 2014; Qu et al., 2015); for example, in cases like (Bansal et al., 2014), word clusters outperform other kinds of representations, including word embeddings. In the literature, it is often observed that word clusters seem to encode a considerable amount of morphosyntactic and semantic knowledge (Brown et al., 1992; Derczynski et al., 2015). However, it has not yet been studied to which extent such knowledge is encoded, as previous work on Brown and Exchange clusters focuses mostly on algorithmic improvements and on applications to different NLP tasks.

In this work, we present a principled study of the morphosyntactic information encoded in flat word clusters induced exclusively from class-based language models via Brown Clustering and Exchange algorithm. In particular, we focus on how well these approaches derive clusters that represent Parts of Speech as a measure of the morphosyntactic information encoded.

We find that Brown and Exchange clusters are highly effective at representing morphosyntactic information, even when hyper-parameters are set such that they match only the number of Parts of Speech, thereby grouping into relatively few word clusters only. Our results provide empirical evidence for the observed performance gains when including Brown and Exchange word clusters as features in NLP systems that rely on morphosyntactic information.

Furthermore, we find that there is a strong correlation between the optimization goal of Brown clustering and the Exchange Algorithm (i.e., Average Mutual Information), and performance at Parts of Speech separation, which again confirms the appropriateness of choosing AMI in word clustering for morphosyntactic information.

2 Background

Class-based language models address the problem of brittleness in classic n-gram language models by trading precision for performance stability over different text styles (Brown et al., 1992).

Brown Clustering (Brown et al., 1992) and Exchange (Kneser and Ney, 1993) are greedy algorithms that construct word classes by optimizing for higher Average Mutual Information (AMI). Maximizing Average Mutual Information is a proxy for maximizing the log-likelihood of the underlying class-based language model on the given corpus (Martin et al., 1998). Despite their age, most research on Brown or Exchange clusters has so far followed two major directions: algorithm improvements and applications in Natural Language Processing. In contrast little focus has been placed on understanding and evaluating the information content of the clusters.

In the direction of algorithm improvements, work has been done on the effect of greedy merge choices in Brown Clustering (Derczynski and Chester, 2016; Ciosici, 2015) and extension of AMI to n-grams (Martin et al., 1998). Model relaxations, particularly to Exchange, aim to improve computational performance by reducing the effect of words swapping clusters (Dehdari et al., 2016; Uszkoreit and Brants, 2008).

As mentioned earlier, both Brown and Exchange clusters have seen many applications in Natural Language Processing (NLP) systems: PoS tagging (Swain and Cole, 2016; Owoputi et al., 2013; Derczynski et al., 2015), dependency parsing (Koo et al., 2008; Bansal et al., 2014), Chinese Word Segmentation (Liang, 2005), and Named Entity Recognition (Swain and Cole, 2016; Derczynski et al., 2015; Liang, 2005). Most of this work, like (Swain and Cole, 2016) uses the word clusters as sources of features which are combined with hand-designed ones. While word clusters derived using Exchange and Brown clustering have found wide use in NLP systems, their use has been based on the assumption that they encode morphosyntactic and semantic information rather that a principled use.

In relation to Parts of Speech, early on Martin et al. (1998) concluded that initializing Exchange with PoS-homogeneous clusters has no effect on final clustering AMI, but that it does help accelerate convergence. More recently, Christodoulopoulos et al. (2010) found that Brown clusters match the performance of more sophisticated clustering methods, despite their simple algorithmic construction. The study focused on using word clustering algorithms as sources of prototypal information to prototype-driven learning models for classification. In this paper, we study the amount of morphosyntactic information encoded in Brown and Exchange word clusters with the goal of providing empirical results for a principled use of such clusters in downstream tasks.

3 Metric selection

In order to determine the amount of morphosyntactic information encoded in Brown and Exchange word clusters, we measure their ability to separate word types by their Parts of Speech. For this, we require cluster quality measures. Brown and Exchange clusters do not exist in a metric space; therefore unsupervised cluster quality measures relying on distances between points or clusters, such as the Silhouette coefficient (Rousseeuw, 1987), cannot be used. Instead, we focus on two quality measures that compare clusters with a ground truth partitioning. We work under the hypothesis that Brown and Exchange clusters represent parts of speech and thus, we consider parts of speech as the ground-truth partitioning of the data. This makes it possible to use cluster quality measures that require as input an existing ground-truth partitioning. We use PoS tags resulting from manual or automatic annotation. We evaluate using a widespread and easy to interpret measure based on overlap (*purity*), and an infor-

mation theoretical measure (*Adjusted Mutual Information*).

3.1 Clustering Purity

Cluster purity measures how many points in a clustering (in our case words) have been assigned to a cluster whose predominant label they share (e.g. adjectives clustered with other adjectives, nouns with other nouns etc). Intuitively, it measures the percentage of points properly classified (via their cluster membership). Formally, cluster purity is defined as:

$$purity(C_i) = \frac{1}{|C_i|} \max_{l=1}^{|L|} |label(C_i, l)| \quad (1)$$

$$purity(C) = \sum_{i=1}^{k} \frac{|C_i|}{|V|} purity(C_i) \quad (2)$$

$$= \frac{1}{|V|} \sum_{i=1}^{k} \max_{l=1}^{|L|} |label(C_i, l)| \quad (3)$$

Where the function $label(C_i, l)$ provides the number of elements from C_i with label l and L is the set of labels. *Purity* reaches a value of 1 when the clustering is identical to the ground-truth partitioning, or each point is allocated to its own cluster (i.e. $k = |V|$). When $k = 1$, *purity* is equal to the fraction of points labeled with the most popular label, and thus provides a baseline. In our case, that is equal to the percentage of vocabulary allocated to the most popular PoS class, usually nouns. For values of $k \in (1, |V|)$, it varies depending on cluster quality. If $k > |L|$, *purity* can take a value of 1 if each cluster is a subset of a ground partition. Thus, *purity* is expected to increase as k grows higher than $|L|$.

In our experiments, *purity* measures the percentage of vocabulary that is labeled correctly. In other words, *purity* does not depend on word frequency. Thus, it is not an approximation of PoS tagging accuracy, like the M-1 measure used by Bansal et al. (2014). Since we focus on morphosyntactic information encoded in word clusters, we do not want a measure that takes into account word frequency in the given corpus (i.e, one that is a good approximation of PoS tagging performance), but one that focuses exclusively on the clusters and their content.

3.2 Adjusted Mutual Information

Since the vocabulary size is fixed, as the number of clusters k increases, *purity* can increase even if cluster membership is randomly assigned, as it is easier for smaller clusters to randomly achieve label agreement. *Adjusted Mutual Information (AdjMI)* (Vinh et al., 2009), not to be confused with *AMI* (Brown and Exchange's optimization goal), measures the amount of information shared by the ground truth partitioning U and a clustering C. In our evaluation that corresponds to the amount of information shared by the PoS ground-truth partitioning and the clustering resulting from Brown or Exchange. AdjMI corrects for the mutual information expected to exist between the ground truth partitioning U and a random clustering. Formally, it is defined as:

$$AdjMI(U, C) = \frac{MI(U, C) - E\{MI(U, C)\}}{avg\{H(U), H(C)\} - E\{MI(U, C)\}} \quad (4)$$

Where MI and H stand for Mutual Information and Entropy, respectively. Intuitively, *AdjMI* measures how much information we gain about a point's membership in to a cluster in the ground-truth partitioning U, when we know its membership to a cluster in an induced clustering C, and the other way around. As k increases over the number of ground partitions L, *AdjMI* has the opposite effect to *purity*, i.e., it scores lower due to the higher effect attributed to randomness. Just like *purity*, *AdjMI* takes values in the interval $[0, 1]$. An *AdjMI* value of 1 corresponds to a clustering identical to the ground-truth partitioning, while a value of 0 corresponds to a clustering that is not better than a random allocation of points to clusters. Unlike *purity*, values of *AdjMI* cannot be interpreted to say anything about the percentage of points that have been properly allocated. In other words, a value of, say 0.3, does not indicate that 30% of the points have been properly separated.

Data Set	Vocabulary	Length
EN UD	33 904	488 579
EN EuroParl	159 716	73 577 783
FR UD	51 670	575 887
FR EuroParl	204 166	65 878 206
CZ UD	164 483	2 226 848
CZ EuroParl	177 631	15 193 309

Table 1: Overview of data sets

4 Experiments

4.1 Data and preprocessing

We use manually annotated data from Universal Dependencies (UD) (Leung et al., 2017) for English, French and Czech[1]. We chose the group of languages so that it represents different language families. Our choice of languages is based on the amount of manually labeled data, and the presence of each language in the larger, not annotated, EuroParl corpus. We append the manual or automatic PoS tags and convert text to lowercase. Therefore, a sentence such as "Words have meaning." is transformed into "words_NOUN have_VERB meaning_NOUN ._PUNCT". Both word clustering algorithms studied in this paper are insensitive to the appended PoS tags as they operate at word and not character level. The appended tags allow us to evaluate the quality of word clusters using the measures described in the previous section. We replace all numbers, dates, times, URLs and emails with placeholders in order to reduce vocabulary size. Universal Dependencies is the largest manually annotated corpus we have access to. For experiments on larger corpora, we use the unlabeled EuroParl corpus (Koehn, 2005). More specifically, the English-French and English-Czech pairs. Since manually annotated PoS tags are not available for Europarl, we append automatically assigned PoS tags, obtained by using UDPipe (Straka and Straková, 2017) pretrained on manually annotated corpora from Universal Dependencies.

We use flat clusters from the Exchange clustering algorithm for all experiments reported in this section as they outperform the flat clustering resulting from Brown in terms of *Average Mutual Information* (their optimization goal), *Adjusted Mutual Information (AdjMI)* and cluster *purity*. All observations in the following section also apply to the flat clusters resulting from Brown. For interested readers, we include all experiments with Brown Clustering as supplementary material. The fact that Exchange outperforms Brown Clustering in terms of AMI is well-understood (Brown et al., 1992), but its effect on cluster content is not.

[1]We use data from release 2.2 of Universal Dependencies.

4.2 Morphosyntactic content in Exchange and Brown clusters

Using Exchange, we induce flat clusterings with k in the range 18 to 800. We start with $k = 18$ as it matches the observed number of distinct PoS tags in the Universal Dependencies corpora (17 distinct tags and one catch-all tag). When setting the hyper-parameter k to be higher than 18, if Exchange separates clusters by Parts of Speech, then the expectation is that clusters are subsets of words sharing the same PoS tags, and that *purity* for such clusterings will be high. In Figure 1a, we show *purity* measured on the aforementioned clusters. We can see that, even when the number of clusters is equal to that of PoS tags ($k = 18$), between 55% and 62% of the vocabulary is properly separated. *Purity* increases as k increases towards 100. At $k = 500$ and $k = 800$, between 64% and 70% of the vocabulary is grouped based on PoS, not that much more than at $k = 100$.

Increasing the number of clusters k to high values is not guaranteed to improve *purity*, for any of the languages studied. This is contrary to the expectation that *purity* increases when $k > 18$. This indicates that Exchange and Brown do not exclusively optimize for Part of Speech separation. We believe the clustering algorithms might be striking a balance between encoding semantic and morphosyntactic information, since at higher values of k we usually see more clusters with a coherent semantic theme such as names of geographic locations, names of men, names of women, nouns determining times, similar to the clusters observed in previous literature (Brown et al., 1992). For example, when using $k = 18$ in English, the token "cat_NOUN" appears in the same cluster as the plural version "cats_NOUN". At $k = 800$, the clusters distinguish between the two tokens and "cat_NOUN" is placed together with a number of nouns in the singular such as "budget_NOUN, computer_NOUN, pet_NOUN, wheel_NOUN", while "cats_NOUN" is placed in a cluster of mostly pluralized nouns like "children_NOUN, rooms_NOUN, dogs_NOUN, families_NOUN".

Adjusted Mutual Information (AdjMI) for the same clusterings, Figure 1b, shows a considerable decrease as the hyper-parameter k increases, especially at high values of 500 and 800. This is in line with the expected punishment due to the effect attributed to randomness (see the term for expected

value of Mutual Information in Equation (4)). At values of k closer to the number of PoS tags in the data, *AdjMI* varies little from one clustering to the other. More interestingly, the relative order of separation performance between the languages studied is maintained going from *purity* to *AdjMI*, suggesting that no measure-specific effects are at play.

By studying the frequency of incorrectly classified words types (i.e., of those whose PoS tag does not match the most popular one in their cluster), we find that most (about 85%) occur less than 5 times in the corpora. Such few observations likely do not provide enough information for Brown or Exchange to properly place those words. Therefore, from the already computed clusterings, we remove words with a frequency less than 5 and recalculate the two quality measures.

In Figures 1c and 1d, we can see that both *purity* and *AdjMI* improve considerably. Even in the most difficult case ($k = 18$), where the number of clusters matches that of distinct PoS tags, between 68% and 78% of words are properly placed, an increase of $21\% - 28\%$ compared to the values in Figure 1a. For *AdjMI*, the scores more than double. These results show that even for small corpora, a large amount of morphosyntactic information can be encoded, completely unsupervised, using the Exchange clustering algorithm. (The same behavior can be observed for clusters derived using the Brown Clustering algorithm, see supplementary material.) It also shows that, for low frequency terms, there is not enough contextual information for a proper clustering.

One disadvantage of thresholding by frequency is that, due to the zipfian distribution of word frequencies in natural language, only a fraction of the original vocabulary remains after filtering out words with a frequency less than 5: English ($8\,143\ words - 24.01\%$), French ($9\,020\ words - 17.45\%$), Czech ($37\,026\ words - 22.51\%$). In order to benefit from more reliable word usage estimates, it is necessary to perform the same experiment on larger corpora. Unfortunately bigger manually annotated data sets do not exist. We therefore turn to automatic PoS tagging.

We use UDPipe (Straka and Straková, 2017) with models pretrained on the Universal Dependencies corpora to automatically tag text from the EuroParl multi-language corpus containing transcriptions of European Parliament proceedings (Koehn, 2005). Automated annotations introduce labeling noise that should lead to a decrease in separation performance. Despite this, we expect to still be able to observe good PoS separation.

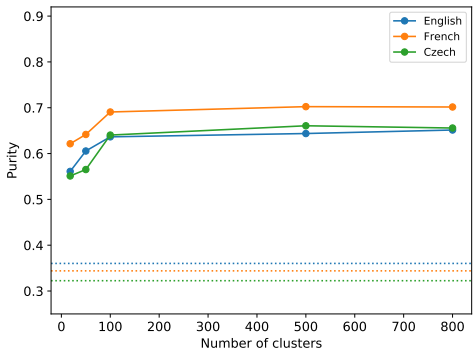
After filtering the EuroParl corpora, the size of remaining vocabulary is considerably larger: English ($60\,373\ words - 37.80\%$), French ($78\,822\ words - 38.60\%$), Czech ($62\,512\ words - 35.19\%$). In Figure 2a we can see that there is a drop in performance that varies with language, but when looking at purity, even in the worst performing clustering (French at $k = 50$), 60% of the vocabulary is still properly separated according to Parts of Speech. A drop in performance can also be observed for *AdjMI* in Figure 2b, with the value dropping for all languages, in some cases reducing by half.

More interestingly, the relative performance order of the languages is changed. PoS separation for Czech outperforms that of the other languages. Actually, PoS separation for Czech on EuroParl data (Figure 2) is scored higher than that of Czech on Universal Dependencies (Figures 1c and 1d). The source of this improvement requires more study for a proper attribution, but could be due to "beneficial" noise introduced by the automatic tagging, or due to the introduction of more sentence structure by human translators.

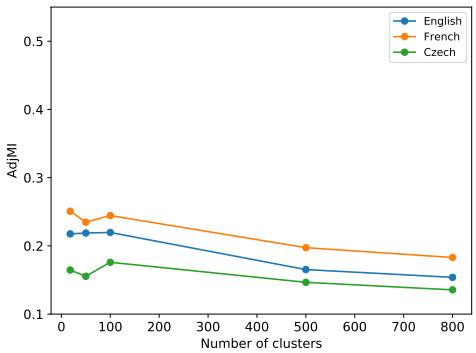
The fact that even at low values of k, for all languages studied, on both corpora, Exchange Word clusters (and also Brown word clusters, see supplementary material) can successfully separate by Parts of Speech, helps understand why word clusters have had such success at PoS tagging whether coupled with Markov Models (Derczynski et al., 2015), Markov Models and morphological features (Owoputi et al., 2013), or just by themselves via M-1 (Bansal et al., 2014).

4.3 The relation between AMI and PoS

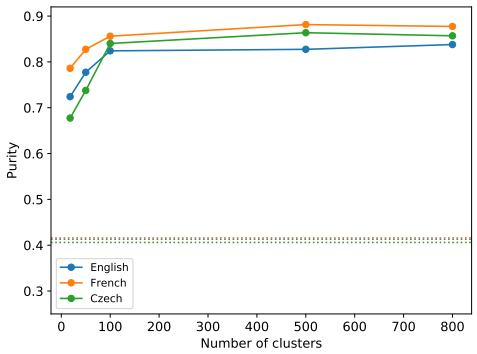
Neither Exchange, nor Brown are guaranteed to converge to a global optimum. Both are greedy algorithms that optimize for high Average Mutual Information (AMI). As we have mentioned earlier, word clusters resulting from Exchange outperform those induced using the Brown clustering algorithm in terms of both *AMI* (the algorithm's optimization goal), *PoS purity* and *Adjusted Mutual Information (AdjMI)*. A natural question to ask is: can one improve the morphosyntactic content of word clusters by obtaining higher AMI, maybe by



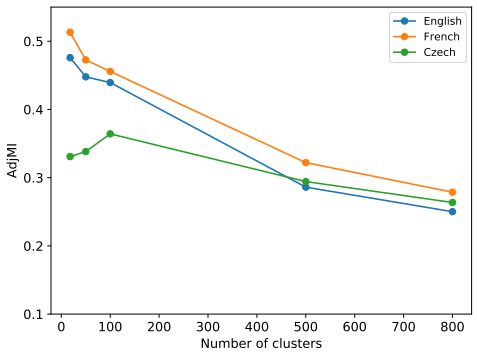
(a) Cluster purity with baselines for $k = 1$. No frequency thresholding.



(b) Adjusted Mutual Information. No frequency thresholding.



(c) Cluster purity with baselines for $k = 1$. Only words with frequency minimum 5.



(d) Adjusted Mutual Information. Only words with frequency minimum 5.

Figure 1: Cluster agreement with manual labels from UD.

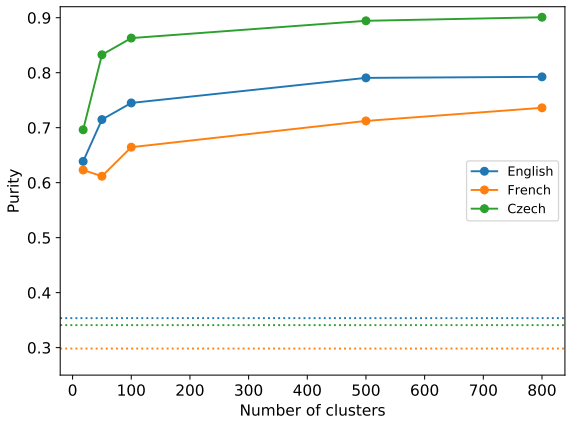
developing new and better AMI-based clustering algorithms?

We answer this question by studying the correlation between *Average Mutual Information* and the two cluster quality measures used earlier: *purity* and *AdjMI*. Brown clustering is a predictable, bottom-up, agglomerative, hard clustering algorithm that for the same hyper-parameter k, generates the same clusters and therefore only one data sample.[2] However, the Exchange algorithm is an iterative clustering algorithm that has a complete and valid cluster partitioning at the end of each iteration. Thus, we can also measure morphosyntactic content in each of these clusterings. In our experiments, we only obtain 10 different data samples from each run of the algorithm, not enough for a correlation analysis.

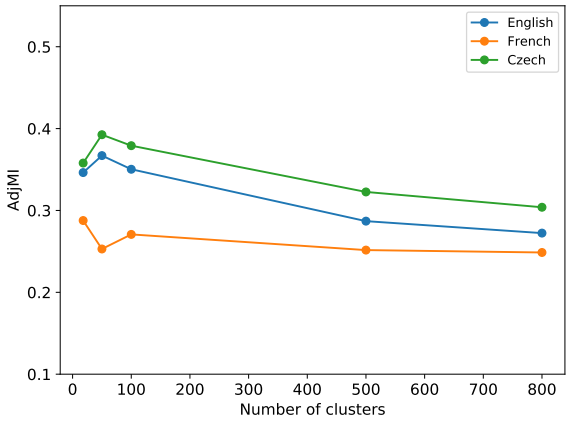
In order to collect more data samples (i.e. more clusterings), we suggest using a stochastic version of Exchange where a percentage of all swaps are performed at random, rather than with the goal of improving AMI. This version of Exchange terminates based on the number of iterations, and generates valid word partitionings of varying quality (from an *AMI* perspective) at the end of each iteration. In this manner, it provides us with more data points (i.e. more different clusterings) for analysis. Due to the small amount of random swaps, at varying *AMI*, we obtain a sufficient number of distinct clusterings to perform a correlation study with sufficient data.

With the stochastic implementation of Exchange, we run 50 iterations for all languages and k combinations studied earlier. In Tables 2 and 3, we show the Pearson and Spearman correlation coefficients between AMI of all clusterings generated by StochasticExchange for a given run, and the two scores used earlier: *purity* and *AdjMI*. Due to space considerations, we only show results for $k = 18$ (i.e., same number of clusters as the number of PoS tags). Correlation coefficients for other combinations are included in the supplementary material. $p < 0.01$ for all correlation experiments here and in the supplementary material and for both correlation coefficients. The analysis presented below also holds for all the correlation re-

[2] Assuming a stable and repeatable tie-breaking process; this is undefined in the literature.



(a) Cluster purity. Dotted lines are baselines for $k = 1$



(b) Adjusted Mutual Information.

Figure 2: Cluster agreement with automatically generated labels from EuroParl. Only words with frequency minimum 5.

Data Set	Pearson	Spearman
EN UD	0.9776	0.7173
FR UD	0.9863	0.3976
CZ UD	0.9883	0.7378

Table 2: Correlation between Average Mutual Information and PoS purity of the clustering resulted from Exchange with $k = 18$. Words with frequency < 5 have been filtered. $p < 0.01$ for all coefficients.

Data Set	Pearson	Spearman
EN UD	0.9897	0.7464
FR UD	0.9845	0.7192
CZ UD	0.9859	0.8930

Table 3: Correlation between Average Mutual Information and AdjMI of the clustering resulted from Exchange with $k = 18$. Words with frequency < 5 have been filtered. $p < 0.01$ for all coefficients.

sults included in the supplementary material.

For both *purity* and *AdjMI*, there is a strong Pearson correlation between higher AMI and better values of the evaluation score. This is independent of the language studied or the number of clusters derived. For Spearman, except for one case, in all combinations studied, there is a high correlation, although to a slightly less extreme degree as with Pearson.

Our experiments show that there is strong correlation between *AMI* and performance in separation of Parts of Speech as measured by *purity* and *AdjMI*. The strong correlation provides grounding for research into new AMI-maximizing word clustering algorithms that can achieve higher AMI than Exchange, or Brown, as such algorithms might be able to separate Parts of Speech even better.

4.4 Effect of polysemy on cluster purity

In previous sections, we studied the ability of word clusters to encode morphosyntactic information. We clustered word types from unstructured text, where each token had its Part of Speech tag appended. The post-pended PoS tags are not used by either Brown, or Exchange. They are essentially invisible to the algorithms, since the both Brown and Exchange recognize words exclusively by internally assigned integer IDs and do not operate at character level.

However, post-pending PoS tags does introduce some information into the text by providing PoS-role disambiguation for each word occurrence.

For example, without post-pended PoS tags, both Exchange, and Brown algorithms, would conflate the two distinct grammatical roles of *show* in the sentence: "Everyone must show their show tickets at the entrance". In this section, we study PoS separation effects caused by such polysemy on Brown and Exchange word clusters.

Both Exchange and Brown construct hard clusters, i.e. each word can be assigned to exactly one word class. Thus, words with multiple roles, such as denominal verbs or deverbal nouns, cannot be differentiated by the algorithms when operating on corpora from languages where the such morphological derivations are performed without employing suffixes or prefixes. In other words, if the lexical form does not change, neither Brown nor Exchange can identify which tokens represent what grammatical role.

The extent of this effect is dependent on language. In English, for example, nouns are often turned into verbs without changing the lexical form through morphological derivation, e.g. *show* as a verb vs *show* as a noun. On the other hand, Czech is highly inflected accounting for gender, case, number and person. This property of each language was not problematic in the experiments we have performed so far due to the fact that postpending the PoS tag from ground-truth (or automatic tags) effectively provides disambiguation of grammatical role. Measuring on the Universal Dependencies corpora, we find that the percentage of polyclass words (i.e. word types that are assigned more than one PoS class tag throughout the corpus) varies by language and increases (as percentage of remaining vocabulary) as we raise the minimum frequency threshold, see Table 4. For English and French, up to 43% of the vocabulary words have more than one tag, while only $5,5\%$ of the Czech vocabulary shares the same property. Part of the reason why so many words have multiple PoS tags has to do with how the various language families derive new words, and part of the reason stems from errors in PoS tagging of large text corpora (Silberztein, 2018).

From a practical point of view, polysemous words create an upper bound on the effectiveness of hard clustering for Part of Speech separation (PoS). In Figure 3, we show PoS purity for clusters induced over Universal Dependencies (UD) corpora, where we consider all polyclass words as clustered incorrectly. We also show the minimum purity (when $k = 1$) as well as the upper bound given by the polysemy of each language as observed from the manual labels. The evaluation strictly penalizes multiclass polysemy and ignores errors in labeling, such as those identified by Silberztein (2018). For example, in the UD English corpus, even though only 3 occurrences of the word "them" are incorrectly tagged as adverb, while the remaining 750 are correctly labeled as pronoun. We defer to the data and consider the word to be impossible to correctly allocate to a cluster. We use such a strict evaluation as it provides a lower bound on what can be expected from Exchange and Brown clusters given the current data. Correcting PoS tags in the data would probably improve PoS separation, however, such corrections are outside the scope of the work in this paper.

We should point out that this evaluation is not representative of the expected PoS tagging performance of word clusters on any given corpus, as for such taggers one would employ a different strategy, such as, for instance, always outputting the most popular PoS tag for any given word type. On top of that, our evaluation here does not take into account the frequency of tokens, which would be highly relevant for PoS taggining performance, but not for our evaluation.

As expected, the most affected language is English, due to the high level of polysemy in the data. Here *purity* drops from 72.4 to 42.32 for $k = 18$, when compared with results in Figure 1c. It is followed by a 20 point drop for French, and only a few points for Czech, the most morphologically rich of the three and with the least amount of ambiguity in grammatical role. The results suggest that even in the presence of language ambiguity, and considering the strictest evaluation, Exchange and Brown clusters successfully encode a considerable amount of morphosyntactic information, which varies by language. These, together with results presented earlier in this paper provide empirical evidence for using word clusters as word representations in downstream NLP systems addressing tasks that rely on morphosyntactic knowledge of the language targeted (e.g. dependency parsing), or for use in new paradigms such as data programming (Ratner et al., 2016), where cluster membership can be a strong signal for probabilistic data labeling, even when considering language ambiguity.

Data Set	Min 1	Min 5
EN UD	15.39	43.02
FR UD	9.09	41.04
CZ UD	1.81	5.51

Table 4: Percentage of vocabulary with multiple PoS tags. Values are calculated relative to the vocabulary remaining after application of threshold.

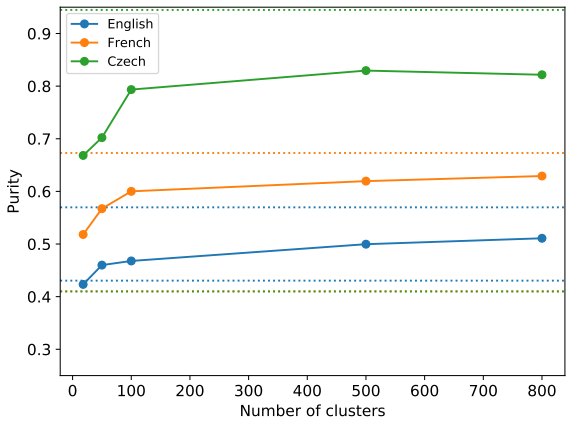


Figure 3: Cluster purity for manually annotated corpora from UD. Only words with frequency minimum 5. Dotted lines are baselines for $k = 1$ and highest achievable purity given polysemy in corpus.

5 Conclusion

In this paper, we quantified the amount of morphosyntactic information encoded in Brown and Exchange word clusters, in a number of languages, from different language families. Our empirical quantification helps explain the success of word clusters as word representations in NLP tasks that rely on morphosyntactic information, such as PoS tagging and Named Entity Recognition. It further provides empirical evidence for using word clusters as word representations in other NLP tasks that require morphosyntactic knowledge of the language targeted (e.g. dependency parsing), or for use in new paradigms such as data programming (Ratner et al., 2016), where cluster membership can be a strong signal for probabilistic data labeling. We have also shown that there is a strong correlation between *AMI* (Brown and Exchange's optimization goal) and performance in PoS separation. The strong correlation demonstrated provides grounding for research into new AMI-maximizing word representation algorithms that can achieve even better AMI optimization than Exchange or Brown.

References

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring Continuous Word Representations for Dependency Parsing. In *Acl*, pages 809–815.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2010. Two Decades of Unsupervised POS Induction: How Far Have We Come? In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 575–584, Stroudsburg, PA, USA. Association for Computational Linguistics.

Manuel Ciosici. 2015. Improving quality of hierarchical clustering for large data series. Master's thesis, Aarhus Universitet, Datalogisk Institut.

Jon Dehdari, Li Ling Tan, and Josef van Genabith. 2016. BIRA: Improved Predictive Exchange Word Clustering. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL). Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2016), Ju*, pages 1169–1174.

Leon Derczynski and Sean Chester. 2016. Generalized Brown Clustering and Roll-up Feature Generation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 1533–1539.

Leon Derczynski, Sean Chester, and Kenneth S Bøgh. 2015. Tune Your Brown Clustering, Please. In *Proceedings of the conference on Recent Advances in Natural Lang Processing (RANLP)*.

Reinhard Kneser and Hermann Ney. 1993. Improved clustering techniques for class-based statistical language modelling. In *Eurospeech*, volume 93, pages 973–976.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Terry Koo, Xavier Carreras Pérez, and Michael Collins. 2008. Simple semi-supervised dependency parsing. *46th Annual Meeting of the Association for Computational Linguistics*, 8(June):595–603.

Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, et al. 2017. Universal Dependencies 2.1.

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

Sven Martin, Jörg Liermann, and Hermann Ney. 1998. Algorithms for bigram and trigram word clustering. *Speech Communication*, 24(1):19–37.

Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pages 1–12.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah a Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT 2013*, June, pages 380–390. Association for Computational Linguistics.

Lizhen Qu, Gabriela Ferraro, Liyuan Zhou, Weiwei Hou, Nathan Schneider, and Timothy Baldwin. 2015. Big Data Small Data, In Domain Out-of Domain, Known Word Unknown Word: The Impact of Word Representation on Sequence Labelling Tasks. In *Proceedings of CoNLL*.

Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3567–3575. Curran Associates, Inc.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Max Silberztein. 2018. Using Linguistic Resources to Evaluate the Quality of Annotated Corpora. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 2–11.

Milan Straka and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Matthew C Swain and Jacqueline Manina Cole. 2016. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *Journal of Chemical Information and Modeling*, page acs.jcim.6b00207.

Jakob Uszkoreit and Thorsten Brants. 2008. Distributed Word Clustering for Large Scale Class-Based Language Modeling in Machine Translation. In *Proceedings of ACL*, June, pages 755–762.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2009. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1073–1080, New York, NY, USA. ACM.