A Cross-Linguistic Comparison of Unsupervised Part-of-Speech Induction Systems

Guido Linders
Saarland University

March 17, 2017

1 Introduction

In unsupervised Part-of-Speech (POS) induction the goal is to cluster words according to their similarity in terms of distribution, function and morphology, using no other information than sequences of words that make up sentences and the sequences of letters that make up words. A systematic cross-lingual comparison is difficult to do because of a lack of annotated data to evaluate on. In addition, there are large differences in annotation schemes disallowing cross-linguistic comparisons. A relatively new project is Universal Dependencies (UD) (Nivre et al., 2016). The goal of this project is to provide consistent universal treebank annotations. This includes annotations with a universal POS (UPOS) tag set, which consists of 17 tags. This means we are able to do a cross-linguistic comparison of different POS induction systems.

In this study three unsupervised POS induction systems will be evaluated on several languages in the Universal Dependencies Treebank. Emphasis will be put on comparing the behavior of the different systems cross-linguistically. In Section 2 I will outline the three systems which will be used in the evaluation. In Section 3 the evaluation method and evaluation metrics will be explained. The results are outlined in Section 4 and the conclusion in Section 5.

2 Part-of-Speech Induction Systems

In this Section three systems for unsupervised POS induction are introduced. One of the earliest and simplest approaches to POS induction is using class-based n-grams, popularized as Brown clustering (Brown et al., 1992). This method is outlined in Section 2.1. Another method, evaluated in this study, builds upon Brown clustering and incorporates morphological and frequency information (Clark, 2003). To be consistent with Christodoulopoulos et al. (2010) I will call it class-based n-grams with morphology for now. It is outlined in Section 2.2. The last system uses a Bayesian Hidden Markov Model (HMM) and directly models the constraint that each word type can only have one tag (Lee et al., 2010). I will call this system the "Type Tagger". It is outlined in Section 2.3.

All methods induce clusters based on word types. This means that all word tokens are mapped to the same cluster. In addition, for all models we have to manually select the number of cluster that we want to induce. Lastly, all models use distributional information of words as primary source of information.

2.1 Brown Clustering

In Brown clustering the intuition is to cluster words solely based on their distribution in a text (Brown et al., 1992). Brown clustering uses a class-based bigram language model. Let $w_1, \cdots, w_n$ be a sequence of words

(a corpus) and C be a clustering, mapping each word to a cluster. The model is then defined as in equation 1.

$$P(w_1, \cdots, w_n) = \prod_{i=1}^{n} P(w_i|C(w_i))P(C(w_i)|C(w_{i-1})) \quad (1)$$

As the model is type-based it clusters a text with a vocabulary with V as the number of words in the vocabulary into C clusters. In the first step of the algorithm we assign the C most frequent words their own cluster. In the next step we assign the remaining (V-C) words to one of the existing clusters, starting with the next most frequent word. The word is assigned to the cluster with the smallest loss on average mutual information if that word is added. For the evaluation I use the reimplementation of Brown clustering by Liang (2005).[1]

2.2 Class-based n-grams with Morphology

Clark (2003) uses the same model as Brown clustering, which was defined in Equation 1. He however uses a different clustering algorithm. In addition to the basic model, frequency and morphological information can be added. The clustering algorithm is an exchange algorithm, where each word type is moved to the cluster which leads to the lowest perplexity on the corpus. The initial assignment of word types to a cluster is done pseudo-randomly.

The frequency information tells the model how many word types the cluster has. The intuition behind this is that that rare word types are more likely to have a tag from an open class. Very frequent word types more often have a tag from a closed class. This usually means that open classes contain more word types than closed classes. Morphological information is represented as a first-order Hidden Markov Model (HMM) with the states being the letters of the word. With this approach the model can capture recurring sequences of letters in words that represent morphological inflections.

Both the frequency and morphological information is added to the model through using a Bayesian approach. This means that the basic distributional model does not give only the most likely prediction, but gives a probability distribution over all possibilities with a probability on how certain it is about each prediction. The frequency and morphological information are then added as priors to bias the model. I evaluate both the basic system, which I will call "Clark 1" (named after the author), as the systems including only morphological information ("Clark 2") and both frequency and morphological information ("Clark 3"). I use the implementation used in the original paper by Clark (2003).[2]

2.3 Type Tagger

The main idea behind the approach of Lee et al. (2010) evolves from the fact that Part-of-Speech tag distributions among word tokens are sparse. By directly constraining the model to assigning only one tag per word type they reduce the number of parameters and provide linguistic guidance to the model on how to cluster. The model consists of two phases. In the first phase the model creates a so-called lexicon, listing possible word types for each cluster with probabilities. Next, based on the lexicon, a Bayesian token-based Hidden Markov Model (HMM) is trained with Gibbs sampling. This HMM is constrained to one tag per word type. The HMM has transitions between POS tags and, using the lexicon, emits words from POS tags.

The lexicon can be built using different kinds of information. In the most basic model the clustering is purely based on the distribution of word types in a corpus. This basic model can be augmented with a prior distribution over the number of word types per tag. This distribution is automatically learned during the cluster induction. A third augmentation is the use of different features that are based on the letters and letter sequences of each word type. I evaluate the system with the basic model, which I will call "TT 1" (shortened version of

[1] The source code was retrieved from: `https://github.com/percyliang/brown-cluster`

[2] The source code was retrieved from: `https://github.com/ninjin/clark_pos_induction`

"Type Tagger 1"), as well as the system with the prior frequency distribution (TT 2) and the system with both the prior frequency distribution and the features (TT 3). I use the implementation used in the original paper by Lee et al. (2010).[3]

3 Method

3.1 Data Acquisition

The corpora are acquired through the Universal Dependencies Treebank. In this evaluation version 1.4 is used (Nivre et al., 2016). I used all languages which had more than 200,000 word tokens. This leads to 14 different languages: Arabic, Brazilian Portuguese, Catalan, Czech, English, French, German, Hindi, Latin, Norwegian, Portuguese (Mixture of European and Brazilian Portuguese), Romanian, Russian and Spanish. As the task is unsupervised clustering, I merge the "train", "dev", and "test" set into one large data set. To be able to make cross-linguistic comparisons I equalize the number of word tokens for each language. To avoid cutting off a data set in the middle of a sentence I allowed each sentence to finish. This means that all data sets are a few word tokens larger than 200,000. I assume this difference in word tokens to be negligible.

3.2 Evaluation Metrics

As Christodoulopoulos et al. (2010) noted, there is no consensus on which evaluation measure to use. Therefore I compare the performance of the systems on four different measures, briefly sketched below: many-to-one mapping, one-to-one mapping, V-measure and variation of information. For the evaluation I use a tool from Christodoulopoulos, which was used in (Christodoulopoulos et al., 2010).[4] All metrics range from 0 to 1 (I denote them in percentages), except Variation of information which is measured in bits. Furthermore, for all metrics except one-to-one mapping holds that a higher score is a better performance. For variation of information a lower score is a better performance.

3.2.1 Many-to-one mapping

Many-to-one mapping compares each assigned tag to the tag in the gold standard. Each induced cluster is assigned the tag label from the gold standard that is occurs the most. Finally, the accuracy is measured by dividing the number of correct classifications by the total number of word tokens.

3.2.2 One-to-one mapping

In many-to-one mapping each induced cluster can be assigned any tag label, regardless of whether there is already an induced cluster with that label. In one-to-one mapping tag label can only be assigned to one induced cluster. The process of assigning a tag label to an induced cluster is done greedily. Finally, again the accuracy is measured like in many-to-one mapping.

3.2.3 V-Measure

V-measure is also entropy-based and combines both homogeneity (h) and completeness (c). Homogeneity means that in every induced cluster there are only word tokens with the same gold standard tag. Completeness is the opposite. It means that all word tokens with the same gold standard tag are in the same induced cluster.

[3]The source code was retrieved from: `http://people.csail.mit.edu/yklee/code.html`

[4]The source code was retrieved from: `http://christos-c.com/resources.html`

Both are defined in Equation 2, with G being the set of gold standard tags and C being the set of induced clusters.

$$h = 1 - \frac{H(G|C)}{H(G)} \quad \text{and} \quad c = 1 - \frac{H(C|G)}{H(C)} \tag{2}$$

The V-measure is defined as the balanced combination of homogeneity and completeness and is defined in Equation 3.

$$V = \frac{2 \cdot h \cdot c}{h + c} \tag{3}$$

3.2.4 Variation of Information

Variation of Information (VI) is an entropy-based measure that intuitively measures the amount of information that changed from going from one clustering to another. If n is the total number of word tokens and n_k the number of word tokens in cluster C_k. Then we can define the probability of a word, that is randomly picked, being in cluster C_k as in Equation 4.

$$P(k) = \frac{n_k}{n} \tag{4}$$

The entropy of a set of words, mapped to a cluster in the set of clusters K, is defined in Equation 5.

$$H(C) = -\sum_{k \in K} P(k) \log_2(P(k)) \tag{5}$$

The mutual information between two clusters is defined in Equation 6.

$$I(C, G) = \sum_{k \in K} \sum_{k \in K} P(k, k) \log_2 \frac{P(k, k)}{P(k)P(k)} \quad \text{where} \quad P(k, k) = \frac{n_k \cap n_k}{n} \tag{6}$$

Finally, if G is the set of gold standard tags and C is the set of induced clusters Variation of Information is measured as in Equation 7.

$$VI(C, G) = H(C) + H(G) - 2I(C, G) \tag{7}$$

3.3 Parameter Decisions

To be able to do a cross-linguistic comparison and reliably compare the performance of the different systems, I try to keep as much parameters as possible constant. This includes number of induced clusters, which is set to 17 and number of word tokens, which is set to 200,000. In addition I wanted to keep the parameters of the systems similar for each language evaluated on. Both in the implementation of Brown clustering and the Type Tagger the number of clusters to-be induced was the only parameter that could be set (next to choosing a model for the Type Tagger). The implementation of class-based n-grams with morphology required a small bit of experimenting with different parameter settings. It should be noted that this implementation is not very stable and does not work on all parameter settings for all languages. This experimenting was done by using two languages (Arabic and English) and trying different parameter settings, assuming independence between the parameters. Parameters optimized in this way were the number of iterations, number of states for the HMM, the use of true weights for HMM training and the influence of the prior frequency information.

4 Results

4.1 Corpus Statistics

Statistics about the different corpora have been summarized in Table 4.1. We can observe that for languages we perceive as "morphologically richer"[5] usually have more word types, with examples being Czech, German,

[5]I did not fully specify the notion of morphologically rich languages. In this evaluation I define a language as morphologically richer as another language if there is more marking of morpho-syntactic properties and roles through

Language	#Tokens	#Types	Token/Type	#Hapaxes	#snts	Avg |snt|	#Tags	Entropy
Arabic	200036	29053	6.89	15028	4649	43.03	17	3.16
Br. Portuguese	200006	24492	8.17	14009	8117	24.64	14	3.26
Catalan	200003	17564	11.39	8739	5490	36.43	17	3.46
Czech	200007	33377	6.00	19292	12484	16.02	17	3.28
English	200003	16508	12.12	7739	12322	16.23	17	3.60
French	200005	27373	7.31	16835	7973	25.09	18	3.43
German	200005	37583	5.32	25516	10512	19.03	16	3.49
Hindi	200012	14303	13.98	6282	9599	20.84	16	3.24
Latin	200001	10322	19.38	4694	12902	15.50	15	3.23
Norwegian	200020	24554	8.15	14175	12985	15.40	17	3.47
Portuguese	200007	23389	8.55	13194	7328	27.29	18	3.36
Romanian	200005	30000	6.67	16732	8707	22.97	17	3.35
Russian	200025	39022	5.13	24055	11771	16.99	15	3.24
Spanish	200030	28647	6.98	17713	7337	27.26	17	3.43

Table 1: Statistics of the corpora for the different languages used for the evaluation. Note that the entropy is measured in bits.

Language	Brown	Clark 1	Clark 2	Clark 3	TT 1	TT 2	TT 3
Arabic	**66.18**	65.36	62.69	59.53	62.12	62.33	64.99
Br. Portuguese	**67.89**	61.50	56.47	57.06	59.79	61.97	60.58
Catalan	65.75	65.82	67.38	68.90	66.54	69.14	**70.60**
Czech	57.94	61.39	62.35	65.48	58.87	59.29	**66.91**
English	62.99	63.92	63.17	65.31	64.33	65.18	**65.41**
French	67.81	66.31	69.27	**71.63**	66.78	68.50	71.43
German	57.22	56.32	52.56	53.68	56.58	57.79	**64.42**
Hindi	61.09	65.21	61.99		63.76	61.04	**68.17**
Latin	64.76	63.03	59.18	58.87	63.59	65.11	**65.56**
Norwegian	62.94	64.48	62.34	66.84	62.90	68.97	**70.18**
Portuguese	66.11	65.55	67.35	68.99	65.96	68.16	**70.58**
Romanian	59.30	59.96	47.39	49.76	52.16	55.86	**65.65**
Russian	58.74	57.45	59.01	61.30	55.52	58.17	**68.45**
Spanish	68.50	67.72	70.73	**71.26**	65.92	70.51	68.66
Average	63.37	63.14	61.56	62.97	61.77	63.72	**67.26**
Standard dev.	3.78	3.28	6.19	6.38	4.38	4.70	2.92

Table 2: Evaluation on many-to-one mapping

Romanian and Russian. This also roughly corresponds with the number of words that occur only once in the corpus. We also see that the number of sentences varies widely across languages. Furthermore we can observe that not all languages use all Part-of-Speech (POS) tags. Note that some languages use 18 POS tags despite there being 17 POS tags. This is due to the use of a special tag in case it is unclear what tag should be assigned to a word, for example in contractions of words that have different tags. Lastly we can see that the entropy does not vary greatly across languages, meaning the frequency distribution over tags does not vary greatly.

4.2 Systems Comparison

The systems have been evaluated using the four different evaluation metrics described in Section 3.2. The results have been summarized in a table for each evaluation metric. See Table 2 for the results on many-to-one mapping, Table 3 for one-to-one mapping, Table 4 for V-measure and Table 5 for variation of information. Brown clustering has been shortened to "Brown", the different systems using class-based n-grams with morphology to "Clark 1", "Clark 2" and "Clark 3" and the different Type Tagger systems to "TT 1", "TT 2", and "TT 3". The best performing system for each language is highlighted. The first thing I should note is that Clark 3 failed to generate a clustering for Hindi. This is due to the implementation that was not very robust.

Let me first note that there is not always consistency between the evaluation metrics. When comparing two systems we sometimes see an improvement on one evaluation metric and a decrease in performance on

variation of basic word forms.

Language	Brown	Clark 1	Clark 2	Clark 3	TT 1	TT 2	TT 3
Arabic	43.74	**46.01**	44.75	45.20	36.74	38.81	44.45
Br. Portuguese	**50.73**	44.18	41.94	47.84	45.32	39.90	39.50
Catalan	51.33	46.93	**57.42**	56.29	51.42	54.23	55.49
Czech	**46.47**	44.52	40.52	45.60	35.78	41.66	39.09
English	50.65	52.22	50.39	54.55	52.10	54.09	**55.86**
French	53.51	44.18	53.89	**60.27**	45.04	49.51	50.77
German	39.10	40.30	36.09	36.62	43.71	46.85	**51.10**
Hindi	42.96	44.82	**50.67**		44.87	36.66	48.49
Latin	**45.35**	40.61	39.29	38.28	37.46	38.46	39.63
Norwegian	49.76	50.05	48.35	52.22	47.43	53.45	**56.87**
Portuguese	49.38	42.48	48.60	**51.64**	36.94	38.39	39.69
Romanian	39.77	38.94	28.70	35.54	32.01	33.90	**40.90**
Russian	42.24	38.91	39.54	**49.22**	35.02	37.03	44.68
Spanish	52.78	47.45	57.24	**59.17**	44.35	49.33	50.50
Average	46.98	44.40	45.53	**48.65**	42.01	43.73	46.93
Standard dev.	4.65	3.84	8.03	7.88	6.08	6.96	6.38

Table 3: Evaluation on one-to-one mapping

Language	Brown	Clark 1	Clark 2	Clark 3	TT 1	TT 2	TT 3
Arabic	44.11	45.11	43.72	42.60	40.88	41.76	**45.89**
Br. Portuguese	**51.90**	49.05	44.29	47.45	46.43	47.58	48.73
Catalan	54.26	55.87	58.76	61.15	56.16	59.94	**62.15**
Czech	41.99	43.39	45.11	**51.13**	40.75	42.95	50.26
English	51.43	53.80	53.65	56.48	54.31	55.56	**58.10**
French	54.28	52.26	56.74	**60.52**	52.50	56.76	59.41
German	45.46	44.79	42.94	45.84	45.13	48.92	**54.60**
Hindi	49.77	53.06	53.04		52.15	50.97	**54.64**
Latin	50.66	48.66	46.02	46.93	48.60	51.37	**52.54**
Norwegian	50.24	50.70	49.94	55.54	49.19	55.60	**57.52**
Portuguese	51.39	51.15	52.61	55.72	49.73	54.02	**56.49**
Romanian	42.50	43.13	32.78	37.85	36.21	39.97	**50.72**
Russian	41.10	39.73	45.05	51.73	39.01	42.42	**51.76**
Spanish	54.29	52.21	59.37	**60.99**	52.38	57.46	56.53
Average	48.81	48.78	48.86	51.84	47.39	50.38	**54.24**
Standard dev.	4.61	4.62	7.09	7.11	5.95	6.34	4.33

Table 4: Evaluation on V-measure

Language	Brown	Clark 1	Clark 2	Clark 3	TT 1	TT 2	TT 3
Arabic	3.87	3.82	3.80	3.79	4.18	4.08	**3.77**
Br. Portuguese	**3.36**	3.52	3.88	3.58	3.84	3.81	3.69
Catalan	3.25	3.21	2.94	2.79	3.23	2.92	**2.77**
Czech	3.98	4.01	3.94	**3.43**	4.26	4.06	3.56
English	3.46	3.42	3.47	3.22	3.44	3.34	**3.12**
French	3.24	3.46	3.13	**2.84**	3.52	3.17	2.97
German	3.98	3.98	4.12	3.92	4.08	3.71	**3.34**
Hindi	3.43	3.24	**3.19**		3.41	3.48	3.21
Latin	3.40	3.59	3.82	3.75	3.71	3.49	**3.39**
Norwegian	3.55	3.55	3.68	3.26	3.80	3.31	**3.14**
Portuguese	3.43	3.52	3.37	**3.10**	3.67	3.34	3.17
Romanian	4.10	4.08	4.69	4.19	4.68	4.40	**3.60**
Russian	4.17	4.20	3.86	**3.10**	4.39	4.11	3.44
Spanish	3.29	3.49	2.87	**2.72**	3.50	3.10	3.16
Average	3.61	3.65	3.63	3.36	3.84	3.59	**3.31**
Standard dev.	0.32	0.30	0.48	0.44	0.41	0.43	0.27

Table 5: Evaluation on variation of information

another evaluation metric. On three out of four evaluation metrics TT 3 was the best performing system across the different languages. Only on one-to-one mapping this is not the case. There seem to be large differences in performance measurements by the different evaluation metrics. While on many-to-one mapping and V-measure TT 3 performs best quite consistently, this is not the case for both one-to-one mapping and variation of information. TT 3 also seems to perform most stable across languages, leading to the smallest differences in scores cross-linguistically on all but one-to-one mapping.

We see that on all systems not including morphological information, the morphologically richer languages Czech, German, Romanian and Russian systematically perform worse on all evaluation metrics. Recall that all these languages also had more different word types than all other languages, which is likely to be an explanation for the performance differences. We can also see that including morphological information or information on the word strings (more specifically, comparing Clark 1 to Clark 2 and TT 2 to TT 3) lead to mixed results on Clark 2 and always lead to an improvement on TT 3 on all evaluation metrics except one-to-one mapping. On the Type Tagger on one-to-one mapping leads to better performance on most languages. In addition we can observe that adding morphological information to the Type Tagger has the largest increase in performance on morphologically richer languages like Czech, German, Romanian and Russian for all evaluation metrics except one-to-one mapping. The results on the Type Tagger suggest that, especially for morphologically richer languages, the right morphological information can lead to better performance. Note that, in contrast to my results, Clark (2003) observed substantial improvements for all languages from the Multext East Corpus he evaluated on. This suggests that the parameter settings for class-based n-grams with morphology are not set right in my evaluation.

Including frequency information overall improves the performance for both class-based n-grams with morphology and the Type Tagger. For class-based n-grams with morphology Arabic is the only language consistently performing worse. For the Type Tagger Hindi is the only language which did not show an improvement. Also Arabic showed the least improvement. Unfortunately no results on Clark 3 for Hindi have been generated. Arabic is the only language not from the Indo-European language family and Hindi is also quite distantly related, suggesting that the way frequency information is used in the systems might not be helpful for all languages or language families.

Both Brown and Clark 1 achieve similar performance on all evaluation metrics, except on one-to-one mapping. This is probably due to the two systems using the same model and only differing in the optimization algorithm. Note also that both Clark 2 and Clark 3 seem to be most unstable across languages with the largest differences across languages. This could be due to a larger number of parameter settings that needs to be decided and was kept constant across languages.

We can also observe that the languages with the best performance on most systems and evaluation metrics are Catalan, French and Spanish. Those languages are also closely related to each other and are perceived as relatively morphologically rich. Both the French and Spanish corpora contain relatively a lot of different word types, compared to the other languages. It is therefore perhaps surprising to see that these languages perform best. The fact that these three languages are closely related and they are morphologically richer suggests that it is due to other specific phenomena in these languages that make them perform best on most of the systems. The best overall performing system on all evaluation metrics for both French and Spanish is Clark 3, while for most other languages this is TT 3.

Due to their similarity, we would expect Brazilian Portuguese and Portuguese to perform similarly on the different systems. This is the case for Brown. For all other systems, especially Clark 2 and TT3, the performance differs significantly, with Brazilian Portuguese consistently ranking worse than Portuguese on all evaluation metrics but one-to-one and all other systems than Clark 1. Previously was mentioned that including morphology in class-based n-grams with morphology had mixed effects on the performance. It is interesting to see that on Portuguese including morphology lead to a decrease in performance while on Portuguese the opposite is true. When looking at the statistics of both corpora we see more or less similar statistics except that Brazilian Portuguese uses 14 POS tags, compared to 18 for Portuguese. Even though the POS induction

task is fully unsupervised, which means this information is not known to the systems, the way the corpora are annotated could influence the evaluation itself.

5 Conclusion

First of all we can conclude there is not always consistency among the evaluation measures. Especially the scores on one-to-one mapping often do not correspond well with the other evaluation metrics. For this reason, assuming one-to-one mapping has been implemented correctly, I am questioning the reliability and stability of the one-to-one mapping evaluation metric.

We can also conclude that TT3 is the best performing and most stable system in my evaluation. The way frequency information is used generally improves performance, but this is not the case for all languages. As Christodoulopoulos et al. (2010) found, including morphological information improves performance on most languages, in particular languages with a rich morphology. The related languages Catalan, French and Spanish seem to lead to the best performance on most systems, while being perceived as relatively morphologically rich compared to the other languages evaluated on.

Future research could look into the behavior, reliability and stability of the evaluation metrics, in particular one-to-one mapping. Furthermore this evaluation was limited to only three systems, which were all type-based. This could be extended to a larger scale, including toke-based methods. There could also be looked into other parameters that were not varied in this evaluation, such as data size and number of clusters to induce. Besides, the kind of texts used could alter the results.

Finally, the number of languages could also be increased or altered. I did not evaluate on any language perceived as morphologically poor. In addition most languages are from the same language family. Totally different languages could lead to different results. We could look at different corpora for the same language to see the influence of using different data sources and different kinds of texts. It would also be interesting to look at languages with a more free word order as distributional information is a primary source of information for most models. In particular it would be interesting to know why Catalan, French and Spanish perform best and what caused the difference in performance between Brazilian Portuguese and Portuguese.

References

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D. and Lai, J. C. (1992). Class-based n-gram models of natural language. Computational linguistics *18*, 467–479.

Christodoulopoulos, C., Goldwater, S. and Steedman, M. (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 pp. 575–584, Association for Computational Linguistics.

Clark, A. (2003). Combining distributional and morphological information for part of speech induction. In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1 pp. 59–66, Association for Computational Linguistics.

Lee, Y. K., Haghighi, A. and Barzilay, R. (2010). Simple type-level unsupervised POS tagging. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing pp. 853–861, Association for Computational Linguistics.

Liang, P. (2005). Semi-supervised learning for natural language. PhD thesis, Massachusetts Institute of Technology.

Nivre, J., Agić, Ž., Ahrenberg, L., Aranzabe, M. J., Asahara, M. et al. (2016). Universal Dependencies 1.4. `http://hdl.handle.net/11234/1-1827`.