A chance-corrected measure of inter-annotator agreement for syntax

Arne Skjærholt
Language technology group, dept. of informatics
University of Oslo
arnskj@ifi.uio.no

Abstract

Following the works of Carletta (1996) and Artstein and Poesio (2008), there is an increasing consensus within the field that in order to properly gauge the reliability of an annotation effort, chance-corrected measures of inter-annotator agreement should be used. With this in mind, it is striking that virtually all evaluations of syntactic annotation efforts use uncorrected parser evaluation metrics such as bracket F_1 (for phrase structure) and accuracy scores (for dependencies).

In this work we present a chance-corrected metric based on Krippendorff's α, adapted to the structure of syntactic annotations and applicable both to phrase structure and dependency annotation without any modifications. To evaluate our metric we first present a number of synthetic experiments to better control the sources of noise and gauge the metric's responses, before finally contrasting the behaviour of our chance-corrected metric with that of uncorrected parser evaluation metrics on real corpora.[1]

1 Introduction

It is a truth universally acknowledged that an annotation task in good standing be in possession of a measure of inter-annotator agreement (IAA). However, no such measure is in widespread use for the task of syntactic annotation. This is due to a mismatch between the formulation of the agreement measures, which assumes that the annotations have no or relatively little internal structure, and syntactic annotation where structure is the entire point of the annotation. For this reason efforts to gauge the quality of syntactic annotation are hampered by the need to fall back to simple accuracy measures. As shown in Artstein and Poesio (2008), such measures are biased in favour of annotation schemes with fewer categories and do not account for skewed distributions between classes, which can give high observed agreement, even if the annotations are inconsistent.

In this article we propose a family of chance-corrected measures of agreement, applicable to both dependency- and constituency-based syntactic annotation, based on Krippendorff's α and tree edit distance. First we give an overview of traditional agreement measures and why they are insufficient for syntax, before presenting our proposed metrics. Next, we present a number of synthetic experiments performed in order to find the best distance function for this kind of annotation; finally we contrast our new metric and simple accuracy scores as applied to real-world corpora before concluding and presenting some potential avenues for future work.

1.1 Previous work

The definitive reference for agreement measures in computational linguistics is Artstein and Poesio (2008), who argue forcefully in favour of the use of chance-corrected measures of agreement over simple accuracy measures. However, most evaluations of syntactic treebanks use simple accuracy measures such as bracket F_1 scores for constituent trees (NEGRA, Brants, 2000; TIGER, Brants and Hansen, 2002; Cat3LB, Civit et al., 2003; The Arabic Treebank, Maamouri et al., 2008) or labelled or unlabelled attachment scores for dependency syntax (PDT, Hajič, 2004; PCEDT Mikulová and Štěpánek, 2010; Norwegian Dependency Treebank, Skjærholt, 2013). The only work we know of using chance-corrected metrics

[1] The code used to produce the data in this paper, and some of the datasets used, are available to download at https://github.com/arnsholt/syn-agreement/

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 934–944,
Baltimore, Maryland, USA, June 23-25 2014.

is Ragheb and Dickinson (2013), who use MASI (Passonneau, 2006) to measure agreement on dependency relations and head selection in multi-headed dependency syntax, and Bhat and Sharma (2012), who compute Cohen's κ (Cohen, 1960) on dependency relations in single-headed dependency syntax. A limitation of the first approach is that token ID becomes the relevant category for the purposes of agreement, while the second approach only computes agreements on relations, not on structure.

In grammar-driven treebanking (or parsebanking), the problems encountered are slightly different. In HPSG and LFG treebanking annotators do not annotate structure directly. Instead, the grammar parses the input sentences, and the annotator selects the correct parse (or rejects all the candidates) based on discriminants[2] of the parse forest. In this context, de Castro (2011) developed a variant of κ that measures agreement over discriminant selection. This is different from our approach in that agreement is computed on annotator decisions rather than on the treebanked analyses, and is only applicable to grammar-based approaches such as HPSG and LFG treebanking.

The idea of using edit distance as the basis for an inter-annotator agreement metric has previously been explored by Fournier (2013). However that work used a boundary edit distance as the basis of a metric for the task of text segmentation.

1.2 Notation

In this paper, we mostly follow the notation and terminology of Artstein and Poesio (2008), with some additions. The key components in an agreement study are the *items* annotated, the *coders* who make judgements on individual items, and the *annotations* created for the items. We denote these as follows:

- The set of items $I = \{i_1, i_2, \dots\}$
- The set of coders $C = \{c_1, c_2, \dots\}$
- The set of annotations X is a set of sets $X = \{X_i | i \in I\}$ where each set $X_i = \{x_{ic} | c \in C\}$ contains the annotations for each item. If not all coders annotate all items, the different X_i will be of different sizes.

[2] A discriminant is an attribute of the analyses produced by the grammar where some of the analyses differ, e.g. is the word *jump* a noun or a verb, or does a PP attach to a VP or the VP's object NP.

In the case of nominal categorisation we will also use the set K of possible categories.

2 The metric

The most common metrics used in computational linguistics are the metrics κ (Cohen, 1960, introduced to computational linguistics by Carletta, 1996) and π (Scott, 1955). These metrics express agreement on a nominal coding task as the ratio $\kappa, \pi = {A_o - A_e}/{1 - A_e}$ where A_o is the observed agreement and A_e the expected agreement according to some model of "random" annotation. Both metrics have essentially the same model of expected agreement:

$$A_e = \sum_{k \in K} P(k|c_1)P(k|c_2) \quad (1)$$

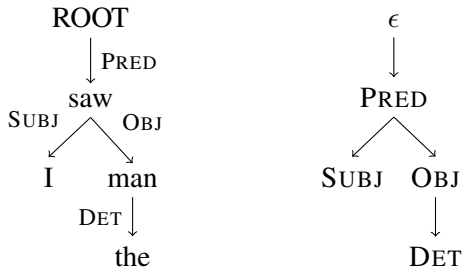
differing only in how they estimate the probabilities: κ assigns separate probability distributions to each coder based on their observed behaviour, while π uses the same distribution for both coders based on their aggregate behaviour.

Now, if we want to perform this same kind of evaluation on syntactic annotation it is not possible to use κ or π directly. In the case of dependency-based syntax we could conceivably use a variant of these metrics by considering the ID of a token's head as a categorical variable (the approach taken in Ragheb and Dickinson, 2013), but we argue that this is not satisfactory. This use of the metrics would consider agreement on categories such as "tokens whose head is token number 24", which is obviously not a linguistically informative category. Thus we have to reject this way of assessing the reliability of dependency syntax annotation. Also, this approach is not directly generalisable to constituency-based syntax.

For dependency syntax we could generalise these metrics similarly to how κ is generalised to κ_w to handle partial credit for overlapping annotations. Let the function $\text{LAS}(t_1, t_2)$ be the number of tokens with the same head and label in the two trees t_1 and t_2, $T(i)$ the set of trees possible for an item $i \in I$, and tokens the number of tokens in the corpus. Then we can compute an expected agreement as follows:

$$A_e = \frac{1}{\text{tokens}} \sum_{i \in I} \sum_{t_1, t_2 \in T(i)^2} \text{LAS}_e(t_1, t_2) \quad (2)$$

$$\text{LAS}_e(t_1, t_2) = P(t_1|c_1)P(t_2|c_2)\text{LAS}(t_1, t_2)$$



(a) The original dependency tree

(b) The tree used in comparisons

Figure 1: Transformation of dependency trees before comparison

We see three problems with this approach. First of all the number of possible trees for a sentence grows exponentially with sentence length, which means that explicitly iterating over all possible such pairs is computationally intractable, nor have we been able to easily derive an algorithm for this particular problem from standard algorithms.

Second, the question of which model to use for $P(t|c)$ is not straightforward. It is possible to use generative parsing models such as PCFGs or the generative dependency models of Eisner (1996), but agreement metrics require a model of *random* annotation, and as such using models designed for parsing runs the risk of over-estimating A_e, resulting in artificially low agreement scores.

Finally, it may be hard to establish a consensus in the field of which particular metric to use. As shown by the existence of three different metrics (κ, π and S (Bennett et al., 1954)) for the relatively simple task of nominal coding, the choice of model for $P(t|c)$ will not be obvious, and thus differing choices of generative model as well as different choices for parameters such as smoothing will result in subtly different agreement metrics. The results of these different metrics will not be directly comparable, which will make the results of groups using different metrics unnecessarily hard to compare.

Instead, we propose to use an agreement measure based on Krippendorff's α (Krippendorff, 1970; Krippendorff, 2004) and tree edit distance. In this approach we compare tree structures directly, which is extremely parsimonious in terms of assumptions, and furthermore sidesteps the problem of probabilistically modelling annotators' behaviour entirely. Krippendorff's α is not as commonly used as κ and π, but it has the advantage of being expressed in terms of an arbitrary *distance function* δ.

A full derivation of α is beyond the scope of this article, and we will simply state the formula used to compute the agreement. Krippendorff's α is normally expressed in terms of the ratio of observed and expected disagreements: $\alpha = 1 - {}^{D_o}/_{D_e}$, where D_o is the mean squared distance between annotations of the same item and D_e the mean squared distance between all pairs of annotations:

$$D_o = \sum_{i \in I} \frac{1}{|X_i| - 1} \sum_{c \in C} \sum_{c' \in C} \delta(x_{ic}, x_{ic'})^2$$

$$D_e = \frac{1}{\sum_{i \in I} |X_i| - 1} \sum_{i \in I} \sum_{c \in C} \sum_{i' \in I} \sum_{c' \in C} \delta(x_{ic}, x_{i'c'})^2$$

Note that in the expression for D_e, we are computing the difference between annotations for *different* items; thus, our distance function for syntactic trees needs to be able to compute the difference between arbitrary trees for completely unrelated sentences. The function δ can be any function as long as it is a metric; that is, it must be (1) non-negative, (2) symmetric, (3) zero only for identical inputs, and (4) it must obey the triangle inequality:

1. $\forall x, y : \delta(x, y) \geq 0$
2. $\forall x, y : \delta(x, y) = \delta(x, y)$
3. $\forall x, y : \delta(x, y) = 0 \Leftrightarrow x = y$
4. $\forall x, y, z : \delta(x, y) + \delta(y, z) \geq \delta(x, z)$

This immediately excludes metrics like PARSEVAL (Black et al., 1991) and Leaf-Ancestor (Sampson and Babarczy, 2003), since they assume that the trees being compared are parses of the same sentence. Instead, we base our work on tree edit distance. The tree edit distance (TED) problem is defined analogously to the more familiar problem of string edit distance: what is the minimum number of edit operations required to transform one tree into the other? See Bille (2005) for a thorough introduction to the tree edit distance problem and other related problems. For this work, we used the algorithm of Zhang and Shasha (1989). Tree edit distance has previously been used in the TEDEVAL software (Tsarfaty et al., 2011; Tsarfaty et al., 2012) for parser evaluation agnostic to both annotation scheme and theoretical framework, but this by itself is still an

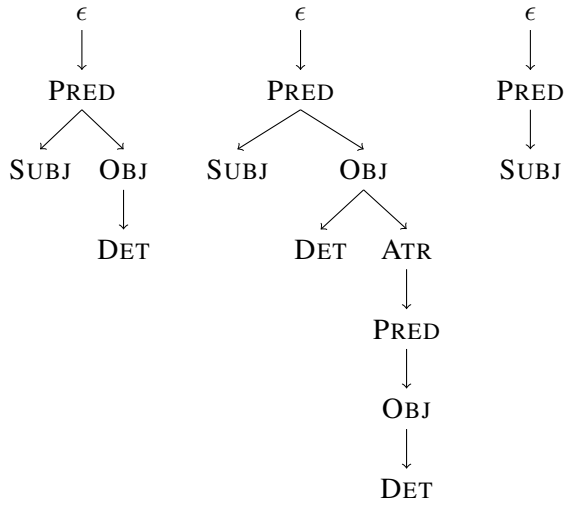


Figure 2: Three trees with distance zero using δ_{diff}

uncorrected accuracy measure and thus unsuitable for our purposes.[3]

When comparing syntactic trees, we only want to compare dependency relations or non-terminal categories. Therefore we remove the leaf nodes in the case of phrase structure trees, and in the case of dependency trees we compare trees whose edges are unlabelled and nodes are labelled with the dependency relation between that word and its head; the root node receives the label ϵ. An example of this latter transformation is shown in Figure 1.

We propose three different distance functions for the agreement computation: the unmodified tree edit distance function, denoted δ_{plain}, a second function $\delta_{diff}(x, y) = \mathrm{TED}(x, y) - \mathrm{abs}(|x| - |y|)$, the edit distance minus the difference in length between the two sentences, and finally $\delta_{norm}(x, y) = {}^{\mathrm{TED}(x,y)}/_{|x|+|y|}$, the edit distance normalised to the range $[0, 1]$.[4]

The plain TED is the simplest in terms of parsimony assumptions, however it may overestimate the difference between sentences, we intuitively find to be syntactically similar. For example the only difference between the two leftmost trees in Figure 2 is a modifier, but δ_{plain} gives them distance 4 and δ_{diff} 0. On the other hand, δ_{diff} might underestimate some distances as well; for example the leftmost and rightmost trees also have distance zero using δ_{diff}, despite our syntactic intuition that the difference between a transitive and an intransitive should be taken account of.

The third distance function, δ_{norm}, takes into account a slightly different concern; namely that when comparing a long sentence and a short sentence, the distance has to be quite large simply to account for the difference in number of nodes, unlike comparing two short or two long sentences. Normalising to the range $[0, 1]$ puts all pairs on an equal footing.

However, we cannot *a priori* say which of the three functions is the optimal choice of distance functions. The different functions have different properties, and different advantages and drawbacks, and the nature of their strengths and weaknesses differ. We will therefore perform a number of synthetic experiments to investigate their properties in a controlled environment, before applying them to real-world data.

3 Synthetic experiments

In the previous section, we proposed three different agreement metrics α_{plain}, α_{diff} and α_{norm}, each involving different trade-offs. Deciding which of these metrics is the best one for our purposes of judging the consistency of syntactic annotation poses a bit of a conundrum. We could at this point apply our metrics to various real corpora and compare the results, but since the consistency of the corpora is unknown, it's impossible to say whether the best metric is the one resulting in the highest scores, the lowest scores or somewhere in the middle. To properly settle this question, we first performed a number of synthetic experiments to gauge how the different metrics respond to disagreement.

The general approach we take is based on that used by Mathet et al. (2012), adapted to dependency trees. An already annotated corpus, in our case 100 randomly selected sentences from the Norwegian Dependency Treebank (Solberg et al., 2014), are taken as correct and then permuted to produce "annotations" of different quality. For dependency trees, the input corpus is permuted as follows:

1. Each token has a probability $p_{relabel}$ of being assigned a different label uniformly at random from the set of labels used in the corpus.

[3] While it is quite different from other parser evaluation schemes, TEDEVAL does not correct for chance agreement and is thus an uncorrected metric. It could of course form the basis for a corrected metric, given a suitable measure of expected agreement.

[4] We can easily show that $|x| + |y|$ is an upper bound on the TED, corresponding to deleting all nodes in the source tree and inserting all the nodes in the target.

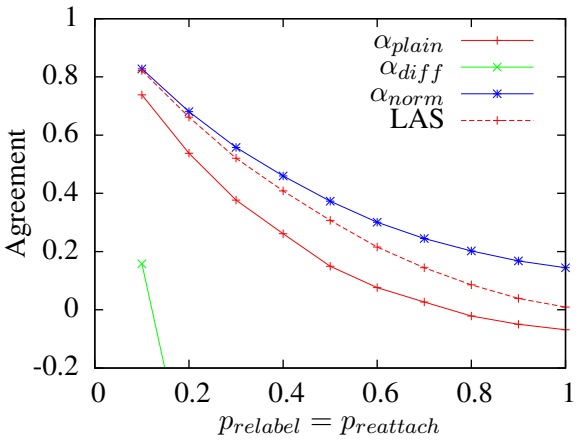


Figure 3: Mean agreement over ten runs

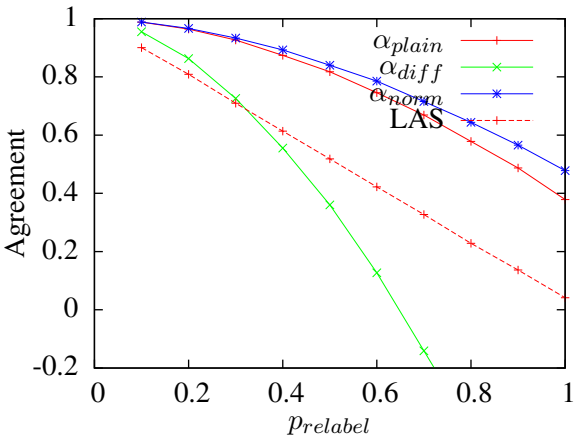


Figure 4: Mean agreement over ten runs, $p_{reattach} = 0$

2. Each token has a probability $p_{reattach}$ of being assigned a new head uniformly at random from the set of tokens not dominated by the token.

The second permutation process is dependent on the order the tokens are processed, and we consider the tokens in the post-order[5] as dictated by the original tree. This way tokens close to the root have a fair chance of having candidate heads if they are selected. A pre-order traversal would result in tokens close to the root having few options, and in particular if the root has a single child, that node has no possible new heads unless one of its children has been assigned the root as its new head first. For example in the trees in figure 2, assigning any other head than the root to the PRED nodes directly dominated by the root will result in invalid (cyclic and unconnected) dependency trees. Traversing the tokens in the linear order dictated by the sentence has similar issues for tokens close to the root and close to the start of the sentence.

For our first set of experiments, we set $p_{relabel} = p_{reattach}$ and evaluated the different agreement metrics for 10 evenly spaced p-values between 0.1 and 1.0. Initial exploration of the data showed that the mean follows the median very closely regardless of metric and perturbation level, and therefore we only report the mean scores across runs in this paper. The results of these experiments are shown in Figure 3, with the labelled attachment score[6] (LAS) for comparison.

The α_{diff} metric is clearly extremely sensitive to noise, with $p = 0.1$ yielding mean $\alpha_{diff} = 15.8\%$, while α_{norm} is more lenient than both LAS and α_{plain}, with mean $\alpha_{norm} = 14.5\%$ at $p = 1$, quite high compared to LAS $= 0.9\%$, $\alpha_{plain} = -6.8\%$ and $\alpha_{diff} = -246\%$. To further study the sensitivity of the metrics to the two kinds of noise, we performed an additional set of experiments, setting one $p = 0$ while varying the other over the same range as in the previous experiment, the results of which are shown in Figures 4 and 5.

The LAS curves are mostly unremarkable, with one exception: Mean LAS at $p_{reattach} = 1$ of Figure 5 is 23.9%, clearly much higher than we would expect if the trees were completely random. In comparison, mean LAS when only labels are perturbed is 4.1%, and since the sample space of trees of size n is clearly much larger than that of relabellings, a uniform random selection of tree would yield a LAS much closer to 0. This shows that our tree shuffling algorithm has a non-uniform distribution over the sample space.

While the behaviour of our alphas and LAS are relatively similar in Figure 3, Figures 4 and 5 show that they do in fact have important differences. Whereas LAS responds linearly to perturbation of both labels and structure, with its parabolic behaviour in Figure 3 being simply the product of these two linear responses, the α metrics respond differently to structural noise and label noise, with label disagreements being penalised less harshly

[5] That is, the child nodes of a node are all processed before the node itself. Nodes on the same level are traversed from left to right.

[6] The *de facto* standard parser evaluation metric in dependency parsing: the percentage of tokens that receive the correct head *and* dependency relation.

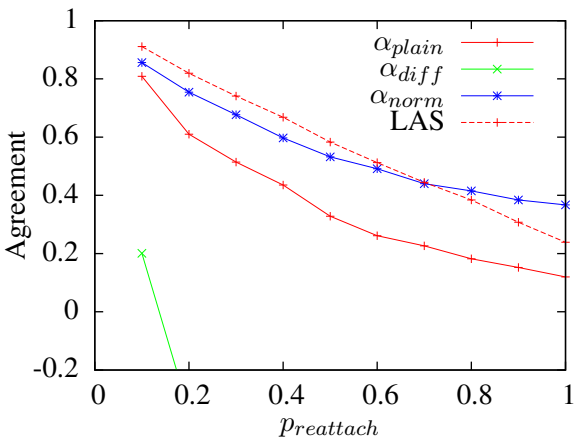


Figure 5: Mean agreement over ten runs, $p_{relabel} = 0$

than structural disagreements.

The reason for the strictness of the α_{diff} metric and the laxity of α_{norm} is the effects the modified distance functions have on the distribution of distances. The δ_{diff} function causes an extreme shift of the distances towards 0; more than 30% of the sentence pairs have distance 0, 1, or 2, which causes D_e^{diff} to be extremely low and thus gives disproportionally large weight to non-zero distances in D_o^{diff}. On the other hand δ_{norm} causes a rightward shift of the distances, which results in a high D_e^{norm} and thus individual disagreements having less weight.

4 Real-world corpora

Synthetic experiments do not always fully reflect real-world behaviour, however. Therefore we will also evaluate our metrics on real-world inter-annotator agreement data sets. In our evaluation, we will contrast labelled accuracy, the standard parser evaluation metric, and our three α metrics. In particular, we are interested in the correlation (or lack thereof) between LAS and the alphas, and whether the results of our synthetic experiments correspond well with the results on real-world IAA sets. Finally, we also evaluate the metric on both dependency and phrase structure data.

4.1 The corpora

We obtained[7] data from four different corpora. Three of the data sets are dependency treebanks (NDT, CDT, PCEDT) and one phrase structure treebank (SSD), and of the dependency treebanks the PCEDT contains semantic dependencies, while the other two have traditional syntactic dependencies. The number of annotators and sizes of the different data sets are summarised in Table 1.

Corpus	Sentences	Tokens
NDT 1[a]	130	1674
NDT 2[a]	110	1594
NDT 3[a]	150	1997
CDT (da)[a]	162	2394
CDT (en)[a]	264	5528
CDT (es)[b]	55	924
CDT (it)[c]	136	3057
PCEDT[d]	3531	61737
SSD[e]	96	1581

[a] 2 annotators
[b] 4 annotators, avg. 2.8 annotators/text (min. 2, max. 4)
[c] 3 annotators, avg. 2.7 annotators/text
[d] 11 annotators, avg. 2.5 annotators/text (min. 2, max. 6)
[e] 3 annotators, avg. 2.9 annotators/sent.

Table 1: Sizes of the different IAA corpora

NDT The Norwegian Dependency Treebank (Solberg et al., 2014) is a dependency treebank constructed at the National Library of Norway. The data studied in this work has previously been used by Skjærholt (2013) to study agreement, but using simple accuracy measures (UAS, LAS) rather than chance-corrected measures. The IAA data set is divided into three parts, corresponding to different parsers used to preprocess the data before annotation; what we term NDT 1 through 3 correspond to what Skjærholt (2013) labels Danish, Swedish and Norwegian, respectively.

CDT The Copenhagen Dependency Treebanks (Buch-Kromann et al., 2009; Buch-Kromann and Korzen, 2010) is a collection of parallel dependency treebanks, containing data from the Danish PAROLE corpus (Keson, 1998b; Keson, 1998a) in the original Danish and translated into English, Italian and Spanish.

PCEDT The Prague Czech-English Dependency Treebank 2.0 Hajič et al. (2012) is a parallel corpus of English and Czech, consisting of English data from the Wall Street Journal Section of the Penn Treebank (Marcus et al., 1993) and

[7] We contacted a number of treebank projects, among them the Penn Treebank and the Prague Dependency Treebank, but not all of them had data available.

Czech translations of the English data. The syntactic annotations are layered and consist of an analytical layer similar to the annotations in most other dependency treebanks, and a more semantic tectogrammatical layer.

Our data set consists of a common set of analytical annotations shared by all the annotators, and the tectogrammatical analyses built on top of this common foundation. A distinguishing feature of the tectogrammatical analyses, vis a vis the other treebanks we are using, is that semantically empty words only take part in the analytical annotation layer and nodes are inserted at the tectogrammatical layer to represent covert elements of the sentence not present in the surface syntax of the analytical layer. Thus, inserting and deleting nodes is a central part of the task of tectogrammatical annotation, unlike the more surface-oriented annotation of our other treebanks, where the tokenisation is fixed before the text is annotated.

SSD The Star-Sem Data is a portion of the dataset released for the *SEM 2012 shared task (Morante and Blanco, 2012), parsed using the LinGO English Resource Grammar (ERG, Flickinger, 2000) and the resulting parse forest disambiguated based on discriminants. The ERG is an HPSG-based grammar, and as such its analyses are attribute-value matrices (AVMs); an AVM is not a tree but a directed acyclic graph however, and for this reason we compute agreement not on the AVM but the so-called *derivation tree*. This tree describes the types of the lexical items in the sentence and the bottom-up ordering of rule applications used to produce the final analysis and can be handled by our procedure like any phrase-structure tree.

4.2 Agreement results

To evaluate our corpora, we compute the three α variants described in the previous two sections, and compare these with labelled accuracy scores.

When there are more than two annotators, we generalise the metric to be the average pairwise LAS for each sentence, weighted by the length of the sentence. Let $\mathrm{LAS}(t_1, t_2)$ be the fraction of tokens with identical head and label in the trees t_1 and t_2; the pairwise labelled accuracy $\mathrm{LAS}_p(X)$ of a set of annotations X as described in section 1.2 is:

$$\mathrm{LAS}_p(X) = \frac{1}{\sum_i |x_{i1}|} \sum \frac{|x_{i1}|\Lambda(X_i)}{|X_i|(|X_i|-1)/2} \quad (3)$$

$$\Lambda(X_i) = \sum_{c=1}^{|C|} \sum_{c'=c+1}^{|C|} \mathrm{LAS}(x_{ic}, x_{ic'})$$

This is equivalent to the traditional metric in the case where there are only two annotators.

As our uncorrected metric for comparing two phrase structure trees we do not use the traditional bracket F_1 as it does not generalise well to more than two annotators, but rather Jaccard similarity. The Jaccard similarity of two sets A and B is the ratio of the size of their intersection to the size of their union: $J(A, B) = |A \cap B| / |A \cup B|$, and we use the Jaccard similarity of the sets of labelled bracketings of two trees as our uncorrected measure. To compute the similarity for a complete set of annotations we use the mean pairwise Jaccard similarity weighted by sentence length; that is, the same procedure as in 3, but using Jaccard similarity rather than LAS.

Since LAS assumes that both of the sentences compared have identical sets of tokens, we had to exclude a number of sentences from the LAS computation in the cases of the English and Italian CDT corpora, and especially the PCEDT. The large number of sentences excluded in the PCEDT is due to the fact that in the tectogrammatical analysis of the PCEDT, inserting and deleting nodes is an important part of the annotation task.

Looking at the results in Table 2, we observe

Corpus	α_{plain}	α_{diff}	α_{norm}	LAS
NDT 1	98.4	93.0	98.8	94.0
NDT 2	98.9	95.0	99.1	94.4
NDT 3	97.9	91.2	98.7	95.3
CDT (da)	95.7	84.7	96.2	90.4
CDT (en)	92.4	70.7	95.0	88.4
CDT (es)	86.6	48.8	85.8	78.9[a]
CDT (it)	84.5	55.7	89.2	81.3[b]
PCEDT	95.9	89.9	96.5	68.0[c]
SSD	99.1	98.6	99.3	87.9[d]

[a] 2 sentences ignored
[b] 15 sentences ignored
[c] 1178 sentences ignored
[d] Mean pairwise Jaccard similarity

Table 2: Agreement scores on real-world corpora

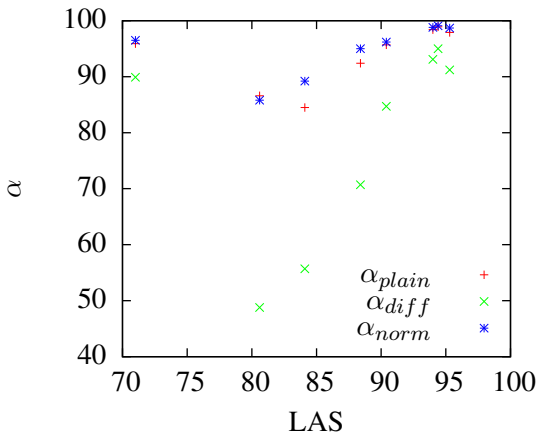


Figure 6: Correlation of LAS with α

two things. Most obvious, is the extremely large gap between the LAS and α metrics for the PCEDT data. However, there is a more subtle point; the orderings of the corpora by the different metrics are not the same. LAS order the corpora NDT 3, 2, 1, CDT da, en, it, es, PCEDT, whereas α_{diff} and α_{norm} gives the order NDT 2, 1, 3, PCEDT, CDT da, en, it, es, and α_{plain} gives the same order as the other alphas but with CDT es and it changing places. Furthermore, as the scatterplot in Figure 6 shows, there is a clear correlation between the α metrics and LAS, if we disregard the PCEDT results.

The reason the PCEDT gets such low LAS is essentially the same as the reason many sentences had to be excluded from the computation in the first place; since inserting and deleting nodes is an integral part of the tectogrammatical annotation task, the assumption implicit in the LAS computation that sentences with the same number of nodes have the same nodes in the same order is obviously false, resulting in a very low LAS.

The corpus that scores the highest for all three metrics is the SSD corpus; the reason for this is uncertain, as our corpora differ along many dimensions, but the fact that the annotation was done by professional linguists who are very familiar with the grammar used to parse the data is likely a contributing factor. The difference between the α metrics and the Jaccard similarity is larger than the difference between α and LAS for our dependency corpora, however the two similarity metrics are not comparable, and it is well known that for phrase structures single disagreements such as a PP-attachment disagreement can result in multiple disagreeing bracketings.

5 Conclusion

The most important conclusion we draw from this work is the most appropriate agreement metric for syntactic annotation. First of all, we disqualify the LAS metric, primarily due to the methodological inadequacies of using an uncorrected measure. While our experiments did not reveal any serious shortcomings (unlike those of Mathet et al., 2012 who in the case of categorisation showed that for large p the uncorrected measure can be *increasing*), the methodological problems of uncorrected metrics makes us wary of LAS as an agreement metric. Next, of the three α metrics, α_{plain} is clearly the best; α_{diff} is extremely sensitive to even moderate amounts of disagreement, while α_{norm} is overly lenient.

Looking solely at Figure 3, one might be led to believe that LAS and α_{plain} are interchangeable, but this is not the case. As shown by Figures 4 and 5, the paraboloid shape of the LAS curve in Figure 3 is simply the combination of the metric's linear responses to both label and structural perturbations. The behaviour of α on the other hand is more complex, with structural noise being penalised harder than perturbations of the labels. Thus, the similarity of LAS and α_{plain} is not at all assured when the amounts of structural and labelling disagreements differ. Additionally, we consider this imbalanced weighting of structural and labelling disagreements a benefit, as structure is the larger part of syntactic annotation compared to the labelling of the dependencies/bracketings. Finally our experiments show that α is a single metric that is applicable to both dependencies and phrase structure trees.

Furthermore, α metrics are far more flexible than simple accuracy metrics. The use of a distance function to define the metric means that more fine-grained distinctions can be made; for example, if the set of labels on the structures is highly structured, partial credit can be given for differing annotations that overlap. For example, if different types of adverbials (temporal, negation, etc.) receive different relations, as is the case in the Swedish Talbanken05 (Nivre et al., 2006) corpus, confusion of different adverbial types can be given less weight than confusion between subject and object. The α-based metrics are also far easier to apply to a more complex annotation task such

as the tectogrammatical annotation of the PCEDT. In this task inserting and deleting nodes is an integral part of the annotation, and if two annotators insert or delete different nodes the all-or-nothing requirement of identical yield of the LAS metric makes it impossible as an evaluation metric in this setting.

5.1 Future work

In future work, we would like to investigate the use of other distance functions, in particular the use of approximate tree edit distance functions such as the pq-gram algorithm (Augsten et al., 2005). For large data sets such as the PCEDT set used in this work, computing α with tree edit distance as the distance measure can take a very long time.[8] This is due to the fact that α requires $O(n^2)$ comparisons to be made, each of which is $O(n^2)$ using our current approach. The problem of directed graph edit distance is NP-hard, which means that to apply our method to HPSG analyses directly approximate algorithms are a requirement.

Another avenue for future work is improved synthetic experiments. As we saw, our implementation of tree perturbations was biased towards trees similar in shape to the source tree, and an improved permutation algorithm may reveal interesting edge-case behaviour in the metrics. A method for perturbing phrase structure trees would also be interesting, as this would allow us to repeat the synthetic experiments performed here using phrase structure corpora to compare the behaviour of the metrics on the two types of corpus.

Finally, annotator modelling techniques like that presented in Passonneau and Carpenter (2013) has obvious advantages over agreement coefficients such as α. These techniques are interpreted more easily than agreement coefficients, and they allow us to assess the quality of individual annotators, a crucial property in crowd-sourcing settings and something that's impossible using agreement coefficients.

Acknowledgements

I would like to thank Jan Štěpánek at Charles University for data from the PCEDT and help with the conversion process, the CDT project for publishing their agreement data, Per Erik Solberg at the Norwegian National Library for data from the NDT, and Emily Bender at the University of Washington for the SSD data.

[8]The Python implementation used in this work, using NumPy and the PyPy compiler, took seven and a half hours compute a single α for the PCEDT data set on an Intel Core i7 2.9 GHz computer. The program is single-threaded.

References

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Nikolaus Augsten, Böhlen Michael, and Johann Gamper. 2005. Approximate Matching of Hierarchical Data Using pq-Grams. In *Proceedings of the 31st international conference on Very large data bases*, pages 301–312, Trondheim. VLDB Endowment.

E. M. Bennett, R. Alpert, and A. C. Goldstein. 1954. Communications Through Limited-Response Questioning. *Public Opinion Quarterly*, 18(3):303–308.

Riyaz Ahmad Bhat and Dipti Misri Sharma. 2012. A Dependency Treebank of Urdu and its Evaluation. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 157–165, Jeju. Association for Computational Linguistics.

Philip Bille. 2005. A survey on tree edit distance and related problems. *Theoretical Computer Science*, 337(1-3):217–239, June.

Ezra Black, Steven Abney, Dan Flickinger, Claudia Gdaniec, Ralph Grishman, Philip Harrison, Donald Hindle, Robert Ingria, Frederick Jelinek, Judith Klavans, Mark Liberman, Mitchell P. Marcus, Salim Roukos, Beatrice Santorini, and Tomek Strzalkowski. 1991. A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars. In *Proceedings of the workshop on Speech and Natural Language*, pages 306–311, Pacific Grove, USA.

Sabine Brants and Silvia Hansen. 2002. Developments in the TIGER Annotation Scheme and their Realization in the Corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 1643–1649.

Thorsten Brants. 2000. Inter-Annotator Agreement for a German Newspaper Corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation*.

Matthias Buch-Kromann and Iørn Korzen. 2010. The unified annotation of syntax and discourse in the Copenhagen Dependency Treebanks. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 127–131, Uppsala. Association for Computational Linguistics.

Matthias Buch-Kromann, Iørn Korzen, and Henrik Høeg Müller. 2009. Uncovering the 'lost' structure of translations with parallel treebanks. In Fabio Alves, Susanne Göpferich, and Inger Mees, editors,

Methodology, Technology and Innovation in Translation Process Research, pages 199–224. Samfundslitteratur, Frederiksberg.

Jean Carletta. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249–254.

Montserrat Civit, Alicia Ageno, Borja Navarro, Núria Bufí, and M. Antònia Martí. 2003. Qualitative and Quantitative Analysis of Annotators' Agreement in the Development of. In Joakim Nivre and Erhard Hinrichs, editors, *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*, pages 21–32, Växjö. Växjö University Press.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, April.

Sérgio Ricardo de Castro. 2011. *Developing reliability metrics and validation tools for datasets with deep linguistic information*. Master's thesis, Universidade de Lisboa.

Jason M. Eisner. 1996. Three New Probabilistic Models for Dependency Parsing: An Exploration. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 340–345, Stroudsburg. Association for Computational Linguistics.

Dan Flickinger. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6(1):15–28, March.

Chris Fournier. 2013. Evaluating Text Segmentation using Boundary Edit Distance. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1702–1712, Sofia. Association for Computational Linguistics.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Prague Czech-English Dependency Treebank 2.0.

Jan Hajič. 2004. Complex Corpus Annotation: The Prague Dependency Treebank. Jazykovedný ústav Ľ. Štúra, SAV.

Britt Keson. 1998a. The Danish Morphosyntactically Tagged PAROLE Corpus. Technical report, Danish Society for Literature and Language, Copenhagen.

Britt Keson. 1998b. Vejledning til det danske morfosyntaktisk taggede PAROLE-korpus. Technical report, Danish Society for Literature and Language, Copenhagen.

Klaus Krippendorff. 1970. Estimating the Reliability, Systematic Error and Random Error of Interval Data. *Educational and Psychological Measurement*, 30(1):61–70, April.

Klaus Krippendorff. 2004. *Content Analysis: An introduction to its methodology*. Sage Publications, Thousand Oaks, 2nd edition.

Mohamed Maamouri, Ann Bies, and Seth Kulick. 2008. Enhancing the Arabic Treebank : A Collaborative Effort toward New Annotation Guidelines. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 3192–3196. European Language Resources Association.

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

Yann Mathet, Antoine Widlöcher, Karën Fort, Claire François, Olivier Galibert, Cyril Grouin, Juliette Kahn, Sophie Rosset, and Pierre Zweigenbaum. 2012. Manual Corpus Annotation : Giving Meaning to the Evaluation Metrics. In *Proceedings of COLING 2012*, pages 809–818, Mumbai.

Marie Mikulová and Jan Štěpánek. 2010. Ways of Evaluation of the Annotators in Building the Prague Czech-English Dependency Treebank. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 1836–1839, Valletta. European Language Resources Association.

Roser Morante and Eduardo Blanco. 2012. *SEM 2012 Shared Task : Resolving the Scope and Focus of Negation. In *The First Joint Conference on Lexical and Computational Semantics*, pages 265–274, Montreal. Association for Computational Linguistics.

Joakim Nivre, Jens Nilsson, and Johan Hall. 2006. Talbanken05 : A Swedish Treebank with Phrase Structure and Dependency Annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*.

Rebecca J. Passonneau and Bob Carpenter. 2013. The Benefits of a Model of Annotation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 187–195, Sofia. Association for Computational Linguistics.

Rebecca J. Passonneau. 2006. Measuring Agreement on Set-valued Items (MASI) for Semantic and Pragmatic Annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pages 831–836.

Marwa Ragheb and Markus Dickinson. 2013. Inter-annotator Agreement for Dependency Annotation of Learner Language. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 169–179, Atlanta. Association for Computational Linguistics.

Geoffrey Sampson and Anna Babarczy. 2003. A test of the leaf-ancestor metric for parse accuracy. *Natural Language Engineering*, 9(4):365–380, December.

William A. Scott. 1955. Reliability of Content Analysis: The Case of Nominal Scale Coding. *Public Opinion Quarterly*, 19(3):321–325, January.

Arne Skjærholt. 2013. Influence of preprocessing on dependency syntax annotation: speed and agreement. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 28–32, Sofia. Association for Computational Linguistics.

Per Erik Solberg, Arne Skjærholt, Lilja Øvrelid, Kristin Hagen, and Janne Bondi Johannesen. 2014. The Norwegian Dependency Treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik. European Language Resources Association.

Reut Tsarfaty, Joakim Nivre, and Evelina Andersson. 2011. Evaluating Dependency Parsing: Robust and Heuristics-Free Cross-Annotation Evaluation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 385–396, Edinburgh. Association for Computational Linguistics.

Reut Tsarfaty, Joakim Nivre, and Evelina Andersson. 2012. Cross-Framework Evaluation for Statistical Parsing. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 44–54, Avignon. Association for Computational Linguistics.

Kaizhong Zhang and Dennis Shasha. 1989. Simple Fast Algorithms for the Editing Distance between Trees and Related Problems. *SIAM Journal on Computing*, 18(6):1245–1262, December.