# Linguistically debatable or just plain wrong?

**Barbara Plank, Dirk Hovy and Anders Søgaard**
Center for Language Technology
University of Copenhagen, Denmark
Njalsgade 140, DK-2300 Copenhagen S
bplank@cst.dk,dirk@cst.dk,soegaard@hum.ku.dk

## Abstract

In linguistic annotation projects, we typically develop annotation guidelines to minimize disagreement. However, in this position paper we question whether we should actually limit the disagreements between annotators, rather than embracing them. We present an empirical analysis of part-of-speech annotated data sets that suggests that disagreements are systematic across domains and to a certain extend also across languages. This points to an underlying ambiguity rather than random errors. Moreover, a quantitative analysis of tag confusions reveals that the majority of disagreements are due to linguistically debatable cases rather than annotation errors. Specifically, we show that even in the absence of annotation guidelines only 2% of annotator choices are linguistically unmotivated.

## 1 Introduction

In NLP, we often model annotation as if it reflected a single ground truth that was guided by an underlying linguistic theory. If this was true, the specific theory should be learnable from the annotated data. However, it is well known that there are linguistically *hard cases* (Zeman, 2010), where no theory provides a clear answer, so annotation schemes commit to more or less arbitrary decisions. For example, in parsing auxiliary verbs may head main verbs, or vice versa, and in part-of-speech (POS) tagging, possessive pronouns may belong to the category of determiners or the category of pronouns. This position paper argues that annotation projects should embrace these hard cases rather than pretend they can be unambiguously resolved. Instead of using overly specific annotation guidelines, designed to minimize inter-annotator disagreement (Duffield et al., 2007), and adjudicating between annotators of different opinions, we should embrace systematic inter-annotator disagreements. To motivate this, we present an empirical analysis showing

1. that certain inter-annotator disagreements are systematic, and
2. that actual errors are in fact so infrequent as to be negligible, even when linguists annotate without guidelines.

The empirical analysis presented below relies on text corpora annotated with syntactic categories or parts-of-speech (POS). POS is part of most linguistic theories, but nevertheless, there are still many linguistic constructions – even very frequent ones – whose POS analysis is widely debated. The following sentences exemplify some of these hard cases that annotators frequently disagree on. Note that we do not claim that both analyses in each of these cases (1–3) are equally good, but that there is some linguistic motivation for either analysis in each case.

| | | | | |
|---|---|---|---|---|
| (1) | Noam | goes | **out** | tonight |
| | NOUN | VERB | ADP/PRT | ADV/NOUN |
| (2) | Noam | likes | **social** | media |
| | NOUN | VERB | ADJ/NOUN | NOUN |
| (3) | Noam | likes | **his** | car |
| | NOUN | VERB | DET/PRON | NOUN |

To substantiate our claims, we first compare the distribution of inter-annotator disagreements across domains and languages, showing that most disagreements are systematic (Section 2). This suggests that most annotation differences derive from hard cases, rather than random errors.

We then collect a corpus of such disagreements and have experts mark which ones are due to actual annotation *errors*, and which ones reflect linguistically hard cases (Section 3). The results show that the majority of disagreements are due



*Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 507–511,
Baltimore, Maryland, USA, June 23-25 2014. 

to hard cases, and only about 20% of conflicting annotations are actual errors. This suggests that inter-annotator agreement scores often hide the fact that the vast majority of annotations are actually linguistically motivated. In our case, less than 2% of the overall annotations are linguistically unmotivated.

Finally, in Section 4, we present an experiment trying to learn a model to distinguish between hard cases and annotation errors.

## 2 Annotator disagreements across domains and languages

In this study, we had between 2-10 individual annotators with degrees in linguistics annotate different kinds of English text with POS tags, e.g., newswire text (PTB WSJ Section 00), transcripts of spoken language (from a database containing transcripts of conversations, Talkbank[1]), as well as Twitter posts. Annotators were specifically *not* presented with guidelines that would help them resolve hard cases. Moreover, we compare professional annotation to that of lay people. We instructed annotators to use the 12 universal POS tags of Petrov et al. (2012). We did so in order to make comparison between existing data sets possible. Moreover, this allows us to focus on really hard cases, as any debatable case in the coarse-grained tag set is necessarily also part of the finer-grained tag set.[2] For each domain, we collected exactly 500 doubly-annotated sentences/tweets. Besides these English data sets, we also obtained doubly-annotated POS data from the French Social Media Bank project (Seddah et al., 2012).[3] All data sets, except the French one, are publicly available at `http://lowlands.ku.dk/`.

We present disagreements as Hinton diagrams in Figure 1a–c. Note that the spoken language data does not include punctuation. The correlations between the disagreements are highly significant, with Spearman coefficients ranging from 0.644 (spoken and WSJ) to 0.869 (spoken and Twitter). Kendall's $\tau$ ranges from 0.498 (Twitter and WSJ) to 0.659 (spoken and Twitter). All diagrams have a vaguely "dagger"-like shape, with the blade going down the diagonal from top left to bottom right, and a slightly curved "hilt" across the counter-diagonal, ending in the more pronounced ADP/PRT confusion cells.

Disagreements are very similar across all three domains. In particular, adpositions (ADP) are confused with particles (PRT) (as in the case of "*get* out"); adjectives (ADJ) are confused with nouns (as in "*stone* lion"); pronouns (PRON) are confused with determiners (DET) ("*my* house"); numerals are confused with adjectives, determiners, and nouns ("2*nd* time"); and adjectives are confused with adverbs (ADV) ("see you *later*"). In Twitter, the X category is often confused with punctuations, e.g., when annotating punctuation acting as discourse continuation marker.

Our analyses show that a) experts disagree on the known hard cases when freely annotating text, and b) that these disagreements are the same across text types. More surprisingly, though, we also find that, as discussed next, c) roughly the same disagreements are also observed when comparing the linguistic intuitions of lay people.

More specifically, we had lay annotators on the crowdsourcing platform Crowdflower re-annotate the training section of Gimpel et al. (2011). They collected five annotations per word. Only annotators that had answered correctly on 4 gold items (randomly chosen from a set of 20 gold items provided by the authors) were allowed to submit annotations. In total, 177 individual annotators supplied answers. We paid annotators a reward of \$0.05 for 10 items. The full data set contains 14,619 items and is described in further detail in Hovy et al. (2014). Annotators were satisfied with the task (4.5 on a scale from 1 to 5) and felt that instructions were clear (4.4/5), and the pay reasonable (4.1/5). The crowdsourced annotations aggregated using majority voting agree with the expert annotations in 79.54% of the cases. If we pre-filter the data via Wiktionary and use an item-response model (Hovy et al., 2013) rather than majority voting, the agreement rises to 80.58%.

Figure 2 presents the Hinton diagram of the disagreements of lay people. Disagreements are very similar to the disagreements between expert annotators, especially on Twitter data (Figure 1b).

[1] `http://talkbank.org/`

[2] Experiments with variation $n$-grams on WSJ (Dickinson and Meurers, 2003) and the French data lead us to estimate that the fine-to-coarse mapping of POS tags disregards about 20% of observed tag-pair confusion types, most of which relate to fine-grained verb and noun distinctions, e.g. past participle versus past in "[..] criminal lawyers speculated/VBD vs. VBN that [..]".

[3] We mapped POS tags into the universal POS tags using the mappings available here: `https://code.google.com/p/universal-pos-tags/`

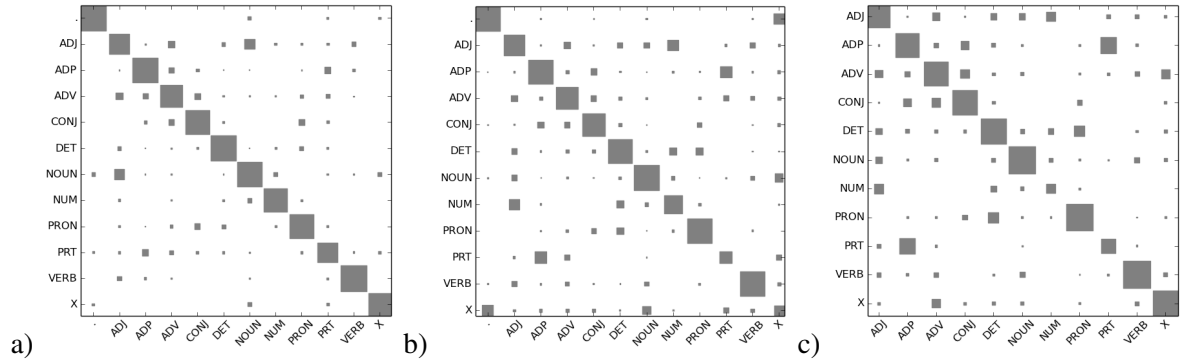


Figure 1: Hinton diagrams of inter-annotator disagreement on (a) excerpt from WSJ (Marcus et al., 1993), (b) random Twitter sample, and (c) pre-transcribed spoken language excerpts from talkbank.org

One difference is that lay people do not confuse numerals very often, probably because they rely more on orthographic cues than on distributional evidence. The disagreements are still strongly correlated with the ones observed with expert annotators, but at a slightly lower coefficient (with a Spearman's $\rho$ of 0.493 and Kendall's $\tau$ of 0.366 for WSJ).

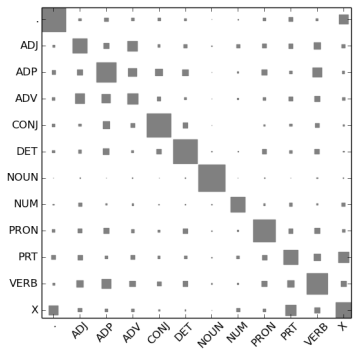


Figure 2: Disagreement between lay annotators

Lastly, we compare the disagreements of annotators on a French social media data set (Seddah et al., 2012), which we mapped to the universal POS tag set. Again, we see the familiar dagger shape. The Spearman coefficient with English Twitter is 0.288; Kendall's $\tau$ is 0.204. While the correlation is weaker across languages than across domains, it remains statistically significant ($p < 0.001$).

## 3 Hard cases and annotation errors

In the previous section, we demonstrated that some disagreements are consistent across domains and languages. We noted earlier, though, that disagreements can arise both from hard cases and from annotation errors. This can explain some

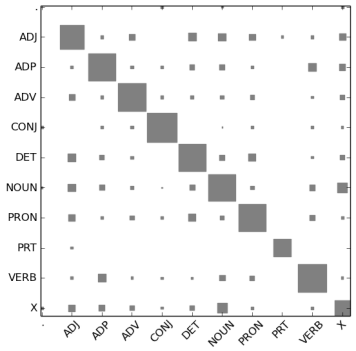


Figure 3: Disagreement on French social media

of the variation. In this section, we investigate what happens if we weed out obvious errors by detecting annotation inconsistencies across a corpus. The disagreements that remain are the truly hard cases.

We use a modified version of the a priori algorithm introduced in Dickinson and Meurers (2003) to identify annotation inconsistencies. It works by collecting "variation $n$-grams", i.e. the longest sequence of words ($n$-gram) in a corpus that has been observed with a token being tagged differently in another occurence of the same $n$-gram in the same corpus. The algorithm starts off by looking for unigrams and expands them until no longer $n$-grams are found.

For each variation $n$-gram that we found in WSJ-00, i.e, a word in various contexts and the possible tags associated with it, we present annotators with the cross product of contexts and tags. Essentially, we ask for a binary decision: Is the tag plausible for the given context?

We used 3 annotators with PhD degrees in linguistics. In total, our data set contains 880 items,

i.e. 440 annotated confusion tag pairs. The raw agreement was 86%. Figure 4 shows how truly hard cases are distributed over tag pairs (dark gray bars), as well as the proportion of confusions with respect to a given tag pair that are truly hard cases (light gray bars). The figure shows, for instance, that the variation $n$-gram regarding ADP-ADV is the second most frequent one (dark gray), and approximately 70% of ADP-ADV disagreements are linguistically hard cases (light gray). NOUN-PRON disagreements are always linguistically debatable cases, while they are less frequent.

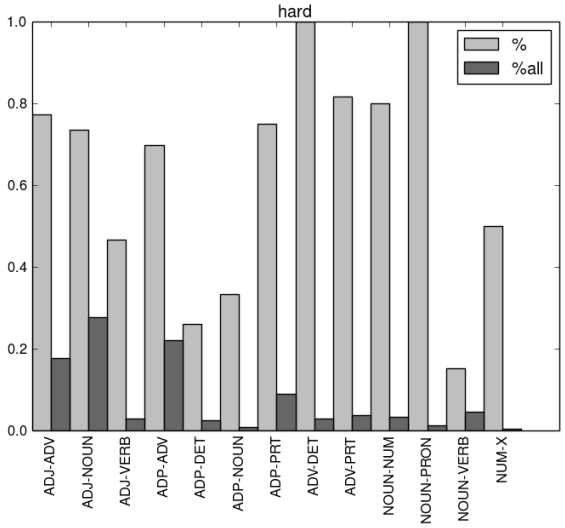


Figure 4: Relative frequency of hard cases

**A survey of hard cases.** To further test the idea of there being truly hard cases that probably cannot be resolved by linguistic theory, we presented nine linguistics faculty members with 10 of the above examples and asked them to pick their favorite analyses. In 8/10 cases, the faculty members disagreed on the right analysis.

## 4 Learning to detect annotation errors

In this section, we examine whether we can learn a classifier to distinguish between hard cases and annotation errors. In order to do so, we train a classifier on the annotated data set containing 440 tag-confusion pairs by relying only on surface form features. If we *balance* the data set and perform 3-fold cross-validation, a L2-regularized logistic regression (L2-LR) model achieves an $f_1$-score for detecting errors at 70% (cf. Table 1), which is above average, but not very impressive.

The two classes are apparently not easily separable using surface form features, as illustrated in the two-dimensional plot in Figure 5, obtained using PCA. The logistic regression decision boundary is plotted as a solid, black line. This is probably also why the nearest neighbor (NN) classifier does slightly better, but again, performance is rather low. While other features may reveal that the problem is in fact learnable, our initial experiments lead us to conclude that, given the low ratio of errors over truly hard cases, learning to detect errors is often not worthwhile.

| $f_1$ | HARD CASES | ERRORS |
|---|---|---|
| L2-LR | 73%(71-77) | 70%(65-75) |
| NN | 76%(76-77) | 71%(68-72) |

Table 1: Classification results

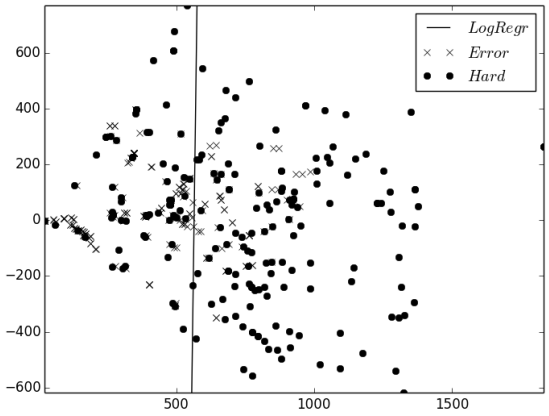


Figure 5: Hard cases and errors in 2d-PCA

## 5 Related work

Juergens (2014) presents work on detecting linguistically hard cases in the context of word sense annotations, e.g., cases where expert annotators will disagree, as well as differentiating between underspecified, overspecified and metaphoric cases. This work is similar to ours in spirit, but considers a very different task. While we also quantify the proportion of hard cases and present an analysis of these cases, we also show that disagreements are systematic.

Our work also relates to work on automatically correcting expert annotations for inconsistencies (Dickinson and Meurers, 2003). This work is very different in spirit from our work, but shares an interest in reconsidering expert annotations, and we made use of their mining algorithm here. There has also been recent work on adjudicat-

ing noisy crowdsourced annotations (Dawid and Skene, 1979; Smyth et al., 1995; Carpenter, 2008; Whitehill et al., 2009; Welinder et al., 2010; Yan et al., 2010; Raykar and Yu, 2012; Hovy et al., 2013). Again, their objective is orthogonal to ours, namely to collapse multiple annotations into a gold standard rather than embracing disagreements.

Finally, Plank et al. (2014) use small samples of doubly-annotated POS data to estimate annotator reliability and show how those metrics can be implemented in the loss function when inducing POS taggers to reflect confidence we can put in annotations. They show that not biasing the theory towards a single annotator but using a cost-sensitive learning scheme makes POS taggers more robust and more applicable for downstream tasks.

## 6 Conclusion

In this paper, we show that disagreements between professional or lay annotators are systematic and consistent across domains and some of them are systematic also across languages. In addition, we present an empirical analysis of POS annotations showing that the vast majority of inter-annotator disagreements are competing, but valid, linguistic interpretations. We propose to embrace such disagreements rather than using annotation guidelines to optimize inter-annotator agreement, which would bias our models in favor of some linguistic theory.

## Acknowledgements

We would like to thank the anonymous reviewers for their feedback, as well as Djamé Seddah for the French data. This research is funded by the ERC Starting Grant LOWLANDS No. 313695.

## References

Bob Carpenter. 2008. Multilevel Bayesian models of categorical data annotation. Technical report, LingPipe.

A. Philip Dawid and Allan M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, pages 20–28.

Markus Dickinson and Detmar Meurers. 2003. Detecting errors in part-of-speech annotation. In *EACL*.

Cecily Duffield, Jena Hwang, Susan Brown, Dmitriy Dligach, Sarah Vieweg, Jenny Davis, and Martha Palmer. 2007. Criteria for the manual grouping of verb senses. In *LAW*.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. In *ACL*.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *NAACL*.

Dirk Hovy, Barbara Plank, and Anders Søgaard. 2014. Experiments with crowdsourced re-annotation of a POS tagging data set. In *ACL*.

David Juergens. 2014. An analysis of ambiguity in word sense annotations. In *LREC*.

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

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *LREC*.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *EACL*.

Vikas C. Raykar and Shipeng Yu. 2012. Eliminating Spammers and Ranking Annotators for Crowdsourced Labeling Tasks. *Journal of Machine Learning Research*, 13:491–518.

Djamé Seddah, Benoit Sagot, Marie Candito, Virginie Mouilleron, and Vanessa Combet. 2012. The French Social Media Bank: a treebank of noisy user generated content. In *COLING*.

Padhraic Smyth, Usama Fayyad, Mike Burl, Pietro Perona, and Pierre Baldi. 1995. Inferring ground truth from subjective labelling of Venus images. In *NIPS*.

Peter Welinder, Steve Branson, Serge Belongie, and Pietro Perona. 2010. The multidimensional wisdom of crowds. In *NIPS*.

Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *NIPS*.

Yan Yan, Rómer Rosales, Glenn Fung, Mark Schmidt, Gerardo Hermosillo, Luca Bogoni, Linda Moy, and Jennifer Dy. 2010. Modeling annotator expertise: Learning when everybody knows a bit of something. In *AIStats*.

Daniel Zeman. 2010. Hard problems of tagset conversion. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources*.