Increasing the Recall of Corpus Annotation Error Detection

Adriane Boyd
Ohio State University
Dept. of Linguistics

Markus Dickinson
Indiana University
Dept. of Linguistics

Detmar Meurers
Ohio State University
Dept. of Linguistics

Abstract

While error detection approaches have been developed for various types of corpus annotation, so far only limited attention has been paid to the recall of those methods. We show how the recall of the so-called variation *n*-gram method can be increased by examining comparable part-of-speech tag sequences instead of the recurring strings themselves. To guide the search for erroneous annotation and to distinguish errors with high precision, we also develop new context reliability indicators.

1 Introduction and Motivation

Linguistically annotated corpora are widely used in computational linguistics for a variety of purposes (see, e.g., Abeillé 2003). Annotation errors can have a profound impact on training and evaluation (van Halteren et al. 2001; Květǒn and Oliva 2002; Dickinson and Meurers 2005b; Padro and Marquez 1998), often prompting researchers to develop work-around techniques to deal with noisy data for a particular task (e.g., Hogan 2007).

Previous research has addressed the question of detecting errors in treebanks (Dickinson and Meurers 2003b, 2005b; Ule and Simov 2004). One issue rarely addressed in this work, however, is the recall of the methods, i.e., the number of errors found, and how it could be increased. Naturally one would like the techniques to maximize the number of errors detected, without opening the floodgates to low error detection precision, requiring extra manual effort.

One method for detecting annotation errors is to find recurring data and compare their analyses in different corpus instances, using shared context as a heuristic to determine when they should be annotated identically. From this perspective, there are two ways to increase recall: one can either extend the set of what constitutes recurring data or one can relax the heuristic which narrows down the set of comparable strings to those which likely are errors.

In this paper, we explore the first option, relaxing the notion of what constitutes recurring data. Instead of relying on recurring identical words, as proposed in the

variation *n*-gram approach (Dickinson and Meurers 2003b), we propose to compare classes of words. The classes of words needed here should be distributionally similar, which makes part-of-speech (POS) tags a natural choice.

When generalizing the variation *n*-gram approach so that it detects errors by comparing the syntactic annotation of recurring strings of POS tags instead of the words themselves, the more general nature of the recurring units will also impact the precision of error detection. With a more general representation, we will need more constraints on the disambiguating contexts. We therefore explore what treebank information can be used to accurately predict the presence of an erroneous variation (as opposed to a legitimate ambiguity). Such an investigation of what local information is sufficient for disambiguating a string is also of interest beyond error detection (cf., e.g., Klein and Manning 2002) and can provide insights into the nature of the corpus annotation schemes.

2 Background

Our approach builds on the variation *n*-gram algorithm introduced in Dickinson and Meurers (2003a,b). The basic idea behind the approach is that a string occurring more than once in a corpus may occur annotated with different labels. Such *variation* in the annotation is caused by one of two reasons: i) *ambiguity*: there is a type of string with multiple possible labels and different corpus occurrences of that string realize the different legitimate options, or ii) *error*: the annotation of a string is inconsistent across comparable occurrences.

The more similar the context of a variation, the more likely the variation is an error. In Dickinson and Meurers (2003a), contexts are composed of words, and identity of the context is required. The term *variation n-gram* refers to an *n*-gram of words in a corpus that contains a string annotated differently in another occurrence of the same *n*-gram in the corpus. The string exhibiting the variation is referred to as the *variation nucleus*.

Dickinson and Meurers (2003a) explore this idea for part-of-speech annotation. For example, in the Wall Street Journal (WSJ) corpus, part of the Penn Treebank (Marcus et al. 1993), the string in (1) is a variation 12-gram since *off* is a variation nucleus that in one corpus occurrence is tagged as a preposition (IN), while in its other occurrence it is tagged as a particle (RP).[1]

(1) to ward off a hostile takeover attempt by two European shipping concerns

Once the variation *n*-grams for a corpus have been computed, heuristics are employed to classify the variations into errors and ambiguities. As Dickinson (2005) reports, the most effective heuristic takes into account the fact that natural languages favor the use of local dependencies over non-local ones: nuclei found at the fringe of an *n*-gram are more likely to be genuine ambiguities than those occurring with at least one word of surrounding context. This *non-fringe* heuristic is

[1] Here and in the following examples, the variation nucleus is shown in grey.

independent of a specific corpus, annotation scheme, or language and receives interesting support both from human category acquisition (cf. Mintz 2003) and from unsupervised grammar induction (cf. Klein and Manning 2002)

Applying the variation *n*-gram method to syntactic annotation, Dickinson and Meurers (2003b) decompose the variation *n*-gram detection for syntactic annotation into a series of passes with different nucleus sizes. This is needed to establish a one-to-one relation between a unit of data and a syntactic category annotation for comparison. Each pass detects the variation in the annotation of strings of a specific length. By performing such passes for strings from length 1 to the length of the longest constituent in the corpus, the approach ensures that all strings which are analyzed as a constituent somewhere in the corpus are compared to the annotation of all other occurrences of that string.

For example, a labeling error in the WSJ involves the nucleus *next Tuesday* as part of the variation 3-gram *maturity next Tuesday*, which appears three times in the WSJ. Twice it is labeled as a noun phrase (NP) and once as a prepositional phrase (PP). As an example for a bracketing error, consider the two WSJ occurrences of *last month* in Figure 1. To make them comparable to constituents, non-constituent occurrences are implicitly given the special label NIL (which essentially are the same as *distituents* in unsupervised grammar induction (Klein and Manning 2002)).

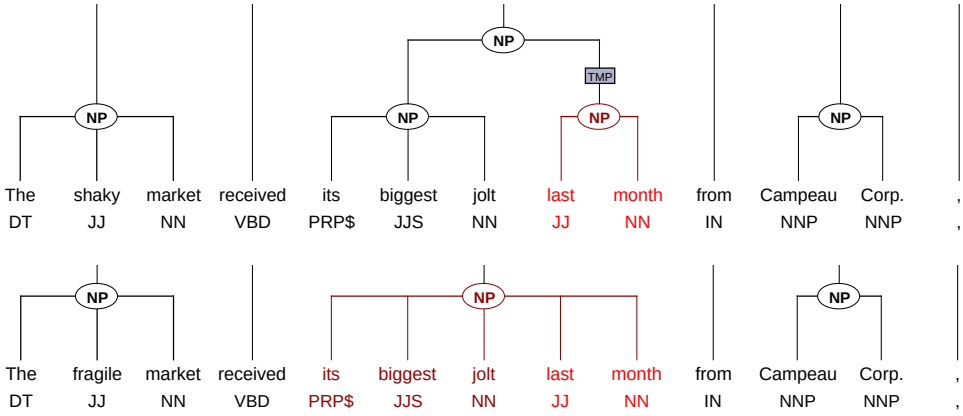


Figure 1: Bracketing differences in the analysis of "last month"

As reported in Dickinson (2005), this error detection method returns 36,859 variation nuclei for the WSJ. With the *shortest* non-fringe heuristic, where by shortest we mean that it contains exactly one word of context on each side, there are 3,769 shortest non-fringe variation nuclei with an estimate of 67% error detection precision. Removing cases in which an empty element (e.g., a trace of a long-distance dependency) is the variation nucleus results in an estimated precision of 75.86% for 3,619 variation nuclei, i.e., 2,745 annotation errors.

3 Approach

3.1 Using part-of-speech nuclei instead of words

While the method described in the previous section works very well, it misses more general errors by insisting on identical words in the variation nucleus. To increase the number of errors found, we can relax the requirements of what constitute comparable strings. For example, in (2)[2] there are comparable strings with variation in the noun phrase (NP)[3] structure, yet this case would not have been identified by examining the strings given the use of *an order* in one and *a contract* in the other.

(2) a. Boeing on Friday said 0 it received [$_{NP}$ an/DT order/NN] *ICH* from Martinair Holl

b. it received [$_{NP}$ a/DT contract/NN *ICH*] from Timken Co.

In order to successfully detect errors like this, we need to abstract the variation nuclei away from the concrete strings to something which still accurately encodes their distributional properties. To this end, we redefine the variation nucleus to consist of POS labels instead of the words. In (2), the nucleus would be DT NN for both examples. This gives us precisely the result we want: instead of using words to identify comparable nuclei, we now use more general distributional classes.

While this definition is simple in concept, there remains the task of determining appropriate disambiguating contexts for such POS nuclei.

3.2 Identifying more reliable contexts

We use the shortest non-fringe heuristic as the basis for our approach, given its relatively high precision with the original word nuclei. However, it is important to develop more reliable contextual indicators for POS nuclei, given that the shortest non-fringe context will often not be enough to reliably tell ambiguity from error (see section 4).

Consider example (3), with a variation nucleus of IN CD in the variation *n*-gram *began IN CD and*. With only one word of context on each side, we do not have enough shared context to tell us about the coordination attachment, i.e., whether it should be consistently internally attached within the PP or not.

(3) a. crippled * by a bitter , decade-long strike that *T*
began [$_{PP}$ in/IN 1967/CD] and cut circulation in half

b. its problems began [$_{PP}$ in/IN [$_{NP}$ 1987/CD and early 1988]] when its ...

[2]Here and in the following examples, the shortest non-fringe *n*-gram is underlined. Note also that the treebank contains empty elements inserted into the text by annotators. 0: null complementizer, *: null subject, *T*: trace, *U*: unit of measurement, *ICH*: interpret constituent here.

[3]In the text, the syntactic categories are shown in SMALL CAPITALS to typographically distinguish them from the terminal POS categories.

But what would constitute sufficient contextual information for determining whether a variation nucleus consisting of POS should be syntactically annotated in the same way? In the next sections, we discuss three new heuristics which can help identify annotation errors among the set of variation nuclei.[4]

3.2.1 Heuristic 1: Shared complete bracketing

First, consider variation between two nuclei which are constituents, i.e., a variation between two non-NIL labels (i.e., XP vs. YP). For variation between two such labels, the fact that both annotations agree on the bracketing makes it significantly more likely that the variation in the label is an error. This yields the first heuristic, to interpret shared complete bracketing as indicating comparable strings. Relatedly, in Dickinson and Meurers (2005b) we determined that variation between two categories for the same POS tag sequence often indicates an error.

For example, TO NN in the WSJ appears as a variation nucleus varying between VP and PP in the context between * (an empty element) and . (the period). The VP label is incorrect. It arises from an error in the POS annotation of the verb *hum*, which in the sentence *Kidder is gonna [= going to] hum* is annotated as NN, as shown in (4). Practically speaking, then, we can evaluate these cases simply with the shortest non-fringe context; i.e., no more context is needed.

(4) Kidder is " gon * [$_{VP}$ na/TO hum/NN] . "

3.2.2 Heuristic 2: Shared partial bracketing with additional word context

The remaining cases all include bracketing errors, i.e., variation between a constituent label and the non-constituent NIL label. In the case of the variation we saw in (3) above, for example, there is a legitimate attachment difference. But the one word of surrounding context is not sufficient to distinguish the two cases. On the one hand, one could introduce a special treatment for conjunctions and other words which turn out not to be reliable indicators of the local distributional environment. On the other hand, it arguably is preferable to determine more general, corpus-independent indicators for where attachment ambiguities may arise.

Looking at example (3) again, if we know that *in 1967* is a phrase which attaches to *began* to form a complete VP, then we can see that it is different from *in 1987*, which does not form a complete VP with *began*. But to determine this requires us to rely on the attachment decisions which may or may not be correctly annotated. While relying only on the data as such avoids any such complications, a practically useful, conservative extension is to rely on points of annotation agreement to help guide the search. This was, of course, also the insight we followed in the treatment of variation with complete shared bracketing in the previous section.

[4]While the shortest non-fringe context works well for identical word contexts, the attachment issue in (3) can, of course, also occur with word nuclei (even though more rarely). The heuristics developed in the following thus are relevant beyond the abstraction from word to POS nuclei.

For partially shared bracketing we can apply the same reasoning and conclude that for example (3), the shared left VP bracket is a good indicator for the two instances being comparable, whereas the difference in the right constituency bracket would require additional shared context on the right to make it likely for the variation in the annotation to be an error. Given that we want to remain as data-driven as possible, we implement this heuristic by requiring one extra word on the side without shared bracketing.

Consider example (5), a case of erroneous variation for the variation nucleus VBG JJ NNS. The constituent and the NIL string share a left (VP) bracket, but not a right one. Requiring an extra word of context on the side with no shared bracketing, i.e., the right side, shows us that these cases are indeed comparable.

(5) a. he stayed inside the Capitol * [$_{VP}$ [$_{VP}$ monitoring/VBG tax-and-budget/JJ talks/NNS] instead of flying to San Francisco . . .]

b. one of the first bids under new takeover rules aimed * at * [$_{VP}$ encouraging/VBG open/JJ bids/NNS instead of gradual accumulation of large stakes] .

3.2.3 Heuristic 3: Shared vertical context

The third heuristic makes use of shared structure involving the word context of a nucleus. Consider example (6), which contains erroneous variation for the nucleus RB JJR IN CD in the context of *to* on the left and % on the right. There is shared bracketing on the left, but not on the right.

(6) a. will be diluted * to [$_{NP}$ [$_{QP}$ slightly/RB less/JJR than/IN 50/CD] %] after . . .

b. will fall to [$_{NP}$ slightly/RB more/JJR than/IN 11/CD %] from slightly more than 14 % .

While the variation nucleus does not share the complete bracketing structure, the surrounding word context does. Once we add one word to the right, we have an NP in both cases. As mentioned in the discussion of the first heuristic in section 3.2.1, shared bracketing is a good indicator of comparable strings. Thus, we base our third heuristic on shared vertical context: a nucleus that is a part of a variation *n*-gram with shared structure is likely to be an annotation error. Given that we start with the shortest non-fringe variation nuclei, the shared vertical context (i.e., the encompassing constituent) can either consist of the nucleus with the context word to the left, the nucleus with the context word to the right, or the nucleus together with both words.

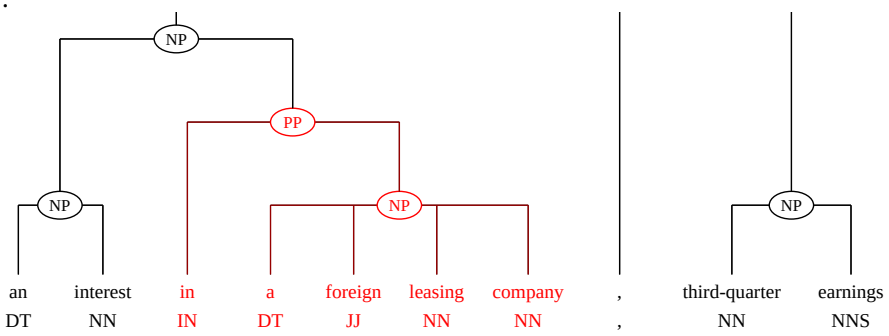
3.3 Defining shared brackets

In determining whether or not two strings are structurally parallel, we have to make two issues concrete. First, do we require the strings to share only the bracketing or

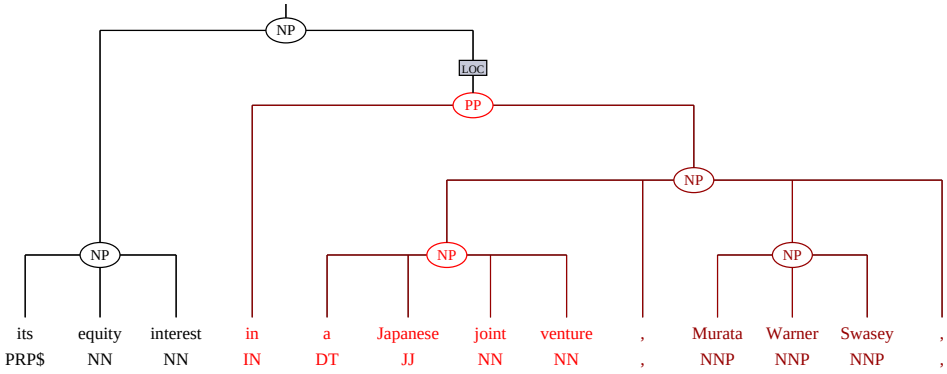
also the category label? To answer this, consider that, in addition to the fact that the label may be wrong, the issue of bracket sharing is an attachment issue not a labeling issue. Once we have determined the presence of a bracket, the issue of the most appropriate label is something which is determined mainly by the string within the bracketing (given the endocentric nature of most constituent structure). Thus, for the above heuristics we require only that the bracketing match, i.e., without requiring identical labels.

Secondly, for the heuristics to be effective in dealing with attachment ambiguities, we stipulate that in order for the bracketing to count for the heuristics, it must be for a constituent not contained within the nucleus. This prevents false positives in cases such as (7).

(7) a. excluding the addition to its reserves , certain tax benefits , and a one-time $ 16 million *U* gain on the sale of an interest [$_{PP}$ in/IN [$_{NP}$ a/DT foreign/JJ leasing/NN company/NN]] , third-quarter earnings were $ 75 million *U* .



b. Cross & Trecker is also selling its equity interest [$_{PP}$ in/IN [$_{NP}$ [$_{NP}$ a/DT Japanese/JJ joint/NN venture/NN] , Murata Warner Swasey]] , to Murata Machinery .



The nucleus IN DT JJ NN NN here varies between a PP constituent analysis, shown in (7a), and a non-constituent (i.e., NIL) analysis, shown in (7b). This is due to the fact that in the second case there is an appositive which attaches within the object

of the preposition (i.e., *Murata Warner Swasey*). We see that the nucleus IN DT JJ NN NN has brackets on both sides in the NIL case (7b): on the left side begins a PP constituent that continues beyond the nucleus, indicating that there is an attachment issue. On the right side (after *venture*) is a closing bracket for an NP. But this NP is completely within the nucleus, which is not informative about how this string is being used in context.

4 Results

After generalizing the nuclei from words to POS, we obtain 50,396 variation nuclei for the Wall Street Journal corpus. 17,251 of those occur in a shortest non-fringe context. After removing nuclei which are single null elements and thus problematic (cf. Dickinson and Meurers 2003b), we are left with 16,598 shortest non-fringe variation nuclei, significantly higher than the 3618 cases with word nuclei (Dickinson 2005).

In order to gauge the overall error detection performance with POS nuclei, we sampled 100 cases from the total set of 16,598 and examined these by hand. We found that 28 point to an error, three of which are POS errors. Note that the result is a significant improvement in terms of recall, with an estimated 4,647 cases[5] of the total set of 16,598 variations being errors, compared to 2,745 for word nuclei (Dickinson 2005).

Focusing our attention on the 28% error detection precision, we applied the three heuristics which in section 3.2 we determined to be potentially more reliable contextual indicators of errors. From the set of 16,598 nuclei, we calculated the cases identified by one of the three new heuristics, and we find 1,339 cases of shared complete bracketing and 1,273 cases of shared vertical context. For the shared partial bracketing cases, we extended the context by one word as described in section 3.2.2, resulting in 3,731 cases of shared partial bracketing with extended context. In total, then, there are 6,343 variation nuclei which are covered by the three new heuristics.

Starting with the 34 variation nuclei from the sample which are covered by the three heuristics introduced in section 3.2, we randomly selected more cases for a total of 33 of each kind. For each variation, we inspected whether the variation in the annotation is an error or a genuine ambiguity, obtaining the results in Figure 2.

shared complete bracketing	61% (20/33)
shared partial bracketing with extra word context	61% (20/33)
shared vertical context	85% (28/33)

Figure 2: Error detection precision using the three heuristics

[5]The 95% confidence interval (CI) for the point estimate of .28 is (0.1920, 0.3680), meaning that we estimate between 3,186 and 6,108 errors being detected.

Overall, then, we find a 68.69% (68/99) error detection precision using these heuristics. Based on this precision, we estimate that the 6,343 total cases contain 4,357 errors – an increase in recall of 59% over the estimated 2,745 errors detected using the word nuclei (Dickinson 2005).

We can see that our additions to the non-fringe heuristic, based on insights about the nature of syntactic annotation, approach the high precision of the original method with word nuclei, while at the same time significantly increasing the number of errors found.

In terms of evaluating the three new heuristics, it is relevant to report that for the 73 cases from the original sample[6] which are not covered by either of the three heuristics, we obtain only 8.22% precision. The three heuristics thus seem to cover most of the error cases included in the variation nuclei.

Insights into the annotation scheme As part of the error detection process, the method can also provide general insights into the annotation scheme and the nature of the evidence it relies on. For example, we discover that the category NX is one which relies on both semantic information and non-local information because it is only used when there are both shared and unshared modifiers in a coordinate structure. For the nucleus JJ NN CC NN NN, as shown in (8), even though the vertical context is the same (i.e., NP), this is a legitimate ambiguity because of scoping considerations. Distinguishing such cases is clearly beyond the bounds of a form-based error detection method.

(8) a. ... could run Pinkerton 's better than [$_{NP}$ an [$_{NX}$ [$_{NX}$ unfocused/JJ conglomerate/NN] or/CC [$_{NX}$ investment/NN banker/NN]]] .

b. Jacobs is [$_{NP}$ an international/JJ engineering/NN and/CC construction/NN concern/NN] .

4.1 Alternatives for increasing recall

As mentioned in the introduction, there are other ways to increase the recall of the error detection method. Instead of relaxing the definition of the nucleus, Dickinson (2005) and Dickinson and Meurers (2005a) experimented with using different, more general types of context, i.e., relaxing the heuristics. Specifically, those methods required the context surrounding the (word) nucleus to consist of identical POS tags instead of identical words. This allowed nuclei which appear next to low-frequency words to be grouped with other strings sharing the same POS labels. The technique works with the original set of variation nuclei, simply allowing more of them to be detected as errors, as opposed to redefining the nucleus to detect a different set of new cases.

[6]Once the additional word of context is added for the partial bracketing case, some of the original nuclei split into multiple cases given that they involve different word contexts. In our 100 word sample, seven variations split in this way, resulting in 107 (=73+34) cases in total.

Also ignoring null element nuclei, the previous generalization to contexts of identical POS tags on the same data gives 8,715 shortest non-fringe variation nuclei. From a sample of 99 cases, 52 pointed to an error (six of which were POS errors). The 95% confidence interval for the point estimate of .53 is (.4269, .6236), which means that the estimate is between 3,720 and 5,434 errors.

Since the methods work in different ways, they complement each other nicely: they extend the variation *n*-gram method in orthogonal (i.e., non-overlapping) ways. Together they increase recall by a significant amount. The problem with both methods, of course, is their precision, but the reliable context heuristics we have employed here indicate that recall can be increased in such a way as to maintain a high precision. Thus, in the future, one could even experiment with treating the corpus completely as a set of POS tags, and using the current heuristics to guide the search for erroneous variation.

Turning to a different, but related method, the immediate dominance (ID) variation method in Dickinson and Meurers (2005b) uses the right hand sides of context-free rules extracted from the treebank as units of comparison. While this opens up the space of possibilities of errors to be found, it is less of a data-driven method, relying solely on annotation to find errors. It overlaps with the current method when the RHS of a rule is a complete sequence of POS tags, linking up with our shared complete bracketing cases. The limitation of the ID variation method is that it mainly handles errors stemming from variation in labeling and not bracketing errors. Exploring more of the convergence between the two methods in the future, however, could lead to a better characterization of the types of errors to be found.

For example, one of the errors not caught by any of our more reliable context heuristics is for the nucleus VB in *to VB among*, with variation between NP and NIL, shown in (9). As we can see, the issue is really not about variation between NP and NIL or about context; it is about the non-endocentric property of an NP dominating an VB. Thus, we should be employing endocentricity-based error detection methods (Dickinson and Meurers 2005b) to find such cases.

(9) ... have returned to [$_{NP}$ favor/VB] among ...

One issue that needs to be mentioned is that, despite the strides made in increasing recall, there are some cases which will always come down to the exact lexical item used. Consider the variation trigram *remains JJ for* in (10).

(10) a. a virus that *T* [$_{VP}$ remains [$_{ADJP}$ active/JJ] [$_{PP}$ for a few days]]

b. remains [$_{ADJP}$ responsible/JJ for the individual policy services department]

This case depends upon particular adjective, in determining how the *for* phrase attaches. Expanding the context on the right will eliminate this case, but almost accidentally, as *a* and *the* could easily be the same. In the future, one could explore a mixture of word and part-of-speech nuclei; however, these cases also seem somewhat rare, so it is not clear how much would be gained in doing so.

5 Summary and Outlook

In this paper, we have discussed how one can increase the recall of an error detection method for syntactic annotation by relaxing the requirement of what constitutes comparable recurring data. More concretely, we generalized the so-called variation nuclei of the variation *n*-gram error detection approach to POS tags instead of relying on identical surface forms.

In order to handle the loss in error detection precision accompanying this generalization, we determined additional contextual heuristics for errors, such as shared vertical context or shared complete bracketing, essentially using annotator agreement to guide the search for disagreements. While the proposed additional heuristics in this paper are used as additional filter on the non-fringe variation nuclei, they arguably are more general in nature. As a way to relax the contextual requirements on recurring units, they can be applied on their own to increase the recall of any variation *n*-gram method. Items with shared vertical context or with shared complete bracketing do not require as much word context in common for them to be comparable in their annotation.

It seems attractive to further explore the space of reliable contexts and to generalize them where possible. While the shared vertical context and complete bracketing heuristics are already defined in general terms, there clearly is room for a more general motivation for others, such as the shared partial bracketing heuristic which currently stipulates an additional word of shared context.

Finally, given the task of defining nuclei as a mixture of POS and lexical items, it would also be fruitful to explore increased recall for annotations such as dependency annotation. Since the head drives the selection, we might be able to keep it as a word while backing off to the POS category for the dependent.

Acknowledgements This paper is based upon work supported by the National Science Foundation under Grant No. IIS-0623837. We would like to thank the anonymous TLT reviewers for their useful comments.

References

Abeillé, A. (ed.) (2003). *Treebanks: Building and using syntactically annotated corpora*. Dordrecht: Kluwer.

Dickinson, M. (2005). Error detection and correction in annotated corpora. Ph.D. thesis, The Ohio State University.

Dickinson, M. and W. D. Meurers (2003a). Detecting Errors in Part-of-Speech Annotation. In *Proceedings of EACL-03*. Budapest, pp. 107–114.

Dickinson, M. and W. D. Meurers (2003b). Detecting Inconsistencies in Treebanks. In *Proceedings of TLT-03*. Växjö, Sweden, pp. 45–56.

Dickinson, M. and W. D. Meurers (2005a). Detecting Errors in Discontinuous Structural Annotation. In *Proceedings of ACL-05*. pp. 322–329.

Dickinson, M. and W. D. Meurers (2005b). Prune Diseased Branches to Get Healthy Trees! How to Find Erroneous Local Trees in a Treebank and Why It Matters. In *Proceedings of TLT-05*. Barcelona, Spain.

Hogan, D. (2007). Coordinate Noun Phrase Disambiguation in a Generative Parsing Model. In *Proceedings of ACL-07*. Prague, Czech Republic, pp. 680–687.

Klein, D. and C. D. Manning (2002). A Generative Constituent-Context Model for Improved Grammar Induction. In *Proceedings of ACL-02*. Philadelphia, PA.

Květǒn, P. and K. Oliva (2002). Achieving an Almost Correct PoS-Tagged Corpus. In P. Sojka, I. Kopeček and K. Pala (eds.), *Proccedings of TSD-02*. Heidelberg: Springer, no. 2448 in Lecture Notes in Artificial Intelligence (LNAI), pp. 19–26.

Marcus, M., B. Santorini and M. A. Marcinkiewicz (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2), 313–330. `ftp://ftp.cis.upenn.edu/pub/treebank/doc/cl93.ps.gz`.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition* 90, 91–117.

Padro, L. and L. Marquez (1998). On the Evaluation and Comparison of Taggers: the Effect of Noise in Testing Corpora. In *Proceedings of ACL/COLING-98*. San Francisco, California, pp. 997–1002.

Ule, T. and K. Simov (2004). Unexpected Productions May Well be Errors. In *Proceedings of LREC-04*. Lisbon, Portugal.

van Halteren, H., W. Daelemans and J. Zavrel (2001). Improving Accuracy in Word Class Tagging through the Combination of Machine Learning Systems. *Computational Linguistics* 27(2), 199–229.