

# Towards Detecting Annotation Errors in Spoken Language Corpora\*

Markus Dickinson  
Department of Linguistics  
The Ohio State University  
dickinso@ling.osu.edu

W. Detmar Meurers  
Department of Linguistics  
The Ohio State University  
dm@ling.osu.edu

**The issue** Consistency of corpus annotation is an essential property for the many uses of annotated corpora in computational and theoretical linguistics. While some research addresses the detection of inconsistencies in part-of-speech and other positional annotation (van Halteren, 2000; Eskin, 2000; Dickinson and Meurers, 2003a), only recently has there been some work in detecting errors in syntactic and other structural annotation (Dickinson and Meurers, 2003b; Ule and Simov, 2004).

Spoken language differs in many respects from written language, but to the best of our knowledge the issue of error detection in spoken language corpora has not yet been addressed. This is significant since spoken data is increasingly relevant for linguistic and computational research—and such corpora are starting to become more readily available. We address this issue in this paper, based on the variation  $n$ -gram error detection approach developed in Dickinson and Meurers (2003a). We use the German Verbmobil treebank (Hinrichs et al., 2000) as an exemplar of a spoken language corpus and discuss properties of such corpora which are relevant when adapting the variation  $n$ -gram approach to spoken language corpora.

**Why detecting annotation errors is relevant** Annotated corpora have at least two kinds of uses: firstly, as training material and as “gold standard” testing material for the development of tools in computational linguistics, and secondly, as a source of data for theoretical linguists searching for analytically relevant language patterns. The high quality annotation present in “gold standard” corpora is generally the result of a manual or semi-manual mark-up process. The annotation thus can contain annotation errors from automatic preprocesses, human post-editing, or human annotation. The presence of errors creates problems for both computational and theoretical linguistic uses, from unreliable training and evaluation of

---

\*A longer version will appear in the Proceedings of the 15th Nordic Conference of Computational Linguistics (NODALIDA 2005), Special Session on Treebanks for Spoken Language and Discourse.

natural language processing technology (see, e.g., Padro and Marquez, 1998; van Halteren, 2000; Květon and Oliva, 2002, and the work mentioned below) to low precision and recall of queries for already rare linguistic phenomena. Investigating the quality of linguistic annotation and improving it where possible thus is a key issue for the use of annotated corpora in computational and theoretical linguistics.

In Dickinson and Meurers (2003a,b), we develop the so-called variation  $n$ -gram approach and show that it can successfully detect a significant number of errors in the part-of-speech and syntactic annotation of typical “gold-standard” newspaper corpora.

**The variation  $n$ -gram approach to error detection** Our approach to error detection is based on the idea that a string occurring more than once can occur with different labels in a corpus, which we refer to as *variation*. Variation 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 options, or ii) *error*: the tagging 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 the simplest case, 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*. As an example, consider figures 1 and 2.

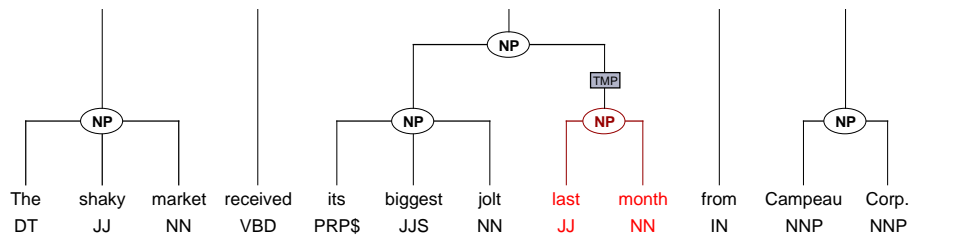


Figure 1: An occurrence of “last month” as a constituent

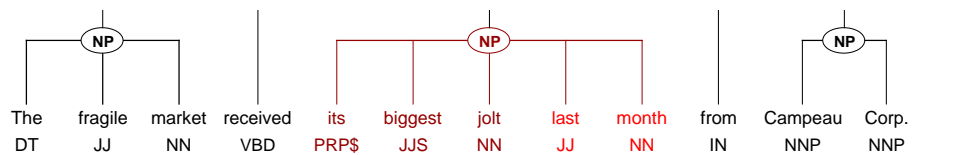


Figure 2: An occurrence of “last month” as a non-constituent

The string *last month* is a variation nucleus in this 12-gram because in one instance in the corpus it is analyzed as a noun phrase (NP), as in Figure 1, while in another

it does not form a complete constituent on its own, as shown in Figure 2, and is given the special label `NL`. An example with two syntactic categories 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).

Once the variation  $n$ -grams for a corpus have been computed, heuristics are employed to classify the variations into errors and ambiguities. The first heuristic encodes the basic fact that the label assignment for a nucleus is dependent on the context: variation nuclei in long  $n$ -grams are likely to be errors. The second takes into account 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. Both of these heuristics are independent of a specific corpus, annotation scheme, or language. Using such heuristics, we have obtained error detection precisions of 97% for pos annotation and approximately 80% for syntactic annotation.

**This study and its results** As far as we are aware, no systematic error detection research has been carried out for spoken language corpora, and it is an open question whether an error detection method such as the variation  $n$ -grams method is as effective for spoken language data as it is for written. To test this, we used 24,901 sentences (248,922 tokens) of the German Verbmobil corpus (Hinrichs et al., 2000)<sup>1</sup>. This corpus is domain-specific, consisting of transcripts of appointment negotiation, travel planning, hotel reservation, and personal computer maintenance scenarios. The speech was segmented into *dialog turns*, in order to take into account repetitions, hesitations, and false starts (cf. Stegmann et al., 2000).

While many structures annotated using crossing branches in other corpora (e.g., the TIGER corpus Brants et al., 2002) are encoded in the Verbmobil corpus using edge labels, the Verbmobil corpus still does contain discontinuous structures, i.e., category labels applying to a non-contiguous string. Thus, we developed and ran a version of the variation  $n$ -grams method for syntactic annotation (Dickinson and Meurers, 2003b) that is suitable for handling discontinuous constituents.

Turning to the results we have obtained so far, we first used sentence boundaries as stopping points for  $n$ -gram expansion and obtained 9174 total variation nuclei, as shown in figure 3. We extracted from this set only the shortest nonfringe variation  $n$ -grams<sup>2</sup> With this method, we obtained 1426 variation nuclei.

It is useful to compare this result to our previous work on written newspaper corpora. The Verbmobil corpus is roughly one-third the size of the TIGER corpus, but we obtained more  $n$ -grams for the Verbmobil corpus (1426) than for TIGER

---

<sup>1</sup>More specifically, we used the treebank versions of the following Verbmobil CDs: CD15, CD20, CD21, CD22, CD24, CD29, CD30, CD32, CD38, CD39, CD48, and CD49.

<sup>2</sup>Occurrences of the same strings within larger  $n$ -grams are thereby ignored, so as not to artificially increase the resulting set of  $n$ -grams (cf. Dickinson and Meurers, 2003a).

(500), indicating that the former is more repetitive and has more variation in the annotation of the repeated  $n$ -grams. The variation  $n$ -gram approach thus appears well-suited for domain-specific spoken language corpora, such as the Verbmobil corpus.

size	nuclei	nonfringe nuclei	size	nuclei	nonfringe nuclei
1	1808	897	8	47	2
2	2777	252	9	26	1
3	2493	135	10	12	1
4	1223	80	11	6	0
5	482	35	12	3	0
6	200	13	13	1	0
7	95	10	14	1	0

Figure 3: Number of variation nuclei in the Verbmobil corpus

Taking a closer look at the source of the repetition, there are two main sources: one which helps the variation  $n$ -grams method, and one which hurts. The first is that, because people engaged in a dialogue on a specific topic tend to express the same contents, we encounter the same strings again and again. This is the kind of repetition readily exploited by the variation  $n$ -gram approach. A different kind of recurrence, however, is that of identical words appearing next to each other, often caused by hesitations and false starts. For example, we find the unigram *und* 'and' in the middle of the trigram *und und Auto*. The problem with such examples is that with the same word being repeated, the surrounding context is no longer informative. However, given that such false starts and hesitations follow a set pattern, they can be easily identified and filtered out prior to error detection.

In another experiment, we explored the relevance of sentences (i.e., dialog turn boundaries) for detecting variation  $n$ -grams. Allowing variation  $n$ -grams to extend beyond a dialog turn resulted in 1720 cases, i.e., 20% more than for the case when variation detection is limited to a single sentence. Repeated segments frequently go beyond one dialog turn, and so the increase in detected variation outweighs the slight efficiency gain obtained with the use of dialog turns as boundaries.

Finally, we investigated the role of punctuation, which was inserted into transcribed speech of the Verbmobil corpus. We removed all punctuation from the corpus and reran the error detection code (in the version ignoring dialog turn boundaries). This resulted in 1056 examples, a loss of almost 40% of the detected cases. It thus seems to be the case that the punctuation inserted in speech corpora such as the Verbmobil corpus provides useful context for detecting variation  $n$ -grams.

Summing up, given that repetitions are prevalent in domain-specific speech, the variation  $n$ -gram method seems well-suited for detecting errors in the annotation

of such corpora. At the same time, error detection in spoken language corpora requires special attention to the role of segmentation, inserted punctuation, and particularly the nature of repetition and its causes.

## References

- Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius and George Smith (2002). The TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*. Sozopol, Bulgaria.
- Dickinson, Markus and W. Detmar Meurers (2003a). Detecting Errors in Part-of-Speech Annotation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*. Budapest, Hungary, pp. 107–114. .
- Dickinson, Markus and W. Detmar Meurers (2003b). Detecting Inconsistencies in Treebanks. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*. Växjö, Sweden.
- Eskin, Eleazar (2000). Automatic Corpus Correction with Anomaly Detection. In *Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-00)*. Seattle, Washington. .
- Hinrichs, Erhard, Julia Bartels, Yasuhiro Kawata, Valia Kordoni and Heike Telljohann (2000). The Tübingen Treebanks for Spoken German, English, and Japanese. In Wolfgang Wahlster (ed.), *Verbmobil: Foundations of Speech-to-Speech Translation*, Berlin: Springer, Artificial Intelligence, pp. 552–576.
- Květón, Pavel and Karel Oliva (2002). Achieving an Almost Correct PoS-Tagged Corpus. In Petr Sojka, Ivan Kopeček and Karel Pala (eds.), *Text, Speech and Dialogue 5th International Conference, TSD 2002, Brno, Czech Republic, September 9-12, 2002*. Heidelberg: Springer, no. 2448 in Lecture Notes in Artificial Intelligence (LNAI), pp. 19–26.
- Padro, Lluís and Lluís Marquez (1998). On the Evaluation and Comparison of Taggers: the Effect of Noise in Testing Corpora. In *COLING-ACL*. pp. 997–1002.
- Stegmann, Rosmary, Heike Telljohann and Erhard W. Hinrichs (2000). *Stylebook for the German Treebank in VERBMOBIL*. Verbmobil-Report 239, Universität Tübingen, Tübingen, Germany. .
- Ule, Tylman and Kiril Simov (2004). Unexpected Productions May Well be Errors. In *Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC 2004)*. Lisbon, Portugal.
- van Halteren, Hans (2000). The Detection of Inconsistency in Manually Tagged Text. In Anne Abeillé, Thorsten Brants and Hans Uszkoreit (eds.), *Proceedings of the Second Workshop on Linguistically Interpreted Corpora (LINC-00)*. Luxembourg.