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

1 Introduction

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), more recently work has also started to address errors in syntactic and other structural annotation (Dickinson and Meurers, 2003b, 2005; Ule and Simov, 2004; Dickinson, 2005).

Spoken language differs in many respects from written language, but to the best of our knowledge the issue of detecting errors in the annotation of spoken language corpora has not yet been systematically 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, as illustrated by the holdings of the Linguistic Data Consortium (http://www.ldc.upenn.edu). This paper addresses the issue, 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 for detecting errors in syntactic annotation of 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 in a variety of ways, from unreliable training of natural language processing technology (see, e.g., van Halteren et al., 2001; Květŏn and Oliva, 2002) and evaluation of such technology (e.g., Padro and Marquez, 1998; van Halteren, 2000) to low precision and recall of queries for already rare linguistic phenomena (e.g., Meurers, 2005). 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.

2 The variation *n*-gram approach

In Dickinson and Meurers (2003a,b, 2005), 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. 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,[1] 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*.

Consider the examples in Figures 1 and 2, taken from the Wall Street

[1] For example, the word *can* is ambiguous between being an auxiliary, a main verb, or a noun and thus there is variation in the way *can* would be tagged in *I can play the piano*, *I can tuna for a living*, and *Pass me a can of beer, please*.

Journal corpus (WSJ, Marcus et al., 1993) as annotated in the Penn Treebank 3 (Marcus et al., 1999).

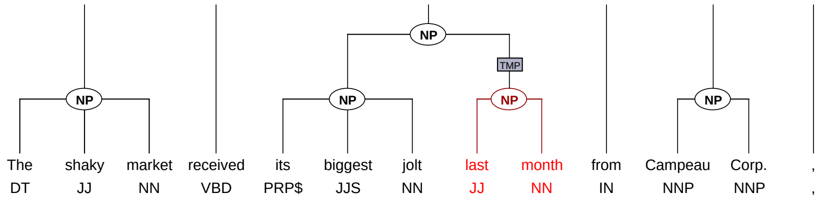

Figure 1: An occurrence of “last month” as a constituent

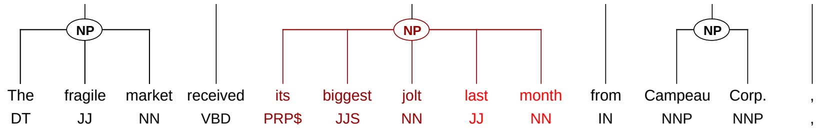

Figure 2: An occurrence of “last month” as a non-constituent

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, for which we assign the special label NIL. The two annotations of the recurring string thus differ with respect to the syntactic bracketing, which is reflected in the category label.

Another example, illustrating a basic category variation, 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).

While for part-of-speech annotation there is a one-to-one correspondence between tokens and labels, for syntactic annotation the annotated structures can range in size, from a size of one word to the length of the longest constituent in the corpus. Correspondingly, variation nuclei of size one are used for detecting errors in part-of-speech annotation (Dickinson and Meurers, 2003a), but a sequence of runs for all constituent sizes is used to detect variation nuclei in syntactic annotation (Dickinson and Meurers, 2003b). For discontinuous structural annotation, the nuclei in addition are not required to consist of continuous stretches of material (Dickinson and Meurers, 2005).

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 these heuristics, we obtained error detection precisions of 93% for part-of-speech annotation (Dickinson and Meurers, 2003a; Dickinson, 2005) and approximately 80% for syntactic annotation, in case studies for English with continuous constituents (Dickinson and Meurers, 2003b) and for German with discontinuous constituents (Dickinson and Meurers, 2005).

3 Error detection for spoken language corpora

The work reported in the previous section has concentrated on written language, and there are many challenges to deal with when attempting to apply such a method to spoken language data. For example, spoken language is characterized by repetitions, false starts, and other speech errors which make it appear quite different from written (see, e.g., Brennan, 2000). Additionally, people typically speak in shorter sentences, especially in dialogues with others, where a single "utterance" may not even be a complete sentence (see, e.g., Traum and Heeman, 1997). Furthermore, while written language typically includes punctuation, any punctuation included in the transcription of spoken language was inserted by the transcriber—a difference which can have a significant impact, considering the complex role punctuation plays in written language (Nunberg, 1990). All of these factors can be expected to have an effect on an error detection procedure that is driven by a search for inconsistently-annotated recurring patterns.

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*-gram method is as effective for spoken language data as it is for written. To address this question, we used 24,901 sentences (248,922 tokens) of the German Verbmobil corpus (Hinrichs et al., 2000).[2] 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

[2] 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.

starts; these are akin to sentences found in written language, but dialog turns "may consist of one or more sentences in the grammatical sense" (Stegmann et al., 2000, p. 22).

The annotation of the Verbmobil corpus consists of tree structures with node and edge labels (Stegmann et al., 2000). The node labels refer to one of four different levels of syntactic annotation: the highest level of nodes specify the turn type; field-level nodes give topological field names; phrase-level nodes indicate syntactic categories; and lexical-level nodes encode the part-of-speech using the Stuttgart-Tübingen TagSet (STTS, Schiller et al., 1995; Thielen and Schiller, 1996). Thus, the node labels in the tree encode both syntactic category and topological field information. Edge labels on the phrase level encode grammatical functions.

Figure 3 on the next page shows the annotation for the short example sentence given with its English translation in (1).

(1) und wann wollen wir uns nach der Reise auf ein Glas Wein treffen
and when want we us after the trip on a glass wine meet
'And when do you want to meet over a glass of wine after the trip?'

In the figure, the node labels are shown as circles and the edge labels as boxes. Note that the different levels of annotation are all encoded in the same way (with the part-of-speech being displayed differently). For example, both LK and PX are node labels in the tree, but LK (left sentence bracket) is a topological field whereas PX (prepositional phrase) is a syntactic category. We will return to the issue of different layers of annotation and their effect below.

While many structures annotated using crossing branches in other corpora, such as TIGER (Brants et al., 2002) are encoded in the Verbmobil corpus using edge labels, the Verbmobil corpus does contain some discontinuous structures, i.e., category labels applying to a non-contiguous string. The discontinuities were often over punctuation, which is unattached in the corpus. Thus, we developed and ran a version of the variation *n*-grams method for syntactic annotation that is suitable for handling discontinuous constituents (Dickinson and Meurers, 2005).

Before turning to a discussion of the results of running the resulting algorithm on the Verbmobil corpus, there are two interesting aspects of the corpus that should be discussed, given that they are typical for such a spoken language corpus and directly affect the variation *n*-gram error detection method. The first is repetition, arising because people engaged in a dialogue on a specific topic tend to express the same contents; thus, one encounters the same strings again and again in a corpus. For example, one finds 35 instances of (2) in the Verbmobil corpus, with *guten Tag* labeled 33 times as

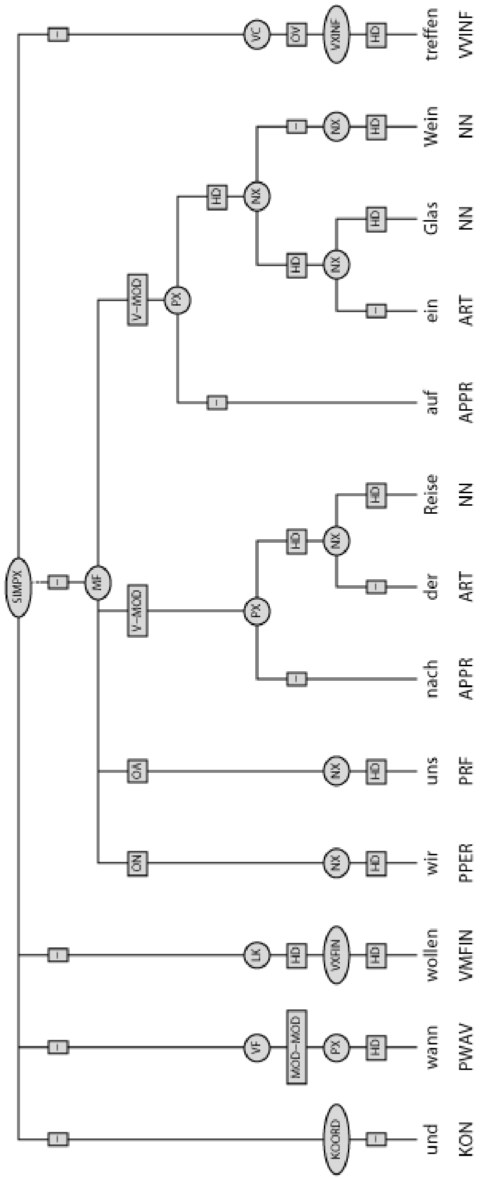

Figure 3: A simple example from the Verbmobil corpus

DM/NX (discourse marker dominating a noun phrase)[3] and twice as NIL (non-constituent).

(2) , guten Tag , Frau
, good day , woman
'Hello, Ms. ...'

This kind of repetition is 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*, as in (3).

(3) und und Auto
and and car

The problem with such examples is that with the same word being repeated, the surrounding context is no longer informative: it results in differences in context for otherwise comparable variation *n*-grams, as well as the opposite of making contexts comparable which otherwise would not be. False starts and hesitations involving single words can be identified and filtered out prior to error detection, but longer false starts are difficult to detect and can be confusable with ordinary sentences that should not be filtered out, such as in the English *he said he said hello*.

3.1 Results

Turning to the results we have obtained so far, in our first experiment we started with a version of the variation *n*-gram algorithm for discontinuous constituents (Dickinson and Meurers, 2005) that uses the boundaries of the largest syntactic units as stopping points for *n*-gram expansion to ensure efficient processing. As shown in Figure 4, this resulted in 9174 total variation nuclei. From this set, we extract the shortest nonfringe variation nuclei, just as in Dickinson and Meurers (2005). Occurrences of the same strings within larger *n*-grams are thereby ignored, so as not to artificially increase the resulting set of nuclei. This resulted in 1426 shortest nonfringe variation nuclei.

It is useful to compare this result to that obtained in Dickinson and Meurers (2005) for the German newspaper corpus TIGER (Brants et al., 2002). The Verbmobil corpus is roughly one-third the size of the TIGER corpus, but we obtained significantly more shortest nonfringe variation nuclei for the

[3] As discussed at the end of section 3.1.3, we collapse unary branches into a single label.

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 4: Number of variation nuclei in the Verbmobil corpus

Verbmobil corpus (1426) than for TIGER (500), indicating that the Verbmobil corpus is more repetitive and/or includes more variation in the annotation of the repeated strings. This supports the reasoning in the previous section that the variation *n*-gram approach is well-suited for domain-specific spoken language corpora, such as the Verbmobil corpus.

3.1.1 The effect of dialog turn boundaries

In another experiment, we explored the effect of using the boundaries of the largest syntactic units in the corpus, i.e., the dialog turn boundaries, as stopping points for *n*-gram expansion. Allowing variation *n*-grams to extend beyond a dialog turn resulted in 1720 cases, i.e., 20% more than in our first experiment, where variation detection was limited to a single dialog turn; a complete breakdown is given in Figure 5.

size	nuclei	nonfringe nuclei	size	nuclei	nonfringe nuclei
1	1808	1081	8	47	2
2	2777	307	9	26	1
3	2493	169	10	12	1
4	1223	95	11	6	0
5	482	39	12	3	0
6	200	14	13	1	0
7	95	11	14	1	0

Figure 5: Number of variation nuclei, ignoring dialog turn boundaries

In conclusion, the second experiment shows that repeated segments frequently go beyond a dialog turn so that error detection for spoken language corpora should ignore dialog turn boundaries.

3.1.2 The effect of punctuation

Finally, in a third experiment, we investigated the role of punctuation, which had been inserted into the 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 shortest nonfringe variation nuclei, a loss of almost 40% of the detected cases compared to the second experiment. The punctuation inserted into the corpus thus seems to provide useful context for detecting variation *n*-grams.

However, even though punctuation appears to be useful in finding more variations, punctuation symbols are not always reliable indicators of identical context. To illustrate this, let us examine some different uses of the comma. In (4), we find commas delimiting elements of an enumerated list, and thus *Freitag* ('Friday') forms a noun phrase (NX) by itself.

(4)	das	wäre	Donnerstag	,	Freitag	,	Samstag	.
	that	would be	Thursday	,	Friday	,	Saturday	.

In example (5), on the other hand, we have a comma being used in a date expression. In this case, *Freitag* correctly forms part of the larger noun phrase *Freitag den achten Mai* ('Friday, the eighth of May'), where the comma is used to separate the day from the month.

(5)	ab	achten	Mai	,	Freitag	,	den	achten	Mai	,	hätte	ich	für	vier
	from	eighth	May	,	Friday	,	the	eighth	May	,	would've	I	for	four
	Tage	Zeit												
	days	time												

The comma and other punctuation thus are potentially ambiguous tokens, with different uses and meanings, essentially on a par with ordinary word tokens—an observation which is not specific to spoken language but equally applies to written language corpora.

3.1.3 The effect of the annotation scheme

Turning to the annotation scheme used in the Verbmobil corpus and its effect on error detection using the variation *n*-gram detection method, there are two issues that deserve some attention: non-local category distinctions and the role of topological fields.

Non-local distinctions Clearly the most serious problem for our error detection method (and for most algorithms for corpus annotation) are category distinctions which are inherently non-local. We use the nonfringe heuristic to isolate variation *n*-grams for analysis, but some nuclei cannot be reliably disambiguated using the local context. For example, the nucleus *fahren* ('drive') in the 4-gram given in (6) is ambiguous; in (7) it is a finite verb (VXFIN), but it is a non-finite verb (VXINF) in (8).

(6) nach Hannover fahren .
to Hannover drive .
'to drive to Hannover'

(7) daß wir am Mittwoch und Donnerstag nach Hannover fahren .
that we on Wednesday and Thursday to Hannover drive .
'that we drive to Hannover on Wednesday and Thursday.'

(8) Wir wollten nach Hannover fahren .
we wanted to Hannover drive .
'We wanted to drive to Hannover.'

In (7), *fahren* ('drive') is a finite verb in third person singular, occurring in a dependent clause. In (8), on the other hand, the verb *wollten* ('wanted') is the finite verb in a declarative sentence and it selects the infinitival *fahren* ('to drive'). The problem is that looking solely at *fahren* and its local context, it is not possible to determine whether one is dealing with the finite or the non-finite form since the finite verb in a declarative sentence like (8) typically occurs as the second constituent of a sentence, which is arbitrarily far away from the non-finite verbs typically found at the right edge of a sentence. To be able to distinguish such cases, the variation *n*-gram detection method thus needs to be extended with a more sophisticated notion of disambiguating context that goes beyond the local environment.

Topological fields When we introduced the annotation of the Verbmobil corpus in section 3, we mentioned that the annotation includes topological field information. The topological field labels encode the general word order properties of the entire sentence, not the properties of a word and its local context. As a result, they can cause problems similar to the just discussed non-local category distinctions. For example, the complementizer field C is described as follows in the manual: "The C-position only occurs in verb-final clauses" (Stegmann et al., 2000, p. 11), but whether a clause is verb-final or not is a property of the sentence, not of the C field itself.

A second issue involving the topological fields arises from the fact that the annotation scheme includes two kinds of non-terminal nodes: field-level nodes that bear topological field labels and phrase-level nodes with syntactic category labels. This has the effect that some phrases are dominated by phrases, whereas others are dominated by fields—but clearly one does not want to compare field labels with category labels.

A case where this becomes directly relevant to the detection of annotation errors is the treatment of unary branches in the syntactic annotation. Since both nodes in a unary branching tree dominate the same terminal material, we propose in Dickinson and Meurers (2003b) that such unary branches are best dealt with by folding the two category labels into a single category. For example, an NP (noun phrase) dominating a QP (quantifier phrase) in the WSJ corpus is encoded as a node of category NP/QP. Such folding can involve any non-terminal node dominating a single non-terminal daughter, so that in the Verbmobil corpus it can also combine a topological label with a syntactic category label. For instance, we find NF/NX for a Nachfeld (NF, used for extraposed material) dominating a noun phrase (NX).

Such labels combining field and syntactic category information can cause problems by introducing artificial variation. Consider, for example, a variation nucleus involving *je nachdem* ('depending on') in (9), which varies between PX (prepositional phrase) and VF/PX (Vorfeld, i.e., fronted material, dominating a prepositional phrase).

(9) und je nachdem ,
and depending on ,

This sort of variation is perfectly acceptable, i.e., not an error, since the topological field (Vorfeld) refers to where the prepositional phrase is placed in the sentence and has nothing to do with the internal properties of the PX. For the purpose of our error detection approach, we thus need to keep the topological field nodes clearly distinct from the syntactic category nodes.

More generally, the issue of the topological field information included in the annotation of the Verbmobil corpus highlights that it is important to understand the nature of the corpus annotation scheme when porting the variation *n*-gram method to a new type of annotation scheme.

4 Summary and Outlook

As our pilot study on the German Verbmobil corpus indicates, the variation *n*-gram method seems well-suited for detecting errors in the annotation of such corpora given that repetitions are prevalent in domain-specific speech. 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.

In the future, we plan on fully evaluating the number of errors detected by the method, after identifying and removing the problematic patterns mentioned above. We would also like to apply the method to other layers of annotation in the German Verbmobil corpus, such as part-of-speech annotation, and to test the general applicability of the insights we gained from working with the Verbmobil corpus by applying the method to other spoken language corpora.

Finally, the special nature of punctuation in spoken language corpora creates an interesting opportunity for error detection, and we are experimenting with adapting our error detection method to find inconsistencies in the insertion of punctuation.

Acknowledgments

We would like to thank the reviewers and participants of the NODALIDA'05 special session on Treebanks for Spoken Language and Discourse, the participants of the 2005 Midwestern Computational Linguistics Colloquium, and the CLippers discussion group at OSU for helpful feedback and discussion of our work. We are also grateful to Sandra Kübler and Erhard Hinrichs for making the German Verbmobil treebank available.

References

Brants, S., S. Dipper, S. Hansen, W. Lezius and G. Smith (2002). The TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*. Sozopol, Bulgaria. http://www.bultreebank.org/proceedings/paper03.pdf.

Brennan, S. E. (2000). Processes that shape conversation and their implications for computational linguistics. In *Proceedings of the 38th Annual Meeting of the Association of Computational Linguistics*. http://www.aclweb.org/anthology/P00-1001.

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

Dickinson, M. and W. D. 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. http://www.aclweb.org/anthology/E03-1068.

Dickinson, M. and W. D. Meurers (2003b). Detecting Inconsistencies in Treebanks. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT-03)*. Växjö, Sweden. http://ling.osu.edu/˜dm/papers/dickinson-meurers-tlt03.html.

Dickinson, M. and W. D. Meurers (2005). Detecting Errors in Discontinuous Structural Annotation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*. Ann Arbor, MI, USA. http://ling.osu.edu/˜dm/papers/dickinson-meurers-05.html.

Eskin, E. (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. http://www.cs.columbia.edu/˜eeskin/papers/treebank-anomaly-naacl00.ps.

Hinrichs, E., J. Bartels, Y. Kawata, V. Kordoni and H. Telljohann (2000). The Tübingen Treebanks for Spoken German, English, and Japanese. In W. Wahlster (ed.), *Verbmobil: Foundations of Speech-to-Speech Translation*, Berlin: Springer, Artificial Intelligence, pp. 552–576.

Květŏn, P. and K. Oliva (2002). Achieving an Almost Correct PoS-Tagged Corpus. In P. Sojka, I. Kopeček and K. 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.

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.

Marcus, M., B. Santorini, M. A. Marcinkiewicz and A. Taylor (1999). Treebank-3 Corpus. Linguistic Data Consortium. Philadelphia. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99T42.

Meurers, W. D. (2005). On the use of electronic corpora for theoretical linguistics. Case studies from the syntax of German. *Lingua* 115(11), 1619–1639. http://ling.osu.edu/˜dm/papers/meurers-03.html.

Nunberg, G. (1990). *The Linguistics of Punctuation*. No. 18 in Lecture Notes. Stanford, CA: CSLI Publications.

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

Schiller, A., S. Teufel and C. Thielen (1995). *Guidlines für das Taggen deutscher Textcorpora mit STTS*. Tech. rep., IMS, Univ. Stuttgart and SfS, Univ. Tübingen. http://www.cogsci.ed.ac.uk/˜simone/stts_guide.ps.gz.

Stegmann, R., H. Telljohann and E. W. Hinrichs (2000). *Stylebook for the German Treebank in VERBMOBIL*. Verbmobil-Report 239, Universität Tübingen, Tübingen, Germany. http://verbmobil.dfki.de/cgi-bin/verbmobil/htbin/decode.cgi/share/VM-depot/FTP-SERVER/vm-reports/report-239-00.ps.

Thielen, C. and A. Schiller (1996). Ein kleines und erweitertes Tagset fürs Deutsche. In H. Feldweg and E. W. Hinrichs (eds.), *Lexikon und Text: wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen*, Tübingen: Max Niemeyer Verlag, vol. 73 of *Lexicographica: Series maior*, pp. 215–226.

Traum, D. and P. Heeman (1997). Utterance Units in Spoken Dialogue. In E. Maier, M. Mast and S. LuperFoy (eds.), *Dialogue Processing in Spoken Language Systems, Lecture Notes in Artificial Intelligence*, Heidelberg: Springer-Verlag, pp. 125–140.

Ule, T. and K. Simov (2004). Unexpected Productions May Well be Errors. In *Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC 2004)*. Lisbon, Portugal. http://www.sfs.uni-tuebingen.de/˜ule/Paper/us04lrec.pdf.

van Halteren, H. (2000). The Detection of Inconsistency in Manually Tagged Text. In A. Abeillé, T. Brants and H. Uszkoreit (eds.), *Proceedings of the Second Workshop on Linguistically Interpreted Corpora (LINC-00)*. Luxembourg.

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.