

Detecting Errors in Corpus Annotation

On Variation Detection
(<http://decca.osu.edu>)

Detmar Meurers
University of Tübingen

CLARA Thematic Training Course on Methods and Technologies
for Consolidating and Harmonising Treebank Annotation
UFAL, Charles University, Prague
December 13–16, 2010

Detecting Errors in Corpus Annotation
Detmar Meurers
University of Tübingen

Introduction
What is Annotation Error?
How to obtain high quality
Part of Speech
Verbler annotation
Comparing annotated corpora
Independent annotation
Multiple annotation
Results for the NLI
Annotation scheme feedback

Consistency
Verbler annotation
Comparing annotated corpora
NLI case study
Results
Consistency
Comparing annotated corpora
Results for T02B
Annotation tool
Results

Dependency
Verbler annotation
Nature of dependencies
Inflectional dependencies
Annotation tool
Annotation tool
Summary

UNIVERSITÄT TÜBINGEN

Introduction

Corpora with "gold standard" annotation are used

- as **training** and **testing** material for NLP algorithms/tools
- for searching for linguistically relevant patterns

Such annotation generally results from a semi-automatic markup process, which can include errors through

- automatic processes
- human annotation or post-editing

Detecting Errors in Corpus Annotation
Detmar Meurers
University of Tübingen

Introduction
What is Annotation Error?
How to obtain high quality
Part of Speech
Verbler annotation
Comparing annotated corpora
Independent annotation
Multiple annotation
Results for the NLI
Annotation scheme feedback

Consistency
Verbler annotation
Comparing annotated corpora
NLI case study
Results
Consistency
Comparing annotated corpora
Results for T02B
Annotation tool
Results

Dependency
Verbler annotation
Nature of dependencies
Inflectional dependencies
Annotation tool
Annotation tool
Summary

UNIVERSITÄT TÜBINGEN

Effects of Annotation Errors

- Less reliable **training** of NLP technology

- van Halteren et al. (2001): a tagger trained on WSJ (Marcus et al. 1993) performs significantly worse than one trained on LOB (Johansson 1986)

- Less reliable **evaluation** of NLP technology

- van Halteren (2000): 13.6%–20.5% of cases where WPDV tagger disagrees with BNC-sampler annotation, cause is error in BNC-sampler (0.3% error, Leech 1997). Error rates for other corpora much higher.
- Padro & Marquez (1998): because of errors in the testing data, cannot tell which of two taggers is better

- Low precision** and **recall** of queries for already rare linguistic phenomena

- Meurers (2005): low precision of queries for verbal complex patterns since certain finite and non-finite verb forms are not reliably distinguished by German taggers

Detecting Errors in Corpus Annotation
Detmar Meurers
University of Tübingen

Introduction
What is Annotation Error?
How to obtain high quality
Part of Speech
Verbler annotation
Comparing annotated corpora
Independent annotation
Multiple annotation
Results for the NLI
Annotation scheme feedback

Consistency
Verbler annotation
Comparing annotated corpora
NLI case study
Results
Consistency
Comparing annotated corpora
Results for T02B
Annotation tool
Results

Dependency
Verbler annotation
Nature of dependencies
Inflectional dependencies
Annotation tool
Annotation tool
Summary

UNIVERSITÄT TÜBINGEN

Effects of Annotation Errors

Searching for linguistic phenomena: The role of precision

- By precision of search we are referring to:

- Of the results to the query, how many represent the learner language patterns searched for?
- False positives can result in two ways:
 - Expression used in query also characterizes patterns other than the ones we are interested in.
 - Some of the annotations the query refers to are incorrect.

- Requirements on precision of search

- for **qualitative** analysis: Needs to be high enough to find relevant examples among the false positives.
- for **quantitative** analysis: For reliable results, very high precision is required, in particular where specific rare language phenomena are concerned.
 - As known from Zipf's curse, most things occur rarely ...

Effects of Annotation Errors

Searching for linguistic phenomena: The role of recall

- By recall of search we are referring to:

- How many of the intended examples that in principle are in the corpus are in fact found by the query?

- Requirements on recall of search

- for **qualitative** analysis: Any results found useful, but danger of partial blindness where subcases are not captured by query approximating larger phenomenon.
- for **quantitative** analysis: Maximizing recall is crucial for reliable quantitative results.

⇒ Where a query characterizing a target phenomenon is expressed in terms of annotation, high annotation quality is important, and essential for quantitative analysis.

How to obtain high quality annotation

- Annotate corpus independently several times, then test interannotator agreement (Brants & Skut 1998; Arstein & Posio 2009)

- Interannotator agreement: Can the distinctions made in the annotation scheme can be applied consistently based on the information available in the corpus?

- Define adequate annotation scheme, with explicit documentation and a list of problematic cases to achieve maximal agreement (Voutilainen & Järvinen 1995; Sampson & Barbaczy 2003).

- keep only distinctions which can be reliably and consistently identified and annotated uniquely
- appendix of difficult cases and how to resolve them crucial

Detecting Errors in Corpus Annotation
Detmar Meurers
University of Tübingen

Introduction
What is Annotation Error?
How to obtain high quality
Part of Speech
Verbler annotation
Comparing annotated corpora
Independent annotation
Multiple annotation
Results for the NLI
Annotation scheme feedback

Consistency
Verbler annotation
Comparing annotated corpora
NLI case study
Results
Consistency
Comparing annotated corpora
Results for T02B
Annotation tool
Results

Dependency
Verbler annotation
Nature of dependencies
Inflectional dependencies
Annotation tool
Annotation tool
Summary

UNIVERSITÄT TÜBINGEN

Our research questions

- How about automatic methods for error detection?
 - Detection can feed into repair as second stage of correction (cf. also Oliva 2001; Bhaeta 2002).
 - What can be done for annotation of language in general?
 - Can errors be found in common "gold standard" corpora regarding their
 - part-of-speech annotation (Dickinson & Meurers 2003a)
 - syntactic annotation (Dickinson & Meurers 2003b; Boyd, Dickinson & Meurers 2007)
 - discontinuous syntactic annotation (Dickinson & Meurers 2004)
 - dependency annotation (Boyd, Dickinson & Meurers 2008)
- including spoken language corpora (Dickinson & Meurers 2005a).
- ⇒ Detection of annotation errors through automatic analysis of comparable data recurring in the corpus
- DECCA NSF project (<http://decca.osu.edu>)
 - Dickinson (2005)

Variation Detection for POS Annotation

(Dickinson & Meurers 2003a)

- POS tagging** reduces the set of lexically possible tags to the correct tag for a specific corpus occurrence.

- A word occurring multiple times in a corpus can occur with more than one annotation.

- Variation**: material occurs multiple times in corpus with different annotations

- Variation can result from
 - genuine **ambiguity**
 - inconsistent, **erroneous tagging**

- How can one find such variation and decide whether it's an ambiguity or error?

Classifying variation

- The key to classifying variation lies in the context:

- The more similar the context of the occurrences, the more likely the variation is an error.

- A simple way of making "similarity of context" concrete is to say it consists of

- words
- which immediately surround the variation, and
- require identity of contexts.

⇒ Extract all n-grams containing at least one token that is annotated differently in another occurrence of that n-gram in the corpus.

- variation nucleus**: recurring unit with different annotation
- variation n-gram**: variation nucleus with identical context

Detecting Errors in Corpus Annotation
Detmar Meurers
University of Tübingen

Introduction
What is Annotation Error?
How to obtain high quality
Part of Speech
Verbler annotation
Comparing annotated corpora
Independent annotation
Multiple annotation
Results for the NLI
Annotation scheme feedback

Consistency
Verbler annotation
Comparing annotated corpora
NLI case study
Results
Consistency
Comparing annotated corpora
Results for T02B
Annotation tool
Results

Dependency
Verbler annotation
Nature of dependencies
Inflectional dependencies
Annotation tool
Annotation tool
Summary

UNIVERSITÄT TÜBINGEN

Detecting Errors in Corpus Annotation
Detmar Meurers
University of Tübingen

Introduction
What is Annotation Error?
How to obtain high quality
Part of Speech
Verbler annotation
Comparing annotated corpora
Independent annotation
Multiple annotation
Results for the NLI
Annotation scheme feedback

Consistency
Verbler annotation
Comparing annotated corpora
NLI case study
Results
Consistency
Comparing annotated corpora
Results for T02B
Annotation tool
Results

Dependency
Verbler annotation
Nature of dependencies
Inflectional dependencies
Annotation tool
Annotation tool
Summary

UNIVERSITÄT TÜBINGEN

Detecting Errors in Corpus Annotation
Detmar Meurers
University of Tübingen

Introduction
What is Annotation Error?
How to obtain high quality
Part of Speech
Verbler annotation
Comparing annotated corpora
Independent annotation
Multiple annotation
Results for the NLI
Annotation scheme feedback

Consistency
Verbler annotation
Comparing annotated corpora
NLI case study
Results
Consistency
Comparing annotated corpora
Results for T02B
Annotation tool
Results

Dependency
Verbler annotation
Nature of dependencies
Inflectional dependencies
Annotation tool
Annotation tool
Summary

UNIVERSITÄT TÜBINGEN

Feedback for revising annotation scheme

For 140 of the 2436 erroneous variation nuclei, the variation was clearly incorrect, but which tag is the correct one is unclear from the guidelines (Santorini 1990).

Example: Salomon Brothers Inc

Brothers is tagged

- ▶ 27 times as proper noun (NWP)
- ▶ 24 as plural proper noun (NPPs).

⇒ Variation n-gram error detection helps identify error-prone distinctions, which need to be documented more explicitly or possibly eliminated, e.g.:

- proper vs. common nouns
- certain types of noun-adjective homographs

UNIVERSITÄT TüBINGEN

Related work on POS error detection

- ▶ Work with another focus, which could be combined with our consistency-checking approach:
 - Deriving and searching for bigrams of tags which should never be allowed (Kvétón & Oliva 2002). → Inconsistencies are mostly possible bigrams.
 - Sparse Markov transducers used to detect anomalies, i.e., rare local tag patterns (Eskin 2000). → Inconsistencies are mostly recurrent, not rare.
- ▶ Using parsing failures to detect ill-formed annotation serving as parser input (Hirakawa et al. 2000; Müller & Ule 2002). → Language specific resources.
- ▶ Searching and correcting with hand-written rules (Oliva 2001; Blaheta 2002)
- ▶ Related to consistency of annotation:
 - Comparing tagger output with gold standard (van Halleren 2000; Abney et al. 1999). Taggers detect consistent behavior in order to replicate it.

UNIVERSITÄT TüBINGEN

Summary for POS error detection

- ▶ We discussed a detection methods for POS annotation errors in gold-standard corpora:
 - detect variation within comparable contexts
 - classify such variation as error or ambiguity using general heuristics
- ▶ Idea relies on multiple corpus occurrences of a particular word with different annotations
 - particularly useful for hand-corrected gold-standard corpora
- ▶ Evaluation showed the method detects errors in the WSJ with
 - 92.8% precision
 - 17% estimated recall
- ▶ Qualitative inspection of the detected variation can provide valuable feedback for annotation scheme (re)design and documentation.

UNIVERSITÄT TüBINGEN

Variation Detection for Syntactic Annotation

(Dickinson & Meurers 2003b, 2004; Boyd, Dickinson & Meurers 2007)

- ▶ Let's try to apply variation detection to the syntactic annotation in treebanks!
 - How can two syntactically annotated sentences be compared for this?
- ▶ Variation detection is closely related to interannotator agreement testing for multiply annotated corpus.
 - How are multiple annotations of the same sentences compared for testing interannotator agreement?
 - Calder (1997) and Brants & Skut (1998) present algorithm for detecting differences in annotation.
 - algorithm is annotation-driven, asymmetric, and sentence-based

⇒ We are looking for a data-driven, symmetric, string-based approach.

UNIVERSITÄT TüBINGEN

Defining variation nuclei for syntactic annotation

How can we obtain a data-driven definition of a variation nucleus as the unit of data on which the comparison of syntactic annotation can be based?

Problem: No one-to-one mapping between word and label, as with part of speech.

Idea: Decompose variation nucleus detection into series of runs for all relevant string lengths, more specifically

- ▶ define one-to-one mapping between string of a given length and the label for that string
- ▶ perform runs for strings from length 1 to longest constituent in corpus

UNIVERSITÄT TüBINGEN

Defining variation nuclei for syntactic annotation

How to compare annotation for syntactic variation nuclei

- ▶ To obtain a uniform mapping from strings to labels
 - assign all non-constituent occurrences of a string the special label ω .
- ▶ Only compare categories assigned to the entire nucleus.
 - This intentionally ignores the internal structure, which is taken into account when shorter strings are checked.

UNIVERSITÄT TüBINGEN

Examples from the WSJ corpus

- ▶ Variation between two syntactic category labels:

(4) *maturity* next Tuesday

labeled as **NP** twice
PP once
- ▶ Variation between constituent and non-constituent:

UNIVERSITÄT TüBINGEN

Computing the variation nuclei of a treebank

A simple way to calculate all variation nuclei:

1. step through corpus:
 - store all stretches of length i with category label or ω .
2. eliminate the non-varying stretches

Problem: Inefficient generate-and-test method considering all stretches of strings starting at any position in the corpus.

Insight:

- ▶ The way we have set things up, variation involves at least one constituent occurrence of a nucleus.
- Only strings analyzed as constituent somewhere in corpus need to be compared to annotation of other occurrences of that string.

UNIVERSITÄT TüBINGEN

Computing variation n-grams for a treebank

Algorithm

For each constituent length i ($1 \leq i \leq$ longest-constituent):

1. Compute the set of nuclei:
 - a) Find all constituents of length i : store them with their category label
 - b) For each type of string stored as constituent of length i , add ω for each non-constituent occurrence
2. Compute variation nuclei set as:
 - all nuclei from step 1 with more than one label
3. Generate variation n-grams for these variation nuclei, just as defined for part of speech annotation

UNIVERSITÄT TüBINGEN

A case study: Applying the method to the WSJ

Detecting Errors in Corpus Annotation
 Erman, Mueser, University of Tübingen
 Introduction
 What is Annotation Error
 How to obtain high quality
 Part of Speech
 Verbalizer
 Noun
 Verb
 Adjective
 Adverb
 Preposition
 Conjunction
 Participle
 Pronoun
 Interjection
 Negation
 Whitespace
 Dependency
 Relation of dependence
 Relation of modification
 Relation of coreference
 Summary

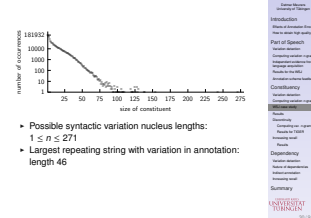
- Two types of syntactic information in the PennTreebank (Marcus, Santorini, Marcinkiewicz & Taylor 1999):
 - syntactic category generally determined by
 - lexical material in the covered string and
 - the way this material is combined
 - syntactic function (also) determined by
 - material outside constituent
- We focus on the syntactic category.
- Technical realization:
 - TIGERRegistry (Lezuis et al. 2002) converts, e.g., temporal noun phrase (NP-TMP) to noun phrase node (NP) under temporal (TMP) edge
 - variation n-gram test based on node labels only

Dealing with unary trees
 Erman, Mueser, University of Tübingen
 Introduction
 What is Annotation Error
 How to obtain high quality
 Part of Speech
 Verbalizer
 Noun
 Verb
 Adjective
 Adverb
 Preposition
 Conjunction
 Participle
 Pronoun
 Interjection
 Negation
 Whitespace
 Dependency
 Relation of dependence
 Relation of modification
 Relation of coreference
 Summary

- unary branch causes same string to be annotated by two distinct categories
 - would be detected as variation in annotation
- eliminate unary branches and relabel with mother/daughter category label, adding 70 new labels to original 27.
- Example:

$$\begin{array}{c}
 \text{NP} \\
 \text{QP} \\
 \text{10 million}
 \end{array}
 \Rightarrow
 \begin{array}{c}
 \text{NP/QP} \\
 \text{10 million}
 \end{array}$$

Constituent lengths in the WSJ
 Erman, Mueser, University of Tübingen
 Introduction
 What is Annotation Error
 How to obtain high quality
 Part of Speech
 Verbalizer
 Noun
 Verb
 Adjective
 Adverb
 Preposition
 Conjunction
 Participle
 Pronoun
 Interjection
 Negation
 Whitespace
 Dependency
 Relation of dependence
 Relation of modification
 Relation of coreference
 Summary



Error detection results
 Erman, Mueser, University of Tübingen
 Introduction
 What is Annotation Error
 How to obtain high quality
 Part of Speech
 Verbalizer
 Noun
 Verb
 Adjective
 Adverb
 Preposition
 Conjunction
 Participle
 Pronoun
 Interjection
 Negation
 Whitespace
 Dependency
 Relation of dependence
 Relation of modification
 Relation of coreference
 Summary

Misclassified Ambiguities I: Null elements

- 6277 distinct, non-fringe variation nuclei
 - distinct: each corpus position is only taken into account for longest variation n-gram it occurs in
 - non-fringe: nucleus is surrounded by at least one word of identical context
- We inspecting 100 randomly sampled examples:
 - 71% errors, with 95% confidence interval for point estimate of .71 being (.6211, .7989)
 - between 3898 and 5014 erroneous variation nuclei, each corresponding to at least one token error
- What are the reasons for the misclassified ambiguities?

Misclassified Ambiguities I: Null elements
 Erman, Mueser, University of Tübingen
 Introduction
 What is Annotation Error
 How to obtain high quality
 Part of Speech
 Verbalizer
 Noun
 Verb
 Adjective
 Adverb
 Preposition
 Conjunction
 Participle
 Pronoun
 Interjection
 Negation
 Whitespace
 Dependency
 Relation of dependence
 Relation of modification
 Relation of coreference
 Summary

Misclassified Ambiguities I: Null elements

- 10 of the 29 ambiguous nuclei in sample are null elements varying between two different categories.
- WSJ annotators inserted markers for arguments and adjuncts realized non-locally, or unstated units of measurement (cf. Bies et al. 1995, p.59).
- Example: *EXP* (expletive) annotated as S or SBAR
 - ... it **[S*EXP*]** may be a wise business investment * [S to help "keep those jobs and sales taxes within city limits"] .
 - ... it **[SBAR*EXP*]** may be impossible [SBAR for the broker to carry out the order] because ...
- ⇒ Ambiguity arises where null items occur in place of element non-locally realized.

Misclassified Ambiguities I: Null elements
 Erman, Mueser, University of Tübingen
 Introduction
 What is Annotation Error
 How to obtain high quality
 Part of Speech
 Verbalizer
 Noun
 Verb
 Adjective
 Adverb
 Preposition
 Conjunction
 Participle
 Pronoun
 Interjection
 Negation
 Whitespace
 Dependency
 Relation of dependence
 Relation of modification
 Relation of coreference
 Summary

Misclassified Ambiguities I: Null elements

- Effect of eliminating variation detection for null elements
- remove null elements from set of variation nuclei of length 1
- resulting number of non-fringe distinct variation nuclei: 5584
- 78.9% of sample are errors, with 95% conf. interval (.7046, .8732):
 - between 3934 and 4875 erroneous variation nuclei, each corresponding to at least one error instance

Misclassified Ambiguities II: Coordination
 Erman, Mueser, University of Tübingen
 Introduction
 What is Annotation Error
 How to obtain high quality
 Part of Speech
 Verbalizer
 Noun
 Verb
 Adjective
 Adverb
 Preposition
 Conjunction
 Participle
 Pronoun
 Interjection
 Negation
 Whitespace
 Dependency
 Relation of dependence
 Relation of modification
 Relation of coreference
 Summary

Coordinate structure example
 Erman, Mueser, University of Tübingen
 Introduction
 What is Annotation Error
 How to obtain high quality
 Part of Speech
 Verbalizer
 Noun
 Verb
 Adjective
 Adverb
 Preposition
 Conjunction
 Participle
 Pronoun
 Interjection
 Negation
 Whitespace
 Dependency
 Relation of dependence
 Relation of modification
 Relation of coreference
 Summary

- 6 of the 29 ambiguities deal with coordinate structures.
- Annotation scheme distinguishes simple (i.e., non-modified) and complex coordinate elements.
 - Even if an element is simple, it is annotated like a complex element when conjoined with one.

- interest in a flat coordinate structure:

$$\begin{array}{c}
 \text{NP} \\
 \text{The amount covers taxes - interest and penalties owed} \\
 \text{DT NN VBZ NNS - NN CC and NNS VBN}
 \end{array}$$
- interest in a complex coordinate structure:

$$\begin{array}{c}
 \text{NP} \\
 \text{is lot of back taxes - interest and civil fraud penalties} \\
 \text{DT NN IN JJ JJ NNS - NN CC and JJ NN NN NNS}
 \end{array}$$
- ⇒ Annotation scheme makes a distinction externally motivated

- Related work on syntactic error detection
- CCGbank (Hockenmaier & Steedman 2005): derived from Penn Treebank, fixing some errors:
 - e.g.: "Under ADVP, if the adverb has only one child, and it is tagged as NNP, change it to RB."
- Blaheta (2002): discusses types of errors and some rules to identify them
 - e.g.: "If an IN is occurring somewhere other than a PP, it is likely to be a mistag."
- Ule & Simov (2004) search for unexpected rules, using information about a node and its mother
 - Discrepancies between mother and daughter annotation can point to errors

Summary for constituency error detection

▶ We showed how one can extend the POS-error detection approach to syntactic annotation.

▶ Illustrated with a case study based on WSJ treebank that the method is successful (71% precision) in detecting inconsistencies in syntactic category annotation.

▶ Approach supports two aspects of treebank improvement:

- makes it possible to find and correct erroneous variation in corpus annotation
- provides feedback for development of empirically adequate standards for syntactic annotation, identifying distinctions difficult to maintain over entire corpus

Detecting Errors in Corpus Annotation
 Error Message
 University of Tübingen

Introduction
 About an Annotation Error
 How to obtain high quality
 Part of Speech
 Variation detection
 Consistency (variation n-grams)
 Results for the WSJ
 Annotated sentence breakdown
 Consistency
 Variation detection
 Consistency (variation n-grams)
 WSJ case study
 Dependency
 Variation detection
 Consistency (variation n-grams)
 Results for TIGER
 Annotated word
 Summary

UNIVERSITÄT TÜBINGEN

Discontinuous constituents

▶ Discontinuous constituents (or equivalents) have been proposed in a wide range of syntactic frameworks, e.g.:

- Tree Adjoining Grammar (Koch & Joshi 1987; Rambow & Joshi 1994)
- Categorical Grammar (Dowty 1996; Hapelle 1994; Morrill 1995)
- linearization-based Head-Driven Phrase Structure Grammar (Reape 1993; Kathol 1995; Richter & Sailer 2001; Müller 1999; Penn 1999; Donohue & Sag 1999; Bonami et al. 1999)
- non-projective Dependency Grammar (Böker 1998; Platak et al. 2001)
- approaches positing tangled trees (McCawley 1982; Huck 1985; Ojeda 1987; Blavins 1990)

▶ They are also used in German treebanks (NEGRA, Skut et al. 1997, 1998; TIGER, Brants et al. 2002)

Some examples for discontinuous constituents

▶ An English extraposition example:
 (7) *The man came into the room who everybody loved.*

▶ An English particle verb example:
 (8) a. *I called up John.*
 b. *I called John up.*

▶ German extraposition example (Brants et al. 2002):
 (9) *Ein Mann kommt, der lacht*
 a man comes, who laughs
 'A man who laughs comes.'

Here **Ein Mann der lacht** is an NP constituent.

Treebanks and discontinuous constituents

▶ Treebanks which have been developed for languages with relatively free constituent order often represent discontinuous constituents (one way or another).

▶ For German, we take a closer look at:

- NEGRA Treebank (Skut et al. 1997, 1998)
 - ▶ written language: *Frankfurter Rundschau*, a national newspaper
 - ▶ 20,000 sentences (350,000 tokens)
 - ▶ flat structures as encoding of argument structure
- TIGER Treebank (Brants et al. 2002)
 - ▶ Extension of the NEGRA project
 - ▶ > 35,000 sentences (700,000 tokens)

The NEGRA/TIGER Treebanks

▶ annotation consists of tree structures with node and edge labels

▶ tree structure:

- encodes argument structure
- properties:
 - ▶ crossing branches used extensively
 - ▶ no empty terminal nodes
 - ▶ each daughter has one mother (but some secondary edges)

▶ node and edge labels encode:

- phrase level: syntactic categories
- lexical level: STTS part-of-speech (Schiller, Teufel & Thielen 1995)

An extraposition example (NEGRA corpus)

Error detection for discontinuous constituents

▶ The variation *n*-gram method relies on the assumption that a continuous string can be mapped to a category.

▶ Extend it to account for the fact that

- the variation nuclei, and
- their contexts

are no longer required to be continuous strings, and

▶ adapt the variation classification heuristics accordingly.

Adapting the algorithm to discontinuity

Error detection for syntactic annotation is broken down into runs for all constituent lengths ($1 \leq i \leq \text{longest-constituent}$):

▶ Constituent size includes only the tokens that are a part of the constituent, not possibly intervening material.

1. Compute the set of nuclei:
 - a) Find all constituents of size *i*; store them with their category label
 - b) For each type of string stored as constituent of length *i*, add each non-constituent occurrence with label NIL
2. The variation nuclei set is the set of all nuclei with more than one label
3. Generate variation *n*-grams for these variation nuclei.

Notes on Variation Nuclei

Discontinuous non-constituent occurrences

To find all strings that match a constituent in the corpus, we need to take discontinuous strings into account. The strings to be found may not be constituents:

(10) *in diesem Punkt seien sich Bonn und London on this point are self Bonn and London nicht einig not agreed.*
 'Bonn and London do not agree on this point.'

(11) *in diesem Punkt seien sich Bonn und London offensichtlich on this point are self Bonn and London clearly nicht einig not agreed.*

Here, **sich einig** is an AP in (10), but a discontinuous non-constituent (= NIL) in (11).

Detecting Errors in Corpus Annotation
 Error Message
 University of Tübingen

Introduction
 About an Annotation Error
 How to obtain high quality
 Part of Speech
 Variation detection
 Consistency (variation n-grams)
 Results for the WSJ
 Annotated sentence breakdown
 Consistency
 Variation detection
 Consistency (variation n-grams)
 WSJ case study
 Dependency
 Variation detection
 Consistency (variation n-grams)
 Results for TIGER
 Annotated word
 Summary

UNIVERSITÄT TÜBINGEN

Detecting Errors in Corpus Annotation
 Error Message
 University of Tübingen

Introduction
 About an Annotation Error
 How to obtain high quality
 Part of Speech
 Variation detection
 Consistency (variation n-grams)
 Results for the WSJ
 Annotated sentence breakdown
 Consistency
 Variation detection
 Consistency (variation n-grams)
 WSJ case study
 Dependency
 Variation detection
 Consistency (variation n-grams)
 Results for TIGER
 Annotated word
 Summary

UNIVERSITÄT TÜBINGEN

<h3>Notes on Variation Nuclei</h3> <p>Limiting the occurrence of non-constituents</p> <p>Constituents can overlap with non-constituent occurrences:</p> <p>(12) Ohne diese Ausgaben, so die Weltbank, seien without these expenses so the world bank are die Menschen totes Kapital the people dead capital "according to the worldbank, without these expenses the people are dead capital"</p> <ul style="list-style-type: none"> The string die Menschen occurs twice: <ul style="list-style-type: none"> once as a continuous constituent once as a discontinuous non-constituent <p>⇒ If a constituent overlaps with a non-constituent string, ignore the non-constituent string.</p>	<h3>Computing variation nuclei efficiently</h3> <p>Use tries for storage</p> <p>Task: Find all potentially discontinuous strings that match a string occurring as a constituent in the corpus.</p> <p>Determine a tractable domain for the search: Syntactic annotation: only consider strings within a sentence.</p> <p>How to do the search:</p> <ul style="list-style-type: none"> Inefficient generate-and-test method: <ol style="list-style-type: none"> Generate every (potential discontinuous) substring of a sentence ($= 2^n - 1$ cases for sentence length n) Test to see which ones match a constituent. Incremental method using a trie as a guide: <ol style="list-style-type: none"> Store all constituents in a trie with words as nodes Incrementally match every (potentially discontinuous) substring of a sentence with a path in the trie. <p>⇒ Incremental matching significantly reduces search space.</p>	<h3>Which contexts for discontinuous constituents</h3> <p>Idea: the more similar the context, the more likely variation in the annotation of a nucleus is an error.</p> <ul style="list-style-type: none"> Previously: expanded context to left and right Now: also expand into <i>internal context</i>, i.e., material contained within span of discontinuous constituent but not part of constituent itself <p>How to do it:</p> <ul style="list-style-type: none"> Incrementally add context adjacent to the nucleus. Why? The most local context helps the most with disambiguation. <p>⇒ Require surrounding context for every terminal element of the nucleus in order for it to be <i>non-fringe</i>.</p>
<h3>Results on the TIGER corpus</h3> <h4>The Setup</h4> <ul style="list-style-type: none"> Used TIGER treebank (Brants et al. 2002), a German newspaper corpus with 712.332 tokens in 40.020 sentences Evaluation of whether a detected variation points to an error was carried out by George Smith and Robert Langner of the TIGER project. 	<h3>Results on the TIGER corpus</h3> <p>Baseline, without context:</p> <ul style="list-style-type: none"> Method detects 10.964 variation nuclei. 13% pointed to at least one token error in sample of 100. (95% conf. interv.: 702 (6.4%) – 2.148 (19.6%) are errors) <p>Using word contexts (non-fringe nuclei):</p> <ul style="list-style-type: none"> Resulted in 500 shortest non-fringe variation nuclei, shortest non-fringe = rely solely on non-fringe heuristic 80% pointed to at least one token error in sample of 100. (95% conf. interv.: 361 (72.2%) – 439 (87.8%) are errors) Precision comparable to regular syntactic annotation (71% in Dickinson & Meurers 2003b) 	<h3>Increasing recall</h3> <p>The variation n-gram method for detecting annotation errors</p> <ul style="list-style-type: none"> Finds recurring data and compares analyses in different corpus instances Uses shared context as a heuristic to determine when analyses should be annotated identically <p>Two ways to increase recall:</p> <ul style="list-style-type: none"> Redefine <i>variation nuclei</i>, to extend the set of what counts as recurring data for which annotation is compared. Redefine <i>context and heuristics</i>, to obtain more variation n-grams predicted to be errors.
<h3>Approach explored</h3> <p>Using part-of-speech nuclei to increase recall</p> <ul style="list-style-type: none"> To increase the number of errors found, relax the requirements of what constitutes comparable strings <ul style="list-style-type: none"> Redefine <i>variation nuclei</i>: POS instead of words <p>Example (WSJ corpus, PennTreebank3 tagset, 45 tags):</p> <p>(13) a. <i>Boeing on Friday said 0</i> it received [_{NP} <u>an</u>DT <u>order</u>NN] "ICH" from Martinair Holl b. <i>it received</i> [_{NP} <u>at</u>DT <u>contract</u>NN "ICH"] from Timken Co.</p> <ul style="list-style-type: none"> But more general recurring units (POS tags) may negatively impact the precision of error detection. <ul style="list-style-type: none"> To use a more general representation, we also need more constraints on the disambiguating contexts. 	<h3>Approach explored</h3> <p>Identifying reliable contexts to maintain high precision</p> <ul style="list-style-type: none"> Example illustrating problem that shortest non-fringe heuristic does not ensure sufficient context: <p>(14) a. <i>crippled * by a bitter , decade-long strike that "T" began</i> [_{NP} <u>in</u>IN <u>1967</u>/CD] and cut circulation in half b. <i>its problems began</i> [_{NP} <u>in</u>IN [_{NP} <u>1967</u>/CD and early 1988] when its ...</p> <p>Here, the variation in structure is a correct ambiguity.</p> <ul style="list-style-type: none"> What treebank information can accurately distinguish erroneous variation from legitimate ambiguity? <p>→ We explore three new heuristics, based on annotation:</p> <ul style="list-style-type: none"> Heuristic 1: Shared complete bracketing Heuristic 2: Shared partial bracketing Heuristic 3: Shared vertical context 	<h3>Heuristic 1: Shared complete bracketing</h3> <p>Target 1: Variation with bracketing agreement, i.e., between nuclei which are constituents (XP vs. YP)</p> <ul style="list-style-type: none"> Both annotations agree on the bracketing → Significantly more likely that variation in label is an error <p>Ex.: RB JJ varies between NP and wrong ADJP in 4-gram:</p> <p>(15) a. This was [_{NP} <u>too</u>RB <u>much</u>/JJ] for James Oakes , the court's chief judge b. <i>Avondale was notified " by Louisiana officials in 1986 that it was</i> [_{NP} <u>powerfully</u>RB <u>responsible</u>/JJ] for a cleanup at an oil-recycling plant .</p> <p>Heuristic 1: Shared complete bracketing is comparable context</p>

Heuristic 2: Shared partial bracketing

Target 2: Variation between constituent and non-constituent

- (16) a. *crippled* * by a bitter, decade-long strike that *T* began [_{NP} *in/IN* 1967/CD] and cut circulation in half
- b. *its problems* began [_{PP} *in/IN*] [_{VP} 1987/CD and early 1988/] when its ...

- Legitimate attachment difference here because
 - "in 1967" forms a complete VP with began, but "in 1987" does not
 - one word of surrounding context is not sufficient to distinguish the two cases
- Can we define a heuristic to reduce risk of attachment ambiguities?

Detecting Errors in Corpus Annotation

Samuel Meurers
University of Tübingen

Introduction
What is Annotation Error
How to obtain high quality
Part of Speech
Verb
Verb-Modifier
Verb-Object
Verb-Complement
Verb-Adpositional
Verb-Prepositional
Verb-Particle
Verb-Infinitive
Verb-Subjunctive
Verb-Relative
Verb-Interjection
Verb-Other
Verb-None
Verb-Unknown

Constituency
Verb-Modifier
Verb-Object
Verb-Complement
Verb-Adpositional
Verb-Prepositional
Verb-Particle
Verb-Infinitive
Verb-Subjunctive
Verb-Relative
Verb-Interjection
Verb-Other
Verb-None
Verb-Unknown

Dependency
Verb-Modifier
Verb-Object
Verb-Complement
Verb-Adpositional
Verb-Prepositional
Verb-Particle
Verb-Infinitive
Verb-Subjunctive
Verb-Relative
Verb-Interjection
Verb-Other
Verb-None
Verb-Unknown

Summary

UNIVERSITÄT TüBINGEN

Heuristic 3: Shared partial bracketing

Heuristic 2: Require one extra word on side(s) without shared bracket.

Example: Erroneous variation for the variation nucleus VBG JJ NNS

- (17) *he stayed inside the Capitol*
 * [_{VP} *in/IN* monitoring/VBG tax-and-budget/JJ talks/NNS] instead of flying to San Francisco ...]
- (18) *one of the first bids under new takeover rules aimed **
 * [_{VP} *in/IN* encouraging/VBG open/LL bids/NNS] instead of gradual accumulation of large stakes] .

- The constituent and the *xx* string share a left (vp) bracket, but not a right one.
- Requiring extra word of context on right side (of supports that these cases are indeed comparable.

Heuristic 3: Shared vertical context

Target 3: Variation in bracketing of nucleus, but shared bracketing in n-gram

Example: erroneous variation for nucleus RB JJR IN CD:

- (19) a. *will be diluted **
 to [_{NP} [_{CP} *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 % .

- Considering n-gram with additional token (%) to right of nucleus provides shared complete bracketing (NP).

Heuristic 3: Shared bracketing of n-gram resulting from adding a word of context to left and/or right of nucleus.

Results for POS nuclei

After generalizing the nuclei from words to POS, we obtain

- 50,396 variation nuclei for the WSJ
- 16,598 of which remain after removing nuclei which are single null elements (cf. Dickinson & Meurers 2003b)
 - Significantly higher than the 3,619 comparable cases with word nuclei

To gauge performance of POS nuclei:

- Sampled 100 cases from 16,598 to examine by hand
- 28% point to an error
- 4,647 estimated cases of errors, which is a significant improvement in recall over 2,745 for word nuclei

Results with heuristics

- Heuristics select 6,343 variation nuclei of 16,598:
 - Heuristic 1 (Shared complete bracketing): 1,339
 - Heuristic 2 (Shared partial bracketing): 3,731
 - Heuristic 3 (Shared vertical context): 1,273
- Inspected random sample of 100 cases to judge precision:

Overall	68.69%
Heuristic 1	61%
Heuristic 2	61%
Heuristic 3	85%

- Estimate 4,357 errors from 6,343 cases, 59% increase in recall over estimated 2,745 errors with word nuclei
- 73 cases not covered by heuristics: 8.22% precision
- ⇒ New heuristics cover most cases, approaching high precision of word nucleus method while increasing recall.

Limitations of POS nuclei

- Generalizing from word to POS nuclei is not always successful, i.e., POS class not fine grained enough.
- Example: variation trigram "remains JJ for"

- (20) a. *a virus that *T* [_{VP} remains [_{AD,VP} active/LL] [_{VP} for a few days]]*
- b. *remains [_{AD,VP} responsible/LL] for the individual policy services department]*

- Depends upon particular adjective in determining how the for phrase attaches
- One could explore refining or lexicalizing some part-of-speech classes to account for such differences.

Alternatives ways to increase recall

- Use more general types of context (e.g., POS tags, Dickinson 2005; Dickinson & Meurers 2005b)
 - 8,715 shortest non-fringe variation nuclei, with an estimated 53% error detection precision
 - Could be combined with the POS nucleus approach using the new heuristics.
 - Impeadite dominance variation method (Dickinson & Meurers 2005c) based on RHSs of treabank rules
 - Overlaps with shared complete bracketing cases when RHS is complete sequence of POS tags
 - Mainly only handles errors stemming from variation in labeling and not bracketing errors
- Separate slides on exploring endocentricity

Summary for increasing recall of constituency error detection

- Increased error detection recall for syntactic annotation by generalizing nature of comparable recurring unit
- Generalized variation nuclei of variation n-gram method to POS tags instead of using identical surface forms
- Determined additional contextual heuristics for errors

Variation Detection for Dependency Annotation (Boyd, Dickinson & Meurers 2008)

- A range of high-quality dependency treebanks for a variety of different languages are available, e.g.:
 - Prague Dependency Treebank (PDT) of Czech (Hajič et al. 2001)
 - Alpino Dependency Treebank of Dutch (van der Beek et al. 2001)
 - Talbanken05 corpus of Swedish (Nivre et al. 2006)
 - Arboretum treebank for Danish (Blick 2003)
 - Danish Dependency Treebank (Kromann et al. 2004)
- Multi-lingual dependency parsing highlighted by 2006 CoNLL-X Shared Task
- As far as we are aware, little work has been done on automatically detecting errors in dependency treebanks.

Detecting Errors in Corpus Annotation

Samuel Meurers
University of Tübingen

Introduction
What is Annotation Error
How to obtain high quality
Part of Speech
Verb
Verb-Modifier
Verb-Object
Verb-Complement
Verb-Adpositional
Verb-Prepositional
Verb-Particle
Verb-Infinitive
Verb-Subjunctive
Verb-Relative
Verb-Interjection
Verb-Other
Verb-None
Verb-Unknown

Constituency
Verb-Modifier
Verb-Object
Verb-Complement
Verb-Adpositional
Verb-Prepositional
Verb-Particle
Verb-Infinitive
Verb-Subjunctive
Verb-Relative
Verb-Interjection
Verb-Other
Verb-None
Verb-Unknown

Dependency
Verb-Modifier
Verb-Object
Verb-Complement
Verb-Adpositional
Verb-Prepositional
Verb-Particle
Verb-Infinitive
Verb-Subjunctive
Verb-Relative
Verb-Interjection
Verb-Other
Verb-None
Verb-Unknown

Summary

UNIVERSITÄT TüBINGEN

Detecting Errors in Corpus Annotation

Samuel Meurers
University of Tübingen

Introduction
What is Annotation Error
How to obtain high quality
Part of Speech
Verb
Verb-Modifier
Verb-Object
Verb-Complement
Verb-Adpositional
Verb-Prepositional
Verb-Particle
Verb-Infinitive
Verb-Subjunctive
Verb-Relative
Verb-Interjection
Verb-Other
Verb-None
Verb-Unknown

Constituency
Verb-Modifier
Verb-Object
Verb-Complement
Verb-Adpositional
Verb-Prepositional
Verb-Particle
Verb-Infinitive
Verb-Subjunctive
Verb-Relative
Verb-Interjection
Verb-Other
Verb-None
Verb-Unknown

Dependency
Verb-Modifier
Verb-Object
Verb-Complement
Verb-Adpositional
Verb-Prepositional
Verb-Particle
Verb-Infinitive
Verb-Subjunctive
Verb-Relative
Verb-Interjection
Verb-Other
Verb-None
Verb-Unknown

Summary

UNIVERSITÄT TüBINGEN

Detecting Errors in Corpus Annotation

Samuel Meurers
University of Tübingen

Introduction
What is Annotation Error
How to obtain high quality
Part of Speech
Verb
Verb-Modifier
Verb-Object
Verb-Complement
Verb-Adpositional
Verb-Prepositional
Verb-Particle
Verb-Infinitive
Verb-Subjunctive
Verb-Relative
Verb-Interjection
Verb-Other
Verb-None
Verb-Unknown

Constituency
Verb-Modifier
Verb-Object
Verb-Complement
Verb-Adpositional
Verb-Prepositional
Verb-Particle
Verb-Infinitive
Verb-Subjunctive
Verb-Relative
Verb-Interjection
Verb-Other
Verb-None
Verb-Unknown

Dependency
Verb-Modifier
Verb-Object
Verb-Complement
Verb-Adpositional
Verb-Prepositional
Verb-Particle
Verb-Infinitive
Verb-Subjunctive
Verb-Relative
Verb-Interjection
Verb-Other
Verb-None
Verb-Unknown

Summary

UNIVERSITÄT TüBINGEN

Detecting Errors in Corpus Annotation

Samuel Meurers
University of Tübingen

Introduction
What is Annotation Error
How to obtain high quality
Part of Speech
Verb
Verb-Modifier
Verb-Object
Verb-Complement
Verb-Adpositional
Verb-Prepositional
Verb-Particle
Verb-Infinitive
Verb-Subjunctive
Verb-Relative
Verb-Interjection
Verb-Other
Verb-None
Verb-Unknown

Constituency
Verb-Modifier
Verb-Object
Verb-Complement
Verb-Adpositional
Verb-Prepositional
Verb-Particle
Verb-Infinitive
Verb-Subjunctive
Verb-Relative
Verb-Interjection
Verb-Other
Verb-None
Verb-Unknown

Dependency
Verb-Modifier
Verb-Object
Verb-Complement
Verb-Adpositional
Verb-Prepositional
Verb-Particle
Verb-Infinitive
Verb-Subjunctive
Verb-Relative
Verb-Interjection
Verb-Other
Verb-None
Verb-Unknown

Summary

UNIVERSITÄT TüBINGEN

Detecting Errors in Corpus Annotation

Samuel Meurers
University of Tübingen

Introduction
What is Annotation Error
How to obtain high quality
Part of Speech
Verb
Verb-Modifier
Verb-Object
Verb-Complement
Verb-Adpositional
Verb-Prepositional
Verb-Particle
Verb-Infinitive
Verb-Subjunctive
Verb-Relative
Verb-Interjection
Verb-Other
Verb-None
Verb-Unknown

Constituency
Verb-Modifier
Verb-Object
Verb-Complement
Verb-Adpositional
Verb-Prepositional
Verb-Particle
Verb-Infinitive
Verb-Subjunctive
Verb-Relative
Verb-Interjection
Verb-Other
Verb-None
Verb-Unknown

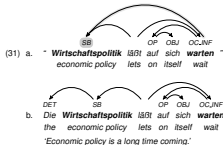
Dependency
Verb-Modifier
Verb-Object
Verb-Complement
Verb-Adpositional
Verb-Prepositional
Verb-Particle
Verb-Infinitive
Verb-Subjunctive
Verb-Relative
Verb-Interjection
Verb-Other
Verb-None
Verb-Unknown

Summary

UNIVERSITÄT TüBINGEN

Heuristic 1: NIL internal context heuristic

Example for case predicted to be an error



Detecting Errors in Corpus Annotation

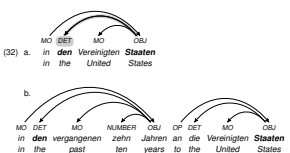
Center Module
University of Tübingen

Introduction
What is Annotation Error
How to detect high quality
Part of Speech
Verbler annotation
Constituency
Verbler annotation
Results
Summary

UNIVERSITÄT TÜBINGEN

Heuristic 1: NIL internal context heuristic

Example for case predicted not to be an error



Heuristic 2: Dependency context heuristic

- If the head of a variation nucleus is being used in the same function in all instances, the variation in the labeling of the nucleus is more likely to be an error.
- Conversely, when the head is used differently, it is more likely a genuine ambiguity.

Detecting Errors in Corpus Annotation

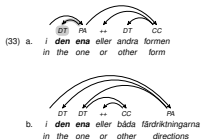
Center Module
University of Tübingen

Introduction
What is Annotation Error
How to detect high quality
Part of Speech
Verbler annotation
Constituency
Verbler annotation
Results
Summary

UNIVERSITÄT TÜBINGEN

Heuristic 2: Dependency context heuristic

Example: head with two different functions → not an error



Detecting Errors in Corpus Annotation

Center Module
University of Tübingen

Introduction
What is Annotation Error
How to detect high quality
Part of Speech
Verbler annotation
Constituency
Verbler annotation
Results
Summary

UNIVERSITÄT TÜBINGEN

Results

Talbanken 05

- 197,123 tokens in 11,431 sentences in sections P and G
- 210 different variation nuclei using non-fringe heuristic
- 92.9% precision (195 error nuclei) (thanks to Joakim Nivre, Mattias Nilsson and Eva Pettersson for the evaluation)
- 274 error instances:
 - 145 labelling confusion
 - 129 dependency identification
- observations:
 - common problems: determiner (DT), preposition (PA)
 - more errors with adverbials (73) than arguments (31)

Detecting Errors in Corpus Annotation

Center Module
University of Tübingen

Introduction
What is Annotation Error
How to detect high quality
Part of Speech
Verbler annotation
Constituency
Verbler annotation
Results
Summary

UNIVERSITÄT TÜBINGEN

Results

PDT 2.0

- 38,482 sentences (670,544 tokens) in full/amw section
- 553 different variation nuclei using non-fringe heuristic
 - 426 cases after removing errors involving punctuation
 - 354 cases after reocoding indirect AuxP and AuxC deps.
- 59.7% precision (205 error nuclei) (thanks to Jirka Hana and Jan Štěpánek for evaluation)
- 251 error instances
 - 152 labelling confusion (60.6%)
 - 99 dependency identification
- observations:
 - 49% of false positives due to other indirect annotation scheme decisions (coordination)
 - common problem with AdvAttr vs. AttrAdv, preference for adverbial of predicate vs. attribute of lower node

Detecting Errors in Corpus Annotation

Center Module
University of Tübingen

Introduction
What is Annotation Error
How to detect high quality
Part of Speech
Verbler annotation
Constituency
Verbler annotation
Results
Summary

UNIVERSITÄT TÜBINGEN

Results

TigerDB

- Only used sentences with lexically-rooted dependency structures, ignoring abstract and sublexical nodes.
- 1,567 sentences (29,373 tokens)
- 276 variation nuclei, NIL internal context heuristic
- 48.1% precision (133 error nuclei)
- 149 error instances
 - 46 labeling errors
 - 103 dependency identification
- observations:
 - consistent tokenization is a problem for multi-word expressions and proper names, e.g. *Den Haag* (The Hague), *zur Zeit* (at that time)
 - prepositional argument vs. modifier distinction difficult, e.g. *Bedarf an X* (demand for X).
 - false positives due to ambiguous tokens, for which POS disambiguation would help

Detecting Errors in Corpus Annotation

Center Module
University of Tübingen

Introduction
What is Annotation Error
How to detect high quality
Part of Speech
Verbler annotation
Constituency
Verbler annotation
Results
Summary

UNIVERSITÄT TÜBINGEN

Outlook: Increasing recall

The issue

- word-word dependencies are highly specific.
- How can they be generalized to increase the number of recurring dependency pairs limiting the recall of the variation detection method?

Specific lexical properties of head important, e.g.:

- Lexical information is known to improve PCFGs through head-lexicalization (e.g., Collins 1996)

To characterize the dependent, POS class may be sufficient (cf. subcategorization frame in lexicalized theories of grammar).

- Generalize from word-word to word-POS dependencies
 - For nuclei not annotated as a dependency (NIL), use head-dependent orientation of string we compare it to.

Detecting Errors in Corpus Annotation

Center Module
University of Tübingen

Introduction
What is Annotation Error
How to detect high quality
Part of Speech
Verbler annotation
Constituency
Verbler annotation
Results
Summary

UNIVERSITÄT TÜBINGEN

Outlook: Increasing recall

Tagset dependency

- Use of word-POS dependencies is dependent on the granularity of the POS tagset used.
 - Talbanken05 corpus has 40 coarse-grained POS tags
 - PDT 2.0 distinguishes 4290 POS tags (Hajič 2004)
- For positional tagsets, one can decide which positions of the tagset to use, e.g.,
 - including case information is likely to increase precision
 - distinguishing comparative and superlative adjectives could decrease recall

Detecting Errors in Corpus Annotation

Center Module
University of Tübingen

Introduction
What is Annotation Error
How to detect high quality
Part of Speech
Verbler annotation
Constituency
Verbler annotation
Results
Summary

UNIVERSITÄT TÜBINGEN

- We motivated the need for error detection in annotated corpora, and introduced the variation n-gram approach as an automatic error detection method.
- Research on category learning in humans provides independent evidence for the notion of context used.
- The method successfully detects errors in
 - part of speech
 - constituency,
 - discontinuous constituency,
 - and dependency annotation
- We showed that the method can provide significant feedback on annotation scheme distinctions which
 - are not sufficiently documented,
 - rely on representational choices not locally motivated,
 - or cannot be relied to be based on the evidence found in the corpus.

Chemia, E., T. H. Mints, S. Bernal & Christophe (2009). Categorizing words using 'treasure frames': what cross-linguistic analysis reveal about distributional acquisition strategies. *Developmental Science* 12(3). URL: <http://dx.doi.org/10.1111/j.1467-7687.2009.00825.x>

Collins, M. J. (1996). A New Statistical Parser Based on Bigram Lexical Dependency. In J. F. Allen & J. P. Fisiadis (eds.), *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*. San Francisco: Morgan Kaufmann Publishers, pp. 184–191. URL: <http://www.cis.upenn.edu/~collins96new.html>

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

Dickinson, M. & 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. 102–114. <http://linguistics.utoronto.edu/~dm/papers/dickinson-meurers-03.html>

Dickinson, M. & W. D. Meurers (2003b). Detecting Inconsistencies in Treebanks. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT-03)*. Växjö, Sweden, pp. 45–56. <http://linguistics.utoronto.edu/~dm/papers/dickinson-meurers-1033.html>

Dickinson, M. & W. D. Meurers (2004). Error detection with discontinuous constituents. In R. Rodrigues, D. Caver & J. Herring (eds.), *Proceedings of the First Midwest Computational Linguistics Colloquium*. Bloomington, Indiana.

Dickinson, M. & W. D. Meurers (2005a). Detecting Annotation Errors in Spoken Language Corpora. In *The Special Session on treebanks for spoken language and second- and non-adjacent contexts*. URL: <http://linguistics.utoronto.edu/~dm/papers/dickinson-meurers-nodal05.html>

King, T. H., R. Crouch, S. Riezler, M. Dainyem & R. M. Kaplan (2003). The PARC 700 Dependency Bank. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora*, held at the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03). Budapest, URL: <http://www2.parc.utoronto.ca/lingroups/itl75bank/>

Kroch, A. S. & A. K. Janda (1987). Analyzing Extrapolation in a Tree Adjoining Grammar. In Huck & Ojeda (1987)

Kromann, M. T., L. Mikkelson & S. L. Kunde (2004). Danish Dependency Treebank: Annotation Guide. <http://www.idi.uu.dk/~mkb/treebank/guide1.txt>

Kvitén, P. & K. Ojeda (2002). Achieving an Almost Correct POS-Tagged Corpus. In P. Fisiadis, L. Kaplan & K. Pata (eds.), *Text Speech and Dialogue 5th International Conference, TSD 2002*. Brest, Czech Republic, September 9–12, 2002. Heidelberg: Springer, no. 2448 in Lecture Notes in Artificial Intelligence (LNAI), pp. 19–26.

Leach, G. (1997). *A Brief Overview of Tools to the Grammatical Tagging of the British National Corpus*. UCLRE, Lancaster University, Lancaster. <http://www.hcu.ac.uk/BNC/whatgramtag.htm>

Luzius, W., H. Blasinger & C. Genstinger (2002). *TIGER/Analyzing*. MS, University of Tübingen.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates, third edition ed.

Marcus, M., B. Santorini & M. A. Marcinkiewicz (1999). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2), 313–330. <http://www.cis.upenn.edu/pub/treebank/doc/c093.ps.gz>

Marcus, M., B. Santorini, M. A. Marcinkiewicz & Taylor (1999). Treebank-3 Corpus. *Linguistic Data Consortium*, Philadelphia. <http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LD99942>

- Abella, A. (ed.) (2003). *Treebanks: Building and using syntactically annotated corpora*. Dordrecht: Kluwer Academic Publishers. <http://treebank.linguist.uva.nl/>
- Ahney, S., R. E. Schapire & Y. Singer (1999). Applying Applied to Tagging and PP Attachment. In P. Fung & J. Zhou (eds.), *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 30–45.
- Argente, R. & R. Schaure (1994). Fast Algorithms for Finding Association Rules in Large Databases. In J. B. Bocca, M. Jarke & C. Zaniolo (eds.), *VLDB 1994*. Morgan Kaufmann, pp. 487–499.
- Arstén, R. & M. Posio (2009). Survey Article: Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34(4), 1–42. URL: <http://www.cis.upenn.edu/~arsten/pubs/11.11.09.pdf>
- Bisk, E. (2003). Arboretum, a Hybrid Treebank for Danish. In *Proceedings of TLT 2003 (2nd Workshop on Treebanks and Linguistic Theory)*. Växjö, Sweden, pp. 9–20.
- Bies, A., M. Ferguson, K. Katz & R. MacIntyre (1995). Bracketing Guidelines for Treebank II Style Penn Treebank Project. University of Pennsylvania. <http://ftp.cis.upenn.edu/pub/treebank/doc/manual/inst.pdf>
- Blaheša, D. (2002). Handling noisy training and testing data. In *Proceedings of the 7th conference on Empirical Methods in Natural Language Processing*, pp. 111–116. <http://www.cs.cmu.edu/~dpb/papers/emnp-02nlp.html>

Dickinson, M. & W. D. Meurers (2005). Detecting Errors in Discontinuous Structural Annotation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL05)*, pp. 322–329. <http://www.aclweb.org/anthology/P/P05/P05-1040>

Dickinson, M. & W. D. Meurers (2005c). Puncte Disasas: Branches and Wh Why Matters. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*. Barcelona, Spain. URL: <http://linguistics.utoronto.edu/~dm/papers/mtl05.html>

Donohue, C. & A. I. Sag (1999). Domains in Warlpiri. In *Abstracts of the 51th Int. Conference on HPSG*. Edinburgh: University of Edinburgh, pp. 101–116. <http://www.ccl.utoronto.edu/~sag/papers/warlpiri.ps>

Down, D. C. (1998). Towards a Minimalist Theory of Syntactic Structure. In B. Burtt & A. van Hank (eds.), *Discontinuous Constituents*. New York, NY: Mouton de Gruyter, vol. 6 of *Natural Language Processing*

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, URL: <http://www.cs.columbia.edu/~eskin/papers/treebank-naacl00.ps>

Forst, M., N. Bertomeu, E. Crymawn, F. Fouvy, S. Hansen-Schäfer & V. Kordoni (2004). Towards a Dependency-Based Gold Standard for German Parsers. In *The TIGER Dependency Bank*, in S. Hansen-Schäfer, S. Opper & J. Kopeček (eds.), *5th International Workshop on Linguistically Interpreted Corpora (LINC-04) at COLING*. Geneva, Switzerland: COLING, pp. 31–38. URL: <http://icclweb.org/coling04/W04-1105/>

Hajčič, J. (2004). *Investigation of WPS-R Reflexion (Computational Morphology of Czech)*. Karolinum, Charles University Press, Prague, Czech Republic.

Hajčič, J. & A. Šimková, E. Hajičová & B. Vidová-Hadžiková (2003). The Prague Dependency Treebank: A Three-Level Annotation Scenario. In Abella (2003), chap. 7, pp. 103–127. URL: <http://rufl.net/~cunl.cdu/2003/publications/HajcHajcovaA2003.pdf>

Hajčič, J., B. Hajičová & P. Pajas (2001). The Prague Dependency Treebank: Annotation Structure and Support. In *IFCS Workshop on Linguistic Databases*.

Hopfle, M. (1994). Discontinuity and the Lambert Calculus. In *Proceedings of the 5th Conference on Computational Linguistics (COLING-94)*, Kyoto. <http://dcs.shu.ac.uk/home/people/papers/coling94.pdf>

Hirakawa, H., K. Ono & Y. Yoshimura (2000). Automatic Refinement of a POS Tagger Using a Reliable Parser and Plain Text Corpus. In *Proceedings of the 8th International Conference on Computational Linguistics (COLING)*. Saarbrücken, Germany: ICCL.

Hockenmaier, J. & M. Steedman (2003). *CCGbank: User's Manual*. Tech. Rep. MS-CIS-03-09, Department of Computer Science and Information Science, University of Pennsylvania.

Huck, G. (1985). Exclusivity and discontinuity in phrase structure grammar. In *West Coast Conference on Formal Linguistics (WCCFL)*. Stanford University, CSLI Publications, vol. 4, pp. 99–98.

Huck, G. & A. Ojeda (eds.) (1987). *Continuous Constituency*. No. 20 in *Syntax and Semantics*. Dordrecht: Reidel.

Johansson, S. (1986). *The Tagged LOB Corpus: Users' Manual*. Norwegian Computing Centre for the Humanities, Bergen.

Kathol, A. (1995). *Linearization-Based German Syntax*. Ph.D. thesis, Ohio State University, Columbus, OH. Revised version published by Oxford University Press.

McCawley, J. D. (1982). Parentheticals and discontinuous constituent structure. *Linguistic Inquiry* 13(1), 91–100.

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://linguistics.utoronto.edu/~dm/papers/meurers-03.html>

Mints, T. H. (2002). Category induction from distributional cues in an artificial language. *Memory & Cognition* 30, 678–686.

Mints, T. H. (2003). Frequent frames as a cue for grammatical categories in child second language. *Linguistics* 41(2), 91–117.

Morill, G. (1995). Discontinuity in categorial grammar. *Linguistics and Philosophy* 18, 175–219.

Müller, F. H. & U. Lieke (2002). Annotating topological fields and chunks – and how POS tags do the same thing. In *Proceedings of COLING*. <http://linguistics.utoronto.edu/~dm/papers/7956/muller.pdf>

Müller, S. (1999). *Deutsche Syntax deklarativ*. Hand-Driven Phrase Structure Grammar for the Datas Deutsche. No. 394 in *Linguistische Arbeiten*. Tübingen: Max Niemeyer Verlag.

Nivre, J., J. Nilsson & J. Hall (2006). *Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation*. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC2006)*. Genova, Italy.

Ojeda, A. (1987). Discontinuity, multidominances and unbounded dependency in Generalized Phrase Structure Grammar. In Huck & Ojeda (1987).

Olive, K. (2001). The Possibilities of Automatic Detection/Correction of Errors in Tagged Corpora: A Pilot Study on a German Corpus. In V. Matuszek,

Blewins, J. (1990). *Syntactic Complexity: Evidence for Discontinuity and Multidominance*. Ph.D. thesis, University of Massachusetts, Amherst, MA.

Bonami, D., G. Dodard & J.-M. Marandin (1999). Constituency and word order in French subject inversion. In G. Bouma, E. W. Honing, G. M. Kruijf & T. Ojeda (eds.), *Constraints and Resources in Natural Language Syntax and Semantics*. Stanford, CA: CSLI Publications, Studies in Constraint-Based Lexicology, pp. 21–40.

Boyd, A., M. Dickinson & C. Meurers (2007). Increasing the Recall of Corpus Annotation Error Detection. In *Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories (TLT-07)*. Bergen, Norway. URL: <http://purl.org/odp/papers/boyd-ai-07nlp.html>

Boyd, A., M. Dickinson, C. Meurers & J. Kopeček (2008). Detecting Errors in Dependency Treebanks. *Research in Language and Computation* 6(2), 113–137. URL: <http://purl.org/odp/papers/boyd-ai-08nlp.html>

Brants, S., S. Dipper, S. Hansen, W. Lezius & G. Smith (2002). The TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*. Saarbrücken, Germany: Saarland University, <http://colweb.org/p03.pdf>

Brants, T. & W. Skut (1998). Automation of Treebank Annotation. In *Proceedings of New Methods in Language Processing (NAMLP-98)*. Sydney. <http://www.col.ucl.ac.uk/~thbrants/papers/naamlp98.pdf>

Böker, N. (1998). Separating Surface Order and Syntactic Relations in a Dependency Grammar. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING) and the 38th Annual Meeting of the ACL*. Los Angeles, California.

Calder, J. (1997). On aligning trees. In *Proceedings of the Second Conference of Empirical Methods in Natural Language Processing*. Brown University. <http://xanati.lanl.gov/cmp-1p/9707016.html>

Dickinson, M. & W. D. Meurers (2005). Detecting Errors in Discontinuous Structural Annotation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL05)*, pp. 322–329. <http://www.aclweb.org/anthology/P/P05/P05-1040>

Dickinson, M. & W. D. Meurers (2005c). Puncte Disasas: Branches and Wh Why Matters. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*. Barcelona, Spain. URL: <http://linguistics.utoronto.edu/~dm/papers/mtl05.html>

Donohue, C. & A. I. Sag (1999). Domains in Warlpiri. In *Abstracts of the 51th Int. Conference on HPSG*. Edinburgh: University of Edinburgh, pp. 101–116. <http://www.ccl.utoronto.edu/~sag/papers/warlpiri.ps>

Down, D. C. (1998). Towards a Minimalist Theory of Syntactic Structure. In B. Burtt & A. van Hank (eds.), *Discontinuous Constituents*. New York, NY: Mouton de Gruyter, vol. 6 of *Natural Language Processing*

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, URL: <http://www.cs.columbia.edu/~eskin/papers/treebank-naacl00.ps>

Forst, M., N. Bertomeu, E. Crymawn, F. Fouvy, S. Hansen-Schäfer & V. Kordoni (2004). Towards a Dependency-Based Gold Standard for German Parsers. In *The TIGER Dependency Bank*, in S. Hansen-Schäfer, S. Opper & J. Kopeček (eds.), *5th International Workshop on Linguistically Interpreted Corpora (LINC-04) at COLING*. Geneva, Switzerland: COLING, pp. 31–38. URL: <http://icclweb.org/coling04/W04-1105/>

Hajčič, J. (2004). *Investigation of WPS-R Reflexion (Computational Morphology of Czech)*. Karolinum, Charles University Press, Prague, Czech Republic.

Hajčič, J. & A. Šimková, E. Hajičová & B. Vidová-Hadžiková (2003). The Prague Dependency Treebank: A Three-Level Annotation Scenario. In Abella (2003), chap. 7, pp. 103–127. URL: <http://rufl.net/~cunl.cdu/2003/publications/HajcHajcovaA2003.pdf>

Hajčič, J., B. Hajičová & P. Pajas (2001). The Prague Dependency Treebank: Annotation Structure and Support. In *IFCS Workshop on Linguistic Databases*.

Hopfle, M. (1994). Discontinuity and the Lambert Calculus. In *Proceedings of the 5th Conference on Computational Linguistics (COLING-94)*, Kyoto. <http://dcs.shu.ac.uk/home/people/papers/coling94.pdf>

Hirakawa, H., K. Ono & Y. Yoshimura (2000). Automatic Refinement of a POS Tagger Using a Reliable Parser and Plain Text Corpus. In *Proceedings of the 8th International Conference on Computational Linguistics (COLING)*. Saarbrücken, Germany: ICCL.

Hockenmaier, J. & M. Steedman (2003). *CCGbank: User's Manual*. Tech. Rep. MS-CIS-03-09, Department of Computer Science and Information Science, University of Pennsylvania.

Huck, G. (1985). Exclusivity and discontinuity in phrase structure grammar. In *West Coast Conference on Formal Linguistics (WCCFL)*. Stanford University, CSLI Publications, vol. 4, pp. 99–98.

Huck, G. & A. Ojeda (eds.) (1987). *Continuous Constituency*. No. 20 in *Syntax and Semantics*. Dordrecht: Reidel.

Johansson, S. (1986). *The Tagged LOB Corpus: Users' Manual*. Norwegian Computing Centre for the Humanities, Bergen.

Kathol, A. (1995). *Linearization-Based German Syntax*. Ph.D. thesis, Ohio State University, Columbus, OH. Revised version published by Oxford University Press.

Dickinson, M. & W. D. Meurers (2005). Detecting Errors in Discontinuous Structural Annotation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL05)*, pp. 322–329. <http://www.aclweb.org/anthology/P/P05/P05-1040>

Dickinson, M. & W. D. Meurers (2005c). Puncte Disasas: Branches and Wh Why Matters. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*. Barcelona, Spain. URL: <http://linguistics.utoronto.edu/~dm/papers/mtl05.html>

Donohue, C. & A. I. Sag (1999). Domains in Warlpiri. In *Abstracts of the 51th Int. Conference on HPSG*. Edinburgh: University of Edinburgh, pp. 101–116. <http://www.ccl.utoronto.edu/~sag/papers/warlpiri.ps>

Down, D. C. (1998). Towards a Minimalist Theory of Syntactic Structure. In B. Burtt & A. van Hank (eds.), *Discontinuous Constituents*. New York, NY: Mouton de Gruyter, vol. 6 of *Natural Language Processing*

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, URL: <http://www.cs.columbia.edu/~eskin/papers/treebank-naacl00.ps>

Forst, M., N. Bertomeu, E. Crymawn, F. Fouvy, S. Hansen-Schäfer & V. Kordoni (2004). Towards a Dependency-Based Gold Standard for German Parsers. In *The TIGER Dependency Bank*, in S. Hansen-Schäfer, S. Opper & J. Kopeček (eds.), *5th International Workshop on Linguistically Interpreted Corpora (LINC-04) at COLING*. Geneva, Switzerland: COLING, pp. 31–38. URL: <http://icclweb.org/coling04/W04-1105/>

Hajčič, J. (2004). *Investigation of WPS-R Reflexion (Computational Morphology of Czech)*. Karolinum, Charles University Press, Prague, Czech Republic.

Hajčič, J. & A. Šimková, E. Hajičová & B. Vidová-Hadžiková (2003). The Prague Dependency Treebank: A Three-Level Annotation Scenario. In Abella (2003), chap. 7, pp. 103–127. URL: <http://rufl.net/~cunl.cdu/2003/publications/HajcHajcovaA2003.pdf>

Hajčič, J., B. Hajičová & P. Pajas (2001). The Prague Dependency Treebank: Annotation Structure and Support. In *IFCS Workshop on Linguistic Databases*.

Hopfle, M. (1994). Discontinuity and the Lambert Calculus. In *Proceedings of the 5th Conference on Computational Linguistics (COLING-94)*, Kyoto. <http://dcs.shu.ac.uk/home/people/papers/coling94.pdf>

Hirakawa, H., K. Ono & Y. Yoshimura (2000). Automatic Refinement of a POS Tagger Using a Reliable Parser and Plain Text Corpus. In *Proceedings of the 8th International Conference on Computational Linguistics (COLING)*. Saarbrücken, Germany: ICCL.

Hockenmaier, J. & M. Steedman (2003). *CCGbank: User's Manual*. Tech. Rep. MS-CIS-03-09, Department of Computer Science and Information Science, University of Pennsylvania.

Huck, G. (1985). Exclusivity and discontinuity in phrase structure grammar. In *West Coast Conference on Formal Linguistics (WCCFL)*. Stanford University, CSLI Publications, vol. 4, pp. 99–98.

Huck, G. & A. Ojeda (eds.) (1987). *Continuous Constituency*. No. 20 in *Syntax and Semantics*. Dordrecht: Reidel.

Johansson, S. (1986). *The Tagged LOB Corpus: Users' Manual*. Norwegian Computing Centre for the Humanities, Bergen.

Kathol, A. (1995). *Linearization-Based German Syntax*. Ph.D. thesis, Ohio State University, Columbus, OH. Revised version published by Oxford University Press.

Dickinson, M. & W. D. Meurers (2005). Detecting Errors in Discontinuous Structural Annotation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL05)*, pp. 322–329. <http://www.aclweb.org/anthology/P/P05/P05-1040>

Dickinson, M. & W. D. Meurers (2005c). Puncte Disasas: Branches and Wh Why Matters. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*. Barcelona, Spain. URL: <http://linguistics.utoronto.edu/~dm/papers/mtl05.html>

Donohue, C. & A. I. Sag (1999). Domains in Warlpiri. In *Abstracts of the 51th Int. Conference on HPSG*. Edinburgh: University of Edinburgh, pp. 101–116. <http://www.ccl.utoronto.edu/~sag/papers/warlpiri.ps>

Down, D. C. (1998). Towards a Minimalist Theory of Syntactic Structure. In B. Burtt & A. van Hank (eds.), *Discontinuous Constituents*. New York, NY: Mouton de Gruyter, vol. 6 of *Natural Language Processing*

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, URL: <http://www.cs.columbia.edu/~eskin/papers/treebank-naacl00.ps>

Forst, M., N. Bertomeu, E. Crymawn, F. Fouvy, S. Hansen-Schäfer & V. Kordoni (2004). Towards a Dependency-Based Gold Standard for German Parsers. In *The TIGER Dependency Bank*, in S. Hansen-Schäfer, S. Opper & J. Kopeček (eds.), *5th International Workshop on Linguistically Interpreted Corpora (LINC-04) at COLING*. Geneva, Switzerland: COLING, pp. 31–38. URL: <http://icclweb.org/coling04/W04-1105/>

Hajčič, J. (2004). *Investigation of WPS-R Reflexion (Computational Morphology of Czech)*. Karolinum, Charles University Press, Prague, Czech Republic.

Hajčič, J. & A. Šimková, E. Hajičová & B. Vidová-Hadžiková (2003). The Prague Dependency Treebank: A Three-Level Annotation Scenario. In Abella (2003), chap. 7, pp. 103–127. URL: <http://rufl.net/~cunl.cdu/2003/publications/HajcHajcovaA2003.pdf>

Hajčič, J., B. Hajičová & P. Pajas (2001). The Prague Dependency Treebank: Annotation Structure and Support. In *IFCS Workshop on Linguistic Databases*.

Hopfle, M. (1994). Discontinuity and the Lambert Calculus. In *Proceedings of the 5th Conference on Computational Linguistics (COLING-94)*, Kyoto. <http://dcs.shu.ac.uk/home/people/papers/coling94.pdf>

Hirakawa, H., K. Ono & Y. Yoshimura (2000). Automatic Refinement of a POS Tagger Using a Reliable Parser and Plain Text Corpus. In *Proceedings of the 8th International Conference on Computational Linguistics (COLING)*. Saarbrücken, Germany: ICCL.

Hockenmaier, J. & M. Steedman (2003).

Santorini, B. (1990). Part-Of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision, 2nd printing). Ms., U.Penn.

Schiller, A., S. Teufel & C. Thielke (1995). Guidelines für das Taggen deutscher Textkorpora mit STTS. Tech. rep., IMS-CL, Univ. Stuttgart and SFS, Univ. Tübingen. <http://www.cogsci.ed.ac.uk/~simon/sfts.gu/95.ps.gz>.

Skut, W., T. Brants, B. Krenn & H. Uszkoreit (1998). A Linguistically Interpreted Corpus of German Newspaper Text. In Proceedings of the ESSLLI Workshop on Recent Advances in Corpus Annotation. Saarbrücken, Germany. <http://www.coli.uni-sb.de/~thorsten/publications/Skut-ss-ESSLLI-Corpus98.ps.gz>

Skut, W., B. Krenn, T. Brants & H. Uszkoreit (1997). An Annotation Scheme for Free Word Order Languages. In Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP), Washington, D.C. <http://www.coli.uni-sb.de/~thorsten/publications/Skut-ss-ANLP97.ps.gz>

Ulla, T. & 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/~ula/Paper/ula4lrec.pdf>.

van der Baek, L., G. Bouma, R. Malou & G. van Noord (2001). The Alpino Dependency Treebank. In Computational Linguistics in the Netherlands (CLIN) 2001, Amsterdam: Rodopi.

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

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

Correcting Errors in
Corpus Annotation

Simon Morice
University of Tübingen

Introduction

Part of Speech

Verbosity

Constituency

Dependency

Summary

University of Tübingen

30 / 35

Voutilainen, A. & T. Järvinen (1995). Specifying a shallow grammatical representation for parsing purposes. In Proceedings of the 7th Conference of the EACL, Dublin, Ireland. <http://www.aclweb.org/anthology/E95-1.029>.

Xiao, L., X. Cai & T. Lee (2006). The development of the verb category and verb argument structures in Mandarin-speaking children before two years of age. Paper presented at The Seventh Tokyo Conference on Psycholinguistics, Keio University.

Correcting Errors in
Corpus Annotation

Simon Morice
University of Tübingen

Introduction

Part of Speech

Verbosity

Constituency

Dependency

Summary

University of Tübingen

30 / 35