

Analyzing learner language – a computational linguistic perspective

Why analyze learner language? modeling what?
using which information sources?

Detmar Meurers
Universität Tübingen

Séminaire LPL
Aix-en-Provence, January 21, 2011

Analyzing learner language
Detmar Meurers

Introduction

Learner Corpora
Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case
NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language
Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion

Overview

- ▶ Motivations behind analyzing learner language and points of contact with computational linguistics
- ▶ A case study on parts-of-speech
 - ▶ sources of evidence
 - ▶ nature of categories
- ▶ Issues in linguistic analysis and error annotation
 - ▶ Target hypotheses
 - ▶ Inter-annotator agreement and available gold-standards
 - ▶ Relevance of the task and learner modeling

Analyzing learner language
Detmar Meurers

Introduction

Learner Corpora
Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case
NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language
Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion

Why Analyze Learner Language?

Second Language Acquisition (SLA)

- ▶ SLA research is aimed at understanding how second languages are acquired (and how language works)
- ▶ research on **instructed** SLA studies the effect of instructional interventions
 - ▶ targeting different aspects of language, and
 - ▶ supporting different kinds of feedback or interaction

Foreign Language Teaching (FLT)

- ▶ develop, adapt, advance, and test effectiveness of intervention methods from SLA for teaching practice
- ▶ design and use language tests

Analyzing learner language
Detmar Meurers

Introduction

Learner Corpora
Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case
NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language
Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion

Contact Points with Computational Linguistics

- ▶ **Learner corpora:** representation and annotation
 - ▶ can provide empirical evidence for SLA research
 - ▶ can provide insights into typical student needs in FLTannotation = off-line analysis
- ▶ **Writer's aid tools:** on-line analysis of learner language to provide immediate feedback *aimed at producing text*.
- ▶ **Intelligent Tutoring Systems:** on-line analysis
 - ▶ to provide immediate individualized feedback, e.g.:
 - ▶ meta-linguistic feedback in a form-focused activity
 - ▶ incidental focus-on-form in a meaning-based activity
 - ▶ feedback on meaning (very rare in ITS)
 - ▶ to determine progression through pedagogical material *aimed at supporting language acquisition*.

Analyzing learner language
Detmar Meurers

Introduction

Learner Corpora
Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case
NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language
Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion

Data in SLA research

An example: Clahsen & Muysken (1986)

- ▶ They studied word order acquisition in German by native speakers of Romance languages.
 - ▶ Stages of acquisition:
 1. S (Aux) V O
 2. (AdvP/PP) S (Aux) V O
 3. S V[+fin] O V[-fin]
 4. XP V[+fin] S O
 5. S V[+fin] (Adv) O
 6. *dass* S O V[+fin]
- Stage 2 example: *Früher ich kannte den Mann*
 earlier_{AdvP} I_S knew_V [the man]_O
- Stage 4 example: *Früher kannte ich den Mann*
 earlier_{AdvP} knew_{V[+fin]} I_S [the man]_O
- ▶ How is the data characterized?
 - ▶ lexical and syntactic categories and functions
 - ▶ some acquisition stages are well-formed, others ill-formed

Analyzing learner language
Detmar Meurers

Introduction

Learner Corpora

Data in SLA Research

Error annotation & beyond

Linguistic Annotation

Annotation Quality

A Concrete Case

NOCE Corpus

Automatic POS-Tagging

Three Sources of Evidence

Mismatching Evidence

Categories for Learner Language

Systematic POS for Learner Language

On the nature of interlanguage categories

Comparative fallacy

Error annotation

Target hypotheses

Importance of activity and learner modeling

Task-specific learner corpora

Conclusion

EBERHARD KARLS UNIVERSITÄT TUBINGEN

5 / 32

Annotation: Error Annotation and Beyond

- ▶ The annotation of learner corpora has focused on errors made by the learners (cf., e.g., Granger 2003; Díaz-Negrillo & Fernández-Domínguez 2006).
- ▶ Yet, SLA research essentially observes correlations of linguistic properties, whether erroneous or not.
- ▶ Even research focusing on learner errors needs to identify correlations with linguistic properties, e.g., to determine
 - ▶ overuse/underuse of particular patterns
 - ▶ measures of language development
 - ▶ CAF (Complexity, Accuracy, Fluency), Developmental Sentence Scoring, Index of Productive Syntax, ...

Analyzing learner language
Detmar Meurers

Introduction

Learner Corpora

Data in SLA Research

Error annotation & beyond

Linguistic Annotation

Annotation Quality

A Concrete Case

NOCE Corpus

Automatic POS-Tagging

Three Sources of Evidence

Mismatching Evidence

Categories for Learner Language

Systematic POS for Learner Language

On the nature of interlanguage categories

Comparative fallacy

Error annotation

Target hypotheses

Importance of activity and learner modeling

Task-specific learner corpora

Conclusion

EBERHARD KARLS UNIVERSITÄT TUBINGEN

6 / 32

Annotation of Linguistic Properties

- ▶ Annotation schemes have been developed for a wide range of linguistic properties, including
 - ▶ part-of-speech and morphology
 - ▶ syntactic constituency or lexical dependency structures
 - ▶ semantics (word senses, coreference), discourse structure
- ▶ Each type of annotation typically requires an extensive manual annotation effort → gold standard corpora
- ▶ Automatic annotation tools learning from such gold standard annotation are becoming available, but
 - ▶ quality of automatic annotation drops significantly for text differing from the gold standard training material
- ▶ Interdisciplinary collaboration between SLA and CL crucial to adapt annotation schemes and methods to learner language
 - ▶ Surprisingly little research on this (Meunier 1998; de Haan 2000; de Mönnink 2000; van Rooy & Schäfer 2002, 2003).

Analyzing learner language
Detmar Meurers

Introduction

Learner Corpora

Data in SLA Research

Error annotation & beyond

Linguistic Annotation

Annotation Quality

A Concrete Case

NOCE Corpus

Automatic POS-Tagging

Three Sources of Evidence

Mismatching Evidence

Categories for Learner Language

Systematic POS for Learner Language

On the nature of interlanguage categories

Comparative fallacy

Error annotation

Target hypotheses

Importance of activity and learner modeling

Task-specific learner corpora

Conclusion

EBERHARD KARLS UNIVERSITÄT TUBINGEN

7 / 32

Annotation quality

- ▶ An annotation scheme is only as good as the distinctions it reliably supports making based on available evidence.
 - ▶ E.g., particle vs. preposition dropped in Penn Treebank tagset since often not enough evidence available.
 - ▶ Note: More classes can be more reliable if they are more coherent (cf. CLAWS7 annotation, followed by mapping to CLAWS5 in BNC Tag Enhancement Project).
- ▶ How can high quality gold standards be obtained?
 - ▶ Keep only reliably and consistently identifiable distinctions, described in detailed manual, including appendix on hard cases (Voutilainen & Järvinen 1995; Sampson & Babarczy 2003)
 - ▶ Annotate corpus several times and independently, then test interannotator agreement (Brants & Skut 1998)
 - ▶ Detection of annotation errors through automatic analysis of comparable data recurring in the corpus → DECCA (Dickinson & Meurers 2003a,b, 2005; Boyd et al. 2008)

Analyzing learner language
Detmar Meurers

Introduction

Learner Corpora

Data in SLA Research

Error annotation & beyond

Linguistic Annotation

Annotation Quality

A Concrete Case

NOCE Corpus

Automatic POS-Tagging

Three Sources of Evidence

Mismatching Evidence

Categories for Learner Language

Systematic POS for Learner Language

On the nature of interlanguage categories

Comparative fallacy

Error annotation

Target hypotheses

Importance of activity and learner modeling

Task-specific learner corpora

Conclusion

EBERHARD KARLS UNIVERSITÄT TUBINGEN

8 / 32

A Case Study in Linguistic Annotation of Learner Language

- ▶ The NOCE learner corpus (Díaz-Negrillo 2009)
- ▶ POS analysis of learner language
 - ▶ Exploring automatic POS annotation
 - ▶ What does it mean to POS-annotate learner language?
- ▶ Reporting on joint work:
 - ▶ Díaz-Negrillo, Meurers, Valera & Wunsch (2010)

Analyzing learner language
Detmar Meurers

Introduction

Learner Corpora
Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case

NOCE Corpus

Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language
Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion

EBERHARD KARLS UNIVERSITÄT TUBINGEN

9 / 32

The NOCE Learner Corpus (Díaz-Negrillo 2009)

- ▶ Participants
 - ▶ Writing by 1st and 2nd year students of English at the universities of Granada and Jaén
 - ▶ Corpus includes meta-information on learner: age, level, L2 exposure, motivation, etc.
- ▶ Task
 - ▶ Written text, around 250 words
 - ▶ Topics chosen from 3 suggestions or free writing
- ▶ Corpus structure and size
 - ▶ 3 text collections per academic year, for 4 years
 - ▶ 998 texts, 337.332 tokens (149.256 types)
- ▶ Annotation:
 - ▶ Editorial (struck-out units, insertions, reordering)
 - ▶ Error (179 texts, 39.165 tokens, 5.285 errors, 357 types)

⇒ How about adding linguistic information?

Analyzing learner language
Detmar Meurers

Introduction

Learner Corpora
Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case

NOCE Corpus

Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language
Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion

EBERHARD KARLS UNIVERSITÄT TUBINGEN

10 / 32

Automatic POS-Tagging of NOCE

Setup

- ▶ Used 3 POS taggers trained on WSJ newspaper text, using Penn Treebank tagset
 - ▶ TreeTagger, TnT tagger, Stanford tagger
- ▶ Tagged the error-annotated section in NOCE

Results

- ▶ Manually evaluated POS tags assigned by taggers to 10 texts by 10 different participants (1.850 words)
 - ▶ Evaluation performed by Ana Díaz Negrillo
- ▶ Accuracy of automatically assigned tags
 - ▶ TreeTagger: 94.95%
 - ▶ TnT Tagger: 94.03%
 - ▶ Stanford Tagger: 88.11%

Analyzing learner language
Detmar Meurers

Introduction

Learner Corpora
Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case

NOCE Corpus

Automatic POS-Tagging

Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language
Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion

EBERHARD KARLS UNIVERSITÄT TUBINGEN

11 / 32

Aspects of a qualitative analysis

We found lower performance for expressions which do not exist in English (cf. also de Haan 2000; van Rooy & Schäfer 2002).

Spelling

(1) *I think that university **teachs** to people [. . .]*

Word boundaries

(2) *They can't pay their studies and **more over** they have to pay a flat [. . .]*

- ▶ But is tagging learner language really just a robustness issue, like adapting taggers to another domain?
- ▶ What does it mean to use POS tags developed for native language for the interlanguage of learners?
- ▶ What research questions can such “native POS” tags answer?

Analyzing learner language
Detmar Meurers

Introduction

Learner Corpora
Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case

NOCE Corpus

Automatic POS-Tagging

Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language
Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion

EBERHARD KARLS UNIVERSITÄT TUBINGEN

12 / 32

Three Sources of Evidence for POS analysis

Lemma/Lexical entry: *of* ⇒ preposition

(3) *I was surprised by the word **of** the day.*

Morphology: *-ion* ⇒ noun

(4) *There is a lot of **construction** going on here.*

Distribution: *adj __ verb* ⇒ noun

(5) *The old **man** left.*

Analyzing learner language
Detmar Meurers

Introduction

Learner Corpora
Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case
NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence

Mismatching Evidence

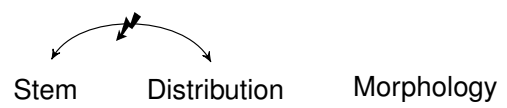
Categories for Learner Language
Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion

EBERHARD KARLS UNIVERSITÄT TUBINGEN

13 / 32

Case 1: Stem-Distribution mismatch



(6) *[...] you can find a big **vary** of beautiful beaches [...]*

Stem	Distribution	Morphology
verb	noun	?

(7) *[...] they are very kind and **friendship**.*

Stem	Distribution	Morphology
noun	adjective	?

Analyzing learner language
Detmar Meurers

Introduction

Learner Corpora
Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case
NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence

Mismatching Evidence

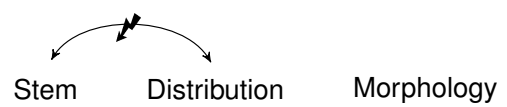
Categories for Learner Language
Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion

EBERHARD KARLS UNIVERSITÄT TUBINGEN

14 / 32

Case 1: Stem-Distribution mismatch



(8) *[...] that's the reason **because** I went to Tunisia twice.*

Stem	Distribution	Morphology
conjunction	wh-pronoun	?

(9) *RED helped him **during** he was in the prison.*

Stem	Distribution	Morphology
preposition	conjunction	?

Analyzing learner language
Detmar Meurers

Introduction

Learner Corpora
Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case
NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence

Mismatching Evidence

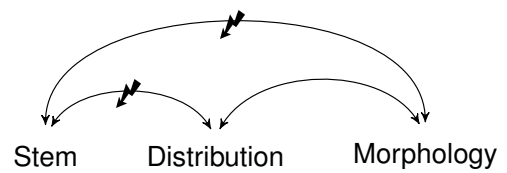
Categories for Learner Language
Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion

EBERHARD KARLS UNIVERSITÄT TUBINGEN

15 / 32

Case 2: Stem-Distrib./Stem-Morph. mismatch



(10) *[...] one of the favourite places to visit for many **foreigns**.*

Stem	Distribution	Morphology
adjective	noun	noun / verb 3 rd sg

(11) *[...] to be **choiced** for a job [...]*

Stem	Distribution	Morphology
noun / adjective	verb	verb

Analyzing learner language
Detmar Meurers

Introduction

Learner Corpora
Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case
NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence

Mismatching Evidence

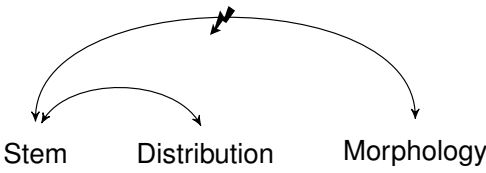
Categories for Learner Language
Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion

EBERHARD KARLS UNIVERSITÄT TUBINGEN

16 / 32

Case 3: Stem-Morphology mismatch



(12) [...] this film is one of the **bests** ever customes [...]

Stem	Distribution	Morphology
adjective (noun / verb)	adjective	noun / verb 3 rd sg

(13) [...] television, radio are very **subjectives** [...]

Stem	Distribution	Morphology
adjective / noun	adjective	noun / verb 3 rd sg

Analyzing learner language
Detmar Meurers

Introduction

Learner Corpora
Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case
NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

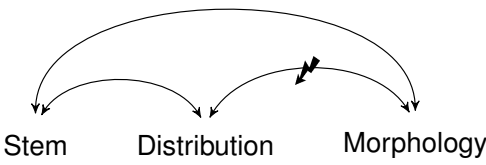
Categories for Learner Language
Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion

EBERHARD KARLS UNIVERSITÄT TUBINGEN

17/32

Case 4: Distribution-Morphology mismatch



(14) [...] for almost every **jobs** nowadays [...]

Stem	Distribution	Morphology
noun	noun sg	noun pl / verb 3 rd sg

(15) [...] it has **grew up** a lot specially after 1996 [...]

Stem	Distribution	Morphology
verb	verb past participle	verb past tense

Analyzing learner language
Detmar Meurers

Introduction

Learner Corpora
Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case
NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

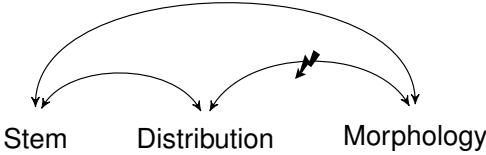
Categories for Learner Language
Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion

EBERHARD KARLS UNIVERSITÄT TUBINGEN

18/32

Case 4: Distribution-Morphology mismatch



(16) [...] if he **want** to know this [...]

(17) This first year **have** been wonderful [...]

Stem	Distribution	Morphology
verb	verb 3 rd person sg	verb non-3 rd sg

Analyzing learner language
Detmar Meurers

Introduction

Learner Corpora
Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case
NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language
Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion

EBERHARD KARLS UNIVERSITÄT TUBINGEN

19/32

Systematic POS for Learner Language

- ▶ A single, standard POS tag fails to systematically identify properties of learner language.
 - ▶ Mismatching evidence from distribution, stem, morphology
- ▶ Alternative: The annotation schemes for learner language should encode minimal observations.
 - ▶ POS: tripartite encoding of distribution, stem, morphology
- ▶ Some errors in learner language are **epiphenomena of mismatches** in linguistic encoding.
 - Identify such errors through linguistic annotation, not error annotation.

Analyzing learner language
Detmar Meurers

Introduction

Learner Corpora
Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case
NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language
Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion

EBERHARD KARLS UNIVERSITÄT TUBINGEN

20/32

On the nature of categories for learner language

- ▶ Where do the categories abstracted to come from?
- ▶ Categories result from generalizations, which require a significant amount of comparable data to be made.
- ▶ How fine grained are they?
 - ▶ In NLP, *robustness* is the ability to *ignore* variation in the realization of a category to be identified.
 - ▶ Whenever we want to identify and analyze a distinction, the category system used must be more fine grained than the variation we look for.
- ▶ Example: breaking down constituency in terms of
 - ▶ overall topology of a sentence (Hirschmann et al. 2007)
 - ▶ chunks and chunk-internal word order (Abney 1997)
 - ▶ dependency
 - ▶ canonical, as interface to meaning (MacWhinney 2008; Rosén & Smedt 2010; Ott & Ziai 2010; Hirschmann et al. 2010)
 - ▶ not standardized (Dickinson & Ragheb 2009)

Analyzing learner language
Detmar Meurers

Introduction

Learner Corpora
Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case
NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language
Systematic POS for Learner Language

On the nature of interlanguage categories

Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion

EBERHARD KARLS UNIVERSITÄT TUBINGEN

21 / 32

Comparative fallacy

- ▶ Comparative fallacy is “the mistake of studying the systematic character of one language by comparing it to another.” (Bley-Vroman 1983, p. 6)
 - ▶ extended to include bias towards native language (Lakshmanan & Selinker 2001)
- ▶ Essentially trying to analyze a “non-canonical variety” using a “robust” version of the canonical grammar.
 - ▶ divergences from norm annotated as errors
 - ▶ but: the research question is the issue here, not corpus error annotation as such (Tenfjord et al. 2006)
- ▶ Issue more general than language acquisition research:
 - ▶ Eurocentrism in field work, e.g., Gil (2001)
 - ▶ Variationist sociolinguistics:
 - ▶ Importance of defining variation to be studied and when an instance is counted as one of the variants.

Analyzing learner language
Detmar Meurers

Introduction

Learner Corpora
Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case
NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language
Systematic POS for Learner Language

On the nature of interlanguage categories

Comparative fallacy

Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion

EBERHARD KARLS UNIVERSITÄT TUBINGEN

22 / 32

Error annotation

- ▶ Error annotation involves (implicitly or explicitly):
 - a) Determining what the learner wanted to say (**target**).
 - b) Identifying
 - i. the **location** of the error, and
 - ii. the **nature** of the error corresponding to the difference between the learner sentence and the target hypothesis.
 - c) **Annotating** the error in the corpus
- ▶ Each of these steps can present ambiguity:
 - a) multiple possible target hypotheses
 - b)
 - i. different locations in which the error can be rooted
 - ii. different types of errors divergence can be attributed to
 - c) different ways to mark an error location & type in corpus

Analyzing learner language
Detmar Meurers

Introduction

Learner Corpora
Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case
NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language
Systematic POS for Learner Language

On the nature of interlanguage categories

Comparative fallacy

Error annotation

Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion

EBERHARD KARLS UNIVERSITÄT TUBINGEN

23 / 32

Error annotation schemes: Desiderata

Inter-annotator agreement

- ▶ An annotation is only relevant and useful if it provides a uniform, reliable index to relevant classes of data.
- ▶ Traditionally every researcher develops their own error annotation scheme. (cf. Díaz-Negrillo & Fernández-Domínguez 2006)
- ▶ Alarmingly, there does not seem to be a single study showing what level of inter-annotator agreement can be reached for which type of distinctions.

Analyzing learner language
Detmar Meurers

Introduction

Learner Corpora
Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case
NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language
Systematic POS for Learner Language

On the nature of interlanguage categories

Comparative fallacy

Error annotation

Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion

EBERHARD KARLS UNIVERSITÄT TUBINGEN

24 / 32

Error annotation schemes: Desiderata

Gold standard annotation

- ▶ Freely available gold standard annotations for error annotation schemes supporting high inter-annotator agreement levels are crucially needed.
- ▶ Without an available gold standard annotation,
 - ▶ no reliable quantitative evaluation possible for research
 - ▶ no training, evaluation and comparison of NLP tools for error analysis is possible.
- ▶ Promising progress for some subclasses (det, prep). (e.g., Lee & Seneff 2006; Tetreault & Chodorow 2008; De Felice 2008)
 - ▶ but it is important to establish a tool-independent, transparent definition of the markables to be annotated.

Target hypotheses

- ▶ Anke Lüdeling has argued for making the target hypothesis an explicit part of error annotation (Lüdeling et al. 2005; Hirschmann et al. 2007; Lüdeling 2008).
- ▶ Fitzpatrick & Seegmiller (2004) report unsatisfactory levels of agreement in determining the learner targets.
 - ▶ Keeping the target hypothesis implicit results in error annotation which diverge even more unsatisfactorily.
- ▶ If target hypothesis is explicit, at least the second step from target hypothesis to error tag might be realizable with high reliability.

Difficulty of determining target hypotheses

- ▶ What are the target forms for the sentences taken from the Hiroshima English Learners' Corpus (Miura 1998):

(18) *I didn't know*
(19) *I don't know his lives.*
(20) *I know where he lives.*
(21) *I know he lived*

They are taken from a translation task, for the Japanese of
(22) *I don't know where he lives.*
- ▶ How can one obtain a better handle on target hypotheses?
 - ▶ focus on more advanced learners
 - ▶ take explicit task context into account
 - ▶ support targets other than fully explicit surface forms
 - ▶ take more learner strategies into account
 - ▶ Learners sometimes use known L2 chunks instead of trying to express appropriate meaning!

Constraining the search space of interpretation

Importance of activity and learner modeling

- ▶ All approaches to modeling errors, such as
 - ▶ *mal*-rules
 - ▶ constraint relaxation
 - ▶ statistical modelingmust model the space of **well-formed and ill-formed variation** that is possible given
 - ▶ a particular activity, and
 - ▶ a given learner.
- ▶ For example, without task and speaker context, how would you interpret the following?

(23) *I will not buy this record it is scratched*

(24) *My hovercraft is full of eels.*

Exemplifying interpretation in context

Analyzing learner language
Detmar Meurers

Introduction


Learner Corpora
Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case
NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language
Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling

Task-specific learner corpora

Conclusion



29 / 32

Monty Python: Hungarian Phrase Book sketch
http://www.youtube.com/watch?v=akbflkF_1zY

Towards task-specific learner corpora

- ▶ Explicit task and learner models included as meta-information in a corpus can provide crucial constraining information for interpreting learner language.
 - ▶ E.g., it's easier to infer what a learner wanted to say if one knows the text they are answering questions about.
 - ▶ Related to taking task, strategic competence, and L1 into account in learner models of Intelligent Tutoring Systems (Amaral & Meurers 2008).
- ▶ Most current learner language corpora consist of essays, yet learners produce language in a wide range of contexts, naturalistic or instructed, e.g.,
 - ▶ email and chat messages
 - ▶ answering reading or listening comprehension questions
 - ▶ asking questions in information gap activities
- ▶ To obtain corpora which are interpretable and representative of learner language, we need language from a variety of contexts, including longitudinal data.

Analyzing learner language
Detmar Meurers

Introduction


Learner Corpora
Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case
NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language
Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling

Task-specific learner corpora

Conclusion



30 / 32

Conclusion

- ▶ We discussed the different motivations for analyzing learner language in SLA, FLT, and their connection to CL
- ▶ We motivated linguistic annotation to support effective querying for SLA patterns and discussed an approach to the POS analysis of learner language separating
 - ▶ lexical, morphological, and distributional information

Goal: Corpus annotation systematically characterizing language (native-like as well as learner innovations).
- ▶ Turning to error annotation, we argued for inter-annotator agreement as crucial for establishing which distinctions are replicable based on the available information.
- ▶ We explored the nature of target hypotheses and argued for explicit task and learner modeling to constrain the search space of interpretation.
- ▶ Well-defined analysis subtasks on widely available corpora are needed for sustained progress.

Analyzing learner language
Detmar Meurers

Introduction


Learner Corpora
Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case
NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language
Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling

Task-specific learner corpora

Conclusion



31 / 32

Our Background

Analyzing language for learners

- ▶ Input enhancement of texts for learners (Meurers et al. 2010b)
- ▶ Search engine for language learners (Ott & Meurers 2010)
- ▶ Prediction of functional elements (Elghafari, Meurers & Wunsch 2010)

Analyzing learner language

- ▶ Intelligent Tutoring System TAGARELA for Portuguese (Amaral & Meurers 2008, 2009, 2011; Amaral et al. 2011)
- ▶ Linguistic analysis of NOCE corpus of English written by Spanish learners (Díaz-Negrillo, Meurers, Valera & Wunsch 2010)
- ▶ Automatic analysis of learner language (Meurers 2009)
- ▶ Word order errors (Metcalfe & Meurers 2006b; Boyd & Meurers 2008)
- ▶ Content assessment of answers to reading comprehension questions (Bailey & Meurers 2008) → SFB 833 A4 (CoMIC)
 - ▶ Longitudinal corpus collection using WELCOME (Meurers, Ott & Ziai 2010a) → KU/OSU collaboration
 - ▶ Dependency parsing of learner language (Ott & Ziai 2010)

Analyzing learner language
Detmar Meurers

Introduction


Learner Corpora
Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case
NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language
Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling

Task-specific learner corpora

Conclusion



32 / 32

References

- Abney, S. (1997). Partial Parsing via Finite-State Cascades. *Natural Language Engineering* 2, 337–344. URL <http://www.vinartus.net/spa/97a.pdf>.
- Amaral, L., V. Metcalf & D. Meurers (2006). Language Awareness through Re-use of NLP Technology. Pre-conference Workshop on NLP in CALL – Computational and Linguistic Challenges. CALICO 2006. May 17, 2006. University of Hawaii. URL <http://purl.org/net/icall/handouts/calico06-amaral-metcalf-meurers.pdf>.
- Amaral, L. & D. Meurers (2008). From Recording Linguistic Competence to Supporting Inferences about Language Acquisition in Context: Extending the Conceptualization of Student Models for Intelligent Computer-Assisted Language Learning. *Computer-Assisted Language Learning* 21(4), 323–338. URL <http://purl.org/dm/papers/amaral-meurers-call08.html>.
- Amaral, L. & D. Meurers (2009). Little Things With Big Effects: On the Identification and Interpretation of Tokens for Error Diagnosis in ICALL. *CALICO Journal* 27(1). URL <http://purl.org/dm/papers/amaral-meurers-09.html>.
- Amaral, L. & D. Meurers (2011). On Using Intelligent Computer-Assisted Language Learning in Real-Life Foreign Language Teaching and Learning. *ReCALL* 23(1). URL <http://purl.org/dm/papers/amaral-meurers-10.html>.
- Amaral, L., D. Meurers & R. Ziai (2011). Analyzing Learner Language: Towards A Flexible NLP Architecture for Intelligent Language Tutors. *Computer-Assisted Language Learning* URL <http://purl.org/dm/papers/amaral-meurers-ziai-10.html>.

- Díaz-Negrillo, A. (2009). *EARS: A User's Manual*. Munich, Germany: LINCOM Academic Reference Books.
- 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. 107–114. URL <http://purl.org/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. URL <http://purl.org/dm/papers/dickinson-meurers-tlt03.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 (ACL'05)*. pp. 322–329. URL <http://aclweb.org/anthology/P05-1040>.
- Dickinson, M. & M. Ragheb (2009). Dependency Annotation for Learner Corpora. In *Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories (TLT-8)*. Milan, Italy. URL <http://jones.ling.indiana.edu/~mdickinson/papers/dickinson-ragheb09.html>.
- Díaz-Negrillo, A. & J. Fernández-Domínguez (2006). Error Tagging Systems for Learner Corpora. *Revista Española de Lingüística Aplicada (RESLA)* 19, 83–102. URL http://dialnet.unirioja.es/servlet/fichero_articulo?codigo=2198610&orden=72810.
- Díaz-Negrillo, A., D. Meurers, S. Valera & H. Wunsch (2010). Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum* URL <http://purl.org/dm/papers/diaz-negrillo-et-al-09.html>.

Analyzing learner language

Detmar Meurers

Introduction

Learner Corpora

Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case

NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language

Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion



32 / 32

Analyzing learner language

Detmar Meurers

Introduction

Learner Corpora

Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case

NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language

Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion



32 / 32

- Bailey, S. & D. Meurers (2008). Diagnosing meaning errors in short answers to reading comprehension questions. In J. Tetreault, J. Burstein & R. D. Felice (eds.), *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications (BEA-3) at ACL'08*. Columbus, Ohio, pp. 107–115. URL <http://aclweb.org/anthology/W08-0913>.
- Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning* 33(1), 1–17.
- Boyd, A., M. Dickinson & D. Meurers (2008). On Detecting Errors in Dependency Treebanks. *Research on Language and Computation* 6(2), 113–137. URL <http://purl.org/dm/papers/boyd-et-al-08.html>.
- Boyd, A. & D. Meurers (2008). On Diagnosing Word Order Errors. Poster presented at the CALICO Pre-Conference Workshop on Automatic Analysis of Learner Language. URL <http://purl.org/net/calico-workshop-abstracts.html#6>.
- Brants, T. & W. Skut (1998). Automation of Treebank Annotation. In *Proceedings of New Methods in Language Processing*. Sydney, Australia. URL <http://wing.comp.nus.edu.sg/acl/W/W98/W98-1207.pdf>.
- Clahsen, H. & P. Muysken (1986). The availability of Universal Grammar to adult and child learners: A study of the acquisition of German word order. *Second Language Acquisition* 2, 93–19. URL <http://slr.sagepub.com/cgi/reprint/2/2/93.pdf>.
- De Felice, R. (2008). Automatic Error Detection in Non-native English. Ph.D. thesis, St Catherine's College, University of Oxford.
- de Haan, P. (2000). Tagging non-native English with the TOSCA-ICLE tagger. In Mair & Hundt (2000), pp. 69–79.
- de Mönnink, I. (2000). Parsing a learner corpus. In Mair & Hundt (2000), pp. 81–90.

- Elghafari, A., D. Meurers & H. Wunsch (2010). Exploring the Data-Driven Prediction of Prepositions in English. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*. Beijing, China, pp. 267–275. URL <http://aclweb.org/anthology/C10-2031>.
- Fitzpatrick, E. & M. S. Seegmiller (2004). The Montclair electronic language database project. In U. Connor & T. Upton (eds.), *Applied Corpus Linguistics: A Multidimensional Perspective*, Amsterdam: Rodopi. URL <http://chss.montclair.edu/linguistics/MELD/rodopipaper.pdf>.
- Gil, D. (2001). Escaping Eurocentrism: Fieldwork as a Process of Unlearning. In P. Newman & M. Ratliff (eds.), *Linguistic Fieldwork*, Cambridge University Press, pp. 102–132.
- Granger, S. (2003). Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal* 20(3), 465–480. URL <http://purl.org/calico/granger03.pdf>.
- Hirschmann, H., S. Doolittle & A. Lüdeling (2007). Syntactic annotation of non-canonical linguistic structures. In *Proceedings of Corpus Linguistics 2007*. Birmingham. URL <http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/neu2/mitarbeiter-innen/anke/pdf/HirschmannDoolittleLuedelingCL2007.pdf>.
- Hirschmann, H., A. Lüdeling, I. Rehbein, M. Reznicek & A. Zeldes (2010). Syntactic Overuse and Underuse: A Study of a Parsed Learner Corpus and its Target Hypothesis. Presentation given at the Treebanks and Linguistic Theory Workshop.
- Lakshmanan, U. & L. Selinker (2001). Analysing interlanguage: how do we know what learners know? *Second Language Research* 17(4), 393–420. URL <http://proxy.lib.ohio-state.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=utf&AN=7393417&site=ehost-live>.

Analyzing learner language

Detmar Meurers

Introduction

Learner Corpora

Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case

NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language

Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion



32 / 32

Analyzing learner language

Detmar Meurers

Introduction

Learner Corpora

Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case

NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language

Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion



32 / 32

Lee, J. & S. Seneff (2006). Automatic Grammar Correction for Second-Language Learners. In *INTERSPEECH 2006 – ICSLP*. URL <http://groups.csail.mit.edu/sls/publications/2006/IS061299.pdf>.

Lüdeling, A. (2008). Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In P. Grommes & M. Walter (eds.), *Fortgeschrittene Lernervarietäten*, Tübingen: Niemeyer, pp. 119–140.

Lüdeling, A., M. Walter, E. Kroymann & P. Adolphs (2005). Multi-level error annotation in learner corpora. In *Proceedings of Corpus Linguistics*. Birmingham. URL <http://www.corpus.bham.ac.uk/PCLC/Falko-CL2006.doc>.

MacWhinney, B. (2008). Enriching CHILDES for morphosyntactic analysis. In H. Behrens (ed.), *Corpora in Language Acquisition Research: History, Methods, Perspectives*, Amsterdam and Philadelphia: John Benjamins, vol. 6 of *Trends in Language Acquisition Research*, pp. 165–197. URL <http://childes.psy.cmu.edu/grasp/morphosyntax.doc>.

Mair, C. & M. Hundt (eds.) (2000). *Corpus Linguistics and Linguistic Theory*. Amsterdam: Rodopi.

Metcalfe, V. & D. Meurers (2006a). Generating Web-based English Preposition Exercises from Real-World Texts. URL <http://purl.org/net/icall/handouts/eurocall06-metcalfe-meurers.pdf>. EUROCALL 2006. Granada, Spain. September 4–7, 2006.

Metcalfe, V. & D. Meurers (2006b). When to Use Deep Processing and When Not To – The Example of Word Order Errors. URL <http://purl.org/net/icall/handouts/calico06-metcalfe-meurers.pdf>. Pre-conference Workshop on NLP in CALL – Computational and Linguistic Challenges. CALICO 2006. May 17, 2006. University of Hawaii.

Rosén, V. & K. D. Smedt (2010). Syntactic Annotation of Learner Corpora. In H. Johansen, A. Golden, J. E. Hagen & A.-K. Helland (eds.), *Systematisk, variert, men ikke tilfeldig. Antologi om norsk som andrespråk i anledning Kari Tenfjords 60-årsdag [Systematic, varied, but not arbitrary. Anthology about Norwegian as a second language on the occasion of Kari Tenfjord's 60th birthday]*, Oslo: Novus forlag, pp. 120–132.

Sampson, G. & A. Babarczy (2003). Limits to annotation precision. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*. pp. 61–68. URL <http://www.sfs.uni-tuebingen.de/~zinsmeis/AnnotCorp05/materials/sampson-barbarczy03.pdf>.

Tenfjord, K., J. E. Hagen & H. Johansen (2006). The Hows and Whys of coding categories in a learner corpus (or "How and Why an error-tagged learner corpus is not ipso facto one big comparative fallacy"). *Rivista di psicolinguistica applicata* 6, 93–108.

Tetreault, J. & M. Chodorow (2008). The Ups and Downs of Preposition Error Detection in ESL Writing. In *Proceedings of COLING-08*. Manchester, UK. URL <http://www.ets.org/Media/Research/pdf/r3.pdf>.

van Rooy, B. & L. Schäfer (2002). The Effect of Learner Errors on POS Tag Errors during Automatic POS Tagging. *Southern African Linguistics and Applied Language Studies* 20, 325–335.

van Rooy, B. & L. Schäfer (2003). An Evaluation of Three POS Taggers for the Tagging of the Tswana Learner English Corpus. In D. Archer, P. Rayson, A. Wilson & T. McEnery (eds.), *Proceedings of the Corpus Linguistics 2003 conference Lancaster University (UK), 28 – 31 March 2003*. vol. 16 of *University Centre For Computer Corpus Research On Language Technical Papers*, pp. 835–844.

Analyzing learner language

Detmar Meurers

Introduction

Learner Corpora

Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case

NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language

Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion

ERHARD KARLS
UNIVERSITÄT
TÜBINGEN

32 / 32

Meunier, F. (1998). Computer Tools for Interlanguage Analysis: A Critical Approach. In G. Sylviane (ed.), *Learner English on Computer*, London and New York: Addison Wesley Longman, pp. 19–37.

Meurers, D. (2009). On the Automatic Analysis of Learner Language: Introduction to the Special Issue. *CALICO Journal* 26(3), 469–473. URL <http://purl.org/dm/papers/meurers-09.html>.

Meurers, D., N. Ott & R. Ziai (2010a). Compiling a Task-Based Corpus for the Analysis of Learner Language in Context. In *Proceedings of Linguistic Evidence*. Tübingen, pp. 214–217. URL <http://purl.org/dm/papers/meurers-ott-ziai-10.html>.

Meurers, D., R. Ziai, L. Amaral, A. Boyd, A. Dimitrov, V. Metcalf & N. Ott (2010b). Enhancing Authentic Web Pages for Language Learners. In *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-5) at NAACL-HLT 2010*. Los Angeles: Association for Computational Linguistics. URL <http://purl.org/dm/papers/meurers-ziai-et-al-10.html>.

Miura, S. (1998). Hiroshima English Learners' Corpus: English learner No. 2 (English I & English II). URL <http://home.hiroshima-u.ac.jp/d052121/eigo2.html>. Last Modified 14 May, 1998.

Ott, N. & D. Meurers (2010). Information Retrieval for Education: Making Search Engines Language Aware. *Themes in Science and Technology Education. Special issue on computer-aided language analysis, teaching and learning: Approaches, perspectives and applications* URL <http://purl.org/dm/papers/ott-meurers-10.html>.

Ott, N. & R. Ziai (2010). Evaluating Dependency Parsing Performance on German Learner Language. In *Proceedings of TLT9*. NEALT Proceeding Series. URL <http://www.sfs.uni-tuebingen.de/~rziai/papers/Ott.Ziai-10.pdf>. TO APPEAR.

Voutilainen, A. & T. Järvinen (1995). Specifying a shallow grammatical representation for parsing purposes. In *Proceedings of the 7th Conference of the EACL*. Dublin, Ireland. URL http://portal.acm.org/ft_gateway.cfm?id=977003&type=pdf&coll=GUIDE&dl=GUIDE&CFID=47108142&CFTOKEN=71182750.

Analyzing learner language

Detmar Meurers

Introduction

Learner Corpora

Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case

NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language

Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion

ERHARD KARLS
UNIVERSITÄT
TÜBINGEN

32 / 32

Analyzing learner language

Detmar Meurers

Introduction

Learner Corpora

Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case

NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language

Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion

ERHARD KARLS
UNIVERSITÄT
TÜBINGEN

32 / 32

Analyzing learner language

Detmar Meurers

Introduction

Learner Corpora

Data in SLA Research
Error annotation & beyond
Linguistic Annotation
Annotation Quality

A Concrete Case

NOCE Corpus
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence

Categories for Learner Language

Systematic POS for Learner Language
On the nature of interlanguage categories
Comparative fallacy
Error annotation
Target hypotheses
Importance of activity and learner modeling
Task-specific learner corpora

Conclusion

ERHARD KARLS
UNIVERSITÄT
TÜBINGEN

32 / 32