

Towards Linguistic Annotation of Interlanguage

Conceptual and practical issues in POS analyzing the NOCE corpus

Detmar Meurers, Holger Wunsch

based on joint research with
Luiz Amaral, Ana Díaz-Negrillo, Salvador Valera; cf. also:

Díaz-Negrillo/Meurers/Valera/Wunsch (2010): "Towards interlanguage POS annotation for effective learner corpora in SLA and FLT". Language Forum.
<http://purl.org/dm/papers/diaz-negrillo-et-al-09.html>

Berlin, October 16, 2009

Towards Linguistic Annotation of Interlanguage
Detmar Meurers
Holger Wunsch

Learner Corpora
Why they're useful
On compiling learner corpora
Why annotate corpora
Data in SLA research
Error annotation & beyond

Annotation quality
Why it's important
DECCA: Variation n-gram error detection

A Concrete Case
NOCE Corpus
Linguistic Information
Tokenization
POS-Tagging
Automatic POS-Tagging

Analyzing learner language
Sources of Evidence
Mismatching Evidence
Mismatch-free errors

Conclusion

Roadmap

- ▶ Which role can learner corpora play in Foreign Language Teaching & Second Language Acquisition (SLA) research?
- ▶ Why is linguistic annotation relevant?
- ▶ How can high quality annotation be obtained?
- ▶ Corpus Representation: A Concrete Case
 - ▶ The NOCE (NON-native Corpus of English) learner corpus
 - ▶ Towards linguistic annotation of NOCE
- ▶ Analyzing learner language:
 - ▶ sources of evidence for POS annotation
 - ▶ mismatches in combining evidence

Towards Linguistic Annotation of Interlanguage
Detmar Meurers
Holger Wunsch

Learner Corpora
Why they're useful
On compiling learner corpora
Why annotate corpora
Data in SLA research
Error annotation & beyond

Annotation quality
Why it's important
DECCA: Variation n-gram error detection

A Concrete Case
NOCE Corpus
Linguistic Information
Tokenization
POS-Tagging
Automatic POS-Tagging

Analyzing learner language
Sources of Evidence
Mismatching Evidence
Mismatch-free errors

Conclusion

Learner Corpora

- ▶ Learner corpora can serve
 - ▶ as a teaching resource for Foreign Language Teaching materials design,
 - ▶ provide insights into typical student needs, and
 - ▶ contribute an empirical basis for theories of Second Language Acquisition.
- ▶ Depending on the corpus composition, it can support **qualitative** and **quantitative** analysis of examples found

Towards Linguistic Annotation of Interlanguage
Detmar Meurers
Holger Wunsch

Learner Corpora
Why they're useful
On compiling learner corpora
Why annotate corpora
Data in SLA research
Error annotation & beyond

Annotation quality
Why it's important
DECCA: Variation n-gram error detection

A Concrete Case
NOCE Corpus
Linguistic Information
Tokenization
POS-Tagging
Automatic POS-Tagging

Analyzing learner language
Sources of Evidence
Mismatching Evidence
Mismatch-free errors

Conclusion

On compiling learner corpora

- ▶ Many 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 representative of learner language, it is important to include language produced in a variety of contexts, ideally also including longitudinal data.
- ▶ Including explicit task contexts in the meta-information of a corpus can also provide constraining information useful 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.

Towards Linguistic Annotation of Interlanguage
Detmar Meurers
Holger Wunsch

Learner Corpora
Why they're useful
On compiling learner corpora
Why annotate corpora
Data in SLA research
Error annotation & beyond

Annotation quality
Why it's important
DECCA: Variation n-gram error detection

A Concrete Case
NOCE Corpus
Linguistic Information
Tokenization
POS-Tagging
Automatic POS-Tagging

Analyzing learner language
Sources of Evidence
Mismatching Evidence
Mismatch-free errors

Conclusion

Annotation of Learner Corpora

- ▶ Effective querying of corpora for specific phenomena often requires reference to corpus annotation.
- ▶ To find relevant classes of examples, the terminology used to single out learner language aspects of interest needs to be mapped to instances in the corpus (Meurers 2005; Meurers & Müller 2009).
- ▶ Annotations function as an index to classes of data which cannot easily be identified in the surface form.

Towards Linguistic Annotation of Interlanguage
Dietmar Meurers
Hilger Wurch

Learner Corpora
Why they're useful
On compiling learner corpora
Why annotate corpora
Data in SLA research
Error annotation & beyond

Annotation quality
Why it's important
DECCA: Variation n-gram
error detection

A Concrete Case
NOCE Corpus
Linguistic Information
"Taxonomic"
POS-Tagging
Automatic POS-Tagging

Analyzing learner language
Sources of Evidence
Mismatching Evidence
Mismatch-free errors

Conclusion

Annotation of Learner Corpora (cont.)

- ▶ Example: Finding all sentences containing modal verbs using only the surface forms is possible, but involves specifying a long list of all forms of all modal verbs.
 - ▶ Even so, sentences where *can* is not actually a modal would be wrongly identified:
 - (1) *Pass me a **can** of beer.*
 - (2) *I **can** tuna for a living.*
- ▶ Many search patterns cannot be specified in finite form, e.g. finding all sentences with past participle verbs.
- ▶ What type of learner language annotations are needed to support the searches for the data which are important for FLT and SLA research?

Towards Linguistic Annotation of Interlanguage
Dietmar Meurers
Hilger Wurch

Learner Corpora
Why they're useful
On compiling learner corpora
Why annotate corpora
Data in SLA research
Error annotation & beyond

Annotation quality
Why it's important
DECCA: Variation n-gram
error detection

A Concrete Case
NOCE Corpus
Linguistic Information
"Taxonomic"
POS-Tagging
Automatic POS-Tagging

Analyzing learner language
Sources of Evidence
Mismatching Evidence
Mismatch-free errors

Conclusion

Data in SLA research

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

Towards Linguistic Annotation of Interlanguage
Dietmar Meurers
Hilger Wurch

Learner Corpora
Why they're useful
On compiling learner corpora
Why annotate corpora
Data in SLA research
Error annotation & beyond

Annotation quality
Why it's important
DECCA: Variation n-gram
error detection

A Concrete Case
NOCE Corpus
Linguistic Information
"Taxonomic"
POS-Tagging
Automatic POS-Tagging

Analyzing learner language
Sources of Evidence
Mismatching Evidence
Mismatch-free errors

Conclusion

Data in SLA research

Kanno (1997), Pérez-Lerroux & Glass (1997)

- ▶ They studied the use of overt and null pronouns by non-native speakers of Japanese and Spanish.
- ▶ Examples:
 - (3) *Nadie dice que **él** ganará el premio.*
nobody says that he will win the prize
'Nobody_i says that he_{i/jj} will win the prize.'
 - (4) *Nadie dice que **__** ganará el premio.*
nobody says that pro will win the prize
'Nobody_i says that he_{i/jj} will win the prize.'
- ▶ How is the data characterized?
 - ▶ syntactic functions and semantic relations
 - ▶ not overtly expressed but interpreted elements

Towards Linguistic Annotation of Interlanguage
Dietmar Meurers
Hilger Wurch

Learner Corpora
Why they're useful
On compiling learner corpora
Why annotate corpora
Data in SLA research
Error annotation & beyond

Annotation quality
Why it's important
DECCA: Variation n-gram
error detection

A Concrete Case
NOCE Corpus
Linguistic Information
"Taxonomic"
POS-Tagging
Automatic POS-Tagging

Analyzing learner language
Sources of Evidence
Mismatching Evidence
Mismatch-free errors

Conclusion

Annotation: Error annotation and beyond

- ▶ The annotation of learner corpora has focused on *errors* made by the learners (e.g., Granger 2003; Díaz-Negrillo & Fernández-Domínguez 2006; Lüdeling et al. 2008).
- ▶ 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 identify
 - ▶ overuse/underuse of certain patterns
 - ▶ measures of language development (Developmental Sentence Scoring, Index of Productive Syntax, ...)

Towards Linguistic Annotation of Interlanguage
Dexter Meurers
Hilger Wurch

Learner Corpora
Why they're useful
On compiling learner corpora
Why erroneous corpora
Data in SLA research
Error annotation & beyond

Annotation quality
Why it's important
DECCA: Variation n-gram error detection

A Concrete Case
NOCE Corpus
Linguistic Information
Tokenization
POS-Tagging
Automatic POS-Tagging

Analyzing learner language
Sources of Evidence
Mismatching Evidence
Mismatch-free errors

Conclusion

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 FLT, SLA and Computational Linguistics crucial to **adapt annotation schemes and methods** to learner language corpora
 - ▶ Very little research on this so far (but cf. Meunier 1998; de Haan 2000; de Mönnink 2000; van Rooy & Schäfer 2002, 2003)

Towards Linguistic Annotation of Interlanguage
Dexter Meurers
Hilger Wurch

Learner Corpora
Why they're useful
On compiling learner corpora
Why erroneous corpora
Data in SLA research
Error annotation & beyond

Annotation quality
Why it's important
DECCA: Variation n-gram error detection

A Concrete Case
NOCE Corpus
Linguistic Information
Tokenization
POS-Tagging
Automatic POS-Tagging

Analyzing learner language
Sources of Evidence
Mismatching Evidence
Mismatch-free errors

Conclusion

The importance of high-quality annotation

Precision of search

- ▶ 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:
 - ▶ Term used for 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 (and as known from Zipf's curse, most things occur rarely).

Towards Linguistic Annotation of Interlanguage
Dexter Meurers
Hilger Wurch

Learner Corpora
Why they're useful
On compiling learner corpora
Why erroneous corpora
Data in SLA research
Error annotation & beyond

Annotation quality
Why it's important
DECCA: Variation n-gram error detection

A Concrete Case
NOCE Corpus
Linguistic Information
Tokenization
POS-Tagging
Automatic POS-Tagging

Analyzing learner language
Sources of Evidence
Mismatching Evidence
Mismatch-free errors

Conclusion

The importance of high-quality annotation

Recall of search

- ▶ 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 are useful, but danger of partial blindness if example subclasses are not captured by query approximating target phenomenon.
 - ▶ for **quantitative** analysis: Maximizing recall is crucial for reliable quantitative results.
- ⇒ Where the query characterizing the target phenomenon is expressed in terms of the annotation, quality and consistency of the annotation is important.

Towards Linguistic Annotation of Interlanguage
Dexter Meurers
Hilger Wurch

Learner Corpora
Why they're useful
On compiling learner corpora
Why erroneous corpora
Data in SLA research
Error annotation & beyond

Annotation quality
Why it's important
DECCA: Variation n-gram error detection

A Concrete Case
NOCE Corpus
Linguistic Information
Tokenization
POS-Tagging
Automatic POS-Tagging

Analyzing learner language
Sources of Evidence
Mismatching Evidence
Mismatch-free errors

Conclusion

Annotation quality

Methods for obtaining quality

- ▶ How can a high quality gold standard be obtained?
 - ▶ Annotate corpus several times and independently, then test interannotator agreement (Brants & Skut 1998)
 - ▶ Keep only reliably and consistently identifiable distinctions, described in detailed manual, including appendix on hard cases (Voutilainen & Järvinen 1995; Sampson & Barbaczy 2003)
 - ▶ Detection of annotation errors through automatic analysis of comparable data recurring in the corpus → DECCA (Dickinson & Meurers 2003a,b, 2005; Boyd et al. 2008)

Towards Linguistic Annotation of Interlanguage
 Dettmer Meurers
 Higer Wurach

Learner Corpora
 Why they're useful
 On compiling learner corpora
 Why annotate corpora
 Data in SLA research
 Error annotation & beyond

Annotation quality
 Why it's important
 DECCA: Variation n-gram error detection

A Concrete Case
 NOCE Corpus
 Linguistic Information
 Tokenization
 POS-Tagging
 Automatic POS-Tagging

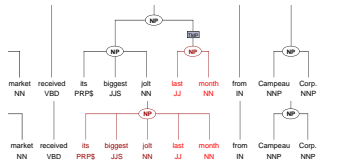
Analyzing learner language
 Sources of Evidence
 Mismatching Evidence
 Mismatch-free errors

Conclusion

DECCA: Variation n-gram error detection

- ▶ **Variation:** multiple occurrences, with different annotations
 - ambiguity:** different annotations correctly label the same material used in different contexts
 - annotation error:** annotation is inconsistent across comparable occurrences

▶ Variation between constituent and non-constituent:



Towards Linguistic Annotation of Interlanguage
 Dettmer Meurers
 Higer Wurach

Learner Corpora
 Why they're useful
 On compiling learner corpora
 Why annotate corpora
 Data in SLA research
 Error annotation & beyond

Annotation quality
 Why it's important
 DECCA: Variation n-gram error detection

A Concrete Case
 NOCE Corpus
 Linguistic Information
 Tokenization
 POS-Tagging
 Automatic POS-Tagging

Analyzing learner language
 Sources of Evidence
 Mismatching Evidence
 Mismatch-free errors

Conclusion

DECCA: Variation n-gram error detection (cont.)

- ▶ Variation between two syntactic category labels:
 - (5) *maturity* next Tuesday
 - labeled as **NP** twice
PP once
- ▶ Efficient methods for detecting such annotation errors have been developed for a range of annotation types (Dickinson & Meurers 2003a,b, 2005; Boyd et al. 2008):
 - ▶ positional: words, part-of-speech
 - ▶ binary relations: lexical dependencies
 - ▶ structural domains: chunks, constituents
- ▶ Python code is freely available from our project website: <http://decca.osu.edu>

Towards Linguistic Annotation of Interlanguage
 Dettmer Meurers
 Higer Wurach

Learner Corpora
 Why they're useful
 On compiling learner corpora
 Why annotate corpora
 Data in SLA research
 Error annotation & beyond

Annotation quality
 Why it's important
 DECCA: Variation n-gram error detection

A Concrete Case
 NOCE Corpus
 Linguistic Information
 Tokenization
 POS-Tagging
 Automatic POS-Tagging

Analyzing learner language
 Sources of Evidence
 Mismatching Evidence
 Mismatch-free errors

Conclusion

A Concrete Case

- ▶ The NOCE learner corpus (Díaz-Negrillo 2009)
- ▶ Towards linguistic annotation
- ▶ Exploring automatic POS annotation of learner language
- ▶ What does it mean to POS-annotate learner language?

Towards Linguistic Annotation of Interlanguage
 Dettmer Meurers
 Higer Wurach

Learner Corpora
 Why they're useful
 On compiling learner corpora
 Why annotate corpora
 Data in SLA research
 Error annotation & beyond

Annotation quality
 Why it's important
 DECCA: Variation n-gram error detection

A Concrete Case
 NOCE Corpus
 Linguistic Information
 Tokenization
 POS-Tagging
 Automatic POS-Tagging

Analyzing learner language
 Sources of Evidence
 Mismatching Evidence
 Mismatch-free errors

Conclusion

The NOCE Learner Corpus

Participants

- Writing by 1st/2nd year students of English at the universities of Granada and Jaén
- Learner information included: age, level, L2 exposure, motivation, etc.

Task

- Written texts (argumentative, descriptive, narrative)
- Around 250 words per text
- Topics chosen from 3 suggestions or free writing

Internal structure

- 3 text collections per academic year
- 4 years (2003-2005; 2007-2009)

Towards Linguistic Annotation of Interlanguage

Detmar Meurers
Hilger Wurch

Learner Corpora

Why they're useful
On compiling learner corpora
Why annotate corpora
Data in SLA research
Error annotation & beyond

Annotation quality

Why it's important
DECCA: Variation n-gram
error detection

A Concrete Case

NOCE Corpus

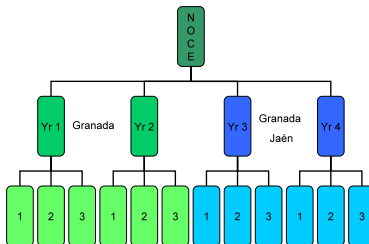
Linguistic Information
Tokenization
POS-Tagging
Automatic POS-Tagging

Analyzing learner language

Sources of Evidence
Mismatching Evidence
Mismatch-free errors

Conclusion

NOCE: Corpus Structure



Towards Linguistic Annotation of Interlanguage

Detmar Meurers
Hilger Wurch

Learner Corpora

Why they're useful
On compiling learner corpora
Why annotate corpora
Data in SLA research
Error annotation & beyond

Annotation quality

Why it's important
DECCA: Variation n-gram
error detection

A Concrete Case

NOCE Corpus

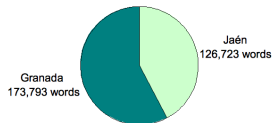
Linguistic Information
Tokenization
POS-Tagging
Automatic POS-Tagging

Analyzing learner language

Sources of Evidence
Mismatching Evidence
Mismatch-free errors

Conclusion

NOCE: Corpus Size



Overall figures
300,516 words 994 texts 438 participants

Towards Linguistic Annotation of Interlanguage

Detmar Meurers
Hilger Wurch

Learner Corpora

Why they're useful
On compiling learner corpora
Why annotate corpora
Data in SLA research
Error annotation & beyond

Annotation quality

Why it's important
DECCA: Variation n-gram
error detection

A Concrete Case

NOCE Corpus

Linguistic Information
Tokenization
POS-Tagging
Automatic POS-Tagging

Analyzing learner language

Sources of Evidence
Mismatching Evidence
Mismatch-free errors

Conclusion

NOCE: Annotation

- **EYES (Explicitly Encoded Surface modifications)**
100% of corpus annotated
 - Struck-out units
 - Late insertions
 - Reordering of units
 - Missing/unreadable text
- **EARS (Error Annotation and Retrieval System)**
≈25% of corpus annotated
 - Spelling
 - Punctuation
 - Word, phrase and clause grammar
 - Lexis
- How about adding linguistic information?

Towards Linguistic Annotation of Interlanguage

Detmar Meurers
Hilger Wurch

Learner Corpora

Why they're useful
On compiling learner corpora
Why annotate corpora
Data in SLA research
Error annotation & beyond

Annotation quality

Why it's important
DECCA: Variation n-gram
error detection

A Concrete Case

NOCE Corpus

Linguistic Information
Tokenization
POS-Tagging
Automatic POS-Tagging

Analyzing learner language

Sources of Evidence
Mismatching Evidence
Mismatch-free errors

Conclusion

First Step: Tokenization

- ▶ Maps input string into a series of tokens (words)
- ▶ Tokenization is
 - ▶ language dependent: e.g., English uses spaces to delimit words (vs. Chinese) (but: **in spite of, insofar as**)
 - ▶ character-set dependent: e.g., accented characters
 - ▶ application dependent: e.g., are there 1 or 2 tokens in
 - ▶ pronunciation vs. named entity: **US**
 - ▶ abbreviation vs. sentence-ending: **Mass.**
 - ▶ hyphenized words: **text-based**
 - ▶ contractions: **I'm, gonna, cannot**
- ▶ Learner spelling mistakes such as additional or missing spaces can create problems for tokenization, e.g.:

(6) *I, saw, John, inthe, park, the, other, day.*

Second Step: POS-Tagging

- ▶ Automatic assignment part-of-speech tags to each token
- ▶ Three freely available taggers
 - ▶ Stanford Tagger (Stanford University NLP Group)
 - ▶ TnT (Universität des Saarlandes, Saarbrücken)
 - ▶ TreeTagger (University of Stuttgart)
- ▶ All three taggers use Penn Treebank tagset
 - ▶ Fairly general tag inventory, limited number of categories
- ▶ All three taggers come with models trained on the same newspaper texts (Wall Street Journal)
 - ▶ Comparable results
- ▶ Performance is known to degrade on other text genres
 - ▶ Learner essays ≠ newspaper text

POS tagging of NOCE: An experiment

Setup

- ▶ Used 3 POS taggers trained on newspaper text
 - ▶ TreeTagger, TnT tagger, Stanford tagger
- ▶ Tagged the error-annotated section in NOCE
 - ▶ 179 texts ≈ 44 000 words

Results

- ▶ Manually evaluated POS tags assigned by taggers to 10 texts by 10 different participants (1850 words)
- ▶ Accuracy of automatically assigned tags
 - ▶ TreeTagger: 94.95%
 - ▶ TnT Tagger: 94.03%
 - ▶ Stanford Tagger: 88.11%

POS tagging of NOCE: Some issues

Spelling

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

Word boundaries

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

- ▶ Found lower performance for expressions which do not exist in English (in line with de Haan 2000; van Rooy & Schäfer 2002)
- ▶ But is tagging learner language really just a robustness issue, like adapting taggers to another domain?
- ▶ What does it mean for a POS tag to be correct for learner language?!

► POS analysis based on evidence in the text:

- information in **lexical entries**

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

- information encoded in **morphological information**

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

- information conveyed by **distribution**

(11) *The old **man** the boat.*

Dietmar Meurers
Hilger Wurch

Learner Corpora

Why they're useful
On compiling learner corpora
Why annotate corpora
Data in SLA research
Error annotation & beyond

Annotation quality

Why it's important
DECCA: Variation n-gram
error detection

A Concrete Case

NOCE Corpus

Linguistic Information

Tokenization
POS-Tagging
Automatic POS-TaggingAnalyzing learner
language

Sources of Evidence

Mismatching Evidence
Mismatch-free errors

Conclusion

- POS tagging learner language usually handled as a domain transfer (robustness) problem
 - train/develop on **native** language
 - apply post-correction
- Are POS tags designed for native language suitable for **systematically** describing learner language?
- Can they make interesting properties of learner language explicit?
- We argue for developing a new POS category system that can better represent learner language

Dietmar Meurers
Hilger Wurch

Learner Corpora

Why they're useful
On compiling learner corpora
Why annotate corpora
Data in SLA research
Error annotation & beyond

Annotation quality

Why it's important
DECCA: Variation n-gram
error detection

A Concrete Case

NOCE Corpus

Linguistic Information

Tokenization
POS-Tagging
Automatic POS-TaggingAnalyzing learner
language

Sources of Evidence

Mismatching Evidence
Mismatch-free errors

Conclusion

Case 1: Stem-Distribution mismatch



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

Stem	Distribution	Morphology
verb	noun	?

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

Stem	Distribution	Morphology
noun	adjective	?

Dietmar Meurers
Hilger Wurch

Learner Corpora

Why they're useful
On compiling learner corpora
Why annotate corpora
Data in SLA research
Error annotation & beyond

Annotation quality

Why it's important
DECCA: Variation n-gram
error detection

A Concrete Case

NOCE Corpus

Linguistic Information

Tokenization
POS-Tagging
Automatic POS-TaggingAnalyzing learner
language

Sources of Evidence

Mismatching Evidence
Mismatch-free errors

Conclusion

Case 1: Stem-Distribution mismatch



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

Stem	Distribution	Morphology
conjunction	wh-pronoun	?

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

Stem	Distribution	Morphology
preposition	conjunction	?

Dietmar Meurers
Hilger Wurch

Learner Corpora

Why they're useful
On compiling learner corpora
Why annotate corpora
Data in SLA research
Error annotation & beyond

Annotation quality

Why it's important
DECCA: Variation n-gram
error detection

A Concrete Case

NOCE Corpus

Linguistic Information

Tokenization
POS-Tagging
Automatic POS-TaggingAnalyzing learner
language

Sources of Evidence

Mismatching Evidence
Mismatch-free errors

Conclusion

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



(16) [...] *one of the favourite places to visit for many foreigners.*

Stem	Distribution	Morphology
adjective	noun	noun / verb 3 rd sg

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

Stem	Distribution	Morphology
noun / adjective	verb	verb

Towards Linguistic Annotation of Interlanguage

Center: Meurers, Higger, Wurch
 Learner Corpora
 Why they're useful
 On compiling learner corpora
 Why annotate corpora
 Data in SLA research
 Error annotation & beyond
 Annotation quality
 Why it's important
 DECCA: Variation n-gram error detection
 A Concrete Case
 NOCE Corpus
 Linguistic Information
 Tokenization
 POS-Tagging
 Automatic POS-Tagging
 Analyzing learner language
 Sources of Evidence
 Mismatching Evidence
 Mismatch-free errors
 Conclusion

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



(18) [...] *and dark politicals will be defeated.*

(19) [...] *internet have some "pages" that contents something so horrible [...]*

Derivational morphology and inflectional morphology point to different POS: Further splitting within slots?

Towards Linguistic Annotation of Interlanguage

Center: Meurers, Higger, Wurch
 Learner Corpora
 Why they're useful
 On compiling learner corpora
 Why annotate corpora
 Data in SLA research
 Error annotation & beyond
 Annotation quality
 Why it's important
 DECCA: Variation n-gram error detection
 A Concrete Case
 NOCE Corpus
 Linguistic Information
 Tokenization
 POS-Tagging
 Automatic POS-Tagging
 Analyzing learner language
 Sources of Evidence
 Mismatching Evidence
 Mismatch-free errors
 Conclusion

Case 3: Stem-Morphology mismatch



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

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

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

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

Towards Linguistic Annotation of Interlanguage

Center: Meurers, Higger, Wurch
 Learner Corpora
 Why they're useful
 On compiling learner corpora
 Why annotate corpora
 Data in SLA research
 Error annotation & beyond
 Annotation quality
 Why it's important
 DECCA: Variation n-gram error detection
 A Concrete Case
 NOCE Corpus
 Linguistic Information
 Tokenization
 POS-Tagging
 Automatic POS-Tagging
 Analyzing learner language
 Sources of Evidence
 Mismatching Evidence
 Mismatch-free errors
 Conclusion

Case 4: Distribution-Morphology mismatch



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

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

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

Stem	Distribution	Morphology
verb	verb past participle	verb past tense

Towards Linguistic Annotation of Interlanguage

Center: Meurers, Higger, Wurch
 Learner Corpora
 Why they're useful
 On compiling learner corpora
 Why annotate corpora
 Data in SLA research
 Error annotation & beyond
 Annotation quality
 Why it's important
 DECCA: Variation n-gram error detection
 A Concrete Case
 NOCE Corpus
 Linguistic Information
 Tokenization
 POS-Tagging
 Automatic POS-Tagging
 Analyzing learner language
 Sources of Evidence
 Mismatching Evidence
 Mismatch-free errors
 Conclusion

Case 4: Distribution-Morphology mismatch



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

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

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

Towards Linguistic Annotation of Interlanguage

Danier Mueurs
Hélène Waroch

Learner Corpora

Why they're useful
On compiling learner corpora
Why annotate corpora
Data in SLA research
Error annotation & beyond

Annotation quality

Why it's important
DECCA: Variation n-gram error detection

A Concrete Case

NOCE Corpus
Linguistic Information
Tokenization
POS-Tagging
Automatic POS-Tagging

Analyzing learner language

Sources of Evidence

Mismatching Evidence
Mismatch-free errors

Conclusion

Mismatch-free learner language

Realization using wrong allomorph

(26) *The majority of people that die in Irak are **childs** [...]*

(27) *He **runned** to buy one [...]*

Towards Linguistic Annotation of Interlanguage

Danier Mueurs
Hélène Waroch

Learner Corpora

Why they're useful
On compiling learner corpora
Why annotate corpora
Data in SLA research
Error annotation & beyond

Annotation quality

Why it's important
DECCA: Variation n-gram error detection

A Concrete Case

NOCE Corpus
Linguistic Information
Tokenization
POS-Tagging
Automatic POS-Tagging

Analyzing learner language

Sources of Evidence

Mismatching Evidence
Mismatch-free errors

Conclusion

Mismatch-free learner language

Realization using wrong stem

(28) [...] the 11th March **comes** to our minds.

Towards Linguistic Annotation of Interlanguage

Danier Mueurs
Hélène Waroch

Learner Corpora

Why they're useful
On compiling learner corpora
Why annotate corpora
Data in SLA research
Error annotation & beyond

Annotation quality

Why it's important
DECCA: Variation n-gram error detection

A Concrete Case

NOCE Corpus
Linguistic Information
Tokenization
POS-Tagging
Automatic POS-Tagging

Analyzing learner language

Sources of Evidence

Mismatching Evidence
Mismatch-free errors

Conclusion

Mismatch-free learner language

Duplicate inflection

(29) ***Childrens** spend so much time [...]*

(30) [...] it **stresseses** me a lot.

Towards Linguistic Annotation of Interlanguage

Danier Mueurs
Hélène Waroch

Learner Corpora

Why they're useful
On compiling learner corpora
Why annotate corpora
Data in SLA research
Error annotation & beyond

Annotation quality

Why it's important
DECCA: Variation n-gram error detection

A Concrete Case

NOCE Corpus
Linguistic Information
Tokenization
POS-Tagging
Automatic POS-Tagging

Analyzing learner language

Sources of Evidence

Mismatching Evidence
Mismatch-free errors

Conclusion

Mismatch-free learner language

Inappropriate word-formation rules

(31) [...] *internet can **modificate** [...]*

(32) [...] *different **socialities** and ways of life.*

Towards Linguistic Annotation of Interlanguage

Dexter Meurers
Hilger Wurch

Learner Corpora

Why they're useful
On compiling learner corpora
Why annotate corpora
Data in SLA research
Error annotation & beyond

Annotation quality

Why it's important
DECCA: Variation n-gram error detection

A Concrete Case

NOCE Corpus
Linguistic Information
Tokenization
POS-Tagging
Automatic POS-Tagging

Analyzing learner language

Sources of Evidence
Mismatching Evidence
Mismatch-free errors

Conclusion



37 / 38

Mismatch-free learner language

Creative lexis

(33) [...] *people shouldn't be **menospreciated** because of the music they listen to [...]*
(*menospreciados (span.): undervalued*)

(34) [...] *for many **raison**s.*

Towards Linguistic Annotation of Interlanguage

Dexter Meurers
Hilger Wurch

Learner Corpora

Why they're useful
On compiling learner corpora
Why annotate corpora
Data in SLA research
Error annotation & beyond

Annotation quality

Why it's important
DECCA: Variation n-gram error detection

A Concrete Case

NOCE Corpus
Linguistic Information
Tokenization
POS-Tagging
Automatic POS-Tagging

Analyzing learner language

Sources of Evidence
Mismatching Evidence
Mismatch-free errors

Conclusion



38 / 38

Conclusion

- ▶ Data collected in learner corpora in principle can provide empirical insights for development & validation of theories
- ▶ We discussed
 - linguistic annotation of learner corpora to support effective querying for example patterns discussed in SLA research
 - design criteria for an error annotation scheme
- ▶ We argued for an approach to the POS analysis of learner language, which distinguishes
 - lexical information
 - morphological information
 - distribution

to obtain a systematic classification of POS properties capturing native-like text as well as learner innovations.

⇒ The (automatic) analysis of learner language collected in corpora provides many interesting challenges and opportunities.

Towards Linguistic Annotation of Interlanguage

Dexter Meurers
Hilger Wurch

Learner Corpora

Why they're useful
On compiling learner corpora
Why annotate corpora
Data in SLA research
Error annotation & beyond

Annotation quality

Why it's important
DECCA: Variation n-gram error detection

A Concrete Case

NOCE Corpus
Linguistic Information
Tokenization
POS-Tagging
Automatic POS-Tagging

Analyzing learner language

Sources of Evidence
Mismatching Evidence
Mismatch-free errors

Conclusion



39 / 38

References

- 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>.
- 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/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, 93–119. URL <http://slr.sagepub.com/cgi/reprint/2/2/93.pdf>.
- de Haan, P. (2000). Tagging non-native English with the TOSCA-ICLE tagger. In Mair & Hundt (2000), pp. 69–79.
- de Mönnik, I. (2000). Parsing a learner corpus. In Mair & Hundt (2000), pp. 81–90.
- 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>.

Towards Linguistic Annotation of Interlanguage

Dexter Meurers
Hilger Wurch

Learner Corpora

Why they're useful
On compiling learner corpora
Why annotate corpora
Data in SLA research
Error annotation & beyond

Annotation quality

Why it's important
DECCA: Variation n-gram error detection

A Concrete Case

NOCE Corpus
Linguistic Information
Tokenization
POS-Tagging
Automatic POS-Tagging

Analyzing learner language

Sources of Evidence
Mismatching Evidence
Mismatch-free errors

Conclusion



39 / 38

- 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. URL <http://www.aclweb.org/anthology-new/P05-1040>.
- 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.
- 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>.
- Kanno, K. (1997). The acquisition of null and overt pronominals in Japanese by English speaker. *Second Language Research* 13, 265–287. URL <http://slr.sagepub.com/cgi/reprint/13/3/265>.
- Lüdeling, A., S. Doolittle, H. Hirschmann, K. Schmidt & M. Walter (2008). Das Lernerkorpus Falko. *Deutsch als Fremdsprache* 45(2), 67–73.
- Mair, C. & M. Hundt (eds.) (2000). *Corpus Linguistics and Linguistic Theory*. Amsterdam: Rodopi.
- 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, W. D. (2005). On the use of electronic corpora for theoretical linguistics. Case studies from the syntax of German. *Lingua* 115(11), 1619–1639. URL <http://purl.org/dm/papers/meurers-03.html>.
- Meurers, W. D. & S. Müller (2009). Corpora and Syntax (Article 42). In A. Lüdeling & M. Kytö (eds.), *Corpus linguistics*, Berlin: Mouton de Gruyter, vol. 2 of *Handbooks of Linguistics and Communication Science*, pp. 920–933. URL <http://purl.org/dm/papers/meurers-mueller-09.html>.

Towards Linguistic Annotation of Interlanguage

Detlev Meurers
Hilger Wenzsch

Learner Corpora

Why they're useful

On compiling learner corpora

Why annotate corpora

Data in SLA research

Error annotation & beyond

Annotation quality

Why it's important

DECCA: Variation n-gram error detection

A Concrete Case

NOCE Corpus

Linguistic Information

Tagmatrices

POG-Tagging

Automatic POG-Tagging

Analyzing learner language

Sources of Evidence

Mismatching Evidence

Mismatch-free errors

Conclusion



- Pérez-Lerroux, A. & W. Glass (1997). OPC effects in the L2 acquisition of Spanish. In A. Pérez-Lerroux & W. Glass (eds.), *Contemporary Perspectives on the Acquisition of Spanish*, Somerville, MA: Cascadia Press, vol. 1, pp. 149–165.
- 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>.
- 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.
- Voutlainen, 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.

Towards Linguistic Annotation of Interlanguage

Detlev Meurers
Hilger Wenzsch

Learner Corpora

Why they're useful

On compiling learner corpora

Why annotate corpora

Data in SLA research

Error annotation & beyond

Annotation quality

Why it's important

DECCA: Variation n-gram error detection

A Concrete Case

NOCE Corpus

Linguistic Information

Tagmatrices

POG-Tagging

Automatic POG-Tagging

Analyzing learner language

Sources of Evidence

Mismatching Evidence

Mismatch-free errors

Conclusion

