On Annotating Learner Corpora: Some Recent Developments

Ana Díaz Negrillo
Universidad de Jaén

Detmar Meurers, Holger Wunsch
Universität Tübingen

Corpus Annotation Workshop, Paris 13
May 28, 2009

Roadmap of Talk

- Data in Second Language Acquisition (SLA) Research
  - How are the relevant sets of examples characterized?
- Learner Corpora
  - Which role can they play in SLA research?
  - Which types of annotation are relevant?
  - How can high quality annotation be obtained?
- Current work
  - The NOCE (NOn-native Corpus of English) learner corpus
  - Towards linguistic annotation of NOCE
    - Tokenization, POS tagging
  - XML and TEI representation of the annotated corpus
  - Automatic POS tagging learner language: First insights
- Conclusion

Data in Second Language Acqusition research

- Learner data is essential empirical basis of SLA research
- Questions for our work:
  - How do SLA researchers characterize the data relevant to their theories of language acquisition?
  - What linguistic categories and properties do they refer to?
  - Can example data for the relevant patterns be found in learner corpora?
  - How does the data need to be annotated to provide direct access to the relevant example classes?

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

Learner corpora

- As collections of data, learner corpora can in principle
  - help validate generalizations about language acquisition
  - provide a broad empirical basis for the development of new hypotheses and theories
- Depending on the corpus composition, it can support *qualitative* and *quantitative* analysis of examples found
- 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
  - Effective querying of corpora often requires reference to annotations – what kind of annotations are needed?
  - SLA research essentially observes correlations of linguistic properties, whether erroneous or not

  ⇒ Learner corpora should ideally provide annotation of linguistic properties, including but not limited to errors

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
- 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 & Babarczy 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)

Automatic annotation and required collaboration

- Automatic annotation techniques learning from such gold standard annotation are becoming available
  - 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. de Haan 2000; de Mönnink 2000; van Rooy & Schäfer 2002, 2003)

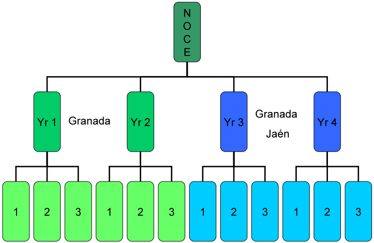
Current Work: Outline

- The NOCE learner corpus (Díaz Negrillo 2009)
- Towards linguistic annotation
- Corpus representation
  - XML
  - TEI
- Exploring automatic POS annotation of learner language

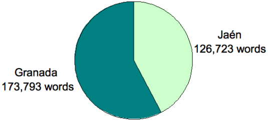
The NOCE Learner Corpus

- 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)

NOCE: Corpus Structure



NOCE: Corpus Size



Overall figures		
300,516 words	994 texts	438 participants

NOCE: Annotation

- EYES (ExplicitlY Encoded Surface modifications) 100% of corpus annotated
  - Struckout 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?

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 tokenziation, e.g.:

(1) *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

Representing rich information: XML

- Many different types of information:
  - Learner information
  - Learner text
  - Error tags and editorial tags
  - Tokenization of the text
  - POS tags
- How can we keep the information in the same file, but still clearly separated?

⇒ Use XML

XML: Representation of annotation

- Primary data: everything between a <w> tag
- Edited out data: enclosed in <C> tags
- POS-tags: attributes on each token

```
<?xml version="1.0" encoding="ISO-8859-15"?>
<corpus>
  <w id='w520' pos-stt='IN' pos-tnt='IN' pos-tt='IN'>inside</w>
  <C>
   <w id='w521' pos-stt='NN' pos-tnt='(' pos-tt='('>(</w>
   <w id='w522' pos-stt='DT' pos-tnt='DT' pos-tt='DT'>the</w>
   <w id='w523' pos-stt='NN' pos-tnt='NN' pos-tt='NN'>cassette</w>
   <w id='w524' pos-stt='NN' pos-tnt=')' pos-tt=')'>)</w>
  </C>
  <w id='w525' pos-stt='DT' pos-tnt='DT' pos-tt='DT'>a</w>
  <w id='w526' pos-stt='JJ' pos-tnt='JJ' pos-tt='JJ'>small</w>
  <w id='w527' pos-stt='NN' pos-tnt='NN' pos-tt='NN'>cassette</w>
  <w id='w528' pos-stt='.' pos-tnt='.' pos-tt='SENT'
    sb='true'>.</w>

</corpus>
```

XML: TEI header

- TEI: Text Encoding Initiative (http://www.tei-c.org)
- TEI headers in NOCE contain information about:
  - Who compiled the corpus and where
  - The tasks the learners carried out
  - The learners (proficiency level, their reasons for learning English, native language(s), location, . . .)
  - The tools used to produce the corpus
  - . . .
- Particularly important for interdisciplinary research as it provides comprehensive and standardized information

XML: More on the benefits

- Standard XML tools help quickly find cases where
  - annotators forgot to type in closing error tags
  - accidentally interleaving error tags were annotated
  - error tags were mistyped

```
<?xml version="1.0" encoding="ISO-8859-15"?>
<corpus>
  To <LX.VR.IT.CC.MS>practice basketball, football
  <PN.CM.OM></PN.CM.OM> tennis <PN.EP.OV>...
  </PN.EP.OV> </LX.VR.IT.CC.MS> is a form
  <PG.CS.CP.NN.RE.NF.MS> to
    <LX.VR.IT.CC.MS> delete
    </PG.CS.CP.NN.RE.NF.MS> fats and sugars
  </LX.VR.IT.CC.MS>.
</corpus>
```

XML Schema: definition of annotation schemes

- Provide exact definition of annotation scheme
- Typos and confusions can be automatically detected while you type
  - e.g., <VBB> instead of <VBP> (verb, present, sg, ¬3rd)
- Essentially a formalized kind of documentation

POS tagging of NOCE: First 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: Examples

Spelling

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

Word boundaries

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

Morphology

(4) [. . .] *american's customes are totally different* [. . .]

⇒ Found lower performance for expressions which do not exist in English (in line with de Haan 2000; van Rooy & Schäfer 2002)

Issues in POS tagging learner language

Goal

- POS tagging of learner language: description of learner language in terms of the target language categories
- What are the issues?

Issue 1
How to adapt TL linguistic classifications to learner language?

(5) *You will not need a guide that translate you everything* [. . .]
i. **VB** (verb, base form)
ii. **VBP** (verb, present, sg, ¬3rd)

(6) *And also it creats sometimes a shock situation* [. . .]
i. **NN** (noun, sg or mass)
ii. **VB** (verb, base form)

Issues in POS tagging learner language

Issue 2
What information do we want to have in POS tags?

- Distribution in TL
- Lexical look-up of word in TL
- Lexical look-up or morphological analysis of word in LL

(7) [. . .] *it will be very importance because* [. . .]
i. **NN** (TL lexical look-up)
ii. **ADJ** (TL distribution)

(8) *They have more oportunities to be choiced for a job* [. . .]
i. **NN+ed** (LL lex. look-up)
ii. **VVN** (TL distribution)

Issues in POS tagging learner language

- Potential solution to these issues we are exploring:
  - Underspecify wherever mismatches arise, so that all subsumed POS classifications are encoded
- Another open issue: How can units problematic for automatic taggers be identified?
  - So far: error-tagged section of the corpus identifies some problematic non-words
  - Explore other detection mechanisms

Conclusion

- Data collected in learner corpora in principle can provide empirical insights for development & validation of theories (cf. Meurers 2005; Meurers & Müller 2008)
- In this talk, we argued for
  - linguistic annotation of learner corpora to support effective querying for example patterns discussed in SLA research
  - description of learner language using TL categories: mismatches can help define specific properties of learner language
  - usage of XML/TEI for data representation
- There is a clear need for interdisciplinary collaboration between applied and computational linguistics to develop
  - annotation schemes, gold standard corpora, and automatic annotation methods for learner language

Appendix: Some tools we use

Tools for detecting errors in corpus-annotation:

- Decca project: http://decca.osu.edu

POS-Taggers

- Stanford POS Tagger (free, University of Stanford) http://nlp.stanford.edu/software/tagger.shtml
- TnT POS Tagger (free, University of the Saarland) http://www.coli.uni-saarland.de/˜thorsten/tnt
- TreeTagger (free, University of Stuttgart, Germany) http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

XML processing

- xmllint: XML checking and formatting (free, in LibXML2) http://www.xmlsoft.org
- SyncRO Soft Oxygen: XML Editor & Validator (commercial) http://www.oxygenxml.com

References

Boyd, A., M. Dickinson & D. Meurers (2008). On Detecting Errors in Dependency Treebanks. *Research on Language and Computation* URL http://purl.org/dm/papers/boyd-et-al-09.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/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.

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.

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. Http://www.aclweb.org/anthology-new/E/E03/.

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://www.aclweb.org/anthology-new/P05-1040.

Díaz Negrillo, A. (2009). *EARS - A User's Manual*. Munich: Lincom.

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

Meurers, D. & S. Müller (2008). Corpora and Syntax (Article 44). In A. Lüdeling & M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, Berlin: Mouton de Gruyter, Handbooks of Linguistics and Communication Science. URL http://purl.org/dm/papers/meurers-mueller-07.html.

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.

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.grsampson.net/Alta.html.

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.

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://www.aclweb.org/anthology/E95-1029.