On the Annotation and Use of Learner Language Corpora

Detmar Meurers
Universität Tübingen

Luiz Amaral
UMass Amherst

AAAL-09 Colloquium
Bridging Computational and Applied Linguistics:
Implementation Challenges and Benefits

March 21, 2009

Roadmap of Talk

- Data in SLA Research
  - How are the relevant sets of examples characterized?
- Learner Corpora
  - Corpora of what?
- Corpus Annotation
  - Which types of annotation are relevant?
- Annotation Quality
  - Why is it important?
  - How can it be obtained?

Data in SLA research

Learner data is the essential empirical basis of SLA research

- 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]
- Examples:

(1) *Früher* *ich kannte* *den Mann* (Stage 2)
earlier$_{AdvP}$ I$_S$ knew$_V$ [the man]$_O$

(2) *Früher* *kannte* *ich den Mann* (Stage 4)
earlier$_{AdvP}$ knew$_{V[+fin]}$ I$_S$ [the man]$_O$

- How is the data characterized?
  - lexical and syntactic categories and functions

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$_{*i/j}$ will win the prize.'

(4) *Nadie dice que __ ganará el premio.*
nobody says that *pro* will win the prize
'Nobody$_i$ says that he$_{i/j}$ will win the prize.'

- How is the data characterized?
  - syntactic functions and semantic relations
  - not overtly expressed but interpreted elements

Data in SLA research

Amaral (forthcoming)

- Paper investigates acquisition of subcategorization and selectional restrictions in Spanish by English speakers.
- Examples:

(5) a. * ***Ella*** *gusta el pastel.*
she likes the cake

b. ✓ ***Le*** *gusta el pastel.*
to her pleases the cake

(6) a. * *Ella conoce* ***Juan****.*
she knows Juan

b. ✓ *Ella conoce* ***a Juan.***
she knows a-personal *Juan*

Data in SLA research

Amaral (forthcoming) cont.

- Hypotheses of the study:
  - Selectional restrictions are the driving force in the acquisition of verbal lexical properties.
  - L1 subcategorization frames are transferred and their reanalysis only occurs later.
- How are the data and the hypotheses characterized?
  - lexical subcategorization requirements
  - selectional restrictions
  - syntax-semantics mapping

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.
  - Some SLA studies using learner corpora (e.g., in Ortega & Byrnes 2008)
- To find relevant classes of examples, the terminology used to single out the learner language aspects of interest needs to be mapped to instances in the corpus.
  - Effective querying of corpora often requires reference to annotated linguistic abstractions instead of extensionally characterizing individual strings.

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.

Annotation: Beyond errors

- The annotation of learner corpora has typically focused on *errors* made by the learners.
- Yet, SLA research essentially observes correlations of linguistic properties, whether erroneous or not.
  - SLA research discussed earlier
  - Research focusing on
    - overuse/underuse of specific patterns
    - measures of language development (Developmental Sentence Scoring, Index of Productive Syntax, . . . , cf. also Lu 2008)

⇒ 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
  - Annotation schemes: as **theory-neutral** as possible
- 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
- Lack of research into linguistic annotation of L2 corpora (but cf. Lüdeling et al. 2005)
  - Interdisciplinary collaboration between SLA and CL crucial to adapt **annotation schemes** and **methods** from L1 corpora to interlanguage collected in L2 corpora

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).

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.

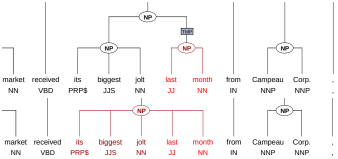
⇒ Where the query characterizing the target phenomenon is expressed in terms of the annotation, quality and consistency of the annotation is important.

How to obtain high quality annotation

- Annotate corpus several times and independently, then test interannotator agreement (Brants & Skut 1998)
  - Interannotator agreement is an essential measure of whether the annotation scheme distinctions can be applied consistently based on the information in the corpus.
- Define adequate annotation scheme with good manual to allow for 100% agreement (Voutilainen & Järvinen 1995; Sampson & Babarczy 2003)
  - keep only distinctions which can be reliably and consistently identified and annotated uniquely
  - appendix of difficult cases and how to resolve them
- Detection of annotation errors through automatic analysis of comparable data recurring in the corpus
  → NSF project DECCA

DECCA: Variation n-gram error detection

- **Variation**: multiple occurrences, with different annotations
  a) **ambiguity**: different annotations correctly label the same material used in different contexts
  b) **annotation error**: annotation is inconsistent across comparable occurrences
- Variation between constituent and non-constituent:



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

- Variation between two syntactic category labels:

(7) *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
- All code is freely available from our project website

http://decca.osu.edu

Conclusion

- Data collected in learner corpora in principle can
  - help validate generalizations about language acquisition
  - provide a broad empirical basis for the development of new hypotheses and theories

  (cf. also Meurers 2005; Meurers & Müller 2008)
- In this talk, we argued for
  - the creation of learner corpora stemming from a variety of contexts and tasks
  - linguistic annotation of learner corpora to support effective querying for example patterns discussed in SLA research
  - the importance of annotation quality
- There is a clear need for interdisciplinary collaboration between applied and computational linguistics to develop
  - annotation schemes for learner language
  - gold standard corpora and automatic annotation methods for such interlanguage

References

Amaral, L. (forthcoming). The lack of parallelism in the acquisition of subcategorization frames and selectional restrictions in Spanish by English speakers .

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 (NeMLaP-98)*. Syndey. http://www.coli.uni-sb.de/˜thorsten/publications/Brants-Skut-NeMLaP98.ps.gz

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.

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. 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. 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. http://www.aclweb.org/anthology/P/P05/P05-1040.

Kanno, K. (1997). The acquisition of null and overt pronominals in Japanese by English speaker. *Second Language Research* 13, 265–287.

Lu, X. (2008). Automatic measurement of syntactic complexity using the revised developmental scale. In *Proceedings of the 21st International Florida Artificial Intelligence Research Society Conference (FLAIRS-08)*. Coconut Grove, FL: AAAI Press.

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.

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. 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. http://purl.org/dm/papers/meurers-03.html.

Ortega, L. & H. Byrnes (eds.) (2008). *The longitudinal study of advanced L2 capacities*. Routledge.

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: Cascadilla 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. http://www.grsampson.net/Alta.html.

Voutilainen, A. & T. Järvinen (1995). Specifying a shallow grammatical representation for parsing purposes. In *Proceedings of the 7th Conference of the EACL*. Dublin, Ireland. http://www.aclweb.org/anthology/E95-1029.