NLP for Non-Canonical Language and Learner Language

Detmar Meurers
Universität Tübingen

Syntactic Analysis of Non-Canonical Language Workshop
NAACL-HLT, Montreal, 8. June 2012

Why is Learner Language analyzed?

- *Annotation of learner corpora*
  - for research into how languages are acquired
    → Second Language Acquisition (SLA)
  - to identify typical student needs
    → Foreign Language Teaching and Learning (FLTL)
- *Analysis of form or meaning of learner responses to tasks*
  - provide feedback to support *acquisition*
    → Intelligent Tutoring Systems
  - assess learner abilities
    → Language Testing
- *Analysis of form of free text*
  - provide feedback to support *text production*
    → Writer's aids

(cf. survey article: Meurers 2012)

On the nature of categories for learner language

- Where do linguistic categories come from?
  - Categories result from generalizations, which *require a significant amount of comparable data* to be made.
  - What constitutes useful categories characterizing learner language is subject of SLA research.
- In NLP, *robustness* is the ability to *ignore variation* in the realization of a category to be identified.
  - Robustness is based on assumption of an *intended target!*
  - Danger of *comparative fallacy*: “the mistake of studying the systematic character of one language by comparing it to another.” (Bley-Vroman 1983, p. 6)

$\Rightarrow$ Pre-theoretic classes close to the empirical observations are best-suited for the emergent nature of interlanguage.

Appropriate categories for learner language

Parts-of-speech (Díaz Negrillo, Meurers, Valera & Wunsch 2010)

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

- stem: preposition
- distribution: conjunction

(2) *to be **choiced** for a job*

- stem: noun or adjective
- distribution, morphology: verb

⇒ A single category from a standard POS tagset fails to systematically identify properties of learner language.

On the nature of categories for learner language

Consequences for syntactic annotation

- Idea: break down constituency in terms of
  - overall topology of a sentence (Hirschmann et al. 2007)
  - chunks and chunk-internal word order (Abney 1997)
  - dependency
- What is the empirical basis of dependency analysis?
  - distinguish morphological, syntactic, and semantic dependencies (cf. also Meaning Text Theory, Mel'čuk 1988)
- Some work on dependency analysis of learner language:
  - surface-evidence based (Dickinson & Ragheb 2009)
    - fine-grained record of morphological & syntactic evidence
  - semantic dependencies (MacWhinney 2008; Rosén & Smedt 2010; Ott & Ziai 2010; Hirschmann et al. 2010)
    - robustly abstract away from learner specific forms

    e.g. CoMiC project: comparing meaning of answers to reading comprehension questions (Hahn & Meurers 2011, 2012)

The importance of tasks and learners

- Targets are assumed for any kind of robust classification.
- What are the targets for the sentences taken from the Hiroshima English Learners' Corpus (Miura 1998)?

(3) *I didn't know*
(4) *I don't know his lives.*
(5) *I know where he lives.*
(6) *I know he lived*

They are taken from a translation task, for the Japanese of

(7) *I don't know where he lives.*

$\Rightarrow$ Cannot be determined just by the learner sentences alone!

- Task information crucial
- Learner information relevant (L1, past interaction, learner strategies used to accomplish tasks)

Summary

- Learner language is analyzed for a range of purposes.
- For analyzing learner language, we need to
  - identify the appropriate categories for a given purpose
  - determine the empirical basis of these categories
  - and what kind of robustness (= variation in realizing target categories) is appropriate given the purpose
- Pre-theoretic classes close to the empirical observations are best-suited for the emergent nature of interlanguage.
- Multiple levels of analysis needed to identify the right level of abstraction for different purposes.
  - Distinct POS categories for distribution, lemma, morphology
  - Syntactic analysis in terms of topology, chunks, dependency
- Explicit task and learner models can provide crucial constraining information for interpreting learner language.

References

Abney, S. (1997). Partial Parsing via Finite-State Cascades. *Natural Language Engineering* 2, 337–344. URL http://www.vinartus.net/spa/97a.pdf.

Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning* 33(1), 1–17. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1467-1770.1983.tb00983.x/pdf.

Díaz Negrillo, A., D. Meurers, S. Valera & H. Wunsch (2010). Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum* 36(1–2), 139–154. URL http://purl.org/dm/papers/diaz-negrillo-et-al-09.html. Special Issue on Corpus Linguistics for Teaching and Learning. In Honour of John Sinclair.

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.

Hahn, M. & D. Meurers (2011). On deriving semantic representations from dependencies: A practical approach for evaluating meaning in learner corpora. In *Proceedings of the Intern. Conference on Dependency Linguistics (DEPLING 2011)*. Barcelona, pp. 94–103. URL http://purl.org/dm/papers/hahn-meurers-11.html.

Hahn, M. & D. Meurers (2012). Evaluating the Meaning of Answers to Reading Comprehension Questions: A Semantics-Based Approach. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-7) at NAACL-HLT 2012*. Montreal, pp. 94–103. URL http://purl.org/dm/papers/hahn-meurers-12.html.

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.

Krivanek, J. & D. Meurers (2011). Comparing Rule-Based and Data-Driven Dependency Parsing of Learner Language. In *Proceedings of the Intern. Conference on Dependency Linguistics (DEPLING 2011)*. Barcelona, pp. 310–317.

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.

Mel'čuk, I. (1988). *Dependency Syntax: Theory and Practice*. State University of New York Press. URL http://books.google.com/books?id=diq29vrjAa4C&lpg=PR13&ots=ZcCJBmEA7g&dq=Dependency20Syntax3A20Theory20and20Practice&lr&pg=PR13#v=onepage&q&f=false.

Meurers, D. (2012). Natural Language Processing and Language Learning. In C. A. Chapelle (ed.), *Encyclopedia of Applied Linguistics*, Oxford: Wiley-Blackwell. URL http://purl.org/dm/papers/meurers-12.html. To appear.

Miura, S. (1998). Hiroshima English Learners' Corpus: English learner No. 2 (English I & English II). Department of English Language Education, Hiroshima University. http://purl.org/icall/eigo1.html, http://purl.org/icall/eigo2.html,.

Ott, N. & R. Ziai (2010). Evaluating Dependency Parsing Performance on German Learner Language. In M. Dickinson, K. Müürisep & M. Passarotti (eds.), *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories*. vol. 9 of *NEALT Proceeding Series*, pp. 175–186. URL http://hdl.handle.net/10062/15960.

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.

Exemplifying importance of context

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