Stefano Rastelli (University of Pavia)

Why tagging learners' errors when errors are only in a native speaker's mind?

The *SLA – Tagging Project*

1. Theoretical background

In 2006 a group of researchers and Ph.D students working at the Department of Theoretical and Applied Linguistics at the University of Pavia thought to establish a feasible method in order to tag consistently learner data and to build a morphosyntactic parser designed to run on a subset of Italian learner corpora collected over more than twenty years. The starting assumption was that error tagging is not suitable for SLA research because it fails to account for the inner systematicity of the Interlanguage. It was thought that an alternative way was possible and it was called "SLA-tagging" (henceforth SLAt: Second Language Acquisition tagging). In our opinion, the best way to completely give up premature over-interpretation of Interlanguage data is to run a Treetagger designed for L1 Italian on L2 data (Rastelli and Frontini 2008; Rastelli, 2009; Rastelli and Frontini, 2011). In spite of the apparent paradox, a L1 Treetagger - precisely in virtue of its evident flaws and weaknesses - can shed light over the Interlanguage more than any L2-tailored error grid can do. The SLA-tagging approach is radically different from the Error-tagging approach because: (a) it discards errors; (b) it substitutes errors with 'virtual categories'; (c) it tags only form and position of items, but not their morhosyntactic function; (d) it excludes human interpretation from tagging and saves it for successive data analysis. SLAt is not meant to disclose areas where learners show under-use or over-use of linguistic features nor to know which errors learners commit more. The SLAt procedure, despite being more complex, is also more rewarding for SLA research because it helps to reveal unexpected category assignments to items. These assignations may reveal how grammatical functions are gradually attributed to forms by learners

2. Roadmap of the workshop

(a) Learner corpora without error-tagging: the counting of learners' errors is of no use for second language acquisition studies.

(b) The SLA-tagging approach: a treetagger's failure to assign a POS is much more indicative for Second Language Acquisition research than any error-grids.

(c) "Virtual POS-tagging": what it is meant with "virtual" tag and the advantages of using virtual POS tags.

(d) How to run XML queries by using virtual tags: unexpected data are the most valuable thing

3. Tutorial and downloadable papers

Instructions for practicing SLA-tagging on English and Italian Learner corpora and downloadable papers will be available at http://sla-tagging.unipv.it by mid July. Participants at the Summer School are strongly encouraged to download, run the scripts and practice before coming to class.

4. References

Rastelli, Stefano, (2007): "Going beyond errors: position and tendency tags in a learner corpus". In Sansò, Andrea (ed.): *Language Resources and Linguistic Theories*. Milano, Franco Angeli: 96-109.

Rastelli, Stefano and Frontini, Francesca (2008): "SLA meets FLT research: the form/function split in the annotation of Learner Corpora". In *Proceedings of TaLC 8*. Lisbona: 446-451.

Rastelli, Stefano (2009): "Learner corpora without error tagging". *Linguistic online*, 38, II.

Rastelli, Stefano / Frontini, Francesca (2011): "The education of sight: POS-tagging of Italian Learner corpora for Second Language Acquisition research", Rassegna Italiana di Linguistica Applicata (RILA), 1-2, 85-96.