FLT meets SLA research: the form/function split in the annotation of learner corpora

Stefano Rastelli

University of Pavia - Italy

Francesca Frontini

University of Pavia - Italy

Our work[1] explores the advantages of adopting a strict form-to-function perspective when annotating learner corpora. Hopefully, such a perspective provides both Foreign Language Teaching (FLT) and Second Language Acquisition (SLA) researchers with insights not relating to learners' errors, but to some systematic features of interlanguage (IL). A split between forms and functions (or categories) is desirable in order to avoid both the "closeness fallacy" and the "comparative fallacy". In fact - especially in basic learner varieties - forms (or "functors") may precede functions and in their turn, functions may show up in unexpected forms. In the computer-aided error analysis (CEA) tradition, all items produced by learners are traced to a grid of error tags, which is based on the categories of the target language (TL). In a different way, we believe it is preferable to account for IL features in terms of "virtual" TL categories. For this purpose, a preliminary project-study for the tagging of L2 Italian (PIL2) has been completed at the University of Pavia. The project concluded that it is possible to use a tree-tagger designed for L1 Italian also for learner data on condition that the tagging system retrieves separately four levels of annotation: (a) the information about how a word is actually spelled / uttered by learners; (b) its position in the sentence; (c) the virtual categories attributed to that form on the basis of formal resemblance with TL items; (d) the level of confidence in recognizing both the category and the lemma. The aim of PIL2 project is not to disclose areas where learners show under-use or overuses of linguistic features nor to know which errors learners commit more. Using a tree-tagger designed for L1 Italian on data of learner Italian may reveal unexpected IL phenomena and allows us to see how the functions of the TL are gradually acquired by learners.

Interlanguage, learner corpora, error-tagging, comparative fallacy, L2 Italian.

[1] Stefano Rastelli wrote the first five paragraphs and the conclusions while Francesca Frontini wrote the sixth, the seventh and the eighth paragraph.

Is error tagging really inherent to learner corpora?

Far from neglecting or minimizing the tremendous importance of error tagging, especially for teaching purposes and for lexicography, we would like to propose a different way of pursuing the annotation of learner corpora. Our proposal gives up error tagging and consequently our answer to the question posed in the title of this paragraph (that is taken from Díaz-Negrillo and Fernández-Domínguez, 2006:84) is assumed to be "no". Error tagging should neither be considered as the Pillars of Hercules, beyond which the world ends, nor the only means available to teachers of becoming aware of learners' performance. The reason in twofold. First of all, it is possible that the nature of errors does not make them the best candidate possible for SLA research (see next paragraph). Secondly, researchers have yet to agree about general error-taxonomy, the standardization of error tagset still a long way from being at hand (Tono, 2003:801). According to Díaz-Negrillo and Fernández-Domínguez (2006:89), the number of tags in different error-tagging projects varies from 31 to 100. As far as the layers of analysis are concerned, phonetic, pragmatic and discourse errors are treated rarely and inconsistently, while the textual dimension do not seem to be considered at all (see also Rastelli, 2007: 99).

The error tagging and the "comparative/closeness fallacy"

A Chinese beginner student of L2 Italian, describing a house suddenly catching fire, says: *la casa di loro c'è fuoco* [lit. "The house of them there is fire"]. None of the items of this sentence taken individually is wrong, nor is it straightforward to pinpoint the source of the ill-formedness. Despite the fact that this scene is clear, it is not enough in order to label the possible errors unambiguously because there are at least three ways to correct the "wrong" sentence. Far from being an exception in learner data, sentences like the one above show that - unfortunately for us - many interesting IL features are not proper "errors", that is, they do not show up as "incorrect forms" each having one or more correct equivalent in a native speakers' mind. First of all in learner data it is not always possible even to isolate the form responsible for the sentence becoming incorrect or to define what this form, once singled out, stands for (that is, which is its "correct version" in the TL provided that it has just one, see Rastelli, 2007). Secondly, errors are often seen as token-based, whilst they often entail (or are embedded in) other errors (this problem has been recently addressed by adopting a multi-level standoff annotation, see Lüdeling et al., 2005). Finally, especially in basic varieties, learners often produce not just “lacking”, "wrong" or “mispelled” items, but rather "impossible" ones (the issue of the existence of different layers of "grammaticality" is partially addressed also in Foster, 2007:131). Here “impossible” is meant as unclassifiable and unpredictable. “Unclassifiable” is a combination of a number of *per se* well-formed items, that a native-speaker perceives as being wrong as a whole, despite not knowing the precise rule being violated. “Unpredictable” is a combination of characters whose nature is not capturable by using a pre-fabricated, closed set of errors, no matters its size. It has been pointed out that the practice of error tagging rests on native speaker's intuition. The elaboration of an error manual is usually meant to avoid or at least minimize taggers' subjectivity when dealing with deviant phenomena. While, in everyday life judgements, subjectivity is not necessarily a flaw, when it plays a decisive role in annotation of learner corpora it is at risk of committing "comparative fallacy" and "closeness fallacy", as far as these two concepts are intended by Huebner (1979), Bley-Vroman (1983), Klein and Perdue (1992), Cook (1997), Lakshmanan and Selinker (2001) (see also a special issue of TESOL &Applied Linguistics, 2004). The comparative fallacy emerges when a researcher studies the systematic character of one language by comparing it to another or (as often happens) to the TL. The "closeness fallacy" occurs "in cases where an utterance produced bore a superficial resemblance to a TL form, whereas it was in fact organised along different principles" (Klein & Perdue, 1992: 333). The comparative fallacy represents an attitude, while the closeness fallacy the most likely case of its practical application, that is, when the TL coincides with the language of the researcher. Failure to avoid the comparative fallacy will result in "incorrect or misleading assessments of the systematicity of the learner's language". Bley-Vroman's criticism (1983: 2) applies also "to any study in which errors are tabulated [...] or to any system of classification of IL production based on such notions as *omission*, *substitution* or the like". The logic of "correct-incorrect" binary choice which is so peculiar to errors, hides the fact that the surface contrast in IL may be determined by no single factor, but by a multiplicity of interacting principles, some of which unknown (8). For all these reasons, it is the analysis of unexpected and "spurious" items sorted out by the system also in non obligatory contexts that is likely to reveal the systematicity of some IL features.

Since using error-tags means to get exactly what one expects and to hide developing and provisional non target-like learner grammars, in our project it was decided to find an alterative way to run queries on learner corpora. Since this query system should have been TL rule-oriented and not TL rule-governed, it was thought that the best way to deal with learner data without error tagging would have to focus on some kind of xml treatment of the outcome of a Treetagger designed for L1 Italian.

“Unexpected” data and patterned queries

The fact that, according to our view, "unpredictable data" is so important for SLA research does not mean that we should give up using TL categories and that all queries on the learner data should be carried out randomly. Also "unexpected/unpredictable" data should be looked for systematically when testing a hypothesis about developing learner grammars. The following example is taken from the Pavia Corpus. A Chinese beginner student of L2 Italian, when asked to report about his education, said: *Cinese fato media* ("Chinese done middle [school]" that is assumed to mean: "In China I attended the middle school"). A few days later, when asked about holidays, the same learner said that: *Sì, in Cina festa pasqua anche* ("Yes, in China holiday Easter too", that is assumed to mean: "Yes, in China there are Easter Holidays as well"). Following the bracketed and provisional interpretation and under an error-driven perspective, only in the first sentence is the learner blurring the distinction between the category of adjectives ("Chinese") and the category of nouns (here placed into a locative expression "in China"). We thus could label this as an "error" following the appropriate category of FRIDA tagset. It would belong to the subset of errors named "class" <CLA> (exchange of class) and to the higher set of Grammar <G> errors (Granger, 2003: 4). If we adopt a different perspective, we might compare the two items *cinese* and *Cina* in order to test the hypothesis that the learner in question is not lacking a rule, nor is he/she wild-guessing or even backsliding in his/her developmental path, but simply that she’s/he’s applying some kind of rule that affects both occurrences. We don't know this rule yet nor can we easily figure out what kind of rule it is. Using any tag based on binary opposition (correct vs. incorrect) would be misleading. The solution is to sort out all "virtual adjectives" and "virtual nouns" (for a detailed meaning of "virtual", see next paragraph) containing similar strings of characters (in our case, *c-i-n* or the like) in different positions of the sentence. By repeating this query pattern throughout the sentences in the corpus, we might find out that "virtual" adjectives (like *cinese*) rather than "virtual nouns" (like *Cina*) are likely to be placed to the initial place, at the left periphery of the sentence (the typical topic-position in Chinese) and that this preferably happens when a noun (like *media*) occurs somewhere rightwards. Or we might find out that the differences in suffixation that we expect to be between adjectives and nouns (-*ese* vs. -*a* or zero-suffix) are systematically blurred when there is what we interpret as being a locative expression. If either of these combinations of facts recurs systematically in the corpus, then the grammar of the learner might contain a rule of the kind "position of items counts more than their eventual suffix" or "items in locative expressions agree, regardless their category". If, on the contrary, these combinations do not recur systematically, it is likely that the learner's grammar does not contain such rules or that our interpretation of the learner's sentences was wrong under some respects. Whatever the answer, since this procedure prevents researcher's interpretation from affecting the annotation of the sentence, sooner or later other unexpected linguistic features will surface from the corpus and new hypotheses will be made available to be systematically tested out on data.

TL Rule-motivated vs. Form-motivated, "virtual" categories

As Nicholls (2003: 572) pointed out, error tags are not an end in themselves, "but rather act as a bookmark" for queries, that is, they should give the researcher the information they are looking for. Contrarily, our point is that error tags are likely to commit comparative/closeness fallacy and to obstacle - instead of allowing - the retrieval of important IL phenomena because what they are likely to annotate is taggers' TL-governed interpretation (often just one among other possible interpretations), not the structural value of the item in the IL. In everyday experience, human interpretation is called into action to unpredictable extent when trying to make sense of learners' utterances. We can include it in the annotation consistently or completely exclude it from annotation at the cost of losing usability in the query system. The solution provided is a compromise between transparency of data and usability. On one hand we decided to exclude all interpretation based on taggers' judgements, on the other hand we encoded all interpretations based on automatic and successful matching between the item in question and all TL items. In our view, this would prevent running the risk of "ontologizing" errors, that is, to treat them as if they were really psychological *realia*, sort of holes or gaps existing in learners' mind. Functional interpretation is thus excluded and "virtual", formal-motivated (TL-oriented) tags substitute rule-motivated (TL-governed) tags by allowing different levels of annotation, as will be shown in next two paragraphs.

When a L1 tagger is run on a learner corpus

The key idea is to use a L1 tagger on the L2 corpus as a means of detecting virtual categories corresponding to each L2 item. In our opinion, far from being a step back, this would help minimize the risk of comparative fallacy and gain deep insights into learners' IL. Using a strictly formal definition we can identify a category by lexical root, by morphology or by context. There are formal hints that must be taken into account in order to recognize, say a verb in a sentence like "Loro andavano a scuola" (They went to school): post-pronominal position, a verbal root like "and-" (go), verbal inflection "-avano" (3 person plural imperfect). In L1 the criteria normally converge and tend to be redundant. Rule-based taggers for instance generally rely on morphology and lemma in conjunction, so they will only recognize known lemmas with the right morphology attached. In IL, on the contrary, not all criteria are always satisfied at the same time. So ideally we need a much more flexible tagger that takes into account all hints and expresses a possible tagging together with its level of confidence. We chose to use Treetagger (Schmid, 1994), with the standard tagset and the standard training for Italian L1 and obtained encouraging results. Being built on a probabilistic algorithm, Treetagger will recognize, say, a verb by the presence of either a verbal position, a verbal root or a verbal morphology. These levels are independent: the tagger recognizes a verbal ending even if this is attached to an unknown lemma. Therefore, once each word is analyzed, the tagger issues a tag, a lemma (which can be <unknown>) and a confidence probability, which is determined by the convergence of the different hints. A verbal tag with lemma <unknown> and a low level of confidence means that the lexical criteria failed and that the tagging was performed on the basis of position and (possibly) morphology.

Annotation sample

Once the annotation per category, lemma and probability is translated in xml tags, queries can be performed on the corpus, mixing the virtual categories level with positional information and formal data at the source level (via regular expressions matching). The tagset at word level is defined as follows:

<token> – grammatical word

attributes: **tag** – part of speech; **lemma**; **prob** – Treetagger confidence level

Here is a sample of annotated text:

"è un bambino che in la camera sua ha un cane e una rana..." (it's a child that in his room has a dog and a frog).

```
<token tag="VER:pres" lemma="essere" prob="1.000000">è</token>
<token tag="DET:indef" lemma="un" prob="0.998249">un</token>
<token tag="NOM" lemma="bambino" prob="1.000000">bambino</token>
<token tag="PRO:rela" lemma="che" prob="0.594519">che</token>
<token tag="PRE" lemma="in" prob="1.000000">in</token>
<token tag="DET:def" lemma="il" prob="0.999939">la</token>
<token tag="NOM" lemma="camera" prob="1.000000">camera</token>
<token tag="PRO:poss" lemma="suo" prob="1.000000">sua</token>
<token tag="VER:pres" lemma="avere|riavere" prob="1.000000">ha</token>
<token tag="DET:indef" lemma="un" prob="0.998249">un</token>
<token tag="NOM" lemma="cane" prob="1.000000">cane</token>
<token tag="CON" lemma="e" prob="1.000000">e</token>
<token tag="DET:indef" lemma="una" prob="1.000000">una</token>
<token tag="NOM" lemma="rana" prob="0.694963">rana</token>
<token tag="ADV" lemma="dentro" prob="0.830941">dentro</token>
<token tag="PRE" lemma="di" prob="1.000000">di</token>
<token tag="DET:indef" lemma="un" prob="0.997119">un</token>
<token tag="NOM" lemma="barattolo" prob="1.000000">barattolo</token>
<token tag="SENT" lemma="." prob="1.000000">.</token>
```

Basic queries

We give here just one example of how to query the tagged corpus in order to find IL features (including the so-called "errors") without any need of error tags, just by using the following information from Treetagger: (a) (form-motivated) virtual categories; (b) the level of confidence in the tagging and in recognizing the lemma; (c) strings and positional context. Note how here that the possible weakness in analysing IL with a TL tagger, with all recognition problems involved, turn out to become an advantage for the end user. Let's imagine we want to investigate the *transition from indiscriminate to* *selective verbal suffixation*, this being our starting hypothesis on the learner developing grammar. Here are some useful and very simple queries, using first lemma information, then adding confidence level information and finally position:

Query 1: search tokens with lemma <unknown> that have been tagged as verbs (at this stage the level of confidence is ignored). The query outputs contexts such as:

(1.a) il ragazzo **pienere** su la roccia per gritare
the boy <unknown>-verb:infinite on the rock to cry

(1.b) ogni giorno **conoscia** dieci persone
every day (he) meets$^{?}$ ten people

In (1.a) the system recognizes something that could resemble the infinite suffix "-ere" even if it is attached to an unknown stem. Maybe the learner is trying to categorise the token as verb by using verbal morphology: if this is the case, the tagger recognizes it. In (1.b) both root and morphological agreement are target like, but the lemma is not recognised.

Query 2: search all verbs with lemma NOT <unknown> which have been tagged with confidence less then 1.0. This captures all virtual verbs that have been recognised by Treetagger with some degree of uncertainty, like:

(2.a) quando si **sveglia** il bambino
when (refl) wakes up the child

(2.b) salì a la cime de una rocca. Continua **chiamandola**
climbed to the top of a rock. keeps calling (ger+clit.)

(2.c) è sotto una nave che si **sta** costruggendo
is under a ship that (imp.) is being built

Here we get a broader spectrum of phenomena, some of them unexpected and really interesting. We have target like sentences (2.a) in which a form that presents a categorial ambiguity in isolation (sveglia_noun, “alarm-clock” vs sveglia_verb3ps, “wake up”) is correctly disambiguated by context; well formed items in unexpected and possibly non target-like contexts, as in (2.b), where the presence of the verb “continuare” normally requires “a”+infinitive; ill-formed items like “costruggendo”, which apparently stems from the root of the TL verb “costruendo” ("build") in (2.c). Note that these contexts are retrieved without previously tagging them with any error category on purpose.

Query 3: search all sequences of token 1 and token 2 such as token 1 is a virtual verb with confidence < 1 (some degree of uncertainty) and with lemma <unknown> and token 2 is a virtual verb of any kind.

(3.a)	e	corri	corri	corri	il	bambino	sulla	testa
	and	run	run	run	the	child	on the	head

(3.b)	ho	dovuto	parlare	l'	inglese
	had	must	speak	the	English

(3.c)	e	quando	il	furgone	era	andato
	and	when	the	truck	was	gone

Here too a variety of phenomena is present in the output: conversational traits such a repetitions and false starts (3.a); target like compound verbs and verbal periphrasis (3.b, modal) in what one may judge being appropriate or inappropriate context. Again, since our point is that a certain amount of spurious results is proof of the absence of comparative fallacy, also the transparency of the data is thus being respected. Queries like these should be run on portions of the corpus divided by level (and by learner) in order to study the evolution of the phenomena in object.

Future Developments

Using XSL-transformations on XML allows us not only to query the corpus, but also to add further tags “online”. This can be implemented to allow the researchers to assign their own further levels of annotations, like tagging functions related to the sistematicity they might have found in the IL. These tags could be later combined with the others to perform "patterned queries", that restrict the search in a more fine grained and specific way without using any error tag.

References

Bley-Vroman, R. 1983. The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning, 33*: 1-17.

Díaz-Negrillo, A., Fernández-Domínguez, J., 2006. Error tagging system for learner corpora. *RESLA, 19*: 83-102.

Cook, V., 1997, Monolingua Bias in Second Language Acquisition Research. *Revista Canaria de Estudios Ingleses, 34*: 35-50.

Foster, J., 2007. Treebanks gone bad. Parser evaluation and retraining using a trebank of ungrammatical sentences. *International Journal on Document Analysis and Recognition, 10*: 129-145.

Granger, S., 2003, Error-tagged learner corpora and CALL: a promising synergy. *CALICO 20/3*: 465-480.

Huebner, T., 1979, Order-of-Acquisition vs. dynamic paradigm: A comparison of method in interlanguage research, *TESOL Quarterly, 13*: 21-28

Klein, W., Perdue, C., 1992, “Utterance structure”. In *Adult language acquisition: cross-linguistic perspectives. Vol.2: The results*, C.Perdue (ed.), Cambridge, Cambridge University Press.

Lakshmanan, U., Selinker, L. 2001. Analysing interlanguage: How do we know what learners know? *Second Language Research, 17*: 393-420.

Lüdeling, A., Walter, M., Kroymann, E., Adolphs, P. 2005. "Multi-level error annotation in Learner Corpora". Paper presented at the *Corpus Linguistics 2005 Conference*, Birmingham, U.K., www.corpus.bham.ac.uk/PCLC/Falko-CL2006.doc [access date 25/04/2008]

Nicholls, D., 2003, "The Cambridge Learner Corpus: error coding and analysis for lexicography and ELT". In *Proceedings of the Corpus Linguistics 2003 Conference*, Lancaster, United Kingdom.

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

Schmid, H., 1994, "Probabilistic Part-Of-Speech Tagging Using Decision Trees". *Proceedings of International Conference on New Methods in Language Processing,* Manchester, United Kingdom.

Tono Y., 2003. "Learner corpora: design, development and applications". In *Proceedings of the Corpus Linguistics 2003 Conference (CL 2003). Technical Papers 16*, D. Archer, P.Rayson, A.Wilson, T.McEnery (eds.), University Centre for Computer Corpus Research on Language, Lancaster, 800-809.

The authors

Stefano Rastelli is post-doc research fellow at the University of Pavia where he also works as Italian language coordinator. He has been teaching Italian as a foreign language since 1988. His main areas of interest are: Second Language Acquisition (the acquisition of tense-Aspect system), Syntactic Theory, Corpus Linguistics and Foreign Language Teaching.

Francesca Frontini is a Ph.D student at the University of Pavia. She is currently dealing with measuring the performances of stocastic algorithms on learner corpora. Her main areas of interest are: Computational Linguistics, Corpus Linguistics, Second Language Acquisition.