

Learner Corpora without Error Tagging

Stefano Rastelli (Pavia)

Abstract

The article explores the possibility of adopting a form-to-function perspective when annotating learner corpora in order to get deeper insights about systematic features of interlanguage. A split between forms and functions (or categories) is desirable in order to avoid the "comparative fallacy" and because – especially in basic varieties – forms may precede functions (e.g., what resembles to a "noun" might have a different function or a function may show up in unexpected forms). In the computer-aided error analysis tradition, all items produced by learners are traced to a grid of error tags which is based on the categories of the target language. Differently, we believe it is possible to record and make retrievable both words and sequence of characters independently from their functional-grammatical label in the target language. For this purpose at the University of Pavia we adapted a probabilistic POS tagger designed for L1 on L2 data. Despite the criticism that this operation can raise, we found that it is better to work with "virtual categories" rather than with errors. The article outlines the theoretical background of the project and shows some examples in which some potential of SLA-oriented (non error-based) tagging will be possibly made clearer.

1 Research background

1.1 POS annotation and error tagging

The topic of this article is the Part-of-Speech (POS) annotation of learner corpora (LC). The research question is whether it is feasible and convenient to instruct an automatic tagger which is capable of recognizing and annotating the grammatical categories in learner data. If we assume that – in each one of its different stages – a learner's interlanguage (IL) always "targets" the target language (TL), a problem arises when an external observer (teacher or researcher) tries to represent consistently and steadily the inherently unstable nature of learner data. Traditionally, it is believed that L1 taggers are useless because they are unable to capture the divergent phenomena occurring in LC. In fact, misspelled, badly uttered, incomprehensible and not interpretable items are destined to escape the formal requirements of automatic analyzers and of robust parsers. To face this issue, two different solutions are at hand. The first one, which is widely accepted in literature and which is consequently adopted in many European projects is that learner data is best viewed in terms of errors. Most known LC (e.g. the ICLE project, cf. Granger 2003 and 2005) are error-tagged. The POS error-tagging procedure is made up of three steps: (a) collecting learners' typical mistakes all together in a list (typical mistakes/errors with respect to homogeneous groups of learners); (b) turning this list into errors related to traditional linguistic categories (such as errors in nouns, adjectives, verbs etc); (c) tagging the items in the list using a markup language (for instance, XML). In order to turn typical mistakes into tags more layers of annotation are often established. For instance, the error-tagging procedure adopted in the FRIDA (French Interlanguage Database) project comprises three levels: error-domain, error-category and word-category. In more abstract terms, in an error-based tagging system, each well-formed item

produced by learners is recorded in terms of a L1 function or grammatical category and in terms of its attributes or properties as well (e.g. *rana* 'frog' is a noun, its attributes being "feminine", "singular"). Instead, if one item is badly uttered, misspelled, not recognized or – in a word – wrong, this information is recorded as error and the item is either normalized (altered to the correct form) or recorded as a whole in a list of incorrect items (structured on error-types and error-categories) ready to be retrieved as such by purposely designed software. For instance, the item **rano* (instead of *rana*) would possibly be listed (by an error-editor) under the category "error in gender" or more specifically "masculine instead of feminine gender" if software for error-retrieval is trained to associate the endings *-o* and *-a* to most masculine and feminine nouns in L1 Italian. Since the automatization of error-tagging is a (maybe unnecessarily) complicated issue, most projects require a procedure of manual tagging and usually provide researchers engaged in this task with a tagger's manual or with guidelines of some kind. After a LC has been tagged with error-tags by using, for instance, a markup language, all occurrences are retrievable with software (for instance, with WordSmith Tools or Xaira).

1.2 Surface phenomena and acquisitional facts

Another possible solution for the treatment of LC – which has been adopted at the University of Pavia – is called "SLA-tagging" (cf. Rastelli 2006; Rastelli 2007; Andorno and Rastelli 2007; Rastelli and Frontini 2008). This representation of learner data is meant to be an alternative to error-tagging. In fact it starts from the assumption that a too strict (L1-governed) view on learners' data is inadequate for SLA research because the primary interest of SLA research is the IL. From the IL perspective, both wrong and correct items are learners' attempts to map tentatively forms and functions. The mere fact that a certain form appears to be wrong or correct does not say enough to us about neither learners' difficulties nor about learners' mental grammars. Instead, SLA-tagging is aimed at helping researchers discover systematicity (or lack) in how learners map forms and functions and in how they gradually develop their knowledge of the TL categories out of the available input. To these purposes, the procedure of error-tagging is unreliable for at least three reasons: (a) it often fails to restrain the boundaries of errors and to detect the source of errors in a learner's mental representation; (b) it is often inconsistent and unreliable because it is subject to tagger's interpretations; (c) it upgrades surface phenomena to the rank of acquisitional facts. As an example for the situation (a), sentence (1) is written by an American beginner learner of Italian (cf. Rastelli 2006):

- (1) *la ragazza detto fare passeggiare*
 the girl said-PastPART make-INF stroll-INF
 *'the girl said making strolling'

In this sentence it is possible neither to detect nor to isolate the mistake. It seems that the whole sentence is ill-formed and that – at the same time – the source of this ill-formedness is undetectable with precision. In learner data often it is not possible 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, cf. Rastelli 2007). One reason could be that that 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, cf. Lüdeling et al. 2005). Another reason could be that, especially in basic varieties, learners often produce not just lacking, wrong or misspelled 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" means unclassifiable and unpredictable. Unclassifiable is a combination of a number of *per se* well-formed items,

which 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 matter its size. Sentence (1) is both unclassifiable and unpredictable: as a matter of fact we do not have a single clue to decide what this sentence means. Even if one can guess the meaning of each single word taken in isolation, the overall sense of the sentence remains totally obscure. In such a situation, how could a researcher project the items into one or another suitable error category? As an example for the situation (b) is represented by sentence (2), where a Chinese beginner student of L2 Italian, describing a house (with some people inside) suddenly catching fire, says:

- (2) *la casa di loro c' è fuoco*
 the house of them there is fire
 'in their house there is a fire'
 'their house is catching fire'

Here again none of the items of this sentence taken individually is wrong, nor – again – is it straightforward to pinpoint the source of the ill-formedness. Differently from (1), in this case the scene described is clear. Yet this is not enough in order to label the possible errors unambiguously because there are at least two different ways to correct the sentence and there are two possible interpretations. Far from being an exception in learner data, sentences like the one above show that 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. That's why – as it has been often pointed out – the practice of error tagging rests on native speaker's intuition, that is, on just the tagger's peculiar interpretation among other possible interpretations. As an example for the situation described in (c), sentence (3) is written by another Chinese student of Italian:

- (3) *visto ragazze uscita porta*
 seen-PastPART girls gone out-PastPART door
 'I saw (that) two girls went out of the door'
 'I saw two girls out of the door'

Since in L1 Italian compound tenses mandatorily display either auxiliary *avere* 'have' or *essere* 'be', the surface phenomenon in (3) is that the verbs *visto* and *uscita* are lacking their auxiliaries. The procedure of labeling "missing auxiliaries" as an error is at risk of being misleading for SLA research purposes. In fact the error-tag "missing auxiliary" seems to indicate that the learner violates a real rule of the TL. In which sense do we say that a 'missing auxiliary' violates a rule? Does such a rule exist in a learner's mind? Is it a rule of the TL or a rule of the IL? The procedure of error-tagging is risky because it ontologizes TL grammar rules as if they were psychological *realia* in a learner's mind. For this reason the error-tagging procedure is unlikely to help disclose how IL works and to help identify some possible IL rules. At most, it just helps to give an idea of how far IL is from the TL in respect to a restricted number of surface phenomena, which – in their turn – often represent only the mere consequences and not the deep causes of the changes occurring in the IL and the dynamics of the learning process. In the case outlined above, the mere presence or absence of auxiliaries in learner data does not constitute evidence that auxiliaries are – respectively – ignored or represented in a learner's mental grammar. It might be that spelled out auxiliaries are simply unanalyzed chunks and formulas or – conversely – that they are purposely omitted or dropped by learners in order to promote the aspectual value and to demote other values (temporal, personal etc.) of verbs. In SLA tagging, a learner's mental grammar is not looked up and accounted for in terms of its distance from the TL grammar because, even though IL targets

the TL, the rules of the former do not overlap with those of the latter. Since SLA research is concerned about how second languages are learned, SLA-tagging is concerned about the systematicity of learners' IL (its rules), not about the distance between IL and TL.

2 The SLA perspective and the comparative fallacy

2.1 SLA tagging and the rules of interlanguage

Examples (4) and (5) are taken from the Pavia Corpus. A Chinese beginner student of L2 Italian, when asked to report about his education, says:

- (4) *Cinese fatto media*
 Chinese-ADJ done-PastPART middle
 'In China I attended the middle school.'

A few days later, when asked about holidays, the same learner says:

- (5) *Sì, in Cina festa pasqua anche*
 Yes in China holiday Easter too
 'Yes, in China there are Easter Holidays as well.'

Following the bracketed interpretation and under an error-driven perspective, only in (4) 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 – for instance – the appropriate error category outlined in the 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). But in the SLA-tagging perspective, we might compare the two items *cinese* and *in 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 over both occurrences. In SLA research, the mapping between form and function is not taken for granted: it has to be reconstructed inductively.

As a matter of fact, we don't know this rule yet nor can we easily figure out what kind of rule it is. The procedure of using tags which are based on a binary opposition (correct vs. incorrect) would be misleading. The proper solution to us could be 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 to the right. Or we might find out that the differences in suffixation that we expect to be between adjectives and nouns (*-ese* vs. *-a* or \emptyset -suffix for nouns) 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 "only position of items counts and disregard the rest (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 in 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.

2.2 The risk of comparative fallacy

SLA tagging is aimed at accounting for the IL of this Chinese student and to avoid comparative fallacy. Avoiding "comparative fallacy" and "closeness fallacy" (Huebner 1979; Bley-Vroman 1983; Klein and Perdue 1992; Cook 1997; Lakshmanan and Selinker 2001) is the real asset of SLA-tagging methodology. 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 and 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 (1983: 2) criticism 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 still unknown, that is, hidden to any L1-governed observation. 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 (see §4) and to hide developing and provisional non target-like learner grammars, in our project it was decided to find an alternative way to run queries on learner corpora. This query system requires being TL rule-oriented and not TL rule-governed, thus 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 mature languages. By doing so we try to separate drastically forms and functions: L2 forms are tagged automatically, while L2 functions should be restored inductively by researchers.

3 Defining L2 "virtual" TL categories by using probabilistic criteria

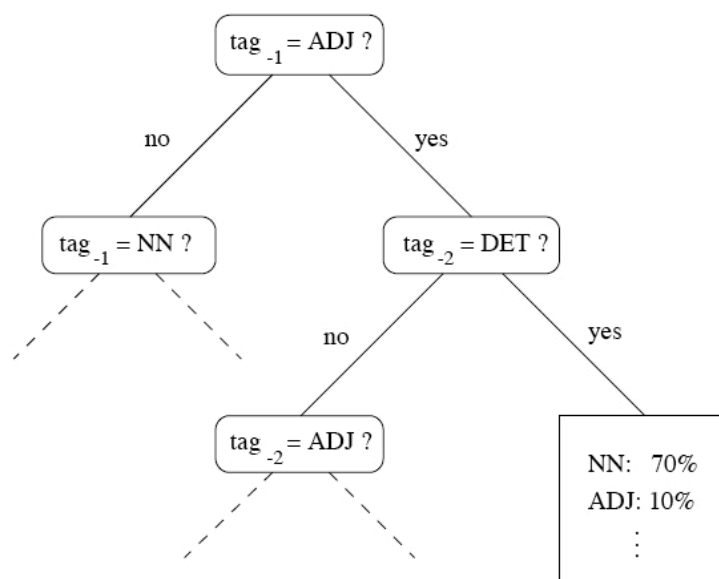
3.1 Running a L1 tagger on L2 data

In the previous paragraph it has been maintained that, in order to ensure the transparency of IL, L2 forms should be looked at without a L1-governed view, that is, without superimposing the corresponding TL functions. One possible solution – in order to avoid comparative fallacy – is to look at both correct and divergent items in LC the same way, that is, completely disabling functional interpretation and allowing only a classification based on the possible misalignments between the position of items in the sentence, their nearness (adjacency) and their formal features (lexical root + morphology). In first languages, the alignment of these three criteria results in functional tagging, that is, in automatically attributing items to one grammatical category or another (nouns, verbs etc.). In second languages, the misalignment of position, adjacency and formal features turns out to be a clue that the grammatical categories of the TL are in the process of being learned. When there is a divorce between position, adjacency and form, it cannot be captured by super-imposing a functional interpretation. In fact, when we say – for instance – that *fare* "make" in sentence (1) is a "wrong verb", how can we be sure that the learner meant it to be a verb? SLA research would take much more advantage if it could identify all that works as a verb despite not resembling formally L1 verbs or despite being misplaced in the sentence, in respect to where TL verbs are expected to be placed. This is something that would offer us a unique opportunity to peep at a learner's mental grammar. Consequently, our proposal is to run a L1 probabilistic Treetagger (cf. Schmid 1994) on L2 data and to start SLA research just from the (sometimes) fuzzy results

obtained in this way. Running a L1 tagger on L2 data may sound like a paradox: how can it be that an evident weakness (the inability of L1 parser to deal with and classify divergent phenomena) may become a strong asset for SLA research? Probabilistic taggers designed for L1 (for instance, the Italian version of Schmid's Treetagger) are not designed to help researchers' accounting for L2 functions and for deviant phenomena. Nevertheless, when run on learner data, they proved to be capable to assign what we can consider "virtually" linguistic categories (Part-of-Speech categories) also to L2 items only by means of distributional constraints, formal resemblance, co-occurrence patterns and confidence rates. The basic idea is that L2 researchers can afford giving up functional algorithms (normally used to define what are L2 verbs, adjectives in terms of government) and limit themselves to adopt only distributional, frequency and formal algorithms. By doing so, they can assign a certain virtual category to any L2 item (for instance, to the item *cinese* in (4)) basing just on the values of three attributes: its position, its co-occurrence patterns, its formal features (the resembling of a part of the characters it is made of to one or more TL items). Whether a L2 item is – say – a "real" adjective (whether it behaves like a TL adjective, for instance, adding quantitative or qualitative information to a noun) is something that still needs to be demonstrated. In fact, since establishing form/function mapping is the aim of SLA research, it cannot be its starting point. The ground for attributing the belonging of an item to a virtual category is not looking at what that item does in the sentence, but looking at its position, at its neighbors and at whether the characters it is made of match one or more forms in the machine-lexicon. Since the matching between an item and a virtual category is established on a probabilistic basis, a researcher can raise or lower the confidence rate, and – for instance – asks the software to sort out all adjectives which are not likely to be adjective (low confidence rate) because somehow they do not resemble TL adjective, lacking the proper suffix or root or being placed in an unexpected position or being sided by an unexpected element. It is precisely under a low confidence rate that a researcher would obtain also highly heterogeneous items which share one or more properties with TL adjectives. In this way one gets an approximate idea of which L2 items work as an adjective in the IL and of how the category "adjective" shapes in a learner's mental grammar. Furthermore, since our starting point is not the target (form+function) categories but virtual categories, we are enabled to rely on our native competence (or our linguistic competence) to sort out data without committing comparative fallacy.

3.2 L2 researchers can take advantage of how the Treetagger works

Helmut Schmid's Treetagger (Schmid 1994) uses a decision tree to assign a certain category to a word and to estimate the probabilities that ambiguous words fall within one category or another. The principle is that any assignation is predictable from the context. The decision tree is different from other methods (such as probabilistic or rule-based) because it determines differently the size of the context (how many words before and after) needed for assignation and because it copes with ungrammaticalities in the input. The probability of a word to be paired with tag (category) is determined through a number of tests from the higher nodes to the leaves of a decision tree like the one below (taken from Schmid 1994: 3).



This decision tree can determine the probability that the third element of a trigram is a noun through recursion steps which are made of alternative "yes-no" paths. If the immediate antecedent word (tag_{-1}) is an adjective and if the previous tag_{-2} is a determiner, then the percentage that the third tag is a noun is 70%. Crucially, a "yes" answer occurs when the tag is identical to a form which is stored in the machine lexicon. The lexicon is threefold: there is a *fullform lexicon*, a *suffix lexicon*, and a *default entry*. The fullform lexicon is searched first, if the tag is not found there, then the suffix is compelled. If also this step has not been successful, the default entry of the lexicon is returned (this is the case for ungrammaticalities, for errors in the tagging process and also for underrepresented items). The Italian version of the Treetagger contains 860.332 fullforms, 747 suffix nodes and 52.338 ambiguous forms (which may occur in various contexts). There are different degrees of probabilities in which a given word can match a fullform or a suffix form stored in the machine lexicon. The lexicon displays the probabilities that a certain tag matches a certain category. These probabilities are returned as "tag probability vector", which is a number associated to each tag. As a consequence, once a learner corpus has been xml tagged with the Treetagger, one can ask all occurrences which are more or less likely to be – say – "nouns" or "verbs" because the confidence rate is an attribute of each tag in the markup language. One can also choose to disable the fullform lexicon and operate only on suffixes, thus obtaining occurrences which are likely to be – say – nouns or verbs in a learner's interlanguage despite the fact that their root is unidentifiable ("lemma unknown"). Both the fullform lexicon and the suffix lexicon are built from a (previously) tagged training corpus, that is, on a L1 corpus where each occurrence has been paired with a tag (category). Admittedly, when adopted for learner corpora, this perspective might turn out to be too restrictive. In fact, if we want to keep the idea of "virtual categories", more permissive conditions need to be introduced in the decision tree and high/low probability vectors need to be consequently adjusted and fine-tuned. Particularly, it is very likely that SLA researchers – in order to reconstruct the rules of interlanguage – have to deal even with very low confidence rates and have to keep track and store in the threefold lexicon also deviant forms which tend to recur systematically. The tagset need to be widened in order to encompass also stable, although tentative, form/function mappings.

3.3 Examples of "virtual" categories

Let's do an example of how Treetagger extracts occurrences on a corpus of learner Italian by American undergraduate students (Rastelli 2006). Let's suppose that we want to sort out all occurrences of virtual adjectives preceded by a modifier (whatever quantifier, determiner or

adverb) and according to a low rate of confidence. Among other occurrences, we obtain sentences (7) and (8) where virtual adjectives are in bold:

- (7) *lui non è **simpatico** e la donna molta **trista***
 he not is funny and the woman much-ADJ-Fem sad-ADJ-Fem
 'he's not funny and the woman is very sad'

What counts more to us is that these two virtual adjectives are listed together in virtue of different criteria. The item *simpatico* 'funny' is recognized as being an adjective because it matches perfectly the corresponding TL form both in the lexical root and in the suffix. The item *trista* 'sad' is preceded by the item *molta* with whom it seems to agree in gender (they share the typical Italian feminine ending *-a*). Differently from *simpatico*, *trista* is signaled as being an adjective in virtue of its lexical root but in spite of its wrong suffix (*trist-a* instead of *triste-e*) and because it is preceded by *molta*. Interestingly, *molta* is recognized as being an adverb in virtue of the frequency of its co-occurrence pattern even if the proper form would be *molto* (ADV, 'very'). This heterogeneous list of virtual adjectives is obtained because – when set on a low confidence rate default value – the Treetagger, in order to assign the category to an item, scales down or promotes formal requirements (e.g. correct suffixes) or – alternatively – utilizes distributional and frequency criteria. The advantage of having collected together in the same sorting in virtue of different criteria *simpatico*, and *trista* enables us to do something we probably could not do if we stick to the traditional correct/incorrect perspective. For instance, by looking carefully at the list, maybe we are put in the ideal conditions to tackle the process through which the category of virtual adverbs (*molta*) emerges along with (or after) the category of virtual adjectives and the whole process of acquisition of the properties of adjectival phrases. In sentence (8) the item *sbagliato* is formally equivalent to the TL adjective 'wrong' but it is in the typical post-determiner position of a noun and also displays the co-occurrence pattern of a noun:

- (8) *quando la donna è in bagno lei faccia un **sbagliato***
 when the woman is in bathroom she does-SUBJ a wrong-ADJ
 'when she is in the bathroom she makes a mistake'

Also in sentence (8), formal features conflict with distributional features, but this time the former outranks the latter. Virtual adjective (and virtual categories in general) are often the result of a conflict in assignation criteria. An item which fits perfectly a TL category in virtue of its position, might not match a formal feature (e.g. the lexical root or the suffix) or the opposite. Instead of considering *sbagliato* 'wrong' an adjective used in place of the noun *sbaglio* 'mistake', we find it preferable to avoid the risk that a superficial labeling sets apart items that could be accounted for as being together. The theoretical assumption which underlies the choice of virtual categories – beside the need to avoid comparative fallacy – is that adjacency more than government could be what drives basic learners in the process of constructing a sentence (cf. Tomasello 2003; Ellis 2003) and that "chunks" (rather than phrases) are possibly more appropriate units for the analysis of initial learners' IL.

4 The unexpected data

Have a L1 Treetagger do (the first part of) the job on L2 data is feasible and convenient for the purposes of SLA research. Instead of getting only the items that we expect, we could have the whole lot of them together at a glance also items that we don't expect. For instance, instead of sorting out only occurrences of wrong/correct adjectives identified and collected just from the positions in the sentence where we – as native speakers – expect adjectives to be, we could find out that there are other items that work as adjectives despite not resembling

TL adjectives. In fact, a L1 Treetagger can also gather apparently unrelated and scattered items much better than us. While we are naturally pushed to superimpose the TL functional interpretation, the Treetagger runs different algorithms at the same time and compares and balances three different weights: the form, the position and the nearness of each item. The possible conflicts in these different algorithms in a certain sense reflect the process of reconstruction of learners' grammar where categories (form/function mappings) are only tentative categories, subject to refinements over time. It is feasible and convenient for SLA research if we can get a list of virtual (also fuzzy, heterogeneous and often spurious) adjectives instead of a list of wrong or correct adjectives seen from the TL perspectives. On the other hand, an evident flaw of the probabilistic tagger (the impossible matching of deviant forms with the machine-dictionary) may again turn out to be an advantage for SLA research. In fact, if we disable the access to the machine-lexicon and select the option that allows us to sort out occurrences with the attribute "lemma" on the value "unknown", virtual adjectives are sorted only in virtue of their position, regardless of their form. Sentence (9) illustrates a case of a "purely virtual" adjective (in bold)

- (9) *quando lei aspetta, incontra una donna che era una **disastra***
 when she waits-PRES for meets a woman which was a disaster
 'while she is waiting, she runs into a woman which was a disaster'

In L1 Italian there exists the fixed expression *essere un disastro* 'be a disaster' where *disastro* is the noun which is a part in a collocation and thus it is not variable in gender and number. Instead, in (9) *disastra* is sorted as an adjective only in virtue of its collocation and of the final vowel *-a* (which is a candidate to agree with the noun *donna* 'woman' which is close to it) but not in virtue of its lexical root. By looking at this sentence, we can obtain much information about how this learner represents the category of adjectives in its mental grammar. This information would be basically lost (or at least much more hard to retrieve) if we had decided that *disastra* is a wrong noun (instead of the correct item *disastro*). Unexpected data may bring us closer to IL. It may represent and reflect the interaction and the sometimes unpredictable outcomes of different organizing principles, which compete even for a long time in a learner's mind.

5 Future research

The algorithms which compose the probabilistic tagger might be fine-tuned to L2 data, not in order to run up deviant (impossible or unpredictable) forms with a more complicated TL-governed grid of errors, but in order to encompass and admit a larger array of phenomena. The joined work of computational linguists and SLA researchers could be to loosen the constraints which orient the automatic classification of L2 forms in order to allow multiple POS assignments.

References

- Andorno, Cecilia/Rastelli, Stefano (2007): "The road not taken: on the way to a parser of Learner Italian". In: Sansò, Andrea (ed.): *Language resources and linguistic theories*. Milano, Franco Angeli: 83–95.
- Bley-Vroman, Robert (1983): "The comparative fallacy in interlanguage studies: The case of systematicity". *Language Learning* 33: 1–17.
- Cook, Vivian,(1997): "Monolingua Bias in Second Language Acquisition Research". *Revista Canaria de Estudios Ingleses* 34: 35–50.

- Ellis, Nick (2003): "Construction, Chunking and Connectionism: the emergence of second language structure". In: Doughty, Catherine/Long, Michael (eds.): *The handbook of second language acquisition*. London, Blackwell: 63–103.
- Foster, Jennifer (2007): "Treebanks gone bad. Parser evaluation and retraining using a treebank of ungrammatical sentences". *International Journal on Document Analysis and Recognition* 10: 129–145.
- Granger, Sylviane (2003): "Error-tagged learner corpora and CALL: a promising synergy". *CALICO* 20(3): 465–480.
- Granger, Sylviane (2005): "Computer Learner Corpus Research: current status and future prospects". In: Connor, Ulla/Upton, Thomas (eds.): *Applied Corpus Linguistics: a multidimensional Perspective*. Amsterdam/Atlanta, Rodopi: 123–145.
- Huebner, Thom (1979): "Order-of-Acquisition vs. dynamic paradigm: A comparison of method in interlanguage research". *TESOL Quarterly* 13: 21–28.
- Klein, Wolfgang/ Perdue, Clive (1992): "Utterance structure". In: Perdue, Clive (ed.): *Adult language acquisition: cross-linguistic perspectives*. Vol. 2: *The results*. Cambridge: Cambridge University Press.
- Lakshmanan, Usha/ Selinker, Larry (2001): "Analysing interlanguage: How do we know what learners know?". *Second Language Research* 17: 393–420.
- Lüdeling, Anke et al. (2005): "Multi-level error annotation in Learner Corpora". *Corpus Linguistics 2005 Conference*. Birmingham, U.K. Available at www.corpus.bham.ac.uk/PCLC/Falko-CL2006.doc, accessed 25/04/2008.
- Rastelli, Stefano (2006): "ISA 0.9 – Written Italian of Americans: syntactic and semantic tagging of verbs in a learner corpus". *Studi Italiani di Linguistica Teorica e Applicata (SILTA)* I: 73–100.
- 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 /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.
- Schmid, Helmut (1994): "Probabilistic Part-Of-Speech Tagging Using Decision Trees". *Proceedings of International Conference on New Methods in Language Processing*, Manchester, United Kingdom.
- Tomasello, Michael (2003): *Constructing a language: a usage-based theory of language acquisition*. Cambridge MA: Cambridge University Press.