Linguistically Annotated Learner Corpora: Aspects of a Layered Linguistic Encoding and Standardized Representation

Detmar Meurers, Holger Wunsch
Seminar für Sprachwissenschaft, Universität Tübingen
{dm,wunsch}@sfs.uni-tuebingen.de

1 Introduction

Linguistically annotated corpora that are stored in standardized digital form can be a valuable source of empirical insight. They can help verify linguistic generalizations and support the formulation of new hypotheses. The linguistic annotation of such corpora often is crucial for their effective exploration from a linguistic perspective. The annotation essentially serves as an index to the linguistic classes and phenomena realized in the corpus (cf., e.g., Meurers 2005).

The situation in principle is parallel in the field of Second Language Acquisition (SLA) research, where an increasing number of corpora consisting of language as written by language learners have been compiled (Granger 2008). Yet, the linguistic annotation of learner data has received virtually no attention so far, apart from the so-called error annotation marking language properties which differ from native language patterns (Díaz-Negrillo and Fernández-Domínguez 2006). This is surprising given that prominent strands of SLA research are concerned with researching linguistic regularities in the stages of the acquisition process (cf., e.g., Pienemann 1998), irrespective of whether they are erroneous or not. Learner language is typically viewed as a linguistic system worth characterizing in its own right, so-called interlanguage. Thus learner corpora require systematic linguistic annotation of both correct and incorrect structures for them to effectively support the empirical questions addressed by SLA research (Díaz-Negrillo et al. (2009), Meurers 2009, Rastelli 2009).

In this paper, we report on work in progress investigating the linguistic annotation of learner corpora in terms of two aspects. We first motivate a new perspective on the part-of-speech (POS) categories of learner language and report on its implications for automatic POS tagging. Secondly, addressing a technical prerequisite of this work, we argue for a standardized representation format for annotated learner data.

2 Learner Language: Diverging Sources of Evidence

Part-of-speech analysis relies on three major sources of evidence, which are lexical evidence, morphological evidence, and evidence drawn from the distribution. For example, in the sentence *'I was surprised by the word **of** the day.'*, looking up the token '*of*' in a lexicon reveals that it can unambiguously be classified as a preposition. For words not listed in a given lexicon, morphological clues can still provide POS information. For instance, in '*His son **brachiated** along the monkey bars.'* the verbal past tense suffix *-ed* provides morphological evidence that the token '*brachiated'* is likely to be a verb. Where lexis and morphology do not unambiguously identify the POS of a word, evidence from the distribution of the word in the sentence can help resolve the ambiguity. In the sentence '*The old **man** the boat.'*, the distributional context of *'man'* identifies it as a verb, even though the more common category for this word is noun.[1]

The three examples above show the three sources of evidence separately. For native language, the three are assumed to be compatible sources of evidence jointly contributing to a single classification. Correspondingly, automatic POS taggers generally make use of all three sources of evidence to decide on the annotation.

For learner language, we argue in Díaz-Negrillo et al. (2009) that the assumption of the three sources of evidence contributing to and converging on a single part-of-speech annotation is rather problematic. We identified four types of mismatches in NOCE (Díaz-Negrillo 2007), a corpus of texts written by Spanish undergraduates at an upper-intermediate to advanced level of English. One of these classes, involving a mismatch between the stem and the inflectional morphology as well as between the stem and the distributional context of the word, is illustrated by (1) and (2).

(1) […] *one of the favourite places to visit for many **foreigns***. GR-1-C-EN-024-F

(2) […] *to be **choiced** for a job* […] GR-1-A-EN-003-X

In (1), a token which stem lookup identifies as an adjective appears in a nominal distribution slot following a determiner. The nominal distribution is compatible with the plural morpheme '*-s'* (which alternatively could also be the verbal third person singular morpheme). In sum, the token *foreigns* in (1) is classified as an adjective according to its lexical stem, but as a noun according to its distribution and morphology. In (2), the word *choiced* distributionally appears in a verbal slot and morphologically it carries verbal inflection (*'-ed'*), but lexically the stem is a noun (or adjective).

Complementing the theoretical question we raise in Díaz-Negrillo et al. (2009) of what constitutes part-of-speech categories which are empirical adequate and useful for characterizing learner language, we were curious to investigate the practical

[1] This specific ambiguity is hard to resolve automatically given that *old* is equally ambiguous between adjective and noun so that the local distributional context is not a clear indicator.

impact of such mismatches on automatic part-of-speech assignment. We therefore tagged the NOCE corpus with three POS taggers: *TreeTagger* (Schmid 1994), *TnT* (Brants 2000), and the *Stanford POS tagger* (Toutanova and Manning 2000), all of which were trained on the same native English corpus, the Wall Street Journal section of the Penn Treebank. The interest in using these three taggers is that they employ different fallback strategies and heuristics for dealing with unknown lexical forms and contexts. As expected, the tokens for which the three taggers differ in their tag assignment often turn out to be points where the learner language provide diverging evidence. On the one hand, this means that such a multi-tagger setup can be used to automatically detect certain learner errors. The next step in that direction would be a quantitative analysis of the nature and quality of such error detection. On the other hand, as we argued in the beginning of the paper, such a focus on errors runs counter to the SLA perspective on learner language in terms of systematic interlanguage development. For creating annotated learner corpora as a resource for SLA research we instead plan to develop an automatic tagging approach which separately annotates the three sources of evidence. This allows the lexical, morphological, and distributional annotation layers to encode conflicting information (which indirectly can be interpreted as errors in terms of the traditional POS system of the language to be learned, the L2). It also makes it possible to specify each of the three layers in terms of the targeted L2 properties as well as in terms of selected parts of the native language system of the learner to encode transfer phenomena.

3 Standardized Encoding of Learner Corpora

At first sight, the question of what data format is used to digitally store a linguistic resource may seem a minor and purely technical issue. But in fact, it is an important prerequisite for both *sustainability* and *usability*. For native language corpora, there has been a growing trend towards using XML-based formats for storage. XML is defined only in terms of a *normative specification*, which is strictly machine-independent. The data can thus be accessed across different types of computers with standard software, and it will remain accessible on future generations of machines. In terms of usability, the comparison of linguistic information across different resources can provide valuable insight. By representing information in a standardized form, it can uniformly be accessed in all resources. Such standards exist for native language corpora, particularly in the TEI corpus encoding guidelines (TEI Consortium 2009).

For learner corpora, it is particularly important to be able to explicitly link the texts to information about the learner and the task context in which the texts have been produced (cf., e.g., Meurers, Ott and Ziai 2010). Yet, the TEI guidelines so far do not support the comprehensive specification of the meta-information relevant for learner language. We therefore are compiling an extension to the TEI guidelines in the form of an additional module which bundles existing relevant specifications in other TEI modules and fills in the gaps with new definitions where necessary.

References

Brants, T. (2000). TnT – A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*. pp. 224–231.

Díaz-Negrillo, A. (2007). *A Fine-Grained Error Tagger for Learner Corpora*. Ph.D. thesis, University of Jaén, Spain. Revised version published 2009 as *EARS: A User's Manual.* Munich, Germany: LINCOM Academic Reference Books.

Díaz-Negrillo, A. and Fernández-Domínguez, J. (2006). Error Tagging Systems for Learner Corpora. Revista Española de Lingüística Aplicada (RESLA), 19, 83-102.

Díaz-Negrillo, A., D. Meurers, S. Valera and H. Wunsch (2010). Towards Interlanguage POS Annotation for Effective Learner Corpora in SLA and FLT. *Language Forum*. in press. http://purl.org/dm/papers/diaz-negrillo-et-al-09.html

Granger S. (2008) Learner Corpora. In Lüdeling A. and M. Kytö (eds) *Corpus linguistics: an international handbook, vol. 1.* Mouton de Gruyter. pp. 259–275.

Meurers, 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-09.html

Meurers, D. (2009). On the Automatic Analysis of Learner Language. Introduction to the Special Issue. *CALICO Journal.* 26 (3), 469–473. http://purl.org/dm/papers/meurers-09.html

Meurers, D., N. Ott, R. Ziai (2010). Compiling a Task-Based Corpus for the Analysis of Learner Language in Context. *Proceedings of Linguistic Evidence 2010, Tübingen* [this volume]. http://purl.org/dm/papers/meurers-ott-ziai-10.html

Pienemann, M. (1998). *Language Processing and Second Language Development: Processability Theory*. John Benjamins.

Rastelli, S. (2009). Learner Corpora without Error Tagging. *Linguistik online* 38, 2/2009. http://www.linguistik-online.de/38_09/rastelli.html

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK.

TEI Consortium (2009). *TEI P5: Guidelines for Electronic Text Encoding and Interchange.*

Toutanova, K. and C. Manning (2000). Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In *Proceedings of the Joint SIGDAT Conference on EMNLP and VLC*. pp. 63–70.