LANGUAGE FORUM
VOL. 36, NO. 1-2, JAN-DEC 2010

Towards Interlanguage POS Annotation for Effective Learner Corpora in SLA and FLT

ANA DÍAZ-NEGRILLO
University of Jaén, Spain

DETMAR MEURERS
University of Tübingen, Germany

SALVADOR VALERA
University of Jaén, Spain

HOLGER WUNSCH
University of Tübingen, Germany

ABSTRACT

Learner corpora can serve as a teaching resource for Foreign Language Teaching (FLT) and contribute empirical insights for Second Language Acquisition (SLA) research. To support effective querying for the specific classes of data which are relevant under the FLT and SLA perspectives, learner corpora ideally should include linguistic annotation. We argue for an approach to Part-Of-Speech (POS) tagging of learner corpora that systematically encodes the distributional, morphological, and lexical aspects specific to such interlanguage. Based on NOCE, an English learner corpus by Spanish learners, we characterize areas where the properties of learner language systematically differ from those assumed by POS annotation schemes developed for native language.

INTRODUCTION

Generally speaking, learner data is the empirical basis of Second Language Acquisition (SLA) research, and it exemplifies typical stages and common learner problems in Foreign Language Teaching (FLT). Such data collected in learner corpora can help validate generalizations about language acquisition and support the development of new hypotheses and theories in SLA.[1] Learner corpora can also play a role in identifying areas of relevance for FLT practice and materials design.

To find relevant classes of examples, the terminology used to single out learner language aspects of interest needs to be mapped to instances in the corpus. Effective querying of corpora for specific phenomena often requires reference to annotations (cf., e.g., Meurers 2005; Meurers & Müller 2009). Annotations essentially function as an index to classes of data which cannot easily be identified based on the surface form. For example, finding all sentences containing modal verbs using only the surface forms is possible, but would require a long list of all forms of the modal verbs. Even so, sentences where, for example, "can" is not actually a modal verb (e.g. "Pass me a *can* of beer" or "I *can* tuna for a living") would be wrongly identified. Other search patterns, such as a query for all sentences containing past participle verbs, cannot even be specified in finite form using the surface string alone. The annotation of corpora thus serves an important function, but also raises the question what type of learner language annotations are needed to support the searches for the data which are important for FLT and SLA research?

A traditional focus of research on learner corpora has been the identification and classification of learner errors. As pointed out by Granger (2003), learner corpora can help overcome some of the key problems of the Error Analysis strand of SLA research in the 70s and 80s (cf. Richards 1974; Corder 1981). And indeed accuracy remains an important issue of interest to FLT (e.g., the recent series of remedial books *Common Mistakes at [...]* by Cambridge University Press) and SLA (cf. Skehan 1998). At the same time, prominent strands of SLA research are concerned with the stages of the acquisition process (cf. Pienemann 1998), often independent of the accuracy of the execution of the patterns which are indicative of the different levels. In sum, SLA research essentially observes correlations of linguistic properties, whether erroneous or not. In consequence, learner corpora should ideally provide annotation of linguistic properties, including but not limited to errors.

Annotation schemes have been developed for different types of linguistic analysis, including Part-Of-Speech (POS), syntactic constituency, lexical dependencies, or semantic and discourse properties (cf. Garside et al. 1997). At the same time, annotation of learner language has generally focused on the annotation of learner-language specific errors, while linguistic annotation of learner corpora has received next to no attention. In this paper, we explore what may constitute appropriate linguistic annotation schemes for learner language. We start with POS annotation as a basic building block of more complex, structural linguistic annotation.

LINGUISTICALLY ANNOTATING LEARNER CORPORA

Part-of-speech annotation

a. *Empirical basis*

The task of POS tagging involves assigning to each token in a text its corresponding POS label. Three types of evidence can be identified based on the text: distribution, morphological information, and lexical information.

For Example (1), looking up the token "of" in a *lexicon* shows that it can unambiguously be classified as a preposition. In other cases, the lexicon provides a set of possible POS tags for a given word, which usually is disambiguated by context.

(1) I was surprised by the word *of* the day.

For words not listed in a given lexicon, *morphological clues* can still provide POS information, as in Example (2), where the verbal past tense suffix *-ed* is morphological evidence that the token is likely to be a verb.

(2) His son *brachiated* along the monkey bars.

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. For example, the context of "man" in (3) identifies it as a verb, even though the more common category for this word is noun.[2]

(3) The old *man* the boat.

b. *Automatic POS-tagging*

A number of different approaches have been developed for automatic POS-tagging, such as probabilistic models of tag/token pairs and tag sequences (Schmid 1994; Brants 2000; Toutanova & Manning 2000), local constraint rules (Karlsson 1990), or error-driven transformation-based methods (Brill 1992). Conceptually, the task of any automatic POS tagging system consists of two parts: *tag lookup* and *tag disambiguation*. During the tag lookup step, all possible POS tags for the given token are determined. This requires access to a substantial lexical database that lists the possible POS tags for each token. Such databases are usually extracted from large, manually POS-tagged corpora. Alternatively, morphological analysis can help determine the

set of possible POS tags. In the tag disambiguation step, the list of possible tags for a given word must be reduced to the correct tag for this particular instance by considering contextual information about the distribution of POS tags.

Even with a very large lexical database for tag lookup, POS taggers may encounter unknown words when tagging previously unseen text. Therefore, all POS taggers implement *fallback strategies* which are applied when an unknown token occurs. The fallback strategies employ weaker versions of the same three sources of evidence, such as morphological (suffix) analysis, local contextual clues and, as a last resort, the use of the most frequently occurring tag.

Previous approaches to POS-tagging learner data

The few approaches to POS-tagging learner corpora discussed in the literature generally rely on tagsets and tools developed for native language. Thus, the task of POS-tagging learner language is essentially perceived as an instance of domain transfer: When applied to a new genre of text, taggers perform worse than when applied to the genre they were developed for. In order to make up for this degradation of performance, post-correction steps are usually added to modify tags that are systematically wrongly assigned.

Van Rooy & Schäfer (2003) report on a study for annotating learner language, which employs a domain transfer strategy. They annotate the Tswana Learner English Corpus (TLEC) with the TOSCA-ICLE tagger (Aarts et al. 1998), which is trained on native language. The tagger output is then post-corrected by selecting a set of tags most frequently confused by the tagger and these are manually corrected by student editors. The authors expect to remove about 69% of all tag errors by combining this with a post-correction step for a more extensive tagset to be carried out by expert linguists.

Thouësny (2009) pursues a similar approach for POS-tagging a French learner corpus, involving automatically tagging the corpus with a probabilistic tagger trained on native data (Schmid 1994) and then using a manually developed set of rules to post-correct cases where the tagger systematically goes wrong.

Several authors have analyzed the error types that occur when tagging learner language with a POS tagger trained on native data. Van Rooy & Schäfer (2002), along with de Haan (2000), identify spelling errors as a major source of problems for the POS tagger. They either result in non-words, which can be handled rather straightforwardly since taggers can easily detect unknown words, or they result in so-called real-word errors (e.g. "there" in place of "their"), which are

harder to identify. The second class of typical errors described by van Rooy & Schäfer (2002) comprises other types of learner errors, such as picking a wrong lexical item, incorrectly inflecting a word, omissions, or using non-standard syntactic configurations.

De Haan (2000) proposes a rather fine-grained classification of learner errors ranging from typing errors to L1-transfer errors related to pronunciation (e.g., the use of "improve" instead of "improved" by native speakers of Spanish, who have difficulties recognizing and producing such closed syllables). De Haan (2000) suggests extending the TOSCA-ICLE POS tagset with an additional feature that indicates the type of learner error at the tag's position.

Overall, these approaches treat POS annotation of learner language as a robustness issue: Their goal is to provide as much information as possible for a given standard POS tagset under suboptimal circumstances (similar to trying to interpret mobile phone speech as a deteriorated version of face-to-face speech). As illustrated by the second version of the ICLE corpus (Granger et al. 2009), which was automatically annotated using the standard CLAWS tagger, this can result in relatively high-quality POS annotation. However, it is unclear what exactly it means for learner language to be correctly annotated with a POS scheme developed for native language, especially where learner language involves incompatible distributional, morphological, and lexical stem information.[3]

By treating learner language as a noisy variant of native language, the above mentioned POS-tagging approaches essentially gloss over the differences between native and learner language. Yet, the systematic nature of learner language and how it differs from native language is of interest to FLT and SLA research so that we want to investigate how it could be systematically encoded in the linguistic annotation. In the following, we explore an alternative path, taking one step back from the POS annotation schemes developed for native language to the nature of the evidence one can identify in learner language by looking at distribution, morphology, and lexis. We start the discussion by introducing the learner corpus we are using for our data-driven exploration.

ANNOTATION OF NOCE

Our study is based on the NOCE corpus (NOn-native Corpus of English, Díaz Negrillo 2007), a written corpus of English as a Foreign Language. It contains written texts by Spanish undergraduates, primarily first year students, enrolled in the English degree program at

the Universities of Granada and Jaén. The participants' age is 18-19, and their level of English ranges from upper-intermediate to advanced. Specific information about the participants (gender, L2 exposure, motivation, etc.) and sampling policy (date, task conditions, etc.) is recorded. The texts in the corpus amount to over 300,000 words collected at the beginning of each term: October-November, February-March, and June-July. The participants complete a timed classroom task which involves writing an essay on one of three topics suggested. A fourth option is free writing. The samples average 200 words, with a marked tendency towards shorter essays in the first sampling and towards longer in the third as a result of the students' progress.

Corpus encoding and error annotation

The corpus contains two types of interpretative annotation: editorial and error. The learner texts, originally handwritten as a classroom activity, were typed in and stored in an XML format. TEI-compliant headers encode the meta-information about the students and the corpus, and the entire corpus is annotated with editorial tags for students' editions of their own writing (e.g., struckouts, late insertions, reordering of units and missing/unreadable text). The error tagset EARS (Díaz Negrillo 2009) was designed to identify and classify learner errors at different levels (spelling, punctuation, word grammar, syntax, and lexis). Currently, one quarter of the corpus is annotated using this very fine-grained error tagset (612 tags).

This paper uses the error-tagged section of the corpus, which contains 39,015 words distributed in 179 texts by 108 different participants, and the EARS error annotation for identifying the relevant learner language examples for this study. For all annotation in the corpus (editorial, error, POS), the XML encoding ensures that the corpus text and the annotations can easily be kept apart, in line with the recommendations of Leech (1997, section 1.3).

Interlanguage POS annotation

Learner language differs markedly from native English in the way the three sources of evidence for the classification of tokens into POS categories combine: i) distribution, or a token's linear order with respect to the other tokens; ii) morphological marking, or the prefixes and suffixes added to stems; and iii) lexical stem lookup, or the lexically encoded specific properties of a word.

For native language, the three sources of evidence converge on a single POS classification. As we mentioned in the automatic POS-tagging section above, for POS taggers this means that it is possible to

reliably combine all evidence in a tag lookup and disambiguation process. Yet, in our investigation of learner language we observed systematic cases in which a single, consistent combination of the three sources of evidence is not possible.

Rather than force a resolution based on conflicting evidence, it seems advantageous to aim for a tripartite annotation which provides access to each type of evidence separately, in order to support an analysis of this apparent characteristic of learner language. In the following, we therefore present a data-driven systematization of the three empirical aspects involved in POS classification for learner language, with a focus on where they provide conflicting evidence.

Mismatches in POS classification variables

Case 1: Stem-Distribution mismatch. In the first case, a lexeme of a given word class appears in a distributional slot which is not available to instances of that word class. The token does not exhibit overt morphological marking.

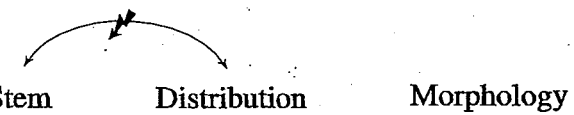

Figure 1. *Stem-Distribution mismatch*

An example for this case is shown in (4), where the lexical entry of "vary" identifies it as a verb but it occurs in a nominal distributional slot, surrounded by an adjective on the left and a preposition on the right.

(4) [...] you can find a big *vary* of beautiful beaches [...] GR-1-B-EN-102-X

Another example is shown in (5), where the lexical entry of "friendship" unambiguously identifies it as a noun, but the distributional slot is that of an adjective.

(5) [...] they are very kind and *friendship*. GR-1-B-EN-102-X

Related cases have been described as "word class transfer" by de Haan (2000: 74). His term seems to apply only to tokens which are derivationally related, as in his example "pride" vs. "proud," here similarly applicable to "vary" vs. "variety" in (4) and "friendship" vs. "friendly" in (5) above. However, a mismatch between distribution and

the word class of the stem can also extend to derivationally unrelated tokens, as in the examples shown in (6) and (7).

(6) [...] that's the reason *because* I went to Tunisia twice. GR-1-A-EN-076-F
(7) RED helped him *during* he was in the prison. GR-1-A-EN-025-F

In (6), the conjunction "because" is found in the distributional slot of a *wh*-pronoun (presumably "why," according to the preceding context and the relation intended between the noun reason and the subordinate clause). In (7), the preposition "during," which in native English combines with noun phrases, here introduces a subordinate clause and takes the distributional slot of a conjunction (presumably "while," by the temporal relationship that the token is intended to set). Two POS classifications can therefore be proposed in each of the cases: conjunction and preposition in accordance with their lexical stem lookup, and pronoun and conjunction in accordance with their distribution, respectively.

Case 2: Stem-Distribution, Stem-Morphology mismatch. As in the first case, a lexeme from a given word class appears in a distributional slot which is not available to instances of that word class. In addition to this stem-distribution mismatch, the token exhibits overt morphology which agrees with the distributional evidence but conflicts with the word class lexically determined for the stem.

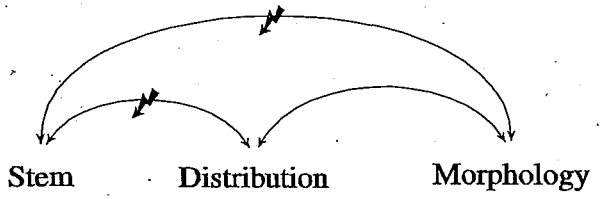

Figure 2. *Stem-Distribution, Stem-Morphology mismatch*

Examples of this case are shown in (8) and (9). Unlike Case 1 above, the additional mismatch between the stem and the morphology results in words which do not exist in native English.

(8) [...] one of the favourite places to visit for many *foreigns*. GR-1-C-EN-024-F
(9) [...] to be *choiced* for a job [...] GR-1-A-EN-003-X

In (8), a token for which stem lookup identifies it 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). Therefore the token "foreigns" in (8) is classified as an adjective according to its lexical stem lookup, but as a noun according to its distribution and morphology. In (9), the word "choiced" distributionally appears in a verbal slot and morphologically it carries verbal inflection (*-ed*), whereas lexically the stem "choice" is a noun (or adjective).

Morphology can sometimes provide evidence for two distinct classifications. For example, in (10) below, the derivational morpheme in "politicals" categorizes the token as an adjective and the inflectional morpheme as a noun (or verb).

(10) [...] and dark *politicals* will be defeated. GR-1-B-EN-073-Y

The example suggests that it is necessary to keep derivation and inflection apart, encoding the former within the lexical lookup dimension and only the latter in the morphology dimension of our tripartite POS classification. Inflectional morphology naturally is often ambiguous. In (11), for example, the affix *-s* of contents can be identified as the third person singular inflection of verbs or as the plural inflection of nouns.

(11) [...] internet have some "pages" that *contents* something so horrible [...] GR-1-A-EN-020-Z

The former interpretation of the suffix would be compatible with the verbal distributional slot it appears in (and the somewhat uncommon lexical lookup of the stem as a transitive verb), whereas the latter interpretation would be consistent with the lexical lookup of the stem as a noun. In the tripartite POS annotation, the morphological dimension will thus need to be disjunctively specified.

Case 3: Stem-Morphology mismatch. The third type of systematic mismatches between the three sources of evidence involves tokens for which the word class determined by lexical lookup agrees with the distributional evidence, but conflicts with the inflectional morphology of the token.

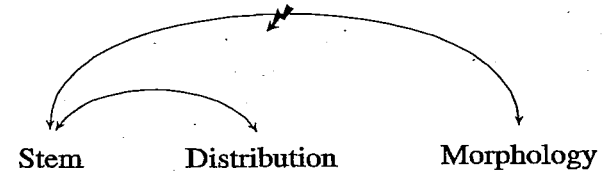

Figure 3. *Stem-Morphology mismatch*

In the examples below, lexically the stems of the words in italics are most likely classified as adjectives,[4] which is in sync with the distributional properties of attributive (12) and predicative adjectives (13). But this classification conflicts with the nominal plural suffix inflection *-s* (or, alternatively, third person singular verbal inflection).

(12) [...] this film is one of the *bests* ever customes [...] GR-1-B-EN-089-F
(13) [...] television, radio are very *subjectives* [...] GR-1-C-EN-041-X

As in Case 2, the words for which such a mismatch between the lexeme and the inflection arises do not exist in native English. The tripartite POS classification makes it possible to label these tokens as adjectives in terms of their distribution and lexical stem lookup, and as nouns according to their inflections.

Case 4: Distribution-Morphology mismatch. Finally, in the fourth case, the lexical word class specification for the stem accords with its distribution and morphology, but the inflectional morphology does not match the distribution, i.e., does not match the grammatical context.

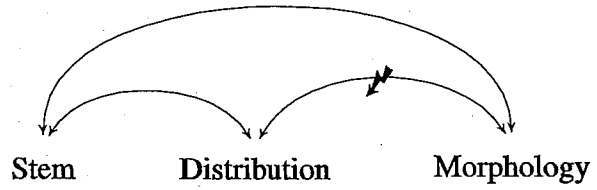

Figure 4. *Distribution-Morphology mismatch*

This is illustrated in (14), where a noun is inflected for plural, but its distributional slot, following the determiner "every," requires singular number. This is a distinction reflected in many POS annotation schemes, such as the Penn tagset (Marcus et al. 1993).

(14) [...] for almost every *jobs* nowadays [...] GR-1-A-EN-040-X

Example (15) shows another example for such a mismatch between distribution and morphology. Here the past tense verb "grow" appears in a distributional context following "has," which requires a past participle.

(15) [...] it has *grew* up a lot specially after 1996 [...] GR-1-A-EN-098-F

The units in italics in (16) and (17) run hand in hand with those in (14) and (15) in that their morphology does not match their distribution, with

the additional problem that "want" in (16) and "have" in the Penn tagset are ambiguous between base form verb and non-third person singular finite verb, which requires them to be disjunctively specified in the morphological dimension.

(16) [...] if he *want* to know this [...] GR-1-A-EN-022-X
(17) This first year *have* been wonderful [...] GR-1-C-EN-103-F

The discussion of the four cases above provides empirical justification for our claim that assigning a single POS tag from an annotation scheme developed for native language is problematic in light of conflicting empirical evidence. To encode this characteristic property of learner language, one can instead use the tripartite POS annotation separating evidence from distribution, from inflectional morphology, and from the lexical stem. Annotating each of the three dimensions of evidence separately naturally also makes it possible to combine the evidence into a single POS classification for those cases where the classifications in the three dimensions are compatible, or to develop explicit weighting methods for resolving all or particular classes of conflicts which arise.

Mismatch-free learner language

While this paper focuses on a systematic POS characterization of learner language, as shown, some types of learner errors are characterized by mismatches between the three dimensions of empirical evidence. Other learner errors do not involve such mismatches; for completeness sake, in the following we characterize some of those orthogonal error types.

a. Realization using wrong allomorph. The allomorph used for the realization of an inflectional morpheme is not available to a particular lexeme:

(18) The mayority of people that die in Irak are *childs* [...] GR-1-C-EN-041-X
(19) He *runned* to buy one [...] GR-1-B-EN-049-F

b. Realization using wrong stem. An inflected form is used as base for additional incompatible inflection. For example, in (20) the past tense form of the verb has been used as base for third person singular inflection.

(20) [...] the 11th March *cames* to our minds. GR-1-C-EN-027-X

c. Duplicate inflection. An inflectional morpheme has been realized twice:

(21) *Childrens* spend so much time [...] GR-1-A-EN-102-F
(22) [...] it *stresseses* me a lot. GR-1-C-EN-094-F

d. Inappropriate word-formation rules. Idiosyncratic word-formation rules are applied, as in "modificate" and "socialities" in (23) and (24).

(23) [...] internet can *modificate* [...] GR-1-A-EN-034-Z
(24) [...] different *socialities* and ways of life. GR-1-A-EN-068-X

e. Creative lexis. Foreign lexis is used, as in (26), or lexical coinage, such as "menospreciated" in (25), arguably a calque from Spanish "*menospreciada*" ("undervalued").

(25) [...] people shouldn't be *menospreciated* because of the music they listen to [...] GR-1-A-EN-086-F
(26) [...] for many *raisons*. GR-1-A-EN-075-X

Learner-specific word-structure errors as those described in this section could also be categorized by an interlanguage POS tagger, but appear to be largely orthogonal to the general issue of systematic POS annotation.

CONCLUSION

Corpus-based FLT and SLA research can benefit from linguistically annotated learner corpora in that annotations provide access to classes of data which cannot easily be characterized based on the surface string alone.

A prerequisite for this is that the linguistic annotation is consistent, comprehensive, and can systematically capture the properties of learner language. In order to develop adequate annotation schemes for learner language corpora and automatic annotation methods for such interlanguage, interdisciplinary collaboration between applied and computational linguists arguably is crucial. In this paper, we discussed the first results of such a collaboration, focusing on the POS analysis of learner language.

The POS annotation of learner language has not received much attention in the literature. Existing approaches essentially view the POS annotation of learner language as a robustness issue in that they apply a POS annotation scheme developed for native language.

Based on an empirical investigation of learner language as collected in NOCE, a Spanish learner corpus of English, we show that the use of native POS annotation schemes for learner language is problematic for several classes of cases in which the evidence from distribution, morphology, and lexis systematically does not converge on a single POS classification. Without explicit conflict resolution procedures, it is unclear which POS tag should be chosen when such a conflict arises. This also makes it somewhat unclear how the accuracy figures for POS tagging of learner language reported in the literature are to be interpreted. Even where resolution determines a single POS classification, its annotation does not provide systematic access to the conflicting evidence as an observable characteristic of learner language.

As an alternative, we propose a tripartite POS analysis which encodes the three separate observations based on the distribution, the morphology, and the lexical stem. On this basis, one can analyse where the three observations are compatible and where they provide conflicting evidence. Such a tripartite POS annotation provides access to characteristic properties of learner language and at the same time makes it possible to uniformly characterize well-formed language patterns as well as erroneous learner language resisting a single POS characterization.

NOTES

1. The use of corpora for obtaining examples is not directly tied to a specific method for evaluating the data thus obtained. Depending on the corpus composition, both quantitative and qualitative analysis of data found in learner corpora are possible.
2. 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.
3. In the German learner corpus FALKO (Lüdeling et al. 2008), a well-formed target hypothesis is provided for each sentence in the corpus. It is this target hypothesis that is POS annotated, which avoids the problem of having to determine POS categories for learner language patterns that do not exist as native language patterns.
4. For "best," there also are lexical entries with POS adverb, noun, or verb; for "subjective" there also is a noun entry, corresponding to nominative.

REFERENCES

Aarts, J., van Halteren, H. & Oostdijk, N. 1998. The linguistic annotation of corpora: The TOSCA analysis system. *International Journal of Corpus Linguistics*, 3/2, 189-210.

Brants, T. 2000. TnT – A statistical part-of-speech tagger. In proceedings of the *Sixth Conference on Applied Natural Language Processing* (pp. 224-231), Association for Computational Linguistics, Morristown, N. J. Available online: <http://aclweb.org/anthology/A00-1031>.

Brill, E. 1992. A simple rule-based part of speech tagger. In proceedings of the *Third Conference on Applied Natural Language Processing* (pp. 152-155), Association for Computational Linguistics, Morristown, N. J. Available online: <http://aclweb.org/anthology/A92-1021>.

Corder, S. P. 1981. *Error Analysis and Interlanguage*. Oxford: Oxford University Press.

de Haan, P. 2000. Tagging non-native English with the TOSCA-ICLE tagger. In C. Mair & M. Hundt (Eds.), *Corpus Linguistics and Linguistic Theory* (pp. 69-79). Amsterdam: Rodopi.

Díaz Negrillo, A. 2007. A Fine-Grained Error Tagger for Learner Corpora. Unpublished Ph.D. thesis, University of Jaén, Jaén.

——. 2009. *EARS: A User's Manual*. Munich: LINCOM Academic Reference Books.

Granger, S. 2003. Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal*, 20/3, 465-480. Available online: <http://purl.org/calico/granger03.pdf>.

——., Dagneaux, E., Meunier, F. & Paquot, M. 2009. *International Corpus of Learner English Version 2*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Karlsson, F. 1990. Constraint Grammar as a framework for parsing running text. In proceedings of the *13th Conference on Computational Linguistics* (pp. 168-173), Association for Computational Linguistics, Morristown, N. J. Available online: <http://aclweb.org/anthology/C90-3030>.

Leech, G. 1997. Introducing corpus annotation. In R. Garside, G. Leech & T. McEnery (Eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora* (pp. 1-18). Harlow: Addison Wesley Longman Limited.

Lüdeling, A., Doolittle, S., Hirschmann, H., Schmidt, K. & Walter, M. 2008. Das Lernerkorpus Falko. *Deutsch als Fremdsprache*, 45/2, 67-73.

Marcus, M. P., Marcinkiewicz, M. A. & Santorini, B. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19/2, 273-290. Available online: <http://aclweb.org/anthology/J93-2004>.

Meurers, W. D. 2005. On the use of electronic corpora for theoretical linguistics. Case studies from the syntax of German. *Lingua*, 115/11, 1619-1639. Available online: <http://purl.org/dm/papers/meurers-03.html>.

——. & Müller, S. 2009. Corpora and syntax (Article 42). In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics* (pp. 920-933). Berlin: Mouton de Gruyter. Available online: <http://purl.org/dm/papers/meurers-mueller-09.html>.

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

Richards, J. C. 1974. *Error Analysis: Perspectives on Second Language Acquisition*. London: Longman.

Schmid, H. 1994. Probabilistic part-of-speech tagging using decision trees. In proceedings of the *International Conference on New Methods in Language Processing*, Manchester, United Kingdom. Available online: <http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf>.

Skehan, P. 1998. *A Cognitive Approach to Learning Language*. Oxford: Oxford University Press.

Thouësny, S. 2009. Increasing the reliability of a part-of-speech tagging tool for use with learner language. Paper presented at the *Automatic Analysis of Learner Language (AALL'09) Workshop*, Tempe, AZ. Available online: <https://calico.org/p-425-Abstracts_AALL09.html>.

Toutanova, K. & Manning, C. D. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In proceedings of the *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora* (pp. 63-70), Association for Computational Linguistics, Morristown, N. J. Available online: <http://www.aclweb.org/anthology/W00-1308>.

van Rooy, B. & Schäfer, L. 2002. The effect of learner errors on POS tag errors during automatic POS tagging. *Southern African Linguistics and Applied Language Studies*, 20, 325-335.

——. & Schäfer, L. 2003. An evaluation of three POS taggers for the tagging of the Tswana Learner English Corpus. In D. Archer, P. Rayson, A. Wilson & T. McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 Conference Lancaster University (UK), 28-31 March 2003. Vol. 16 of University Centre for Computer Corpus Research on Language Technical Papers* (pp. 835-844). Lancaster: UCREL, Lancaster University.

ACKNOWLEDGEMENTS

Research leading to this paper was funded by the Junta de Andalucía Research Project P07-HUM-03028.

ANA DÍAZ-NEGRILLO
UNIVERSITY OF JAÉN, SPAIN.
E-MAIL: <ADINEGRI@UJAEN.ES>

DETMAR MEURERS
UNIVERSITY OF TÜBINGEN, GERMANY.
E-MAIL: <DM@SFS.UNI-TUEBINGEN.DE>

SALVADOR VALERA
UNIVERSITY OF JAÉN, SPAIN.
E-MAIL: <SVALERA@UJAEN.ES>

&

HOLGER WUNSCH
UNIVERSITY OF TÜBINGEN, GERMANY.
E-MAIL: <WUNSCH@SFS.UNI-TUEBINGEN.DE>

LANGUAGE FORUM
VOL. 36, NO. 1-2, JAN-DEC 2010

How Much Wheat is there in the Chaff? Issues Concerning the Use of Concordancers in the Classroom

ROLF KREYER
University of Bonn, Germany

ABSTRACT

Corpus linguistics methods and research have had a huge impact on English Language teaching over the last few decades. Primarily, corpus linguistic findings have contributed to the content and make-up of dictionaries and text books. In addition, starting off with Tim Johns' (1988, 1990, 1991, 2002) plea for data-driving learning, a growing amount of literature has suggested ways of making corpora and corpus-linguistic methods accessible to foreign language teaching and learning. This study focuses on potential problems that the use of concordancers in the classroom might entail. More specifically, on the basis of three case studies, the paper tries to estimate the classroom suitability of un-edited concordance lines by analyzing the amount of junk that concordancers produce. The paper also suggests ways of dealing with this problem.

INTRODUCTION

With the advent of computer corpora in the early sixties, modern corpus linguistics has made its way into the field of linguistics and the field of language teaching. With regard to the latter, two main ways can be distinguished in which corpora have influenced modern language teaching. First, the study of huge amounts of language data has led to new insights into the use of language and has provided new answers to the question of what language actually is. This has had a strong influence on curriculum design and on the shape and content of reference tools (cf. Meunier 2002: 123-130). To name but two examples, the author of the German school grammar *Englische Grammatik Heute* (Ungerer 1999), for instance, emphasizes the fact that the descriptions found therein are (to some extent) corpus based (cf. Mukherjee 2004: 243), and the *Collins Cobuild English Dictionary*