Retrieving Passive Structures from the Secondary-Level Corpus of Learner English (SCooLE) - How Can We Make Part-of-Speech Tagging More Successful?

Verena Möller
Université catholique de Louvain
Universität Hildesheim
verena.moeller@uni-hildesheim.de

1 Learner language in secondary education

In the south-west of Germany, the secondary level of the educational system offers a variety of options to those who learn English as a foreign language.

Having finished primary school, students benefit from six years of EFL (English as a Foreign Language) education until they start preparing for their final examinations. Additionally, a number of schools have introduced CLIL (Content and Language Integrated Learning) programmes. Not only do these lead to an increase in exposure to the English language, but learners are also presented with a different type of written input, adding a scientifically oriented genre to the informative, imaginative and argumentative types of text traditionally used in EFL teaching.

The question therefore arises whether or not written learner language differs according to educational setting. The passive has been chosen as a diagnostic criterion as research suggests that it is one of the characteristics in which scientific text, and therefore presumably the input that CLIL learners receive, differs from other genres (cf. Svartvik 1966).

2 The Secondary-Level Corpus of Learner English (SCooLE)

In order to investigate the research question outlined above, the Secondary-Level Corpus of Learner English (SCooLE) has been compiled from about 850 student essays (around 250,000 words). Participants in the study were learners with a background of at least six years of EFL education, plus at least four years of CLIL experience, if that option was available and chosen.

To be able to account for individual differences between learners from the various backgrounds, the SCooLE was annotated with metadata providing information, amongst others, on language learning/ acquisition experience, cognitive capacities and aspects of motivation.

During data elicitation, learners were asked to type two short argumentative essays, one of their essay topics being formulated in the passive. A large amount of deviance with respect to spelling, vocabulary, morphology and syntax was exhibited in the texts that were produced.

3 Linguistic annotation without normalisation of deviances

Due to the high degree of deviance that was found, the success of automatic part-of-speech tagging was expected to be limited. Hence, three tools were tested in a pilot study on a section of the SCooLE (around 17,000 words), with special reference to the retrieval of passive constructions: The TreeTagger (TT, cf. Schmid 1994), CLAWS (CL, cf. Garside & Smith 1997) and the MATE parser (MA, cf. Bohnet 2010). Both the TreeTagger and CLAWS were shown to be highly accurate in the annotation of target-like instances of *be Ved.* Being the least successful of the three tools in this respect, MATE was eliminated in the course of the pilot study (cf. Table 1).

	TT	CL	MA
be + past participle (n=129)	129	128	123

Table 1: Retrieval of target-like *be Ved*

Erroneous occurrences of *be Ved* presented a problem to all three tools (cf. Table 2). This called for normalisation of the more recognizable flaws within the learner output.

	TT	CL	MA
correct tag for *be* (n=14)	12	12	11
correct tag for the past participle (n=20)	11	15	15
corrects tags for *be* and past participle (n=14)	5	9	9

Table 2: Retrieval of erroneous *be Ved*

4 Linguistic annotation with normalisation of deviances

The limited success of the part-of-speech taggers with erroneous *be Ved* led to the application of several procedures that were supposed to enhance the correct retrieval of passive constructions which are not target-like:

- Replacement of accents used as apostrophes: This affected those forms of *be* that were part of contracted forms and could not be recognized as such (e. g. *it´s* vs. *it’s).* These adjustments were judged to be minor, hence the original was not preserved.

- Normalisation of deviant forms on the basis of output from the Variant Detector, VARD (cf. Rayson & Baron 2011): This involved simple replacement of variants with their normalised version, as well as manual replacement of multi-word expressions wherever the normalisation of the detected variant affected the word's environment. The original was preserved within an XML tag:
  [...] the alcohol can be <vardbased orig= "buyed" type="false">bought</vardbased> [...]

As may be seen in Table 3, the TreeTagger and CLAWS were equally successful in retrieving *be Ved* after these procedures.

	TT	CL
correct tag for *be* (n=14)	14	14
correct tag for the past participle (n=20)	15	15
corrects tags for *be* and past participle (n=14)	11	11

Table 3: Retrieval of erroneous *be Ved* after VARD-based normalisation

Unfortunately, the success rate of both tools was still insufficient to enable reliable observations with respect to the learners' use of passive constructions. Hence, more manual annotation had to be effected to improve the retrieval of *be Ved.* Frequently misspelled homophones or near-homophones were normalised, especially when a form of *be* or a lexical verb were concerned. Again, the original was preserved within an XML tag:

[...] should be <manual orig="aloud" type= "false">allowed</manual> to drink [...]

This procedure increased the correct annotation of past participles by one instance for both the TreeTagger and CLAWS. However, it still did not create a sufficient basis for the retrieval of erroneous *be Ved* from the SCooLE, especially since learners frequently omitted to use a form of *be.*

It seems that this can only be tackled by manually annotating all intended passive constructions. Table 4 shows that both the TreeTagger and CLAWS were able to retrieve almost all previously erroneous instances of *be Ved* after this procedure had been applied. The past participles that received an erroneous part-of-speech tag were not the same in the TreeTagger and the CLAWS output, such that concurrent use of both taggers and the comparison of the respective output seems the most reliable way of retrieving target-like as well as erroneous *be Ved* from learner text in the SCooLE.

	TT	CL
correct tag for *be* (n=20)	20	20
correct tag for the past participle (n=20)	19	19
corrects tags for *be* and past participle (n=20)	19	19

Table 4: Retrieval of errorneous *be Ved* after VARD-based and manual normalisation and normalisation of intended passives

5 Conclusion

In this paper, the linguistic annotation of the Secondary-Level Corpus of Learner English (SCooLE) is presented, with special reference to the retrieval of passive constructions from learner text which exhibits a high degree of deviance from the target with respect to spelling, vocabulary, morphology and syntax.

Various tools for annotation were tested and it was found that the normalisation of variants detected by VARD as well as manual normalisation of homophones/near-homophones and intended passive constructions can lead to reasonable success in part-of-speech tagging with both the TreeTagger and CLAWS.

As of spring 2013, normalisation on the basis of output from VARD as well as normalisation of homophones/near-homophones has been completed for the entire corpus.

References

Bohnet, B. 2010. "Very High Accuracy and Fast Dependency Parsing is not a Contradiction". In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010),* Beijing: 89-97.

Garside R. and Smith, N. 1997. "A hybrid grammatical tagger: CLAWS4". In R. Garside, G. Leech and A. McEnery (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora.* London: Longman.

Rayson, P. and Baron, A. 2011. "Automatic error tagging of spelling mistakes in learner corpora". In F. Meunier, S. De Cock, G. Gilquin and M. Paquot (eds.) *A Taste for Corpora. In honour of Sylviane Granger,* Studies in Corpus Linguistics, 45. Amsterdam: John Benjamins.

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

Svartvik, J. 1966. *On Voice in the English Verb.* The Hague/Paris: Mouton.