Baltic Journal of English Language, Literature and Culture Vol. 4, 2014: 122–129
https://doi.org/10.22364/BJELLC.04.2014.10

TAGGING ERRORS IN NON-NATIVE ENGLISH LANGUAGE STUDENT-COMPOSED TEXTS OF DIFFERENT REGISTERS

ZIGRĪDA VINČELA
University of Latvia, Latvia

Abstract. Research of linguistic features requires part of speech (POS) tagging of texts. The existing POS taggers have been predominantly trained on native speakers' texts to enhance their accuracy. The researchers exploring POS tagging of ELL (English language learners) texts distinguish tagger's and learners' errors and suggest annotation enhancement schemes. However, the frequency and types of CLAWS7 (Constituent Likelihood Automatic Word Tagging System) tagging errors in ELL texts of different communicative purposes have not been sufficiently explored to suggest annotation enhancement solutions in each particular learner corpus building case. This study investigates CLAWS7 tagged texts composed by non-native English philology BA students (English Studies Department, University of Latvia) to uncover the overall precision of the tags having the greatest impact on the error rate and provide an insight into errors to reveal the texts requiring annotation enhancement solutions. Material for the analysis has been selected from the corpus of student-composed texts. The results show that tagging precision varies across the text groups. The texts edited by the students show greater tagging precision, and therefore would not require specific annotation enhancement procedures before their tagging. Tagging precision is lower in such interactional texts as chat messages that could be addressed by the application of an annotation enhancement scheme.

Key words: corpus, annotation, accuracy, tagging error

INTRODUCTION

Scholars (e.g. McEnery et al., 2006; Reppen, 2010) point out that the investigation of a wide range of linguistic features, for example, in the texts of various genres can be explicitly performed on part of speech (POS) annotated texts. This assumption refers also to non-native English language learners' texts. Aarts and Granger (1998) have already drawn attention to the fact that POS annotation of learner corpora can reveal their language use in detail. Meanwhile POS annotation taggers, such as CLAWS7, have been trained and predominantly used to tag texts composed by native speakers. Even if its overall performance on native speakers composed texts is recognised as high, there is comparatively modest research on CLAWS7 performance on non-native student-composed texts of different genres. The goal of the present study is to tag the text samples composed by BA students of English philology (the University of Latvia) with

CLAWS7 and subject them to quantitative and qualitative analysis of examples to uncover the tagger's and students' errors of the most frequent tags leading to an ambiguous assignment. This would reveal the text groups requiring annotation enhancement solutions or editing of CLAWS7 assigned tags in the learner corpus creation process.

THEORETICAL BACKGROUND

POS tagging, also known as morpho-syntactic annotation, is the process during which a POS tag is assigned to each word in a text corpus. Leech explains (1997: 2) that annotation 'enriches the corpus as a source of linguistic information for future research'. Reppen (2010: 35) adds that annotation can substantially relieve parts of speech distinction. For example, in POS annotated corpus, the search of the noun *well* would be more effective than in raw corpus because it would exclude all the instances of the adverb. Linguists (Leech, 1997: 4-5; McEnery, 2003: 454-455 referred to by Mc Enery; Xiao and Tono, 2006: 30-32) have summarised the following main benefits of annotated corpora in linguistic research: (1) the ease of linguistic information extraction; (2) reusability as well as reusability for the purpose that differs from the initial research question; (3) a source of objective record for analysis.

McEnery et al. (2006: 34) note that POS annotation for the English texts is considerably developed to perform it automatically by taggers with the precision rate suitable for various research questions. One of such taggers is CLAWS7 that has been developed at Lancaster University (Leech et al., 1994). Its overall accuracy rate is 96-97 percent for written language texts, and therefore, being acknowledged as a high accuracy tagger, has been widely applied in tagging native texts, e.g. theLancaster-Oslo-Bergen Corpus (LOB) corpus and also c. 100 million words of the British National Corpus (BNC). These corpora have been widely used by researchers in their studies (e.g. Adami, 2009, in research of pronouns).

Learner corpora compilers and researchers have primarily devoted their attention to learner text error tagging, e.g. Granger (2003) as well as developing annotation schemes for English language learners' non-word (spelling and morphological) errors (Hovermale and Martin, 2008) as a precondition for POS tagging of learner corpora towards the development of annotation enhancement schemes. CLAWS7 for its comparatively high accuracy has drawn researchers' attention and has been applied to POS tagged non-native students' essays and letters. Van Rooy and Schaffer (2003) have explored and found promising results on the overall accuracy of CLAWS7 in comparison with two other taggers TOSCA-ICLE and Brill on the sample of non-native students' essays. Twardo (2012) has applied CLAWS7 tagger in the investigation of the learners' essays and letters (levels B1 –C1) with the aim to focus on learners non-word errors. The comparatively promising results on CLAWS7 and its applicability in tagging non-

native students' text samples has called for its application in tagging of a wider range of non-native student-composed text samples in the present study to pre-test the tagger for further annotation enhancement strategy solutions.

MATERIAL AND METHOD

The analysis material has been selected from the corpus of the second year student-composed texts (STUDTEXREG) compiled for investigation of register-based variation of linguistic features at the English Studies Department, University of Latvia. The corpus texts are arranged into six groups (220,012 tokens) according to their communicative purpose: statements, essays, letters, virtual conference messages, chat messages and discussion messages. To make the investigation more feasible, the randomly selected text samples, in total 10,000 tokens, were subjected to CLAWS7 tagging analysis. Samples of 3000 tokens have been selected from each of the three text groups (letters, discussion messages and essays), whereas 1000 tokens from chat messages. The error counts have been calculated on normalized texts (i.e. per 1000 tokens), as the text length differs across and within the genres. For example, chat and discussion messages are considerably shorter than essays.

The tagger's performance evaluation methodology proposed by Van Rooy and Schaffer (2003) has been applied in the present study. These researchers have based their evaluation on Van Halteren's (1999) considerations of a tagger: its tagset, documentation, the tagging process and performance of the tagger. CLAWS7 tagset contains 137 tags (excluding punctuation tags), its documentation is available on University Centre for Computer Corpus Research on Language (UCREL) site (see Leech et al., 1994) and the tagging process is comparatively fast. CLAWS7 overall performance has been investigated by Van Rooy and Schaffer by comparing three taggers on the learner texts and they have found that CLAWS7 is the most accurate on non-native students' texts among all three taggers (CLAWS7 96.26 %, TOSCA-ICLE 88.04 % and Brill 86.34%). These results show that CLAWS7 obtained results on non-native students' texts that correlates with its overall accuracy on native speakers' texts, which is 96-97percent.

However, in order to reveal particular error types and causes, they have identified the tags with generally the lowest precision (*RGR, RRR, DDQ*) and the tags that due to their frequency contribute most significantly to the overall error rate (*NN1, JJ, VV0, ND1*). CLAWS7 tagset information is placed in Appendix 1.

Therefore, in the present study, the frequency of the previously mentioned lowest precision tags and the tags that significantly influence the error rate has been calculated to check their frequency in the selected analysis material – student-composed letters, essays, chat messages and discussion messages. As it is seen in Table 1, the most crucial for tagging precision of the students' texts are *NN1, JJ* and *VV0*, as it was expected in the light of Van Rooy's and Schaffer's

(2003) findings. Finally, the precision, i.e. the number of tokens that have really received *NN1, JJ* and *VV0* correctly, was calculated: the number of the tokens that have received a correct tag was divided by the total number of tokens.

Table 1 **The frequency of tags**

Tags	Chat %	Letters %	Discussion %	Essays %
RGR	0.09	0.04	0.13	0.26
RRR	0.09	0.04	0.07	0.35
DDQ	0.89	0.72	0.40	0.62
NN1	7.58	16.56	13.20	15.20
JJ	4.50	8.88	6.97	6.80
VV0	2.59	0.96	2.44	2.17

RESULTS AND DISCUSSION

Tagging precision of the three selected tags is shown in Table 2. Even if the overall precision is promising, the actual cases of errors differ across the text groups and therefore the error cases from each text group will be exemplified and discussed.

Table 2 **Tagging precision**

Texts	NN1 %	VV0 %	JJ %
Chat	86.40	88.46	88.88
Letters	92.27	98.19	92.30
Discussion	93.13	96.17	96.72
Essays	94.67	92.50	85.65

The lowest precision, in comparison with CLAWS7 overall performance rate, is displayed by chat messages, which can be explained by the fact that they are interactional, instant and, thus, unedited texts. The bulk of the tagging mistakes are caused by the tagger (76.66%) and also by students' (23.40%) errors (non-word and word errors). The non-word errors have been classified by researchers (Hovermale and Martin, 2008: 3) into spelling errors (words where letters are switched, missing or added) and morphological errors (words which are composed of two correctly spelled parts, but the parts themselves are not correct, e.g. *tooked*).

One of the most common tagger's errors is verb/noun confusion as in the case of the word *finish* (example 2) and also noun/adjective confusion as in the case of the word *sausage* (example 1) and hence the faulty assignment of *NN1*.

(1) What_DDQ is_VBZ white_JJ **sausage_JJ/NN1** line_NP1

(2) Lets_VVZ **finish_NN1/VV0** and_CC go_VV1 home_NN1

The second group of errors refers to acronyms that stand for organisation names and have been tagged in a confusing way (see examples 3, 4), in this case by assigning the tag *JJ* or *NN1*, which means that the quality of the proper noun has not been recognised by the tagger. The same refers to fictional proper nouns and authentic proper nouns (see the example in the reference to essays).

(3) EPICTC-_**JJ/NP1**

(4) MUNO_**NN1/NP1**

The third group of errors is due to foreign words in the text, as CLAWS7 is the English text tagger. Example 5 shows that the greeting in the Spanish language *hola* has been wrongly tagged as NN1.

(5) **Hola_NN1/UH**

Such specific features of chat messages, as seen in examples 6 and 7, have been perceived by the tagger as nouns.

(6) yeeeeeee_NN2

(7) Yeeeee_NP1

Letters display higher tagging precision than chat messages, as letters were edited before their submission to the 'addressees'. The tagger's errors are prevailing (tagger's errors 79.43%, students' errors 20.58%) in these texts. For example, there are repeated cases of NN1 and VV0 confusion (see examples 8 and 9) in these texts. Example 10 shows that the students' use of clipping 'biz' for *business* has been recognised by the tagger and tagged correctly; however, the abbreviation 'gov' that in this text stands for *government* is mistaken by the tagger for the common abbreviation that stands for 'preceding the noun of title'.

(8) Should_VM the_AT Eutropen_JJ Commission_NN1 **coordinate_NN1/VV0** or_CC advise_VV0

(9) Elton_NP1 Jackson_NP1,_, **pop_VV0**/NN1 king_NN1

(10) gov_**NNB/NN1** and biz_NN1

Tagging precision of discussion messages is similar to the precision of letters, even if discussion messages are instant, unedited texts, and obviously, therefore, most of the faulty tags are due to the students' errors (57.55%) that is seen in example 11 (spelling caused confusion of the part of speech), example 12 (the introduction of a space between the word *instead*) and example 13 (the omission of an apostrophe). Examples 14 and 15, however, display the confusion of tags *NN1, JJ* and *VV0*. The word *kind* (example 14) is obviously used as an adjective characterising the quality of *words* and the word *work* is used in the function of a noun, whereas in example 16 *identify* is a verb.

(11) activity_NN1 about_II **weather_NN1/whether_CSW** it_PPH1 is_VBZ

(12) in_II **stead_NN1 of**_IO/ **instead_II21 of_II22**

(13) **Im_VV0/I_PPISI 'm_VBM**

(14) about the_AT **kind_NN1/JJ**,_, pleasant_JJ words_NN2

(15) should_VM reread_VVI his/her_PPGE **work_VVO/NN1**

(16) **identify_NN1/VV0** falsification_NN1

Essays that have been the most carefully edited texts generally display similar tagging precision to the other texts apart from highly interactive chat messages. However, they also display the students' errors that cause the assignment of wrong tags. Thus, example 17 shows a non-word error, a spelling mistake that has led to the wrong tag assignment. Examples 18 and 19 also show the tagging result of the fused spelling of *is not* and *cannot*. Examples 20 and 21 demonstrate that the tagger has not recognised the proper nouns, in this case the place names: the fictional place name *Bardland* and also part of the authentic place name *Britain* has not been tagged precisely, obviously because of the spelling mistake in it.

(17) sitting_VVG in_II a_ATI traffic_NN1 **tram_NN1**

(18) People_NN who_PNQS live_VV0 here_RC **cant_NN1/VM ...**

(19) It **isnt_VV0/VBZ**

(20) Great_JJ **Britan_NN1/NP1**...

(21) **Bardland _NN1/NP1** supports ...

CONCLUSIONS

The analysis of unedited samples of CLAWS7 POS tagged texts that were explored reveal a promising tagging precision. However, the particular error cases vary across the texts grouped according to their communicative purpose. Even if the tagger's error analysis displays regularities (e.g. NN1/VV0 confusion, etc.), the specific features of particular text groups, due to their communicative purpose, have to be taken into account because they can lead to specific tagger's errors (e.g. fictional proper names have not been recognised by the tagger as proper names, foreign words, curious abbreviations). Therefore, the samples of each group of the texts envisaged for the inclusion in the corpus should be tested and considered for the following tagging enhancement options. (1) In case the texts, unedited and interactional (e.g. chat messages, email messages), tend to contain students' errors (word/non-word) that could cause faulty tag assignment, a students' error tagging

scheme should be applied in parallel with POS tagging and the tagger assigned tags should be post-edited. (2) In the case of students' edited texts developed on the basis of several drafts (e.g. untimed essays, papers) that hardly contain any language mistakes, manual or semi-automatic editing of the tagger assigned POS tags can be applied. Additional, more exhaustive research of tagging students' transactional and unedited interactional text samples could further contribute to these preliminary conclusions.

REFERENCES

Aarts, J. and Granger, S. (1998) Tag sequence in learner corpora: a key to interlanguage grammar and discourse. In S. Granger (ed.) *Learner English on Computer.* Essex: Longman Limited.

Adami, E. (2009) To each reader his, their or her pronoun: Prescribed, proscribed and disregarded uses of generic pronouns in English. In A. Ronouf and A. Kehoe (eds.) *Corpus Linguistics: Refinements and Reassessments.* Amsterdam, New York: Rodopi.

Granger, S. (2003) Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal,* 20 (3): 465-480.

Hovermale, D. J. and Martin, S. (2008) *Developing an Annotation Scheme for ELL Spelling Errors.* Department of Linguistics, The Ohio State University, Columbus, OH. Available from http://www.ling.ohio-state.edu/~scott/publications/Hovermale-Martin-MCLC05-2008.pdf. [Accessed on 1 August 2013].

Leech, G. (1997) Introducing corpus annotation. In R. Garside, G. Leech and T. McEnery (eds.) *Corpus Annotation Linguistic Information from Computer Text Corpora* (pp. 1-18). London: Longman.

Leech, G., Garside, R. and Bryant, M. (1994) CLAWS4: The tagging of the British National Corpus. In *Proceedings of the 15th International Conference on Computational Linguistics* (COLING 94; pp. 622-628). Kyoto, Japan.

McEnery, A. (2003) Corpus linguistics. In R. Mitkov (ed.) *The Oxford Handbook of Computational Linguistics* (pp. 448-463). Oxford: Oxford University Press.

McEnery, T., Xiao, R. and Tono, Y. (2006) *Corpus Based Language Studies: An Advanced Resource Book*. Routledge: London.

Reppen, R. (2010) Building a corpus. In A O'Keeffe and M. McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics.* London, New York: Routledge.

Twardo, S. (2012) Selected errors in the use of verbs by adult learners of English at B1, B2 and C1 levels. In *Input, Process and Product: Developments in Teaching and Language Corpora* (pp. 273-282). Brno: Masaryk University Press.

UCREL (2010) Available from http://ucrel.lancs.ac.uk/ [Accessed on August 1, 2013].

Van Rooy, B. and 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 and T. McEnery (eds.) *Proceedings of the Corpus Linguistics 2003 Conference*, 28-31 March, Vol. 16, University Centre For Computer Corpus Research on Language Technical Papers (pp. 835-844). Lancaster, UK: Lancaster University.

APPENDIX 1

EXTRACT OF UCREL CLAWS7 TAGSET

AT	article (e.g. the, no)
CS	subordinating conjunction (e.g. if, because, unless, so, for)
CSW	whether (as conjunction)
DDQ	wh-determiner (which, what)
IO	of (as preposition)
JJ	general adjective
NP1	singular proper noun (e.g. London, Jane, Frederick)
PPGE	nominal possessive personal pronoun (e.g. mine, yours)
PPH1	3rd person sing. neuter personal pronoun (it)
PPIS1	1st person sing. subjective personal pronoun (I)
RGR	comparative degree adverb (more, less)
RRR	comparative general adverb (e.g. better, longer)
VBI	be, infinitive (To be or not... It will be ...)
VM	modal auxiliary (can, will, would, etc.)
VV0	base form of lexical verb (e.g. give, work)
VVD	past tense of lexical verb (e.g. gave, worked)
VVG	-ing participle of lexical verb (e.g. giving, working)
VVGK	-ing participle (going in be going to)
VVI	infinitive (e.g. to give... It will work...)

Zigrīda Vincela (Dr. Philol., Assist. Prof.) is currently working at the University of Latvia. Her research interests include written communication, corpus linguistics and English accents. Email: zigrida.vincela@lu.lv.