Inter-rater reliability testing of article error tags in a small learner corpus: an argument for framework simplicity

UNIVERSITY OF BIRMINGHAM

Richard Nickalls,
English for International Students Unit, University of Birmingham.

r.nickalls@birmingham.ac.uk

1. Background

This study is part of a larger doctoral research project investigating frequent English article errors made by Chinese, Mandarin, Japanese and German L1 learners of English for Academic Purposes. Part of this research requires the identification and analysis of article errors in small corpora of essays submitted by students in each L1 group. Summary of original method:

- A completely automatic method of identfying article errors was rejected due to its reported unreliability. Han, Chodorow, & Leacock (2006) report a Kappa coefficient of just 0.48 in choice of article agreement between a computer and human classifiers.
- My preliminary investigation into the English article 'interlanguage' of Mandarin L1 Learners' of English (Nickalls, 2011) used an adaptation of Bickerton's semantic space framework (1981). As shown in Figure 1, this much-used framework (Heubner 1983; Fen Chuan 2001; Humphrey 2007; Diez- Bedmar & Papp 2008) asks researchers to construe whether a referent is specific [± SR] and known [± HK] to the hearer.

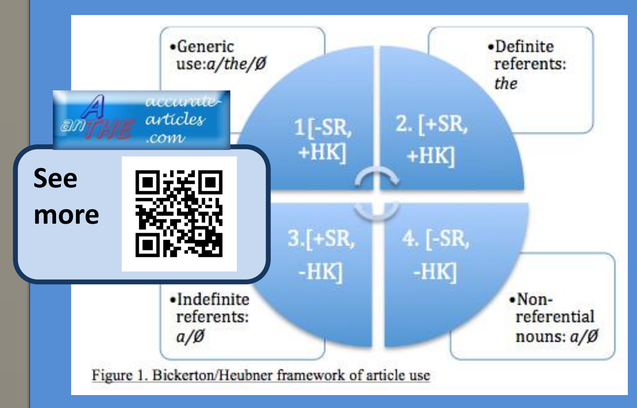

Figure 1. Bickerton/Heubner framework of article use

Research questions

The methodology outlined above rested on the assumption that a single human annotator could reliably tag a corpus not only in terms of language correctness, but also in terms of a complex explanatory classification framework. This poster outlines the use of inter-rater reliability tests to check:

1. To what extent would raters reliably classify article use as 'correct' or 'incorrect'?
2. Would correctness be consistently classified over time?
3. How reliably would the complex classification framework be applied?

2. Methodology

1) A stratified random sample of 112 noun phrases was taken from a small 50,000 word corpus of 30 Mandarin L1 students' essays (Nickalls, 2011).

2) The four classifiers (including the researcher) were all L1 speakers of English with over 10 years of English teaching experience, holding British Council TEFL-Q level teaching equivalent qualifications with either a Masters or Doctorate education level.

The three new raters received identical training and reference materials and each was shown four short tutorial videos. Each also underwent standardisation on a virtual learning environment (VLE) interface, classifying 25 noun phrases and receiving immediate online feedback.

Try a test!
Enrollment key= 'Birmingham'

3) The noun phrases were first tagged for correctness using the online interface, which showed the noun phrases within their immediate sentence and a hyperlink to view the whole essay context. Three weeks later the researchers tagged the same noun phrases again for correctness and then according to the Bickerton/Heubner framework.

3. Results – correctness

The four raters were relatively consistent in their judgements of correctness. In the first session, the only dichotomous choice for raters was of 'correctness'. The raters were instructed to 'classify as incorrect only if stylistically or grammatically impossible within an academic writing context without change'. As shown in table 1, the Cohen's Kappa and Krippendorf's Alpha between rater 1 (the researcher) and other raters were all above 0.7.

Table 1: Coefficient comparisons between first rater and others

	Percent Agreement	Agreements	Disagreements	Cases	Scott's Pi	Cohen's Kappa	Krippendorff's Alpha
rater 1 and 2	86.61	97	15	112	0.73	0.73	0.73
rater 1 and 3	90.18	101	11	112	0.80	0.80	0.80
rater 1 and 4	87.50	98	14	112	0.75	0.75	0.75

Fleiss' Kappa is a reliability measure designed for multiple raters. As outlined in table 2 below , a Fleiss' Kappa of 0.74 again shows relatively consistent agreement about correctness among all raters.

Table 2: Overall correctness ratings in first session of all raters

Coders	Cases	Average pairwise percent agreement	Fleiss' Kappa	FK observed agreement	FK expected agreement	Krippendorff's Alpha
4	112	86.9	0.74	0.87	0.50	0.74

In the second session (three weeks later) the raters were asked to review 28 of the same noun phrases and classify them again in terms of their correctness/the explanatory framework. As shown in table 3, the first rater made identical correctness classifications to the ones made two years previously, while other raters also showed generally consistent reliability coefficients.

Table 3: Consistency of rating over time.

	Percent Agreement	Scott's Pi	Cohen's Kappa	Krippendorff's Alpha
Rater 1	100	100	1.00	1.00
Rater 2	89.29	0.79	0.79	0.79
Rater 3	96.43	0.93	0.93	0.93
Rater 4	89.29	0.79	0.79	0.79

4. Results - testing of framework

The Bickerton-based framework showed much lower inter-rater agreement. As can be seen in Table 4, a Fleiss' Kappa of .26 shows that the raters application of this framework was far less reliable (Fleiss' Kappa takes into account the 0.25 probability of chance agreement).

Table 4: Reliability of the application of the explanatory framework

Average pairwise agreement	Fleiss' kappa	Krippendorff's Alpha
44.64%	0.26	0.26

5. Conclusion

A manual error tagging technique is generally reliable in terms of annotating 'correctness' of articles in Academic English. English L1 raters can generally agree about what is grammatically 'possible' and 'not possible' in the use of articles. They also remain fairly consistent in these judgements over time. These findings confirm assumptions that human raters are more reliable than automated computer methods.

However, in terms of the classifications with the complex (Heubner/Bickerton) explanatory framework, this study shows that four raters could not use the framework consistently. Raters cannot apply such classification frameworks, in which the decision goes beyond a rater's dichotomous intuition, so reliably. It was particularly unreliable in choices between generic, indefinite, non-referential and idiomatic contexts.

The above conclusions must remain tentative, given the small number of raters. However, they raise questions about studies which publish findings without details of inter-rater reliability tests. Further rater training (to standardise use of the Heubner/Bickerton framework) might arguably have resulted in higher reliability coefficients, but probably only marginal improvements among one team of raters, while the implied benefit of such frameworks is that studies can be replicated. There seems to be a need for a simpler framework which has greater reliably for teams investigating English article use.

Find out more about this research at: www.accurate-articles.com

Use your mobile device with the QR code whenever you see the logo above.

References

Bickerton, D. (1981) *Roots of language*. Repr. 1985. Ann Arbor, Mich.: Karoma.

Diez-Bedmar, M.B. and Papp, S. (2008) 'The use of the English Article System by Chinese and Spanish learners'. In Gilquin, G., Papp, S. & Diez-Bedmar, M.B. (Eds.) *Linking up Contrastive and Learner Corpus Research.* Amsterdam, Atlanta, Rodipi, pp 147-175.

Fen Chuan, C. (2001) 'The acquisition of English articles by Chinese learners'. *Second Language Studies*: 20(1): 43-78.

Han, N., Chodorow, M., Leacock, C. (2006) 'Detecting errors in English article usage by non--native speakers'. *Natural Language Engineering*: 12, pp 115--129

Huebner, T. (1983) Longitudinal analysis of the acquisition of English. Ann Arbor: Karoma Pubr.

Humphrey, S.J. (2007). 'Acquisition of the English Article System: Some Preliminary Findings'. *Journal of School of Foreign Languages* [Online] available from http://library.nakanishi.ac.jp/ [accessed on 1/6/2011].

Nickalls, R. (2011) 'How definite are we about articles in English? A study of L2 learners' English article interlanguage during a university presessional English course'. *proceedings from the 2011 Corpus Linguistics Conference,* #92 available from the University of Birmingham [online] Centre for Corpus Research Website.