Inter-annotator Agreement: By Hook or by Crook

John Blake
University of Aizu
Aizu-wakamatsu, Japan.
jblake@u-aizu.ac.jp

Abstract

Through an extended case study, this paper reveals the metaphorical skeletons hidden in statistical cupboards of selective reporting, casting a new light on inter-annotator agreement (IAA) measures. Strategic decisions and their impacts on IAA were tracked in an extended corpus study of rhetorical functions in scientific research abstracts. A search of the research notes of the principal investigator resulted in 142 notes tagged with #IAA that were written between 2013 and 2017. The strategic decisions and their actual or perceived impacts on IAA were logged. A root cause analysis was also conducted to identify the causal factors that reduce IAA. The results show numerous strategic decisions, which using template analysis, were grouped into three categories, namely methodological, statistical and rhetorical. High IAA may be attributed to sound or cogent methodological choices, but it could also be due to manipulating the statistical smoke and rhetorical mirrors. With no standardized convention for reporting IAA in corpus linguistics, researchers can select statistics that portray IAA more or less positively. The metaphorical skeletons hidden in statistical cupboards of selective reporting are revealed, casting a new light on IAA measures of agreement and disagreement. Practical guidelines on best practice are suggested.

Keywords: annotation, reliability, inter-annotator agreement

1. Introduction

This case study focuses on the impact of strategic decisions on inter-annotator agreement (IAA) for manual annotations. Although this paper focusses on the IAA between human annotators; where relevant, automated annotations are also discussed.

1.1 Summary

Corpus linguists working with annotated corpora are often required to prove (or, more accurately, fail to disprove) the veracity of annotations. A multi-year extended corpus research project serves as the vehicle of this case study. During this corpus project, the principal researcher noticed that some decisions could improve the quality of annotation but would result in lower IAA scores. This realization was the impetus for this case study. The results of this study reveal metaphorical skeletons hidden in the statistical cupboards of selective reporting, casting a new light on inter-annotator agreement measures. Inter-annotator reliability and its reporting are problematized as epistemic rather than purely statistical.

1.2 Purpose

The primary purpose of this case study was to investigate the actual or perceived effects on IAA of decisions, which were made during an extended corpus linguistics research project. There is a paucity of pedagogic literature that shows corpus linguists (or other linguists) how to develop an annotated corpus, and even less on how to achieve high IAA. The research literature and conference landscape are also rather sparse. The Association for Computational Linguistics Special Interest Group SIGANN is one of the few organizations that arranges conferences specifically dealing with corpus annotation. This paper aims to address issues pertinent to the needs of corpus linguists working with annotated corpora.

1.3 Overview

The following section provides background details on annotation, annotated corpora, inter-annotator agreement and reported IAA measures. The Method section introduces the extended corpus annotation project from which the data was collected. The two main methods of analysis, namely root cause and template analysis are explained. The results section describes the three types of strategic decisions that affect IAA. Methodological, statistical and rhetorical choices are discussed using examples from the case study. The Results and Discussion sections describe, explain and exemplify issues that are commonplace in corpus annotation projects. The discussion focuses around how IAA can be increased by methodological and statistical choices. The use of rhetorical choices is also addressed from the viewpoints of writers and readers of research articles. This paper concludes with a set of practical guidelines on best practice. These include suggestions, such as the creation of an annotation booklet with clear rules, worked examples and discussion of boundary cases. The final recommendation is to report IAA in sufficient detail to convince skeptical readers of both the rigour and the validity of the reported IAA.

2. Background

In computational linguistics, corpora are often created for machine learning purposes, and so the accuracy of the annotations is of paramount importance. This may explain why most research findings and reference materials on IAA can be found in journals and conferences dealing with natural language processing. Corpus linguists have, according to Gries (2015), started to transition to the use of more sophisticated quantitative methods. This trend may also crossover to measuring and reporting IAA.

2.1 Annotation

Annotation involves assigning labels to language items. The items annotated can range from structural to functional, semantic to pragmatic. Vagueness and ambiguity are prevalent in natural languages (Wasow, Perfors, and Beaver, 2005). This is one of the many possible reasons why annotators may differ on their label assignment.

2.2 Annotated corpora

The addition of layers of annotation adds value to a corpus (Leech, 2005, p.1) by making the linguistic information

explicit, searchable and easily accessible (McEnery and Wilson, 2001, p.32). A tagged corpus frames the contents of the corpus, which was a key objection of critics, such as Sinclair (2004, p.191). However, without annotation, many research questions would remain unanswered, and so the question is not whether to annotate but how to ensure accurate annotation (Hunston, 2002).

2.3 Inter-annotator agreement (IAA) measures

Researchers cannot measure the correctness of annotations directly (Boleda & Evert, 2009), and so resort to reliability as a proxy variable. Reliability of annotations can be evaluated through various IAA measures. The underlying assumption is that high IAA rules out unreliability and allows a claim for validity. Inter-annotation measures are, therefore, used as a proxy for reliability and validity. Interestingly, high IAA does not guarantee accuracy, but simply shows the high degree of agreement between or among annotators.

According to Bayerl and Paul (2011), simple measures, such as observed or raw agreement are the most frequently used. These measures, however, are far from reliable. Simple IAA measures, such as simple ratios often fail to take account of chance agreement (Carletta, 1996; Artstein and Poesio, 2008), which is one reason why more sophisticated measures, such as the Kappa/alpha family (Artstein, 2017) were developed. For the Kappa coefficient, there are three commonly used interpretations, which all differ in their precise ranges. As emphasized by Von Eye (2014), a score of 0.75 can be interpreted as tentative (Krippendorff, 1980) or substantial (Landis and Koch, 1977). Therefore, not only the selection of statistic but its interpretation affects the reported IAA.

```
install.packages("irr")

library(irr)

ds.full <- read.delim("file_name")

ds.iaa <-
data.frame(ds.full$attributive,
ds.full$attributive.anno2)

ds.iaa.sharedobs <- droplevels(
ds.iaa[ds.iaa$ds.full.attributive.a
nno2 !="", ] )

table(ds.iaa.sharedobs)

kappa2(ds.iaa.sharedobs)
```

Fully commented version available on:
https://corpuslinguisticmethods.wordpress.com/2014/01/15/what-is-inter-annotator-agreement/

Figure 1: Example script for R to calculate IAA

Rather than measuring agreement alone, both agreement and disagreement can be considered, for example using Measuring Agreement on Set-valued Items (MASI) and/or Jaccard distance. Both MASI (Passonneau, 2004) and Jaccard distance make use of the union and intersection between sets.

Annotation projects that harness natural language pipelines such as the Natural Language Toolkit (NLTK) (Bird and Loper, 2004) and GATE Teamware (Bontcheva *et al.* ,2013) can calculate IAA measures easily as this functionality is already integrated into the software. In GATE the annotation diff tool can be used to compare two sets of annotations while in NLTK the nltk.metrics.package can be used. IAA measures can be calculated in R for statistics in a few lines of code. An example script is given in Figure 1.

2.4 Reported IAA measures

Claims of annotation accuracy of around 97% are made for part-of-speech (POS) automatic taggers (e.g. Baker, 1997). This percentage is, however, calculated per word, and so when applied to a 20-word sentence, the probability of the whole sentence being tagged accurately drops to slightly over 50% (Manning, 2011). Yet, IAA far higher than 50% is frequently expected for annotations, which may involve far more subjective judgment calls than part-of-speech tagging. Fellbaum *et al.* (1998) provide a detailed discussion of the difficulties annotating word senses by both lay and expert annotators. The POS-tagger example shows that the reported IAA could be 97% when assessed at the unit of word, but falls to around 50% when assessed by sentence. Currently, annotation practices vary greatly, sharing annotation practices and standards will help corpus annotators take their research to greater heights (Gries and Berez, 2017).

3. Method

3.1 Extended corpus project

This case study focusses on the strategic decisions made during a multi-year study of the rhetorical moves in a corpus of 1000 research abstracts from top-tier journals in ten scientific disciplines (100 abstracts per discipline). The corpus specifications are given in Table 1.

Rhetorical moves in abstracts were coded by the principal annotator, and subsets of the corpus were coded by multiple annotators. Annotators included both linguists and subject specialists.

Discipline	Number of words
Evolutionary Computation	17,433
Knowledge and Data Engineering	18,407
Image Processing	16,859
Information Theory	15,982
Wireless Communications	15,971
Advanced materials	6.078
Botany	19,981
Linguistics	13,587
Industrial Electronics	14,569
Medical	29,437

Table 1: Corpus specifications

Moves were defined as "a discoursal or rhetorical unit that performs a coherent communicative function" following Swales (2004, p.228). Abstracts in many of these domains are impenetrable to lay readers due to the accumulation of

highly specialized terminology, creating a particularly challenging corpus to code. This proved particularly problematic for the linguist annotators.

An example of a very short abstract from the Information Theory corpus is given in Figure 2. In this abstract there are two moves: a result and a method. However, most lay readers can understand little of the actual content.

> In this paper, we prove that there does not exist a binary self-dual doubly even code with an automorphism of order 9. To do so, we apply a method for constructing binary self-dual codes possessing an automorphism of order for an odd prime.

Figure 2: Research abstract in Information Theory

Various versions of the UAM Corpus Tool (O'Donnell, 2008) were used to annotate this corpus. Each sentence was coded with a move and a sub-move if appropriate. The <uncertain> tag was used as a temporary label. The annotation schema of the tagset is shown in Figure 3:

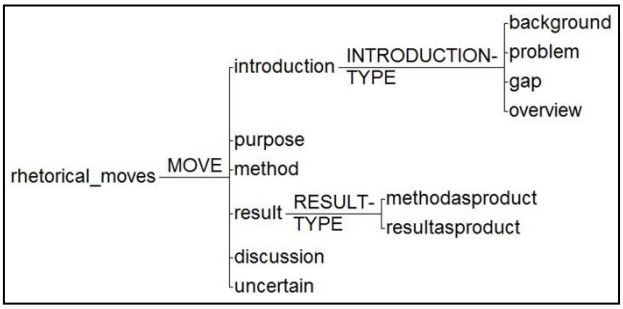

Figure 3: Annotation schema for rhetorical moves

3.2 Research notes

Strategic decisions and their impacts on IAA were tracked in an extended corpus study of rhetorical functions in scientific research abstracts. A search of the research notes of the principal investigator resulted in 142 notes tagged with #IAA that were written between 2013 and 2017.

3.3 Strategic decisions

The strategic decisions and their actual or perceived impacts on IAA of the content of each research note were considered and logged.

3.4 Root cause analysis

A root cause analysis was conducted to identify the causal factors that affect IAA score. This was facilitated using Ishikawa fishbone diagrams. Each cause was traced back using the five-why approach to find the root cause.

3.5 Template analysis

The set of research notes shows numerous strategic decisions and rhetorical choices that had to be made during each stage of the research. Each research note was coded using template analysis (King, 2004). Template analysis is midway between grounded theory in which codes are not determined *a priori* and content analysis in which all codes are predetermined. Codes were grouped by similarity and specific differences. Codes were merged, subsumed or re-classified during the process. The resultant code set comprised three broad categories, namely methodological, statistical and rhetorical.

4. Results

The template analysis resulted in three categories. The first two categories of methodological and statistical choices affect both the quality of the annotations and the associated IAA measures. The final category of rhetorical choices, however, only affects the reported IAA measures. This category shows that since some decisions regarding IAA may be based on techniques of psychological persuasion rather than linguistic science. The following three sub-sections describe and exemplify the methodological, statistical and rhetorical choices in turn. Only the findings that are generalizable to other projects are reported here.

4.1 Methodological choices

The methodological choices aim to affect the judgments of the annotators in such a way that annotators make the same judgment call about which label to assign to a language item. Some of the methodological choices that enhance IAA include ontological unit, size of tagset, clarity of tag demarcation, the presence of catch-all tags, detailed annotation booklet, training and testing, easy-to-use tools, monitoring and pilot studies. The following subsections detail nine methodological choices.

1. Ontological unit

Fixed ontological units simplify the calculation of IAA and may increase IAA since the boundaries of each unit are identical. Variable ontological units provide researchers with additional options on how to calculate (manipulate?) IAA. Identical units, subsumed units and cross-over units need to be considered. Reporting the agreement by word, letter or character (including the white space characters, e.g. U+0020) results in completely different values.

2. Tagset size

Simply put, the more tags, the less agreement. With hindsight, this is obvious, but when corpus linguists develop a tagset, the purpose is to inform their research rather than to secure higher IAA. A tagset of one item will secure total agreement, but no reason to code while a tagset of ten items is going to result in far more agreement than one of hundreds of items. Rissanen (1989, 2018) points out the "mystery of vanishing reliability", i.e. the statistical *un*reliability of annotation that is too detailed.

3. Tag clarity of demarcation

It is not possible to discover problem cases without annotating. In this research, two sets of tags were used before the final version. The first tagset was dropped because of the difficulty in demarcation of boundary cases.

4. Catch-all tags

Archer (2012,n.p.) describes four tag types, all of which increase IAA. These catch-all tags provide easy-to-code options for boundary cases. Fuzzy tags are used when it is difficult to assign a tag from the existing tagset, multiple tags are used when more than one tag applies, portmanteau tags are used when an item transcends two tag domains and

problematic tags are used when it is impossible to assign a tag.

In this case study there is an uncertain tag which was designed purely as a temporary tag. Should IAA measure include moves coded as uncertain, the IAA would likely be higher. This is because the difficult-to-code moves, are likely to be classed as uncertain, and so although this does not inform the research, it does yield higher IAA.

5. Annotation booklet
Codifying a standard operating procedure (SOP) can enhance IAA. The use of guidelines, rules, prototypical examples and especially examples in which borderline cases are disambiguated can help annotators make similar judgments. There is a caveat though: just because annotators allocate the same label, it does not mean the label is correct. The annotation booklet also provides a fixed point of reference, which should help not only inter-annotator reliability, but also help maintain intra-annotator reliability particular when a study spans years.

6. Training course and benchmark test
Requiring annotators to complete a training course ensures that annotators have actually practiced using the guidelines. In this case study, an online course was created on a learning management system (Figure 4). At the end of each stage formative assessment tests were provided with automated feedback. At the end of the course, candidates took a benchmark test. The benchmark test was used to identify the suitability of the candidates. In this study, candidate annotators who scored 90% were assigned annotation tasks, those scoring between 60% and 89% were offered the chance to retrain and retry while those were scored less than 59% were judged as being unsuitable.

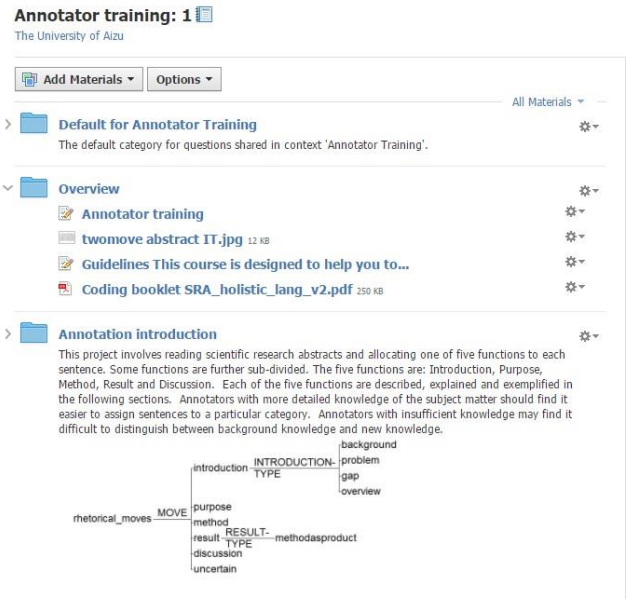

Figure 4: Screenshot of Online training course

7. Easy-to-use tools
This study used the UAM Corpus Tool (Figure 5). The selection of this tool was based primarily on its functionality and ease of use. Although it was easy to use, there were many bugs in the software. Through versions 2 to 3, the severity and frequency of problems reduced, but using the tool was far from stress-free. It was difficult to resolve problems with the tool as the instructions and help forum were limited. This resulted in some qualified annotators dropping out. To go some way to address this, a project-specific instruction booklet was created. With hindsight, it would have been time-saving to use a more sophisticated tool, such as GATE teamware, from the outset.

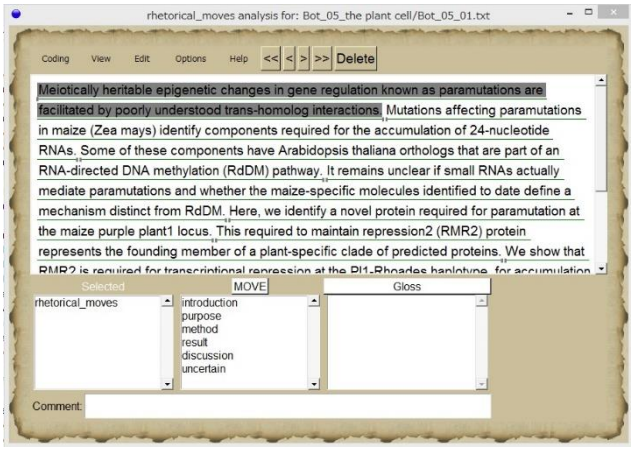

Figure 5: Screenshot of UAM Corpus Tool interface

In the final stage of the case study, an online move visualizer was created to enable specialist informants to alter, comment or confirm the accuracy of the annotations. The aim of the visualizer was not to increase IAA, but to increase the accuracy of the final tagset. This alleviates the need for double annotators to learn how to use a new piece of software as the interface is extremely simple with only three choices: move to next abstract, confirm annotation accuracy or comment on accuracy. This visualizer is used for verifying the accuracy of the final dataset, which was developed after a number of rounds of annotation.

8. Monitoring, feedback and regular meetings
The recruitment, training and retention of annotators is time consuming. Experienced annotators are likely to produce annotations that show higher IAA. This is because they are more familiar with the genre, task, tools and SOP. Three ways to increase retention are by: (1) monitoring carefully in initial stages to identify and solve problems early, (2) providing constructive actionable feedback, and (3) scheduling regular short meetings.

9. Pilot studies and small trials
Using pilot studies and small trials provides the perfect opportunity to test out tags, tagsets, ontological units, annotation guides and software. Using double annotators is expensive in terms of time and/or money, and so it is worthwhile investing time upfront to ensure the annotation procedure is as straightforward as possible.

4.2 Statistical choices

Five decisions relating to statistical choices were found to affect IAA. These choices are the population-sample ratio, method of selection, treatment of outliers, sample selection

timing and granularity. Each of these choices are detailed in the following subsections.

1. Population-sample ratio
In this study double annotators initially coded between 10% and 100% of each discipline within the corpus. The medical abstracts resulted in an exceptionally high degree of annotator agreement, while abstracts in information theory resulted in far less agreement. In disciplines that were easier to code, annotators coded more abstracts. Thus, reporting the true total number of double annotated abstracts regardless of discipline results in a higher IAA than reporting the IAA for 10% of each discipline.

2. Method of selection
When double annotation is conducted for a subset of a corpus, the method of selection can affect IAA. In this case study, annotations by the principal investigator from 2013 showed less reliability than later annotations completed in 2016, and so a random sample shows higher IAA than a subset of the 2013 abstracts, but lower IAA than a subset of the abstracts annotated in 2016. All early abstracts have since been reannotated once this was discovered.

3. Treatment of outliers
When the research aim is to investigate prototypical features, outliers in data can skew results. The inclusion or exclusion of outliers, such as abstracts that show no agreement on any moves could be excluded from the final dataset. However, should the reported IAA include outliers even though they are no longer part of the study?

4. Sample selection timing
When double annotation results differ, meetings may be held to discuss the differences. These meetings may result in annotators agreeing to code in the same way. In this case, is the reported IAA the one that represents the final dataset?

5. Granularity
Granularity and reliability in discourse annotation may be viewed as working in opposition, and so there is a need to achieve the optimum balance (Crible and Degand, 2017).
In this case study, sentences were coded for moves and submoves using a tagset comprising five moves and six submoves. However, when reporting IAA, there is the option to report only for the moves, which is likely to increase IAA. If the moves were reduced to four or five moves, IAA would likely increase (Rissanen, 1989, as cited in Archer, 2012, n.p.). Figure 6 shows a subset of the corpus at different levels of granularity from full code of eleven items to four items.

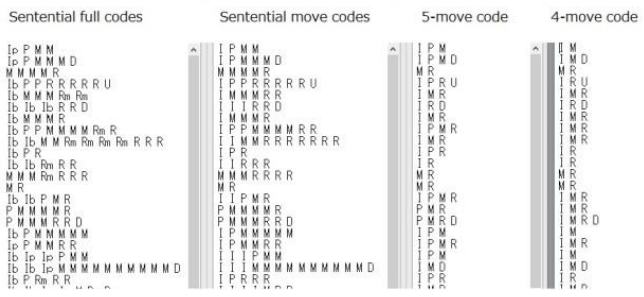

Figure 6: Codes for abstracts at different levels of granularity

4.3 Rhetorical choices

Some rhetorical choices may lead readers to assume or infer higher IAA than was actually achieved. Researchers harness language to portray their research in a positive manner. Researchers are responsible for the selection of which information to emphasize, de-emphasize or omit; and which wording to use to position their research results in the desired manner.

Researchers who wish to hide the actual inter-annotator agreement may rely vagueness, ambiguity and framing to entice readers into inferring that their IAA scores are higher. Many researchers may not report IAA in much details, simply because they do not place much emphasis on providing enough details for readers to judge the annotation procedure and calculation of IAA measures.

Corpus linguists report high IAA in varying degrees of detail. Commonly found options in the research literature are:
1. no further details.
2. simple statistic (e.g. a percentage or ratio)
3. size of doubly-annotated corpus
4. simple statistic and size of doubly-annotated corpus

What is lacking is the specific details on how IAA was calculated. IAA measures are rarely reported in sufficient detail for another researcher to replicate the process leading to the calculation.

5. Discussion

Rhetorical smoke and statistical mirrors can be used to convince non-critical readers of high IAA. This smoke-and-mirrors tactic relies on claiming high IAA while providing, at most, sparse details, allowing readers to infer higher IAA than might have been reported had fuller details been disclosed. With no standardized convention for reporting IAA in applied linguistics, researchers are able to report a high IAA through careful strategic choices (e.g. categorization and ontological units) and statistical analysis (e.g. sampling fraction, outliers and tests). However, it should be remembered that the degree of IAA required depends on what the annotations will be used for (Passonneau, 2006).

Seventy percent of statistics are meaningless. This is an oft-quoted pun, yet the desire to persuade others entices us to use numerical data to support arguments. These data, despite their inherent sources of potential error, tend to be treated as valid (Hammersley and Gomm, 1997). High IAA may be attributed to sound or cogent methodological choices, but it could also be due to manipulating the statistical smoke (i.e. selecting parameters leading to higher IAA) and rhetorical mirrors (i.e. using vagueness/ambiguity to allow the inference of high IAA). In many publications in the field of corpus linguistics, sufficient details are not provided. Lack of details may be due to the lack of the perceived need to declare details, lack of rigour or lack of IAA.

6. Suggested best practice guidelines

Although achieving IAA is not the primary aim of any corpus study, journal reviewers and thesis supervisors are likely to raise the topic when a corpus is annotated. In order to provide support for claims of accuracy, suggested best practice guidelines that are supported by evidence from this case study are:

1. Annotate using tags at one level more finely than the research question requires.
2. Provide clear rules and examples in which boundary cases discussed in an annotation booklet.
3. Develop, trial and require all annotators to complete a training course or session.
4. Require annotators to reach a benchmarked standard.
5. Monitor and provide constructive actionable feedback to annotators.
6. Report IAA in sufficient detail to convince skeptical readers.

7. References

Archer, D. (2012). Corpus annotation: A welcome addition or an interpretation too far? Studies in Variation, Contacts and Change in English. Volume 10. Research Unit for Variation, Contacts and Change in English (VARIENG). Available online: http://www.helsinki.fi/varieng/series/volumes/10/archer/

Artstein, R. (2017). Inter-annotator agreement. In N. Ide & J. Pustejovsky (Eds.) *Handbook of Linguistic Annotation (pp. 297-313).* Springer: Dordrecht.

Artstein, R. and Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, *34* (4), 555–596.

Baker, P. (1997). Consistency and accuracy in correcting automatically tagged corpora. In R. Garside, G. Leech & A. McEnery, (Eds.), *Corpus Annotation: Linguistic Information From Computer Text Corpora*, *(pp.243–250).* London: Longman.

Bayerl, P. S. & Paul, K. I. (2011). What determines inter-coder agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics*, *37*(4),699–725.

Bird, S., & Loper, E. (2004, July). NLTK: the natural language toolkit. *In Proceedings of the ACL 2004 on Interactive poster and demonstration sessions (p. 31).* Association for Computational Linguistics.

Boleda, G. & Evert, S. (2009, July 28). Inter-annotator agreement: Computational lexical semantics. Handout from European Summer School in Logic, Language and Information. Available online: https://clseslli09.files.wordpress.com/2009/07/02_iaa-handout1.pdf

Bontcheva, K., Cunningham, H., Roberts, I., Roberts, A., Tablan, V., Aswani, N., & Gorrell, G. (2013). GATE Teamware: a web-based, collaborative text annotation framework. *Language Resources and Evaluation*, *47*(4), 1007-1029.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, *22*(2), 249–254.

Crible, L., & Degand, L. (2017). Reliability vs. granularity in discourse annotation: What is the trade-off?. *Corpus Linguistics and Linguistic Theory*. Available online: https://doi.org/10.1515/cllt-2016-0046

Gries, S. Th. (2015). The most underused statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora 10*(1). 95-125.

Gries, S. Th. & Berez, A. L. (2017). Linguistic annotation in/for corpus linguistics. In Nancy Ide & James Pustejovsky (eds.), *Handbook of Linguistic Annotation, (pp.379-409).* Berlin & New York: Springer.

Fellbaum, C., Garabowski, J., Landes, S., Baumann, A. (1998). Matching words to senses in WordNet: Naïve versus expert differentiation. In C. Fellbaum, (Ed.) *WordNet: An Electronic Lexical Database, (pp. 217–239).* MIT Press : Cambridge.

Hammersley, M. and Gomm, R. (1997). Bias in Social Research. Sociological Research Online, 2 (1). Available online: http://www.socresonline.org.uk/socresonline/2/1/2.html

Hunston, S. (2002). *Corpora in Applied Linguistics. Cambridge* : Cambridge University Press.

King, N. (2004). Using templates in the thematic analysis of text. In C. Cassell & G. Symon (Eds.), *Essential guide to qualitative methods in organizational research (pp. 256–270).* London: Sage. Available online: http://dx.doi.org/10.4135/9781446280119.n21

Krippendorff, K. (1980). *Content Analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills, CA, USA.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

Leech, G. (2005). Adding Linguistic Annotation. In M. Wynne, (Ed), *Developing Linguistic Corpora: A Guide to Good Practice (pp.17–29).* Oxford: Oxbrow Books.

Manning, C. D. (2011). Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? In Alexander Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing, 12th International Conference, CICLing 2011, Proceedings, Part I. Lecture Notes in Computer Science 6608,(pp. 171—189).* Springer.

McEnery, T. & A. Wilson. (2001). *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

O'Donnell, M. (2008, April). The UAM CorpusTool: Software for corpus annotation and exploration. In *Proceedings of the XXVI Congreso de AESLA (pp. 3-5).* Almeria Spain.

Passonneau, R. (2004). Computing reliability for coreference annotation. *Proceedings of the International Conference on Language Resources and Evaluation (LREC). Portugal*.

Passonneau, R. (2006). Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. *Proceedings of the International Conference on Language Resources and Evaluation.* Paris.

Rissanen, M. (1989). Three problems associated with the use of diachronic corpora. *ICAME Journal 13*, 16–19.

Rissanen, M. (2018). Three problems associated with the use of diachronic corpora [reprinted]. *ICAME Journal 13*, 16–19

Sinclair, J. 2004. *Trust the Text: Language, Corpus and Discourse.* London: Routledge.

Swales, J. M. (2004). *Research Genres: Explorations and Applications.* Cambridge: Cambridge University Press.

Von Eye, A., & Mun, E.Y. (2014). *Analyzing Rater Agreement: Manifest Variable Methods.* NY: Psychology Press.

Wasow, T., Perfors, A., & Beaver, D. (2005). The puzzle of ambiguity. In C. Orhan, and P. Sells (Eds.), *Morphology and the web of grammar: Essays in memory of Steven G. Lapointe, (pp. 265-282).* Chicago: University of Chicago Press.