Inter-Annotator Agreement in linguistica: una rassegna critica

Gloria Gagliardi
FICLIT – Università di Bologna, Italy
gloria.gagliardi2@unibo.it

Abstract

Italiano. I coefficienti di Inter-Annotator Agreement sono ampiamente utilizzati in Linguistica Computazionale e NLP per valutare il livello di "affidabilità" delle annotazioni linguistiche. L'articolo propone una breve revisione della letteratura scientifica sull'argomento.

English. *Agreement indexes are widely used in Computational Linguistics and NLP to assess the reliability of annotation tasks. The paper aims at reviewing the literature on the topic, illustrating chance-corrected coefficients and their interpretation.*

1 Introduzione

La costruzione di risorse linguistiche, e più in generale l'annotazione di dati, implicano la formulazione di giudizi soggettivi. La necessità di stabilire fino a che punto tali giudizi siano affidabili e riproducibili ha assunto crescente importanza, fino a rendere le procedure di validazione prassi consolidata. Ciò è avvenuto in linguistica computazionale (LC) con più di 30 anni di ritardo rispetto alla psicometria: già nel 1960 Cohen, in un celebre articolo, scriveva infatti:

> *"Because the categorizing of the units is a consequence of some complex judgment process performed by a 'two-legged meter' [...], it becomes important to determine the extent to which these judgments are reproducible, i.e., reliable."*
>
> (Cohen, 1960: 37)

È convinzione abbastanza diffusa che un alto livello di *Inter-Annotator Agreement* (da ora in poi: I.A.A.) tra gli annotatori sia indice della bontà e della riproducibilità di un paradigma di annotazione. Come sottolinea Di Eugenio:

> *"This raises the question of how to evaluate the 'goodness' of a coding scheme. One way of doing so is to assess its reliability, namely, to assess whether different coders can reach a satisfying level of agreement with each other when they use the coding manual on the same data."*
>
> (Di Eugenio, 2000: 441)

L'assunto di base è dunque che i dati siano considerabili "attendibili" se due o più annotatori sono in accordo nell'individuare un fenomeno linguistico oppure nell'assegnare una categoria all'item in analisi. In tale prospettiva, la *reliability* si configura perciò come prerequisito per dimostrare la validità di uno schema di codifica, e un ampio consenso tra gli annotatori viene assunto a garanzia della precisione intrinseca del processo di annotazione (Warrens, 2010).

> *"The main reason for the analysis of annotation quality is to obtain a measure of the 'trustworthiness' of annotations. [...] Only if we can trust that annotations are provided in a consistent and reproducible manner, can we be sure that conclusions drawn from such data are likewise reliable and that the subsequent usage of annotations is not negatively influenced by inconsistencies and errors in the data. Inter-annotator (or inter-coder) agreement has become the quasi-standard procedure for testing the accuracy of manual annotations."*
>
> (Bayerl & Paul, 2011: 700)

In ambito computazionale l'I.A.A. è usato come veicolo per passare dal materiale annotato ad un *gold standard*, ovvero un insieme di dati sufficientemente *noise-free* che serva per *training* e *testing* di sistemi automatici. Di prassi i coefficienti di *agreement* vengono usati per assicurare la bontà della procedura di annotazione e del materiale annotato: un alto livello di I.A.A. fa sì che il fenomeno sia considerato consistente e sistematico, e che la risorsa validata sia idonea per addestrare un sistema automatico che svolga il medesimo compito del linguista.
In realtà, l'idea che l'I.A.A. possa indicare in senso assoluto la qualità del *dataset* come risorsa di riferimento è fallace: due osservatori possono, pur sbagliando entrambi, essere in perfetto accordo nel valutare un evento:

> *"However, it is important to keep in mind that achieving good agreement cannot ensure validity: two observers of the same event may well share the same prejudice while still being objectively wrong."*
>
> (Artstein & Poesio, 2008: 557)

È inoltre opportuno considerare che l'*agreement* raggiunto abitualmente dagli annotatori varia in

relazione al livello di esperienza: l'I.A.A. in gruppi omogenei è comparabile a prescindere dai livelli di esperienza, ma si abbassa qualora vengano formati gruppi misti di esperti e non esperti:

> *"Implicit in discussions of inter-annotator agreement is that coders not only agree on which unit belongs to which category, but that if they agree these decisions are also correct with respect to the phenomenon under scrutiny [...]. In our study, this assumption left us with a dilemma. Our data showed that experts and non-experts could achieve comparable levels of agreement, whereas the average agreement for mixed groups was significantly lower. In other words, experts and novices were equally reliable, yet did not agree with each other."*
>
> (Bayerl & Paul, 2011: 721)

Non tutti i task di annotazione linguistica sono valutabili secondo le stesse procedure; dal punto di vista qualitativo, si possono individuare almeno due tipologie generali (Mathet, Widlöcher, A. & Métivier, 2015):

- "individuazione di unità" o "*unitizing*" (Krippendorff, 1980), in cui l'annotatore, dato un testo scritto o parlato, deve identificare posizione e confine degli elementi linguistici (es. identificazione di unità prosodiche o gestuali, *topic segmentation*);
- "categorizzazione": l'annotatore deve attribuire un *tag* a oggetti linguistici preidentificati (es. *PoS Tagging, Word Sense Disambiguation*).

Il paper si propone di presentare una breve rassegna critica delle metriche utilizzate in questa seconda tipologia di *task*, in particolare ponendo attenzione al calcolo dei coefficienti e alla loro interpretazione.

2 I coefficient di agreement

Adottando la notazione proposta da Artstein & Poesio (2008), ogni studio di I.A.A per i *task* di categorizzazione deve prevedere:

- un insieme di item $\{i \mid i \in I\}$;
- un insieme di categorie assegnabili agli item $\{c \mid c \in C\}$;
- un insieme di annotatori, che assegnano ciascun item ad una categoria $\{r \mid r \in R\}$.

Verrà convenzionalmente indicato con A l'*agreement* e con D il *disagreement*. Allo scopo di illustrare le modalità di calcolo dei coefficienti, è stato creato *ad hoc* un esempio fittizio: la situazione immaginata prevede che due annotatori assegnino 20 item a 3 categorie.

		rater 1			
		*c*1	*c*2	*c*3	tot
rater 2	*c*1	9	2	0	11
	*c*2	0	6	0	6
	*c*3	1	0	2	3
	tot	10	8	2	20

Tab. 1: Esempio di tabella di contingenza

2.1 *Agreement* senza correzione del caso

L'indice più rudimentale è quello percentuale, detto anche **"Index of crude agreement"** (Goodman & Kruskal, 1954) o **"Observed Agreement" (A_o)**: la misura corrisponde, banalmente, al rapporto tra il numero di item su cui i *rater* sono d'accordo ed il numero totale di item. Nell'esempio proposto in tab.1, A_o ha un valore di 0.85.

La misura non solo non tiene in considerazione il ruolo che potrebbe giocare il caso, per cui i *rater* potrebbero trovarsi in accordo "tirando ad indovinare", ma deve fare i conti con un fenomeno già notato in Scott (1955) e Artstein & Poesio (2008): dati due diversi schemi di codifica per lo stesso task, quello con il minor numero di categorie registrerebbe una più alta percentuale di I.A.A. Il valore è fortemente influenzato anche dal problema della "prevalenza", ovvero la maggior concentrazione di item in una delle categorie: come avremo modo di discutere in § 2.2.1, una simile distribuzione influenza in negativo la possibilità di raggiungere alti livelli di I.A.A., indipendentemente dalla grandezza del campione.

2.2 Misure "kappa"

Il livello di I.A.A. nell'espressione di giudizi categoriali deve perciò necessariamente essere esplicitato nei termini di eccedenza rispetto all'accordo ottenibile casualmente, pena la mancanza di effettiva informatività. In ambito psicometrico sono stati introdotti numerosi coefficienti statistici in grado di correggere tale aspetto: questi indici, a cui si farà riferimento con il nome di "**misure kappa**", si fondano su tre assunti (Soeken & Prescott, 1986):

- gli *item* soggetti a valutazione sono indipendenti l'uno dall'altro;
- i *rater* che giudicano gli item operano in autonomia ed in modo completamente indipendente;
- le categorie usate sono mutualmente esclusive ed esaustive.

2.2.1 2 rater

Il caso base è rappresentato dai coefficienti per la valutazione dei giudizi prodotti da due soli *rater*, indice noto ai più come "**k di Cohen**". Prima di passare alla presentazione della misura è però necessaria una piccola premessa terminologica. Il celebre articolo di Carletta (1996), a cui va il merito di aver stabilito la valutazione dell'*agreement* come standard *de facto* in LC, ha introdotto una piccola inconsistenza in letteratura (Artstein & Poesio, 2008): la studiosa, nel suggerire l'utilizzo di un coefficiente definito "kappa", fa infatti riferimento non all'originale k proposta in Cohen (1960), ma ad una misura molto simile, introdotta cinque anni prima da Scott. La questione non si esaurisce in un mero problema terminologico: esistono infatti tre indici che, pur condividendo la medesima formula, sono fondati su ipotesi diverse riguardo la distribuzione degli item nelle categorie, ovvero **S di Bennett *et al.***, **π di Scott** e **k di Cohen**. Le differenti ipotesi soggiacenti comportano diverse modalità di calcolo e quindi risultati non coincidenti, seppure in misura minima. La formula di base è la seguente:

1) $S, \pi, k = \frac{A_0 - A_e}{1 - A_e}$

dove A_e è l'*agreement* dovuto al caso ("*Expected Agreement by chance*"); $A_o - A_e$ stima perciò l'*agreement* effettivamente raggiunto al di sopra della soglia della casualità, mentre $1 - A_e$ misura quanto accordo eccedente il caso è ottenibile. Mentre A_o è estremamente semplice da calcolare (§ 2.1) e ha lo stesso valore nelle tre misure, A_e richiede invece un modello del comportamento degli annotatori. Tutti i coefficienti assumono l'indipendenza dei due annotatori che valutano gli item: la probabilità che due *rater* (r_1 ed r_2) siano d'accordo su una determinata categoria *c* è dunque data dal prodotto della probabilità che ciascun *rater* assegni un item a quella categoria, ovvero:

2) $P(c|r1) \cdot P(c|r2)$

A_e è dato dalla sommatoria di tale probabilità congiunta per tutte le categorie dello schema di codifica.

3) $A_e^S = A_e^\pi = A_e^k = \sum_{c \in C} P(c|r_1) \cdot P(c|r_2)$

La differenza tra S, π e k risiede negli assunti che sono alla base del calcolo di $P(c|r_i)$.

S (Bennett *et al.*, 1954) assume che un'annotazione totalmente casuale determini una distribuzione uniforme degli item nelle categorie, ovvero che tutte le categorie dello schema di codifica siano ugualmente probabili; la probabilità che ogni *rater* assegni un item alla categoria *c* è dunque $1/c$.

4) $A_e^S = \sum_{c \in C} \frac{1}{c} \cdot \frac{1}{c} = c \cdot \left(\frac{1}{c}\right)^2 = \frac{1}{c}$

Nell'esempio di tab.1 A_e^S=0.333 e S=0.775.

L'assunto dell'uniformità è un prerequisito estremamente vincolante: per tale ragione non risultano, ad oggi, studi di I.A.A. in LC in cui sia stato impiegato questo coefficiente. In aggiunta, come è stato notato da Scott (1955: 322-323) e riportato da Artstein & Poesio (2008: 561), il valore dell'indice può essere aumentato semplicemente inserendo nello schema di codifica categorie vuote.

Il coefficiente π (Scott, 1955), noto anche col nome di K di Siegel & Castellan (1988), assume che se l'attribuzione degli item alle categorie avviene in modo casuale, la distribuzione sarà uguale per entrambi gli annotatori. $P(c|r_i)$ corrisponderà perciò al rapporto tra il numero totale di assegnazioni alla categoria *c* da parte di entrambi i rater , n_c , e il numero totale di assegnazioni compiute, $2i$.

5) $A_e^\pi = \sum_{c \in C} \left(\frac{n_c}{2i}\right)^2$

Nel caso in oggetto, A_e^π= 0.414 e π=0.744.

k (Cohen, 1960) prevede infine una distribuzione degli item nelle categorie distinta ed unica per ciascun annotatore, rappresentata nelle frequenze marginali della tabella di contingenza.

6) $P(c|r_i) = \frac{n_{r_i c}}{i}$

7) $A_e^k = \sum_{c \in C} \frac{n_{r_1 c}}{i} \cdot \frac{n_{r_2 c}}{i}$

Nell'esempio oggetto di discussione, pertanto, A_e^k=0.41 e k=0.764.

La corretta scelta dell'indice non può prescindere dalla considerazione che i coefficienti sono fortemente influenzati da disomogeneità nella distribuzione dei dati (Feinstein & Cicchetti, 1990; Cicchetti & Feinstein 1990; Di Eugenio & Glass, 2004; Artstein & Poesio, 2008), classificabili in due tipologie principali: la già ricordata "prevalenza" (tab. 2) e il "*bias*", cioè il grado con cui gli annotatori sono in accordo/disaccordo nelle loro valutazioni complessive, ossia le loro "tendenze" nell'esprimere giudizi (tab. 3 e 4).

		rater 1			
		*c*1	*c*2	*c*3	tot
rater 2	*c*1	18	0	1	19
	*c*2	0	0	0	0
	*c*3	1	0	0	1
	tot	19	0	1	20

Tab. 2: Distribuzione affetta da prevalenza.

		rater 1			
		*c*1	*c*2	*c*3	tot
rater 2	*c*1	4	1	1	6
	*c*2	1	3	3	7
	*c*3	1	2	4	7
	tot	6	6	8	20

Tab. 3: Distribuzioni marginali simili.

		rater 1			
		*c*1	*c*2	*c*3	tot
rater 2	*c*1	4	3	1	8
	*c*2	0	3	0	3
	*c*3	1	4	4	9
	tot	5	10	5	20

Tab. 4: Esempio di *bias*, evidente dalle distribuzioni marginali dissimili ("*skewed*").

Nell'esempio di tab. 2, la forte prevalenza in favore della categoria *c1* fa sì che $A_e^{\pi} = A_e^{k} = 0.905$. Di conseguenza, nonostante A_o sia molto alto (0.9), $\pi = k = -0.053$, al di sotto della soglia della pura casualità.

Si confrontino quindi i dati delle tabelle 3 e 4: sebbene entrambe registrino un A_o di 0.55, nel caso in cui le distribuzioni marginali siano molto simili (tab.3) $A_e^{\pi} = 0.335$, $A_e^{k} = 0.336$, $\pi = 0.322$, $k = 0.323$; l'effetto di *bias* (tab.4), invece, affligge la k di Cohen, in ragione delle modalità di calcolo di $P(c|r_1)$: $A_e^{\pi} = 0.334$, $A_e^{k} = 0.287$, $\pi = 0.326$, $k = 0.368$. La differenza tra π e k è empiricamente minima: $A_e^{\pi} \geq A_e^{k}$, perciò $\pi \leq k$. I due coefficienti assumono lo stesso valore nel caso (limite) in cui le distribuzioni marginali dei due *rater* siano identiche, come in tab. 2.

A fronte di ciò, laddove non sia possibile effettuare uno studio che coinvolga più di due *rater*, sembrerebbe pertanto da preferire il coefficiente π di Scott, in grado di generalizzare il comportamento dei singoli annotatori. In letteratura sono state fatte varie proposte riguardo la modalità di presentazione dei risultati dell'I.A.A per due annotatori: allo stato dell'arte sembrerebbe preferibile adottare la soluzione suggerita da Byrt *et al.* (1993) e adottata da Di Eugenio & Glass (2004), ovvero presentare congiuntamente diversi coefficienti:

- k, che in linea di principio meglio si adatta alla valutazione di annotazioni che coinvolgono dati linguistici, e rende conto di eventuali tendenze dei *rater*;
- π, immune all'effetto di *bias*;
- una terza misura, $2A_o$-1, in grado di neutralizzare l'effetto di prevalenza (Byrt *et al.*, 1993).

2.2.2 Possibili estensioni

Sono state proposte moltissime generalizzazioni dei coefficienti presentati, per assicurare maggiore flessibilità ed adattabilità agli specifici task:[1] tra le più note vi è la "***weighted kappa***" (Cohen, 1968), $\mathbf{k_{(w)}}$, indice che consente di esprimere delle gradazioni di disaccordo mediante una tabella di "pesi" di valore compresi tra 0 e 1 ("*weighting scheme*"), come nell'esempio:

	*c*1	*c*2	*c*3
*c*1	1	0	0.5
*c*2	0	1	0.5
*c*3	0.5	0.5	1

Tab.4: Esempio di *weighting scheme*

$A_{o(w)}$ e $A_{e(w)}$ vengono calcolati in modo affine alla k di Cohen (1960), moltiplicando però, in aggiunta, ogni cella della tabella di contingenza per il corrispettivo peso.

$$8) \quad k_{(w)} = \frac{A_{0(w)} - A_{e(w)}}{1 - A_{e(w)}}$$

Se applicata ai dati di Tab.1, $k_{(w)} = 0.774$.

Sono stati inoltre introdotti indici in grado di quantificare l'I.A.A. tra tre o più annotatori: in primis la cosiddetta k di Fleiss (1971), che estende l'indice π di Scott ("***multi-π***"), ed il coefficiente presentato in Davies & Fleiss (1982) che generalizza la k di Cohen ("***multi-k***");[2] ma soprattutto il coefficiente **α di Krippendorff** (1980), che esprime l'I.A.A. in termini di *disagreement*, osservato (D_o) e dovuto al caso (D_e):

$$9) \qquad \alpha = 1 - \frac{D_o}{D_e}$$

La formula, pur essendo stata derivata dalla misura della varianza, non fa esplicito riferimento alle medie dei campioni e può pertanto essere generalizzata ad una moltitudine di schemi di codifica in cui le categorie non siano interpretabili come valori numerici; come per la *weighted kappa* si possono inoltre attribuire pesi alle di-

[1] Alcune estensioni delle misure "kappa", troppo complesse per essere descritte esaurientemente in questa sede, consentono ad esempio di valutare l'I.A.A nel caso in cui i *rater* effettuino osservazioni multiple, e non necessariamente di ugual numero, oppure di gestire gli schemi di annotazione che prevedono la possibilità di attribuire più di una classificazione agli item (Kraemer, 1980).

[2] Le modalità di calcolo sono affini ai coefficienti già descritti. Per i dettagli si rinvia perciò a Fleiss (1971), Davies & Fleiss (1982) e all'ottima sintesi di Artstein & Poesio (2008) e Artstein (2017). Si noti che A_o non potrà essere definito come "percentuale di item su cui c'è accordo", visto che con altissima probabilità ci saranno nei dati item su cui alcuni *rater* saranno d'accordo e altri no: la soluzione proposta in letteratura a partire da Fleiss (1971) è di misurare l'I.A.A. "*pairwise*", ovvero "a coppie".

verse tipologie di *disagreement*, utilizzando *weighting scheme* oppure introducendo nel calcolo delle metriche, ad esempio l'indice statistico MASI (Passonneau, 2006; Dorr *et al.*, 2010).[3] α è equivalente a multi-π per campioni numerosi, ma è in grado, non imponendo un numero minimo di *item*, di mitigare gli effetti statistici di *dataset* a bassa numerosità campionaria; inoltre, consentendo la gestione di *dataset* incompleti, è utilizzabile (o addirittura preferibile) nel caso in cui l'annotazione si svolga in maniera collaborativa e distribuita, ad esempio su piattaforme di *crowdsourcing*.

3 *Reliability*: *agreement* o correlazione?

In letteratura, in particolare in ambito clinico (Bishop & Baird, 2001; Van Noord & Prevatt, 2002; Massa *et al.*, 2008; Gudmundsson & Gretarsson, 2009), non è infrequente che, nella stima dell'I.A.A, vengano preferiti o affiancati alle misure presentate la statistica χ^2 oppure gli indici statistici di correlazione (coefficiente R di Pearson *in primis*, ma anche i non parametrici ρ di Spearman e τ di Kendall).
Come già notato da Cohen (1960), l'utilizzo del χ^2 è una prassi da considerarsi scorretta, poiché la statistica, applicata alla tavola di contingenza, misura casualità e grado di associazione tra i set di giudizi, non l'*agreement* (Banerjee *et al.*, 1999).

> *"[...] Many investigators have computed χ^2 over the table for use as a test of the hypothesis of chance agreement, and some have gone on to compute the contingency coefficient (C) as a measure of degree of agreement. [...] It is readily demonstrable that the use of χ^2 (and therefore the C which is based on it) for the evaluation of agreement is indefensible. When applied to a contingency table, χ^2 tests the null hypothesis with regard to association, not agreement.*
>
> (Cohen, 1960: 38)

Altrettanto scorretta dal punto di vista metodologico è l'applicazione di coefficienti di correlazione inter-/intra- classe, che ugualmente non quantificano l'I.A.A. ma la forza di associazione tra gruppi di valori (Bland & Altman, 1986; Kottner *et al.*, 2011; Stolarova *et al.*, 2014). Si noti inoltre che, dal punto di vista empirico, un'ottima correlazione tra annotazioni può essere raggiunta anche in caso di completa mancanza di accordo, se due set di giudizi differiscono sistematicamente.
La ragione di tali fraintendimenti deve probabilmente essere rintracciata nell'uso sostanzialmente sinonimico dei termini "*reliability*" e "*agreement*" (Stemler, 2004); come puntualizzato da Krippendorff (2004), in realtà:

> *"To be clear, agreement is what we measure; reliability is what we wish to infer from it."*
>
> (Krippendorff, 2004: 413)

Le correlazioni statistiche possono senza dubbio costituire un'informazione interessante nella valutazione globale dell'affidabilità di un *dataset*, a patto però che tale nozione sia tenuta distinta dall'I.A.A. in senso stretto.

4 La valutazione dei coefficienti

La valutazione dei valori assunti dai coefficienti *chance-corrected* rappresenta, ad oggi, un aspetto critico: gli indici possono assumere valori compresi tra -1 e 1, dove k = 1 corrisponde ad un I.A.A. perfetto, k = 0 ad un I.A.A. completamente casuale e k = -1 ad un perfetto disaccordo. Non è però soddisfacente sapere che k abbia un valore superiore alla totale casualità, ma occorre assicurarsi, piuttosto, che gli annotatori non si discostino troppo dall'*agreement* assoluto (Cohen, 1960; Krippendorff, 1980).
A prescindere dal mero valore numerico, va rilevato come i vari studiosi che hanno tentato di indicare delle soglie di riferimento abbiano sottolineato l'arbitrarietà delle loro proposte: in primis Landis & Koch (1977), a cui si deve la più nota griglia per l'interpretazione dei coefficienti:

Kappa Statistic	Strength of Agreement
< 0.0	Poor
0.00 - 0.20	Slight
0.21 - 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81- 1.00	Almost Perfect

Tab. 5: Griglia per l'interpretazione delle misure k (Landis & Koch, 1977).

Così anche Krippendorff, la cui proposta di rifiutare valori di k inferiori a 0.67, accettare quelli superiori a 0.8 e considerare incerti quelli compresi nel *range* costituisce uno dei principali punti di riferimento in letteratura sull'argomento.

> *"Except for perfect agreement, there are no magical numbers, however."*
>
> (Krippendorff, 2004: 324)

Va infine rilevato come il *disagreement* non sia necessariamente indice di bassa qualità

[3] MASI è basato sul coefficiente di Jaccard (1908) e quindi stabilisce la somiglianza/diversità tra insiemi campionari in termini di distanza.

dell'annotazione, scarso *training* degli annotatori o di *guideline* mal definite (Aroyo & Welty, 2015), soprattutto nei task di natura semantica; ed anche che, per aumentare l'affidabilità del *dataset* annotato, non debba necessariamente essere evitato o eliminato: in LC la sua presenza può infatti essere sfruttata esplicitamente, per migliorare le performance di sistemi automatici (come ad esempio in Chklovski & Mihalcea, 2003; Plank, Hovy & Søgaard, 2014).

5 Conclusioni

Come suggerito nei paragrafi iniziali, un alto livello di I.A.A. non costituisce un risultato in sé, ma soltanto uno fra gli indicatori della reale affidabilità dell'annotazione sottoposta a validazione. È perciò auspicabile che un sempre maggior numero di dati sull'I.A.A. nei diversi task di annotazione sia condiviso dai ricercatori, in modo da facilitare l'emergere per confronto dei valori di riferimento.

Bibliografia

Aroyo, L. & Welty, C. (2015). Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine,* 36 (1):15–24.

Artstein, R. (2017). Inter-annotator Agreement. In: Ide, N. & Pustejovsky, J. (eds.), *Handbook of Linguistic Annotation*. Springer, Dordrecht, pp. 297–314.

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

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.

Banerjee, M., Capozzoli, M., McSweeney, L. & Sinha, D. (1999). Beyond Kappa: A Review of Interrater Agreement Measures. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 27(1):3–23.

Bennett, Alpert, R. & Goldstein, A. (1954). Communications through limited response questioning. *Public Opinion Quarterly*, 18:303–308.

Bishop, D.V. & Baird, G. (2001). Parent and teacher report of pragmatic aspects of communication: use of the children's communication checklist in a clinical setting. *Developmental medicine and child neurology*, 43:809–818.

Bland, M. J. & Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 327:307–310.

Byrt, T., Bishop, J. & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46(5):423–9.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

Chklovski, T. & Mihalcea, R. (2003). Exploiting Agreement and Disagreement of Human Annotators for Word Sense Disambiguation. In: *Proceedings of the Conference on Recent Advances in Natural Language Processing* (RANLP 2003).

Cicchetti, D.V. & Feinstein, A.R. (1990). High Agreement but low Kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 43:551–558.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

Davies, M. & Fleiss, J. L. (1982). Measuring Agreement for Multinomial Data. *Biometrics*, 38(4):1047–1051.

Di Eugenio, B. (2000). On the usage of Kappa to evaluate agreement on coding tasks. In: Calzolari N. *et al.* (eds): *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, ELRA - European Language Resources Association, Paris, pp. 441–444.

Di Eugenio, B. & Glass, M. (2004). The Kappa statistic: a second look. *Computational Linguistics*, 30(1):95–101.

Dorr, B.J., Passonneau, R.J., Farwell, D., Green, R., Habash, N., Helmreich, S., Hovy, E., Levin, L., Miller, K. J., Mitamura, T., Rambow, O. & Siddharthan, A. (2010). Interlingual annotation of parallel text corpora: A new framework for annotation and evaluation. *Journal of Natural Language Engineering*, 16(3):197–243.

Feinstein, A.R. & Cicchetti, D.V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43:543–549.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Goodman, L. A. & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of*

the American Statistical Association, 49(268): 732–764.

Gudmundsson, E. & Gretarsson, S. J. (2009). Comparison of mothers' and fathers' ratings of their children's verbal and motor development. *Nordic Psycholgy*, 61:14–25.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 44:223–270.

Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B.J., Hróbjartsson, A., Roberts, C., Shoukri, M. & Streiner, D.L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *International journal of nursing studies*, 48:661–671.

Kraemer, H. C. (1980). Extension of the kappa coefficient. *Biometrics*, 36(2):207–16.

Krippendorff, K. (1980). *Content Analysis: an introduction to its Methodology*. Sage Publications, Thousand Oaks, CA.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3):411–433.

Landis, J. R. & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Massa, J., Gomes, H., Tartter, V., Wolfson, V. & Halperin, J.M. (2008). Concordance rates between parent and teacher clinical evaluation of language fundamentals observational rating scale. *International Journal of Language & Communication Disorders*, 43:99–110.

Mathet, J., Widlöcher, A. & Métivier, J. (2015). The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment. *Computational Linguistics*, 41(3):437–479.

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

Plank, B., Hovy, D. & Søgaard, A. (2014). Learning part-of-speech taggers with inter-annotator agreement loss. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 742–751.

Scott, W.A. (1955). Reliability of Content Analysis: The Case of Nominal Scale Coding. *Public Opinion Quarterly*, 19(3):321–325.

Siegel, S. & Castellan, J. (1988). *Nonparametric statistics for the behavioral sciences*. McGraw-Hill, Boston, MA.

Soeken, K.L. & Prescott, P.A. (1986). Issues in the use of kappa to estimate reliability. *Medical Care*, 24(8):733–41.

Stemler, S.E. (2004). A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability. *Practical Assessment, Research & Evaluation*, 9:66–78.

Stolarova, M., Wolf, C., Rinker, T. & Brielmann, A. (2014). How to assess and compare inter-rater reliability, agreement and correlation of ratings: an exemplary analysis of mother-father and parent-teacher expressive vocabulary rating pairs. *Frontiers in psychology*, 5, 509.

Van Noord, R.G. & Prevatt, F.F. (2002). Rater agreement on IQ and achievement tests: effect on evaluations of learning disabilities. *Journal of School Psychology,* 40(2):167–176.

Warrens, M. J. (2010). Inequalities between multi-rater kappas. *Advances in Data Analysis and Classification*, 4(4):271–286.