AgreeCalc: Uma Ferramenta para Análise da Concordância entre Múltiplos Anotadores

Alexandre Rossi Alvares[1], **Norton Trevisan Roman**[1]

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (USP)
São Paulo, SP – Brazil

{alexandre.alvares,norton}@usp.br

Abstract. *In this paper we describe the prototype of a graphical tool – AgreeCalc – designed for the automatic calculation of agreement between multiple annotators, with no need for advanced knowledge by the user on the implemented coefficients of agreement. Although being attractive for the implemented coefficients, AgreeCalc also allows for (i) the application of filters on the data used to calculate them; (ii) the identification of the annotator pairs with the highest and lowest agreement (along with their mean); and (iii) the construction of a gold standard, with the opinion of the majority of annotators.*

Resumo. *Neste artigo descrevemos o protótipo de uma ferramenta gráfica – AgreeCalc – projetada para o cálculo automático do grau de concordância entre múltiplos anotadores, sem a necessidade de conhecimentos avançados do usuário sobre os coeficientes de concordância nele implementados. Ainda que seu maior atrativo seja o conjunto de coeficientes implementados, o AgreeCalc também permite (i) a aplicação de filtros aos dados usados para o cálculo desses coeficientes; (ii) a identificação do par de anotadores com maior e menor concordância (além da média); e (iii) a criação de um* gold standard*, com a opinião da maioria dos anotadores.*

1. Introdução

Nos últimos tempos, boa parte da pesquisa realizada em Linguística Computacional tem sido feita através do uso de *corpora*, com a finalidade de, por exemplo, observar e propor hipóteses, otimizá-las e por fim avaliá-las [Mitkov et al. 1999], ou mesmo aplicar alguma técnica de aprendizado de máquina para o modelamento de fenômenos linguísticos [Pustejovsky and Stubbs 2012]. Tais *corpora*, contudo, para que sejam úteis, devem receber um tratamento inicial, na forma de metadados que forneçam alguma informação sobre o fenômeno de interesse, ou seja, devem ser anotados [Pustejovsky and Stubbs 2012]. Por tratar-se de um processo de interpretação por parte de seus executores [Grouin et al. 2011], tais anotações invariavelmente apresentam diferenças, diferenças estas que devem ser quantificadas, para que seu efeito possa então ser avaliado.

Uma forma de quantificar as diferenças encontradas é medir o grau de concordância entre os anotadores, ou seja, a medida na qual anotadores diferentes atribuem o mesmo rótulo à mesma porção do texto anotado, determinando assim a consistência da anotação entre eles [Grouin et al. 2011]. A ideia por trás dessa medida é a de que nossa

Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, pages 1–10,
Fortaleza, CE, Brazil, October 21–23, 2013.

confiança na qualidade dos dados aumenta quanto maior for o número de pessoas concordando com sua classificação [Craggs and Wood 2005]. Mais do que isso, uma alta concordância também permite inferir que os anotadores de fato obtiveram um entendimento comum sobre as diretrizes do esquema usado para a anotação [Artstein and Poesio 2008]. Assim, essa concordância, usada como medida da confiabilidade tanto do esquema de anotação quanto dos dados produzidos, torna-se um pré-requisito para a demonstração da validade desse esquema [Artstein and Poesio 2008].

Considerando a importância atribuída à concordância entre anotadores, não é de se espantar que haja um número razoável de maneiras diferentes de quantificá-la, indo desde cálculos mais simples, como a fração das anotações em que os anotadores concordam (*e.g.* [Petasis 2012]), até coeficientes mais elaborados, como o κ de Cohen [Cohen 1960] e o α de Krippendorff [Krippendorff 2004]. De fato, ainda hoje há um grande debate acerca de quão apropriados esses índices são (*e.g.* [Geertzen and Bunt 2006, Artstein and Poesio 2008, Bayerl and Paul 2011]), com experimentos sendo feitos com os mais diferentes índices (*e.g.* [Grouin et al. 2011, Mathet et al. 2012, Fort et al. 2012]), e padrões defendidos para serem então refutados (*e.g.* [Carletta 1996, Eugenio and Glass 2004]), sendo que um consenso ainda parece estar longe de ser obtido.

Essa indefinição no que tange à adoção de um padrão, por sua vez, apresenta a indesejável consequência de que pesquisas diferentes reportam resultados usando índices diferentes, dificultando assim a comparação entre elas. Para se ter uma ideia da magnitude desse problema, em uma pesquisa compreendendo publicações em apenas três domínios da linguística computacional (transcrições prosódicas, fonéticas e desambiguação por extração do sentido da palavra[1]), foi encontrado que 56,1% dos índices reportados tratavam-se de porcentagem de concordância, 39,3% referiam-se ao κ (em alguma de suas várias versões), enquanto que 4,6% correspondiam a outros índices, como o α de Krippendorff, precisão, cobertura e medida-F[2], dentre outros [Bayerl and Paul 2011]. Além disso, foram também contabilizados um total de 972 índices de concordância nos 326 estudos analisados, levando a uma média de quase três índices diferentes reportados em cada estudo.

De fato, o relato de mais de um índice de concordância parece ser a estratégia atualmente adotada por muitos pesquisadores na área da Linguística Computacional (*e.g.* [Petasis 2012, Fort et al. 2012]). Naturalmente, isso implica o cálculo desses índices – algo nem sempre fácil de executar. Ainda que alguns deles estejam disponíveis em muitas ferramentas para anotação de *corpora* de uso geral (*e.g.* [Apostolova et al. 2010, Verhagen 2010, Petasis 2012]), pesquisadores na área encontram dificuldades em que (i) não necessariamente têm suas necessidades satisfeitas por uma dessas ferramentas, devendo assim construir seu próprio sistema e, consequentemente, seu próprio módulo para cálculo dos índices de concordância desejados; e (ii) a maioria das ferramentas existentes parece fornecer não mais que dois índices (*e.g.* [Müller and Strube 2006, Petasis 2012, Bontcheva et al. 2013]), reduzindo assim a chance de comparação com outros estudos, caso esses optem por índices diferentes.

Diante desse cenário, seria interessante haver uma ferramenta dedicada ao cálculo

[1] Word-sense disambiguation.

[2] Precision, recall e F-measure.

dos índices de concordância mais comumente usados, a partir de anotações fornecidas a ela. Dessa forma, bastaria que as anotações obedecessem ao formato aceito pela ferramenta para o pesquisador obter os resultados da concordância entre anotadores, conforme os índices de sua escolha. Neste artigo descrevemos o protótipo de uma ferramenta gráfica – AgreeCalc – projetada para esse fim, ou seja, dar ao pesquisador acesso aos índices de concordância mais comuns. Mais do que permitir ao pesquisador a escolha do índice mais apropriado, dentre uma lista de quatro até agora implementados, a ferramenta também permite o cálculo desses índices usando subconjuntos dos dados fornecidos, dando assim uma maior liberdade de exploração desses dados. Distribuída sob a GPL, essa ferramenta está disponível no endereço http://www.each.usp.br/norton/resdial/.

O restante desse trabalho está organizado como segue. A Seção 2 descreve algumas das atuais ferramentas que fornecem módulos para cálculo de coeficientes de concordância, comparando-as ao AgreeCalc. Na Seção 3 o sistema é descrito em maiores detalhes, sendo os resultados discutidos na Seção 4. Por fim, na Seção 5 são apresentadas as conclusões a esse trabalho, bem como futuras melhorias a serem implementadas no sistema.

2. Trabalhos Relacionados

Ferramentas ligadas à anotação de *corpora* são relativamente comuns no meio da Linguística Computacional, especialmente aquelas projetadas para uso geral, em que o pesquisador define seu próprio esquema de anotação, deixando a cargo da ferramenta a elaboração de uma interface gráfica que implemente esse esquema (*e.g.* [Müller and Strube 2006, Bontcheva et al. 2013]). Nesse cenário, cabe ao anotador final, por sua vez, o uso dessa interface para fazer a marcação do texto-alvo. Dadas essas características, é de se esperar que algumas dessas ferramentas apresentem algum módulo de suporte à análise da concordância entre múltiplos anotadores. Estas, contudo, ainda que bastante úteis, em geral apresentam de um a dois índices diferentes, limitando assim a escolha de seus usuários.

A questão, no entanto, não é somente a quantidade de índices apresentados, mas também sua escolha. Por exemplo, no MMAX2 [Müller and Strube 2006] é usado tanto o κ (*cf.* [Cohen 1960]) quanto uma medida estatística definida em [Vilain et al. 1995]. GATE Teamware [Bontcheva et al. 2013], por outro lado, fornece valores tanto para o κ quanto para a medida-F (*cf.* [Hripcsak and Rothschild 2005]). Implementado como um componente da plataforma Ellogon [Petasis et al. 2002], outro sistema disponível – SYNC3 [Petasis 2012] – herda as funcionalidades dessa plataforma, inclusive seu módulo para cálculo da concordância entre anotadores. Nesse caso, os índices fornecidos pelo SYNC3 são κ e porcentagem de concordância. Por fim, há também o Djangology [Apostolova et al. 2010], que apresenta os valores de precisão, cobertura e medida-F.

Essa disjunção no conjunto de índices fornecidos pelas ferramentas, sem mencionar a inexistência de índices cuja adoção foi aconselhada (*cf.* [Artstein and Poesio 2008]), como o α de Krippendorff [Krippendorff 2004], por exemplo, dificulta em muito a comparação dos resultados de diferentes pesquisas. Nesse ponto, o sistema aqui descrito busca amenizar esse problema apresentando quatro dos principais índices, a saber, κ de Cohen, κ de Fleiss [Davies and Fleiss 1982], α de Krippendorff e porcentagem de concordância (a serem tratados separadamente na

Seção 3). Ao fazer isso, o AgreeCalc permite que pesquisas possam ser comparadas aos resultados gerados pela maioria das ferramentas disponíveis, ainda que não cubra precisão, cobertura e medida-F, ou seja, as medidas de concordância emprestadas do campo da recuperação de informação.

Por fim, na ausência do índice pretendido nas ferramentas disponíveis, ou mesmo na impossibilidade do uso dessas ferramentas, resta ao pesquisador o uso de pacotes estatísticos, como o R[3], que contém funções implementando esses coeficientes. O uso desse tipo de sistema, contudo, implica não somente a formatação dos dados (*i.e.* das anotações por múltiplos anotadores) de forma aceitável pelo R, mas também o conhecimento, por parte do pesquisador, de como cada uma dessas funções deve ser usada. Conforme será visto no que segue, o AgreeCalc aproveita essas definições, usando o R como seu motor para cálculo de concordância, unindo assim o poder que esse pacote fornece, com a simplicidade de uma interface gráfica em que tudo que o pesquisador deve fazer é escolher qual índice usar.

3. Descrição do Sistema

Tomando como entrada anotações feitas em um *corpus*, codificadas como descrito em [Roman 2012], o AgreeCalc apresenta ao pesquisador uma interface gráfica para avaliação de esquemas de anotação, sem exigir deste um conhecimento profundo acerca dos cálculos envolvidos. Para tal, e usando do fato de ter sido desenvolvido em Java, o sistema faz uso das bibliotecas *irr*[4] e *RCaller*[5] para comunicação com o pacote estatístico R, pacote este usado para os cálculos de concordância. Dessa forma, o AgreeCalc une as funcionalidades presentes no R a uma interface em que os detalhes do uso de cada coeficiente tornam-se transparentes ao pesquisador.

Tão logo inicia a ferramenta, o pesquisador visualiza uma tela, em que deve indicar um diretório contendo as demarcações feitas pelos anotadores, codificadas conforme o padrão aceito. Após realizada a validação dos arquivos ali contidos, o AgreeCalc exibe sua tela principal, em forma de abas (Figura 1)[6]. Na primeira aba, o pesquisador tem acesso a uma breve descrição do *corpus* selecionado, contendo o nome do esquema de anotação, sua localização em disco, e o nome do *corpus* de origem, sobre o qual o esquema foi aplicado. Além disso, também são apresentados o total de unidades e documentos presentes no *corpus*, bem como o número de anotadores.

Na segunda aba, por sua vez, o pesquisador pode determinar quais anotadores farão parte do cálculo de concordância (Figura 2). Para tal, uma lista dos anotadores disponíveis é mostrada à direita, enquanto que à esquerda situa-se a lista daqueles excluídos pelo pesquisador. Dessa forma, é possível ao pesquisador limitar seu campo de avaliação. Ao clicar em *Aplicar Mudanças*, o AgreeCalc filtrará os documentos a serem usados para os cálculos de concordância levando em conta apenas os anotadores selecionados.

Na sequência, na aba *Documentos* (Figura 3) o pesquisador pode refinar mais os dados, selecionando os documentos que serão levados em conta quando do cálculo da

[3]http://www.r-project.org/

[4]http://cran.r-project.org/web/packages/irr/irr.pdf

[5]https://code.google.com/p/rcaller/

[6]As figuras correspondem a telas do programa, editadas por questões de espaço.

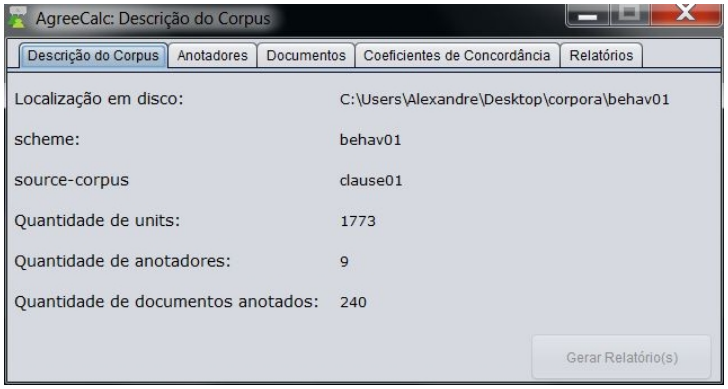

Figura 1. Tela principal da ferramenta.

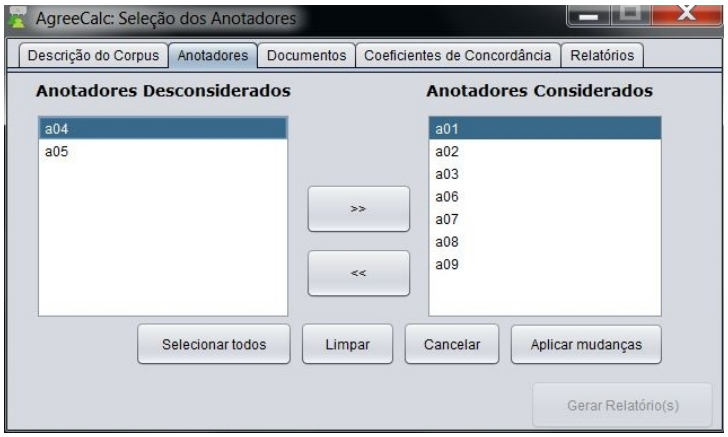

Figura 2. Definição do sub-conjunto de anotadores.

concordância (por padrão, todos os documentos já estão selecionados). Também aqui, ao clicar em *Aplicar Mudanças*, o AgreeCalc irá filtrar novamente os dados para o cálculo dos coeficientes de concordância, levando em conta tanto o subconjunto de dados quanto o de anotadores selecionados. Vale salientar, contudo, que um documento refere-se, de fato, a um texto completo, e não a uma unidade de anotação (*i.e.* a unidade mínima à qual é aplicada uma anotação). Em sua versão atual, o AgreeCalc permite a exclusão apenas de documentos inteiros, caso em que estes são retirados do cálculo para todos os anotadores envolvidos (selecionados na etapa anterior, conforme ilustrada na Figura 2).

A aba seguinte, por sua vez, apresenta ao pesquisador a lista dos coeficientes de concordância implementados (Figura 4). Nesse ponto, vale frisar que, a depender do número de anotadores selecionados para o estudo, alguns dos coeficientes podem não estar disponíveis (por limitações nos próprios coeficientes), como é o caso do κ de Cohen com mais de dois anotadores (ilustrado na figura). Outra funcionalidade útil da ferramenta é o cálculo da concordância entre pares de anotadores. Caso escolha essa opção, o sistema irá calcular o valor dos coeficientes selecionados para toda permutação dos anotadores escolhidos, tomados de dois em dois. Como resultado, a ferramenta apresentará, para cada índice, o par com a menor e a maior concordância, dando também a concordância média entre os pares.

No que diz respeito aos coeficientes implementados, a atual versão do AgreeCalc

Figura 3. Definição do sub-conjunto de documentos.

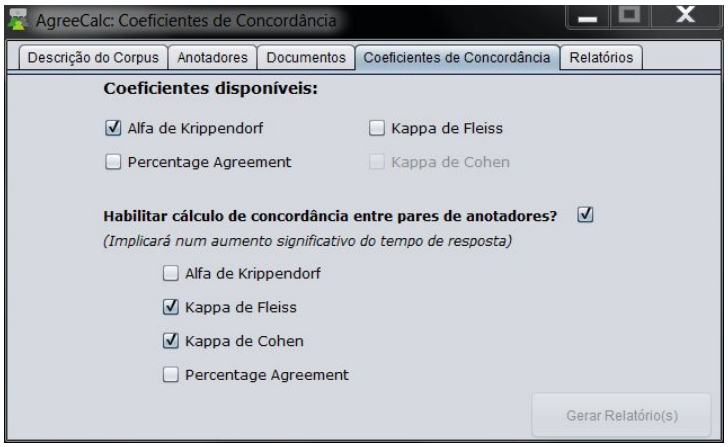

Figura 4. Definição dos coeficientes de concordância.

contempla os seguintes[7]:

- α *de Krippendorf*: Sem restrição quanto ao número de anotadores, e corrigindo os cálculos em relação ao esperado aleatoriamente, o α varia de 0 a 1, em que 0 indica a ausência de confiabilidade e 1 concordância perfeita [Krippendorff 2004] (valores negativos também são possíveis, devido a erros de amostragem ou discordância sistemática). Nesse sistema, o α é usado em sua versão nominal.
- κ *de Cohen*: Limitado a apenas dois anotadores [Cohen 1960], e corrigindo os cálculos em relação ao esperado aleatoriamente, também o κ varia de 0 a 1 (com a mesma interpretação do α).
- κ *de Fleiss*: Generalização do κ de Cohen para mais de dois anotadores [Artstein and Poesio 2008].
- *Porcentagem de Concordância*: Maneira mais simples de determinar a concordância entre anotadores, é calculada como a fração das classificações nas quais os anotadores concordam sobre a mesma unidade. Sofre, contudo, por não levar em conta o valor da concordância esperada aleatoriamente [Artstein and Poesio 2008].

[7]Para uma comparação detalhada dos índices consulte [Artstein and Poesio 2008].

Por fim, na aba *Relatórios* o usuário escolhe o formato em que os coeficientes calculados estarão apresentados, exportando o documento como HTML, XML ou PDF (Figura 5). Além disso, o pesquisador tem a possibilidade de obter dois conjuntos de dados extras, que nada mais são que documentos de anotação contendo, para cada unidade do corpus, respectivamente, a classificação mais popular entre os anotadores (moda), ou seja, a categoria mais frequentemente associada a cada unidade, independentemente de número mínimo de anotadores; ou a classificação atribuída pela sua maioria absoluta (ou seja, acima de 50%), caso em que unidades onde não se atingiu uma maioria permanecem como não anotadas. Tais anotações aparecem também codificadas como as demais, segundo o padrão definido em [Roman 2012]. Com isso, o pesquisador não somente tem os resultados numéricos dos índices de concordância, mas também um *gold standard*, construído conforme uma dessas duas métricas, que pode ser usado para pesquisas futuras.

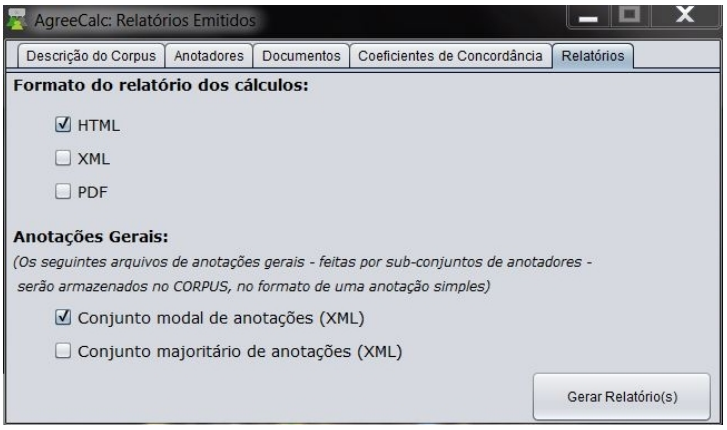

Figura 5. Definição dos relatórios a serem gerados.

4. Discussão

Em sua atual versão, o AgreeCalc permite não somente o cálculo de quatro dos mais populares coeficientes de concordância (*cf.* [Bayerl and Paul 2011]), mas também, semelhante ao sistema apresentado em [Ogren 2006], a aplicação de filtros aos dados a serem usados para esse cálculo. Esses filtros, por sua vez, são definidos mediante a seleção tanto dos anotadores quanto dos documentos que serão usados no cálculo, seleção esta feita de uma maneira bastante direta, conforme visto, através de uma interface gráfica. Assim, a ferramenta apresenta ao pesquisador um modo direto de identificar porções do *corpus* (ou mesmo subconjuntos de anotadores) em que a concordância é demasiado alta ou baixa, permitindo um estudo mais aprofundado acerca das razões para esse comportamento.

Além disso, o fato do sistema fornecer, juntamente com os valores gerais dos coeficientes, a maior e menor concordância obtida por pares, além da concordância média, permite não apenas um melhor entendimento da variabilidade de cada coeficiente, mas também a comparação com pesquisas que reportam algum desses valores (*e.g.* [Gut and Bayerl 2004]). Presente em alguns sistemas de anotação para uso geral (*e.g.* [Ogren 2006]), a construção de um *gold standard*, contendo a classificação de cada unidade pela maioria dos anotadores (ou alternativamente, a atribuição da classificação mais comum), também é de grande utilidade em casos onde o pesquisador pretenda usar esses resultados para outros fins, como o aprendizado de máquina, por exemplo.

Ainda que bastante útil em sua versão atual, vale ressaltar que um dos principais limitantes do AgreeCalc reside no fato de tratar apenas de classificações com categorias nominais, limitante este compartilhado pelos coeficientes nele implementados. Para o futuro, está em estudo a extensão da ferramenta para outras formas de classificação. Por fim, vale ressaltar que a ferramenta foi testada em um *corpus* de anotações feitas por nove anotadores, responsáveis pela classificação de 1773 unidades, contidas em 240 documentos (consulte [Roman and Carvalho 2010] para mais detalhes). Os resultados obtidos pelo sistema foram idênticos aos obtidos manualmente, aumentando assim a confiança na sua implementação.

5. Conclusão

Neste artigo foi apresentado o AgreeCalc – uma ferramenta desenvolvida para o cálculo automático do grau de concordância entre múltiplos anotadores de um *corpus*, sem a necessidade de conhecimentos avançados do usuário sobre os coeficientes nele implementados. Ainda que seu maior atrativo resida no conjunto de coeficientes apresentados, o AgreeCalc também oferece como vantagem a filtragem dos dados a serem usados para o cálculo desses coeficientes. Essa filtragem, por sua vez, se dá na forma da seleção de quais anotadores e documentos deverão fazer parte desse cálculo.

Juntamente com os valores dos coeficientes de concordância, o AgreeCalc permite o cálculo desses coeficientes par a par, possibilitando, dentre outras coisas, a identificação de comportamentos erráticos por parte dos anotadores. Além disso, a ferramenta apresenta também a possibilidade de criação de um *gold standard*, em que o corpus é anotado ou com a classificação atribuída pela maioria dos anotadores, ou com a mais comumente usada por eles. Tal documento, por sua vez, pode ser usado em outras aplicações, como aprendizado de máquina, por exemplo.

Como trabalhos futuros, pretende-se a inclusão de mais coeficientes de concordância, como o π de Scott (*cf.* [Artstein and Poesio 2008]), por exemplo. Além disso, futuras versões contarão também, a exemplo do que acontece com o MMAX2 [Müller and Strube 2006], com a apresentação de matrizes de confusão, bem como outras tabelas mais detalhadas. Também está em estudo a inserção de uma metodologia de cálculo de concordância para anotações com categorias hierarquicamente estruturadas, ou seja, anotações em que uma determinada categoria depende da escolha feita pelo anotador em outra (*cf.* [Geertzen and Bunt 2006]). Por fim, vale ressaltar que o AgreeCalc está disponível à comunidade acadêmica sob a GPL (GNU Public License), no endereço http://www.each.usp.br/norton/resdial/.

Agradecimentos

Esta pesquisa contou com o apoio do Programa de Educação Tutorial (PET) – MEC/SESu.

Referências

Apostolova, E., Neilan, S., An, G., Tomuro, N., and Lytinen, S. (2010). Djangology: A light-weight web-based tool for distributed collaborative text annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

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

Bontcheva, K., Cunningham, H., Roberts, I., Roberts, A., Tablan, V., Aswani, N., and Gorrell, G. (2013). Gate teamware: a web-based, collaborative text annotation framework. *Language Resources and Evaluation*, pages 1–23.

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Education and Psychological Measurement*, 20(1):37–46.

Craggs, R. and Wood, M. M. (2005). Evaluating discourse and dialogue coding schemes. *Computational Linguistics*, 31(3):289–296.

Davies, M. and Fleiss, J. L. (1982). Measuring agreement for multinomial data. *Biometrics*, 38(4):1047–1051.

Eugenio, B. D. and Glass, M. (2004). The kappa statistic: A second look. *Computational linguistics*, 30(1):95–101.

Fort, K., François, C., Galibert, O., and Ghribi, M. (2012). Analyzing the impact of prevalence on the evaluation of a manual annotation campaign. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 224–230, Istanbul, Turkey. European Language Resources Association (ELRA). ACL Anthology Identifier: L12-1233.

Geertzen, J. and Bunt, H. (2006). Measuring annotator agreement in a complex hierarchical dialogue act annotation scheme. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 126–133, Sydney, Australia. Association for Computational Linguistics.

Grouin, C., Rosset, S., Zweigenbaum, P., Fort, K., Galibert, O., and Quintard, L. (2011). Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In *Proceedings of the Fifth Linguistic Annotation Workshop (LAW V)*, pages 92—100, Portland, Oregon, USA.

Gut, U. and Bayerl, P. S. (2004). Measuring the reliability of manual annotations of speech corpora. In *Proceedings of the Second International Conference for Speech Prosody 2004 (SP2004)*, Nara, Japan. Poster.

Hripcsak, G. and Rothschild, A. S. (2005). Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Associati*, 12(3):296–298.

Krippendorff, K. (2004). *Content Analysis: An Introduction to its Methodology*. SAGE, 2nd edition.

Mathet, Y., Widlöcher, A., Fort, K., François, C., Galibert, O., Grouin, C., Kahn, J., Rosset, S., Zweigenbaum, P., and Zweigenbaum, P. (2012). Manual corpus annotation:

Giving meaning to the evaluation metrics. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012): Posters*, pages 809–818, Mumbai, India. The COLING 2012 Organizing Committee.

Mitkov, R., Orasan, C., and Evans, R. (1999). The importance of annotated corpora for nlp: the cases of anaphora resolution and clause splitting. In *Proceeding of "Corpora and NLP: Reflecting on Methodology Workshop", TALN'99*, pages 60 – 69, Cargese, Corse.

Müller, C. and Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. In Braun, S., Kohn, K., and Mukherjee, J., editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.

Ogren, P. V. (2006). Knowtator: A protégé plug-in for annotated corpus construction. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL 2006)*, volume companion volume: demonstrations, pages 273–275, New York, USA.

Petasis, G. (2012). The sync3 collaborative annotation tool. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 363–370, Istanbul, Turkey. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Petasis, G., Karkaletsis, V., Paliouras, G., Androutsopoulos, I., and Spyropoulos, C. D. (2002). Ellogon: A new text engineering platform. In *In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 72–78, Las Palmas, Canary Islands, Spain. European Language Resources Association.

Pustejovsky, J. and Stubbs, A. (2012). *Natural Language Annotation for Machine Learning*. O'Reilly Media, 1 edition. ISBN: 978-1-4493-0666-3.

Roman, N. T. (2012). Resdial – descrição da codificação (v.1.0). Technical Report 001/2012, PPgSI-EACH-USP, São Paulo, SP – Brazil.

Roman, N. T. and Carvalho, A. M. B. R. (2010). A multi-dimensional annotation scheme for behaviour in dialogues. In *Proceedings of the 12th Ibero-American Conference on Artificial Intelligence (IBERAMIA 2010)*, volume 6433 of *Advances in Artificial Intelligence*, pages 386–395, Bahía Blanca, Argentina. Springer. ISBN: 978-3-642-16951-9.

Verhagen, M. (2010). The brandeis annotation tool. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 3638–3643, Valletta, Malta.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference (MUC-6)*, pages 45–52, San Mateo, CA, USA. Morgan Kaufmann.