A Methodology for the Comparison of Human Judgments With Metrics for Coreference Resolution

Mariya Borovikova[1, 2] **Loïc Grobol**[2, 3] **Anaïs Lefeuvre-Halftermeyer**[2] **Sylvie Billot**[2]

[1]ILPGA, Université Sorbonne Nouvelle
[2]LIFO, Université d'Orléans
[3]Modyco, CNRS et Université Paris Nanterre

mariya.borovikova@sorbonne-nouvelle.fr
lgrobol@parisnanterre.fr
{anais.halftermeyer,sylvie.billot}@univ-orleans.fr

Abstract

We propose a method for investigating the interpretability of metrics used for the coreference resolution task through comparisons with human judgments. We provide a corpus with annotations of different error types and human evaluations of their gravity. Our preliminary analysis shows that metrics considerably overlook several error types and overlook errors in general in comparison to humans. This study is conducted on French texts, but the methodology should be language-independent.

1 Introduction

Coreference resolution is still one of the most challenging tasks in Natural Language Processing. Several metrics have been proposed to evaluate the task, each of them meant to rectify the weaknesses of the previous ones. However, neither their correctness nor their ability to reflect the real quality of algorithms is easily be provable from their mathematical definition. Consequently, some additional tests should be conducted in order to confirm their pertinence. This work aims to compare the evaluation measures used for coreference resolution task with human judgments, i.e. to study them in terms of interpretability. More precisely, B-CUBED (Bagga and Baldwin, 1998), LEA (Moosavi and Strube, 2016), CEAFe and CEAFm (Luo, 2005), CoNLL-2012 (MELA) (Denis and Baldridge, 2009), BLANC (Recasens and Hovy, 2011) and MUC (Vilain et al., 1995) metrics will be analysed.

2 Related work

Although some properties of coreference resolution quality measures have already been studied in Lion-Bouton et al. (2020), Moosavi (2020), Kummerfeld and Klein (2013) and others, to the best of our knowledge, there are no works dedicated to the comparison between automatic measurements and human evaluation of performance for this task. However, very few similar studies were conducted in other domains.

Doshi-Velez and Kim (2017) study the interpretability of machine learning models, in general, using application-grounded, human-grounded, and functionally-grounded approaches.

Foster (2008) describes an experience of evaluating a non-verbal behaviour of an embodied conversational agent. People were asked to choose the most appropriate talking head among the two generated using different strategies. Then β inter-annotator agreement measure (Artstein and Poesio, 2008) was calculated.

In Plank et al. (2015), the correlation between metrics for the dependency parsing task and human judgments was examined. Several models were tested for different languages. The annotators had to choose the best of the two annotations predicted by two different models without knowing the correct option. The obtained results were normalised using Spearman's ρ and compared with standard metrics.

Novikova et al. (2017) explore Natural Language Generation (NLG) evaluation measures. The annotation process is organised as follows: an annotator should score an example using three Likert scales from 0 to 6 based on informativeness, naturalness and quality criteria. The obtained results were normalised using Spearman and intra-class correlation coefficients and compared with NLG metrics.

Considering these studies, for the present research, we will use an approach similar to Novikova et al. (2017), where the annotators evaluate a system on a Likert scale. Despite possible difficulties with the Likert scale treatment (too many mid-point answers, a broad spectrum of responses for one question, etc.), this method seems more appropriate for our purposes.

Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval), pages 16 - 23

Two main reasons make us choose this approach: (1) we do not test particular systems and, therefore, have no alternative annotations and (2) a scaled approach is more accurate and exact while evaluating a system.

3 Methodology

This section is dedicated to the theoretical description of the methods used in the experiments within the scope of this study.

3.1 Errors typology

In order to correctly evaluate the quality of the algorithm, it is necessary to consider all the types of errors it can produce and, therefore, to define those types.

For our purposes, we have chosen the typology of Landragin and Oberle (2018):

1. **Border errors** occur when limits of referential expressions are marked inaccurately;
2. **Type errors** occur when a referential expression is assigned to a false chain;
3. **Noise errors** occur when irrelevant linguistic expressions are marked as a part of a coreference chain;
4. **Silence errors** occur when a system ignores referential expressions which are included in a relevant coreference chain;
5. **Tendency of irrelevant coreference chains construction** occurs when a system composes a new chain from several unrelated mentions.

We use this typology because it is more comprehensive than others and reflects the semantic aspect of the problem. However, we need to introduce an additional error type which we call "chain absence". This error may be regarded as a form of the "silence" error, and it occurs when the whole coreference chain (entity) is missing. The necessity of introducing a new error type arose after the experimentation phase of this study as it allowed to explain some patterns in the behaviour of the metrics. You can find the examples for each error type in the appendix section 1.

3.2 Corpus creation

Our corpus consists of a series of texts, each with two coreference annotations: one is a manual gold annotation, and the other is a purposefully erroneous annotation, one or more manually introduced errors of one of the types defined in section 3.1. There are also a few examples with errors of different types. Two existing coreference resolution corpora for French were used as a basis for the corpus. 52 texts were taken from the DEMOCRAT corpus (Landragin, 2018) and 4 examples - from the ANCOR corpus (Muzerelle et al., 2014). More precisely, we have selected the self-standing passages that are understandable out of context. The corpora are collected in the CoNLL-2012 format (Pradhan et al., 2012)[1]. The final dataset consists of 127 passages of 90-130 words each. 108 examples contain only one error, allowing us to analyse to what extent each error reduces the overall system quality. The rest of the samples are needed to adjust the annotations . Coreference chains lengths vary from 2 to 20 mentions. The mentions to contain an error were chosen at random. The total number of each error in the 108 samples varies between 16 and 28. The total number of each error varies between 44 and 97.

3.3 Evaluation scale

As the primary goal of this study is to evaluate the interpretability of the metrics, it is necessary to compare them to humans opinions about the correctness of the system's responses. Even though the metrics' output values are between 0 and 1, we will not use this range as it is more natural for people to evaluate the quality on an integer scale.

For our study, we use a Likert scale (Likert, 1932) with an even number of choices in order to avoid too many mid-point answers. Usually, coreference resolution is only a part of a pipeline of a more complex system, and the way of evaluation depends on the resolved task. In this study, an information retrieval task has been chosen as a global framework. These conditions require some changes in the classic scale; namely, we introduce a notion of the "importance" of an element. We distinguish two types of elements: peripheral elements and key elements. Peripheral elements can be removed from a text without severe consequences in its general sense. Key elements constitute the core of a text, so their removal will lead to the total loss of meaning. Thus, the gravity of an error and the importance of an element with an error is taken into account.

This scale also contains two points to allow differentiation between similar examples with little nuances: (0) The presumed system's annotation contains significant errors on key elements; (1-2) The presumed system's annotation contains significant

[1] https://github.com/boberle/coreference_databases.git

errors on peripheral elements; (3-4) The presumed system's annotation contains insignificant errors on key elements; (5-6) The presumed system's annotation contains insignificant errors on peripheral elements; (7) The presumed system's annotation does not contain any errors.

3.4 Annotation

Every annotation sample contains a correct annotation and an annotation with mistakes. In order to detect inconsistent annotators, three samples appear twice. The objective given to the annotators is to evaluate coreference resolution samples as a part of an information retrieval system using the Likert scale described in section 3.3. General instructions given before the annotations explain all the necessary concepts[2].

As an inter-annotator agreement measure, Krippendorff's alpha (Krippendorff, 1970) has been chosen and used to identify annotators whose answers differ much from the others using a new algorithm (see algorithm 1 in the appendix). The Krippendorff's alpha is computed for all the possible annotators combinations. Then, these combinations and their scores are sorted by ascending alpha score. We assume that those annotators whose rank is below the others are more important. In order to consider the differences between the alpha scores, the ranks are multiplied by their corresponding alpha scores. The final score is the sum of obtained values for each annotator. These values allow us to understand the annotators' ranking as better annotators have a higher score, but even with these values it remains unclear how to detect the outliers. In order to do this, we divide all the scores by the maximal value.

The coefficients obtained by the algorithm (hereinafter the trust coefficients) allow us to detect outliers (an annotator is considered an outlier if their score is less than or equal to 0.5).

In order to interpret the reasoning of each respondent, regressors have been trained to imitate the annotators' and metrics' behaviours. Each model should predict a score having the number of occurrences of each error type as input features. We have trained one model for each annotator and metric. Once the models are trained, the weights assigned to each feature (error type) are extracted and used for further interpretation.

4 Experiments and results

Human evaluation analysis. Since participation in this study was not rewarded and contained many questions, it involved only 12 participants, 9 of whom were linguists and 8 of whom have already worked with coreference. The analysis of the three duplicated questions showed that no one answered at random among the annotators. Krippendorff's alpha is rather low, so we supposed that some questions in our questionnaire raised more confusion among the respondents than others. Therefore we eliminated the questions that contained more than three different answers from the annotators and computed the results only for the remaining *simple* questions. The total number of questions used in the main analysis is 97. We also decided to compute the inter-annotator agreement on a reduced scale from 0 to 4 points ($0 \rightarrow 0$, 1 and $2 \rightarrow 1$, 3 and $4 \rightarrow 2$, 5 and $6 \rightarrow 3$, $7 \rightarrow 4$) and on the gravity (no errors - insignificant error(s) - significant error(s)) and elements importance (no errors - error(s) on peripheral element - error(s) on key element) scales. These agreements are presented in table 1.

Scale	All examples	*Simple* examples
Standard	0.11	$0.25 \rightarrow 0.27$
Reduced	0.16	0.34
Gravity	0.24	0.48
Importance	0.16	$0.34 \rightarrow 0.42$

Table 1: Krippendorff's alphas. An arrow shows that there are outlier annotators on the particular scale and set of examples. A value on the right of an arrow is an alpha after removing outlier annotators.

Human-machine correlation analysis. In order to compare the obtained scores with human judgments, we calculated an average and a mode of human evaluations having previously transformed to a scale from 0 to 1. Every metric was compared with the annotators' assessment on the standard scale, on the reduced scale and on the scale with errors gravity evaluation only. According to the data distributions, in general, the difference between a metric and humans is about 0.33. The averages of differences for all the examples are given in table 2.

Analysis by error type. In order to analyse the influence of a particular error type on a score, we train a linear regression model with the number of errors of each type as the input features and the reversed scores[3] as the outputs. All the input features were centered and reduced in order to obtain more stable results. The coefficients that were assigned to each input feature (and which correspond to one of the error types) during the training have been used as a

[2]You can find the google form with the instructions at https://forms.gle/cgpsfZvKg5zasnqd6.

[3]We replaced 7 by 0, 6 by 1, 5 by 2, etc.

Scale	Method	MUC	B-CUBED	CEAFm	CEAFe	BLANC	LEA	CoNLL
Standard	Average	0.289	0.321	0.308	0.269	0.281	0.231	0.291
	Mode	0.294	0.326	0.313	0.283	0.285	0.24	0.299
Reduced	Average	0.314	0.346	0.333	0.29	0.305	0.253	0.315
	Mode	0.312	0.347	0.333	0.292	0.304	0.255	0.316
Gravity	Average	0.43	0.463	0.45	0.405	0.422	0.368	0.432
	Mode	0.43	0.464	0.451	0.408	0.422	0.369	0.434

Table 2: Differences between humans evaluations and metrics on the scale from 0 to 1.

Name	Border	Type	Noise	Silence	Irrelevant chains	Chain absence
MUC	−0.242	−0.249	−0.121	−0.58	−0.345	−0.076
B-CUBED	−0.662	−0.15	—	−0.889	—	−0.264
CEAFm	−0.325	−0.34	−0.139	−0.408	−0.353	−0.101
CEAFe	−0.458	−0.283	−0.322	−0.447	−0.222	—
CoNLL	−0.382	−0.217	−0.083	−0.556	−0.179	−0.187
BLANC	−0.174	−0.385	−0.233	−0.973	−0.074	−0.56
LEA	−0.425	−0.22	−0.207	0.73	−0.432	—
Humans	−0.343	−0.629	−0.598	−0.513	−0.467	−0.727

Table 3: Coefficients of errors importances. "*Humans*" is the average of all the coefficients of models trained on humans' evaluations. See a more detailed version in the appendix (table 4).

measure of the importance of an error in the process of deciding the example's score (see tables 3 and 4).

5 Discussion

Human evaluation analysis. Table 1 reports the inter-annotator agreement on different scales, with several interesting properties about the task. Firstly, we may observe that the reduced scale results are better than those on the standard scale. It can be explained by the fact that even if people agree on the characteristics of the suggested categories, all of them have their own bias about the task, so they pay attention to different annotation nuances. Secondly, the inter-annotator agreement increased when we eliminated the annotators indicated as outliers by the trust coefficient.

Human-machine correlation analysis. One may notice that the average scores of all annotators are relatively high (see table 2). The average difference between all metrics and the annotators is usually above 0 and varies from 0.2 to 0.4 after normalisation, which shows that, generally, metrics tend to overestimate the actual quality of a model significantly.

Analysis by error type. In order to perform the analysis regarding the error types, we modified the table 4 by removing all positive and null coefficients as they mean either the absence of answers considering a particular error type or insufficient training quality of some models. These modifications can be justified by the fact that every coefficient of the model should be negative. Otherwise, it would mean that the presence of an error improves a score.

As our analysis shows, the **border**, **silence** and **irrelevant chains construction** errors are treated correctly. It could be proven by the fact that metrics coefficients are similar to the human ones. The **type**, **noise** and **chain absence** errors are underestimated by the metrics, as their scores are usually higher for the metrics than for the humans coefficients (see correspondent columns of the table 3).

We can analyse each metric separately as well. Firstly, we have noticed that the **MUC** metric considerably underestimates all types of errors except for the "silence" and the "irrelevant chains" ones. Secondly, the **B-CUBED** measure put relevant scores only to the examples which contain "border" and "silence" errors. The **CEAFe** score estimates correctly only the examples with "border" and "irrelevant chains" errors. Similarly, the **CEAFm** metric also underestimates all examples where any errors except for "border" and "irrelevant chains" ones were made. The **BLANC** measure treats properly only texts with "silence" errors. We observe that the **CoNLL-2012** metric tends to overstate the results of a model when the examples contain any errors except for "border" errors. Likewise, the **LEA** metric considerably underestimates all error types except for "border", "silence" and "irrelevant chains" errors (see correspondent lines of the table 3).

6 Conclusion

This study aims to investigate the extent to which we may understand the results produced by the coreference resolution metrics. The preliminary

results on the limited corpus show that metrics underestimate errors gravity compared to humans and add approximately 0.33 points to the final score on the scale from 0 to 1. However, these results need to be proven on a more significant number of annotators.

This work's contribution consists in creating the corpus with various errors types and its annotation with the human judgments about the gravity of these errors, the proposal of the new automatic outlying annotator identification algorithm and the suggestion of a methodology of comparison of human evaluations with automatic metrics. All the code and corpus are available at `https://github.com/project178/coref-metrics-vs-humans`.

Possible future work directions may consist in involving more people in the annotation process of the proposed corpus in order to verify the obtained results and in the development of a new metric that will take into consideration the identified shortcomings of the existing measures.

7 Acknowledgements

This work was funded by Région Centre-Val-de-Loire through the RTR DIAMS.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Citeseer.

Pascal Denis and Jason Baldridge. 2009. Global joint models for coreference resolution and named entity classification. *Procesamiento del lenguaje natural*, 42.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Mary Ellen Foster. 2008. Automated metrics that agree with human judgements on generated output for an embodied conversational agent. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 95–103, Salt Fork, Ohio, USA. Association for Computational Linguistics.

Klaus Krippendorff. 1970. Bivariate agreement coefficients for reliability of data. *Sociological methodology*, 2:139–150.

Jonathan K. Kummerfeld and Dan Klein. 2013. Error-driven analysis of challenges in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 265–277, Seattle, Washington, USA. Association for Computational Linguistics.

Frédéric Landragin and Bruno Oberle. 2018. Identification automatique de chaînes de coréférences : vers une analyse des erreurs pour mieux cibler l'apprentissage. In *Journée commune AFIA-ATALA sur le Traitement Automatique des Langues et l'Intelligence Artificielle pendant la onzième édition de la plate-forme Intelligence Artificielle (PFIA 2018)*, Nancy, France.

Frédéric Landragin. 2018. LIVRABLE L2 "Manuel d'annotation du corpus et organisation de formations sur l'annotation" du projet DEMOCRAT. Research report, Lattice and LiLPa and ICAR and IHRIM.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Adam Lion-Bouton, Loïc Grobol, Jean-Yves Antoine, Sylvie Billot, and Anaïs Lefeuvre-Halftermeyer. 2020. Comment arpenter sans mètre : les scores de résolution de chaînes de coréférences sont-ils des métriques ? (do the standard scores of evaluation of coreference resolution constitute metrics ?). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). 2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)*, pages 10–18, Nancy, France. ATALA et AFCP.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Nafise Sadat Moosavi. 2020. *Robustness in Coreference Resolution*. Ph.D. thesis.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany. Association for Computational Linguistics.

Judith Muzerelle, Anaïs Lefeuvre, Emmanuel Schang, Jean-Yves Antoine, Aurore Pelletier, Denis Maurel, Iris Eshkol, and Jeanne Villaneau. 2014. ANCOR_Centre, a large free spoken French coreference corpus: description of the resource and reliability measures. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 843–847, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017*

Conference on Empirical Methods in Natural Language Processing, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Barbara Plank, Héctor Martínez Alonso, Željko Agić, Danijela Merkler, and Anders Søgaard. 2015. Do dependency parsing metrics correlate with human judgments? In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 315–320, Beijing, China. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Marta Recasens and Eduard Hovy. 2011. BLANC: Implementing the Rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485.

Marc Vilain, John D Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

A Appendix

1. **Borders errors.** *Whales are marine mammals*. instead of *Whales are marine mammals*.
2. **Type errors.** *John likes his brother because he is funny* instead of *John likes his brother because he is funny*.
3. **Noise errors.** *The dog barked. It's time to go.* instead of *The dog barked. It's time to go.*
4. **Silence errors.** *A phone is on the table. It rings. I pick it up* instead of *A phone is on the table. It rings. I pick it up.*
5. **Tendency of irrelevant coreference chains construction.** *A cat and a dog are playing together* instead of *A cat and a dog are playing together.*
6. **Chain absence.** *A phone is on the table. It rings. I pick it up* instead of *A phone is on the table. It rings. I pick it up.*

Figure 1: Error types examples.

Name	Border	Name	Type	Name	Noise	Name	Silence	Name	Irrelevant chains	Name	Chain absence
REDUCED_A6	-0,033	**B-CUBED**	**-0,15**	**CoNLL-2012**	**-0,083**	GRAVITY_A6	-0,011	GRAVITY_A6	-0,018	**MUC**	**-0,076**
GRAVITY_A10	-0,057	**CoNLL-2012**	**-0,217**	GRAVITY_A7	-0,105	STANDARD_A12	-0,262	**BLANC**	**-0,074**	STANDARD_A10	-0,099
STANDARD_A7	-0,097	**LEA**	**-0,22**	**MUC**	**-0,121**	GRAVITY_A8	-0,265	STANDARD_A11	-0,08	**CEAFm**	**-0,101**
GRAVITY_A7	-0,118	GRAVITY_A7	-0,239	REDUCED_A1	-0,122	STANDARD_A11	-0,274	REDUCED_MEAN	-0,097	GRAVITY_A11	-0,124
GRAVITY_A11	-0,152	GRAVITY_A8	-0,24	**CEAFm**	**-0,139**	STANDARD_A9	-0,329	GRAVITY_A11	-0,164	STANDARD_A7	-0,17
STANDARD_A10	-0,166	**MUC**	**-0,249**	STANDARD_A9	-0,146	GRAVITY_A11	-0,391	STANDARD_A4	-0,167	**CoNLL-2012**	**-0,187**
BLANC	**-0,174**	**CEAFe**	**-0,283**	GRAVITY_A11	-0,167	STANDARD_A10	-0,406	**CoNLL-2012**	**-0,179**	GRAVITY_A7	-0,247
STANDARD_A6	-0,182	STANDARD_A10	-0,31	**LEA**	**-0,207**	**CEAFm**	**-0,408**	**CEAFe**	**-0,222**	**B-CUBED**	**-0,264**
STANDARD_A8	-0,212	**CEAFm**	**-0,34**	**BLANC**	**-0,233**	**CEAFe**	**-0,447**	*GRAVITY_MEAN*	*-0,319*	GRAVITY_A8	-0,265
MUC	**-0,242**	STANDARD_A12	-0,362	STANDARD_A7	-0,316	GRAVITY_A1	-0,49	STANDARD_A8	-0,335	STANDARD_A6	-0,283
CEAFm	**-0,325**	GRAVITY_A3	-0,372	**CEAFe**	**-0,322**	GRAVITY_A7	-0,547	**MUC**	**-0,345**	STANDARD_A11	-0,289
CoNLL-2012	**-0,382**	**BLANC**	**-0,385**	STANDARD_MEAN	-0,356	STANDARD_A6	-0,548	**CEAFm**	**-0,353**	STANDARD_A12	-0,411
GRAVITY_A3	-0,388	STANDARD_A8	-0,458	STANDARD_A11	-0,364	**CoNLL-2012**	**-0,556**	GRAVITY_A7	-0,375	**BLANC**	**-0,56**
LEA	**-0,425**	GRAVITY_A11	-0,507	STANDARD_A10	-0,532	**MUC**	**-0,58**	STANDARD_A6	-0,407	STANDARD_A8	-0,634
CEAFe	**-0,458**	STANDARD_A9	-0,563	GRAVITY_A3	-0,566	GRAVITY_A10	-0,586	GRAVITY_A8	-0,409	GRAVITY_A1	-0,737
B-CUBED	**-0,662**	STANDARD_MEAN	-0,607	GRAVITY_A1	-0,691	STANDARD_A7	-0,629	**LEA**	**-0,432**	STANDARD_A9	-0,756
GRAVITY_MEAN	*-0,848*	STANDARD_A11	-0,625	REDUCED_MEAN	-0,712	**LEA**	**-0,73**	GRAVITY_A3	-0,453	STANDARD_A4	-0,919
REDUCED_A1	-1,529	GRAVITY_A1	-0,663	GRAVITY_A4	-0,911	REDUCED_MEAN	-0,883	STANDARD_A9	-0,491	GRAVITY_A10	-0,932
		STANDARD_A7	-0,686	GRAVITY_A10	-0,954	**B-CUBED**	**-0,889**	STANDARD_A7	-0,64	REDUCED_MEAN	-0,943
		STANDARD_A6	-0,741	*GRAVITY_MEAN*	*-1,184*	**BLANC**	**-0,973**	REDUCED_A6	-0,819	GRAVITY_A6	-1,149
		GRAVITY_A6	-0,741	GRAVITY_A6	-1,197	GRAVITY_MEAN	-1,565	STANDARD_A12	-0,859	REDUCED_A1	-1,419
		STANDARD_A4	-0,818	STANDARD_A6	-1,238			STANDARD_MEAN	-0,904	REDUCED_A6	-1,473
		GRAVITY_A4	-0,869					GRAVITY_A4	-0,911	*GRAVITY_MEAN*	*-2,23*
		GRAVITY_MEAN	*-1,027*					GRAVITY_A1	-0,961		
		REDUCED_MEAN	-1,5								

Table 4: Coefficients of error importances obtained during the regressors training for all metrics and annotators. Values in bold are reported by metrics' regressors. Values in italic are reported by a regressor trained on a mean answer on the gravity scale.

Algorithm 1 Calculate trust coefficients

Input: annotated corpus with k annotators

```
alphas ← empty dictionary
for n = 2 To k+1 do
    for each combination ∈ COMBINATIONS(n,k) do
        alphas[combination] ← KRIPPENDORFFSALPHA(corpus[combination])
    end for
end for
SORT alphas BY alphas.values
coefs ← empty dictionary
coef ← 1
score ← 0
for each annotators_comb,alpha ∈ alphas do
    if score < alpha then
        coef ← coef + 1
        score ← alpha
    end if
    for each annotator ∈ annotators_comb do
        coefs[annotator] ← coefs[annotator] + coef × alpha
    end for
end for
coefs.values ← coefs.values/max(coefs.values)
```

Output: coefs