Human annotation of ASR error regions: Is "gravity" a sharable concept for human annotators?

Daniel Luzzati[1] **Cyril Grouin**[2] **Ioana Vasilescu**[2]
Martine Adda-Decker[2,3] **Eric Bilinski**[2] **Nathalie Camelin**[1]
Juliette Kahn[4] **Carole Lailler**[1] **Lori Lamel**[2] **Sophie Rosset**[2]

[1]Université du Maine, EA 4023, LIUM
72085 Le Mans, France
firstname.lastname@lium.univ-lemans.fr

[2]CNRS, UPR 3251, LIMSI
91403 Orsay, France
firstname.lastname@limsi.fr

[3]Université Sorbonne Nouvelle Paris 3, UMR 7018, LPP
75005 Paris, France
martine.adda-decker@univ-paris3.fr

[4]LNE
78197 Trappes, France
firstname.lastname@lne.fr

Abstract

This paper is concerned with human assessments of the severity of errors in ASR outputs. We did not design any guidelines so that each annotator involved in the study could consider the "seriousness" of an ASR error using their own scientific background. Eight human annotators were involved in an annotation task on three distinct corpora, one of the corpora being annotated twice, hiding this annotation in duplicate to the annotators. None of the computed results (inter-annotator agreement, edit distance, majority annotation) allow any strong correlation between the considered criteria and the level of seriousness to be shown, which underlines the difficulty for a human to determine whether a ASR error is serious or not.

Keywords: Annotation; ASR Seriousness Errors; Speech Recognition

1. Introduction

When addressing the issue of transcription errors in automatic speech recognition (ASR) systems, we may adopt two distinct perspectives:

- We may be interested in etiology, trying to establish a causal relationship between the properties of the speech signal entering the ASR system and the transcription errors in the output. In this case, we argue in a systemic way, trying to relate the errors to causal error categories.

- A different way of looking into ASR errors consists of judging the impact of such errors on further processing, be it automatic or human. Taking an axiological perspective, our aim is then to qualify the seriousness of ASR errors on a seriousness scale ranging from minor to huge mistakes.

In the automatic speech recognition literature, word error rates are regularly reported, however few studies focus on ASR error analyses. In general, such studies aim at identifying major reasons of error (Duta et al., 2006; Adda-Decker, 2006; Nemoto et al., 2008; Goldwater et al., 2010; Dufour et al., 2012) classifying errors according to their phonetic characteristics (Greenberg and Chang, 2000) or at comparing automatic and human performances (Lippmann, 1997; Shen et al., 2008; Vasilescu et al., 2011). Studies on ASR *error seriousness* in link with further processing (Woodland et al., 2000) tend to be lacking. In contrast, error seriousness assessment is very popular subject in foreign language teaching and learning (Vann et al., 1991; Hyland and Anan, 2006).

In this paper, we address the issue of error "gravity" or seriousness in ASR. For this first experiment, we deliberately chose not to give precise guidelines to human judges (no hierarchy concerning linguistic levels involved in errors) nor to specify a precise framework concerning further processing (subtitling, information retrieval, translation, etc.). Instead, error seriousness decisions are individually taken by the independent participating assessors. This procedure raises the following questions:

- Do judges evaluate the seriousness of an error in the same way?

- Are judges consistent during evaluation, or is the evaluation similar to a random trial?

- When judging errors on a common *seriousness* scale, do judges follow different strategies (e.g., are errors harmful w.r.t. global understanding, language syntax, dialog systems, named entity recognition, etc.) depending on their personal competence and interests or is there a sharable *generic* view of the *seriousness* concept?

We would like to mention that this study is not meant to replace perceptual studies, but rather to determine findings which can be helpful in designing further investigations.

2. Material and methods

2.1. Corpora

The data used for this study are part of the French ETAPE corpus (Gravier et al., 2012) for which LIUM provided ASR outputs (Bougares et al., 2013). Three different files were used in this experiment:

- corpus 1: radio show debates (France Inter)
- corpus 2: parliamentary debates (LCP, Top Questions)
- corpus 3: radio show debates (France Inter)

Regions of interest for this study are determined by aligning the reference (manual transcription) with the hypothesis (automatic transcription), and locating *error regions*, which is defined as all the consecutive words in the hypothesis which are different from the reference. The error regions (ER) concern only two or more consecutive words, and do not include any correctly recognized words or single word subsitution errors.
Since these regions are automatically located by aligning the automatic (HYP) transcription with the reference one (REF), a temporal constraint is used to ensure that only temporally close words are associated with one another. An ER is thus the substitution of a sequence of words in the REF by a different word sequence in the HYP. The zone is determined by the time span (or the number of reference words) and not the type of error (deletion, insertion or substitution). Figure 1 illustrates an error region:

REF: on a souvent <ER> enfin en Seine Saint-Denis </ER> malheureusement
we often have well in Seine Saint-Denis unfortunately
HYP: on a souvent <ER> ***** FRANSEN ***** ***** SANI </ER> malheureusement
*we often have ***** FRANSEN ***** ***** SANI unfortunately*

Figure 1: Excerpt from the corpus with error region (ER)

Table 1 shows a few statistics describing error regions in each corpus.

Sources	#words	# ER	% words in ER	Mean ER length
corpus1	1229	192	46.2%	3.0
corpus2	2124	94	7.1%	1.6
corpus3	1475	210	34.2%	2.4

Table 1: General corpus description. **ER** stands for Error Regions

2.2. Method

2.2.1. Annotators

Eight[1] annotators participated in this annotation process, with different scientific background, either linguistics with a specialization in natural or spoken language processing (a1, a2, a7) or computer science (CS) without specialization (a4, a5), or with a specialization in speech recognition (a3) or spoken language processing (a6). Annotator (a5) is considered as a control annotator as this annotator knows very well the data annotated in this study.

Each human annotator annotated the corpora in the same order, and reannotated the first corpus at the end of the annotation process (corpus 1, 2, 3, and then 1 again). All but annotator a1, who prepared the subcorpora, were unaware that the first corpus would be reannotated. The other annotators discovered this repetition when annotating. This procedure allows us to compute inter- and intra-annotator agreement scores.

2.2.2. Annotation tool

We designed an annotation tool to meet the objectives we wanted to achieve. We decided to use a web interface so as to easily be able to save the performed annotations and record the annotation time for each human annotator. We also chose to provide keyboard shortcuts so as to rapidly annotate the corpus, depending on their position on the keyboard: (i) the keys "D", "F" and "G" respectively refer to low, intermediate and high levels of seriousness,[2] and (ii) the arrows keys are used to switch from one error region to another. This configuration allows the user to annotate with the left hand while the right hand is used to move within the corpus.
After logging in, the annotator has to choose the corpus he want to annotate. Then, the annotation tool provides, for each segment, a comparison between the reference transcription (upper part of the interface) and the automatic hypothesized transcription (lower part). Each unannotated error region in the segment is in black. The human annotator has to decide which level of seriousness is relevant for each error region (see Figure 2). After annotation, the color of the annotated region is changed depending on the selected seriousness level (green=low level, orange=intermediate level, red=high level).

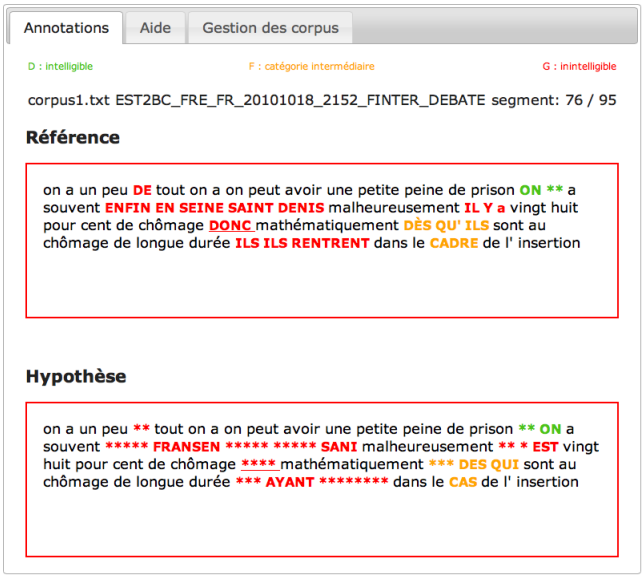

Figure 2: Screenshot of the annotation web interface

[1] During the annotation process, a technical problem occurred for one annotator, but this problem was not detected until the analysis stage. All annotations from this annotator were lost.

[2] These keys were chosen because they form a group on the left side of the keyboard. The letter from the key has no sense.

3. Results

Table 2 shows the distribution of annotations per annotator for each level of seriousness. It is apparent that annotators differ in their error seriousness assessment, with annotators a2, a5 and a6 tending to make stronger judgements (i.e., judging relatively few error regions to have intermediate seriousness levels) compared to the other annotators. Annotator a1 judges significantly more errors to be minor and much fewer severe than the other annotators.

	Annotators						
Level	a1	a2	a3	a4	a5	a6	a7
Low	351	134	248	167	159	109	149
Interm.	121	45	187	143	38	38	104
High	216	509	253	378	491	541	435

Table 2: Distribution of annotations per annotator

Table 3 further explores differences in annotator judgement. The confusion matrix gives the inter-annotator agreements for each pair of annotators based on all of the annotated corpora.

	a1	a2	a3	a4	a5	a6	a7
a1	—	0.24	0.50	0.37	0.31	0.20	0.32
a2	0.24	—	0.28	0.46	0.51	0.56	0.49
a3	0.50	0.28	—	0.49	0.37	0.26	0.41
a4	0.37	0.46	0.49	—	0.52	0.46	0.59
a5	0.31	0.51	0.37	0.52	—	0.56	0.57
a6	0.20	0.56	0.26	0.46	0.56	—	0.52
a7	0.32	0.49	0.41	0.59	0.57	0.52	—

Table 3: Inter-annotator agreement confusion matrix

Table 4 shows the intra-annotator agreements computed for each human annotator. This computation has been made for each annotator using the decisions taken for corpora 1 and 4. The intra-agreements range from 0.60 (annotators a1 and a7) to 0.70 (annotator a5).

	a1	a2	a3	a4	a5	a6	a7
κ	0.60	0.69	0.69	0.69	0.70	0.64	0.60

Table 4: Intra-annotator agreement

Table 5 shows the global inter-annotator agreement computed between each human annotators as a function of the corpus. The inter-annotator agreements range from 0.33 for corpus 3 to 0.47 for corpus 2. We can observe a difference between the agreement on corpus 1 (0.38) and on corpus 4 (0.37) even though they are the same.

	corpus 1	corpus 2	corpus 3	corpus 4
κ	0.38	0.47	0.33	0.37

Table 5: Inter-annotator agreement depending on the corpus

4. Discussion

4.1. Inter- and intra-annotator agreements

4.1.1. Inter-annotator agreements

From a global point of view, we observed very low values of inter-annotator agreements, which is not surprising due to the complexity of the task and our decision to not provide specific guidelines. We computed a Fleiss Kappa of 0.406 on the annotations from all annotators; this Kappa is of 0.388 if we do not take into account the annotations performed by the control annotator (a5).

The matrix confusion shown in Table 3 presents the inter-annotator agreements computed for each pair of annotators. We observed the IAA values on each pair are not really higher than the values computed between all annotators: the higher value is of 0.58 between a4 and a7, the lower value is of 0.19 between a1 and a6.

The IAA computed on each corpus (Table 5) allows us to notice clear differences depending on the considered corpus, even if agreements remain low. Indeed, we obtained similar IAA on corpora 1, 3 and 4 because of their common source (corpora 1 and 3 are debates and corpus 4 is the same than corpus 1) while corpus 2 is of a different genre. We noticed the corpus 2 allows the annotators to achieve higher IAA. Far easier than the other corpora, this corpus includes less error zones than the other corpora (Table 1).

Table 6 presents the confusion matrix of inter-annotator agreements taking into account the scientific background of each annotators. We can observe that the agreement has no direct link with the scientific background. For example, annotator a1 shares a common background with annotator a2 but their agreement is lower than those a1 has with other annotators.

	a1	a2	a3	a4	a5	a6	a7
a1	—	0.24	0.50	0.37	0.31	0.20	0.32
a2	0.24	—	0.28	0.46	0.51	0.56	0.49
a3	0.50	0.28	—	0.49	0.37	0.26	0.41
a4	0.37	0.46	0.49	—	0.52	0.46	0.59
a5	0.31	0.51	0.37	0.52	—	0.56	0.57
a6	0.20	0.56	0.26	0.46	0.56	—	0.52
a7	0.32	0.49	0.41	0.59	0.57	0.52	—

Table 6: Confusion matrix: agreement taking into account the background of each annotators. the following groups of annotators have a similar background: (a1, a2, a7), (a4, a5), (a3...a6)

4.1.2. Intra-annotator agreements

The intra-annotator agreements range from 0.604 to 0.699 (Table 4). Their values are higher than the inter-annotator agreements but still remain low suggesting that humans had difficulties classifying the error seriousness.

In most cases, annotators switched from a low level (L) of seriousness to an intermediate one (I) or from a high level (H) to an intermediate one (I). Table 7 gives an example of such a switching.

	a1	a2	a3	a4	a5	a6	a7
corpus1	L	H	I	**I**	**H**	H	I
corpus4	L	H	I	**H**	**I**	H	I

REF: si mais on <ER> LE DIT </ER> pas trop
yes but we say it not much
HYP: si mais on <ER> NE SAIT </ER> pas trop
yes but we don't know much

Table 7: Example of different annotation between corpus 1 and corpus 4 (the same corpus at two different moments) for the given excerpt

4.2. Equivalent categories

4.2.1. Mean annotation w.r.t. majority annotation

Figure 3 presents the distribution of the majority annotation with respect to the the mean annotation for the four sets of corpus. We can observe that a majority of annotations are clustered on the high level of seriousness.

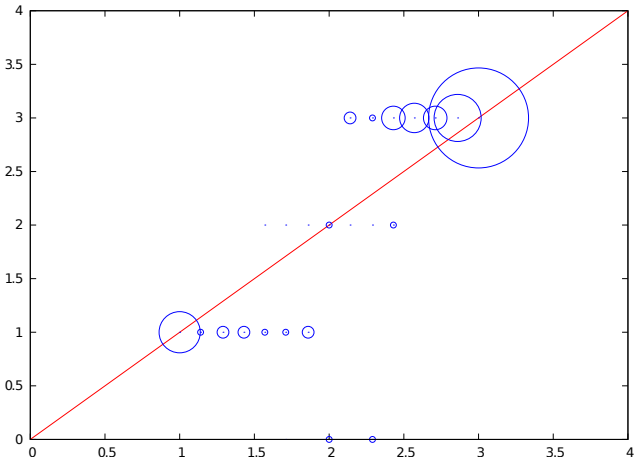

Figure 3: mean annotation w.r.t. majority annotation. Y-axis labels 1, 2 and 3 refer to low, intermediate and high level of seriousness. Y-axis label 0 refers to no majority annotation (i.e., the higher number of annotations per level is shared by two levels of seriousness)

Figure 4 gives a detailed analysis for each set of corpus. Corpus 2 has very different characteristics than the other two corpora.

4.2.2. Perfect consensus

We observed that a perfect consensus[3] (all seven annotators used the same level of seriousness) is only present on the far categories: 77 regions from the lower level of seriousness, no region from the intermediate level, and 174 regions from the higher level of seriousness.

All the human annotators agree more frequently on the higher level of seriousness than on the lower one. There is no consensus on the intermediate level, which seems to be not so surprising due to the quite high number of human annotators involved in this study (see Table 8).

[3]We considered here all the annotations provided by the annotators, i.e. 688 regions for each one of the 7 annotators and not only the 496 primary regions to annotate.

Corpus	Low level	Intermediate	High level
Corpus 1	15	0	54
Corpus 2	34	0	6
Corpus 3	11	0	61
Corpus 4	17	0	53

Table 8: Distribution of perfect consensus in each level of seriousness for each corpus

Out of 688 error regions, we observed a perfect consensus on 251 regions, i.e., 36.5% of all regions. On a three-value scale, and taking into account seven human annotators, this percentage is quite high.

4.2.3. Annotation relevance

While corpora 1 and 4 are the same corpus—annotated at two different times during the annotation process—we observed no discrepancy between the distribution of the perfect consensus within the three-value scale.

A surprising result is observed on the corpus 2, where the perfect consensus is more likely present on the lower level of seriousness category than on the higher level, contrary to other corpora. Nevertheless, this corpus is also the one for which there are fewer error regions to analyze.

Finally, we noticed that the difference of media did not affect the distribution of perfect consensus between categories: the corpora 1 (from "France Inter" radio station) and 3 (from "La Chaîne Parlementaire" parliamentary television) do not provide distinct results in terms of consensus.

4.3. Non-equivalent categories

Instead of looking at perfect agreement between the three classes, we can consider that there can be an equivalence between the low and intermediate levels or the high and the intermediate levels of seriousness in some conditions. Following that idea, we can assume that for each error region, there are three classes of judgments:

- A: majority of *low level* of seriousness judgments; one *high level* of seriousness judgment at most;
- B: majority of *high level* of seriousness judgments; one *low level* of seriousness judgment at most;
- C: others.

Table 9 gives the distribution of the error region within these classes given the different corpus.

Corpus	Class A	Class B	Class C	Class A+B
corpus 1	19.27%	64.06%	16.66%	83.33%
corpus 2	53.19%	25.53%	21.27%	78.72%
corpus 3	14.76%	60.00%	25.24%	74.76%
corpus 4	15.62%	62.50%	21.87%	78.12%

Table 9: Distribution between three classes of judgments

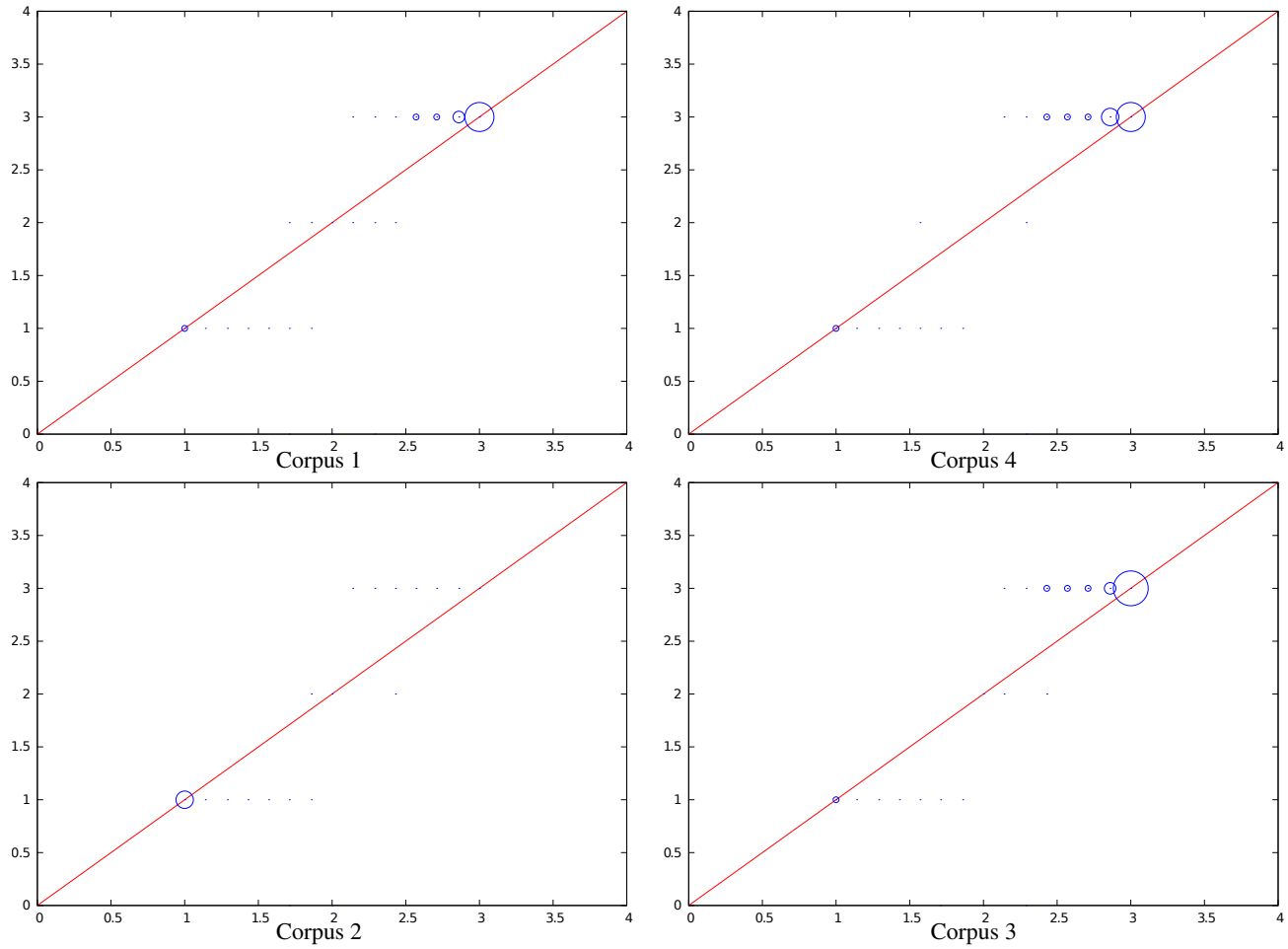

Figure 4: Mean Annotation w.r.t. Majority Annotation. Y-axis labels 1, 2 and 3 refer to low, intermediate and high level of seriousness. Y-axis label 0 refers to no majority annotation (i.e., the higher number of annotations per level is shared by two levels of seriousness)

We can observe that between 78 and 83% of the error region fall within a majority judgment class. The class C represents all the error region with complete undecided judgments. Only between 17% and 25% of the error region to be evaluated fall in this class.

4.4. Distance between hypothesis and reference

One hypothesis is that there may be a strong correlation between seriousness error and edit distance between hypothesis and reference. In order to validate this hypothesis, we computed edit distances for each error zone. Figure 5 shows the density annotation for each level of seriousness with respect to edit distance for all corpora while Figure 6 details the results for each corpus.

We can observe that the highest level of seriousness the highest the edit distance is (as shown with red numbers on these two figures). The distribution of these edit distances differs between Corpus 2 and the other corpora which confirms that Corpus 2 is different from the other ones. Moreover, at the high level of seriousness, we also observe an important number of short edit distances which means that edit distance alone is not enough to give a clear idea of the seriousness error.

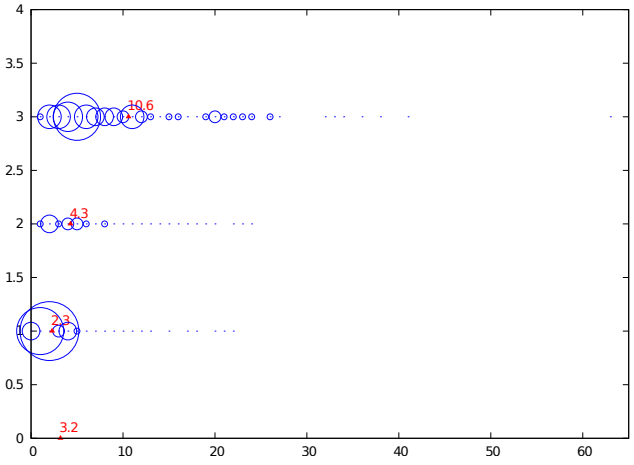

Figure 5: Annotation density w.r.t. Edit Distance for all corpora. Red numbers indicate the mean edit distance. Y-axis labels 1, 2 and 3 refer to low, intermediate and high level of seriousness. Y-axis label 0 refers to no majority annotation (i.e., the higher number of annotations per level is shared by two levels of seriousness)

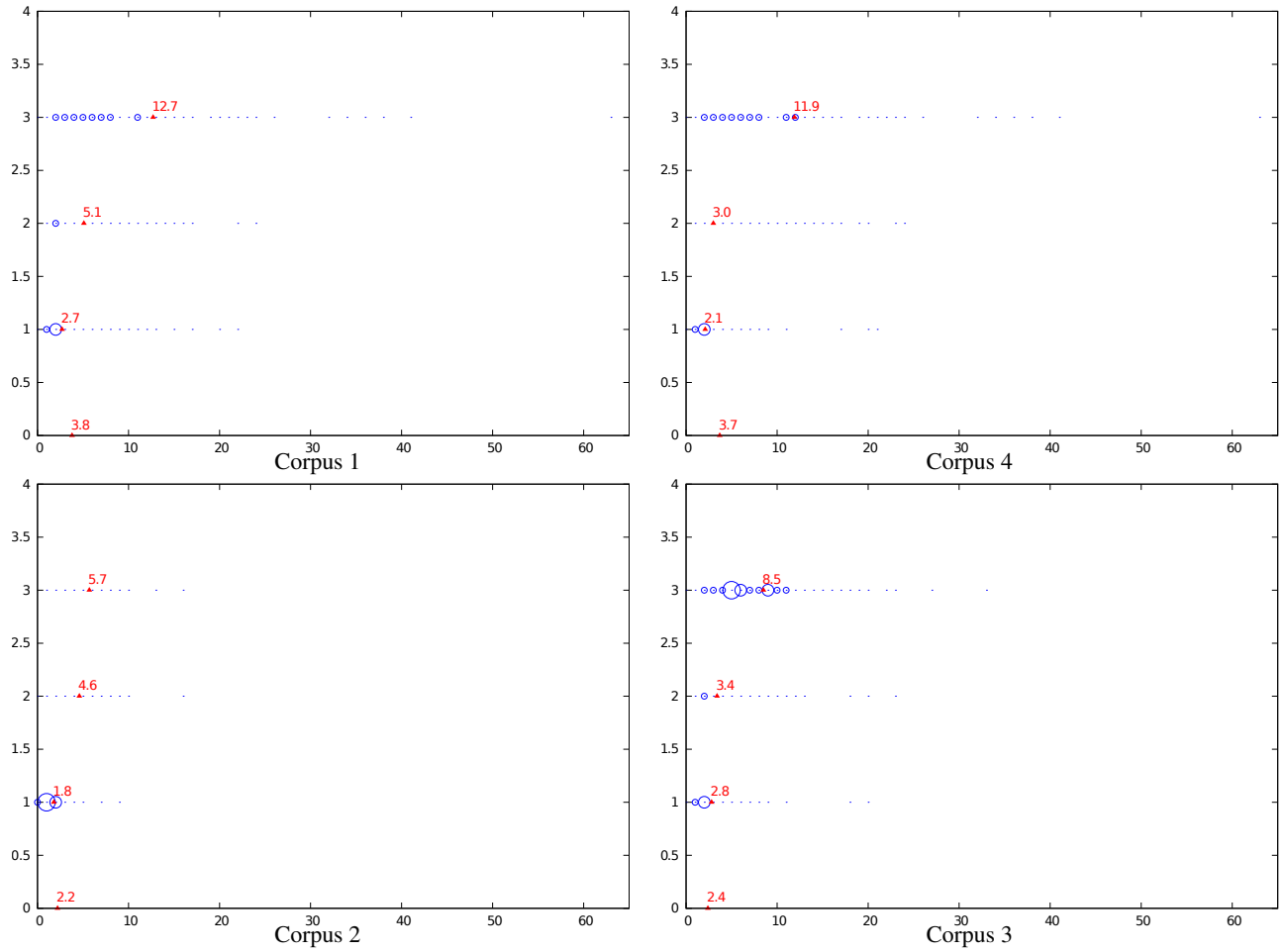

Figure 6: Annotation density w.r.t. Edit Distance for each corpus. Red numbers indicate the mean edit distance. Y-axis labels 1, 2 and 3 refer to low, intermediate and high level of seriousness. Y-axis label 0 refers to no majority annotation (i.e., the higher number of annotations per level is shared by two levels of seriousness)

5. Conclusion

Generally speaking, very low values of inter-annotator agreements are observed, which is not surprising due to the complexity of the task and our decision to not provide specific guidelines. This suggests that humans had difficulties classifying the error seriousness.

The kind of corpora is different enough to produce distinct quality of transcriptions which induces various annotators experiences. While the background of annotators does not seem to play a role in their task understanding, the error region characteristics have an impact on the classification task.

The analysis showed that there is no clear correlation between the considered criteria and the level of seriousness, and that this task was difficult for the human annotators.

6. Acknowledgements

This work was supported by the French National Agency for Research as part of the project VERA (adVanced ERrors Analysis for speech recognition) under grants ANR-2012-BS02-006-04.

We thank Dr Paul Deléglise, Dr Yannick Estève and Dr Olivier Galibert for their help in this work and their useful comments.

7. References

Adda-Decker, M. (2006). De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux. In *Proc. of JEP*, Dinard, France.

Bougares, F., Deléglise, P., Estève, Y., and Rouvier, M. (2013). LIUM ASR system for ETAPE French evaluation campaign: experiments on system combination using open-source recognizers. In *Sixteenth International Conference on TEXT, SPEECH and DIALOGUE (TSD 2013)*, Pilsen, Czech Republic.

Dufour, R., Damnati, G., and Charlet, D. (2012). Automatic error region detection and characterization in LVCSR transcriptions of TV news shows. In *Proc. of IEEE-ICASSP*.

Duta, N., Schwartz, R. M., and Makhoul, J. (2006). Analysis of the errors produced by the 2004 BBN speech recognition system in the DARPA EARS evaluations. *IEEE-TASLP*, 14:1745–1753.

Goldwater, S., Jurafsky, D., and Manning, C. D. (2010). Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52(3):181–200.

Gravier, G., Adda, G., Paulsson, N., Carré, M., Giraudel,

A., and Galibert, O. (2012). The ETAPE corpus for the evaluation of speech-based TV content processing in the french language. In *Proc of LREC*, Istanbul, Turkey.

Greenberg, S. and Chang, S. (2000). Linguistic dissection of switchboard-corpus automatic speech recognition systems. In *ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium, ASR2000*, Paris.

Hyland, K. and Anan, E. (2006). Teachers' perceptions of error: the effects of first language and experience. *System*, 34(4):509–519.

Lippmann, R. P. (1997). Speech recognition by machines and humans. *Speech Communication*, 22(1):99–115.

Nemoto, R., Vasilescu, I., and Adda-Decker, M. (2008). Speech errors on frequently observed homophones in french: perceptual evaluation vs automatic classification. In *Proc. of LREC*, Marrakesh, Morocco.

Shen, W., Olive, J. P., and Jones, D. A. (2008). Two protocols comparing human and machine phonetic recognition performance in conversational speech. In *Proc. of Interspeech*, Antwerp, Belgium.

Vann, R. J., Lorenz, F. O., and Meyer, D. M. (1991). Error gravity: faculty response to errors in written discourse of nonnative speakers of english. In Hamp-Lyons, L., editor, *Assessing second language writing in academic contexts*. Ablex Publishing, Norwood, NJ.

Vasilescu, I., Yahia, D., Snoeren, N. D., Adda-Decker, M., and Lamel, L. (2011). Cross-lingual study of ASR errors: on the role of the context in human perception of near homophones. In *Proc. of Interspeech*, pages 1949–1952, Florence, Italy.

Woodland, P. C., Johnson, S. E., Jourlin, P., and Jones, K. S. (2000). Effects of out of vocabulary words in spoken document retrieval. In *Proc. of SIGIR*, pages 372–374.