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

S. Rosset D. Luzzati C. Grouin I. Vasilescu
M. Adda-Decker E. Bilinski N. Camelin J. Kahn
C. Lailler L. Lamel

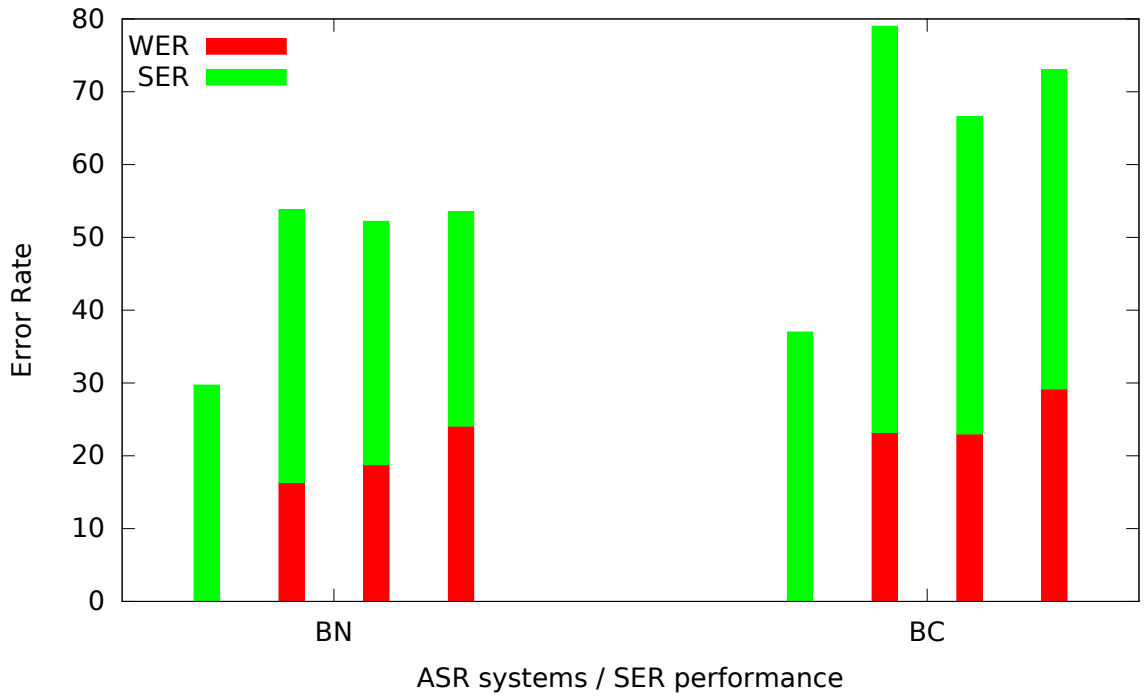
Context

French ANR VERA project

- Partners: LIUM, LNE, LPP, LIMSI
- Objectives
  - ASR errors typology and classification
  - Evaluation metrics within ASR systems (confidence scores) and for task evaluation
  - Integration of error knowledge within systems
    - Named Entity detection
    - Spoken Language Understanding
    - Question-Answering

Classical observations

- ASR systems produce, among other things, errors
- Performance between manual and automatic transcriptions decrease
- we often say *too many errors* but ...
- Example with ENE (Quaero project)



→ Not all errors have the same importance

→ WER is not always a good indicator for a following task success

Related Work

Automatic Speech Recognition literature

- Word error rates are reported
- Few studies focus on error analysis
  - Major causes of errors (Duta et al., 2006; Adda-Decker, 2006; Nemoto et al., 2008; Goldwater et al., 2010; Dufour et al., 2012)
  - Error classification according to phonetic characteristics (Greenberg and Chang, 2000)
  - Comparison between automatic and human performances (Lippmann, 1997; Shen et al., 2008; Vasilescu et al., 2011)
- Few studies on ASR *error gravity* in link with further processing (for example for Spoken Document Retrieval (Woodland et al., 2000))

$\Rightarrow$ Experiment to study human judgments on error gravity

Related work and Error Gravity

Error evaluation in teaching

- Importance of the viewpoint on the error evaluation (Davies, 1983)
  - *Any error evaluation will be coloured by the particular viewpoint from which it is carried out, and thus may not be consistent with evaluation made from other viewpoints.*
- Effects of experience and first language on errors perception (Hyland and Anan, 2006)
  - comparison between three groups of teachers
  - *The relatively consistent patterns of recognition and decision making within each of the three groups compared with those between them suggests that fundamental differences inform their decisions based on their prior experiences.*

Our questions

Study of the judgments on a gravity scale of Error Zones

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

Data and Method

- Data: Three files from the French ETAPE corpus, ASR from LIUM
  - corpus 1: radio show debates (France Inter)
  - corpus 2: parliamentary debates (LCP, Top Questions)
  - corpus 3: radio show debates (France Inter)
- Error Zone: all the consecutive words in the hypothesis which are different from the reference.

REF: enfin en Seine Saint-Denis
HYP: ***** FRANSEN ***** ***** SANI

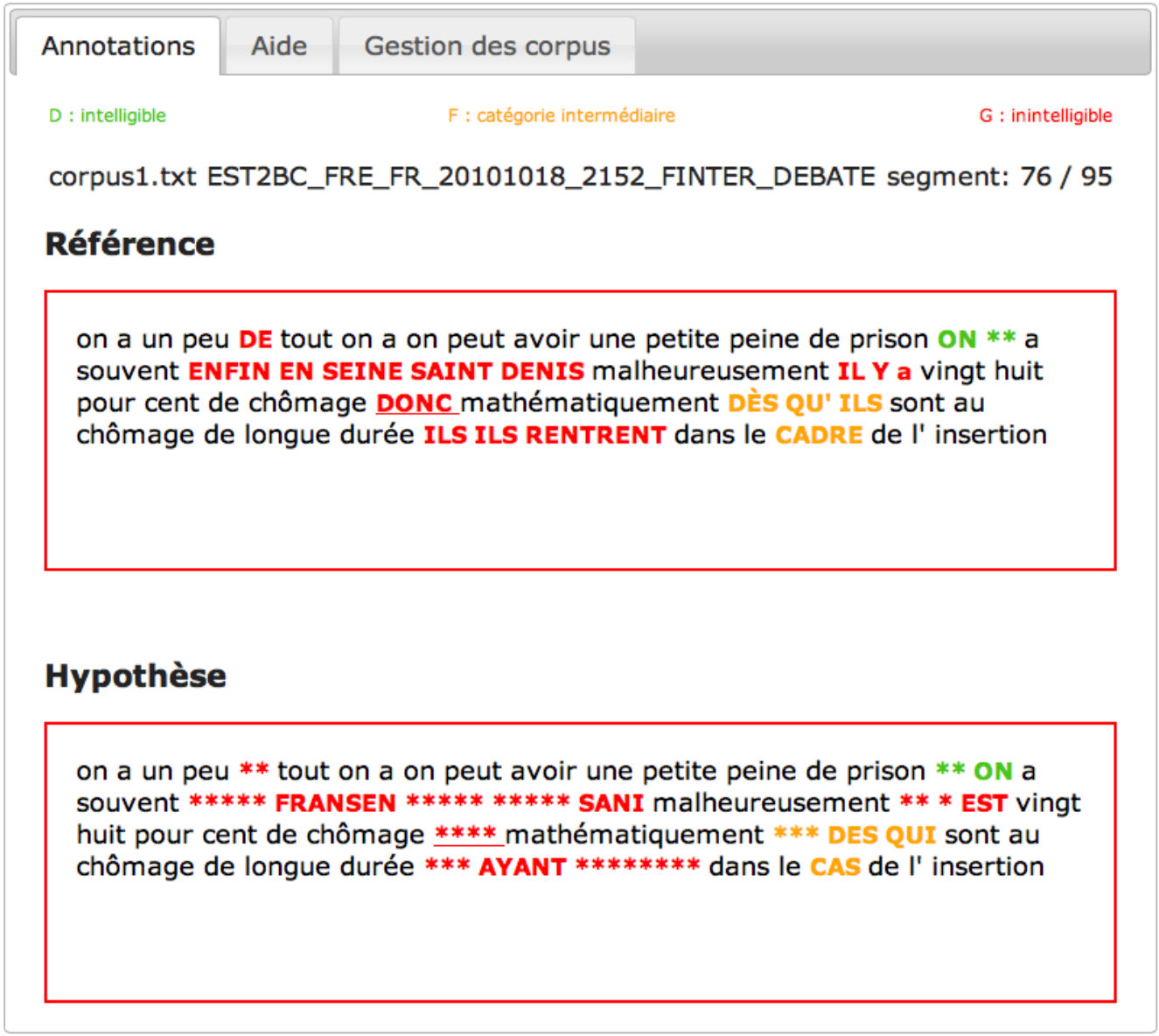
- Corpus

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

Data and Method

- Method
  - 7 annotators with distinct background
    - linguistics with NLP or SLP (a1, a2, a7)
    - linguistics (a4, a5)
    - computer sciences with ASR (a3) or SLP (a6)
  - Same corpus annotation order: corpus1, 2, 3, and then 1 again
    - → inter- and intra-annotator agreement scores
  - No precise guidelines
    - Low: *I still understand, I understand the same thing...*
    - High: *The text is too different and the idea too...*
    - Intermediate: *I don't know*
  - → Do we share a same definition/understanding of gravity?

Data and Method



First results: raw numbers

- Distribution of annotations per annotator: annotators differ in their error gravity assessment

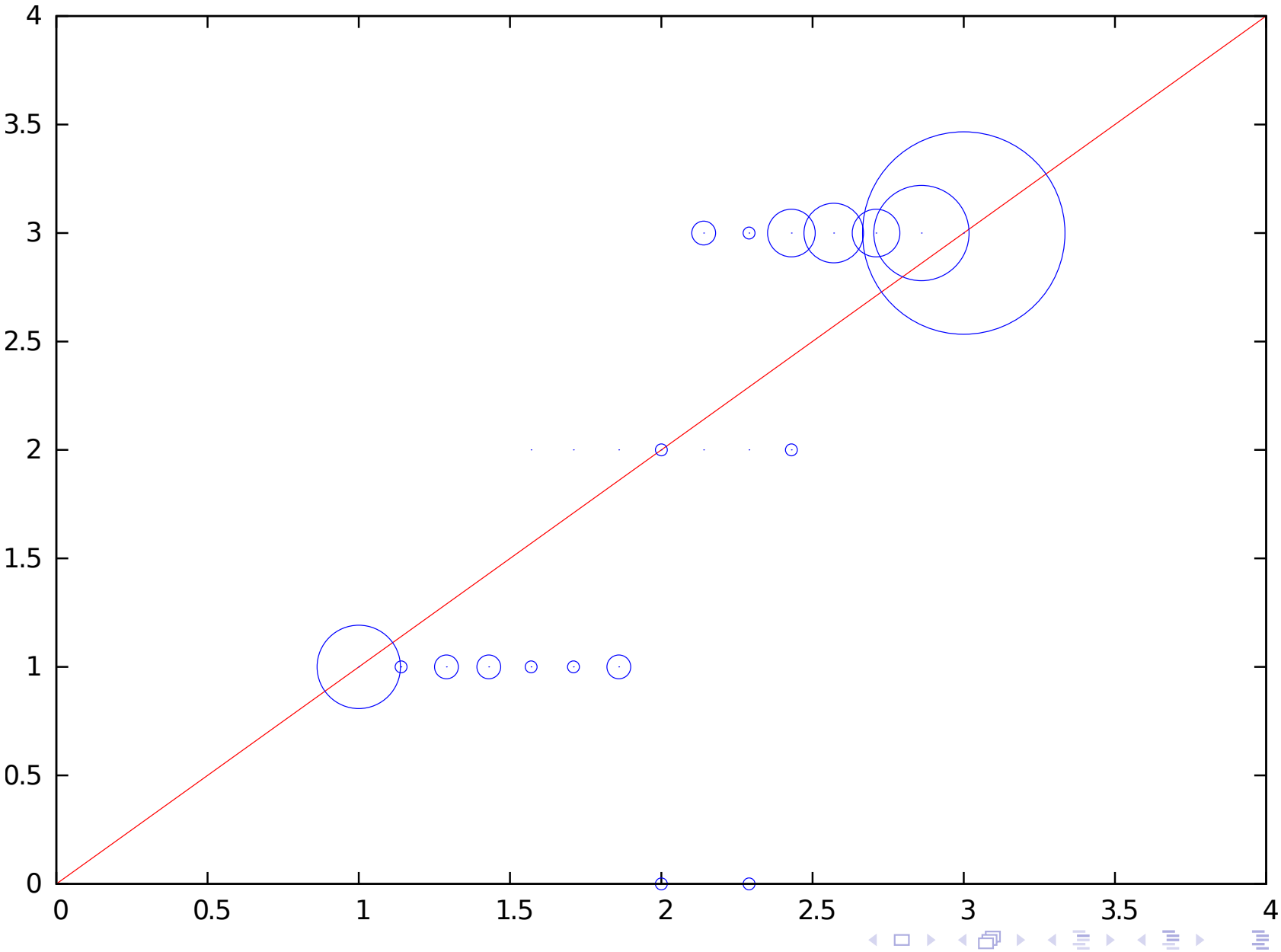
	Annotators						
Level	a1	a2	a3	a4	a5	a6	a7
Low	351	134	248	167	159	109	149
Interm.	121	45	187	143	38	38	104
High	216	509	253	378	491	541	435

- Distribution of perfect consensus in each level for gravity

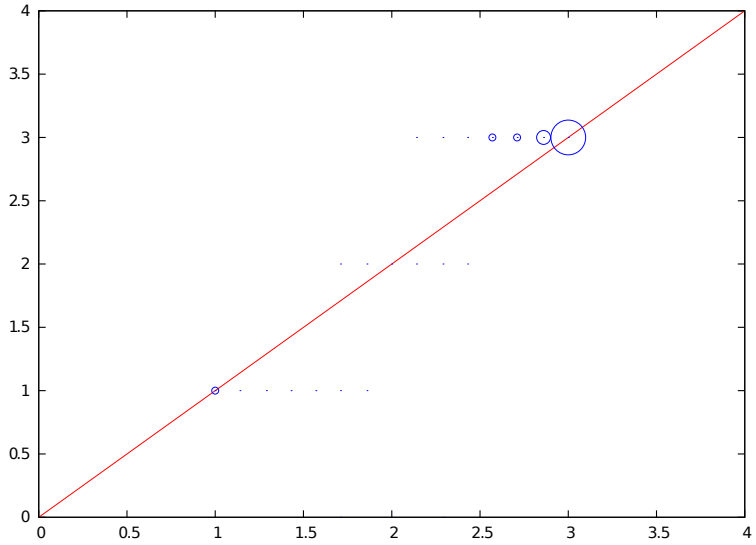
Corpus	Low level	Intermediate	High level	# EZ
Corpus 1	15	0	54	192
Corpus 2	34	0	6	94
Corpus 3	11	0	61	210
Corpus 4	17	0	53	192

- Low general kappa: 0.406

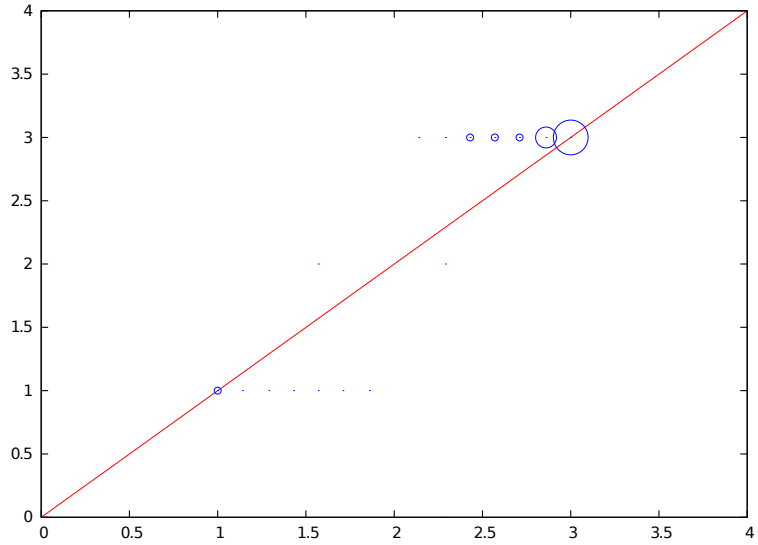
First results: mean annotation w.r.t. majority annotation



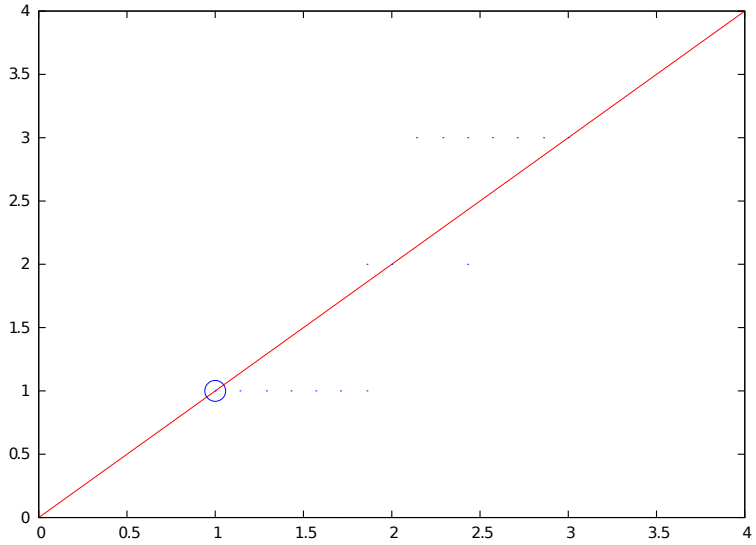
First results: mean annotation w.r.t. majority annotation



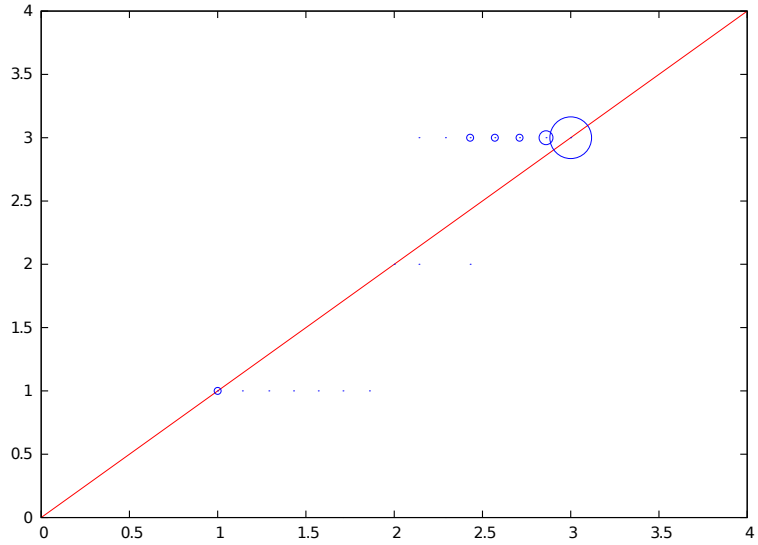
Corpus 1



Corpus 4



Corpus 2



Corpus 3

First results: majority judgment

Three classes:

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

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%

First results: background and tasks

- Confusion matrix: backgrounds does not seem to be important
  (a1, a2, a7), (a4, a5), (a3...a6)

	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	—

- Which task?
  - Understanding task:
    - a1 = slot filling issue
    - a4, a5, a6 = general understanding task
    - a7 = understanding for factual QA
    - a2, a3 = understandability/comprehensibility of the utterance

First results: intra-annotator agreement

- Do the annotators agree with themselves?

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

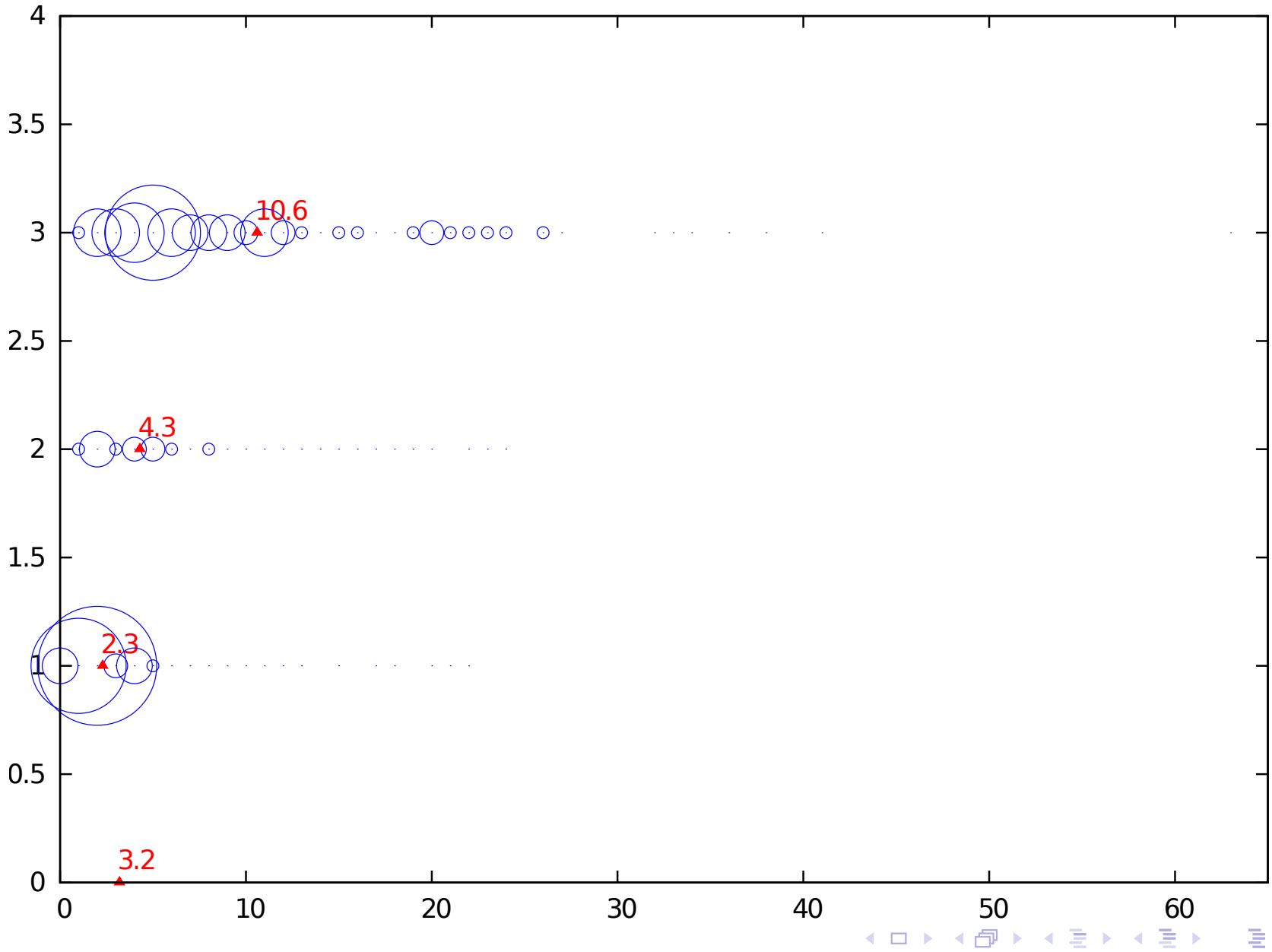
  - from 0.60 to 0.70
  - humans had difficulties classifying the error gravity

- Where are the differences?

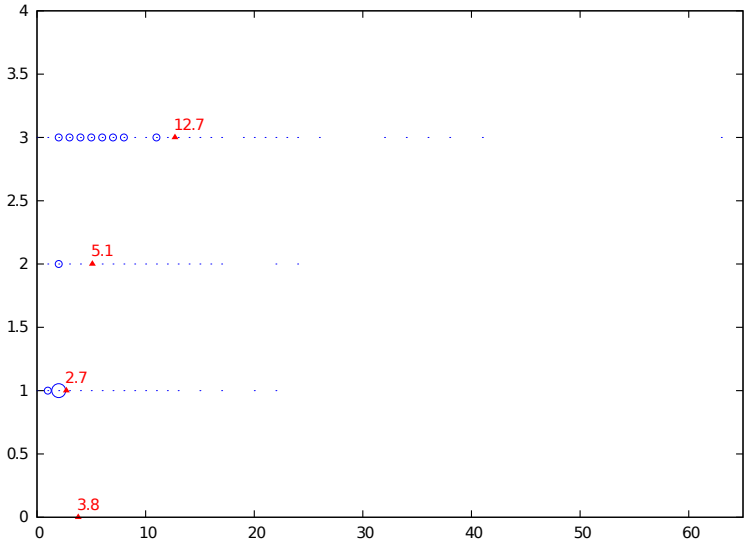
	a1	a2	a3	a4	a5	a6	a7
corpus1	L	H	I	**I**	**H**	H	I
corpus4	L	H	I	**H**	**I**	H	I

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

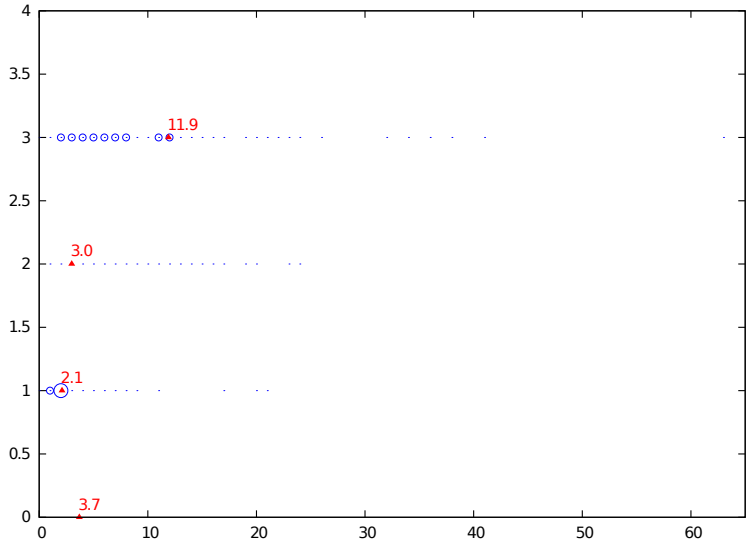
First results: annotation density w.r.t. edit distance



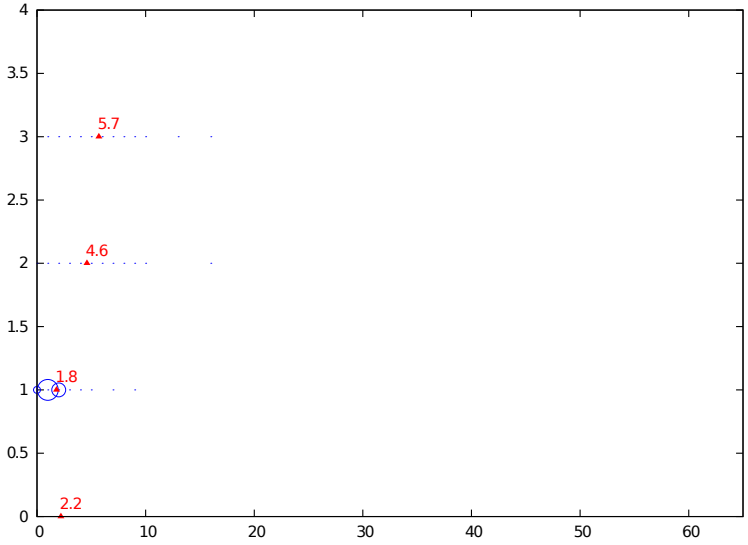
First results: annotation density w.r.t. edit distance



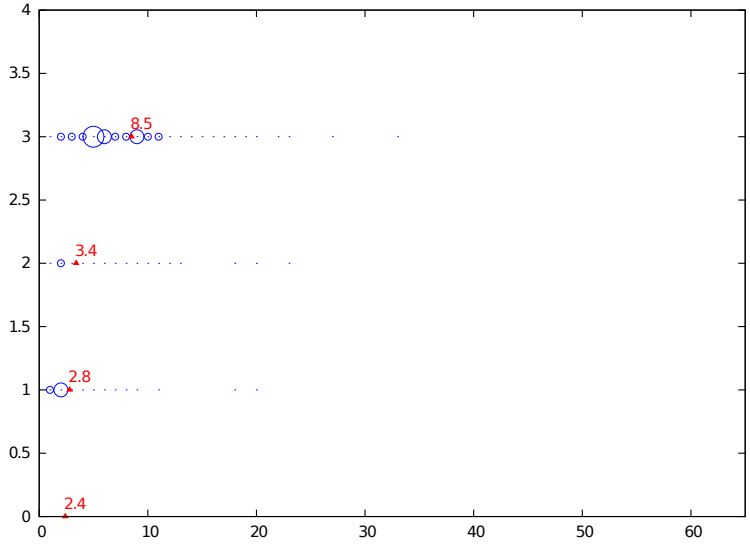
Corpus 1



Corpus 4



Corpus 2



Corpus 3

Examples of annotation and Edit Distance

- Edit Distance = 5
  - Majority on *low level*

    REF: vous chantez pas <EZ> j' ai dit </EZ> parce que j' ai
    you don't sing <EZ> I said </EZ> because I have

    HYP: vous chantez pas <EZ> ** je dis </EZ> parce que j' ai
    you don't sing <EZ> I say </EZ> because I have
  - Majority on *high level*

    REF: <EZ> ** oui </EZ>
    yes

    HYP: <EZ> un oeil </EZ>
    one eye
- Edit Distance = 2
  - Majority on *low level*

    REF: qui <EZ> retrouvent </EZ> un boulot
    who <EZ> find </EZ> a job / PLURAL

    HYP: qui <EZ> retrouve </EZ> un boulot
    who <EZ> find </EZ> a job / SINGULAR
  - Majority on *high level*

    REF: qui <EZ> Mona </EZ>

    HYP: qui <EZ> Monin </EZ>

Conclusion & Perspectives

- First conclusions
  - Each annotator has its own interpretation of the task
  - Not a strong agreement in general
  - But majority votes are found
  - Some expected observations
    - the higher the edit distance between REF and HYP, better the agreement
- Next steps
  - Take into account semantic or morpho-syntactic categories
  - Develop annotation guidelines
    - Examine all cases with a majority and define the error gravity
    - Integrate task knowledge
  - Work on other data: asr systems, tasks etc.
- Study human performance on specific tasks

References I

Adda-Decker, M. (2006).

De la reconnaissance automatique de la parole Ãă l'analyse linguistique de corpus oraux.
In *Proc. of JEP*, Dinard, France.

Davies, E. E. (1983).

Error evaluation: the importance of viewpoint.
ELT Journal, 37(4):304–311.

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., 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.

Greenberg, S. and Chang, S. (2000).

Linguistic dissection of switchboard-corpus automatic speech recognition systems.
Paris.

Hyland, K. and Anan, E. (2006).

TeachersâĂŹ perceptions of error: the effects of first language and experience.
System, 34(4):509–519.

References II

Lippmann, R. (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.

Vasilescu, I., Yahia, D., Snoeren, N., 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., Johnson, S., Jourlin, P., and Jones, K. S. (2000).
Effects of out of vocabulary words in spoken document retrieval.
In *Proc. of SIGIR*, pages 372–374.