# Minutes of the meeting ETAPE on Feb. 6, 2012.

A day dedicated to the ETAPE evaluation was organized on Feb. 6, 2012, at Telecom ParisTech. About 20 persons from various participating sites were present. The day was organized around talks by the organizers to foster discussion so as to:

1. clarify the evaluation plan, rules and protocols
2. get feedback from the participants
3. validate the evaluation plan and calendar

The following presentations were made:

- Introduction et campaign presentation (Guillaume Gravier, AFCP)
- Data and annotation process (Matthieu Carré, ELDA)
- Metrics and scoring software (Olivier Galibert, LNE)
- Phonetic alignments: first steps (Cédric Gendrot, LPP)
- Named entity annotation in Quaero  (Sophie Rosset, LIMSI)

Slides will be available on the ETAPE web site shortly, along with these minutes.

These minutes do not target a detailed description of the discussions that took place. The intention is rather to keep trace of the main points that were raised along with the corresponding decisions.

**About multiple submissions**. A rule inherited from previous evaluations stipulates that a site submitting multiple runs has to identify one of the systems as a primary system which will be considered for official ranking. This is an issue if multiple systems, with different approaches, are submitted as one would (rightfully) like to see all of these systems evaluated and officially ranked. We decide to adopt a different rule where sites are allowed to submit up to 4 runs/systems per task, all systems being ranked with the official metrics. The limit of 4 is there to avoid submissions of many variants with different parameter tweaking so as to choose a posteriori which was the best setting for the test set. Organizers will look at systems descriptions to search for such cases and treat them on a case by case basis.

**About SES-2 evaluation metric**. What metric(s) should be considered for SES-2? As SES-2 is an exploratory task, we will provide several metrics based either on time or on events. Time based metrics are somewhat more challenging but event based ones emphasize too much short overlapping speech regions (which will count the same as big ones). No official ranking of the systems will be made but we will analyze rankings across the various metrics considered. We recall that in all publications, the SES-2 task will be clearly identified as an exploratory one.

**About ASR outputs for entity detection**. What should be made available for the NE-ASR tasks? Organizers will provide a ROVER of all the TRS runs, with confidence measures. Additionally, CTM files will be made available from sites participating to the TRS task and willing to share their CTM (the CTM will be included in the final evaluation package distributed by ELDA with a blind name not identifying the site). **We are looking for sites willing to donate their CTM files and eventually word lattices or confusion networks!** A sample CTM file will be made available shortly for development purposes (mostly for format issues).

**About entity detection evaluation metric**. As of now, the metric is based on a mapping of the reference NE boundaries to the CTM file to be scored, the final SER result being computed on that CTM file regardless of transcription errors. This makes it difficult to compare in "an exact way" two systems with different ASR as the SER is computed on different CTM files. It was agreed (by almost everyone) that this inconvenience remains minor as long as the WERs for the two CTMs are reasonably not too different in their magnitude. It would be interesting in the long run to verify this claim though. Moreover, the current metric is somewhat insensitive to transcription errors, i.e., entities wrongly transcribed but detected as entities being counted as correct (recognizing Banana as a person name instead of Obama is considered as correct from the entity detection point of view). This is both a good thing and a bad thing, depending on what one wants to do (note that this is strongly related with the problem mentioned above of system comparison with different ASR). In the future, using as a complementary performance measure a metric that is sensitive to transcription errors would be helpful since in some application scenarios, both WER and SER impacts the application.

**About the speaker attributed scoring of the TRS task**. The TRS task will be evaluated according to two metrics: Optimal overlap alignment will search for the optimal mapping of hypothesized words to the reference in overlapping speech regions while speaker attributed scoring (as is done in the NIST RT evals) will assume that words are assigned to a particular speaker and make use of that mapping for scoring. Speaker attributed scoring poses some problems to sites not doing speaker diarization. The question is therefore whether all submissions to the TRS task should be considered for speaker attributed scoring or not. The primary metric will obviously be the optimal overlap alignment one. Results from sites not doing speaker diarization will not be considered for speaker attributed WER though we encourage sites to attach speaker tags to words.

It was also pointed out that the DGA data of ESTER are not yet included in ELDA's ESTER package, yet new participants did not receive these data. Matthieu will check at ELDA whether it is possible to "correct" the situation and send the DGA/ESTER data to CRIM.