## Resume of work with Boris

* ~~Reproducing BERT-KPE results on openkp - ok (with bigger kp > 5)~~
* ~~Reproducing BERT-KPE results on openkp20k - ? (kp up to 5)~~

### Discuss the data that we will use in production

* ~~- Recover the (5 dataset) original data - ok~~
* - Have a good way to present the statistics
  + ~~- jupyter notebook for analyzing data~~
  + - Jupyter notebooks results to wiki page on the repo ?
* - See if there are results on other papers about these datasets - ?
  + ~~- Find new SOTA~~
  + - Converge on a metric that we can use maybe to compare to other papers —> on progress
* - Converge to a pre-processing way to use these dataset and potential format Example is Antoine pivot format good enough ? —> on progress
  + ~~Have common format for all dataset~~
  + Code to combine dataset
* ~~- Find a way to store the pivot dataset (jsonl for example) using dvc~~
  + ~~- Using dvc on mlzilla and local laptop~~
  + ~~- Location drive, repo, google storage bucket ?~~

### Training and evaluation of the model

* - Discuss which data we want to use to train our model that will be used in production
  + ~~- Train and eval on each of the "free" dataset~~
  + - Train on one dataset and evaluate on openkp or kp20k test sets (Zero shot) —> on progress
  + - Think about a concatenation method to improve the results on the different test set
* ~~Have a repo github~~
  + ~~Structure the repo~~
  + ~~Combine Antoine code’s and original paper, make it works, clean up~~
* ~~- Keep track of all the trainings using mlflow (ask Boris)~~
  + ~~- Plug mlflow to torch training/code - ?~~
* - Update the wiki benchmark of your repo regularly - ?
* - Try to convert a pythorch model to Onnx - ?
  + - Have a look at the MachineLeanring/OnnxConversionLab repo - ?
  + - Have a way to use saved ONNX model and make prediction

### Production integration

* ~~- Have working Sinequa dev env - ?~~
* - Discuss with Emilie and Loïc about the specs... - ?
* - Have a look at other model configurations in Sinequa - ?
* - Start the integration

## TODO list for next time

* Run on all dataset and test on zero shot
* Preprocess data to concatenate dataset -> to check if there is improvement
* Refactor code to not depend on name data set
* Run [smart bert](https://github.com/victorywys/SMART-KPE), test on all dataset
* Read paper of SOTA
  + Check their metric
  + re -implement their metric in joint kpe
  + Find code or implement
* Script to change from checkpoint model pytorch to ONNX + script to import ONNX model and make prediction
* Test train by remove phrase with punctuation and stop words
* Few shot learning
* Write wiki page or repo