# Charte Éthique et Big Data: pour des corpus sans zone d'ombre

Karën Fort

karen.fort@loria.fr

# Charte Éthique et Big Data: pour des corpus sans zone d'ombre

Karën Fort

karen.fort@loria.fr

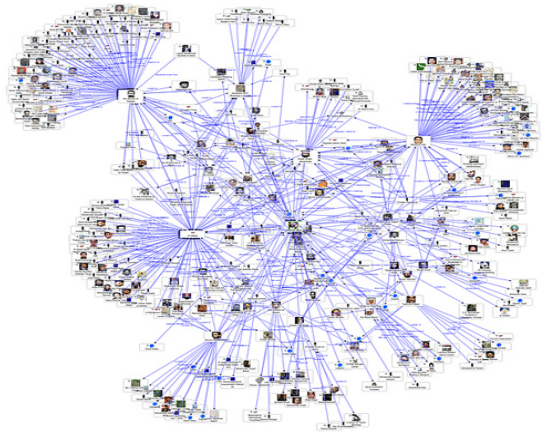
# Big Data ?

Définition (Decideo.fr) :

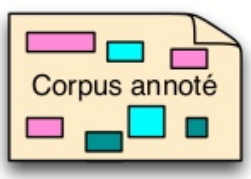
- Volume
- Vitesse de constitution
- Variété

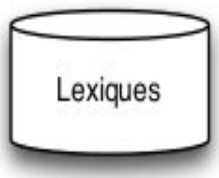
- Complexité

# Accéder aux Big Data



Outils TAL



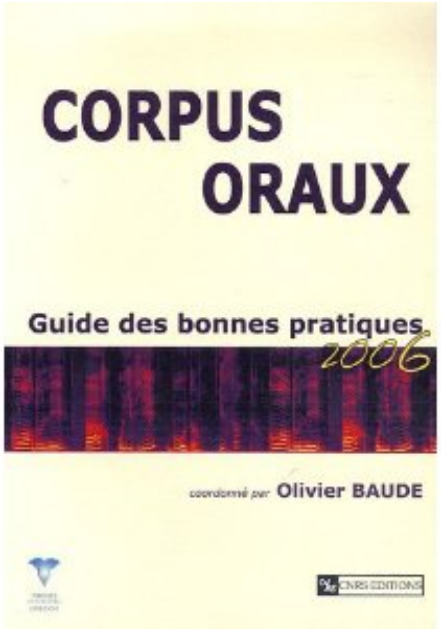



# Beautiful Data et Big Data

Big Data ou pas, mon corpus le vaut bien !

# Malgré d'importants efforts...

Bonnes pratiques générales :





AHDS Guides to Good Practice

**Developing Linguistic Corpora: a Guide to Good Practice**

Edited by Martin Wynne

Méta-données, parfois complexes :

# ... tout n'est pas (bien) documenté

Qui travaille ?

> *<profileDesc> provides a detailed description of non-bibliographic aspects of a text, specifically the languages and sublanguages used, the situation in which it was produced, the participants and their setting. (TEI)*

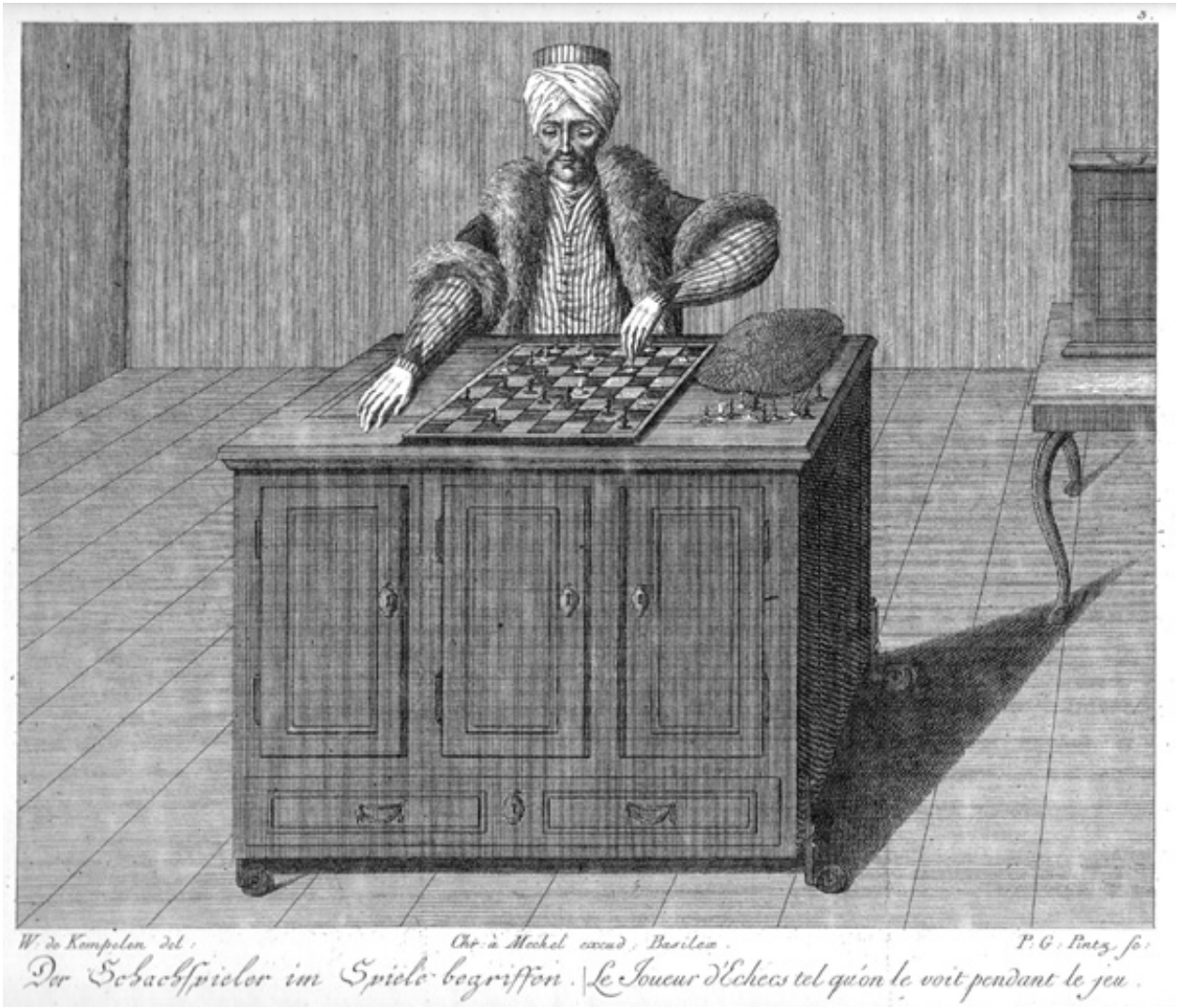
# ... tout n'est pas (bien) documenté

Qui travaille ?

> *<profileDesc> provides a detailed description of non-bibliographic aspects of a text, specifically the languages and sublanguages used, the situation in which it was produced, the participants and their setting. (TEI)*

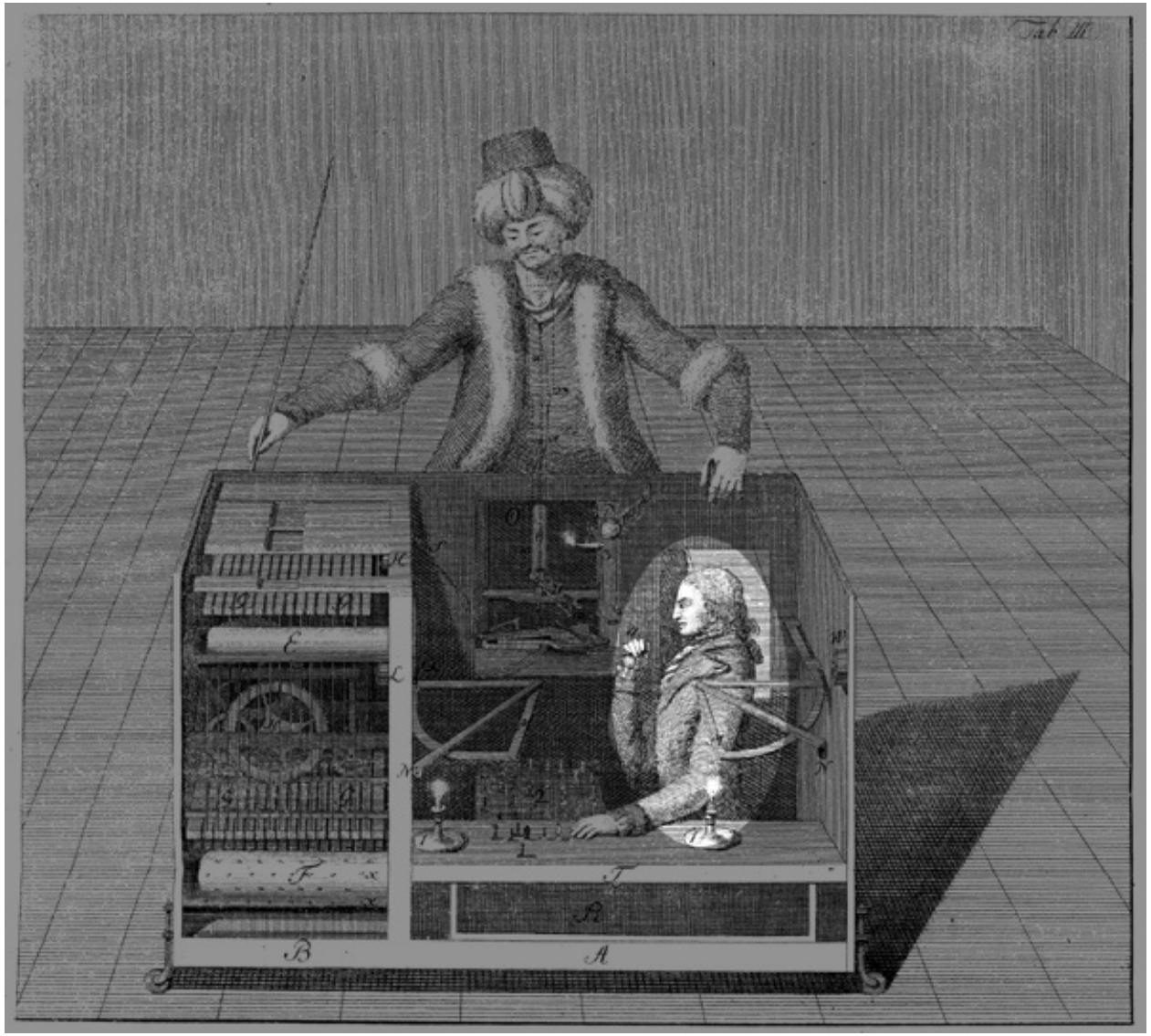
Mode de rémunération ?

# Turc mécanique




Johann Wolfgang von Kempelen (1770)

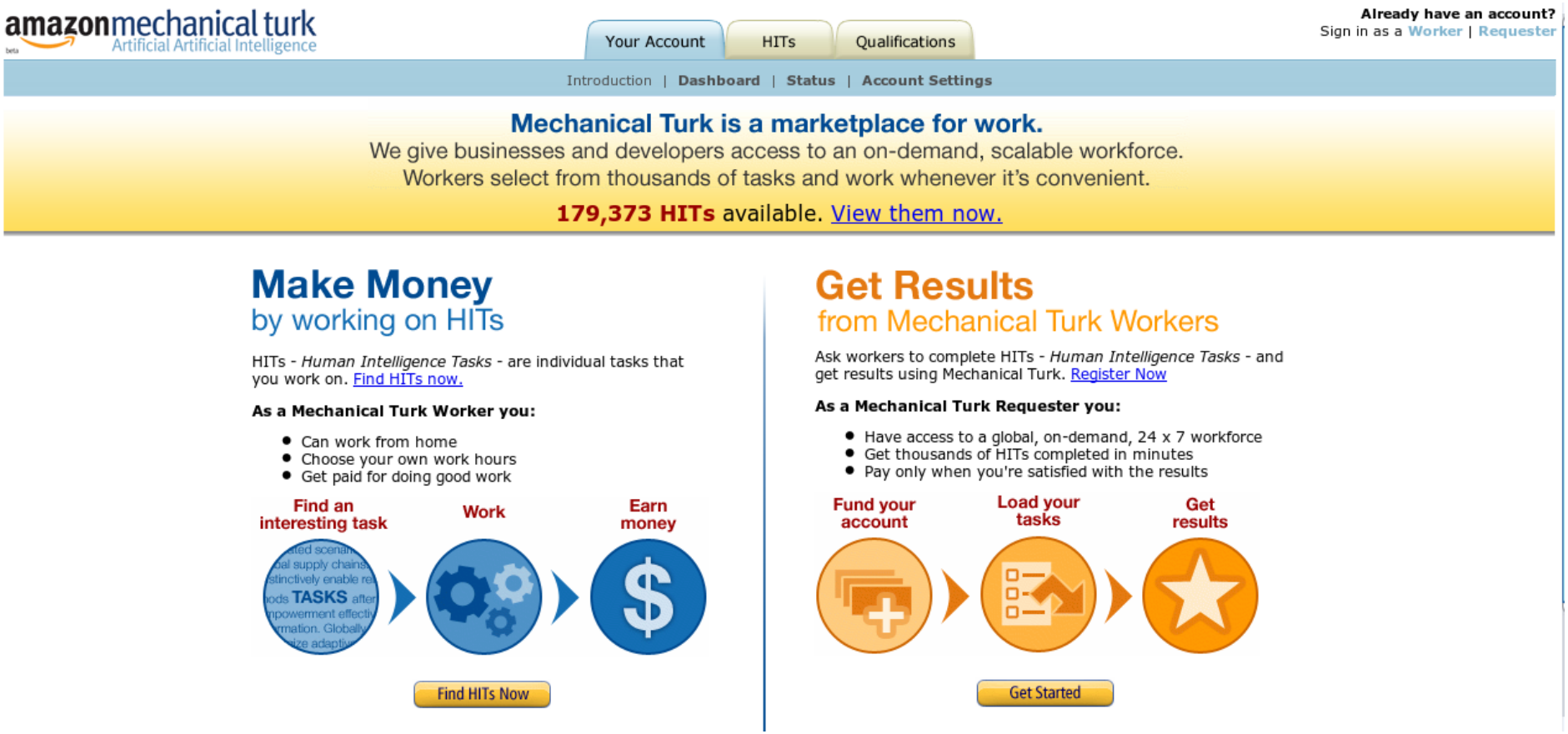
# Intelligence artificielle artificielle




Canular découvert en 1820

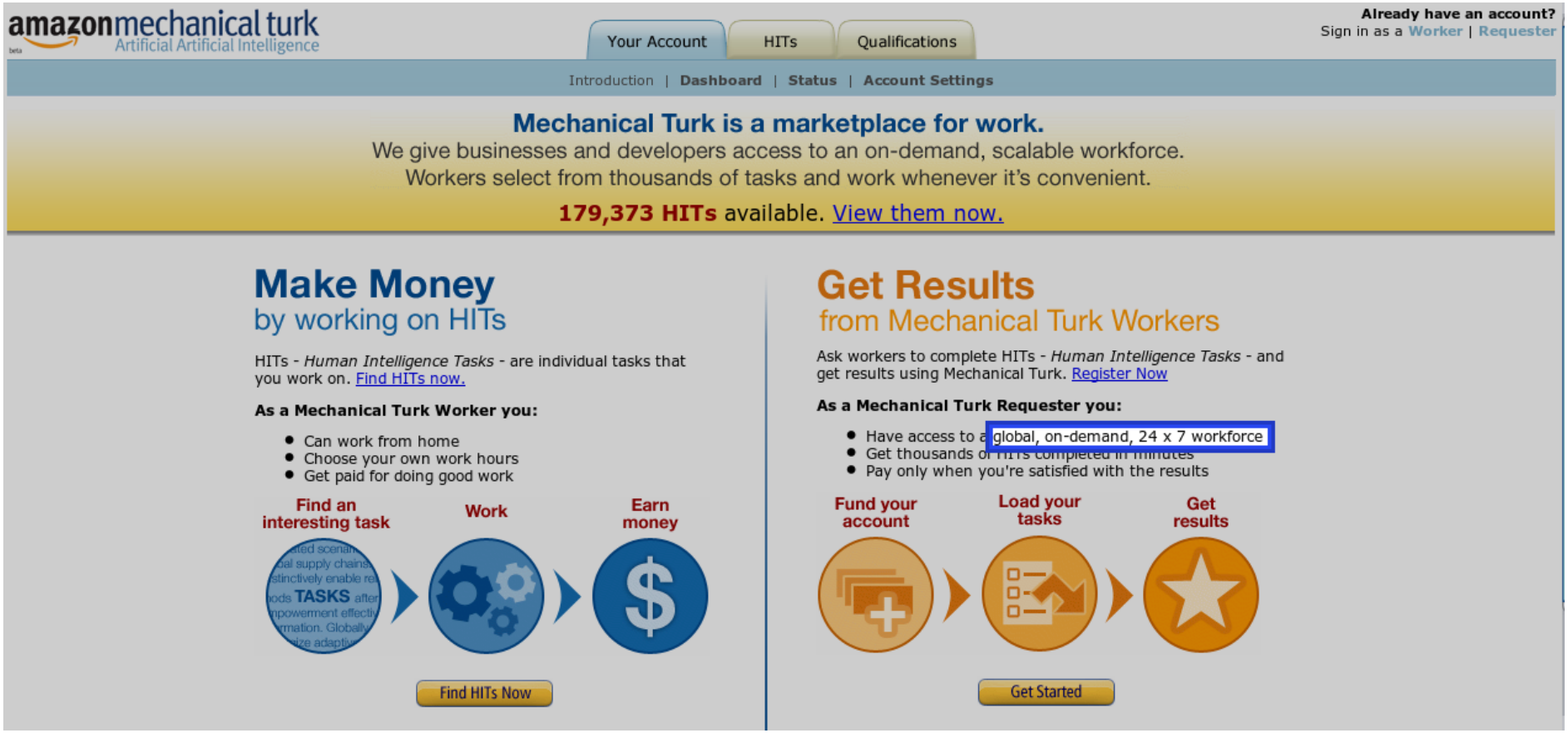
# Amazon Mechanical Turk

## AMT



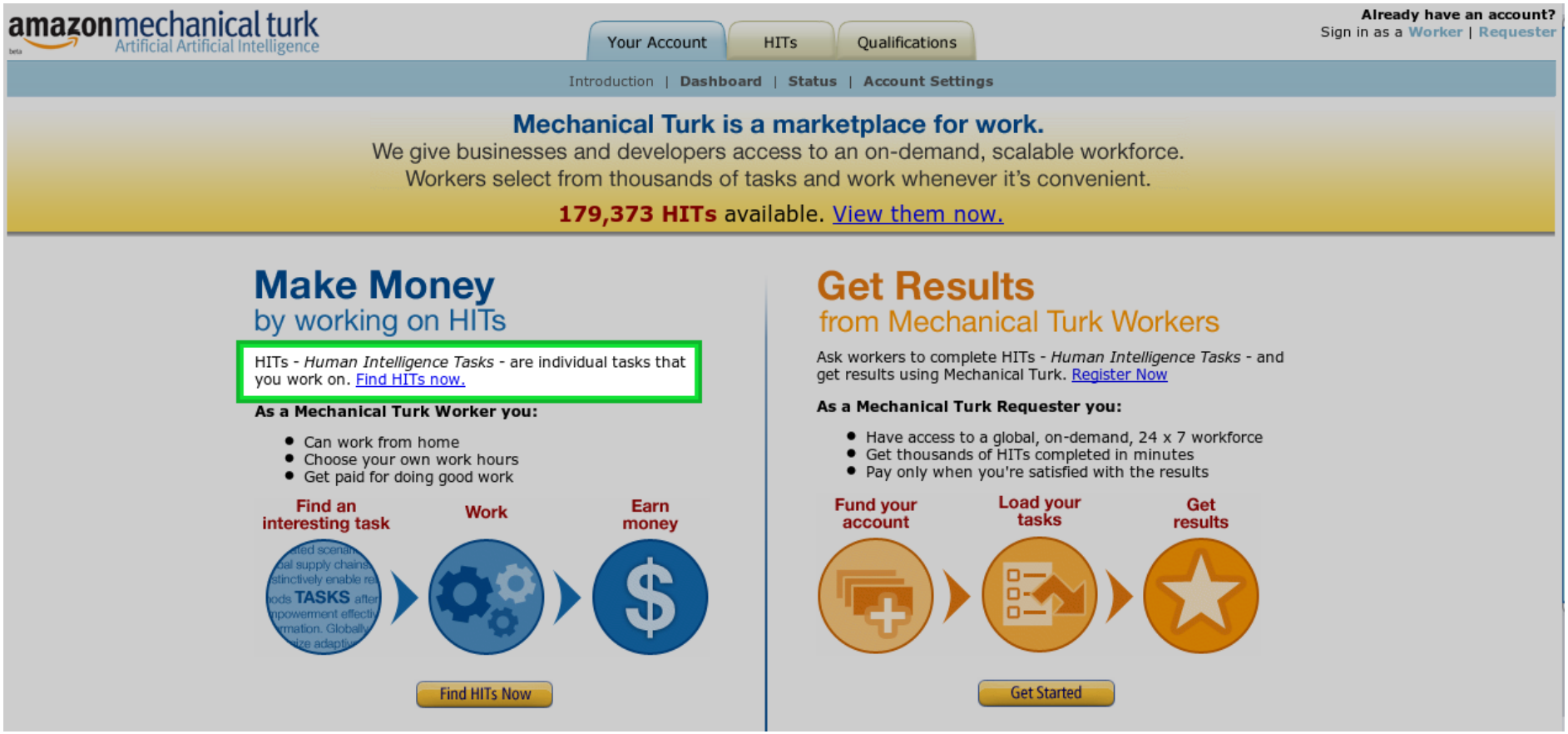
# Amazon Mechanical Turk

AMT est une plate-forme de crowdsourcing (myriadisation) : le travail est *délocalisé (outsourced)* via le Web, et réalisé par une *foule* (crowd) de personnes, appelées ***Turkers***



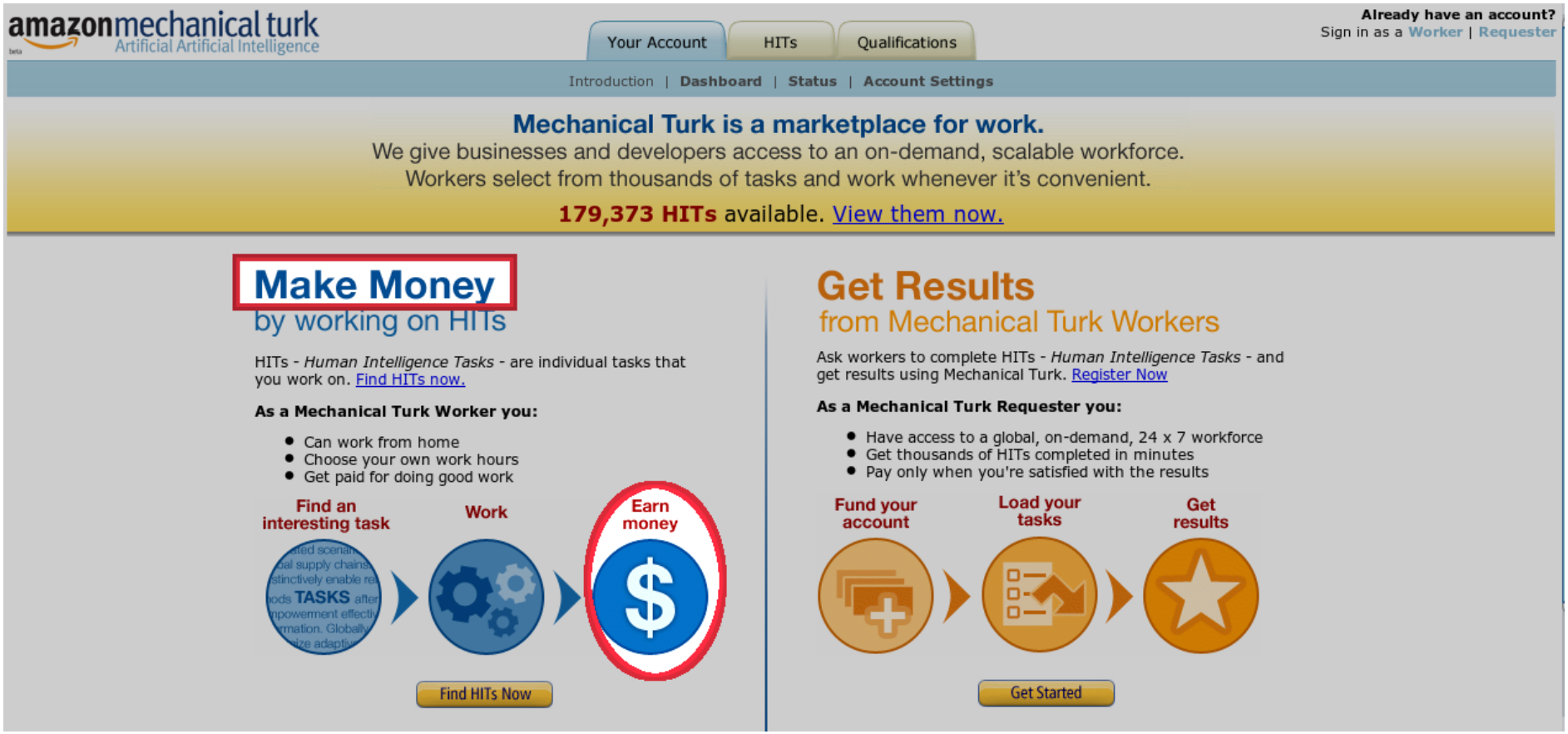
# Amazon Mechanical Turk

AMT est une plate-forme de crowdsourcing (myriadisation), et de microworking (travail parcellisé) : les tâches sont découpées en pièces (HITs) et leur exécution est rémunérée par les ***Requesters***

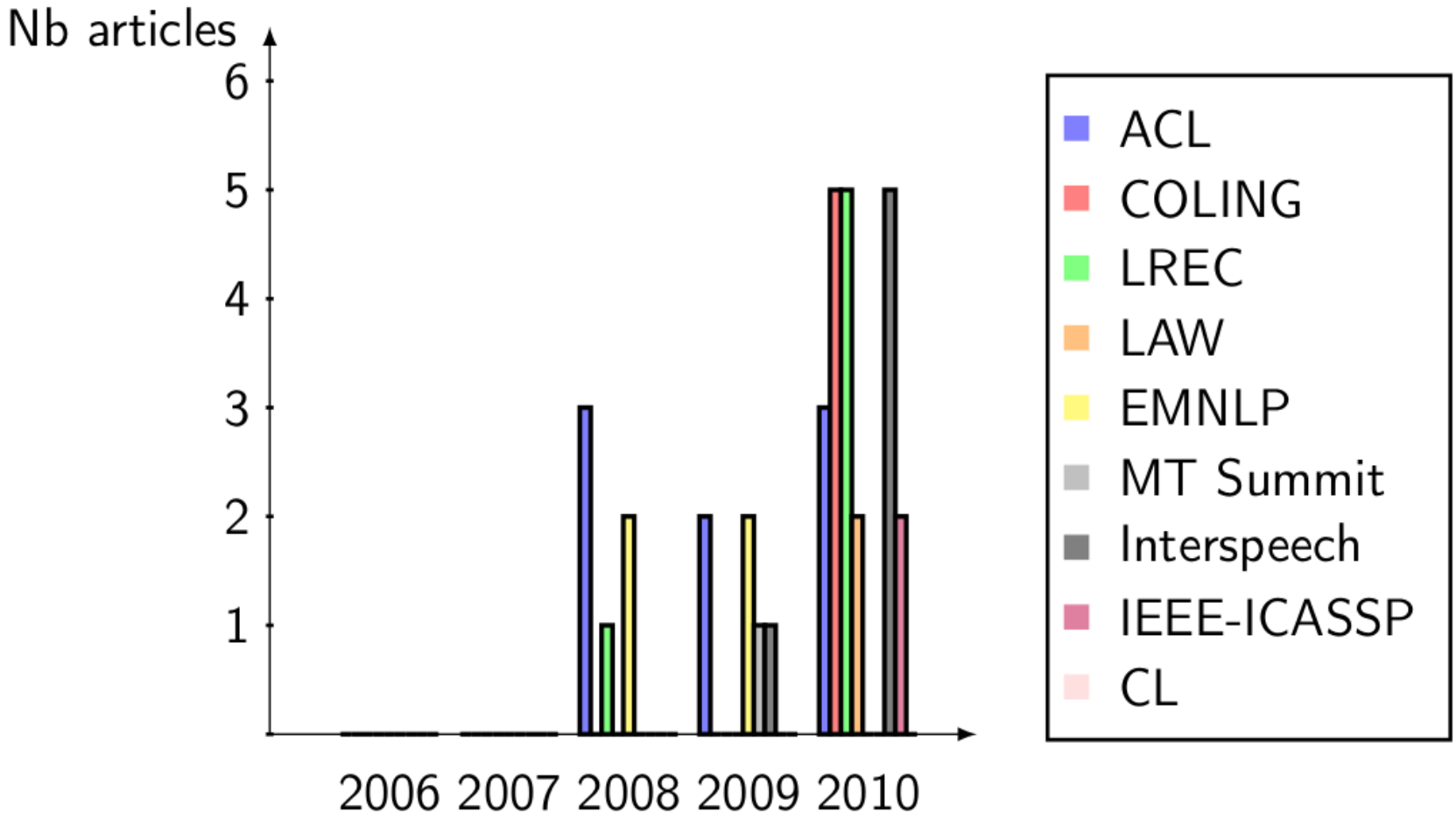


# Amazon Mechanical Turk

AMT est une plate-forme de crowdsourcing (myriadisation), et de microworking (travail parcellisé): les tâches sont découpées en pièces (HITs) et leur exécution est rémunérée.



# Amazon Mechanical Turk (AMT) et le TAL




Évolution de l'utilisation d'AMT dans les publications en TAL

# Amazon Mechanical Turk (AMT) : la légende

Très peu cher, rapide, de bonne qualité [Snow et al., 2008]

et un hobby pour les *Turkers* !

# Amazon Mechanical Turk (AMT) : la légende

Très peu cher, rapide, de bonne qualité [Snow et al., 2008]

et un hobby pour les *Turkers* !

Voyons...

# Combien de *Turkers* ?

[Fort et al., 2011]: si **500 000** personnes sont enregistrées comme *Turkers* sur la plate-forme, ils ne sont qu' entre 15 059 et 42 912 à participer.

# AMT : un hobby pour les *Turkers* ?

[Ross et al., 2010, Ipeirotis, 2010] montrent que :

- les *Turkers* sont avant tout motivés par **l'argent** (91 %):
  - pour 20 % AMT est la principale source de revenus ;
  - pour 50 % la deuxième source de revenus ;
  - c'est un loisir pour une minorité (30 %).
- 20 % des *Turkers* passent plus de 15 h/semaine sur AMT, et contribuent à 80 % des tâches.
- le salaire horaire moyen observé est de moins de 2 $ US.

# AMT et la qualité ?

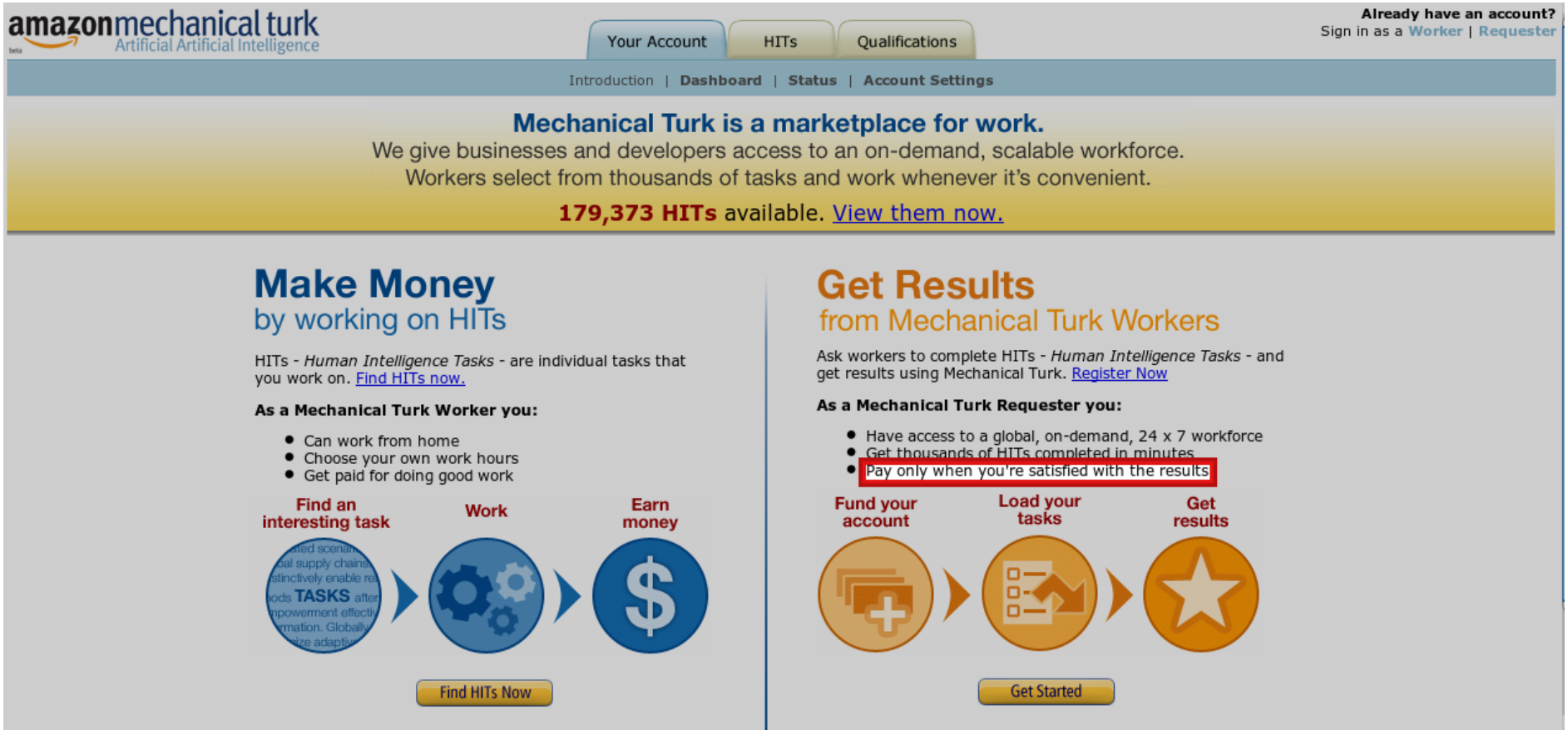
- Possibilité de produire des ressources de qualité dans certains cas précis (par ex. la transcription)
- Mais :
  - la qualité baisse lorsque la tâche devient complexe (par ex. le résumé [Gillick and Liu, 2010])
  - problèmes d'interface [Tratz and Hovy, 2010]
  - *Turkers* (tricheurs, spammers)
  - modèle de rémunération à la tâche (illégal en France) [Kochhar et al., 2010]
- Pour certaines tâches simples, les outils de TAL donnent de meilleurs résultats qu'AMT [Wais et al., 2010].

# Éthique et/ou légalité ?

Éthique :

- Aucune identification : aucune relation entre les *Requesters* et les *Turkers* et entre *Turkers*
- Aucune possibilité de créer un syndicat, pour protester ou ester en justice.
- Pas de salaire minimum ($< 2$ \$/h en moyenne)
- Possibilité de refuser de payer les *Turkers*

# Éthique et/ou légalité ?



# Éthique et/ou légalité ?

Légalité :

- accord de licence d'Amazon : les *Turkers* sont considérés comme des travailleurs indépendants $\Rightarrow$ ils sont supposés payer des cotisations.
- Illusoire, étant donné les salaires proposés.

$\Rightarrow$ Les états perdent une source de revenus légitime.

# Conséquences sur la production de ressources langagières

Cercle vicieux :

- Utilisation de systèmes *low cost* (à-la AMT) dans les projets ⇒
- Les agences de moyens constatent une énorme réduction des coûts ⇒
- Les agences de moyens deviennent réticentes à payer pour des projets développés ailleurs ⇒
- Les coûts AMT deviennent la norme ⇒
- D'autres systèmes, plus chers, disparaissent.

⇒ Nécessité d'une charte.

# Des rédacteurs et des contributeurs






# Une collaboration étroite

1 réunion par mois, de **juin à décembre 2012**

une **validation** par chaque organisme

# Une charte auto-déclarée

Formulaire destiné à accompagner le dossier de demande de financement préalable.

# Points clefs

- **Traçabilité** : historique du corpus
- **Qualité** : moyens mis en œuvre pour assurer la qualité de la ressource
- **Éthique** : statut et le mode de rémunération des personnes
- **Aspects juridiques** : licence et législation

# But

Faire adopter la charte par les agences de moyens, afin que celles-ci puissent définir une politique de sélection.

# TCOF-POS [Benzitoun et al., 2012]

À partir de TCOF (Traitement de Corpus Oraux en Français) :

- corpus de parole spontannée
- transcription sans ponctuation ni majuscules
- avec Transcriber

TCOF-POS :

- annotation en morpho-syntaxe
- correction de pré-annotations
- deux annotatrices + un validateur
- sur un tableur
- méthodologie : calcul régulier de l'accord inter-annotateurs

# TCOF-POS : extrait

| | | |
|---|---|---|
| L2 | LOC | L2 |
| ok | FNO | ok |
| L3 | LOC | L3 |
| il | PRO:clsi | il |
| y | PRO:clo | y |
| aura | VER:futu | avoir |
| il | PRO:clsi | il |
| y | PRO:clo | y |
| aura | VER:futu | avoir |

# TCOF-POS : détails de la charte

# Charte pour TCOF-POS : rédaction

Un processus (relativement) léger :

- 2h de travail : A. Couillault (Aproged) et K. Fort
- révision : C. Benzitoun

Une documentation à compléter :

- Premier exemple fait et disponible
- Aide des rédacteurs prévue

# Charte : disponibilité et adoption

La charte :

- a été annoncée lors de Documation mercredi dernier
- est en cours de relecture par un juriste

est disponible :

- sous forme de Wiki (`http://wiki.ethique-big-data.org/`)
- sous forme de document pdf
- en anglais

et adoptée par Cap Digital

# Perspectives

Forme :

- proposer un formulaire pour générer sa charte en ligne
- proposer des instanciations “types” de la charte
- ajouter des éléments de pédagogie (exemples, suggestions, liens, etc.)

Contacts plus ou moins avancés pour une adoption par :

- l’ANR
- la DGLFLF

La charte devrait être un **critère de sélection** aussi pour l’intégration dans le corpus de référence.


Baude, O., Blanche-Benveniste, C., Calas, M.-F., Cappeau, P., Cordereix, P., Goury, L., Jacobson, M., De Lamberterie, I., Marchello-Nizia, C., and Mondada, L. (2006).
*Corpus oraux, guide des bonnes pratiques 2006*.
CNRS Editions, Presses Universitaires Orléans.


Benzitoun, C., Fort, K., and Sagot, B. (2012).
TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe.
In *Actes de Traitement Automatique des Langues Naturelles (TALN)*, pages 99–112, Grenoble, France.


Fort, K., Adda, G., and Cohen, K. B. (2011).
Amazon Mechanical Turk: Gold mine or coal mine?
*Computational Linguistics (editorial)*, 37(2):413–420.


Gillick, D. and Liu, Y. (2010).
Non-expert evaluation of summarization systems is risky.

In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 148–151, Stroudsburg, PA, USA. Association for Computational Linguistics.


Ipeirotis, P. (2010).
The new demographics of mechanical turk.
http://behind-the-enemy-lines.blogspot.com/2010/03/new-demographics-of-mechanical-turk.html.


Kochhar, S., Mazzocchi, S., and Paritosh, P. (2010).
The anatomy of a large-scale human computation engine.
In *Proceedings of Human Computation Workshop at the 16th ACM SIKDD Conference on Knowledge Discovery and Data Mining, KDD 2010*, Washington D.C.


Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., and Tomlinson, B. (2010).
Who are the crowdworkers?: shifting demographics in mechanical turk.

In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, CHI EA '10, pages 2863–2872, New York, NY, USA. ACM.

Snow, R., O'Connor, B., Jurafsky, D., and Ng., A. Y. (2008).
Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks.
In *Proceedings of EMNLP 2008*, pages 254–263.

Tratz, S. and Hovy, E. (2010).
A taxonomy, dataset, and classifier for automatic noun compound interpretation.
In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 678–687, Uppsala, Suède. Association for Computational Linguistics.

Wais, P., Lingamneni, S., Cook, D., Fennell, J., Goldenberg, B., Lubarov, D., Marin, D., and Simons, H. (2010).
Towards building a high-qualityworkforce with mechanical turk.

In *Proceedings of Computational Social Science and the Wisdom of Crowds (NIPS)*.

Wynne, M., editor (2005).
*Developing Linguistic Corpora: a Guide to Good Practice*.
Oxford: Oxbow Books.