# Experts ou (foule de) non-experts ? la question de l'expertise des annotateurs vue de la myriadisation (crowdsourcing)

Karën Fort

# Experts ou (foule de) non-experts ? la question de l'expertise des annotateurs vue de la myriadisation (*crowdsourcing*)

Karën Fort
Université Paris-Sorbonne / EA STIH, Paris, France
karen.fort@paris-sorbonne.fr

**Résumé.** L'annotation manuelle de corpus est de plus en plus réalisée par myriadisation : produite par une masse de gens, *via* le Web, bénévolement ou à très faible coût. Nos expériences remettent en question une vision communément répandue : la myriadisation ne consiste pas à faire produire des données à une foule de non-experts, mais plutôt à identifier des experts de la tâche (en l'occurrence, d'annotation) dans la foule. Elles contribuent en ce sens à la réflexion sur l'expertise des annotateurs de corpus.

**Abstract.**
**Experts or (crowd of) non-experts ? the question of the annotators' expertise viewed from crowdsourcing.**
Manual corpus annotation is more and more performed using crowdsourcing : produced by a crowd of persons, through the Web, for free, or for a very small remuneration. Our experiments question the commonly accepted vision : crowdsourcing a task is not having a crowd of non-experts performing it, but rather identifying experts of the (annotation) task in the crowd. Those experiments therefore contribute to the reflection on the corpus annotators' expertise.

**Mots-clés :** annotation de corpus, experts, myriadisation.

**Keywords:** corpus annotation, experts, crowdsourcing.

## 1 Le flou de l'expertise des annotateurs

La question de l'expertise des annotateurs revient régulièrement dans les articles de recherche traitant d'annotation manuelle, mais elle est finalement peu considérée en tant que telle. Les auteurs parlent d'annotateurs « experts », les opposant parfois à des annotateurs « naïfs » (voir, par exemple (Péry-Woodley *et al.*, 2011) ou (Tellier, 2014)), sans pour autant définir ces termes, sans doute parce qu'ils semblent évidents. Parfois, le détail de la formation des annotateurs est donné (Péry-Woodley *et al.*, 2011) ou il est précisé qu'ils sont des « experts du domaine » (Candito *et al.*, 2014).

Cette ambiguïté de la définition de l'expertise des annotateurs est évidente dans les domaines de spécialité, comme le biomédical. Un débat nourri a ainsi eu lieu début octobre 2012 sur la liste de diffusion BioNLP [1] concernant le type d'expert le plus efficace pour annoter des éléments « linguistiques » (ou pour le traitement automatique des langues) dans un corpus biomédical, par exemple des entités nommées (noms de protéines, de gènes) : un spécialiste en biomédical ou en linguistique ?

Pour illustrer la question, considérons l'exemple suivant, provenant du corpus *Sequoia* (Candito & Seddah, 2012), annoté en syntaxe :

> Pour les SCA, la durée de la perfusion dépend de la manière dont le SCA doit être traité : elle peut durer jusqu'à 72 heures au maximum chez les patients devant recevoir des médicaments.

Que serait un expert dans ce cas ? Le sous-corpus (*EMEA*) est du domaine de la pharmacologie et l'annotation est de type linguistique (syntaxe). Faudrait-il un linguiste ? Un pharmacien ? Est-ce qu'un locuteur du français, sans connaissances en syntaxe ou en pharmacologie, mais formé à la tâche, pourrait annoter ce type de phrase ? et avec quelle qualité ? Serait-il alors un expert ? un non-expert ?

1. Le fil de discussion avait pour titre : « Trends in Clinical NLP (Jon Patrick) ».

Il est selon nous fondamental de distinguer (i) l'expertise du domaine du corpus (ici, la pharmacologie), (ii) celle du domaine de l'annotation (ici, la syntaxe) et (iii) celle de la tâche (ici, annoter des relations syntaxiques avec tel ou tel outil, selon tel guide d'annotation). Nous montrons dans les sections suivantes une illustration de l'utilité de cette distinction dans la myriadisation pour l'annotation de corpus.

## 2 L'annotation de corpus par myriadisation

La myriadisation [2] est l'activité qui consiste à faire produire (des annotations sur un corpus, un dessin, un vote,...) à une masse de gens, aujourd'hui principalement *via* le Web, bénévolement ou à très bas prix. Étant donné le coût très élevé de l'annotation manuelle [3], on peut s'attendre à ce que la myriadisation devienne une méthode de production de données privilégiée par les agences de moyens. Il est donc urgent de mieux comprendre ce phénomène pour mieux le maîtriser.

Il existe différents types de myriadisation, que l'on peut considérer selon des taxinomies variées (voir (Geiger *et al.*, 2011)). Nous proposons de prendre en compte comme critères (i) la rémunération (ou non) et (ii) la connaissance (ou non) de la donnée produite. Cela nous permet de distinguer la myriadisation de type bénévole et en connaissance, comme `Wikipédia` ou `le projet Gutenberg`, les « jeux ayant un but » (*Games With a Purpose*, GWAP), qui restent bénévoles mais ne dévoilent pas forcément de manière immédiate le type de données produit, et la myriadisation du travail parcellisé, autrement dit le travail (micro-rémunéré) dont la production est évidente, comme sur la plate-forme `Amazon Mechanical Turk` (`AMT`) [4].

Si `AMT` prétend depuis longtemps proposer les services de plus de 500 000 travailleurs [5], les *Turkers* (travailleurs) vraiment actifs n'étaient en 2010 qu'entre 15 059 et 42 912 (Fort *et al.*, 2011). Le jeu `Phrase Detectives`, dont le but est d'annoter des relations anaphoriques dans des textes anglais, a quant à lui réuni plus de 2 000 joueurs en un an, mais 13 d'entre eux ont produit la majorité des données annotées (voir Figure 1). Dans tous les cas observés, peu de participants produisent beaucoup de données. La foule (*crowd*) n'est donc pas celle qu'on imagine.

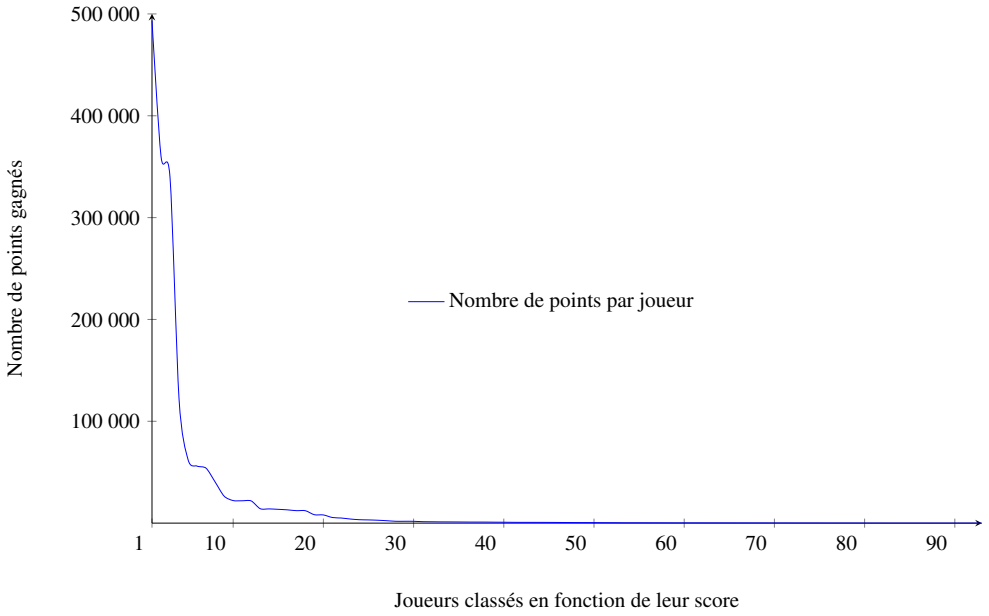


FIGURE 1 – Nombre de joueurs sur `Phrase Detectives` en fonction de leur classement en points (février 2011 - février 2012) (Chamberlain *et al.*, 2013)

2. Cette traduction de *crowdsourcing* a été proposée par Gilles Adda dans (Sagot *et al.*, 2011). Nous la reprenons ici à notre compte.
3. L'annotation du *Prague Dependency Treebank* a ainsi été évaluée à 600 000 dollars (Böhmová *et al.*, 2001).
4. Cette dernière pose de nombreux problèmes, dont des problèmes d'éthique, que nous avons mis au jour dans (Fort *et al.*, 2011).
5. Voir `https://requester.mturk.com/tour`.

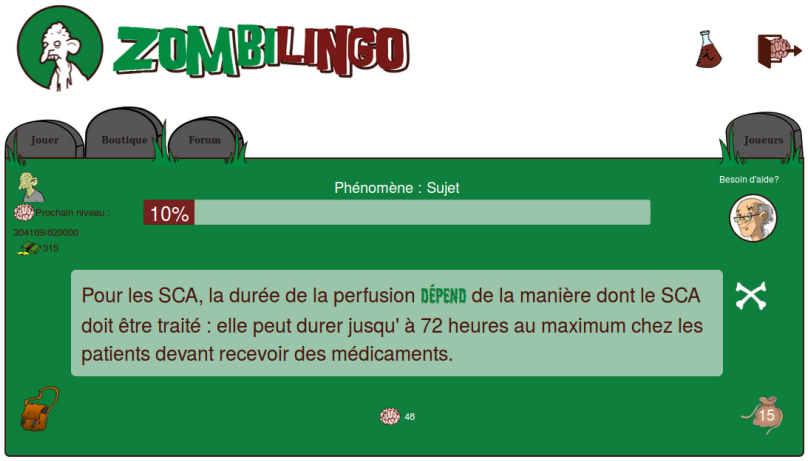


FIGURE 2 – Annotation d'un phénomène avec `Zombilingo` (Fort *et al.*, 2014)

## 3 Redéfinir la myriadisation, élargir l'expertise

La qualité de la production des participants aux jeux ayant un but, quand elle est évaluable, est remarquable [6]. Ainsi, l'accord inter-annotateur entre la référence et les joueurs sur `Phrase Detectives` est proche de 84 % d'accord observé. Ces chiffres sont très proches de ceux que nous obtenons sur le jeu `Zombilingo` (plus de 86 % d'exactitude), qui permet d'annoter des corpus du français en syntaxe de dépendances (voir Figure 3).

Si l'expertise des participants est donc rarement celle de spécialistes du domaine ou de l'annotation portée, les joueurs, les travailleurs ou les bénévoles deviennent rapidement des experts **de la tâche**, mettant en œuvre toutes sortes de stratégies et créant même leurs propres outils, afin d'être plus efficaces (Gupta *et al.*, 2014).

L'analyse de cas réels montre donc que myriadiser une production de données, notamment des corpus annotés, ne signifie pas faire annoter une foule de non-experts mais trouver et former des experts –de la tâche– dans la foule.

## Références

BÖHMOVÁ A., HAJIČ J., HAJIČOVÁ E. & HLADKÁ B. (2001). The prague dependency treebank : Three-level annotation scenario. In A. ABEILLÉ, Ed., *Treebanks : Building and Using Syntactically Annotated Corpora*. Kluwer Academic Publishers.

CANDITO M., PERRIER G., GUILLAUME B., RIBEYRE C., FORT K., SEDDAH D. & DE LA CLERGERIE E. (2014). Deep syntax annotation of the sequoia french treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Islande.

CANDITO M. & SEDDAH D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Actes de Traitement Automatique des Langues Naturelles (TALN)*, Grenoble, France.

CHAMBERLAIN J., FORT K., KRUSCHWITZ U., LAFOURCADE M. & POESIO M. (2013). Using games to create language resources : Successes and limitations of the approach. In I. GUREVYCH & J. KIM, Eds., *The People's Web Meets NLP*, Theory and Applications of Natural Language Processing, p. 3–44. Springer Berlin Heidelberg.

FORT K., ADDA G. & COHEN K. B. (2011). Amazon Mechanical Turk : Gold mine or coal mine ? *Computational Linguistics (editorial)*, **37**(2), 413–420.

FORT K., GUILLAUME B. & CHASTANT H. (2014). Creating Zombilingo, a Game With A Purpose for dependency syntax annotation. In *Gamification for Information Retrieval (GamifIR'14) Workshop*, Amsterdam, Pays-Bas.

GEIGER D., SEEDORF S., SCHULZE T., NICKERSON R. C. & SCHADER M. (2011). Managing the crowd : Towards a taxonomy of crowdsourcing processes. In *AMCIS 2011 Proceedings*.

6. La qualité des productions sur `AMT` est très variable et médiocre dès que les tâches sont un tant soit peu complexes.

GUPTA N., MARTIN D., HANRAHAN B. V. & O'NEILL J. (2014). Turk-life in india. In *Proceedings of the 18th International Conference on Supporting Group Work*, GROUP '14, p. 1–11, New York, NY, USA : ACM.

PÉRY-WOODLEY M.-P., AFANTENOS S., HO-DAC L.-M. & ASHER N. (2011). La ressource ANNODIS, un corpus enrichi d'annotations discursives. *Traitement Automatique des Langues*, **52**(3), 71–101.

SAGOT B., FORT K., ADDA G., MARIANI J. & LANG B. (2011). Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé. In *Actes de Traitement Automatique des Langues Naturelles (TALN)*, Montpellier, France. 12 pages.

TELLIER M. (2014). Quelques orientations méthodologiques pour étudier la gestuelle dans des corpus spontanés et semi-contrôlés. *Discours*, **15**.