

Jouir de la foule est un art [Baudelaire, « Les foules »]

Aspects méthodologiques et éthiques de l'annotation manuelle de corpus par myriadisation

Karën Fort

karen.fort@paris-sorbonne.fr

6 mars 2017



Licence et outils utilisés pour cette présentation



L^AT_EX

*"... les questions que soulève la science du langage, dans toutes ses versions, sont des questions fines : une proposition de linguistique concerne peu de données à la fois et elle y fait apparaître généralement ce que l'opinion courante tiendrait pour des détails. **En bref, il paraît peu vraisemblable qu'à partir de la science du langage, on puisse développer l'équivalent d'une sidérurgie ou d'une aéronautique**"*

[Milner, 1989]

*" . . . les questions que soulève la science du langage, dans toutes ses versions, sont des questions fines : une proposition de linguistique concerne peu de données à la fois et elle y fait apparaître généralement ce que l'opinion courante tiendrait pour des détails. **En bref, il paraît peu vraisemblable qu'à partir de la science du langage, on puisse développer l'équivalent d'une sidérurgie ou d'une aéronautique**"*

[Milner, 1989]



Traitement Automatique des Langues (TAL) et annotation manuelle

Des applications dans nos vies

L'annotation manuelle dans le TAL

Un sujet de recherche émergent

Formaliser l'annotation manuelle de corpus

Myriadisation et production de données

Des limites du travail parcellisé

Myriadiser des tâches complexes : ZombiLingo

Un savoir-faire complexe

Conclusion et perspectives

«Traduction» automatique



Translate

EnglishSpanishFrenchFrench - detected▼



EnglishSpanishArabic▼

Translate

les questions que soulève la science du langage, dans toutes ses versions, sont des questions fines : une proposition de linguistique concerne peu de données à la fois et elle y fait apparaître généralement ce que l'opinion courante tiendrait pour des détails. En bref, il



the questions raised by the science of language, in all its versions, are fine questions: a linguistic proposal concerns little data at a time and it makes it appear that generally current opinion held for details. In short, it seems unlikely that from the science of language, we can develop the equivalent of a steel or aerospace



<https://translate.google.com/>

«Traduction» automatique



Translate

EnglishSpanishFrenchFrench - detected



EnglishSpanishArabic

Translate

les questions que soulève la science du langage, dans toutes ses versions, sont des questions fines : une proposition de linguistique concerne peu de données à la fois et elle y fait apparaître généralement ce que l'opinion courante tiendrait pour des détails. En bref, il



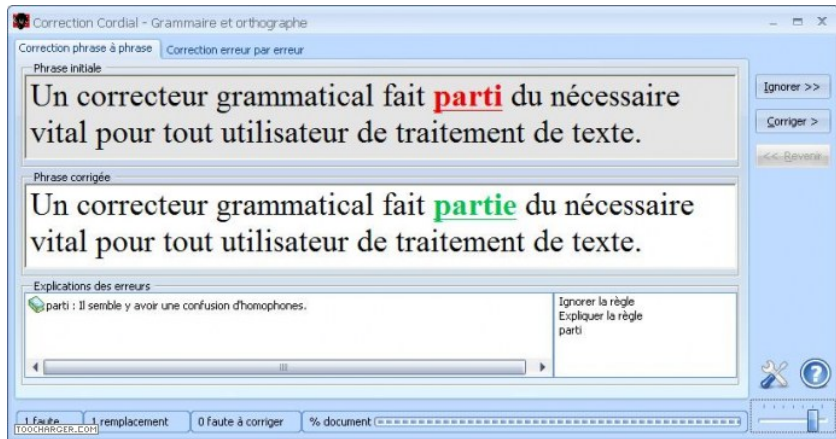
the questions raised by the science of language, in all its versions, are fine questions: a linguistic proposal concerns little data at a time and it makes it appear that generally current opinion held for details. In short, it seems unlikely that from the science of language, we can develop the equivalent of a steel or aerospace industry

Your contribution will be used to improve translation quality and may be shown to users anonymously

ContributeClose

<https://translate.google.com/>

Correction orthographique et grammaticale



<http://www.synapse-fr.com/>

Extraction d'entités nommées

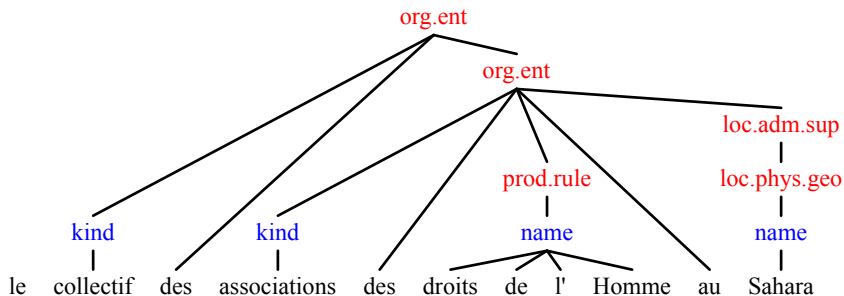
<ENAMEX TYPE="PERSON">Henri</ENAMEX> a acheté **<NUMEX TYPE="QUANTITY">300</NUMEX>** actions de la société **<ENAMEX TYPE="ORGANIZATION">AMD</ENAMEX>** en **<TIMEX TYPE="DATE">2006</TIMEX>**.

http://fr.wikipedia.org/wiki/Reconnaissance_d%27entit%C3%A9s_nomm%C3%A9es

Extraction d'entités nommées

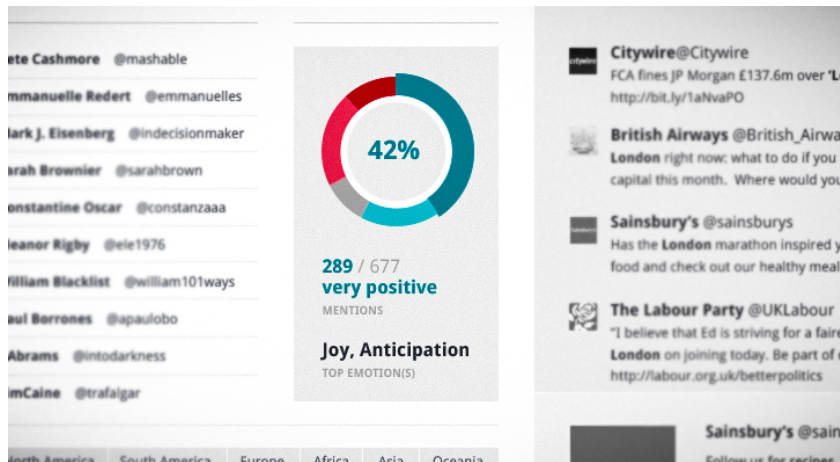
<ENAMEX TYPE="PERSON">Henri</ENAMEX> a acheté **<NUMEX TYPE="QUANTITY">300</NUMEX>** actions de la société **<ENAMEX TYPE="ORGANIZATION">AMD</ENAMEX>** en **<TIMEX TYPE="DATE">2006</TIMEX>**.

http://fr.wikipedia.org/wiki/Reconnaissance_d%27entit%C3%A9s_nomm%C3%A9es



[Grouin et al., 2011]

Analyse de sentiments



<http://www.sentoapp.com/semantic-analysis.php>

Éthique et TAL : une question qui prend corp(u)s

Un questionnement plus large que le TAL :

- ▶ Lettre ouverte sur l'intelligence artificielle (IA), jan. 2015
- ▶ Enjeux éthiques du « Big data » : opportunités et risques (Société Française de Statistique), mai 2014
- ▶ Journée sur l'éthique du numérique (CERNA), juin 2015

Des espaces de réflexions spécifiques :

- ▶ Journée d'études ATALA Éthique et TAL, nov. 2014
- ▶ Atelier ETeRNAL (Éthique et TAL), juin 2015
- ▶ Numéro spécial de la revue TAL « TAL et Éthique », déc. 2016

Des propositions pratiques :

- ▶ Charte Éthique et Big Data [Couillault and Fort, 2013]
- ▶ Blog <http://ethique-et-tal.org/>

Au cœur du TAL : les ressources langagières

- ▶ systèmes à base de règles
 - ▶ définies par l'humain (linguistes)
 - ▶ entrées manuellement

→ grammaires, lexiques

mirador	nc	[pred='mirador____1<Suj:(sn),Objde:(de-sn de-sinf),Objà:(à-sinf)'] ,cat=nc,@ms]	mirador____1	Default ms
miradors	nc	[pred='mirador____1<Suj:(sn),Objde:(de-sn de-sinf),Objà:(à-sinf)'] ,cat=nc,@mp]	mirador____1	Default mp
mirage	nc	[pred='mirage____1<Suj:(sn),Objde:(de-sn de-sinf),Objà:(à-sinf)'] ,cat=nc,@ms]	mirage____1	Default ms
mirages	nc	[pred='mirage____1<Suj:(sn),Objde:(de-sn de-sinf),Objà:(à-sinf)'] ,cat=nc,@mp]	mirage____1	Default mp

[Lefff, [Sagot, 2010]]

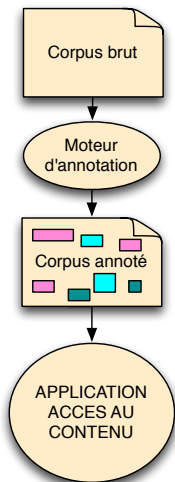
- ▶ systèmes basés sur les données ($\sim 90\%$ aujourd'hui)
 - ▶ apprentissage (en général) supervisé
 - ▶ à partir d'exemples (rédigés et/ou annotés par des humains)
 - ▶ algorithmes statistiques (pensés par des humains)

→ corpus annotés (et lexiques)

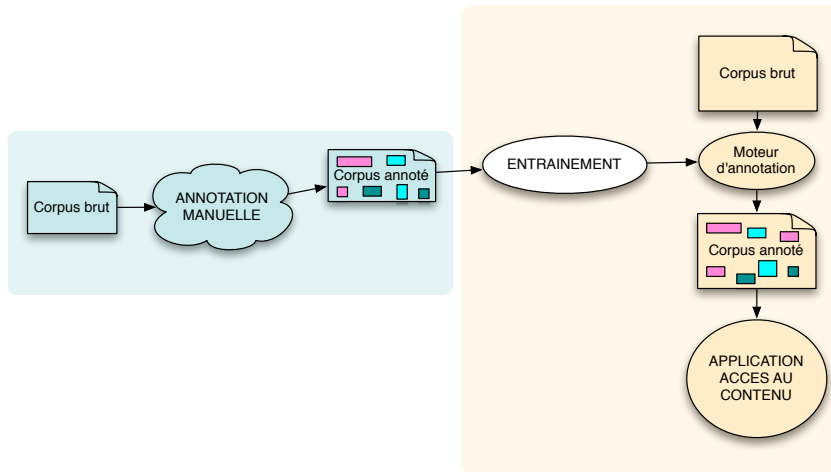


[Annotation en syntaxe de dépendances]

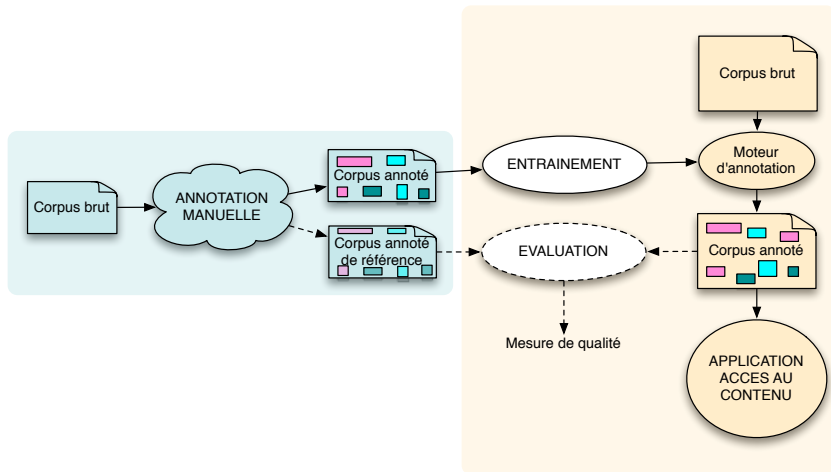
Corpus annotés et Traitement Automatique des Langues



Corpus annotés et Traitement Automatique des Langues



Corpus annotés et Traitement Automatique des Langues



Que sait-on de l'annotation manuelle ?

Un existant **parcellaire** et mal documenté :

- ▶ bonnes pratiques de **haut niveau**
- ▶ solutions au **cas par cas** : biais ? qualité ? éthique ?
- ▶ **bribes** de méthodologies
- ▶ mesures d'évaluation **pas toujours adaptées** et peu explicitées



Que sait-on de l'annotation manuelle ?

Un existant **parcellaire** et mal documenté :

- ▶ bonnes pratiques de **haut niveau**
- ▶ solutions au **cas par cas** : biais ? qualité ? éthique ?
- ▶ **bribes** de méthodologies
- ▶ mesures d'évaluation **pas toujours adaptées** et peu explicitées

Nécessité d'une vision **d'ensemble** :

- **formaliser** l'annotation
- **outiller** l'annotation
- clarifier le rôle des **acteurs**
- définir les **mesures d'évaluation** adaptées



Traitement Automatique des Langues (TAL) et annotation manuelle

Formaliser l'annotation manuelle de corpus

- Définir l'annotation

- Identifier les dimensions de complexité de l'annotation

- Outiller à bon escient

Myriadisation et production de données

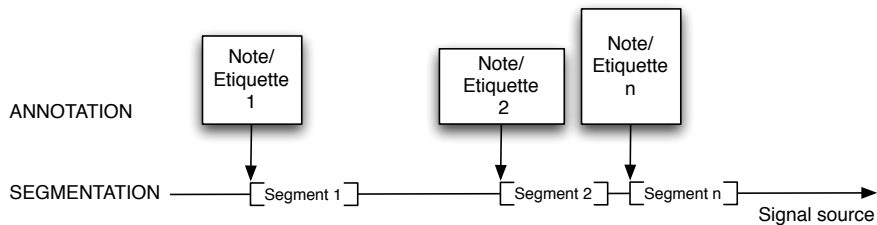
Des limites du travail parcellisé

Myriadiser des tâches complexes : ZombiLingo

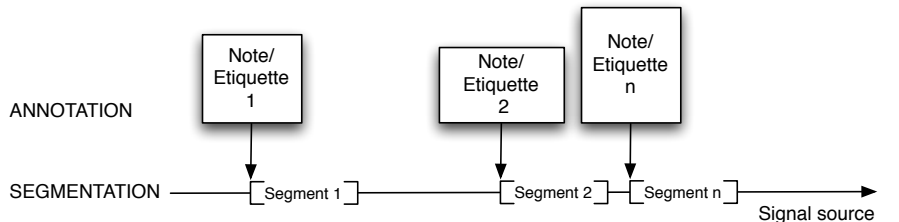
Un savoir-faire complexe

Conclusion et perspectives

Annoter ?



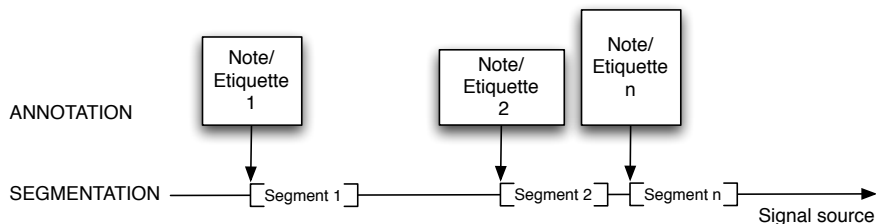
Annoter ?



ouais	ben	je	l'	ai	eu	au	tel
FNO	INT	PRO :cls	PRO :clo	AUX :pres	VER :pper	PRP :det	NOM :trc
ouais	ben	je	le	avoir	avoir	au	téléphone

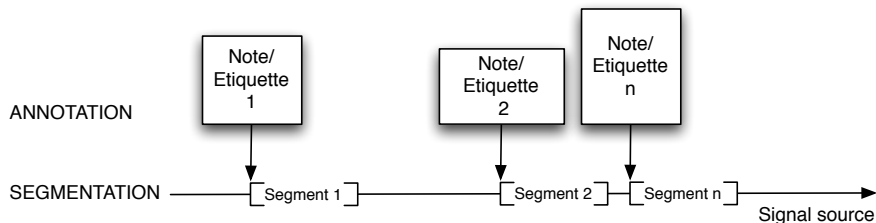
[Benzitoun et al., 2012]

Annoter ?



Ajout d'informations **interprétatives** [Leech, 1997, Habert, 2005]

Annoter ?



Ajout d'informations **interprétatives** [Leech, 1997, Habert, 2005]

Une annotation est toujours **orientée par une tâche** [Habert, 2000]

Exercice : annoter des commentaires de matchs de foot
joueurs, équipes, actions (buts), relations (passes), etc.

Avec une énorme surprise du côté du Bayern Munich puisque Van Bommel, le capitaine, a été écarté. Il n'est même pas sur la liste des remplaçants.

Exercice : annoter des commentaires de matchs de foot
joueurs, équipes, actions (buts), relations (passes), etc.

Avec une énorme surprise du côté du Bayern Munich puisque Van Bommel, le capitaine, a été écarté. Il n'est même pas sur la liste des remplaçants.

Quelle est la tâche orientant l'annotation ?

résumé de match

Van Bommel ?

ne doit pas être annoté

Exercice : annoter des commentaires de matchs de foot
joueurs, équipes, actions (buts), relations (passes), etc.

Fabien Lévêque : C'est bien fait, avec Gouffran maintenant.

Gouffran qui va tenter sa chance, et ça fait le but. Le but !

Xavier Gravelaine : Oh la la la la !

Fabien Lévêque : Et le but du plus breton des Girondins. C'est

Yoann Gourcuff qui vient de mettre un quatrième but ici au stade de France. Le cauchemar continue pour le VOC. Quatre à zéro en faveur des Girondins.

Exercice : annoter des commentaires de matchs de foot

joueurs, équipes, actions (buts), relations (passes), etc.

Fabien Lévêque : C'est bien fait, avec Gouffran maintenant.

Gouffran qui va tenter sa chance, et ça fait le but. Le but !

Xavier Gravelaine : Oh la la la la !

Fabien Lévêque : Et le but du plus breton des Girondins. C'est Yoann Gourcuff qui vient de mettre un quatrième but ici au stade de France. Le cauchemar continue pour le VOC. Quatre à zéro en faveur des Girondins.

Fabien Lévêque : C'est bien fait , avec Gouffran maintenant . Gouffran qui va tenter sa chance , et ça fait le but . Le but !

Xavier Gravelaine : Oh la la la la !

Fabien Lévêque : Et le but du plus breton des Girondins . C'est Yoann Gourcuff qui vient mettre un quatrième but ici au stade de France . Le cauchemar continue pour le VOC . Quatre à zéro en faveur des Girondins .

ID=518

Le consensus, au cœur de l'annotation

Il faut «convenir pour mesurer »[Desrosières, 2008]

L'annotation est de l'ordre de la **quantification**

Mesurer vs quantifier [Desrosières, 2008] :

- ▶ **mesurer** : implique une forme mesurable (par ex. la hauteur du Mont Blanc)
- ▶ **quantifier** : suppose des conventions d'équivalences préalables

Outiller le consensus :

- ▶ guide d'annotation (12 p. pour le football)
- ▶ réunions avec les annotateurs et le questionnaire de la campagne
- ▶ **évaluer** le consensus (la cohérence)

Jugement d'acceptabilité vs annotation

tel Monsieur Jourdain...

- (a') *Certains libraires vendent ces livres*
- (b') *Ces livres, certains libraires les vendent*
- (a'') *?Des libraires vendent ces livres*
- (b'') **Ces livres, certains libraires vendent*

[Guentchéva and Desclés, 1991]

Annotation insérée en début de phrase, 3 catégories possibles :

- ▶ acceptable (aucune note)
- ▶ non acceptable : *
- ▶ incertain : ?

Jugement d'acceptabilité vs annotation

tel Monsieur Jourdain...

- (a') *Certains libraires vendent ces livres*
- (b') *Ces livres, certains libraires les vendent*
- (a'') *?Des libraires vendent ces livres*
- (b'') **Ces livres, certains libraires vendent*

[Guentchéva and Desclés, 1991]

Obtention d'un consensus d'acceptabilité [Habert, 2008] :

- ▶ jugement éduqué, informé, soumis à un apprentissage
- ▶ suppose un travail collectif

Jugement d'acceptabilité vs annotation

cependant...

En France :

- ▶ pas de guide d'acceptabilité : pas de "trace" globale des acceptabilités sur tel ou tel phénomène (sauf LADL/M. Gross)
- ▶ pas de travail en largeur ou systématique (sauf LADL/M. Gross)
- ▶ travail sur des énoncés simplifiés [Milner, 1989]
- ▶ l'annotation traite d'une très (plus ?) large variété de phénomènes (cf football)

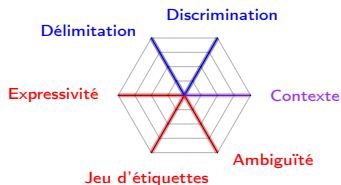
Qu'est-ce qui est complexe ?

dans l'annotation de commentaires de matchs, en syntaxe, etc. et comment les comparer ?



Dimensions de complexité [Fort et al., 2012b]

1. Discrimination des unités à annoter
2. Délimitation des unités à annoter
3. Expressivité du langage d'annotation
4. Dimension du jeu d'étiquettes
5. Ambiguïté
6. Contexte à prendre en compte



- ▶ Métriques associées, calculables a priori ou sur un échantillon
- ▶ Indépendantes du volume à annoter et du nombre d'annotateurs

Quoi annoter ? Discrimination

Parties du discours [Marcus et al., 1993], pré-annotées :

I/PRP do/VBP n't/RB feel/VB very/RB ferocious/JJ ./.

Renommage de gènes [Fort et al., 2012a], non pré-annoté :

The yppB :cat and ypbC :cat null alleles rendered cells sensitive to DNA-damaging agents, impaired plasmid transformation (25- and 100-fold), and moderately affected chromosomal transformation when present in an otherwise Rec+ B. subtilis strain. The yppB gene complemented the defect of the recG40 strain. yppB and ypbC and their respective null alleles were termed recU and "recU1" (recU :cat) and recS and "recS1" (recS :cat), respectively. The recU and recS mutations were introduced into rec-deficient strains representative of the alpha (recF), beta (addA5 addB72), gamma (rech342), and epsilon (recG40) epistatic groups.

Discrimination

Plus les unités à annoter sont « noyées » au milieu des autres, plus le poids de la discrimination est élevé

Définition

$$Discrimination(Flux) = 1 - \frac{|Annotations(Flux)|}{\sum_{i=1}^{nivSeg} |UnitésObtenuesParDécoupage_i(Flux)|}$$

⇒ Nécessité d'une **segmentation de référence**

Parties du discours [Marcus et al., 1993] :

I/PRP do/VBP n't/RB feel/VB very/RB ferocious/JJ ./.

$$Discrimination_{PTB_{POS}} = 0$$

Renommage de gènes [Fort et al., 2012a] :

The yppB :cat and ypbC :cat null alleles rendered cells sensitive to DNA-damaging agents, impaired plasmid transformation (25- and 100-fold), and moderately affected chromosomal transformation when present in an otherwise Rec+ B. subtilis strain. The yppB gene complemented the defect of the recG40 strain. yppB and ypbC and their respective null alleles were termed recU and "recU1" (recU :cat) and recS and "recS1" (recS :cat), respectively. The recU and recS mutations were introduced into rec-deficient strains representative of the alpha (recF), beta (addA5 addB72), gamma (recH342), and epsilon (recG40) epistatic groups.

$$Discrimination_{Renommage} = 0,95$$

Quoi annoter ? Délimitation des frontières

Délimiter les frontières consiste à :

- ▶ **étendre** ou **rétrécir** l'unité discriminée :

Madame Chirac → *Monsieur et Madame Chirac*

Quoi annoter ? Délimitation des frontières

Délimiter les frontières consiste à :

- ▶ **étendre** ou **rétrécir** l'unité discriminée :
Madame Chirac → *Monsieur et Madame Chirac*
- ▶ **décomposer** une unité discriminée en plusieurs éléments :
le préfet Érignac → le *préfet Érignac*

Quoi annoter ? Délimitation des frontières

Délimiter les frontières consiste à :

- ▶ **étendre** ou **rétrécir** l'unité discriminée :
Madame Chirac → *Monsieur et Madame Chirac*
- ▶ **décomposer** une unité discriminée en plusieurs éléments :
le préfet Érignac → *le **préfet Érignac***
- ▶ ou **regrouper** plusieurs unités discriminées en une seule annotation :
Sa Majesté
le roi Mohamed VI → ***Sa Majesté le roi Mohamed VI***

Délimitation

Définition

$$Délimitation(Flux) = \min \left(\frac{Substitutions + Ajouts + Suppressions}{|Annotations(Flux)|}, 1 \right)$$

$$Délimitation_{Renommage} = 0$$

$$Délimitation_{EN_{TypesSoustypes}} = 1$$

Comment annoter ? Expressivité du langage d'annotation

Définition

Les degrés d'expressivité du langage d'annotation sont les suivants :

- ▶ 0,25 : langages de types
- ▶ 0,5 : langages relationnels d'arité 2
- ▶ 0,75 : langages relationnels d'arité supérieure à 2
- ▶ 1 : langages d'ordre supérieur

$$Expressivité_{Renommage} = 0,25$$

$$Expressivité_{PTB_{POS}} = 0,25$$

Comment annoter ? Dimension du jeu d'étiquettes

Person			Function		
<i>pers.ind</i> (individual person)	(individual)	<i>pers.coll</i> (group of persons)	<i>func.ind</i> (individual function)	(individual)	<i>func.coll</i> (collectivity of functions)
Location			Production		
administrative (<i>loc.adm.town</i> , <i>loc.adm.reg</i> , <i>loc.adm.nat</i> , <i>loc.adm.sup</i>)	physical (<i>loc.phys.geo</i> , <i>loc.phys.hydro</i> , <i>loc.phys.astro</i>)	facilities (<i>loc.fac</i>), oronyms (<i>loc.oro</i>), address (<i>loc.add.phys</i> , <i>loc.add.elec</i>)	<i>prod.object</i> (manufactured object)	<i>prod.serv</i> (transportation route)	<i>prod.fin</i> (financial products)
			<i>prod.doctr</i> (doctrine)	<i>prod.rule</i> (law)	<i>prod.soft</i> (software)
			<i>prod.art</i>	<i>prod.media</i>	<i>prod.award</i>
Organization			Time		
<i>org.adm</i> (administration)	(administration)	<i>org.ent</i> (services)	<i>time.date.abs</i> (absolute date), <i>time.date.rel</i> (relative date)		<i>time.hour.abs</i> (absolute hour), <i>time.hour.rel</i> (relative hour)
Amount					
<i>amount</i> (with unit or general object), including duration					

Types et sous-types utilisés pour l'annotation en EN structurées

Comment annoter ? Dimension du jeu d'étiquettes

Person			Function		
<i>pers.ind</i> (individual person)	(individual)	<i>pers.coll</i> (group of persons)	<i>func.ind</i> (individual function)	(individual)	<i>func.coll</i> (collectivity of functions)
Location			Production		
administrative (<i>loc.adm.town</i> , <i>loc.adm.reg</i> , <i>loc.adm.nat</i> , <i>loc.adm.sup</i>)	physical (<i>loc.phys.geo</i> , <i>loc.phys.hydro</i> , <i>loc.phys.astro</i>)	facilities (<i>loc.fac</i>), oronyms (<i>loc.oro</i>), address (<i>loc.add.phys</i> , <i>loc.add.elec</i>)	<i>prod.object</i> (manufactured object)	<i>prod.serv</i> (transportation route)	<i>prod.fin</i> (financial products)
			<i>prod.doctr</i> (doctrine)	<i>prod.rule</i> (law)	<i>prod.soft</i> (software)
			<i>prod.art</i>	<i>prod.media</i>	<i>prod.award</i>
Organization			Time		
<i>org.adm</i> (administration)	(administration)	<i>org.ent</i> (services)	<i>time.date.abs</i> (absolute date), <i>time.date.rel</i> (relative date)	<i>time.hour.abs</i> (absolute hour), <i>time.hour.rel</i> (relative hour)	
Amount					
<i>amount</i> (with unit or general object), including duration					

Niveau 1 : *pers*, *func*, *loc*, *prod*, *org*, *time*, *amount* → 7 possibilités (degré de liberté = 6).

Comment annoter ? Dimension du jeu d'étiquettes

Person			Function		
<i>pers.ind</i> (individual person)		<i>pers.coll</i> (group of persons)	<i>func.ind</i> (individual function)		<i>func.coll</i> (collectivity of functions)
Location			Production		
administrative (<i>loc.adm.town</i> , <i>loc.adm.reg</i> , <i>loc.adm.nat</i> , <i>loc.adm.sup</i>)	physical (<i>loc.phys.geo</i> , <i>loc.phys.hydro</i> , <i>loc.phys.astro</i>)	facilities (<i>loc.fac</i>), oronyms (<i>loc.oro</i>), address (<i>loc.add.phys</i> , <i>loc.add.elec</i>)	<i>prod.object</i> (manufactured object)	<i>prod.serv</i> (transportation route)	<i>prod.fin</i> (financial products)
			<i>prod.doctr</i> (doctrine)	<i>prod.rule</i> (law)	<i>prod.soft</i> (software)
			<i>prod.art</i>	<i>prod.media</i>	<i>prod.award</i>
Organization			Time		
<i>org.adm</i> (administration)		<i>org.ent</i> (services)	<i>time.date.abs</i> (absolute date), <i>time.date.rel</i> (relative date)	<i>time.hour.abs</i> (absolute hour), <i>time.hour.rel</i> (relative hour)	
Amount					
<i>amount</i> (with unit or general object), including duration					

Niveau 1 : *pers*, *func*, *loc*, *prod*, *org*, *time*, *amount* → 7 possibilités (degré de liberté = 6).

Niveau 2 : *prod.object*, *prod.serv*, *prod.fin*, *prod.soft*, *prod.doctr*, *prod.rule*, *prod.art*, *prod.media*, *prod.award* → 9 possibilités (degré de liberté = 8).

Comment annoter ? Dimension du jeu d'étiquettes

Person			Function		
<i>pers.ind</i> (individual person)	(individual)	<i>pers.coll</i> (group of persons)	<i>func.ind</i> (individual function)	(individual)	<i>func.coll</i> (collectivity of functions)
Location			Production		
<i>administrative</i> (<i>loc.adm.town</i> , <i>loc.adm.reg</i> , <i>loc.adm.nat</i> , <i>loc.adm.sup</i>)	physical (<i>loc.phys.geo</i> , <i>loc.phys.hydro</i> , <i>loc.phys.astro</i>)	facilities (<i>loc.fac</i>), oronyms (<i>loc.oro</i>), address (<i>loc.add.phys</i> , <i>loc.add.elec</i>)	<i>prod.object</i> (manufactured object)	<i>prod.serv</i> (transportation route)	<i>prod.fin</i> (financial products)
			<i>prod.doctr</i> (doctrine)	<i>prod.rule</i> (law)	<i>prod.soft</i> (software)
			<i>prod.art</i>	<i>prod.media</i>	<i>prod.award</i>
Organization			Time		
<i>org.adm</i> (administration)	(administration)	<i>org.ent</i> (services)	<i>time.date.abs</i> (absolute date), <i>time.date.rel</i> (relative date)		<i>time.hour.abs</i> (absolute hour), <i>time.hour.rel</i> (relative hour)
Amount					
<i>amount</i> (with unit or general object), including duration					

Niveau 1 : *pers*, *func*, *loc*, *prod*, *org*, *time*, *amount* → 7 possibilités (degré de liberté = 6).

Niveau 2 : *prod.object*, *prod.serv*, *prod.fin*, *prod.soft*, *prod.doctr*, *prod.rule*, *prod.art*, *prod.media*, *prod.award* → 9 possibilités (degré de liberté = 8).

Niveau 3 : *loc.adm.town*, *loc.adm.reg*, *loc.adm.nat*, *loc.adm.sup* → 4 possibilités (degré de liberté = 3).

Dimension du jeu d'étiquettes

Degré de liberté

$$\nu = \nu_1 + \nu_2 + \dots + \nu_m$$

où ν_i est le degré de liberté maximal que l'annotateur a dans le choix de la i^{eme} sous-étiquette ($\nu_i = n_i - 1$).

Dimension du jeu d'étiquettes

$$Dimension(Flux) = \min\left(\frac{\nu}{\tau}, 1\right)$$

où τ est le seuil à partir duquel on considère le jeu d'étiquettes comme arbitrairement grand (déterminé expérimentalement).

$$Dimension_{Renommage} = 0,04$$

$$Dimension_{EN_{TypesSoustypes}} = 0,34$$

Comment annoter ? Degré d'ambiguïté : ambiguïté résiduelle

Utiliser les traces laissées par les annotateurs :



[...] <EukVirus>3CDproM</EukVirus> can process both structural and nonstructural precursors of the <EukVirus
uncertainty-type = "too-generic"><taxon>poliovirus</taxon> polyprotein</EukVirus> [...].

Définition

$$Ambiguité_{Res}(Flux) = \frac{|Annotations_{amb}|}{|Annotations|}$$

$$Ambiguité_{Res_{Renommage}} = 0,02$$

Comment annoter ? Degré d'ambiguïté : ambiguïté théorique

Proportion des unités à annoter qui correspondent à des vocables ambigus.

Définition

$$Ambiguïté_{Th}(Flux) = \frac{\sum_{voc_i=1}^{|Voc(Flux)|} (Ambig(voc_i) * freq(voc_i, Flux))}{|Unités(Flux)|}$$

avec

$$Ambig(voc_i) = \begin{cases} 1 & \text{si } |Étiquettes(voc_i)| > 1 \\ 0 & \text{sinon} \end{cases}$$

→ Ne s'applique pas aux relations de renommage.

Poids du contexte

- **taille de la fenêtre** de signal source à prendre en compte :

- La phrase :

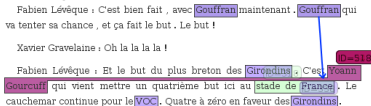
I/PRP do/VBP n't/RB feel/VB very/RB ferocious/JJ ./.

- ... ou plus :

Fabien Lévêque : C'est bien fait , avec Gouffran maintenant . Gouffran qui va tenter sa chance , et ça fait le but . Le but !

Xavier Gravelaine : Oh la la la la !

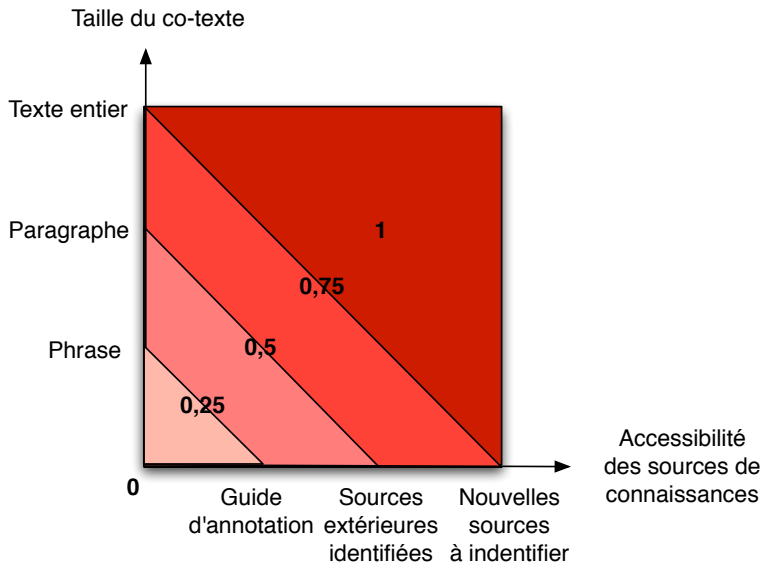
Fabien Lévêque : Et le but du plus breton des Girondins . C'est Yoann Gourcufr qui vient mettre un quatrième but ici au stade de France . Le cauchemar continue pour le VOC . Quatre à zéro en faveur des Girondins .



- nombre de **connaissances** à mobiliser ou degré d'accessibilité des sources de connaissances qui sont consultées :

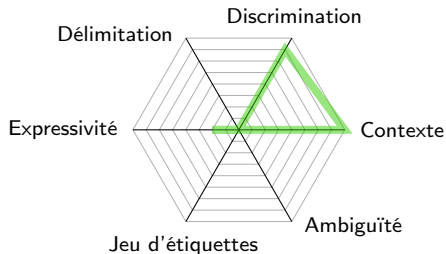
- guide d'annotation
 - nomenclatures (Swiss-Prot)
 - nouvelles sources à trouver (Wikipedia, etc.)

Poids du contexte

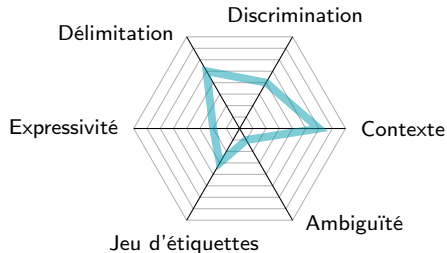


Outiller à bon escient...

Renommage de noms de gènes



Entités nommées structurées



...en fonction du **profil de complexité** de la campagne

Traitement Automatique des Langues (TAL) et annotation manuelle

Formaliser l'annotation manuelle de corpus

Myriadisation et production de données

Chères ressources langagières

La myriadisation comme « solution »

Des limites du travail parcellisé

Myriadiser des tâches complexes : ZombiLingo

Un savoir-faire complexe

Conclusion et perspectives

Produire des données annotées : l'enjeu du coût

Prague Dependency Treebank [Böhmová et al., 2001] :

- ▶ 1,8 millions de mots annotés en morpho-syntaxe et syntaxe
- ⇒ 5 ans, 22 personnes (max. 17 en parallèle), 600 000 dollars

ESTER

- ▶ 100 h de parole transcrite (campagne d'évaluation ESTER, systèmes de transcription, 2008)
- ▶ 1 h de parole = entre 20 et 60 h de travail de transcription

La myriadisation (*crowdsourcing*)

[Sagot et al., 2011] ou *externalisation ouverte*, ou *production participative*

***Crowdsourcing** is "the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call."
[Howe, 2006]*

- **pas d'identification** ou de sélection *a priori* des participants ("open call")

La myriadisation (*crowdsourcing*)

[Sagot et al., 2011] ou *externalisation ouverte*, ou *production participative*

*Crowdsourcing is "the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call."
[Howe, 2006]*

- ▶ **pas d'identification** ou de sélection *a priori* des participants ("open call")
- ▶ **massive** (dans la production et la participation)

La myriadisation (*crowdsourcing*)

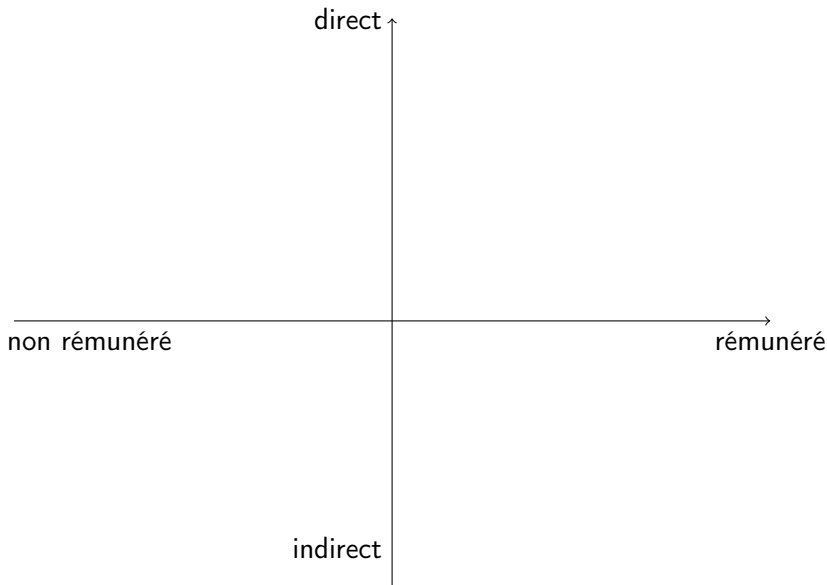
[Sagot et al., 2011] ou *externalisation ouverte*, ou *production participative*

*Crowdsourcing is "the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call."
[Howe, 2006]*

- ▶ **pas d'identification** ou de sélection *a priori* des participants ("open call")
- ▶ **massive** (dans la production et la participation)
- ▶ (relativement) **peu coûteux** (répond à la préoccupation du coût)

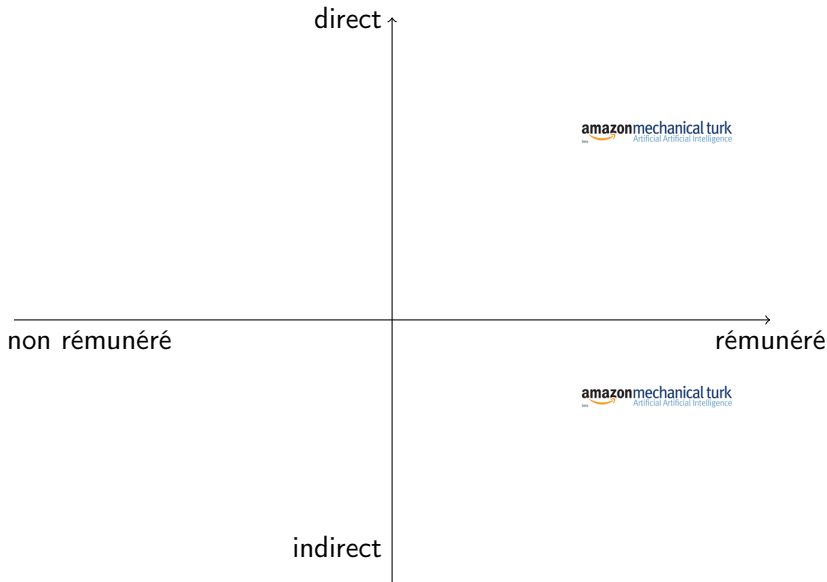
Une taxinomie (simplifiée) de la myriadisation

parmi d'autres, voir [Geiger et al., 2011]



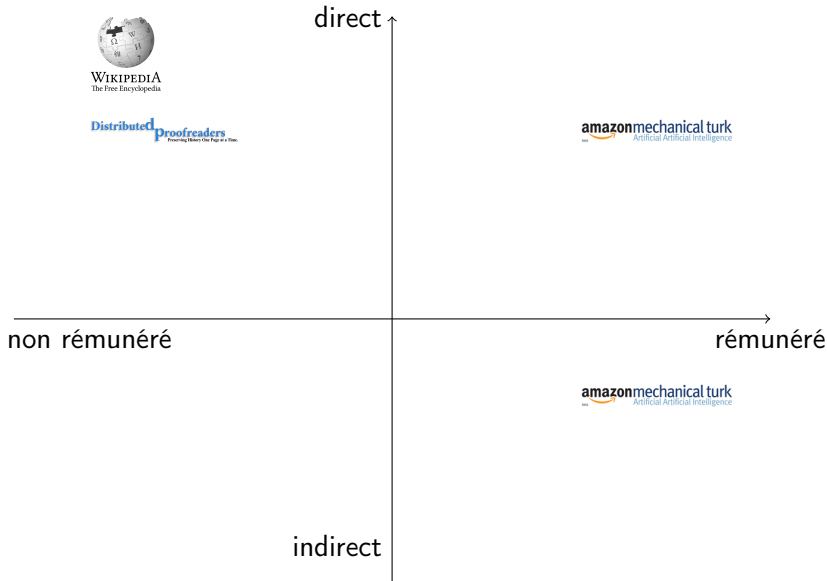
Une taxinomie (simplifiée) de la myriadisation

parmi d'autres, voir [Geiger et al., 2011]



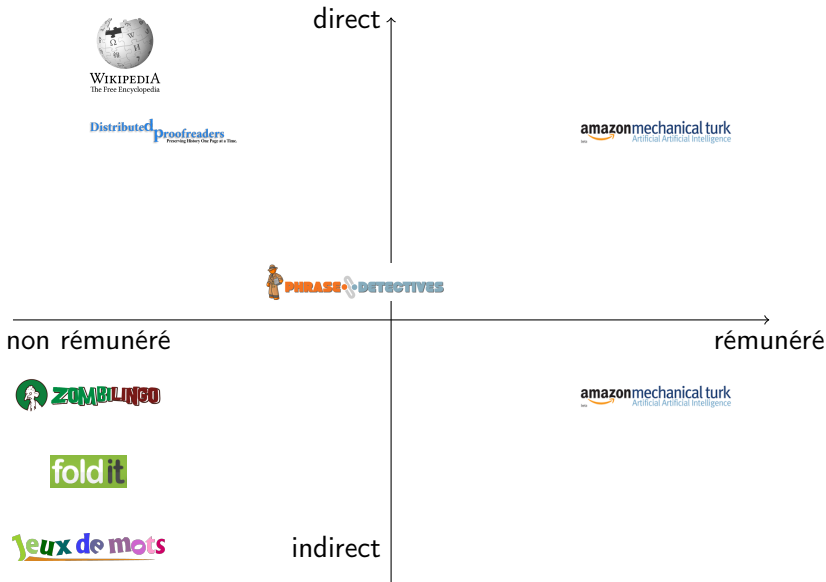
Une taxinomie (simplifiée) de la myriadisation

parmi d'autres, voir [Geiger et al., 2011]



Une taxinomie (simplifiée) de la myriadisation

parmi d'autres, voir [Geiger et al., 2011]



Traitement Automatique des Langues (TAL) et annotation manuelle

Formaliser l'annotation manuelle de corpus

Myriadisation et production de données

Des limites du travail parcellisé

- Amazon Mechanical Turk : une plate-forme de légendes

- Éthique

- Des coûts qui varient

- Complexité des tâches

- Des résultats étonnants

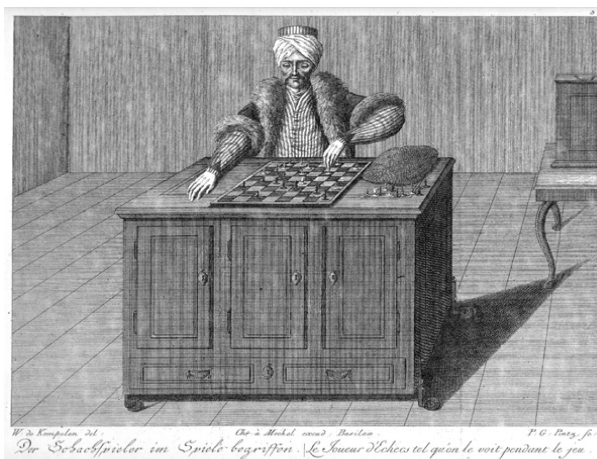
Myriadiser des tâches complexes : ZombiLingo

Un savoir-faire complexe

Conclusion et perspectives

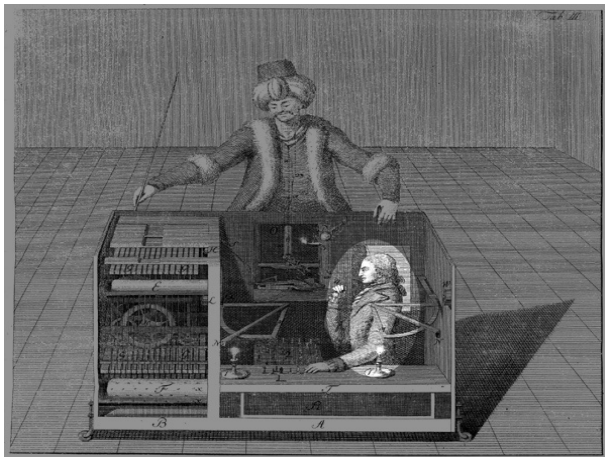
Le «Turc mécanique» de von Kempelen

Un joueur d'échecs mécanique créé par J. W. von Kempelen en 1770 :



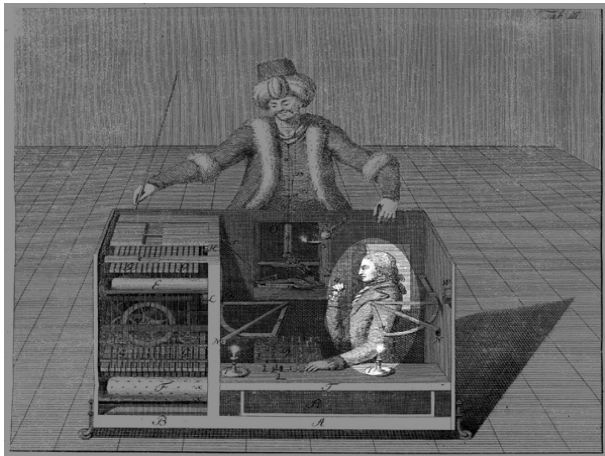
Le «Turc mécanique» de von Kempelen

En fait, un maître d'échecs était caché dans la machine :



Le «Turc mécanique» de von Kempelen

C'est l'intelligence artificielle **artificielle** !



Et Amazon créa AMT

Amazon crée une pour ses propres besoins
plate-forme de travail parcellisé
et en ouvre l'accès en 2005 (moyennant 10 % des transactions)

AMT : le rêve devenu réalité ?

pour le Traitement Automatique des Langues

Cheap and Fast — But is it Good? **Evaluating Non-Expert Annotations for Natural Language Tasks**

Rion Snow[†] Brendan O'Connor[‡] Daniel Jurafsky[§] Andrew Y. Ng[†]

[†]Computer Science Dept.
Stanford University
Stanford, CA 94305
{rion,ang}@cs.stanford.edu

[‡]Dolores Labs, Inc.
832 Capp St.
San Francisco, CA 94110
brendano@doloreslabs.com

[§]Linguistics Dept.
Stanford University
Stanford, CA 94305
jurafsky@stanford.edu

[Snow et al., 2008]

AMT : le rêve devenu réalité ?

pour le Traitement Automatique des Langues

Cheap and Fast — But is it Good? **Evaluating Non-Expert Annotations for Natural Language Tasks**

Rion Snow[†] Brendan O'Connor[‡] Daniel Jurafsky[§] Andrew Y. Ng[†]

[†]Computer Science Dept.
Stanford University
Stanford, CA 94305
{rion,ang}@cs.stanford.edu

[‡]Dolores Labs, Inc.
832 Capp St.
San Francisco, CA 94110
brendano@doloreslabs.com

[§]Linguistics Dept.
Stanford University
Stanford, CA 94305
jurafsky@stanford.edu

[Snow et al., 2008]

C'est très peu cher, rapide, de bonne qualité
et c'est un hobby pour les *Turkers* !

Amazon Mechanical Turk (AMT)

AMT

The screenshot shows the Amazon Mechanical Turk (AMT) website. The browser address bar displays <https://www.mturk.com/mturk/welcome>. The page header includes the Amazon Mechanical Turk logo, navigation tabs for 'Your Account', 'HITS', and 'Qualifications', and a search bar containing the text 'Rechercher'. Below the header, a blue banner reads 'Mechanical Turk is a marketplace for work.' followed by the text 'We give businesses and developers access to an on-demand, scalable workforce. Workers select from thousands of tasks and work whenever it's convenient.' and '410,720 HITS available. [View them now.](#)'

Make Money by working on HITS

HITS - Human Intelligence Tasks - are individual tasks that you work on. [Find HITS now.](#)

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work

Find an interesting task → **Work** → **Earn money**

[Find HITS Now](#)

or [learn more about being a Worker](#)

Get Results from Mechanical Turk Workers

Ask workers to complete HITS - Human Intelligence Tasks - and get results using Mechanical Turk. [Get Started.](#)

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITS completed in minutes
- Pay only when you're satisfied with the results

Fund your account → **Load your tasks** → **Get results**

[Get Started](#)

Amazon Mechanical Turk (AMT)

AMT est une plate-forme de **myriadisation** : le travail est *externalisé* via le Web et réalisé par des internautes, les **Turkers**

The screenshot shows the Amazon Mechanical Turk homepage. At the top, there's a navigation bar with the Amazon Mechanical Turk logo, a search bar, and links for 'Your Account', 'HITS', and 'Qualifications'. Below this is a banner with the text 'Mechanical Turk is a marketplace for work.' and 'We give businesses and developers access to an on-demand, scalable workforce. Workers select from thousands of tasks and work whenever it's convenient.' A yellow box highlights '410,720 HITS available. View them now.'

Make Money by working on HITS

HITS - Human Intelligence Tasks - are individual tasks that you work on. [Find HITS now.](#)

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work

Find an interesting task → **Work** → **Earn money**

[Find HITS Now](#)

or [learn more about being a Worker](#)

Get Results from Mechanical Turk Workers

Ask workers to complete HITS - Human Intelligence Tasks - and get results using Mechanical Turk. [Get Started.](#)

As a Mechanical Turk Requester you:

- Have access to a **global, on-demand, 24 x 7 workforce**
- Get thousands of HITS completed in minutes
- Pay only when you're satisfied with the results

Fund your account → **Load your tasks** → **Get results**

[Get Started](#)

Amazon Mechanical Turk (AMT)

AMT est une plate-forme de **myriadisation** du **travail parcellisé** : les tâches sont découpées en sous-tâches (HIT) et leur exécution est payée par les **Requesters**

The screenshot shows the Amazon Mechanical Turk (AMT) website. At the top, there's a navigation bar with the AMT logo, a search bar, and links for 'Your Account', 'HITs', and 'Qualifications'. Below this, a banner states 'Mechanical Turk is a marketplace for work.' and provides information about the number of HITs available. The main content area is divided into two columns. The left column, titled 'Make Money by working on HITs', explains that HITs are individual tasks and lists benefits for workers. It includes a diagram showing the process from finding a task to earning money. The right column, titled 'Get Results from Mechanical Turk Workers', explains how requesters can use the platform and lists benefits. It includes a diagram showing the process from funding an account to getting results. Both columns have a 'Find HITs Now' or 'Get Started' button.

Make Money by working on HITs

HITs - Human Intelligence Tasks - are individual tasks that you work on. [Find HITs now.](#)

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work

Find an interesting task → **Work** → **Earn money**

[Find HITs Now](#)

or [learn more about being a Worker](#)

Get Results from Mechanical Turk Workers

Ask workers to complete HITs - Human Intelligence Tasks - and get results using Mechanical Turk. [Get Started.](#)

As a Mechanical Turk Requester you:

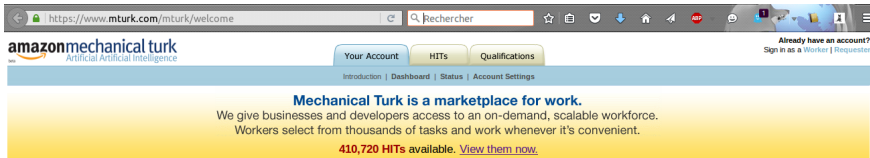
- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results

Fund your account → **Load your tasks** → **Get results**

[Get Started](#)

Amazon Mechanical Turk (AMT)

AMT est une plate-forme de **myriadisation** du travail **parcellisé** : les tâches sont découpées en sous-tâches (HIT) et leur exécution est **payée**.



The screenshot shows the Amazon Mechanical Turk homepage. At the top, there's a navigation bar with the Amazon Mechanical Turk logo, a search bar, and links for 'Your Account', 'HITS', and 'Qualifications'. Below this, a yellow banner reads: 'Mechanical Turk is a marketplace for work. We give businesses and developers access to an on-demand, scalable workforce. Workers select from thousands of tasks and work whenever it's convenient. 410,720 HITs available. View them now.'

Make Money by working on HITs

HITs - Human Intelligence Tasks - are individual tasks that you work on. [Find HITs now.](#)

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work



or [learn more about being a Worker](#)

Get Results from Mechanical Turk Workers

Ask workers to complete HITs - Human Intelligence Tasks - and get results using Mechanical Turk. [Get Started.](#)

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results



Amazon Mechanical Turk (AMT)

AMT est une plate-forme de **myriadisation** du **travail parcellisé** : les tâches sont découpées en sous-tâches (HIT) et leur exécution est **payée (ou pas)**.

The screenshot shows the Amazon Mechanical Turk (AMT) website. The browser address bar displays <https://www.mturk.com/mturk/welcome>. The page header includes the Amazon Mechanical Turk logo, navigation links for "Your Account", "HITS", and "Qualifications", and a search bar. Below the header, a yellow banner states: "Mechanical Turk is a marketplace for work. We give businesses and developers access to an on-demand, scalable workforce. Workers select from thousands of tasks and work whenever it's convenient. 410,720 HITS available. [View them now.](#)"

The main content area is divided into two columns:

- Make Money by working on HITs**: This section explains that HITs (Human Intelligence Tasks) are individual tasks that workers can complete from home on their own schedule. It lists three benefits for workers: "Can work from home", "Choose your own work hours", and "Get paid for doing good work". A diagram shows a flow from "Find an interesting task" (represented by a gear icon) to "Work" (represented by a gear icon) to "Earn money" (represented by a dollar sign icon). A button labeled "Find HITS Now" is at the bottom of this section.
- Get Results from Mechanical Turk Workers**: This section explains that requesters can ask workers to complete HITs and get results. It lists three steps: "Fund your account", "Load your tasks", and "Get results". A diagram shows a flow from "Fund your account" (represented by a plus sign icon) to "Load your tasks" (represented by a gear icon) to "Get results" (represented by a star icon). A button labeled "Get Started" is at the bottom of this section.

Below the "Make Money" section, there is a link: "or [learn more about being a Worker](#)".

Types de tâches les plus courants sur AMT

<http://www.mturk-tracker.com>, février 2016



General

Arrivals

Completions

Top requesters

Active requesters

Demographics

Search

API

Old version

Top-1000 Requesters, report for February 3, 2016 to March 4, 2016

Requester name	hits	reward
Jon Brellg	84646	\$12,939.88
SET Master Account	75842	\$816.35
VDX	68792	\$1,634.16
p9r	58274	\$2,199.26
Amazon Requester Inc - browse classification	40297	\$2,429.52
AI Indoors Project	26934	\$9,669.04
OCMP5	26579	\$1,232.27
NoblisCV	16337	\$7,122.20
PDS_PN1	14808	\$406.49
HITmeUp	14374	\$2,872.14

« Previous

10

25

50

100

Page 1 of 100

Next »

- ▶ identifications diverses sur des tickets de caisse
- ▶ classification binaire d'images (est/n'est pas sur l'image)
- ▶ ... et transcription (*CastingWords*), dialogue (*Dialogue Systems*), etc.

Est-ce qu'AMT est éthique et/ou légal ?

- ▶ pas d'identification : pas de lien officiel entre *Requesters* et *Turkers* et entre *Turkers*
- ▶ pas de possibilité de se syndiquer, pour protester contre des manquements des *Requesters* ou ester en justice
- ▶ pas de salaire minimum (< 2 \$/h en moyenne)
- ▶ possibilité de refuser de payer les *Turkers*

Already have an account?
Sign in as a [Worker](#) | [Requester](#)

HITs - *Human Intelligence Tasks* - are individual tasks that you work on. [Find HITs now.](#)

- Can work from home
- Choose your own work hours
- Get paid for doing good work



Ask workers to complete HITs - Human Intelligence Tasks - and get results using Mechanical Turk. [Get Started.](#)

- Have access to a global, on-demand

-
- Fund your account** → **Load your tasks** → **Get results**
- [Get Started](#)

AMT : un passe-temps pour les *Turkers* ?

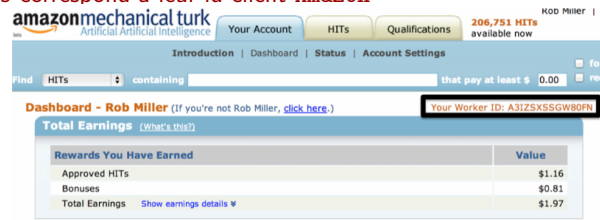
[Ross et al., 2010, Ipeirotis, 2010] montrent que :

- ▶ les *Turkers* sont avant tout motivés par l'**argent** (91 %) :
 - ▶ 20 % considèrent AMT comme leur source de revenu primaire ;
 - ▶ 50 % comme leur source de revenu secondaire ;
 - ▶ l'aspect loisir n'est important que pour une minorité (30 %).
- ▶ 20 % des *Turkers* passent plus de 15 h par semaine sur AMT, et contribuent à 80 % des tâches
- ▶ le salaire horaire moyen observé est **inférieur à 2 \$**

[Gupta et al., 2014] : en l'absence de possibilité de formation, un important **travail caché** est réalisé par les *Turkers*

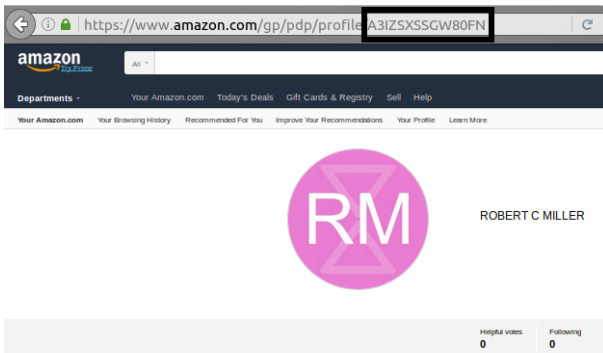
Les *Turkers* ne sont pas anonymes [Lease et al., 2013]

L'id des *Turkers* correspond à leur id client Amazon



The screenshot shows the Amazon Mechanical Turk dashboard for a user named Rob Miller. At the top, there are tabs for 'Your Account', 'HITS', and 'Qualifications'. A notification indicates '206,751 HITS available now'. Below the navigation bar, there's a search bar and a section titled 'Dashboard - Rob Miller'. A box highlights the 'Your Worker ID: A3IZSXSSGW80FN'. Below this, a 'Total Earnings' section shows a table of rewards earned.

Rewards You Have Earned	Value
Approved HITS	\$1.16
Bonuses	\$0.81
Total Earnings	\$1.97



The screenshot shows the Amazon profile page for a user named Rob C Miller. The browser address bar shows the URL 'https://www.amazon.com/gp/pdp/profile/A3IZSXSSGW80FN'. The profile picture is a pink circle with the letters 'RM' and a large 'X' over it. The name 'ROBERT C MILLER' is displayed next to the profile picture. Below the profile picture, there are statistics for 'Helpful votes' and 'Following', both showing a count of 0.

Helpful votes: 0

Following: 0

AMT permet de réduire les coûts

Très basse rémunération $\Rightarrow$ coûts faibles ? Oui, mais. . .

- ▶ coût de mise au point de l'**interface**
- ▶ coût de création de protections contre les **spammers**
- ▶ coût de **validation** et de **post-traitement**

certaines tâches (par exemple, la traduction du coréen vers l'anglais) génèrent des coûts similaires aux coûts habituels dans le domaine, du fait du **manque de *Turkers* qualifiés** [Novotney and Callison-Burch, 2010].

Quand Amazon se sert...

Amazon is doubling the fee it collects from "requesters," those seeking laborers to perform online tasks, to 20% beginning July 21. And for tasks requiring at least 10 people, Amazon will charge an additional 20%, a new fee.

[blog du Wall Street Journal, 23 juin 2015]

Dépendance à une plateforme externe

Impossibilité de maîtriser :

- ▶ les coûts
- ▶ les conditions de travail des participants
- ▶ la sélection des participants
- ▶ les conditions d'expérience

Les HIT (*Human Intelligence Task*) : des tâches simplifiées

Pas de possibilité de se former à la tâche sur AMT :

- ▶ la qualité est insuffisante lorsque la tâche est **complexe** (par exemple, le résumé [Gillick and Liu, 2010])
- ▶ pour certaines tâches simples les outils de TAL produisent de **meilleurs résultats** qu'AMT [Wais et al., 2010].

⇒ **Simplification** des tâches :

- ▶ une « vraie » tâche d'annotation en inférences textuelles (inférence, neutre, contradiction) est réduite à 2 phrases et une question :
« Would most people say that if the first sentence is true, then the second sentence must be true ? » [Bowman et al., 2015]

Jeux ayant un but

Profiter des :

1. connaissances « du monde »
2. connaissances scolaires
3. capacités d'apprentissage

des participants (rarement une foule)

JeuxDeMots : jouer à associer des idées. . .

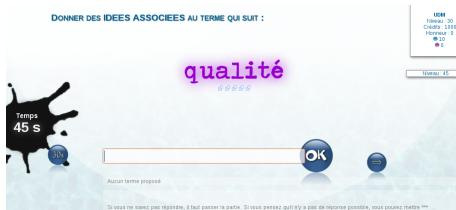
. . . pour créer un réseau lexical [Lafourcade and Joubert, 2008]

Plus de 10 millions de relations créées et constamment mises à jour

Mécanisme très élaboré :

- ▶ jeu par paires
- ▶ défis entre joueurs
- ▶ procès, etc.

→ Profiter des capacités de la foule



Phrase Detectives : jouer au détective...

... pour annoter des anaphores [Chamberlain et al., 2008]

Corpus annoté de 200 000 mots :

- ▶ corpus pré-annoté
- ▶ instructions détaillées
- ▶ phase de formation
- ▶ 2 manières de jouer

DETECTIVES CONFERENCE
Another detective has made a decision about a phrase, either that it refers to another phrase, it has not been mentioned before, it is a property or it does not refer to anything. Do you agree with them?

USERPROFILE
Kaa
22 this week
0 decisions
21 agreements
0 extras
22 this month
82 all time
Level: Apprentice
Your rating: 80%
CASE OPEN
32 tasks remaining
1 completed case
EDIT PROFILE | LOGOUT
1 / 2 items
Soyez le premier de vos amis à indiquer que vous

Knitta (Wikipedia)
PolyCoff and Molyk came up with their own names, then invented names for other members in a brainstorming session they considered "one of the more hilarious meetings". Some former member names include Krotosious N.T.T., SenOfaBitch and P-Kinky.
As of January 2008, the group has two female members and one male, ages 30 to 73, who wish to remain anonymous. Current members' are PolyCoff, MaaCoffKinky, and Gravy SQ. An estimated five to twelve cryptic groups exist around the world.
Usually tagging on Friday nights and Sunday mornings, **Knitta taggers** leave a paper tag on each work, bearing the slogan "knitta please!" or "whaddup knitta?". They tag trees, lamp posts, railings, fire hydrants, monuments and other urban targets, and even get a little "hardcore" with ideas like hanging knitted bagged sneakers over aerial telephone cable. The crew marks holidays by doing themed work, using, for example, pink yarn for their Valentine's Day pieces and sparkly yarn for New Years. When Knitta is not working with a theme, they work on projects, tagging specific targets or specific areas.
The group and their followers consider their graffiti "a method of beautifying public space".

The phrase in blue is the **eldest** phrase that refers to the phrase in orange.

Instructions
FAQ

NAME THE CULPRIT
Has the phrase shown in orange been mentioned before in this text or is it a property? Use your mouse to select the **eldest phrase(s)** if it has been mentioned before.

Knitta (Wikipedia)
PolyCoff and Molyk came up with their own names, then invented names for other members in a brainstorming session they considered "one of the more hilarious meetings". Some former member names include Krotosious N.T.T., SenOfaBitch and P-Kinky.
As of January 2008, the group has two female members and one male, ages 30 to 73, who wish to remain anonymous. Current members' are PolyCoff, MaaCoffKinky, and Gravy SQ. An estimated five to twelve cryptic groups exist around the world.
Usually tagging on Friday nights and Sunday mornings, **Knitta taggers** leave a paper tag on each work, bearing the slogan "knitta please!" or "whaddup knitta?". They tag trees, lamp posts, railings, fire hydrants, monuments and other urban targets, and even get a little "hardcore" with ideas like hanging knitted bagged sneakers over aerial telephone cable. The crew marks holidays by doing themed work, using, for example, pink yarn for their Valentine's Day pieces and sparkly yarn for New Years. When Knitta is not working with a theme, they work on projects, tagging specific targets or specific areas.

Not mentioned before ☒ This is a property

Instructions
Done

SEARCHCLUES
Where like they **are**, her and it are busy to visit to something else in the text. Try to find the closest mention of this phrase.
Where like they or **been** could refer to more than one thing in the text so select more than one phrase if necessary.
Always look for the **eldest** previous mention of the phrase to score maximum agreement points.

SEARCHCLUES
Phrases beginning with a, an or he can refer to two different persons.
1. **As an object**
They can be used to identify an object in the text. For example "The **postman** delivered a letter" or "Jane owns a **laptop**".
2. **As a property**
They can also be used to say something about an object. For example "The **postman** delivered a letter" describes the object "The" as having the property of being "The postman".
If you think the phrase describes a property try to select the closest phrase it refers to.

Feedback

→ Profiter des connaissances scolaires de la foule

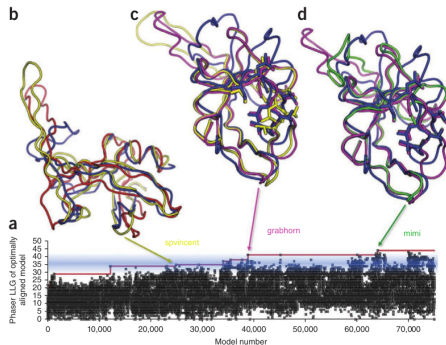
FoldIt : jouer à replier des protéines. . .

. . . pour résoudre des problèmes scientifiques [Khatib et al., 2011]

Résolution de la structure cristalline de la protéine responsable de la propagation du virus du SIDA chez les macaques rhésus

Solution à un problème non résolu depuis une dizaine d'années :

- ▶ trouvée en quelques semaines
- ▶ par étapes
- ▶ par une équipe de joueurs
- ▶ qui permettra la création d'antirétroviraux



→ Profiter des **capacités d'apprentissage** de la foule

Traitement Automatique des Langues (TAL) et annotation manuelle

Formaliser l'annotation manuelle de corpus

Myriadisation et production de données

Des limites du travail parcellisé

Myriadiser des tâches complexes : ZombiLingo

Annoter de la syntaxe en dépendances

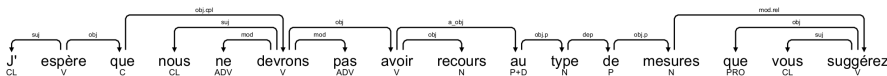
Derrière le rideau

Résultats obtenus (COLING 2016)

Un savoir-faire complexe

Conclusion et perspectives

Un type d'annotation complexe



- ▶ Guide d'annotation :
 - ▶ 29 types de relations
 - ▶ 50 pages
- ▶ décisions contre-intuitives :

[...] une forge qui connut son apogée au XIXe siècle.

sujet = qui

→ décomposer la complexité de la tâche [Fort et al., 2012b]
≠ simplifier la tâche !

ZOMBI LINGO

BIENVENUE à TOI, JEUNE ZOMBIE !

LE MONDE EST
CONDAMNÉ, TA
TRANSFORMATION EN
ZOMBIE A COMMENCÉ.

POUR SURVIVRE,
SUIS MES RÈGLES,
IDENTIFIE LES TÊTES
ET MANGE-LES.

ATTENTION AUX
PIÈGES, ILS SONT
NOMBREUX !



SIGNE

Prof. Frankenscience.

Jouer

*Pas de limite
pour toi !
Tu accèdes à
toutes les
options, bonus
cachés !*

Karen

*Retrouve ici tes
statistiques, et
compare ton
score avec
celui de tes
amis !*

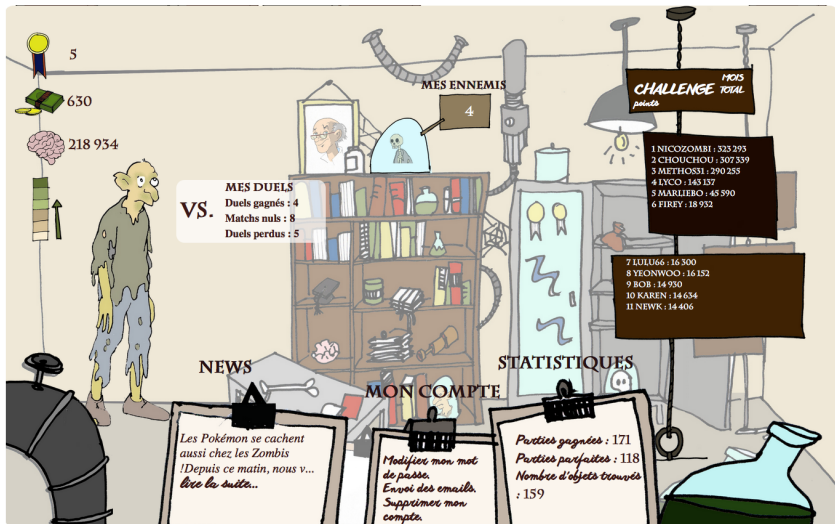
SEMAINE

MOIS
TOTAL

1 nouveau : 38 993
2 nicolef : 16 170

3 Bruno : 11 546
4 testKF : 7 327
5 Karen : 7 012





AUXILIAIRE CAUSATIF
Points : 290 / 998
Annotations : 68

AUXILIAIRE DE TEMPS
Points : 700 / 998
Annotations : 1434

CHOIX DU CORPUS
JEU
FORMATION
Points : 0
Annotations : 0/120

NIVEAU 1
SUJET
JEU
FORMATION
Points : 4980
Annotations : 22/282

NIVEAU 1
AFFIXE
JEU
FORMATION
Points : 0
Annotations : 0/108

NIVEAU 1
DÉTERMINANT
JEU
FORMATION
Points : 0
Annotations : 0/148

NIVEAU 2
COMPLÈMENT EN «DE»
JEU
FORMATION
Points : 0
Annotations : 0/120

NIVEAU 2
COMPLÈMENT EN «À»
JEU
FORMATION
Points : 0
Annotations : 9/127

NIVEAU 2
OBJET DE LA PRÉPOSITION
JEU
FORMATION
Points : 0
Annotations : 0/213

NIVEAU 2
AUXILIAIRE PASSIF
JEU
FORMATION
Points : 598
Annotations : 9/56

Jouer

Boutique

Forum

Joueurs



Niveau

maximum!

165

Trouve le complément (objet indirect introduit par "à") du verbe indiqué !

10%

Besoin d'aide?



Très jeune, il a fait preuve d'initiative et de courage pour **PARTICIPER** à un sauvetage lors d'inondations.



74

43

57

150

Acheter

1

15

Acheter



3

300

Acheter

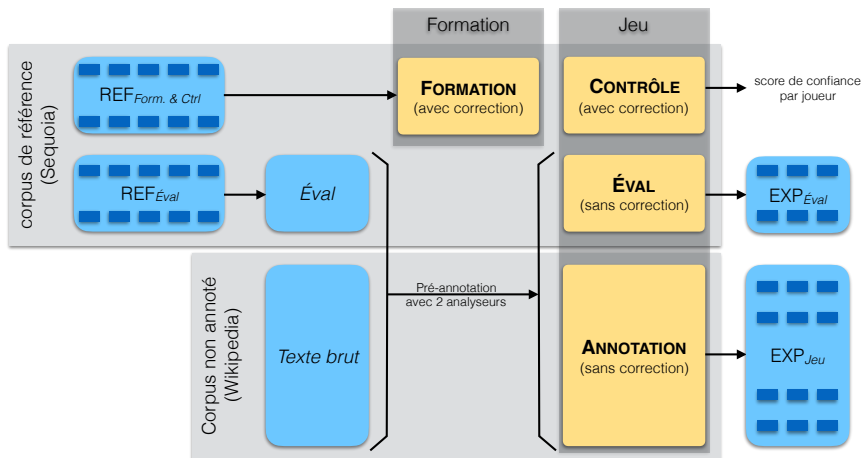


0

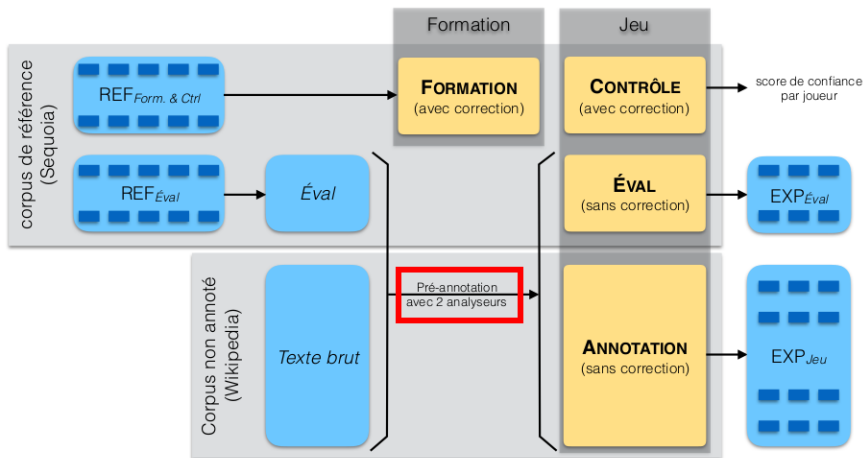
15

Acheter

Organiser une production de qualité



Pré-traitement des données (corpus librement disponibles)



Pré-traitement des données (corpus librement disponibles)

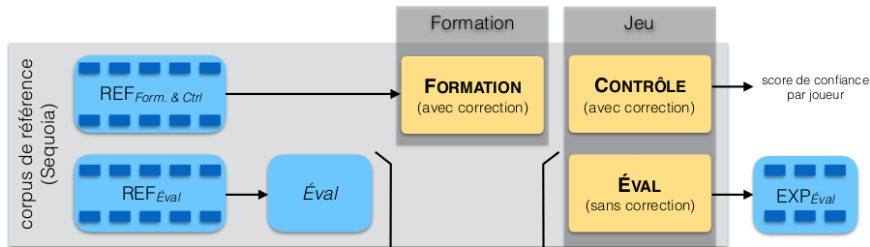
Pré-annotation avec deux analyseurs syntaxiques :

1. un statistique : Talismane [Urieli, 2013]
2. un symbolique, utilisant la ré-écriture de graphes :
FrDep-Parse [Guillaume and Perrier, 2015]

→ les joueurs ne jouent que les éléments pour lesquels les analyseurs donnent des solutions différentes

Formation, contrôle et évaluation

Référence : 3 099 phrases du corpus Sequoia [Candito and Seddah, 2012]



REF _{Formation&Contrôle}	REF _{Éval}	Non utilisées
50 %	25 %	25 %
1 549 phrases	776 phrases	774 phrases

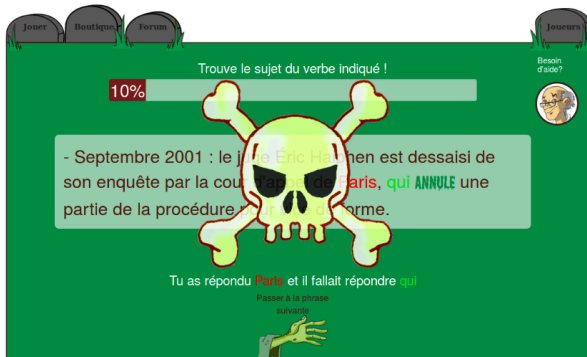
- ▶ REF_{Formation&Contrôle} est utilisé pour former les joueurs
- ▶ REF_{Éval} est utilisé comme un corpus brut, pour évaluer les annotations produites

Formation des joueurs



Obligatoire pour **chaque** relation

- ▶ les phrases sont extraites du corpus REF_{Formation&Contrôle}
- ▶ les erreurs sont signalées et corrigées



Gestion de la fatigue cognitive et des joueurs au long cours

Mécanisme de contrôle

Des phrases du corpus REF_{Formation&Contrôle} sont proposées régulièrement :

- ▶ si le joueur ne trouve pas la bonne réponse, son erreur lui est signalée et la solution montrée

Ils ont été reçus à la boulangerie Leroy **POUR** visiter le fournil et **surtout** pétrir la pâte afin de confectionner de délicieux pains au chocolat qu'ils ont dégustés à l'heure du goûter avec un verre de jus de fruit.

Tu as répondu **surtout** et il fallait répondre **visiter**

Il te reste 2 essais avant de devoir refaire le tutoriel de ce phénomène

⚠ Je ne suis pas d'accord

Passer à la phrase
suivante



- ▶ après un certain nombre d'erreurs sur une même relation, le joueur n'est plus autorisé à jouer cette relation et doit refaire la formation

Gestion de la fatigue cognitive et des joueurs au long cours

Mécanisme de contrôle

Des phrases du corpus REF_{Formation&Contrôle} sont proposées régulièrement :

- ▶ si le joueur ne trouve pas la bonne réponse, son erreur lui est signalée et la solution montrée
- ▶ après un certain nombre d'erreurs sur une même relation, le joueur n'est plus autorisé à jouer cette relation et doit refaire la formation

- 1er **FÉVRIER** 1995 : Jean-Paul Schimpf, un ami intime de Didier Schuller, est **arrêté** sur un parking, alors que la dirigeante d'une entreprise d'assainissement disait vouloir lui remettre **une** somme d'argent en liquide.

Tu as répondu **une** et il fallait répondre **arrête**

Tu as un peu oublié comment jouer ce phénomène. Pour continuer à jouer sur celui-ci, tu vas devoir refaire le tutoriel correspondant.

 Je ne suis pas d'accord

[Retourner au menu](#)



Gestion de la fatigue cognitive et des joueurs au long cours

Mécanisme de contrôle

Des phrases du corpus REF_{Formation&Contrôle} sont proposées régulièrement :

- ▶ si le joueur ne trouve pas la bonne réponse, son erreur lui est signalée et la solution montrée
 - ▶ après un certain nombre d'erreurs sur une même relation, le joueur n'est plus autorisé à jouer cette relation et doit refaire la formation
- nous en déduisons un **niveau de confiance** dans le joueur pour cette relation

Production : taille du corpus obtenu

comparée à celle des autres corpus annotés en syntaxe pour le français

Au 10 juillet 2016 [Guillaume et al., 2016] :

- ▶ 647 joueurs (992 au 2 mars 2017)
- ▶ qui ont produit 107 719 annotations (216 344 au 2 mars 2017)

	Sequoia 7.0	UD-French 1.3	FTB-UC	FTB-SPMRL	ZL
Phrases	3 099	16 448	12 351	18 535	5 221
Tokens	67 038	401 960	350 947	557 149	128 046
Tok./phr.	21,6	24,4	28,4	30,1	24,5

Production : taille du corpus obtenu

comparée à celle des autres corpus annotés en syntaxe pour le français

Au 10 juillet 2016 [Guillaume et al., 2016] :

- ▶ 647 joueurs (992 au 2 mars 2017)
- ▶ qui ont produit 107 719 annotations (216 344 au 2 mars 2017)

	Sequoia 7.0 <i>libre</i>	UD-French 1.3 <i>libre</i>	FTB-UC <i>non libre</i>	FTB-SPMRL <i>non libre</i>	ZL <i>libre</i>
Phrases	3 099	16 448	12 351	18 535	5 221
Tokens	67 038	401 960	350 947	557 149	128 046
Tok./phr.	21,6	24,4	28,4	30,1	24.5

Production : taille du corpus obtenu

comparée à celle des autres corpus annotés en syntaxe pour le français

Au 10 juillet 2016 [Guillaume et al., 2016] :

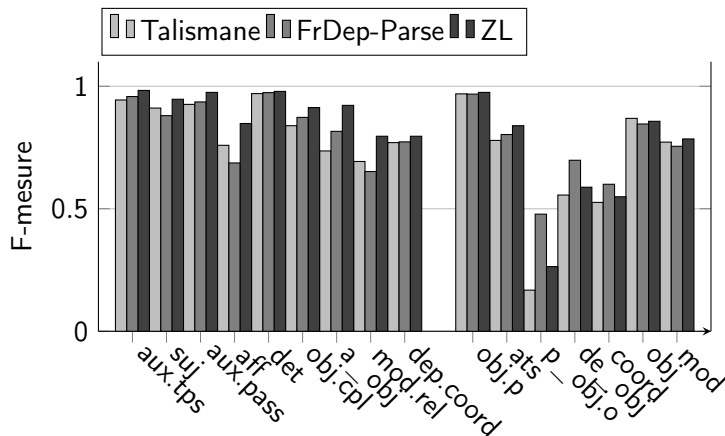
- ▶ 647 joueurs (992 au 2 mars 2017)
- ▶ qui ont produit 107 719 annotations (216 344 au 2 mars 2017)

	Sequoia 7.0 <i>libre</i> validé	UD-French 1.3 <i>libre</i> <i>erreurs</i>	FTB-UC <i>non libre</i> validé	FTB-SPMRL <i>non libre</i> validé	ZL <i>libre</i> validé
Phrases	3 099	16 448	12 351	18 535	5 221
Tokens	67 038	401 960	350 947	557 149	128 046
Tok./phr.	21,6	24,4	28,4	30,1	24.5

+ ressource dynamique (en croissance constante)

Évaluer la qualité

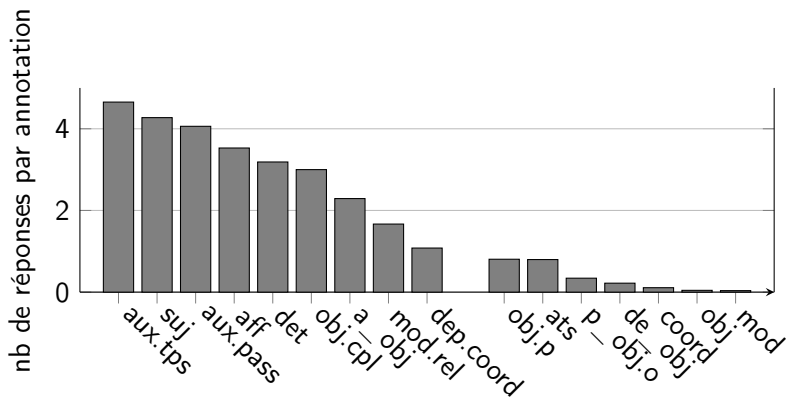
sur le corpus REF_{Eval}



NB : partie gauche de la figure = densité > 1

Densité des annotations

sur le corpus REF_{Eval}



→ besoin de **plus** d'annotations sur certaines relations

Traitement Automatique des Langues (TAL) et annotation manuelle

Formaliser l'annotation manuelle de corpus

Myriadisation et production de données

Des limites du travail parcellisé

Myriadiser des tâches complexes : ZombiLingo

Un savoir-faire complexe

Dépasser les mythes de la myriadisation

Être conscient des contraintes

Conclusion et perspectives

Visibilité...

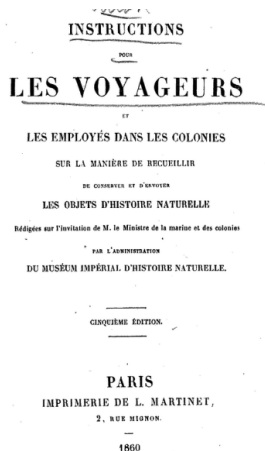
Depuis l'avènement du Web 2.0 :

- ▶ accès facilité à une masse de personne innédite
- ▶ possibilité d'interagir avec la page visitée (Web 2.0, dit «social»)

Exemples : Wikipédia, Projet Gutenberg (*Distributed proofreaders*)

... n'est pas découverte

Instructions pour les voyageurs et les employés des colonies



Science participative :

- ▶ publié par le Museum National d'Histoire Naturelle
- ▶ pour que les voyageurs et les employés des colonies :

"[fassent] connaître les résultats de leurs propres expériences, afin d'en profiter et d'en faire profiter le monde savant"

- ▶ première édition : 1824

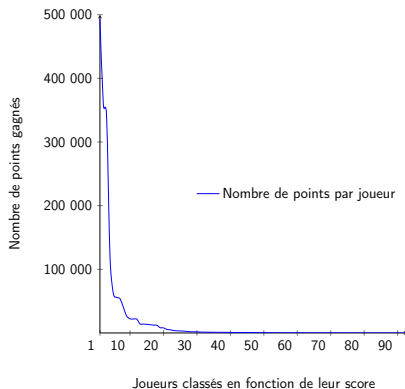
Autres exemples

- ▶ Ligue de Protection des Oiseaux¹ :
 - ▶ suivis des populations d'oiseaux
 - ▶ depuis plus d'un siècle
 - ▶ 5 000 bénévoles actifs
- ▶ Longitude Prize² (1714) :
 - ▶ prix octroyé par le gouvernement britannique à qui inventerait une méthode simple et pratique permettant de déterminer la longitude d'un navire
 - ▶ existe encore : thématique de 2014 = «Global antibiotics resistance»

1. <https://www.lpo.fr/partager-vos-observations/partagez-vos-observations>

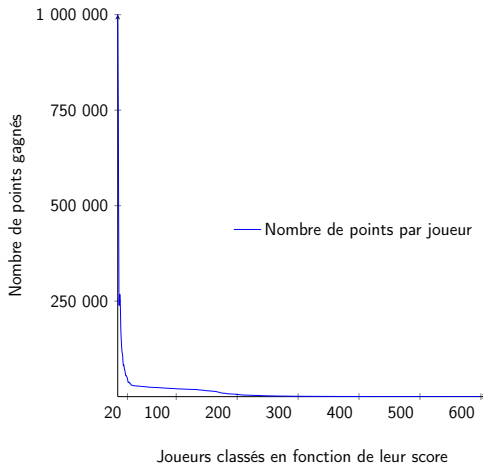
2. <https://longitudeprize.org/>

Une foule de joueurs? Phrase Detectives [Chamberlain et al., 2013]



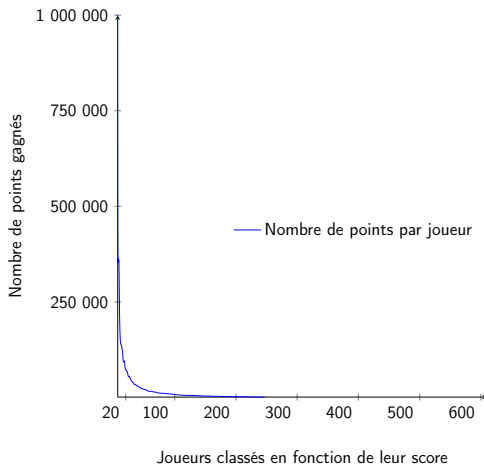
Nombre de joueurs sur *Phrase Detectives* en fonction de leur classement en points (février 2011 - février 2012)

Une foule de joueurs ? JeuxDeMots



Nombre de joueurs sur *JeuxDeMots* en fonction de leur classement en points
(source : <http://www.jeuxdemots.org/generateRanking-4.php>)

Une foule de joueurs ? ZombiLingo [Fort, 2017]



Joueurs sur ZombiLingo en fonction de leur classement en points (mars 2016).

Une foule de travailleurs ? [Fort et al., 2011]

Nombre de *Turkers* actifs sur Amazon Mechanical Turk (MTurk) :

- ▶ nombre enregistré sur le site : plus de 500 000
 - ▶ 80 % des tâches (HIT) sont réalisées par les 20 % de *Turkers* les plus actifs [Deneme, 2009]
- ⇒ réellement actifs : entre 15 059 et 42 912

Experts vs non-experts

Exemple de l'annotation en entités nommées dans un corpus de microbiologie :

- ▶ experts du domaine ?
 - ▶ du corpus (microbiologie) ?
 - ▶ de l'application (TAL) ?

Experts vs non-experts

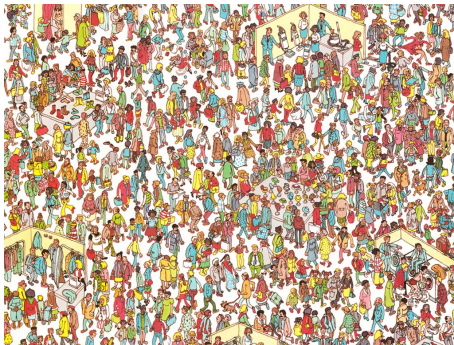
Exemple de l'annotation en entités nommées dans un corpus de microbiologie :

- ▶ experts du domaine ?
 - ▶ du corpus (microbiologie) ?
 - ▶ de l'application (TAL) ?

→ experts de la tâche

Myriadisation

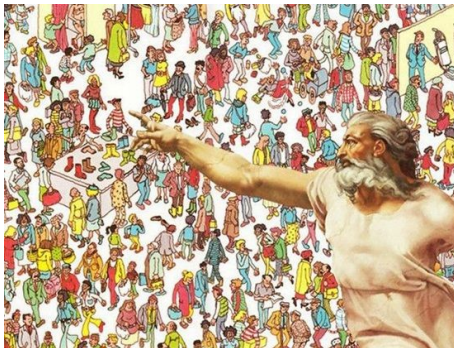
Faire annoter des « non-experts » ?



Myriadisation

~~Faire annoter des « non-experts » ?~~

→ Trouver/former des experts (de la tâche) dans la foule



Créer des données de qualité vs créer des fonctionnalités fun
préserver le cercle vertueux n'est pas si facile



Créer des données de qualité vs créer des fonctionnalités fun

préserver le cercle vertueux n'est pas si facile



phrase qui disparaît soudainement dans ZombiLingo :

- + le joueur est surpris : fun !
- le joueur clique n'importe où : ressource de mauvaise qualité

Créer des données de qualité vs créer des fonctionnalités fun

préserver le cercle vertueux n'est pas si facile



phrase qui disparaît soudainement dans ZombiLingo :

- + le joueur est surpris : fun !
- le joueur clique n'importe où : ressource de mauvaise qualité

joueur qui a trouvé une faille dans JeuxDeMots [Lafourcade and Joubert, 2008] pour obtenir du temps :

- + crée de la meilleure donnée : bonne qualité
- génère de l'envie et de la colère dans la communauté de joueurs : mauvais pour le jeu

Les sciences *participatives* : pas si évident

- ▶ Bisame (alsacien) : sans ludification, les participants ne reviennent pas (article soumis à Games4NLP)
- ▶ ZombiLingo : les gros joueurs sont motivés par le jeu (article soumis à Games4NLP)
- ▶ [Munro, 2013] : le volontariat se tarit rapidement, même pour une action humanitaire (Haïti)

La ludification permet de faire revenir les joueurs,

mais il faut les relancer continuellement

Et l'éthique dans tout ça ?

Discussion pour la seconde partie !

Traitement Automatique des Langues (TAL) et annotation manuelle

Formaliser l'annotation manuelle de corpus

Myriadisation et production de données

Des limites du travail parcellisé

Myriadiser des tâches complexes : ZombiLingo

Un savoir-faire complexe

Conclusion et perspectives

Améliorer la jouabilité

Quelques pistes :

- ▶ proposer plus à explorer et collectionner
- ▶ bâtir un vrai scénario
- ▶ construire une communauté autour du jeu

Améliore la ressource exportée

Tester l'influence :

- ▶ du score attribué à la pré-annotation
- ▶ du niveau du joueur dans le jeu
- ▶ du niveau de confiance que nous avons dans le joueur pour chaque type de relation

Vers d'autres langues

et d'autres types d'annotation

Nouvelles langues :

- ▶ anglais
- ▶ langues peu dotées

Alice Millour (doctorante)

Nouveaux types d'annotation :

- ▶ parties du discours (**POS**),
- ▶ unités polylexicales (jeu Rigor Mortis),
- ▶ récolte de corpus, etc.



Apprendre ET produire



Participation (porteuse pour la France) à l'action Cost enetCollect :

combinaison de l'apprentissage des Langues avec du crowdsourcing explicite ou implicite dans le but de créer du matériel pédagogique mais aussi des ensemble de données tels que des ressources pour le TAL

ZombiLingo, c'est surtout une équipe



Bruno Guillaume (CR Inria)



Nicolas Lefèvre (ingénieur Inria)

Inria



THANK  YOU



<https://github.com/zombilingo>



<http://zombilingo.org/export>

Annexes

Compléments sur AMT

Foule Factory

Motiver les joueurs

Plus gros Requesters (en \$ dépensés)

<http://www.mturk-tracker.com>, février 2016



[General](#)

[Arrivals](#)

[Completions](#)

[Top requesters](#)

[Active requesters](#)

[Demographics](#)

[Search](#)

[API](#)

[Old version](#)

Top-1000 Requesters, report for February 3, 2016 to March 4, 2016

Requester name	hits	reward
Colleen Harmeling	1451	\$50,725.82
Jon Brellig	84646	\$12,939.88
CastingWords	12029	\$10,499.14
AI Indoors Project	26934	\$9,669.04
NobilisCV	16337	\$7,122.20
Groupsolver	3359	\$6,563.45
HITimeUp	14374	\$2,872.14
Amazon Requester Inc - browse classification	40297	\$2,429.52
p9r	58274	\$2,199.26
Dialogue Systems	6899	\$1,935.55

« Previous

10

25

50

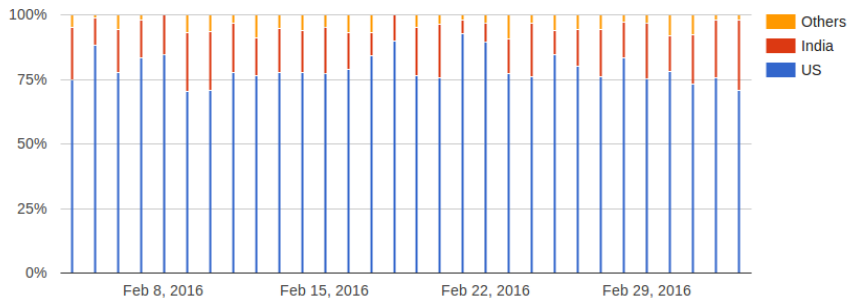
100

Page 1 of 100

Next »

Provenance des *Turkers*

<http://www.mturk-tracker.com>, février 2016



Foule Factory : version française plus éthique d'AMT ?

«[...] nous avons (contrairement à Amazon) mis en place plusieurs leviers pour s'assurer d'une certaine éthique dont l'auteure pense qu'elle nous fait défaut. a/ Foulefactory n'est ouvert qu'aux contributeurs français. b/ Le plafond annuel de rémunération est de 3 000 euros - nous ne souhaitons pas que foulefactory constitue une source primaire de revenus pour nos contributeurs. c/ Le taux horaire moyen est compris entre 10 et 15 euros, c'est le prix minimum que nous demandons à nos clients de respecter. d/ Si un projet rémunère moins, nos fumeurs (contributeurs de foulefactory) nous remontent l'information et nous adressons un avertissement au client. e/ Si une tâche réalisée par un fumeur est refusée (et donc non-rémunérée), ce dernier peut faire une « réclamation » en 1 clic. f/ Si plusieurs fumeurs adressent une réclamation à un client, nous adressons un avertissement à ce dernier. g/ Après 3 avertissements, un client est banni de la plateforme foulefactory.[...] »

[extrait du commentaire laissé par D. Benoïd co-fondateur de FF, sur mon article Journal du CNRS / Libération sur la myriadisation]

Indices extraits du forum <http://www.fouleurs.com/>

« c/ Le taux horaire moyen est compris entre 10 et 15 euros, c'est le prix minimum que nous demandons à nos clients de respecter. »



Mel Anle 10/15 euros de l'heure... n'importe quoi... très très rares sont les projets qui atteignent ce montant....

J'aime · Répondre · 29 juillet 2015, 14:02



Dominique Bonhomme oui mais tu regardes l'info, ils sont dans la légalité, dans la mesure où il mesure ta tâche 10 s

J'aime · Répondre · 29 juillet 2015, 16:13



Mel Anle la plupart du temps, les projets sont plutôt autour de 7-9 euros de l'heure

J'aime · Répondre · 29 juillet 2015, 16:15 · Modifié



foulefactory

5 0 2 days ago

Identifiant: #5326

Titre: Normalization 190220161709

Catégorie: Rédaction

Temps par tâche: 00:00:05

Défraiment: 0.01 €

Taux horaire: 7.2 €

Nombre de Fouleur: 1

Date de creation: 2/19/2016 5:12:48 PM

Indices extraits du forum <http://www.fouleurs.com/>

« f/ Si plusieurs fouleurs adressent une réclamation à un client, nous adressons un avertissement à ce dernier. »

lipqil

9 0 on 9 février 2016



Pour moi aussi, problème sur le 5294. Dans « Mes réalisations » j'ai 326 tâches validées (pour 2.46€), alors que dans le détail du projet il a de mentionné: **Validées: 41, Refusées: 0, En attente: 0**. Alors ou sont passées les 285 tâches restantes. C'EST DE LA PURE ARNAQUE!!

Pour le projet 5234 les tâches sont en attente, la date de fin est 30/11/2017, On peut toujours espérer. Il n'y aura pas de paiement.

Et comme toujours, aucune réponse de FF sur les multiples explications que je leur ai envoyées sur de nombreux problèmes.



Samuel

0 0 on 9 février 2016

J'ai contacté FF j'attends qu'il me réponde.

lipqil

0 0 on 9 février 2016

C'est bien, mais il ne faut pas espérer avoir une réponse. Qui a eu des réponses de leur part ces dernières semaines?

Moi, aucune depuis des mois.



Samuel

0 0 on 9 février 2016

Je vous redit



Samuel

0 0 on 9 février 2016

quand j'aurais une réponse

lipqil

0 0 on 9 février 2016

Oui, oui. J'entends bien. Comme on dit, l'espoir fait vivre

Foule Factory : dépassé par son succès ?

Certains des problèmes sont sans doute dus à l'afflux soudain et massif de travailleurs suite à une émission de TV.

Cependant :

- ▶ Déficit de tâches à réaliser : du fait de la loi de l'offre et de la demande, il paraît peu probable que les déposants soient inquiétés...
- ▶ Taux horaire défini comment ?

Motivation vs volition [Fenouillet et al., 2009]

motiver les gens à **venir jouer** (motivation)

puis

motiver les gens à **continuer à jouer** (volition)

Motiver les gens, en général

par exemple à trahir (leur pays) ou à espionner

Analyse de la CIA, MICE :

- ▶ Money : récompense

Motiver les gens, en général

par exemple à trahir (leur pays) ou à espionner

Analyse de la CIA, MICE :

- ▶ Money : récompense
- ▶ Ideology : intérêt

Motiver les gens, en général

par exemple à trahir (leur pays) ou à espionner

Analyse de la CIA, MICE :

- ▶ Money : récompense
- ▶ Ideology : intérêt
- ▶ Constraint : contrainte, légère et légale

Motiver les gens, en général

par exemple à trahir (leur pays) ou à espionner

Analyse de la CIA, MICE :

- ▶ Money : récompense
- ▶ Ideology : intérêt
- ▶ Constraint : contrainte, légère et légale
- ▶ Ego : place dans la communauté

Types de joueurs selon [Bartle, 1996]

- ▶ *Achievers* : aiment réussir dans le jeu

Types de joueurs selon [Bartle, 1996]

- ▶ *Achievers* : aiment réussir dans le jeu
- ▶ *Explorers* : aiment connaître les coins cachés du jeu

Types de joueurs selon [Bartle, 1996]

- ▶ *Achievers* : aiment réussir dans le jeu
- ▶ *Explorers* : aiment connaître les coins cachés du jeu
- ▶ *Socializers* : aiment interagir avec les autres

Types de joueurs selon [Bartle, 1996]

- ▶ *Achievers* : aiment réussir dans le jeu
- ▶ *Explorers* : aiment connaître les coins cachés du jeu
- ▶ *Socializers* : aiment interagir avec les autres
- ▶ *Killers* : aiment attaquer les autres joueurs

Motiver les joueurs en fonction de leur type



© Gamified UK 2014



Bartle, R. (1996).

Hearts, clubs, diamonds, spades : Players who suit MUDs.
The Journal of Virtual Environments, 1(1).



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.



Böhmová, A., Hajič, J., Hajičová, E., and Hladká, B. (2001).

The prague dependency treebank : Three-level annotation scenario.

In Abeillé, A., editor, Treebanks : Building and Using Syntactically Annotated Corpora. Kluwer Academic Publishers.



Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015).

A large annotated corpus for learning natural language inference.



Candito, M. and 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., and Poesio, M. (2013).

Using games to create language resources : Successes and limitations of the approach.

In Gurevych, I. and Kim, J., editors, The People's Web Meets NLP, Theory and Applications of Natural Language Processing, pages 3–44. Springer Berlin Heidelberg.



Chamberlain, J., Poesio, M., and Kruschwitz, U. (2008).

Phrase Detectives : a web-based collaborative annotation game.

In Proceedings of the International Conference on Semantic Systems (I-Semantics'08), Graz, Autriche.



Couillault, A. and Fort, K. (2013).

Charte éthique et big data : parce que mon corpus le vaut bien !

In Actes de colloque international Corpus et Outils en Linguistique, Langues et Parole : Statuts, Usages et Mésusages, Strasbourg, France.



Deneme (2009).

How many turkers are there ?

[http ://groups.csail.mit.edu/uid/deneme/](http://groups.csail.mit.edu/uid/deneme/).



Desrosières, A. (2008).

Pour une sociologie historique de la quantification :
L'Argument statistique I.

Presses de l'École des Mines de Paris.



Fenouillet, F., Kaplan, J., and Yennek, N. (2009).

Serious games et motivation.

In 4ème Conférence francophone sur les Environnements Informatiques pour l'Apprentissage Humain (EIAH'09), vol. Actes de l'Atelier "Jeux Sérieux : conception et usages", pages 41–52, Le Mans, France.



Fort, K. (2017).

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

CORELA (cognition, représentation, langage), page 12 p.



Fort, K., Adda, G., and Cohen, K. B. (2011).

Amazon Mechanical Turk : Gold mine or coal mine ?

Computational Linguistics (editorial), 37(2) :413–420.



Fort, K., François, C., Galibert, O., and Ghribi, M. (2012a).

Analyzing the impact of prevalence on the evaluation of a manual annotation campaign.

In Proceedings of the International Conference on Language Resources and Evaluation (LREC), Istanbul, Turquie.

7 pages.



Fort, K., Nazarenko, A., and Rosset, S. (2012b).

Modeling the complexity of manual annotation tasks : a grid of analysis.

In Proceedings of the International Conference on Computational Linguistics (COLING), pages 895–910, Mumbai, Inde.



Geiger, D., Seedorf, S., Schulze, T., Nickerson, R. C., and Schader, M. (2011).

Managing the crowd : Towards a taxonomy of crowdsourcing processes.

In AMCIS 2011 Proceedings.



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.



Grouin, C., Rosset, S., Zweigenbaum, P., Fort, K., Galibert, O., and Quintard, L. (2011).

Proposal for an extension of traditional named entities : From guidelines to evaluation, an overview.

In Proceedings of the 5th Linguistic Annotation Workshop, pages 92–100, Portland, Oregon, USA.

Poster.



Guentchéva, Z. and Desclés, J.-P. (1991).

Test et acceptabilité.

Histoire Épistémologie Langage, 13(2) :9–25.



Guillaume, B., Fort, K., and Lefebvre, N. (2016).

Crowdsourcing complex language resources : Playing to annotate dependency syntax.

In Proceedings of the International Conference on Computational Linguistics (COLING), Osaka, Japon.



Guillaume, B. and Perrier, G. (2015).

Dependency Parsing with Graph Rewriting.

In

Proceedings of IWPT 2015, 14th International Conference on Parsing
pages 30–39, Bilbao, Spain.



Gupta, N., Martin, D., Hanrahan, B. V., and O'Neill, J.
(2014).

Turk-life in india.

In Proceedings of the 18th International Conference on
Supporting Group Work, GROUP '14, pages 1–11, New York,
NY, USA. ACM.



Habert, B. (2000).

Corpus. Méthodologie et applications linguistiques, chapter
Détournements d'annotation : armer la main et le regard,
pages 106–120.

Champion and Presses Universitaires de Perpignan.



Habert, B. (2005).

Portrait de linguiste(s) à l'instrument.

Texto !, vol. X(4).



Habert, B. (2008).

Observer, aujourd'hui, c'est manipuler.

In François, J., editor,

Observations et manipulations en linguistique : entre concurrence et c

volume 16 of Mémoires de la Société de linguistique de Paris.

Nouvelle série, pages 33–53. Peeters, Paris, France.



Howe, J. (2006).

The rise of crowdsourcing.

Wired Magazine, 14(6).



Ipeirotis, P. (2010).

The new demographics of mechanical turk.

<http://behind-the-enemy-lines.blogspot.com/2010/03/new-demographics-of-mechanical-turk.html>.



Khatib, F., DiMaio, F., Cooper, S., Kazmierczyk, M., Gilski, M., Krzywda, S., Zabranska, H., Pichova, I., Thompson, J., Popović, Z., et al. (2011).

Crystal structure of a monomeric retroviral protease solved by protein folding game players.

Nature structural & molecular biology, 18(10) :1175–1177.



Lafourcade, M. and Joubert, A. (2008).

JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes.

In Actes de Journées internationales d'Analyse statistique des Données Textuelles (JADT), Lyon, France.



Lease, M., Hullman, J., Bigham, J. P., Bernstein, M. S., Kim, J., Lasecki, W., Bakhshi, S., Mitra, T., and Miller, R. C. (2013).

Mechanical turk is not anonymous.

Technical report.



Leech, G. (1997).

Corpus annotation : Linguistic information from computer text corpora, chapter Introducing corpus annotation, pages 1–18.

Longman, Londres, Angleterre.



Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993).
Building a large annotated corpus of English : The Penn
Treebank.

Computational Linguistics, 19(2) :313–330.



Milner, J. (1989).

Introduction à une science du langage.

Des travaux. Editions du Seuil.



Munro, R. (2013).

Crowdsourcing and the crisis-affected community : lessons
learned and looking forward from mission 4636.

Journal of Information Retrieval, 16(2).



Novotney, S. and Callison-Burch, C. (2010).

Cheap, fast and good enough : automatic speech recognition
with non-expert transcription.

In Proceedings of the Annual Conference of the North
American Chapter of the Association for Computational

Linguistics (NAACL), HLT'10, pages 207–215, Stroudsburg, PA, USA. Association for Computational Linguistics.



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.



Sagot, B. (2010).

The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French.

In

7th international conference on Language Resources and Evaluation (Valletta, Malta.



Sagot, B., Fort, K., Adda, G., Mariani, J., and 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.



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.



Urieli, A. (2013).

Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit.

PhD thesis, Université de Toulouse II le Mirail, France.



Wais, P., Lingamneni, S., Cook, D., Fennell, J., Goldenberg, B., Lubarov, D., Marin, D., and Simons, H. (2010).

Towards building a high-quality workforce with mechanical turk.

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