

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

Faire produire des données langagières
de qualité
en masse

Karën Fort

13 mai 2015



Licence et outils



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]

The Google logo is displayed in its characteristic multi-colored font. The letters are blue, red, yellow, blue, green, and red from left to right. A small trademark symbol (TM) is located at the top right of the letter 'e'.

Traitement Automatique des Langues (TAL) et annotation manuelle

Des applications dans nos vies

L'annotation manuelle dans le TAL

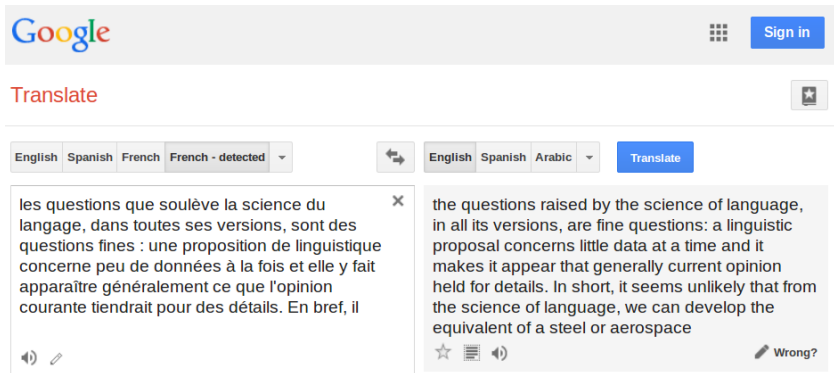
Un sujet de recherche émergent

Formaliser l'annotation manuelle de corpus

Jeux ayant un but et production de données

Conclusion et perspectives

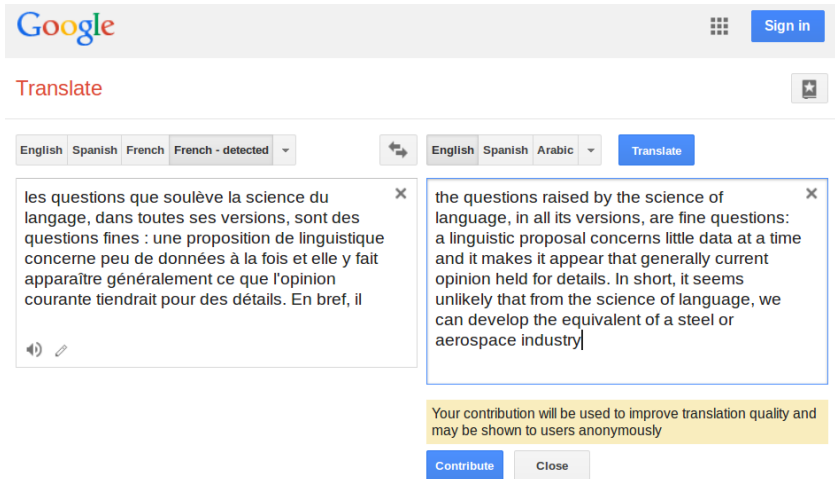
«Traduction» automatique



The screenshot shows the Google Translate web interface. At the top left is the Google logo, and at the top right is a 'Sign in' button. Below the logo is the word 'Translate' in red. The interface features two language selection dropdown menus: the first is set to 'French - detected' and the second to 'English'. A blue 'Translate' button is positioned to the right of the second dropdown. The main content area is split into two columns. The left column contains the original French text: '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'. Below this text are icons for a speaker and a pencil. The right column contains the translated English text: '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'. Below this text are icons for a star, a list, a speaker, and a 'Wrong?' button with a pencil icon.

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

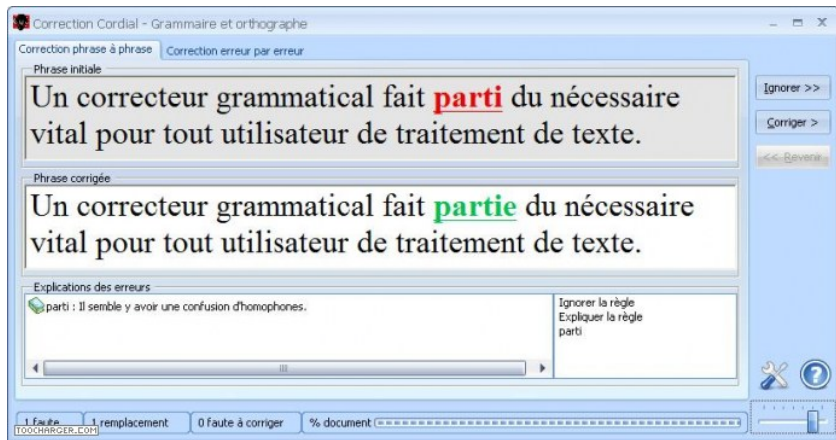
«Traduction» automatique



The screenshot shows the Google Translate web interface. At the top left is the Google logo. To the right is a 'Sign in' button. Below the logo is the word 'Translate' in red. A language selection bar shows 'English', 'Spanish', 'French', and 'French - detected' with a dropdown arrow. To the right of this bar is a bidirectional arrow icon and another language selection bar showing 'English', 'Spanish', and 'Arabic' with a dropdown arrow. A blue 'Translate' button is positioned to the right of the second language bar. Below these elements are two text boxes. The left box contains the French text: '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'. Below this text are icons for a speaker and an eraser. The right box contains the English translation: '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'. Below the translation boxes is a yellow banner with the text: 'Your contribution will be used to improve translation quality and may be shown to users anonymously'. At the bottom of the banner are two buttons: 'Contribute' (blue) and 'Close' (grey).

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

Correction orthographique et grammaticale



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

Extraction d'entités nommées

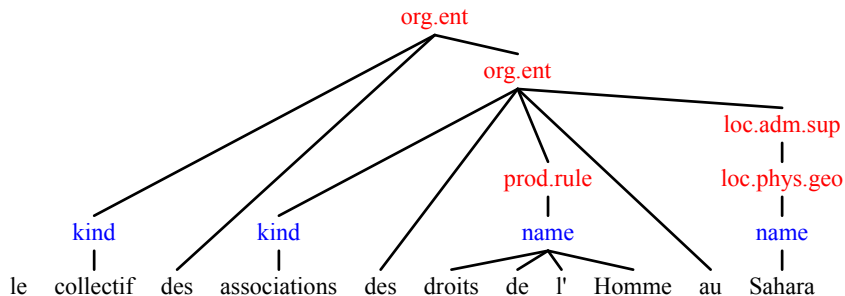
<ENAMEX TYPE="PERSON">Henri</ENAMEX> a acheté **<NUMEX TYPE="QUANTITY">300</NUMEX>** actions de la société **<ENAMEX TYPE="ORGANIZATION">AMD</ENAMEX>** en **<TIMEX TYPE="DATE">2006</TIMEX>**.

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

Extraction d'entités nommées

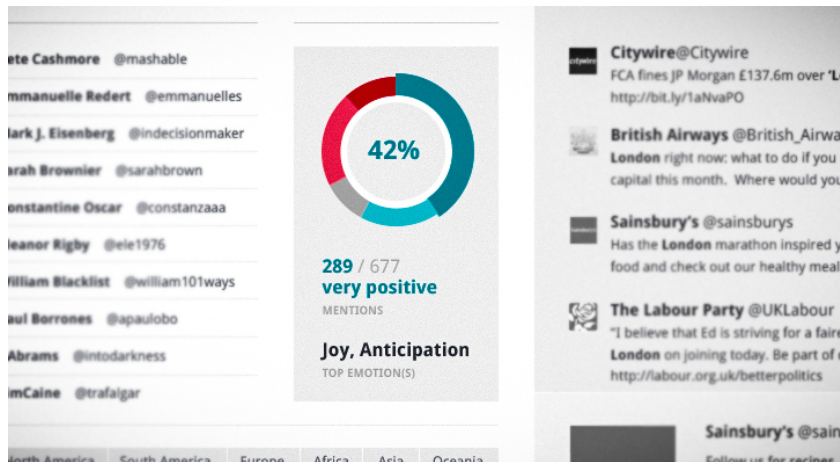
`<ENAMEX TYPE="PERSON">Henri</ENAMEX>` a acheté `<NUMEX TYPE="QUANTITY">300</NUMEX>` actions de la société `<ENAMEX TYPE="ORGANIZATION">AMD</ENAMEX>` en `<TIMEX TYPE="DATE">2006</TIMEX>`.

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

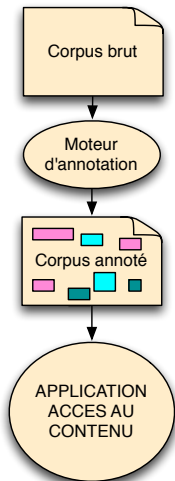
Des espaces de réflexions spécifiques :

- ▶ Journée d'études ATALA Éthique et TAL, nov. 2014
- ▶ Atelier ETeRNAL (Éthique et TAL), juin 2015

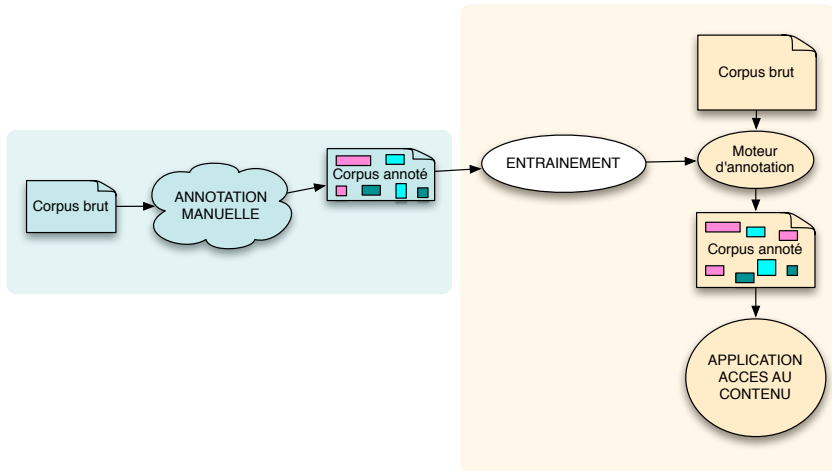
Des propositions pratiques :

- ▶ Charte Éthique et Big Data [Couillault and Fort, 2013]

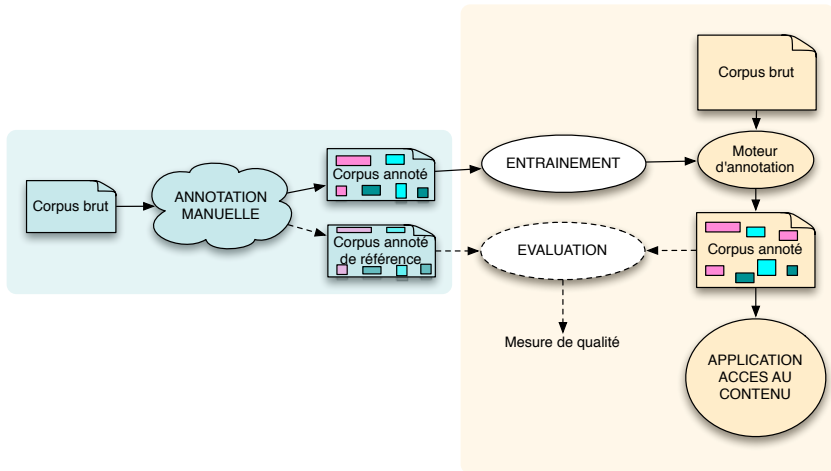
Corpus annotés et Traitement Automatique des Langues



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

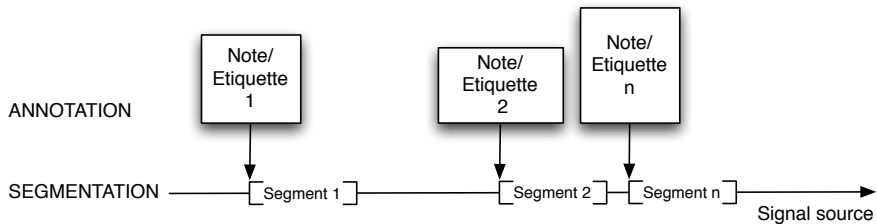
- Identifier les dimensions de complexité de l'annotation

- Outiller à bon escient

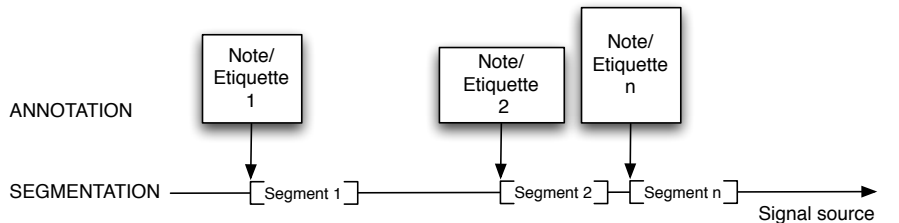
Jeux ayant un but et production de données

Conclusion et perspectives

Annoter ?



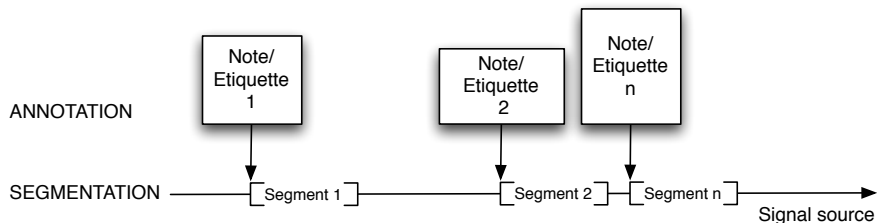
Annoter ?



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

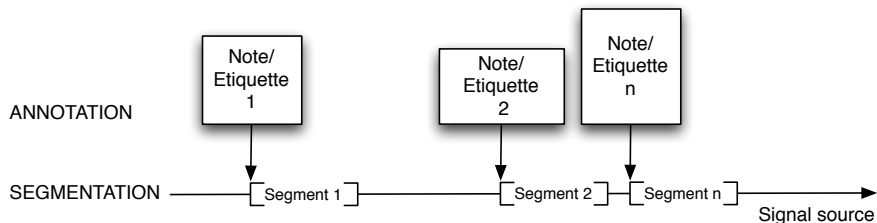
[Benzitoun et al., 2012]

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 de 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 gestionnaire 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)

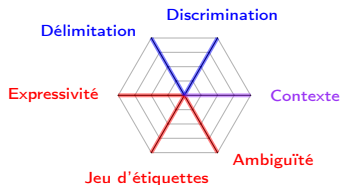
Qu'est-ce qui est complexe ?

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

$$\textit{Expressivité}_{\textit{Renommage}} = 0,25$$

$$\textit{Expressivité}_{\textit{PTB}_{\textit{POS}}} = 0,25$$

Comment annoter ? Dimension du jeu d'étiquettes

Person			Function				
<i>pers.ind</i> (individual person)	(individual)	<i>pers.coll</i> (group of persons)	(group of persons)	<i>func.ind</i> (individual function)	(individual function)	<i>func.coll</i> (collectivity of functions)	(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)	(services)	<i>time.date.abs</i> (absolute date),	<i>time.hour.abs</i> (absolute hour),		
Amount <i>amount</i> (with unit or general object), including duration			<i>time.date.rel</i> (relative date)	<i>time.hour.rel</i> (relative hour)			

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		
<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.hour.abs</i> (absolute hour),	
Amount			<i>time.date.rel</i> (relative date)	<i>time.hour.rel</i> (relative hour)	
<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		
<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)	<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).

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

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

$$Ambiguïté_{Res}_{Renommage} = 0,02$$

Comment annoter ? Degré d'ambiguïté : ambiguïté théorique

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

Définition

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

avec

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

→ 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 !

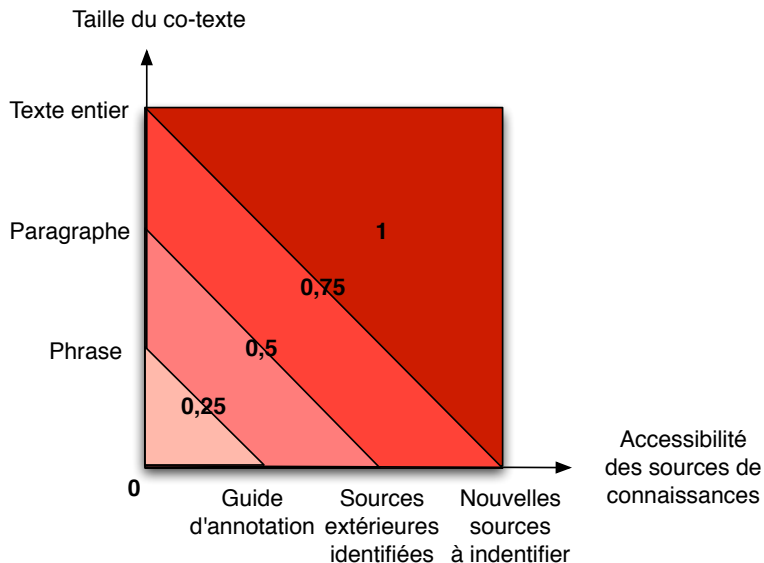
Xavier Gravelaine : Oh la la la la !

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

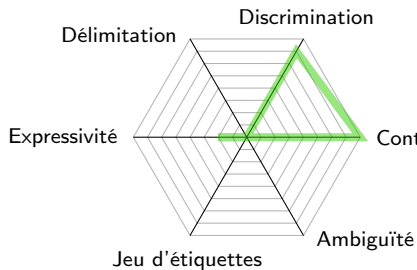
- ▶ guide d'annotation
 - ▶ nomenclatures (Swiss-Prot)
 - ▶ nouvelles sources à trouver (Wikipedia, etc.)

Poids du contexte

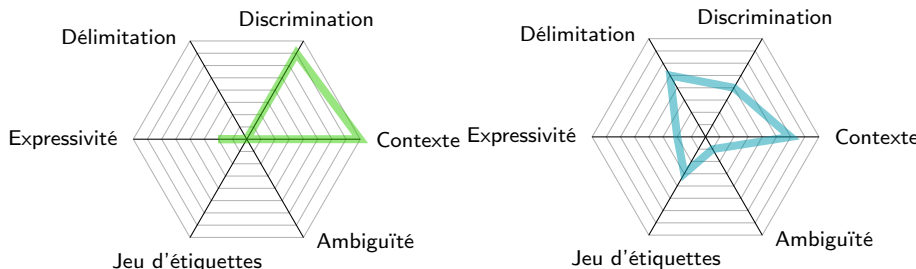


Outiller à bon escient...

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

Jeux ayant un but et production de données

Des résultats étonnants

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

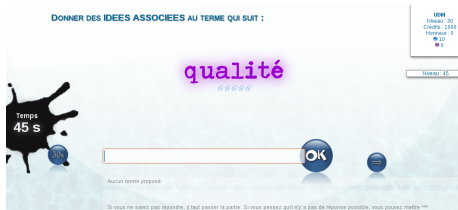
JeuxDeMots : jouer à associer des idées...

... 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
Knitta
22 this week
1 decision
21 agreements
0 extras
22 this month
83 all time
Level: Apprentice
Your rating: 80%

SEARCH CLUES
Where do they live, her and it are likely to refer to something else in the text. Try to find the closest mention of the phrase.
Where do they or them could refer to more than one thing in the text so select more than one phrase if necessary.
Always look for the closest previous mention of the phrase to score maximum agreement points.

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 **closest phrase(s)** if it has been mentioned before.

USERPROFILE
Knitta
21 this week
0 decisions
21 agreements
0 extras
21 this month
81 all time
Level: Apprentice
Your rating: 80%

SEARCH CLUES
Phrases beginning with a, an or Be can refer to two different people.
A, an or subject They can be used to say something about an object. For example 'The policeman delivered a letter' describes the object 'letter' as having the property of being 'the policeman'.
If you think the phrase describes a property try to select the closest phrase it refers to.

→ Profiter des connaissances scolaires de la foule

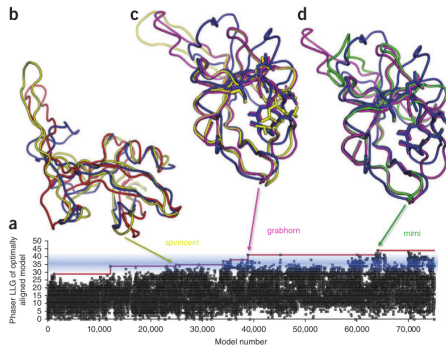
FoldIt : jouer à replier des protéines. . .

. . . 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 :

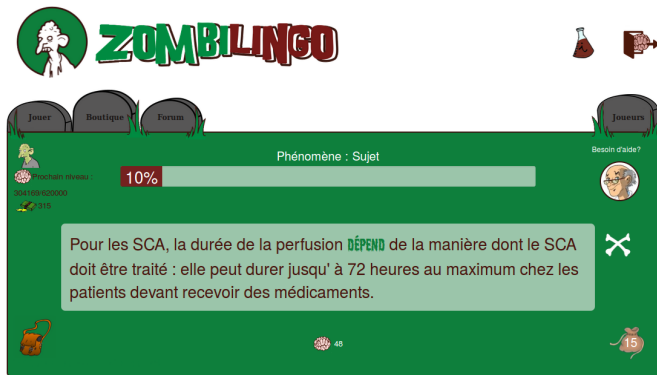
- ▶ trouvée en quelques semaines
- ▶ par étapes
- ▶ par une équipe de joueurs
- ▶ qui permettra la création d'antirétroviraux



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

ZombiLingo : jouer à manger des têtes ...

... pour annoter des corpus en syntaxe de dépendances [Fort et al., 2014]



- ▶ décomposition de la tâche par phénomènes (et non par phrases)
- ▶ tutoriel par phénomène
- ▶ phrases de référence proposées régulièrement

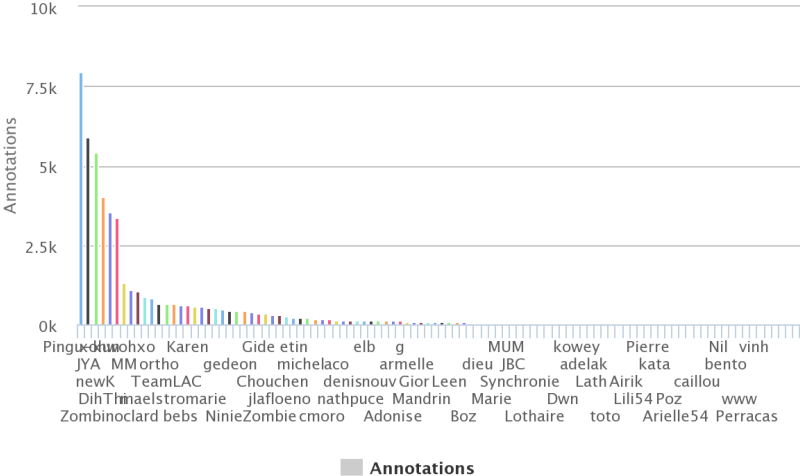
Retour d'expérience : ZombiLingo (après 1 mois)

- ▶ 286 inscrits
- ▶ 6 joueurs ont réalisé plus de 2 500 annotations
- ▶ 11 (sur 18) phénomènes ont plus de 2 500 annotations
- ▶ près de 50 000 annotations en un mois (!)

Un savoir-faire complexe...

Une foule de non-experts d'experts de la tâche

Nombre d'annotations par utilisateur



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

Assurer la qualité de la production

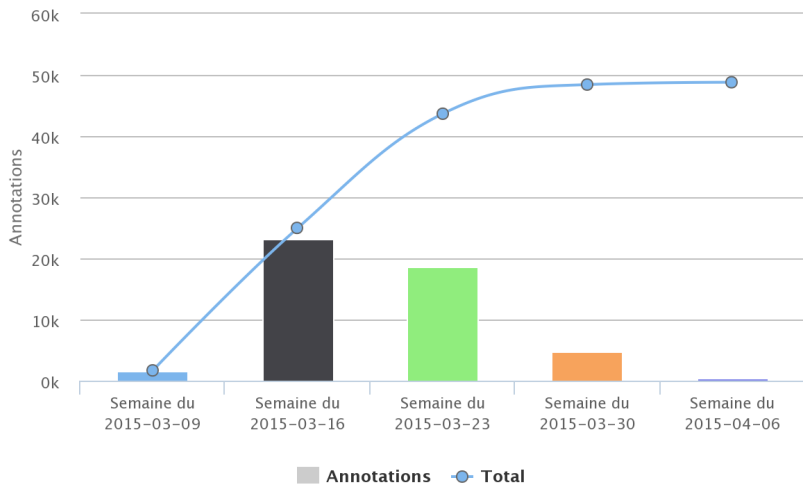
Exemple de ZombiLingo (avant mise en place du retour sur la qualité)

	Nombre d'annotation sur phrases de référence	Nombre d'annotations correctes sur phrases de référence	
Pingu-kun	590	345	58.47%
xohmohxo	909	807	88.78%
JYA	409	361	88.26%
Zombinoclard	220	177	80.45%
newK	1654	1583	95.71%
Bruno	165	154	93.33%

Attirer et faire vivre la communauté de joueurs

Exemple de ZombiLingo (influence de la semaine de la langue française)

Nombre d'annotations par semaine



Highcharts.com

Crowdsourcing éthique : perspectives

Une plate-forme nationale de **sciences participatives** visant à mutualiser les efforts de :

- ▶ communication
- ▶ gestion de la communauté
- ▶ développement
- ▶ hébergement (serveurs)

tout en proposant une **réflexion éthique**

Merci

Merci de jouer !



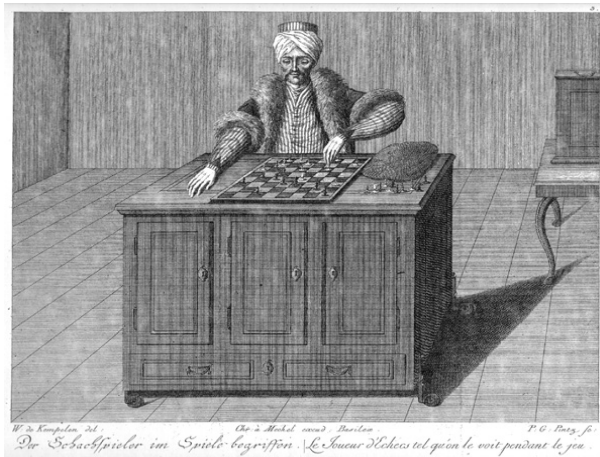
<http://zombilingo.org>

Annexes

Amazon Mechanical Turk : une plate-forme de légendes
Motiver les joueurs

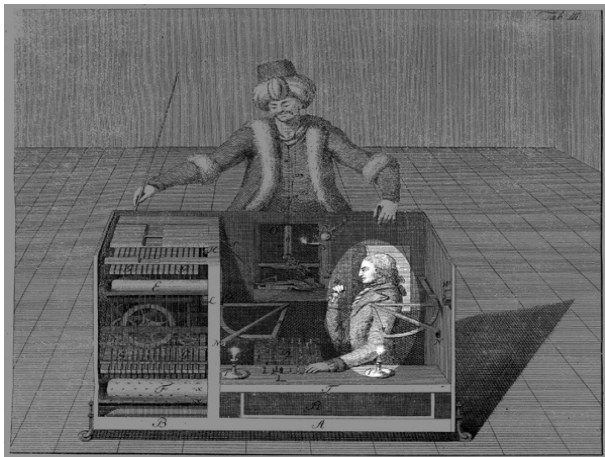
Le «Turc mécanique» de von Kempelen

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



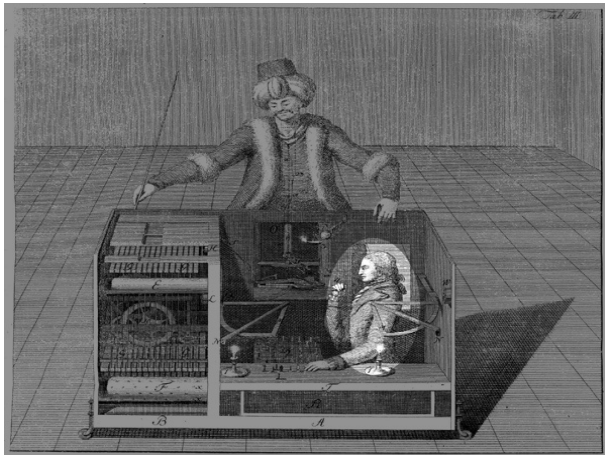
Le «Turc mécanique» de von Kempelen

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)

Amazon Mechanical Turk

MTurk

amazonmechanicalturk
Artificial Intelligence

Your Account

HITS

Qualifications

Already have an account?
Sign in as a Worker | Requester

[Introduction](#) | [Dashboard](#) | [Status](#) | [Account Settings](#)

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.

179,373 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



Get Results from Mechanical Turk Workers

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

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

MTurk est une plate-forme de **myriadisation** : le travail est *externalisé* via le Web et réalisé par de nombreuses personnes (la *foule*), ici les **Turkers**

The screenshot shows the Amazon Mechanical Turk homepage. At the top left is the logo "amazonmechanical turk" with the tagline "Artificial Intelligence". Navigation tabs include "Your Account", "HITS", and "Qualifications". On the top right, there are links for "Already have an account" and "Sign in as a Worker | Request". A central banner reads "Mechanical Turk is a marketplace for work." followed by "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 "179,373 HITS available. View them now." Below this are two main sections: "Make Money by working on HITs" and "Get Results from Mechanical Turk Workers".

amazonmechanical turk
Artificial Intelligence

Your Account | HITS | Qualifications

Already have an account | Sign in as a Worker | Request

Introduction | Dashboard | Status | Account Settings

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.
179,373 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

Get Results
from Mechanical Turk Workers

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

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

MTurk 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 website. At the top left is the logo "amazonmechanical turk" with "Artificial Intelligence" below it. Navigation tabs include "Your Account", "HITS", and "Qualifications". On the top right, it says "Already have an account Sign in as a Worker | Requester". Below the navigation is a banner with the text: "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. 179,373 HITS available. View them now." The main content is split into two columns. The left column is titled "Make Money by working on HITS" and contains a green-bordered box with the text: "HITs - Human Intelligence Tasks - are individual tasks that you work on. Find HITs now." Below this, it says "As a Mechanical Turk Worker you:" followed by a list: "• Can work from home", "• Choose your own work hours", "• Get paid for doing good work". A flow diagram shows "Find an interesting task" (with a circular icon containing text about HITs), "Work" (with a gear icon), and "Earn money" (with a dollar sign icon). A "Find HITs Now" button is at the bottom. The right column is titled "Get Results from Mechanical Turk Workers" and says "Ask workers to complete HITs - Human Intelligence Tasks - and get results using Mechanical Turk. Register Now". Below, it says "As a Mechanical Turk Requester you:" followed by a list: "• 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". A flow diagram shows "Fund your account" (with a plus sign icon), "Load your tasks" (with a task icon), and "Get results" (with a star icon). A "Get Started" button is at the bottom.

Amazon Mechanical Turk

MTurk 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 website interface. At the top, there is a navigation bar with the Amazon Mechanical Turk logo and the text 'Artificial Intelligence'. To the right, there are buttons for 'Your Account', 'HITS', and 'Qualifications'. Further right, it says 'Already have an account? Sign in as a Worker | Requester'. Below the navigation bar, there is a main heading '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. 179,373 HITS available. View them now.'

The main content area is divided into two columns. The left column is titled 'Make Money by working on HITS'. It includes the text 'HITS - Human Intelligence Tasks - are individual tasks that you work on. Find HITS now.' and 'As a Mechanical Turk Worker you:' followed by a list of benefits: 'Can work from home', 'Choose your own work hours', and 'Get paid for doing good work'. Below this is a flow diagram: 'Find an interesting task' (with a circular icon containing text about finding tasks) leads to 'Work' (with a gear icon), which leads to 'Earn money' (with a dollar sign icon). A 'Find HITS Now' button is at the bottom of this section.

The right column is titled 'Get Results from Mechanical Turk Workers'. It includes the text 'Ask workers to complete HITS - Human Intelligence Tasks - and get results using Mechanical Turk. Register Now' and 'As a Mechanical Turk Requester you:' followed by a list of benefits: 'Have access to a global, on-demand, 24 x 7 workforce', 'Get thousands of HITS completed in minutes', and 'Pay only when you're satisfied with the results'. Below this is a flow diagram: 'Fund your account' (with a plus sign icon) leads to 'Load your tasks' (with a task icon), which leads to 'Get results' (with a star icon). A 'Get Started' button is at the bottom of this section.

Amazon Mechanical Turk

MTurk 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 website interface. At the top, there is a navigation bar with the Amazon Mechanical Turk logo and the text 'Artificial Intelligence'. To the right, there are buttons for 'Your Account', 'HITS', and 'Qualifications', and a sign-in prompt for workers. Below the navigation bar, a central banner states 'Mechanical Turk is a marketplace for work.' and provides information about the number of available HITs. The main content area is divided into two columns: 'Make Money by working on HITs' and 'Get Results from Mechanical Turk Workers'. The 'Make Money' section includes a list of benefits for workers and a three-step diagram: 'Find an interesting task', 'Work', and 'Earn money'. The 'Get Results' section includes a list of benefits for requesters and a three-step diagram: 'Fund your account', 'Load your tasks', and 'Get results'.

amazonmechanicalturk
Artificial Intelligence

Your Account | HITS | Qualifications

Introduction | Dashboard | Status | Account Settings

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.
179,373 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

Get Results

from Mechanical Turk Workers

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

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

Caractéristiques d'AMT

Rémunération :

- ▶ à la tâche (*illégal* en France sauf (rares) exceptions) : moins de 2 \$/h
- ▶ pas de relation explicite entre les *Turkers* et les *Requesters*

Tâches :

- ▶ nouveaux usages : par exemple, des créations artistiques, comme <http://www.thesheepmarket.com/>
- ▶ des tâches traditionnellement réalisées par des employés salariés : transcription, traduction (agences LDC, ELDA), etc

AMT : le rêve devenu réalité ?

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 ?

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* !

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 pachto 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].

AMT permet de produire des ressources de qualité ?

- ▶ permet de produire des ressources de qualité dans certains cas précis (par exemple, la transcription simple)
- ▶ mais :
 - ▶ la qualité est insuffisante lorsque la tâche est **complexe** (par exemple, le résumé [Gillick and Liu, 2010])
 - ▶ l'**interface** d'AMT pose parfois problème [Tratz and Hovy, 2010]
 - ▶ les *Turkers* posent parfois problème (tricheurs, **spammers**)
 - ▶ le modèle de rémunération **à la tâche** pose problème [Kochhar et al., 2010]
- ▶ pour certaines tâches simples les outils de TAL produisent de **meilleurs résultats** qu'AMT [Wais et al., 2010].

AMT : un passe-temps pour les Turkers ?

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

- ▶ 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 \$**.

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

Éthique :

- ▶ 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*

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

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.

179,373 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



Get Results from Mechanical Turk Workers

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

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



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

Légalité :

- ▶ accord de licence d'Amazon : les *Turkers* sont considérés comme des travailleurs indépendants ⇒ ils sont supposés se déclarer comme tels et payer les cotisations afférentes
- ▶ illusoire, vu le niveau de rémunération
- ⇒ les États **perdent** une source de revenus légitime

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

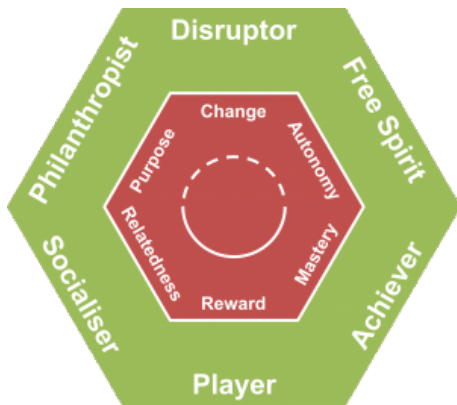
Types de joueurs selon [Bartle, 1996]

- ▶ *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.](#)
-  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.](#)
-  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.



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.



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., Guillaume, B., and Stern, V. (2014).

Zombilingo : eating heads to perform dependency syntax annotation (zombilingo : manger des têtes pour annoter en syntaxe de dépendances) [in french].

In Proceedings of TALN 2014 (Volume 3 : System Demonstrations), pages 15–16, Marseille, France. Association pour le Traitement Automatique des Langues.



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.



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.



Habert, B. (2000).

Corpus. Méthodologie et applications linguistiques, chapitre
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.



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.



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 SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2010, Washington D.C.](#)



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.



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.



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.



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-quality workforce with mechanical turk. In Proceedings of Computational Social Science and the Wisdom of Crowds (NIPS).