

Pépites (algorithmiques) pour le Traitement Automatique des Langues (TAL)

Karën Fort

karen.fort@loria.fr

28 mai 2013



1 Introduction

- Le TAL en 2 minutes
- Le TAL et ses applications
- Le TAL : définitions

2 Ambiguïté à tous les étages

3 Les solutions

4 Au cœur de systèmes : les ressources langagières

5 Conclusion

Что это?

Reoiajr oj earoij reoa o eo ao aeoi oj aeoij aoeir eoaj.

Что это?

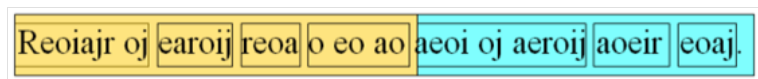
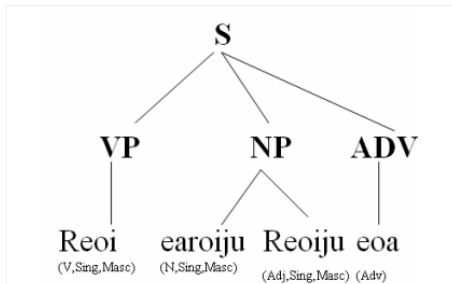
Reoiajr oj earoij reoa o eo ao aeoi oj aroij aoeir eoaj.

Что это?

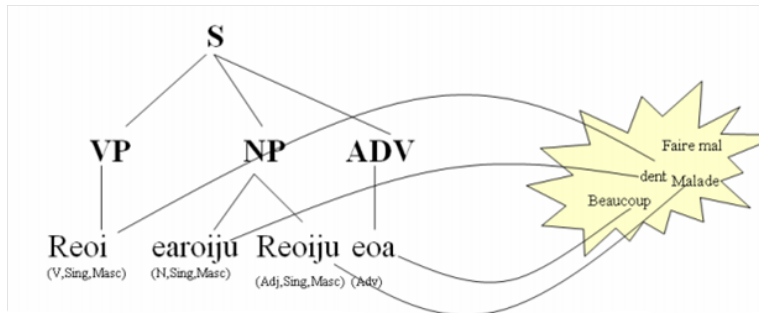
Reoi earoiju Reoiju eoa
 (V,Sing,Masc) (N,Sing,Masc) (Adj,Sing,Masc) (Adv)

Reoiajr oj earoij reoa o eo ao aeoi oj aeoij aoeir eoaj.

Что это?



Что это?



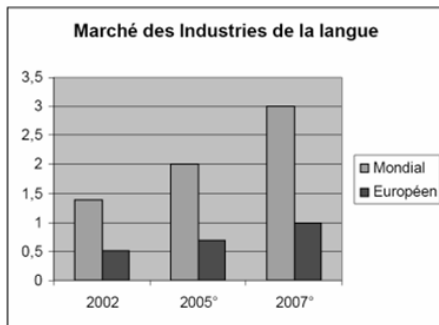
Pour quoi faire ?

- Traduction automatique
- Recherche d'informations
- Indexation
- Traitement d'Emails
- Veille
- ...

Applications



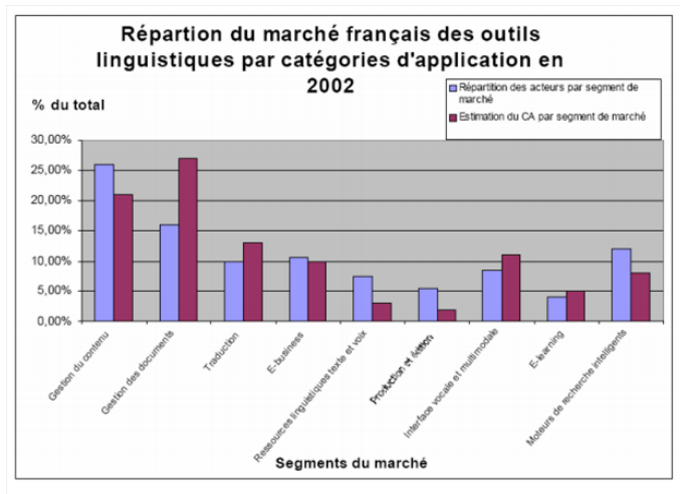
Le marché (Milliards d'€, Source Technolangue)



Source: Bureau Van Dijk



Marché par application



Définitions

Pour les **chercheurs** :

- modéliser une activité langagière pour comprendre comment fonctionne la langue

Pour les **industriels** :

- reproduire l'activité langagière humaine pour optimiser les performances des systèmes

- 1 Introduction
- 2 **Ambiguïté à tous les étages**
 - Exemples
 - Les grandes étapes
- 3 Les solutions
- 4 Au cœur de systèmes : les ressources langagières
- 5 Conclusion

Le TAL : pas si simple. . .

- *La mousse aux fraises est sur la table de l'avocat.*

Le TAL : pas si simple. . .

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer !*

Le TAL : pas si simple. . .

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer!*
- *Le bus a renversé un passant. . .*

Le TAL : pas si simple. . .

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer!*
- *Le bus a renversé un passant. . .*
 - ▶ *. . .je l'ai entendu freiner.*

Le TAL : pas si simple. . .

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer!*
- *Le bus a renversé un passant. . .*
 - ▶ *. . .je l'ai entendu freiner.*
 - ▶ *. . .je l'ai entendu crier.*

Le TAL : pas si simple. . .

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer!*
- *Le bus a renversé un passant. . .*
 - ▶ *. . .je l'ai entendu freiner.*
 - ▶ *. . .je l'ai entendu crier.*
- *Le professeur a envoyé l'élève chez le proviseur. . .*

Le TAL : pas si simple. . .

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer!*
- *Le bus a renversé un passant. . .*
 - ▶ *. . .je l'ai entendu freiner.*
 - ▶ *. . .je l'ai entendu crier.*
- *Le professeur a envoyé l'élève chez le proviseur. . .*
 - ▶ *. . .il faisait trop de bruit.*

Le TAL : pas si simple. . .

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer !*
- *Le bus a renversé un passant. . .*
 - ▶ *. . . je l'ai entendu freiner.*
 - ▶ *. . . je l'ai entendu crier.*
- *Le professeur a envoyé l'élève chez le proviseur. . .*
 - ▶ *. . . il faisait trop de bruit.*
 - ▶ *. . . il était excédé.*

Le TAL : pas si simple. . .

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer!*
- *Le bus a renversé un passant. . .*
 - ▶ *. . . je l'ai entendu freiner.*
 - ▶ *. . . je l'ai entendu crier.*
- *Le professeur a envoyé l'élève chez le proviseur. . .*
 - ▶ *. . . il faisait trop de bruit.*
 - ▶ *. . . il était excédé.*
 - ▶ *. . . il l'avait convoqué.*

Le TAL : pas si simple. . .

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer !*
- *Le bus a renversé un passant. . .*
 - ▶ *. . .je l'ai entendu freiner.*
 - ▶ *. . .je l'ai entendu crier.*
- *Le professeur a envoyé l'élève chez le proviseur. . .*
 - ▶ *. . .il faisait trop de bruit.*
 - ▶ *. . .il était excédé.*
 - ▶ *. . .il l'avait convoqué.*

→ **ambiguïtés** pour les systèmes et/ou pour les humains

Le découpage en " mots" ou *tokenization*

- *l'arbre*

Le découpage en " mots" ou *tokenization*

- *l'arbre*
- *aujourd'hui*

L'analyse morphologique

la *porte*

- *porte* +Nf + Sg

L'analyse morphologique

la *porte*

- *porte* +Nf + Sg
- *porte* +VT + 1/3P + Sg

L'analyse syntaxique

Jean regarde un homme sur la colline avec un télescope.

- **Qui** est sur la colline ?
- **Qui** a un télescope ?

L'analyse sémantique

Tous les hommes aiment une femme.

→ **Chaque** homme aime une femme ou **tous** les hommes aiment la même femme ?

1 Introduction

2 Ambiguïté à tous les étages

3 Les solutions

- Techniques utilisées
- Identification de la langue
- Segmentations
- Analyse morpho-syntaxique
- Désambiguïsation
- Analyse syntaxique : par l'exemple

4 Au cœur de systèmes : les ressources langagières

5 Conclusion

Des techniques variées

- systèmes à base de règles
 - ▶ définies par l'humain (linguistes)
 - ▶ entrées manuellement
- systèmes basés sur les données
 - ▶ apprentissage supervisé ou non supervisé
 - ▶ à partir d'exemples (rédigés et/ou annotés par des humains)
 - ▶ algorithmes (pensés par des humains)

→ et tous les intermédiaires/mélanges possibles

⇒ présentation **partielle/partiale**

1 Introduction

2 Ambiguïté à tous les étages

3 Les solutions

- Techniques utilisées
- **Identification de la langue**
- Segmentations
- Analyse morpho-syntaxique
- Désambiguïsation
- Analyse syntaxique : par l'exemple

4 Au cœur de systèmes : les ressources langagières

5 Conclusion

Exercice : identifier la langue d'un texte

3 groupes

Trouver au moins **2** algorithmes permettant d'identifier la langue d'un texte

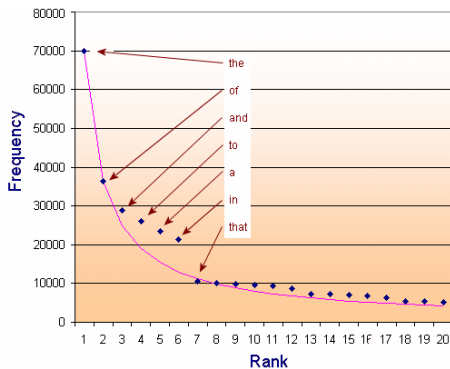
Conseil : connaître les langues et leurs caractéristiques

Loi de Zipf :

- étant donné un corpus d'énoncés en langue naturelle, la fréquence d'un mot est **inversement proportionnelle** à son rang dans la table de fréquence.
- ex. : "the" représente près de 7 % du *Brown Corpus* alors que près de la moitié du vocabulaire total du corpus sont des **hapax**.
- Seuls 135 éléments de vocabulaire sont nécessaires pour couvrir la moitié du *Brown Corpus*

Rang	Mot	Fréquence
1	<i>the</i>	69 970
2	<i>of</i>	36 410
3	<i>and</i>	28 854
20	<i>I</i>	5 180

Loi de Zipf sur le *Brown corpus*



Identifier la langue : solution 1

Méthode des *shortwords* :

- BD de "mots outils" (mots grammaticaux, "petits" mots)
- Compter les occurrences de ces mots outils dans le texte
- Comparer avec les BD

Identifier la langue : solution 2

Méthode des *trigrammes* :

- Rechercher la probabilité qu'un caractère C_i apparaisse après les deux précédents dans la langue l :

$$P(C_i | C_{i-1} : C_{i-2}, l)$$

- Calculer la probabilité résultante pour chaque langue, pour l'ensemble du texte :

$$\prod_{i=1}^{i=n} P(C_i | C_{i-1} : C_{i-2}, l)$$

XRCE (Xerox) language identifier

`http://open.xerox.com/Services/LanguageIdentifier`

XRCE (Xerox) language identifier

`http://open.xerox.com/Services/LanguageIdentifier`

→ *trishort*

Limitations

- Longueur du texte (5 mots mini.)
- textes multilingues :

Natural Language Processing at hand

Many software applications use, create or transform textual data, be them word processors, online reservation applications, electronic messaging, document processing, on-line or off-line watch....

With GramLab, these applications can be enhanced with NLP functions: spell checking, automatic recognition of dates or places, automatic update of email contact, enhance full text search...

Aproged : Nouvelle avancée
dans l'analyse de contenus
et la valorisation de
l'information...

L'Aproged et le Consortium

- et aussi :
 - ▶ Barack Obama → italien
 - ▶ Nicolas Sarkozy → polonais
 - ▶ Barack Obama and Nicolas Sarkozy → anglais
 - ▶ camping caravaning, trekking

Identifier la langue : solutions 3 et 4

- Essayer d'identifier l'encodage
- Regarder les méta-données

1 Introduction

2 Ambiguïté à tous les étages

3 Les solutions

- Techniques utilisées
- Identification de la langue
- **Segmentations**
- Analyse morpho-syntaxique
- Désambiguïsation
- Analyse syntaxique : par l'exemple

4 Au cœur de systèmes : les ressources langagières

5 Conclusion

Segmentations : exercice

*Marchepied à plate forme sécurisée 2.41m hauteur travail
max. 3.16m 5 marches et garde corps fixe PROFORT TUBESCA*

Segmentations : approche simp(ist)e

① Segmentation en phrases (texte brut) :

- ▶ Se termine par une ponctuation de fin de phrase (. ; ? !)
- ▶ La phrase suivante commence par un espace/retour à la ligne suivi d'une majuscule
- ▶ Les ... forment une ponctuation unique

② Segmentation en mots (tokenization) :

- ▶ Séparateurs de mots : espace, CR, tab, apostrophes (avec le mot), ponctuations (considérés comme des mots), les . ne sont pas des séparateurs
- ▶ Les tirets sont des séparateurs s'ils sont suivis d'un pronom (vient-il), et la séquence -t doit être effacée dans certains cas (*envoie-t-elle*, césure de fin de phrase)
- ▶ Appliquer des transformations locales : Au, du, cet, qu', l', m'....

Segmentation en phrases : plus loin

- en allemand ?
- en thaï ?

Tokenization : plus loin

*Je veux **bien que** tu viennes*

→ Conserver l'ambiguïté

Tokenization : encore plus loin

- en allemand ?
- en chinois ?

Segmentations : les questions

Définition théorique du segment :

- Qu'est-ce qu'un **mot** ?
 - ▶ Plate forme, marche-pied marche pied marchepied
 - ▶ Les marchands du temple : du → de le ?
- Qu'est-ce qu'une **phrase** ?
 - ▶ Ah ?
 - ▶ - ?

Ambiguïté des séparateurs :

- Microsoft.com, 23.5, . . . , C.G.T.
- Aujourd'hui, 9'8, jusqu'à
- C&A, R&D
- Sépara-
- teur

1 Introduction

2 Ambiguïté à tous les étages

3 Les solutions

- Techniques utilisées
- Identification de la langue
- Segmentations
- **Analyse morpho-syntaxique**
- Désambiguïsation
- Analyse syntaxique : par l'exemple

4 Au cœur de systèmes : les ressources langagières

5 Conclusion

Morpho-syntaxe : exercice

*Orange, au cours du matin, a permis la prise de bénéfices.
Nous avons fait des paris audacieux sur cette valeur qui est
montée dès l'ouverture du CO*

Article, Nom, Adverbe, Verbe, Adjectif, Pronom, Nom propre, Ppassé,
Préposition, Interjection

Morpho-syntaxe : étiquettes

- Orange [Nom propre, Adjectif, Nom]
- au [Préposition, Article]
- cours [Nom]
- au cours du [Préposition]
- du [Ppassé, Nom]
- matin [interjection, Adverbe, Nom]
- a [Préposition, Nom]
- permis [Ppassé, Nom, Adjectif]
- la [Article, Pronom, Nom]
- prise [Nom, Ppassé]
- de [Préposition]
- bénéfices [Nom]

Morpho-syntaxe : lemmatisation

- Orange → orange — Orange
- au → à + le
- cours → cours
- du → de + le
- matin → matin
- a → avoir
- permis → permis — permettre
- la → le
- prise → prise — prendre
- de → de
- bénéfices → bénéfice

Morpho-syntaxe : solutions en extension

```
<lexicalEntry id="championne_1">
  <feminineVariantOf target="champion_1">champion</feminineVariantOf>
  <formSet>
    <lemmatizedForm>
      <orthography>championne</orthography>
      <grammaticalCategory>commonNoun</grammaticalCategory>
      <grammaticalGender>feminine</grammaticalGender>
    </lemmatizedForm>
    <inflectedForm>
      <orthography>championne</orthography>
      <grammaticalNumber>singular</grammaticalNumber>
    </inflectedForm>
    <inflectedForm>
      <orthography>championnes</orthography>
      <grammaticalNumber>plural</grammaticalNumber>
    </inflectedForm>
  </formSet>
  <originatingEntry target="TLF">CHAMPION, ONNE, subst.</originatingEntry>
</lexicalEntry>
```

Morpho-syntaxe : solutions (?) en extension

Stockage ?

Morphalou 2 = **160 Mo**

Morpho-syntaxe : solutions (?) en extension

Temps d'accès ?

Morpho-syntaxe : solutions (?) en extension

Exemples en finnois :

- *Järjestelmällinen* : organisé
- *järjestelmällistetty* : conçu pour être organisé
- *epäjärjestelmällistetty* : désorganisé ou pas conçu pour être organisé
- *Epäjärjestelmällistämättömyy* : négation et substantivation
- *Epäjärjestelmällistytämättömyydellänsä* : l'objet dénoté par le substantif appartient à quelqu'un (génitif)
- *Epäjärjestelmällistytämättömyydelläsäkään* : nouvelle négation
- *epäjärjestelmällistytämättömyydelläsäkäänkö* : ajout d'un suffixe dénotant l'interrogation
- *epäjärjestelmällistytämättömyydelläsäkäänköhän* : ajout d'une emphase (c'est moi qui ...)
- *epäjärjestelmällistytämättömyydelläänsäkäänköhän* :

The reverse of the reverse of something abstract that is made to be unorganised, which is owned by someone, and is one of the two or more (possibly similar) attributes that have a negative atmosphere or lack of something, and we doubt if it is it at the same time that we ensure that it truly is.

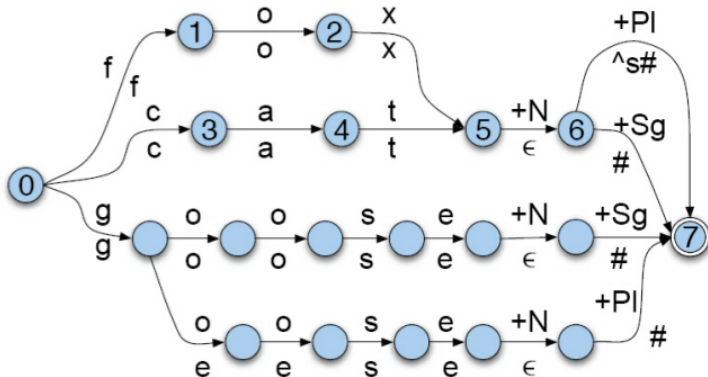
Morpho-syntaxe : solutions en intention

Quelle solution pour :

- optimiser le stockage
- permettre un accès rapide
- dans les deux sens :
 - ▶ leave+NNS → leaves
 - ▶ leaves → leave+NNS

Structure de données **et** accès ?

Morpho-syntaxe : solutions en intention



Morpho-syntaxe : transducteurs à états finis (FST)

XFST (Xerox FST) :

```
[ [l e a v e %+VBZ .x. l e a v e s] |  
[l e a v e %+VB .x. l e a v e] |  
[l e a v e %+VBG .x. l e a v i n g] |  
[l e a v e %+VBD .x. l e f t] |  
[l e a v e %+NN .x. l e a v e] |  
[l e a v e %+NNS .x. l e a v e s] |  
[l e a f %+NNS .x. l e a v e s ] |  
[l e f t %+JJ .x. l e f t] ]
```

APPLY DOWN> leave+VBD

left

APPLY UP> leaves

leave+NNS

leave+VBZ

leaf+NNS

Morpho-syntaxe : transducteurs à états finis (FST)

Lexique du français en FST :

- taille une fois compilé : **800 Ko** (français), 500 Ko (anglais)
- lookup immédiat
- gestion des langues agglutinantes, des cycles, ...

→ importance des **ressources langagières**

1 Introduction

2 Ambiguïté à tous les étages

3 Les solutions

- Techniques utilisées
- Identification de la langue
- Segmentations
- Analyse morpho-syntaxique
- **Désambiguïsation**
- Analyse syntaxique : par l'exemple

4 Au cœur de systèmes : les ressources langagières

5 Conclusion

Désambiguïsation : le problème

permis $\rightarrow$ permettre+V — permis+N :

- un mot = plusieurs analyses
- mots inconnus

$\rightarrow$ Comment faire ? Qu'utilise l'humain ?

Désambiguïsation : le problème

permis → permettre+V — permis+N :

- un mot = plusieurs analyses
- mots inconnus

→ Comment faire ? Qu'utilise l'humain ?

Il m'a permis d'habiter chez lui.

Désambiguïsation

Solutions ?

Désambiguïsation : modèles de Markov

Chaîne de Markov [1913] :

- une variable ne dépend que du présent
- la probabilité d'une variable ne dépend que de celle la précédant
 - ▶ pas de mémoire
 - ▶ homogène (ne dépend pas de la place dans la chaîne)

Désambiguïsation : HMM

- Etant donnée une suite de symboles :
 - ▶ Orange [Nom propre, Adjectif, Nom], au cours du [Préposition], matin [interjection, Adverbe, Nom], a [Préposition, Nom], permis [Ppassé, Nom, Adjectif], la [Article, Pronom, Nom], prise [Nom, Ppassé], de [Préposition], bénéfices [Nom]
- On considère la probabilité qu'un symbole se trouve après deux autres symboles
- On prend la plus grande

1 Introduction

2 Ambiguïté à tous les étages

3 Les solutions

- Techniques utilisées
- Identification de la langue
- Segmentations
- Analyse morpho-syntaxique
- Désambiguïsation
- Analyse syntaxique : par l'exemple

4 Au cœur de systèmes : les ressources langagières

5 Conclusion

Démo : analyseurs syntaxiques

LEOPAR (pour le français) :

`http://wikilligramme.loria.fr/doku.php?id=leopar:demo`

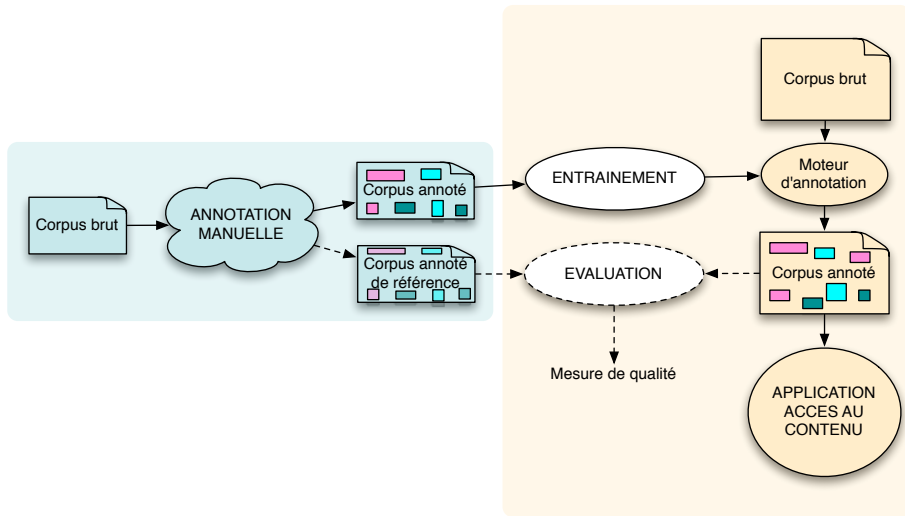
Stanford parser (pour l'anglais) :

`http://nlp.stanford.edu:8080/parser/`

→ observations ?

- 1 Introduction
- 2 Ambiguïté à tous les étages
- 3 Les solutions
- 4 **Au cœur de systèmes : les ressources langagières**
 - Coût et complexité
 - Solutions : pré-annotation
 - Solutions : crowdsourcing
- 5 Conclusion

Annotation manuelle de corpus et TAL



Exemples d'annotation

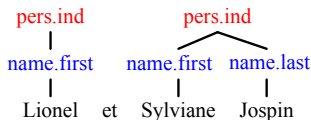
Parties du discours (*Penn Treebank*, [Marcus et al., 1993]), pré-annotées :

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

Renommage de gènes [LREC 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_(recU) and ypbC_(recS) 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.

Entités nommées structurées [LAW 2011] :



Un coût élevé, voire prohibitif

Penn Treebank [Marcus et al., 1993] :

- 4,8 millions de tokens annotés en morpho-syntaxe $\Rightarrow$ phase d'apprentissage d'1 mois, pour atteindre 3 000 mots/h
- 3 millions de tokens annotés en syntaxe $\Rightarrow$ phase d'apprentissage d'2 mois, pour atteindre 475 mots/h

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

- 1,8 millions de tokens annotés en morpho-syntaxe et syntaxe
 $\Rightarrow$ 5 ans, 22 personnes (max. 17 en parallèle), 600 000 dollars

GENIA [Kim et al., 2008] :

- 9 372 phrases annotés en microbiologie (noms de gènes, protéines)
 $\Rightarrow$ 5 annotateurs à temps partiel, 1 coordinateur senior et 1 junior pendant 1,5 an

Solutions ?

Gagner en temps et en qualité ?

Active Learning

Tag Dictionaries

Pre-annotation

Annotation Tools

Crowdsourcing

Methodology

Training

Gagner en temps et en qualité ?

Active Learning

Tag Dictionaries

Pre-annotation

Annotation Tools

Crowdsourcing

Methodology

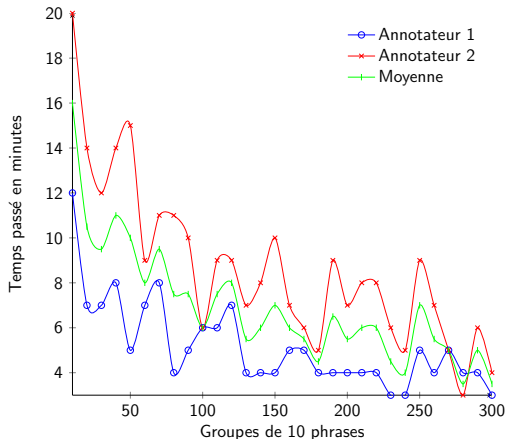
Training

Annotation morpho-syntaxique du *Penn Treebank*

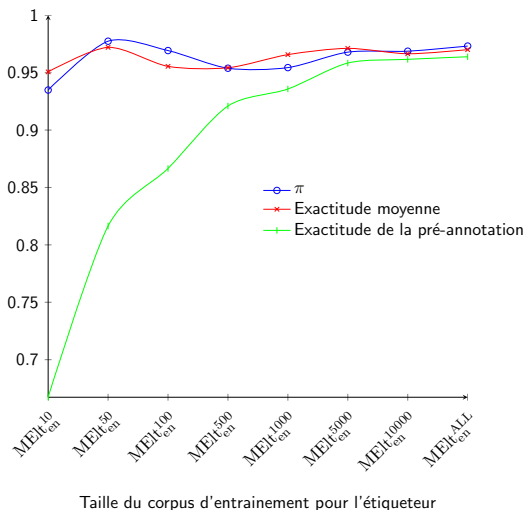
I/PRP do/VBP n't/RB feel/VB very/RB ferocious/JJ ./.

→ tester l'influence de la pré-annotation [Fort and Sagot, 2010]

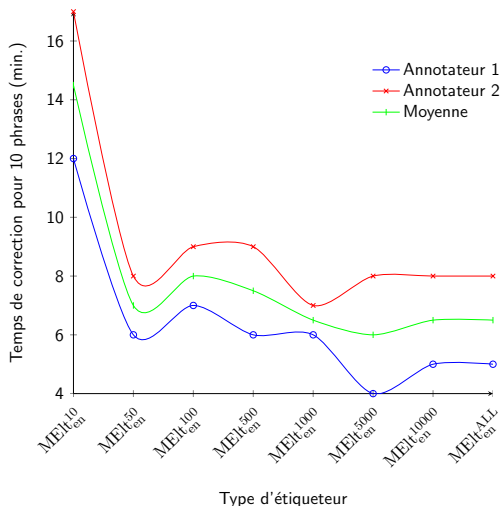
Courbe d'apprentissage : annotation morpho-syntaxique du *Penn Treebank*



Exactitude de l'annotation et π en fonction du type de pré-annotation



Temps d'annotation en fonction de la qualité de la pré-annotation



Gagner en temps et en qualité ?

Active Learning

Tag Dictionaries

Pre-annotation

Annotation Tools

Crowdsourcing

Methodology

Training

Gagner en temps et en qualité ?

Active Learning

Tag Dictionaries

Pre-annotation

Annotation Tools

Crowdsourcing

Methodology

Training

Crowdsourcing : exemple de GWAP (jeu ayant un but)



USERPROFILE

kaa

0 this week
0 decisions
0 agreements
0 extras

0 this month
31 all time

Level: Trainee

Your rating: **80%**

CASE OPEN
50 tasks remaining

1 completed case

[EDIT PROFILE](#) | [LOGOUT](#)

 Instructions

 FAQ

 Restart

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?



The Adventures of Sherlock Holmes - The Man with the Twisted Lip (Sir Arthur Conan Doyle)

"Certainly."

"Then dress. No one is stirring yet, but I know where the stable-boy sleeps, and we shall soon have the trap out." He chuckled to himself as he spoke, his eyes twinkled, and he seemed a different man to the sombre thinker of the previous night.

As I dressed I glanced at my watch. It was no wonder that no one was stirring. It was twenty-five minutes past four. I had hardly finished when Holmes returned with the news that the boy was putting in the horse.

"I want to test a little theory of mine," said he, pulling on his boots. "I think, Watson, that you are now standing in the presence of one of the most absolute fools in Europe. I deserve to be kicked from here to Charing Cross. But I think I have the key of the affair now."

"And where is it?" I asked, smiling.

"In the bathroom," he answered. "Oh, yes, I am not joking," he continued, seeing my look of incredulity. "I have just been there, and I have taken it out, and I have got it in this Gladstone bag. Come on, my boy, and we shall see whether it will not fit the lock."

The phrase in blue is the closest phrase that refers to the phrase in orange.

 Disagree

 Agree

PHRASE DETECTIVES

SEARCHCLUES

Words like they, him, her and it are likely to refer to something else in the text. Try to find the closest mention of this phrase.

Words like 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.

 Feedback

Crowdsourcing : questions

[Fort et al., 2011, Chamberlain et al., 2012] :

- qualité ?
- triche ?
- droits d'auteur ?
- éthique ?

Origines du cours

- Cours d'Alain Couillault, Gestion sémantique des contenus, Master ICONE, La Rochelle.
- http://faculty.washington.edu/ebender/2012_472/0404.pdf
- <http://www.cis.upenn.edu/~cis639/docs/fsexamples.html>



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., Fort, K., Kruschwitz, U., Lafourcade, M., and
Poesio, M. (2012).
Theory and Applications of Natural Language Processing, chapter
Using Games to Create Language Resources : Successes and
Limitations of the Approach.
Theory and Applications of Natural Language Processing. Springer.
À paraître.



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. (2012).

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. and Sagot, B. (2010).

Influence of pre-annotation on POS-tagged corpus development.

In Proceedings of the Fourth ACL Linguistic Annotation Workshop, pages 56–63, Uppsala, Suède.



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.



Kim, J.-D., Ohta, T., and Tsujii, J. (2008).

Corpus annotation for mining biomedical events from literature.
BMC Bioinformatics, 9(1) :10.



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.