

Annotation collaborative de corpus : Motivations et définitions

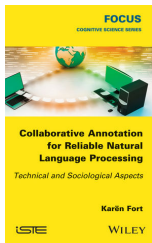
Karën Fort

karen.fort@sorbonne-universite.fr / <https://members.loria.fr/KFort/>

D'où je parle

Voir <https://members.loria.fr/KFort/>

► Création de ressources langagières pour le TAL



► Éthique et TAL



Le TAL aujourd'hui

Le TAL : des applications dans nos vies

Des méthodes par apprentissage omniprésentes



Annotation manuelle et TAL


Parenthèse : présentations

Qu'est-ce qu'annoter ?


Pour finir

«Traduction» automatique

[Sign in](#)




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

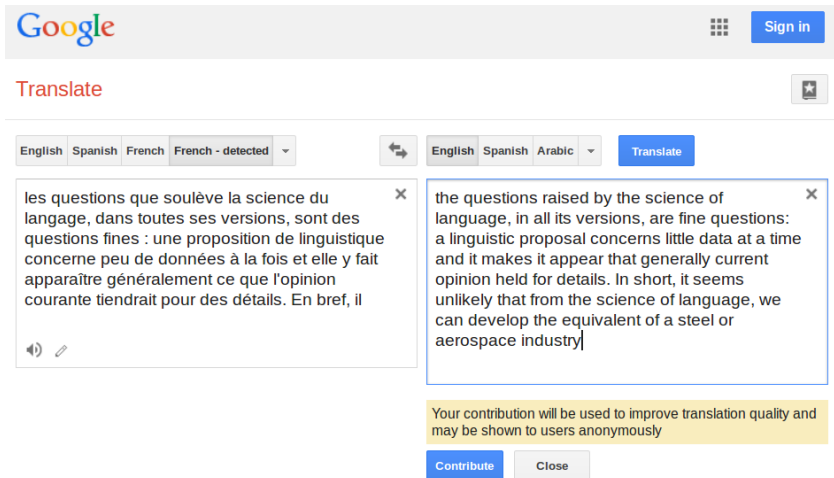


the questions raised by the science of language, in all its versions, are fine questions: a linguistic proposal concerns little data at a time and it makes it appear that generally current opinion held for details. In short, it seems unlikely that from the science of language, we can develop the equivalent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

 Wrong?

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

«Traduction» automatique



The screenshot shows the Google Translate web interface. At the top is the Google logo and a 'Sign in' button. Below is the 'Translate' heading. The language selection bar shows 'English', 'Spanish', 'French', and 'French - detected' (selected). The source text in French is: '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 translated text in English is: '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'. A yellow banner at the bottom states: 'Your contribution will be used to improve translation quality and may be shown to users anonymously'. Below the banner are 'Contribute' and 'Close' buttons.

Google

Sign in

Translate

English Spanish French French - detected

English Spanish Arabic

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

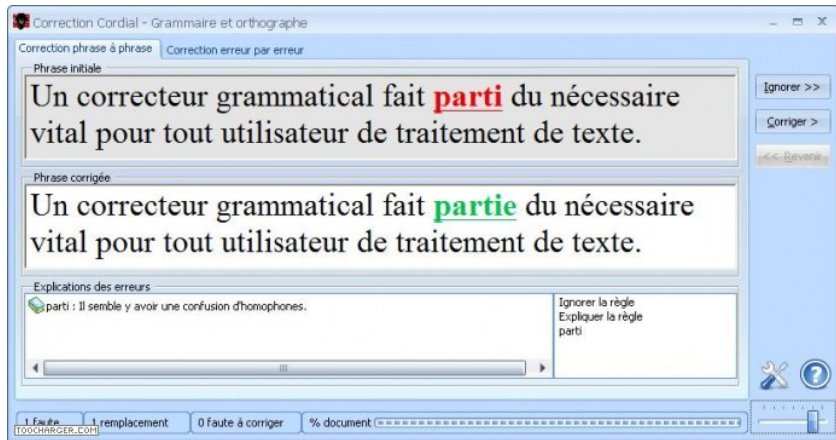
the questions raised by the science of language, in all its versions, are fine questions: a linguistic proposal concerns little data at a time and it makes it appear that generally current opinion held for details. In short, it seems unlikely that from the science of language, we can develop the equivalent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

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

Contribute Close

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

Correction orthographique et grammaticale



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

Extraction d'entités nommées

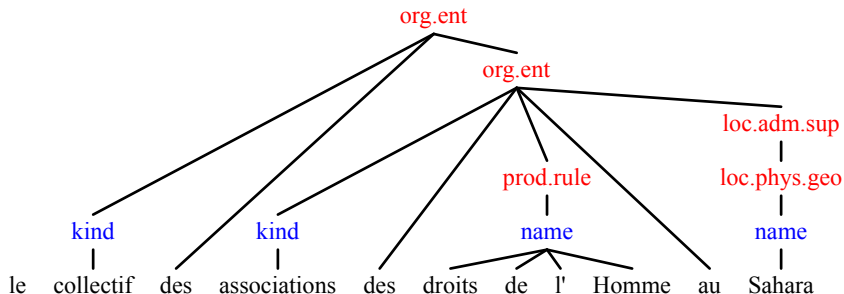
<ENAMEX TYPE="PERSON">Henri</ENAMEX> a acheté **<NUMEX TYPE="QUANTITY">300</NUMEX>** actions de la société **<ENAMEX TYPE="ORGANIZATION">AMD</ENAMEX>** en **<TIMEX TYPE="DATE">2006</TIMEX>**.

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]

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 (~ 99 % aujourd'hui)
 - ▶ apprentissage (en général) supervisé
 - ▶ à partir d'exemples (rédigés et/ou annotés par des humains)
 - ▶ algorithmes statistiques ou neuronaux (pensés par des humains)

→ **corpus annotés** (et lexiques)



[Annotation en syntaxe de dépendances]

Le TAL aujourd'hui

Annotation manuelle et TAL

- L'annotation manuelle en TAL

- Un coût prohibitif

- De la pérennité des ressources langagières

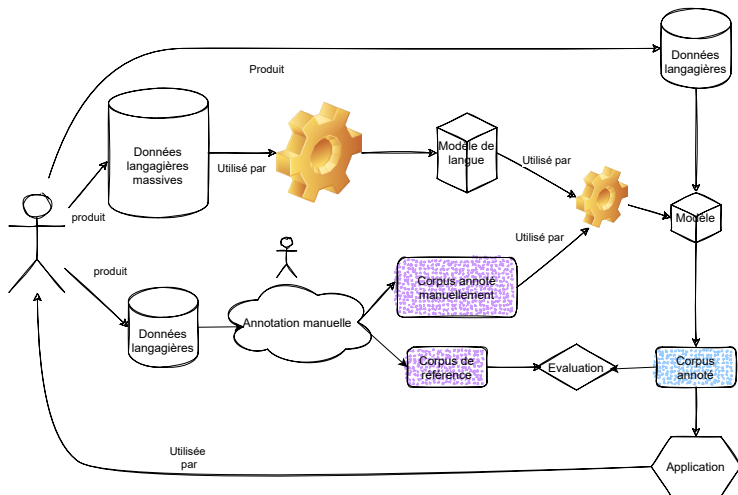
- Un sujet de recherche à part entière

Parenthèse : présentations

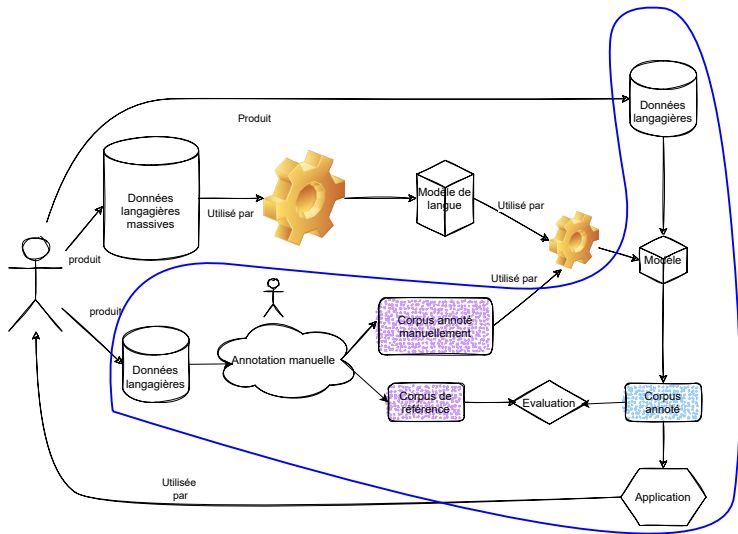
Qu'est-ce qu'annoter ?

Pour finir

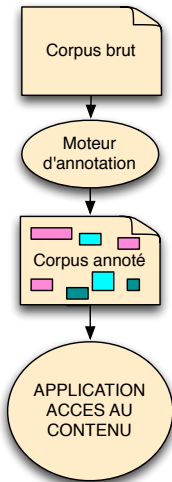
Le TAL aujourd'hui



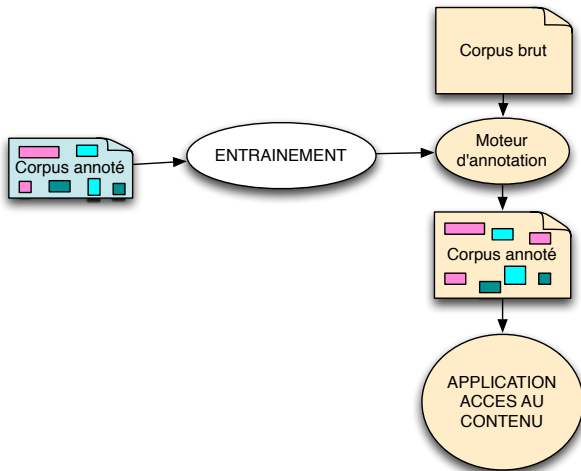
Le TAL aujourd'hui



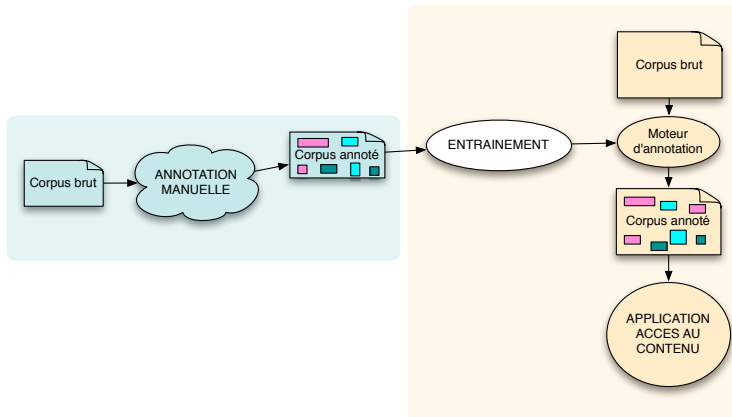
L'annotation manuelle dans|pour le TAL



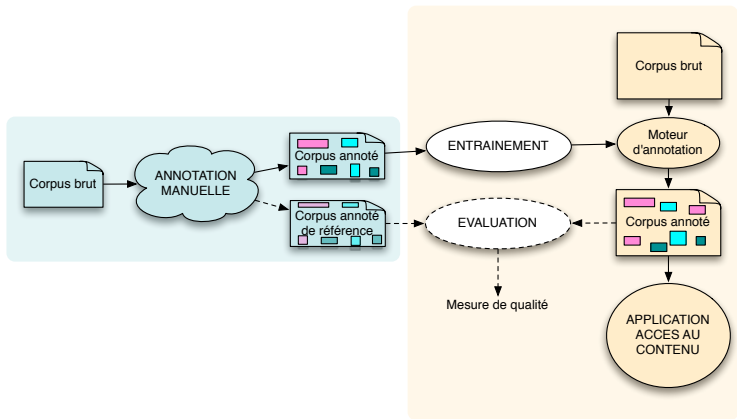
L'annotation manuelle dans|pour le TAL



L'annotation manuelle dans|pour le TAL



L'annotation manuelle dans|pour le TAL



Pourquoi c'est important !



Ben Hamner  @benhamner · Oct 9



Programming: 10% writing code. 90% figuring out why it doesn't work

Analyzing data and ML: 1% writing code. 9% figuring out why code doesn't work. 90% figuring out what's wrong with the data



89



1.9K



8.7K



Des ressources diverses, de complexité croissante

Diversité accrue :

- ▶ des **médias** : texte, parole, musique, vidéo
- ▶ des **champs d'applications** : linguistique, domaines spécialisés

Complexité accrue :

- ▶ 1960-1990 : morpho-syntaxe
- ▶ 1990-2000 : morpho-syntaxe et syntaxe (corpus arborés), entités nommées simples, « sens » (WordNet, FrameNet)
- ▶ Depuis 2000 : annotations sémantiques variées (opinions, émotions, etc.), discours, entités nommées structurées, etc.

Taille accrue :

- ▶ 1961 : Brown Corpus, 1 million de mots
- ▶ 1995 : British National Corpus (BNC), 100 millions de mots
- ▶ Depuis 2014 : Universal Dependencies (2.10 : 130 langues, 228 corpus)

Penn Treebank (PTB) [Marcus et al., 1993]

corpus arboré de l'Université de Pennsylvanie

PTB 1 :

- ▶ correction de l'étiquettage morpho-syntaxique : ? mots à l'heure, ? heures par jour
- ▶ correction de l'étiquettage syntaxique : ? mots à l'heure, ? heures par jour

Penn Treebank (PTB) [Marcus et al., 1993]

corpus arboré de l'Université de Pennsylvanie

PTB 1 :

- ▶ correction de l'étiquetage morpho-syntaxique : 3 000 mots à l'heure, ? heures par jour
- ▶ correction de l'étiquetage syntaxique : ? mots à l'heure, ? heures par jour

Penn Treebank (PTB) [Marcus et al., 1993]

corpus arboré de l'Université de Pennsylvanie

PTB 1 :

- ▶ correction de l'étiquetage morpho-syntaxique : 3 000 mots à l'heure, 3 heures par jour
- ▶ correction de l'étiquetage syntaxique : ? mots à l'heure, ? heures par jour

Penn Treebank (PTB) [Marcus et al., 1993]

corpus arboré de l'Université de Pennsylvanie

PTB 1 :

- ▶ correction de l'étiquetage morpho-syntaxique : 3 000 mots à l'heure, 3 heures par jour
- ▶ correction de l'étiquetage syntaxique : 750 mots à l'heure, ? heures par jour

Penn Treebank (PTB) [Marcus et al., 1993]

corpus arboré de l'Université de Pennsylvanie

PTB 1 :

- ▶ correction de l'étiquetage morpho-syntaxique : 3 000 mots à l'heure, 3 heures par jour
- ▶ correction de l'étiquetage syntaxique : 750 mots à l'heure, 3 heures par jour

Penn Treebank (PTB) [Marcus et al., 1993]

corpus arboré de l'Université de Pennsylvanie

PTB 1 :

- ▶ correction de l'étiquettage morpho-syntaxique : 3 000 mots à l'heure, 3 heures par jour
- ▶ correction de l'étiquettage syntaxique : 750 mots à l'heure, 3 heures par jour
- ▶ + courbe d'apprentissage de 1 mois (étiquettage morpho-syntaxique) à 2 mois (syntaxe) !

Prague Dependency Treebank [Böhmová et al., 2001]

corpus en dépendances de Prague

- ▶ 1996-2004 [Böhmová et al., 2001],
- ▶ construit à partir du CNC (*Czech National Corpus*),
- ▶ 3 niveaux de structure :
 1. morphologique (semi-automatique) : 1,8 millions de tokens
 2. analytique (syntaxe en dépendances, avec un outil adapté)
 3. tectogrammatical (sémantique) : 1 million de tokens

Prague Dependency Treebank [Böhmová et al., 2001]

Version 1.0 :

- ▶ annotation manuelle des niveaux morphologique et analytique
- ▶ temps : ?
- ▶ nombre de personnes : ?
- ▶ Estimation du coût : ?

Prague Dependency Treebank [Böhmová et al., 2001]

Version 1.0 :

- ▶ annotation manuelle des niveaux morphologique et analytique
- ▶ temps : 5 ans
- ▶ nombre de personnes : ?
- ▶ Estimation du coût : ?

Prague Dependency Treebank [Böhmová et al., 2001]

Version 1.0 :

- ▶ annotation manuelle des niveaux morphologique et analytique
- ▶ temps : 5 ans
- ▶ nombre de personnes : 22 personnes, dont 17 simultanément pendant les périodes les plus exigeantes
- ▶ Estimation du coût : ?

Prague Dependency Treebank [Böhmová et al., 2001]

Version 1.0 :

- ▶ annotation manuelle des niveaux morphologique et analytique
- ▶ temps : 5 ans
- ▶ nombre de personnes : 22 personnes, dont 17 simultanément pendant les périodes les plus exigeantes
- ▶ Estimation du coût : 600 000 \$

GENIA [Kim et al., 2008]

GENIA : 400 000 mots annotés en microbiologie.

GENIA [Kim et al., 2008]

GENIA : 400 000 mots annotés en microbiologie.

⇒ 5 annotateurs à mi-temps, 1 coordinateur sénior, 1 coordinateur junior pendant 1,5 an [Kim et al., 2008]

GENIA [Kim et al., 2008]

GENIA : 400 000 mots annotés en microbiologie.

⇒ 5 annotateurs à mi-temps, 1 coordinateur sénior, 1 coordinateur junior pendant 1,5 an [Kim et al., 2008]

⇒ la qualité doit être élevée !

ESTER

- ▶ 100 h de parole transcrite (campagne d'évaluation ESTER, systèmes de transcription, 2008)
- ▶ 1 h de parole = ?

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

Durée de vie des corpus annotés

Penn Treebank [Marcus et al., 1993] :

- ▶ créé au début des années 90
- ▶ encore utilisé

vs tagger PARTS [Church, 1988], qui n'est plus du tout utilisé/connu

→ évolution rapide des outils

⇒ l'annotation ne doit **pas** dépendre d'eux

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



Le TAL aujourd'hui

Annotation manuelle et TAL

Parenthèse : présentations

Qu'est-ce qu'annoter ?

Pour finir

À faire pour la fin du semestre (dernier cours)

Une présentation de 15 min (+5 min), par groupes de 2 sur :

- ▶ un corpus annoté **manuellement**
- ▶ **original** soit :
 - ▶ par la langue traitée,
 - ▶ par le domaine,
 - ▶ par l'annotation portée

Évaluation

Critères :

- ▶ choix du corpus
- ▶ qualité du travail
- ▶ qualité de la présentation
- ▶ réponses aux questions

Attention : la note ne sera pas forcément la même pour les différents membres du groupe

Le TAL aujourd'hui

Annotation manuelle et TAL

Parenthèse : présentations

Qu'est-ce qu'annoter ?

Exercice

DéfinitionS

Annoter, pour quoi ?

Pour finir

Exercice

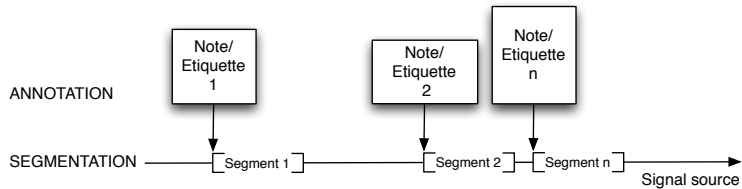
Transcrivez ce que vous entendez du fichier son qui vous est fourni, en utilisant le logiciel Transcriber ou Praat (s'il est présent sur la machine) ou sur papier (oui oui).

Definition

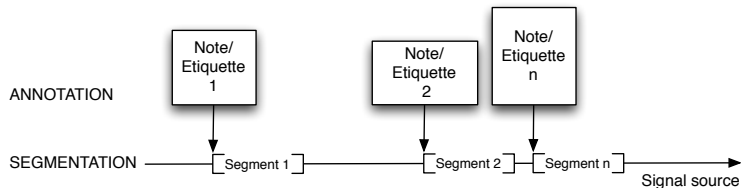
*“[corpus annotation] can be defined as the practice of adding **interpretative**, linguistic information to an electronic corpus of spoken and/or written language data. ‘Annotation’ can also refer to the end-product of this process” [Leech, 1997]*

*“‘Linguistic annotation’ covers any **descriptive** or **analytic** notations applied to raw language data. The basic data may be in the form of time functions - audio, video and/or physiological recordings - or it may be textual.” [Bird and Liberman, 2001]*

Annotation



Annotation



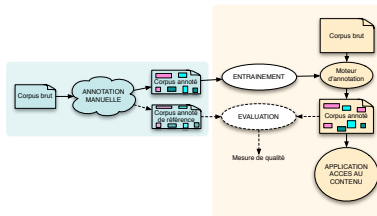
Ajout d'informations **interprétatives** [Leech, 1997, Habert, 2005]

L'application : horizon de l'annotation

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

- ▶ visée applicative directe (résumés de matchs pour la campagne football)
- ▶ application intermédiaire ou interne au TAL (étiquetage morpho-syntaxique)

Une annotation est d'autant plus utile qu'elle a été conçue en fonction d'une application spécifique [Leech, 2005].



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 gestionnaire de la campagne
- ▶ **évaluer** le consensus (la cohérence)

Annoter des unités polylexicales

Exercice : annoter des unités polylexicales

- ▶ allez sur Rigor Mortis : <http://rigor-mortis.org/>
- ▶ faites toute la phase 1, puis la 2 et la 3.

Le TAL aujourd'hui

Annotation manuelle et TAL

Parenthèse : présentations

Qu'est-ce qu'annoter ?

Pour finir

CQFR : Ce Qu'il Faut Retenir
À faire à la maison



Annotation manuelle et TAL :

- ▶ utilisation
- ▶ coût

Annotation manuelle :

- ▶ définition
- ▶ interprétation

Lire un article de recherche (noté)

Lire avec attention et me rendre la fiche de lecture

[Dandapat et al., 2009] :

Complex Linguistic Annotation No Easy Way Out !

A Case from Bangla and Hindi POS Labeling Tasks

<http://www.aclweb.org/anthology/W/W09/W09-3002.pdf>

Comment lire un article de recherche (en anglais) ?

- ▶ Identifier le contexte : qui écrit, d'où, dans le cadre de quel projet ?
- ▶ Lire tout (ne chercher que les mots inconnus qui sont répétés ou posent problème pour la compréhension)
- ▶ Identifier le type d'article :
 - ▶ position paper (prise de position, point de vue)
 - ▶ description d'expérience
 - ▶ état de l'art (point sur le domaine)
 - ▶ autre ?
- ▶ Relire (au moins) l'introduction et les conclusions tirées
- ▶ Vérifier :
 - ▶ avez-vous bien compris les conclusions de l'article ?
 - ▶ quels sont les biais ? sont-ils identifiés par les auteurs ?



Bird, S. and Liberman, M. (2001).

A formal framework for linguistic annotation.

Speech Communication, 33(1-2) :23–60.



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.



Church, K. W. (1988).

A stochastic parts program and noun phrase parser for unrestricted text.

In Proceedings of the Second Conference on Applied Natural Language Processing, ANLC '88, pages 136–143, Stroudsburg, PA, USA. Association for Computational Linguistics.



Dandapat, S., Biswas, P., Choudhury, M., and Bali, K. (2009).

Complex linguistic annotation - no easy way out ! a case from bangla and hindi POS labeling tasks.

In Proceedings of the third ACL Linguistic Annotation Workshop, Singapour.



Desrosières, A. (2008).

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

Presses de l'école des Mines de Paris.



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.



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).



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

Corpus annotation for mining biomedical events from literature.

BMC Bioinformatics, 9(1) :10.



Leech, G. (1997).

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

Longman, Londres, Angleterre.



Leech, G. (2005).

Developing Linguistic Corpora : a Guide to Good Practice, chapter Adding Linguistic Annotation, pages 17–29.

Oxford : Oxbow Books.



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.



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.