Le Traitement Automatique des Langues en France à l'ère du Big Data

À l'aube d'un révolution technologique

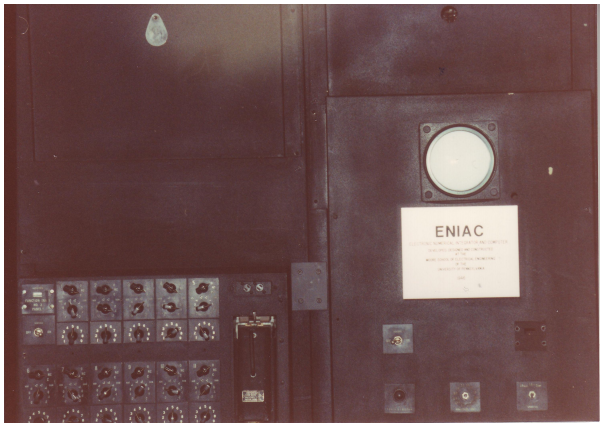
Patrick Paroubek

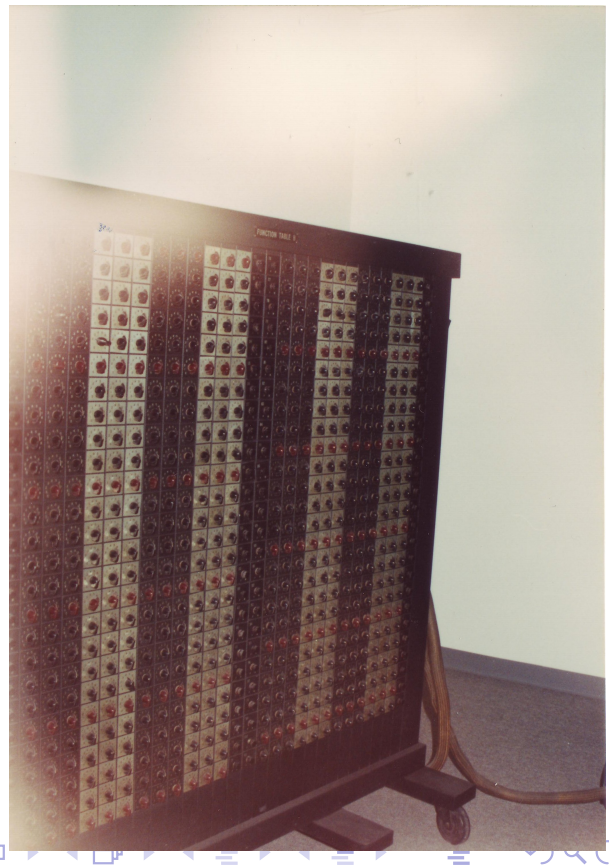
LIMSI-CNRS
Dépt. CHM - Groupe LIR
Bât. 508 Université Paris Sud, 91403 Orsay Cedex
pap@limsi.fr

ATALA - Association pour le Traitement Automatique des Langues
45, rue d'Ulm
75230 PARIS Cedex 5 - France
http://www.atala.org

Naissance de l'informatique

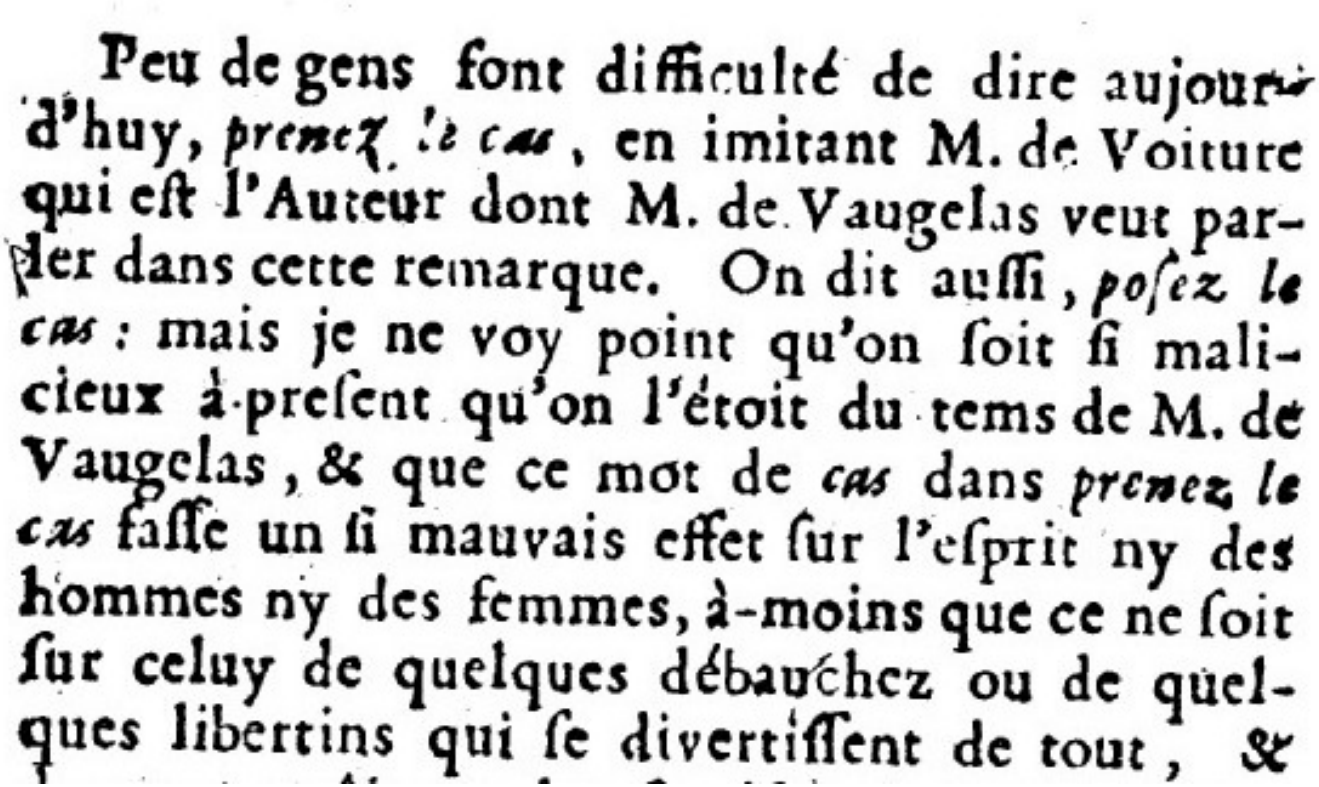
ENIAC - 13 février 1946 - *Moore School of Elec. Engineering* U Penn





La langue naturelle

NOUVELLES REMARQUES *de M. DE VAUGELAS* SUR LA LANGUE FRANÇOISE, Paris, Guillaume Desprez, 1690.



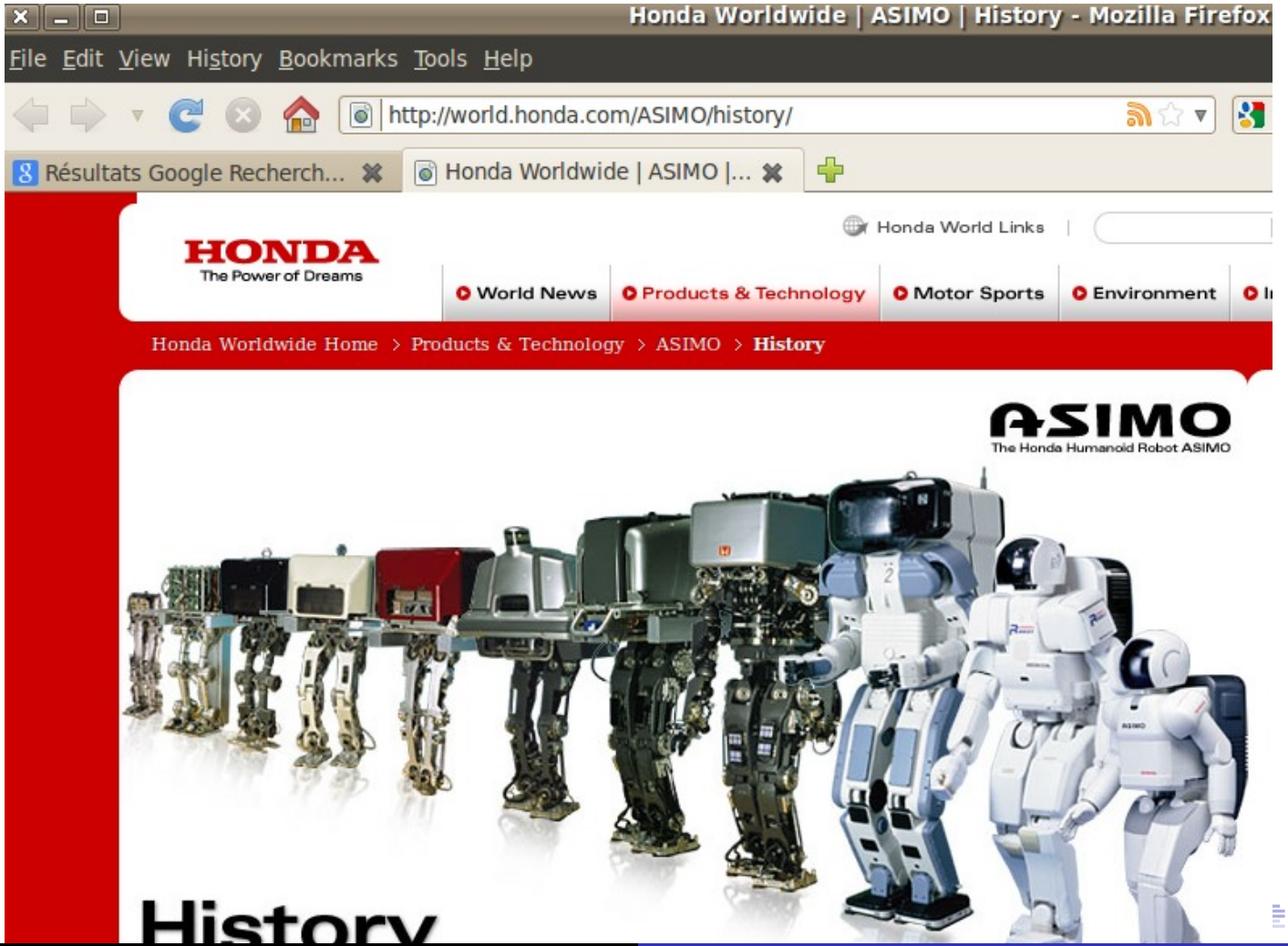
Peu de gens font difficulté de dire aujourd'huy, *prenez le cas*, en imitant M. de Voiture qui eſt l'Auteur dont M. de Vaugelas veut parler dans cette remarque. On dit auſſi, *poſez le cas* : mais je ne voy point qu'on ſoit ſi malicieux à preſent qu'on l'étoit du tems de M. de Vaugelas, & que ce mot de *cas* dans *prenez le cas* faſſe un ſi mauvais effet ſur l'eſprit ny des hommes ny des femmes, à-moins que ce ne ſoit ſur celuy de quelques débauchez ou de quelques libertins qui ſe divertiſſent de tout, &

Naissance du TAL

En France, l'ATALA (Association pour le Traitement Automatique des Langues) est créée en 1959.
`http://www.atala.org`

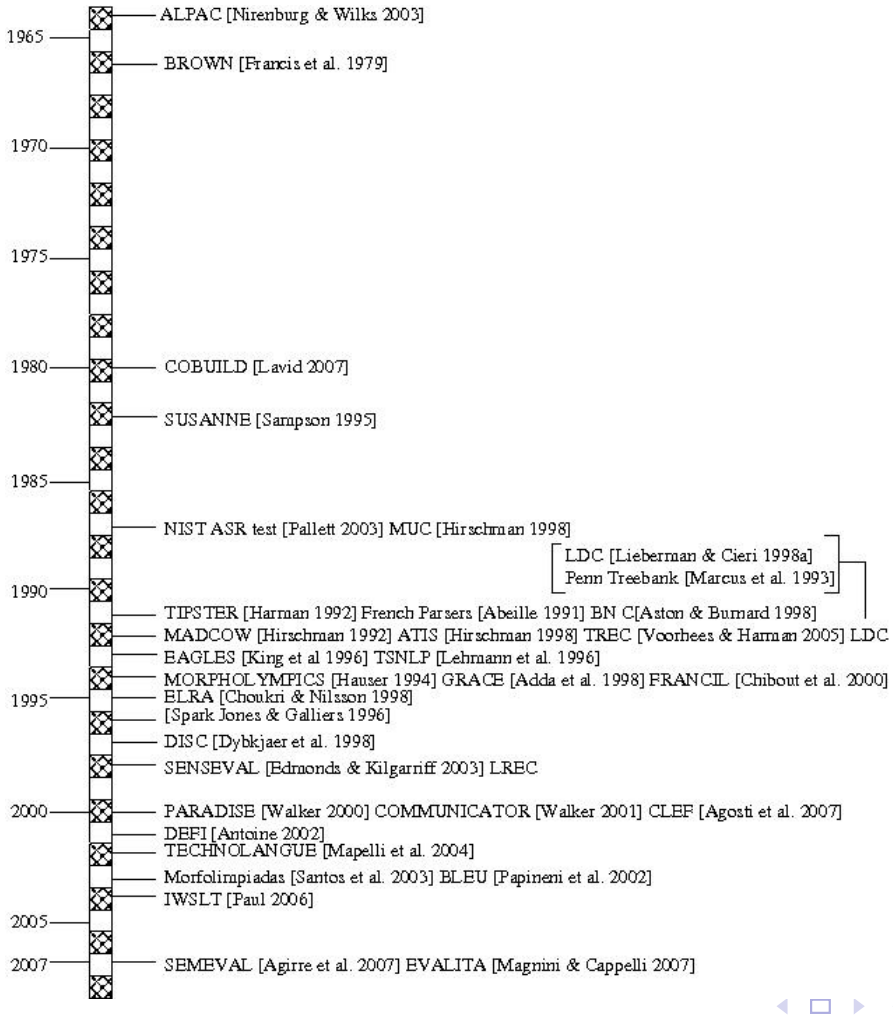
Aux États-Unis, l'ACL (Association for Computational Linguistics) est créée en 1962. `http://www.acl.org`

Evolution
Honda Asimo robot



Evolution

Vers plus de corpus et plus d'ingénieries



LA RECHERCHE EN TAL

Si l'analyse et la production de textes sont naturels pour les êtres humains, il n'en va pas de même pour les ordinateurs : les langues naturelles résistent à la formalisation complète qui constitue le préalable habituel à un traitement algorithmique. Cette situation motive des recherches de fond sur le traitement automatique des langues, recherches menées dans le public (universités, organismes de recherche) comme dans le privé (centres de recherche des grandes entreprises informatiques) en linguistique et en informatique.
En linguistique, les chercheurs modélisent plus complètement et plus finement les propriétés des langues à tous leurs paliers (morphologie, syntaxe, sémantique, pragmatique) en ayant en ligne de mire une formalisation suffisante pour permettre leur implantation en machine. Ce travail inclut la réalisation de grammaires formelles et de grands dictionnaires utilisables par des programmes informatiques.
En informatique, il faut mettre au point des algorithmes d'analyse et de génération de langue naturelle, mais aussi des formalismes de représentation des connaissances linguistiques satisfaisant des contraintes contradictoires d'expressivité, de calculabilité et de robustesse. Par ailleurs, de la même façon que les êtres humains acquièrent des connaissances linguistiques par l'expérience quotidienne et la lecture, certaines méthodes à base statistique parcourent le contenu de grandes collections de textes pour engranger automatiquement des connaissances qui serviront ensuite à analyser ou produire d'autres textes.
Une part importante de ces travaux se situe à la rencontre de ces disciplines et nécessite une double compétence.

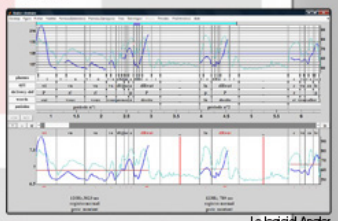
L'ENSEIGNEMENT DU TAL EN FRANCE

L'enseignement du TAL en France

L'enseignement universitaire du Traitement Automatique des Langues se répartit uniformément sur le territoire français, aussi bien au niveau Licence qu'au niveau Master. Le TAL étant une discipline fondamentalement interdisciplinaire entre informatique et linguistique, les UFR de rattachement des formations sont pour moitié des formations en Sciences du Langage, pour l'autre moitié en Informatique, toutes adossées à des laboratoires de recherche ayant une composante forte en TAL.
Selon les formations, une trentaine recensées à ce jour, le TAL est enseigné comme un composant d'une formation professionalisante ou comme une finalité en soi de formation professionnelle et/ou de recherche. Dans un cas comme dans l'autre, les responsables de formation s'accordent sur le fait que, issus des sciences du langage ou de l'informatique, les étudiants ont un fort taux d'insertion dans l'industrie, notamment parce que les enseignements évoluent au gré des avancées de la discipline.
La variété des enseignements et des métiers visés permet de répondre aux besoins du monde professionnel tout en assurant un renouvellement et un accroissement de la communauté scientifique universitaire en traitement automatique des langues. La demande de stagiaires et de professionnels ne décroît pas malgré la crise actuelle, et justifie pleinement la pérennisation des formations en traitement automatique des langues.

Une application du TAL

LE TAL EN QUELQUES CHIFFRES



Le logiciel Analor

Le TAL en France c'est :
- 27 formations universitaires (source REPTIL) ;
- 8 métiers types : ergolinguiste, chef de projet, consultant avant-vente, linguiste informaticien, ingénieur assurance qualité, terminologue, veilleur stratégique, ingénieur support (source TECHNOLANGUE) ;
- 5 secteurs d'activité : moteurs de recherche intelligents, gestion de contenu, applications vocales et multimodales, e-learning et traduction automatique (source TECHNOLANGUE) ;
- Une place dans le trio des leaders Européen sur le marché du TAL (source Bureau van Dijk Information Management 2006) ;
- de 82 à 150 sociétés (sources APIL 2006 et ministère de la Culture 2007) ;
- 500 professionels de haut niveau en ingénierie linguistique appartenant tant au secteur de la recherche publique qu'au secteur industriel (ministère de la Culture 2007).

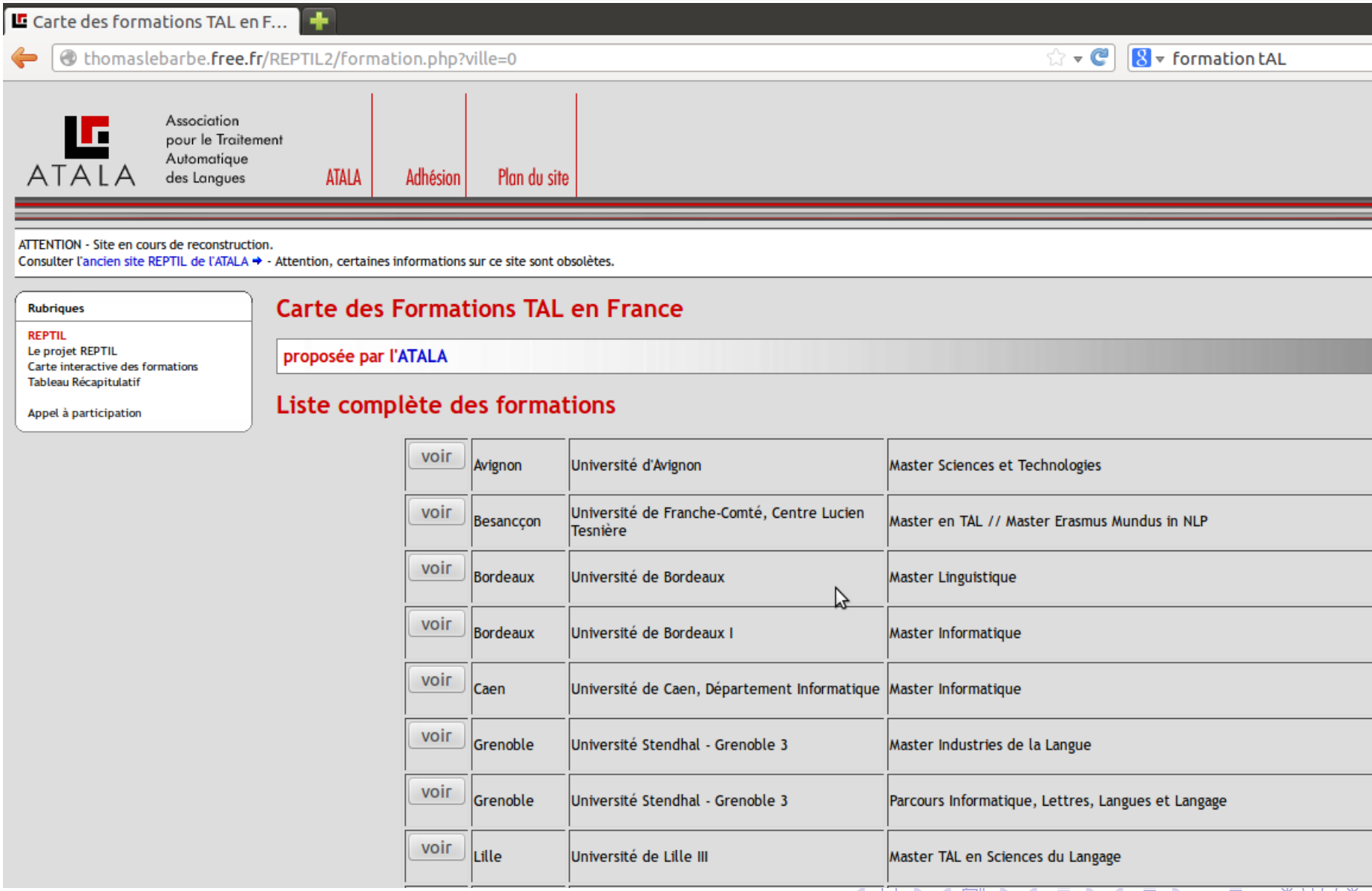
Traduction automatique
Communication homme-machine
Extraction d'information
Analyse de données
Représentation des connaissances

ATALA
Association pour le Traitement Automatique des Langues
www.atala.org

Traitement Automatique des Langues

Le TAL en France

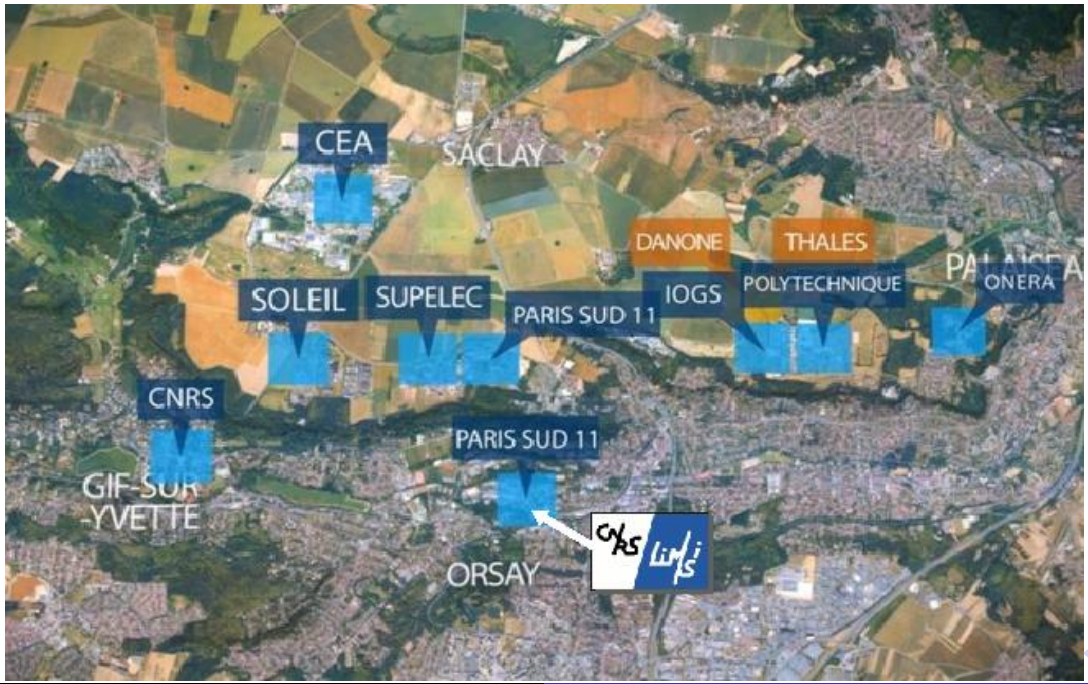
- une trentaine de formations Unversitaires
- 8 métiers types : linguiste informaticien, terminologue, ergolinguiste, chef de projet, consultant avant-vente, ingénieur assurance qualité, veilleur stratégique, ingénieur support,
- 5 secteurs d'activité : moteurs de recherche intelligents, analyse et gestion de contenu, application vocales et multimodales, e-learning et traduction automatique ,
- une place dans le trio des leaders Européens du marché du TAL ,
- une centaine de sociétés,
- plusieurs centaines de professionels de haut niveau en ingénierie linguistique, recherche publique et industrielle



voir	Avignon	Université d'Avignon	Master Sciences et Technologies
voir	Besançcon	Université de Franche-Comté, Centre Lucien Tesnière	Master en TAL // Master Erasmus Mundus in NLP
voir	Bordeaux	Université de Bordeaux	Master Linguistique
voir	Bordeaux	Université de Bordeaux I	Master Informatique
voir	Caen	Université de Caen, Département Informatique	Master Informatique
voir	Grenoble	Université Stendhal - Grenoble 3	Master Industries de la Langue
voir	Grenoble	Université Stendhal - Grenoble 3	Parcours Informatique, Lettres, Langues et Langage
voir	Lille	Université de Lille III	Master TAL en Sciences du Langage

Le LIMSI-CNRS

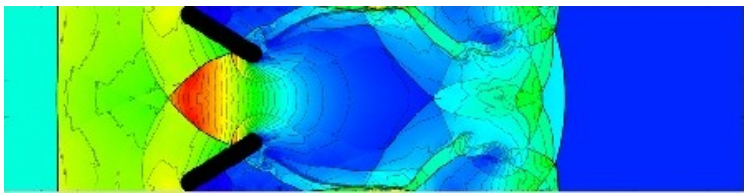
- Laboratoire pour
  la Mécanique et les Sciences de l'Ingénieur
- Unité propre du CNRS
- localisation : campus Paris-Saclay
- effectif : env. 200 permanents + 120 stagiaires

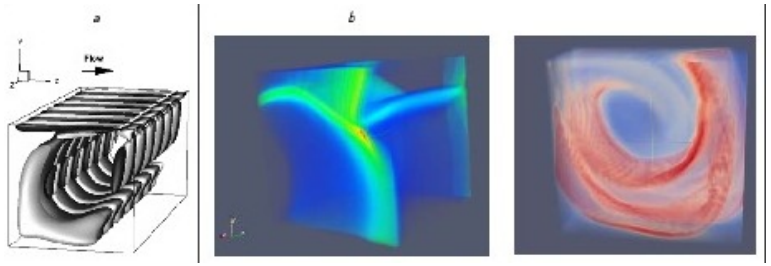


Le LIMSI-CNRS

2 départements :

- Dépt. Mécanique-Énergétique (3 groupes)
  - Aérodynamique Instationnaire : turbulence et contrôle
  - Convection et Rotation : instabilités et turbulence
  - Transferts Solide-Fluide

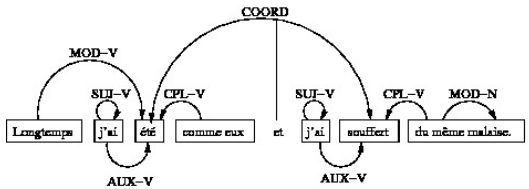


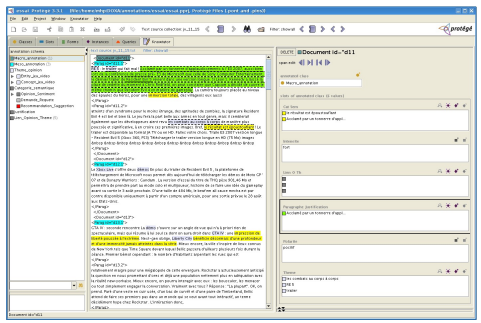


Le LIMSI-CNRS

2 départements :

- **Dépt. Communication Homme-Machine** (6 groupes)
  - Audio & Acoustique
  - Architectures et Modèles pour l'Interaction
  - Réalité Virtuelle & Augmentée
  - Cognition, Perception et Usages
  - Traitement du Langage Parlé
  - **Information, Langue Ecrite et Signée**





Le groupe ILES

Effectif : env. 20 permanents, 15 Doctorants, 15 CDD

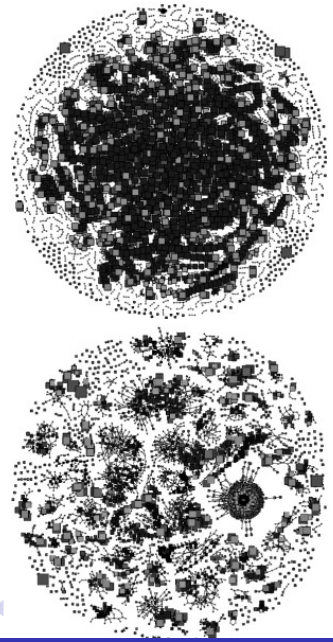
- Corpus & représentations
- Multilinguisme & paraphrases
- Extraction d'information précise & systèmes de Questions-Réponses
- Langues des signes

Langage

Pas de norme

- évolution permanente
- Académie Française (1635)
- langue = consensus émergent au sein d'une population

Expérience d'émergence de lexique d'actions (thèse J.

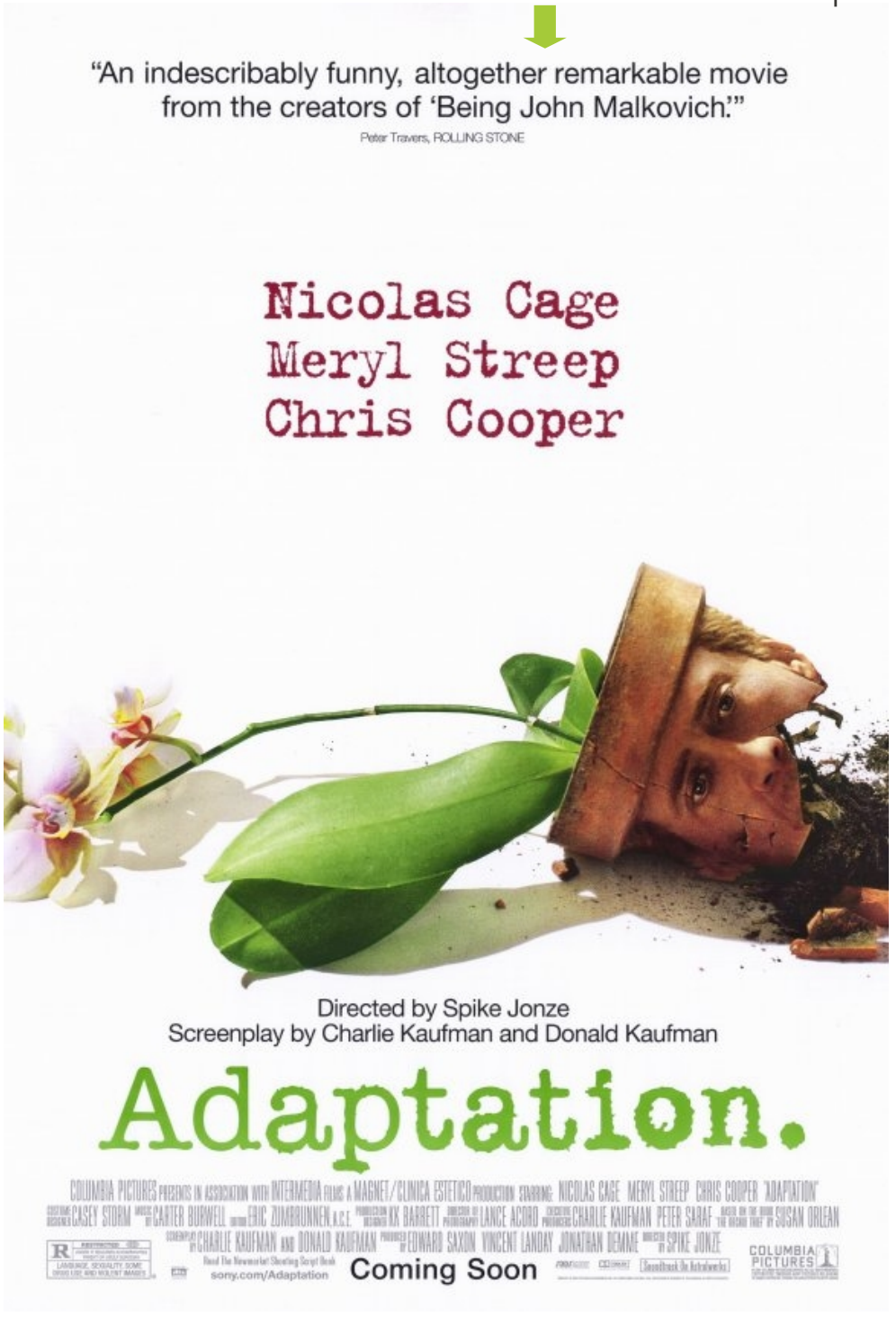


Le rapport du Mac Kinsey Global Institute
http ://www.economie.gouv.fr/files/rapport-mckinsey-company.pdf,
paru en 2012, positionne le Big Data et le Big Analytics parmi les 10 technologies stratégiques pouvant créer un impact fort au sein des entreprises pour les 3 prochaines années à venir.

En 2012, Twitter a publié les statistiques suivantes :

- 465 millions de comptes Twitter,
- 175 millions de Tweets/jour,
- des pointes de 2500 Tweets/seconde,
- +grand nombre de « followers » : 19 millions / Lady Gaga,
- 64% des accès à Twitter se font par Internet,
- 69% des utilisateurs suivent leurs amis,
- 110 nouveaux comptes/seconde,
- près d'un million de nouveaux comptes/jour.
- des un pic record de 24,1 million de tweets envoyés durant la nuit du Super Bowl du 03 févier 2013
- l'anglais +utilisé, 39% des messages en octobre 2011 soit 70 millions de tweets quotidiens,
- MAIS +60% des tweets sont rédigés dans d'autres langues.

How can we determine that this is a positive opinion?



“An indescribably funny, altogether remarkable movie from the creators of 'Being John Malkovich'”

- We can determine positive polarity by finding positive words
- To know which words are positive – we use affective lexicons (ex: ANEW)

ANEW (Affective Norms of English Words) is a list of 1034 English words with assigned scores of polarity, intensity, and control

Description	Valence	(sd)	Arousal		Dominance	
abduction	2.76	2.06	5.53	2.43	3.49	2.38
abortion	3.5	2.3	5.39	2.8	4.59	2.54
absurd	4.26	1.82	4.36	2.2	4.73	1.72
abundance	6.59	2.01	5.51	2.63	5.8	2.16
abuse	**1.8**	1.23	6.83	2.7	3.69	2.94
acceptance	7.98	1.42	5.4	2.7	6.64	1.91
accident	2.05	1.19	6.26	2.87	3.76	2.22
ace	6.88	1.93	5.5	2.66	6.39	2.31
ache	2.46	1.52	5	2.45	3.54	1.73
achievement	**7.89**	1.38	5.53	2.81	6.56	2.35
activate	5.46	0.98	4.86	2.56	5.43	1.84
addict	2.48	2.08	5.66	2.26	3.72	2.54

Valence = polarity takes a value between 1 and 9

Higher value => more positive term

Twitter is the most popular microblogging platform for exchanging short text messages

1. Use emoticons :) to collect positive and negative messages
2. Estimate words polarity based on occurrences in sets

Table VI
AN EXAMPLE OF POSITIVE WORD LIST

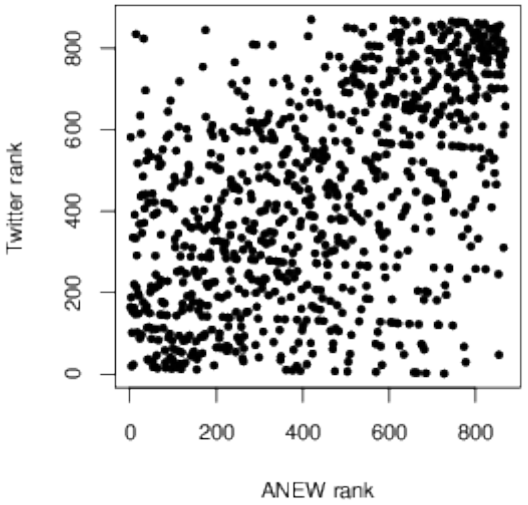
Word	**ANEW**	**Est.**
sexy	8.02	8.18
adorable	7.81	7.67
plaisir	8.28	7.64
salut	5.92	7.32
magnifique	7.6	7.08
anniversaire	7.84	6.68
gentil	7.59	6.45
bien	7.47	6.2
simple	7.1	6.18
beau	6.55	5.88

Table VII
AN EXAMPLE OF NEGATIVE WORD LIST

word	**ANEW**	**Est.**
triste	1.61	1.24
terrible	1.93	1.53
malade	1.9	1.73
dommage	2.49	1.79
mort	1.61	2.82
mal	2.46	2.93
dur	4.74	3
peur	2.76	3.26
seul	2.41	3.61
partie	5.11	4.05

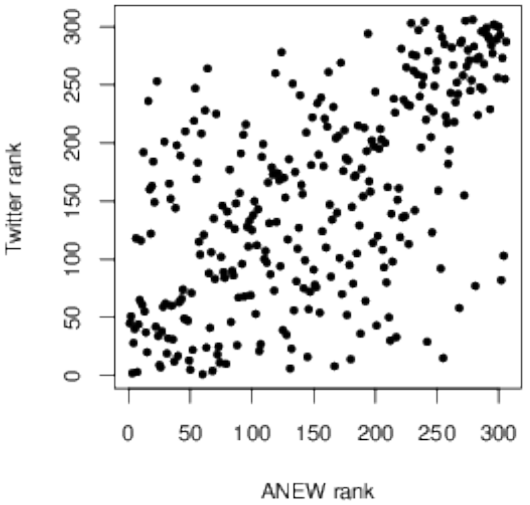
Correlation with ANEW

$\tau = 0.36$



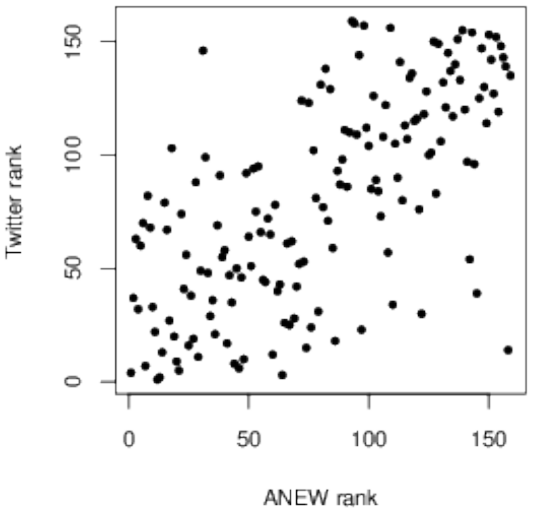
All words

$\tau = 0.45$



Occurring at least
in 100 messages

$\tau = 0.55$



Adjectives

Polarity classification in video game reviews

Model	**Ave. acc**	**Ave. prec**	**Prec$_{pos}$**	**Prec$_{neg}$**
Unigram	73.86	69.57	90.57	48.57
Bigram	72.73	69.11	86.79	51.43
Trigram	64.77	60.08	83.02	37.14
Twitter	71.59	68.16	84.90	51.40

Our naive method worked "almost as good" as a standard N-gram based approach. However, our method does not require training data.

Text representation using dependency tree subgraphs

Traditional approaches use **bag-of-words** text representation model and an **SVM** classifier

Given two sets of texts: **positive** and **negative**

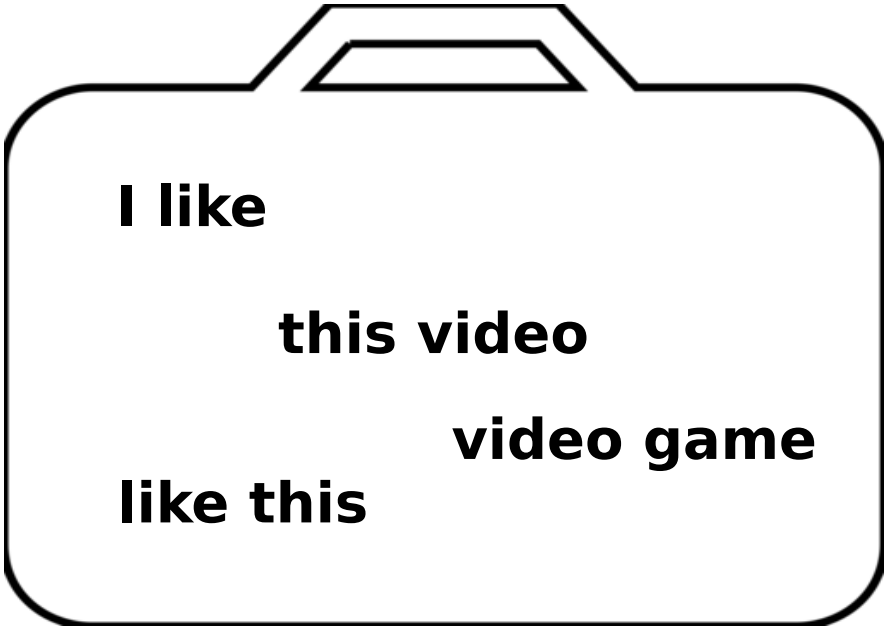
Represent each text using **bag-of-words** model

Construct a **feature vector** for each text

Train an **SVM** classifier (or any other ML)

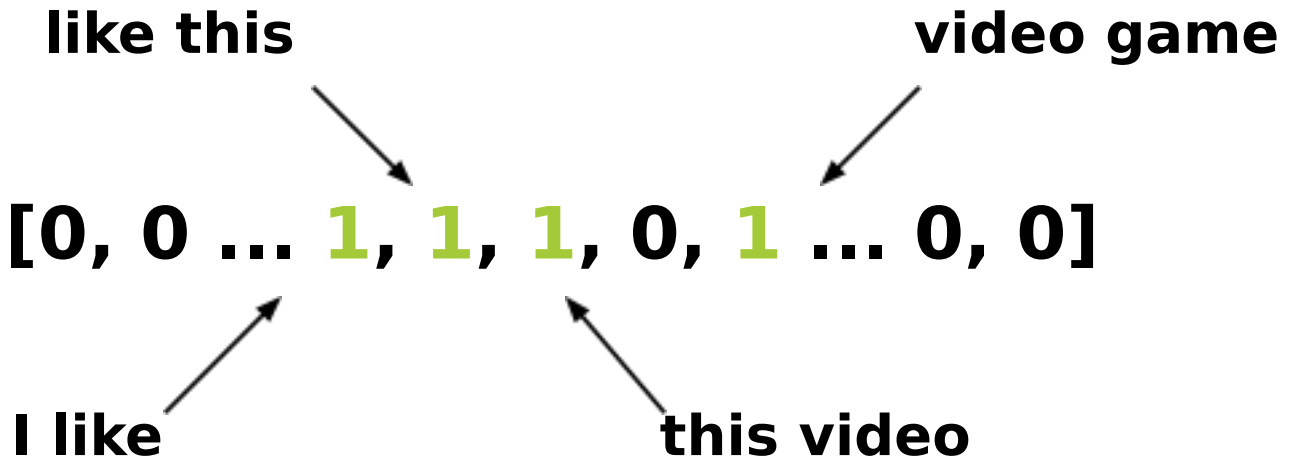
Classify an **unknown** text

Bag-of-words model



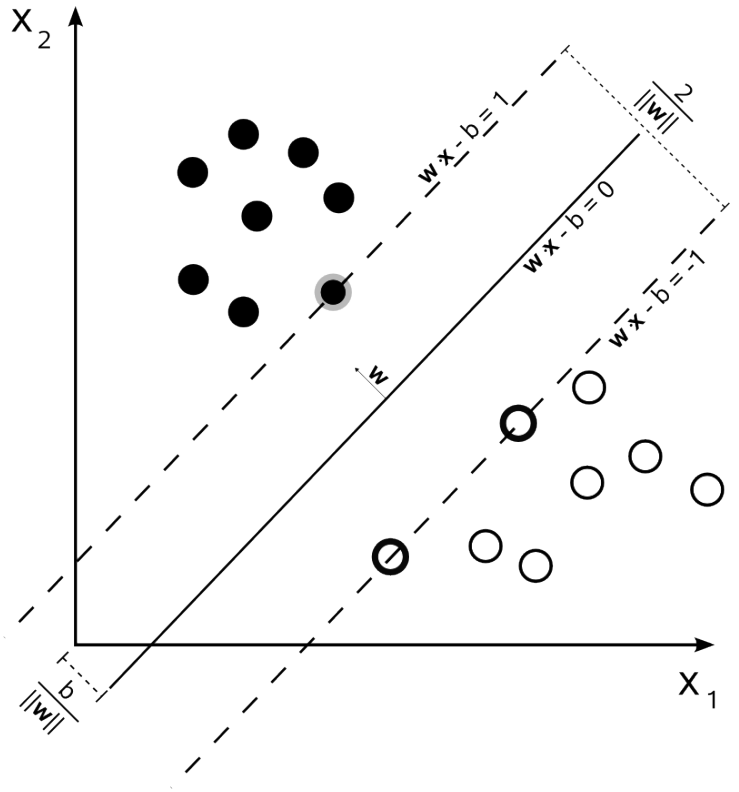
"I like this video game"

Feature vector



"I like this video game"

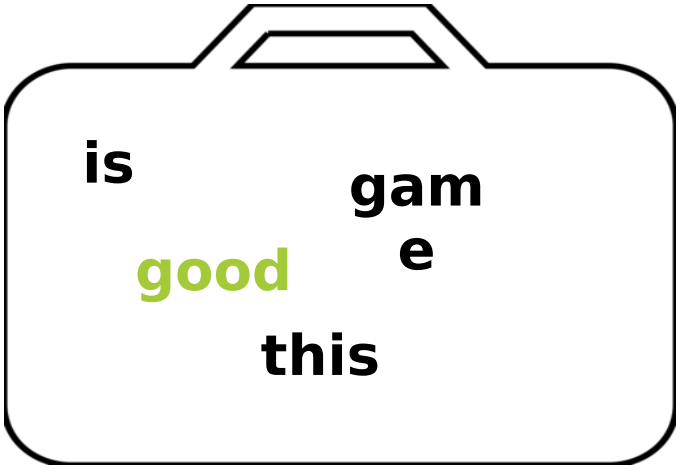
SVM



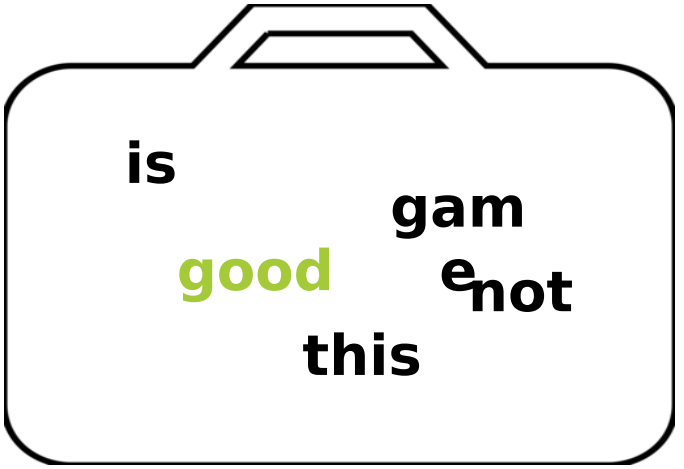
"I like this video game"

Problems of bag-of-words model

This game is good



This game is not good



By considering words as **independent** terms, we lose information from the **links** between words

Problems of bag-of-words model

This game is good

This game is not good

To solve this problem, we can try
to use **n-grams** of a higher order

Problems of bag-of-words model

This game is not good

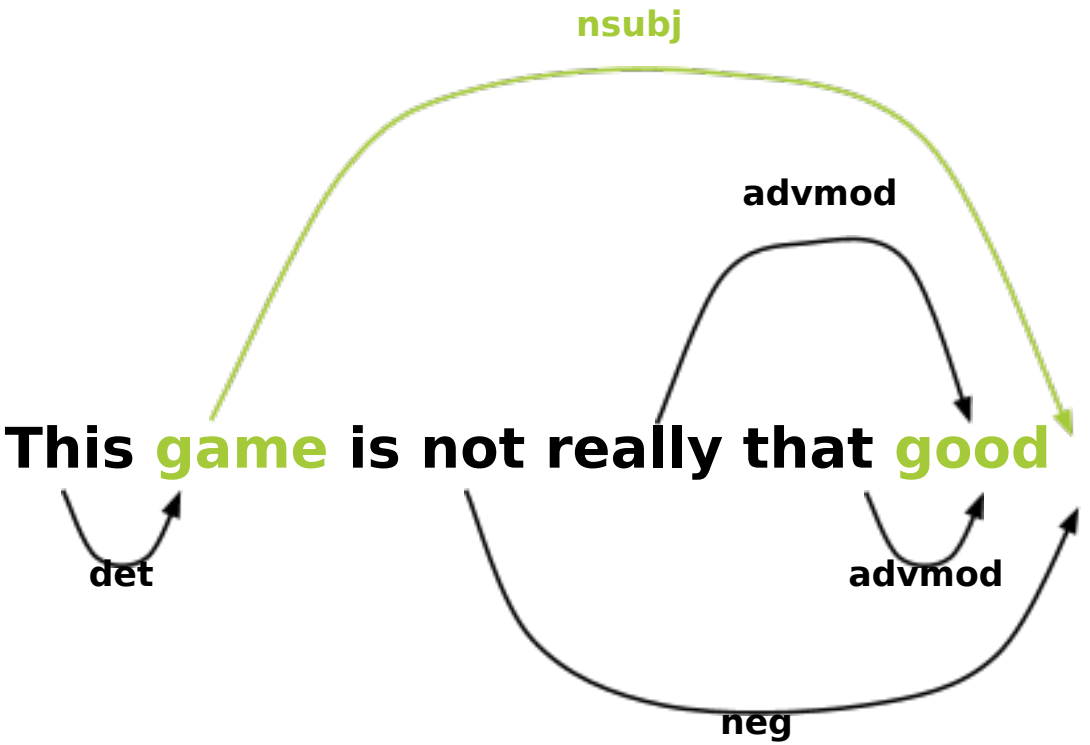
This game is not really that good

But we still have a problem with **long dependencies**

Our method

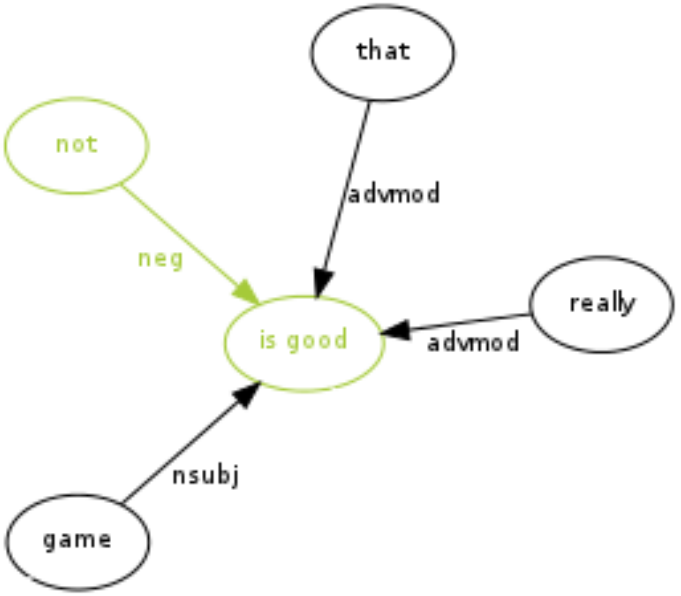
We propose to use sentence **dependency tree** for text representation

Dependencies



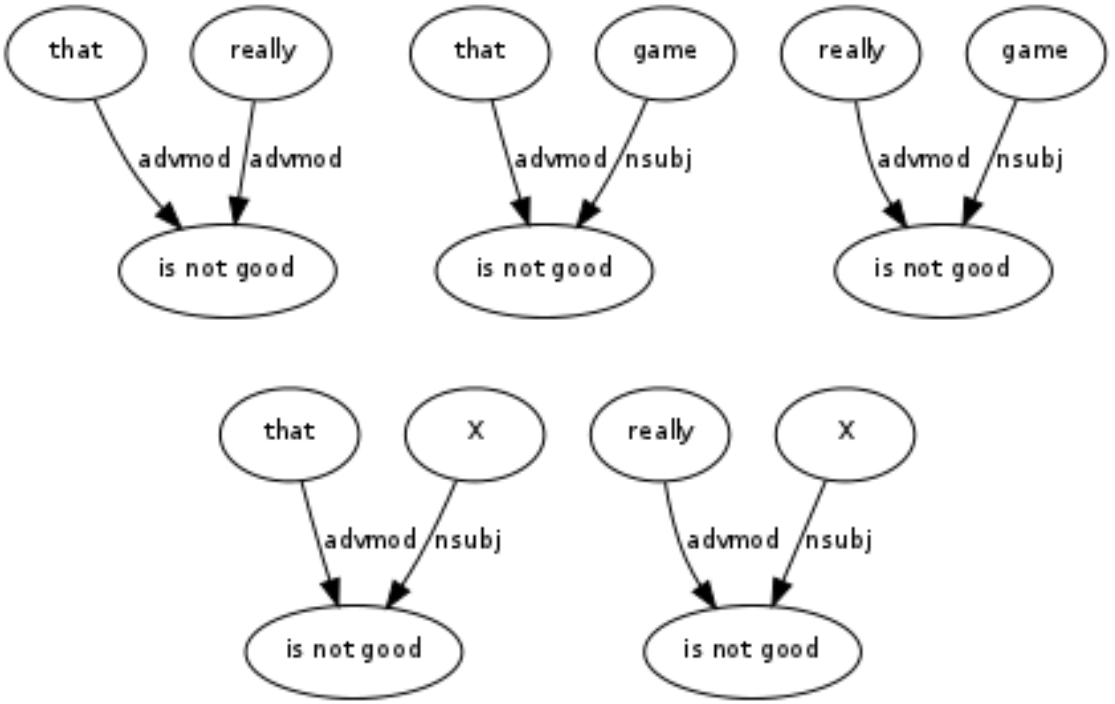
And find the **opinion subject**

Dependency tree



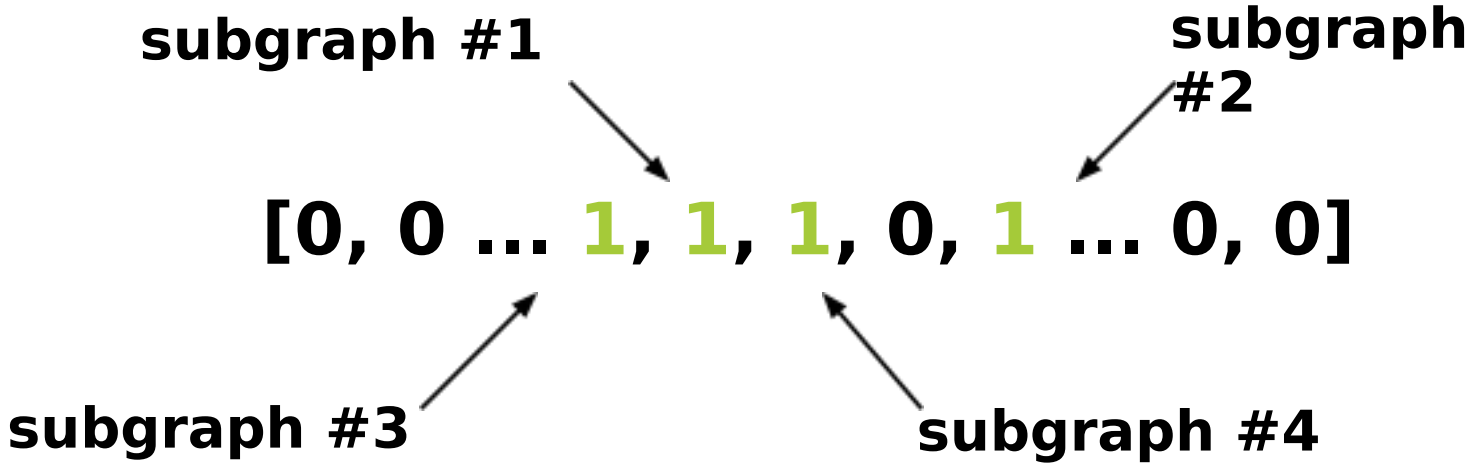
Next, is **combining nodes** with partial meaning

Subgraph text representation



dependency tree subgraph text representation

Feature vector



It can be used the same way as the bag-of-words model,
e.g.: to train an **SVM** classifier

- Le Cloud Computing, rupture technologique dans la gestion des données et la conception des systèmes informatiques,
- Le Cloud Computing substrat technologique et technique facilement assimilable,
- Les approches Big Data sont complexes, multiformes et encore en voie de définition.
- Le Data Mining et la Visualisation de Données préexistants et associés au Big Data n'ont pas encore donné pleinement la mesure de leurs potentialités.

Le rapport du Mac Kinsey Global Institute
http ://www.economie.gouv.fr/files/rapport-mckinsey-company.pdf,
paru en 2012, positionne le Big Data et le Big Analytics parmi les 10 technologies stratégiques pouvant créer un impact fort au sein des entreprises pour les 3 prochaines années à venir.

- Le traitement de telles données imposent de faire appel à des technologies de rupture à tous les niveaux : infrastructure, stockage, analyse et visualisation.
- Pour résoudre le problème de (« scalabilité »), il faut de nouvelles technologies associant **Big Data**, **Big Analytics**, **Visual Analytics** et **Cloud Computing**.

- Le traitement de telles données imposent de faire appel à des technologies de rupture à tous les niveaux : infrastructure, stockage, analyse et visualisation.
- Pour résoudre le problème de (« scalabilité »), il faut de nouvelles technologies associant **Big Data**, **Big Analytics**, **Visual Analytics** et **Cloud Computing**.

Mais il y aura toujours de la place pour les petits...dans des niches qui vont se multiplier...

LA RECHERCHE EN TAL

Si l'analyse et la production de textes sont naturels pour les êtres humains, il n'en va pas de même pour les ordinateurs : les langues naturelles résistent à la formalisation complète qui constitue le préalable habituel à un traitement algorithmique. Cette situation motive des recherches de fond sur le traitement automatique des langues, recherches menées dans le public (universités, organismes de recherche) comme dans le privé (centres de recherche des grandes entreprises informatiques) en linguistique et en informatique.
En linguistique, les chercheurs modélisent plus complètement et plus finement les propriétés des langues à tous leurs paliers (morphologie, syntaxe, sémantique, pragmatique) en ayant en ligne de mire une formalisation suffisante pour permettre leur implantation en machine. Ce travail inclut la réalisation de grammaires formelles et de grands dictionnaires utilisables par des programmes informatiques.
En informatique, il faut mettre au point des algorithmes d'analyse et de génération de langue naturelle, mais aussi des formalismes de représentation des connaissances linguistiques satisfaisant des contraintes contradictoires d'expressivité, de calculabilité et de robustesse. Par ailleurs, de la même façon que les êtres humains acquièrent des connaissances linguistiques par l'expérience quotidienne et la lecture, certaines méthodes à base statistique parcourent le contenu de grandes collections de textes pour engranger automatiquement des connaissances qui serviront ensuite à analyser ou produire d'autres textes.
Une part importante de ces travaux se situe à la rencontre de ces disciplines et nécessite une double compétence.

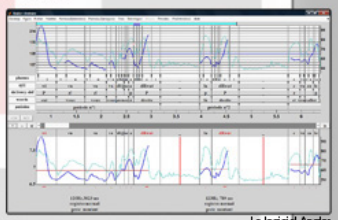
L'ENSEIGNEMENT DU TAL EN FRANCE

L'enseignement du TAL en France

L'enseignement universitaire du Traitement Automatique des Langues se répartit uniformément sur le territoire français, aussi bien au niveau Licence qu'au niveau Master. Le TAL étant une discipline fondamentalement interdisciplinaire entre informatique et linguistique, les UFR de rattachement des formations sont pour moitié des formations en Sciences du Langage, pour l'autre moitié en Informatique, toutes adossées à des laboratoires de recherche ayant une composante forte en TAL.
Selon les formations, une trentaine recensées à ce jour, le TAL est enseigné comme un composant d'une formation professionalisante ou comme une finalité en soi de formation professionnelle et/ou de recherche. Dans un cas comme dans l'autre, les responsables de formation s'accordent sur le fait que, issus des sciences du langage ou de l'informatique, les étudiants ont un fort taux d'insertion dans l'industrie, notamment parce que les enseignements évoluent au gré des avancées de la discipline.
La variété des enseignements et des métiers visés permet de répondre aux besoins du monde professionnel tout en assurant un renouvellement et un accroissement de la communauté scientifique universitaire en traitement automatique des langues. La demande de stagiaires et de professionnels ne décroît pas malgré la crise actuelle, et justifie pleinement la pérennisation des formations en traitement automatique des langues.

Une application du TAL

LE TAL EN QUELQUES CHIFFRES


Le logiciel Analor

Le TAL en France c'est :
- 27 formations universitaires (source REPTIL) ;
- 8 métiers types : ergolinguiste, chef de projet, consultant avant-vente, linguiste informaticien, ingénieur assurance qualité, terminologue, veilleur stratégique, ingénieur support (source TECHNOLANGUE) ;
- 5 secteurs d'activité : moteurs de recherche intelligents, gestion de contenu, applications vocales et multimodales, e-learning et traduction automatique (source TECHNOLANGUE) ;
- Une place dans le trio des leaders Européen sur le marché du TAL (source Bureau van Dijk Information Management 2006) ;
- de 82 à 150 sociétés (sources APIL 2006 et ministère de la Culture 2007) ;
- 500 professionels de haut niveau en ingénierie linguistique appartenant tant au secteur de la recherche publique qu'au secteur industriel (ministère de la Culture 2007).

Traduction automatique
Communication homme-machine
Extraction d'information
Analyse de données
Représentation des connaissances

ATALA
Association pour le Traitement Automatique des Langues
www.atala.org

Traitement Automatique des Langues

Adhésion

http ://www.atala.org/-Adhesion-

20 €	Etudiants / Chômeurs	Présentation d'un justificatif
40 €	Individuels	Moyen de paiement personnel
60 €	Professionnels	Moyen de paiement professionnel
200 €	Institutionnels	Adhésion de soutien