

# Introduction au Traitement Automatique des Langues Naturelles (TALN)

**Philippe Langlais**

`felipe@iro.umontreal.ca`

**October 9, 2012**

# Plan

**TALN & RALI**

**Exemples d'applications du TALN**

**Plan (approximatif)**

**Repères historiques**

**Pipeline du TALN**

**Les mains dans le camboui**

# TALN & RALI

Exemples d'applications du TALN

Plan (approximatif)

Repères historiques

Pipeline du TALN

Les mains dans le camboui

# Traitement Automatique des Langues Naturelles (TALN)

- ▶ Activité multidisciplinaire à caractère applicatif (et éventuellement explicatif) regroupant linguistes, cogniciens, mathématiciens et informaticiens.
- ▶ Natural Language Processing aims at "making computers talk" and more precisely, at endowing them with the linguistic ability of humans (d'après (Gardent 2007)).
- ▶ Buts de ce cours:
  - ▶ problématiques de base du TALN
  - ▶ techniques de bases pour les résoudre
  - ▶ techniques de base de l'intelligence artificielle: apprentissage modèles génératifs ou discriminant, recherche de solutions, techniques d'alignement, etc.

# Tour d'horizon du RALI

<http://rali.iro.umontreal.ca>

- ▶ 3 profs:
  - ▶ Guy Lapalme: question-réponse, résumé, etc.
  - ▶ Jian-Yun Nie: recherche d'information, etc.
  - ▶ Philippe Langlais: traduction, etc.
- ▶ 1 assistant de recherche: Fabrizio Gotti
- ▶  $i \in [0, 5]$  chercheurs invités
- ▶  $m \in [3, 15]$  étudiants à la maîtrise
- ▶  $d \in [3, 10]$  étudiants au doctorat
- ▶  $p \in [0, 5]$  post-doctorants
- ▶ Des séminaires RALI-OLST tous les mercredis à 11h30  
<http://rali.iro.umontreal.ca/rali/?q=fr/node/1222>

# TALN & RALI

## Exemples d'applications du TALN

Plan (approximatif)

Repères historiques

Pipeline du TALN

Les mains dans le camboui

# Quelques exemples d'applications langagières

## Gestion/traitement de documents

- ▶ Moteurs de recherche (google, yahoo!, DuckDuckGo, etc.)
- ▶ Résumé automatique de textes, indexation automatique
- ▶ Extraction d'information
- ▶ Classification de textes (spams, opinions)
- ▶ Stylo-métrie, extraction terminologique, veille technologique
- ▶ Traduction automatique (MÉTÉO, Babel Fish:  
<http://world.altavista.com/tr>, Google Translate:  
<http://translate.google.fr/>)
- ▶ Notation automatique de copies d'étudiants (le rêve de tout prof)

# Quelques exemples d'applications langagières

## Production de documents

- ▶ Réponse automatique
  - ▶ aux questions “ouvertes”  
(ex: Ask Jeeves: <http://www.ask.com>)
  - ▶ aux courriels
- ▶ Aide à la rédaction (correcteurs, accélération de la saisie, etc.)



# Quelques exemples d'applications langagières

## Interfaces

- ▶ Traduction assistée (<http://www.tsrali.com>),
- ▶ Aide à la lecture, à l'apprentissage d'une langue seconde
- ▶ Interrogation de bases de données en langue naturelle
- ▶ Agents de dialogue:
  - ▶ Reconnaissance de la parole (How May I help you ?  
<http://www.research.att.com/~algor/hmihy/>)
  - ▶ Synthèse de la parole (MBROLA:  
<http://tcts.fpms.ac.be/synthesis/mbrola.html>)

# Applications développées au RALI

## Identification de la langue



Recherche appliquée en linguistique informatique

english

Recherche

Technologies

Demos

Séminaires

Partenaires

rali

[istes](#) d'expérience dans le traitement automatique de la langue. Il est le plus important labo

**SILC** (*Système d'Identification de la Langue et du Codage*) détermine automatiquement la langue dans laquelle un document est écrit, de même que le jeu de caractères employé. L'actuelle version reconnaît près d'une trentaine de langues, et une moyenne de trois encodages par langue.

Il est maintenant possible d'intégrer SILC dans votre application sur l'une des plateformes suivantes: [Windows](#), [Linux](#), [Solaris](#), [MAC](#), [HP-UX](#), [AIX](#) et [SGI](#). SILC existe également en version [Java](#).

La liste des langues et des encodages connus

Tapez du texte dans l'espace ci-dessous, dans la langue de votre choix. Notez que le système a besoin d'au moins quelques mots pour effectuer une identification fiable

Anglais cp1252 Chinois utf8 Japonais utf8 Espagnol cp1252 Allemand cp1252 Coréen utf8 Français cp1252 Italien cp1252 Portuguais cp1252 Néerlandais cp1252	Jag talar inte bra.
--	---------------------

Soumettre

texte

Choisir le fichier

aucun fichier sélectionné

Analyser

Afficher détails

# Applications développées au RALI

## Identification de la langue



Recherche appliquée en linguistique informatique

english

Recherche

Technologies

Demos

Séminaires

Partenaires

rali

[listes](#) d'expérience dans le traitement automatique de la langue. Il est le plus important lab

**SILC** (*Système d'Identification de la Langue et du Codage*) détermine automatiquement la langue dans laquelle un document est écrit, de même que le jeu de caractères employé. L'actuelle version reconnaît près d'une trentaine de langues, et une moyenne de trois encodages par langue.

Il est maintenant possible d'intégrer SILC dans votre application sur l'une des plateformes suivantes: [Windows](#), [Linux](#), [Solaris](#), [MAC](#), [HP-UX](#), [AIX](#) et [SGI](#). SILC existe également en version [Java](#).

La liste des langues et des encodages connus

Tapez du texte dans l'espace ci-dessous, dans la langue de votre choix. Notez que le système a besoin d'au moins quelques mots pour effectuer une identification fiable

Suédois cp1252 Suédois cp850 Suédois macintosh Suédois utf8 Thaï tis620 Thaï utf8 Turc cp853 Turc iso-8859-9 Turc utf8 Chinois big5	Jag talar inte bra.
--	---------------------

Soumettre  Choisir le fichier aucun fichier sélectionné

Le langue est Suédois, l'encodage est cp1252

Analyser

Afficher détails

# Applications développées au RALI

## Identification de la langue



Recherche appliquée en linguistique informatique

english

Recherche

Technologies

Demos

Séminaires

Partenaires

rali

[istes](#) d'expérience dans le traitement automatique de la langue. Il est le plus important la

**SILC** (*Système d'Identification de la Langue et du Codage*) détermine automatiquement la langue dans laquelle un document est écrit, de même que le jeu de caractères employé. L'actuelle version reconnaît près d'une trentaine de langues, et une moyenne de trois encodages par langue.

Il est maintenant possible d'intégrer SILC dans votre application sur l'une des plateformes suivantes: [Windows](#), [Linux](#), [Solaris](#), [MAC](#), [HP-UX](#), [AIX](#) et [SGI](#). SILC existe également en version [Java](#).

La liste des langues et des encodages connus

Tapez du texte dans l'espace ci-dessous, dans la langue de votre choix. Notez que le système a besoin d'au moins quelques mots pour effectuer une identification fiable

Malais macintosh Malais utf8 Néerlandais cp850 Néerlandais macintosh Néerlandais utf8 Norvégien cp1252 Norvégien cp850 Norvégien macintosh Norvégien utf8 Polonais cp1250	manao ahoana ianao
--	--------------------

Soumettre

texte

Choisir le fichier

aucun fichier sélectionné

Le langue est Malais, l'encodage est macintosh

Analyser

Afficher détails

# Applications développées au RALI

Lexicum: 207 M. de mots



base de données  
textuelles

Université  
de Montréal



Requête

Recherche

Aide

## Corpus

- Assemblée nationale et commissions
- Éditions Leméac, 1991-1993
- Interface
- La Presse
- Presse canadienne-française
- Université de Montréal

## Critères de sélection

Nombre de résultats

50  100  200  500

Longueur d'un contexte (caractères)

50  100  200

Références bibliographiques

---

Webmestre

# Applications développées au RALI

Lexicum: 207 M. de mots



base de données  
textuelles

## Assemblée nationale et commissions

2 résultat(s) pour la requête «république .. Madagascar»



Nouvelle recherche

Éditions Leméac, 1991-1993>>

**Présence du ministre du Développement du secteur privé et de la Privatisation de la république de Madagascar**, M. Simon Constant Horace J'ai également le plaisir de souligner la présence dans nos tribunes de M

1

*Les travaux parlementaires 36e législature, 2e session Journal des débats DÉBATS DE L'ASSEMBLÉE NATIONALE Le mardi 6 novembre 2001*

e M. Simon Constant Horace, ministre du Développement du secteur privé et de la Privatisation de la **république de Madagascar**. Affaires courantes Alors, nous abordons maintenant les affaires courantes. Il n'y a pas de déclara

2

*Les travaux parlementaires 36e législature, 2e session Journal des débats DÉBATS DE L'ASSEMBLÉE NATIONALE Le mardi 6 novembre 2001*

Éditions Leméac, 1991-1993>>

# Applications développées au RALI

## Réacc



Recherche appliquée en linguistique informatique

english

Recherche

Technologies

Demos

Séminaires

Partenaires

rali

Le RALI réunit des [informaticiens et des linguistes](#) d'expérience dans le traitement automatique de la langue. Il est le plus grand laboratoire dans le domaine au Canada.

**Réacc est un système capable de réintroduire automatiquement les accents et autres marques diacritiques dans un texte qui en est privé.**

Tapez dans l'espace ci-dessous du texte en français, sans accents:

La ou le francais n'est pas accentue,  
il y a de la gene,  
mais quand le systeme m'accentue,  
je suis moins gene!

réaccentuer ce texte

Pour appliquer Réacc sur un fichier:  aucun fichier sélectionné

# Applications développées au RALI

Réacc



Recherche appliquée en linguistique informatique

english

Recherche

Technologies

Demos

Séminaires

Partenaires

rali

Le RALI réunit des [informaticiens et des linguistes](#) d'expérience dans le traitement automatique de la langue. Il est le plus important laboratoire dans le domaine au Canada.

## Entrée:

La ou le français n'est pas accentué,  
il y a de la gêne,  
mais quand le système m'accentue,  
je suis moins gêné!

## Sortie:

Là où le français n'est pas accentué,  
il y a de la gêne,  
mais quand le système m'accentue,  
je suis moins gêné!

[Soumettre une nouvelle requête](#)

Webmeister: (11/7/2005)



# Applications développées au RALI

TransSearch/TSRALI.com <http://transsearch.iro.umontreal.ca/>

- ▶ mis en service sur le web en 1996 sans publicité
- ▶ plus de 20 000 requêtes par mois en 2000
- ▶ profil des utilisateurs:
  - ▶ 51% traducteurs
  - ▶ 32% étudiants
  - ▶ 12% terminologistes et rédacteurs professionnels
- ▶ concepteur: Michel Simard

# Applications développées au RALI

TransSearch/TSRALI.com <http://transsearch.iro.umontreal.ca/>

- ▶ TransSearch est maintenant un service offert en ligne par abonnement: TSRALI.com (Terminotix Inc.)
  - ▶ ~ 1 500 abonnés
  - ▶ ~ 75 000 requêtes par mois
  
- ▶ Bitextes offerts:
  - ▶ **hansard** débats à la chambre des communes depuis 1986 (235 M. de mots)
  - ▶ **cours canadiennes** décisions de la Cour suprême du Canada, de la Cour fédérale et de la Cour canadienne de l'impôt (88 M. de mots)

# Applications développées au RALI

TSRALI.com

## TransSearch

[TERMINOTIX](#)

[RALI](#)

Utilisateur : felipe

[Requêtes](#) | [Mon compte](#) | [Préférences](#) | [Aide](#) | [Quitter](#)

Collection de documents :

**Expression :**

Signet [TransSearch](#)  
([qu'est-ce que c'est?](#))

[Requête bilingue](#)

Soumettez un mot ou une expression, en français ou en anglais : TransSearch cherchera des contextes où cette expression apparaît, de même que le contexte correspondant dans l'autre langue.

*[Pour un service plus rapide, veuillez communiquer avec le webmestre](#)*

Copyright © 2001, 2003. Université de Montréal.  
Tous droits réservés.

# Applications développées au RALI

TSRALI.com

TransSearch

[TERMINOTIX](#)

[RALI](#)

Utilisateur : felipe

[Requêtes](#) | [Mon compte](#) | [Préférences](#) | [Aide](#) | [Quitter](#)

Signet [TransSearch](#)  
(qu'est-ce que c'est?)

Collection de documents :

Expression :

[Requête bilingue](#)

Chercher

- |   |   |  |
|---|---|--|
| 1 | Toutefois, lorsqu'ils ont remporté les élections, c'était alors une toute autre <b>paire de manches</b> .   | However, when they won the elections, it was a different kettle of fish.   |
| 2 | Quant à savoir si la décision vise les publications que la Chambre imprime au nom des membres, c'est une autre <b>paire de manches</b> .  | Whether it applies to any publications that the House may print on behalf of members is another matter.  |
| 3 | La qualité de la gestion est une autre <b>paire de manches</b> , notamment en ce qui concerne la morue de l'Atlantique nord et les problèmes survenus dans la région du fleuve Fraser l'an dernier. | Whether it was well managed or not is another question when one considers the problem with the North Atlantic cod, not to mention the problems on the Fraser River over the last year. |
| 4 | Cependant, comparer cela à la question de savoir si la définition traditionnelle du mariage devrait être maintenue, c'est une toute autre <b>paire de manches</b> .                                 | However, to compare that to the issue of whether the traditional definition of marriage should be maintained is something all together different.                                      |

## Contexte

Expression: **paire de manches**

Collection de documents : **Hansard canadien (1986-2005)**

Document: **hans.2005.0067.fr**

Résultat no. 2

[Nouvelle requête](#)

[Retour aux résultats](#)

[Page précédente](#)

**Des voix: Bravo!**

Le Président: Je signale également à la Chambre la présence de 12 représentants des Forces canadiennes qui sont ici pour prendre part aux activités de la Journée annuelle des Forces canadiennes.

La Journée des Forces canadiennes est l'occasion, pour l'ensemble des Canadiens, de prendre conscience des sacrifices que font pour eux les hommes et les femmes des forces armées.

**Des voix: Bravo!**

**Some hon. members: Hear, hear!**

The Speaker: I am also pleased to draw to the attention of the House the presence of 12 representative members of the Canadian Forces here to take part in annual Canadian Forces Day events.

Canadian Forces Day is an opportunity for Canadians from across the country to recognize the sacrifices that our men and women in uniform make on our behalf.

**Some hon. members: Hear, hear!**

Collection de documents :

Chercher

Expression :

Requête bilingue

français/anglais

1

These changes are taking place in the context of political and economic reforms, as well as an increasing decentralization of the **health services**.

Esos cambios se presentan dentro de un marco de reformas políticas y económicas, y de mayor descentralización de los servicios de salud.

2

It will identify the areas and the population groups that need specific poli- cies, sustained intervention programs, and **health services**.

Permitirá también identificar las áreas y los grupos de población que necesitan políticas específicas, programas de intervención sostenible y servicios de salud.

3

Recommendations issued at this meeting will serve as a platform for developing guidelines and indicators to monitor the impact of macrodeterminants that govern the public health situation and the access, utilization, and financing of **health services**.

Basándose en las recomendaciones de esa reunión, se podrán elaborar pautas e indicadores para monitorear el impacto de los macrodeterminantes de la situación sanitaria y del acceso, la utilización y el financiamiento de los servicios de salud.

Collection de documents :

Expression anglaise :

Expression française :

①

Le Québec se souvient et salue son indéfectible **attachement** à la société québécoise.

Quebec remembers and salutes his unwavering **commitment** to Quebec society.

②

Par le passé, le Canada a appliqué des consignes en matière d'immigration qui contredisaient notre **attachement** commun envers la justice humaine.

In the past Canada enforced some immigration practices that were at odds with our shared **commitment** to human justice.




③

Mme Jean Crowder: Madame la Présidente, laissant de côté les questions commerciales, je dirai que le projet de loi C-39 constitue certes, mais partiellement, un pas dans la bonne direction en réaffirmant notre **attachement** à un régime d'assurance-maladie public au Canada.

Ms. Jean Crowder: Madam Speaker, leaving the trade issues aside, Bill C-39 in part certainly is moving in the right direction in terms of reaffirming our **commitment** to a public health care system in Canada.

# TSRALI.com nouvelle mouture (TS3)

<http://isasli.iro.umontreal.ca:8080/Main.aspx>

**TRANSEARCH<sup>3</sup> BETA**    **TERMINOTIX**            

UTILISATEUR : felipe    REQUÊTES    MON COMPTE    PRÉFÉRENCES    AIDE    QUITTER

Signet / Favori personnalisé : **TransSearch** (ou'est-ce que c'est ?)    Requête bilingue

Collection de documents : Les Hansards canadiens

Expression : paire de manches    Chercher

**46 traductions de *paire de manches* dans 74 occurrences**

different kettle of fish	8	different kettle of fish	8
matter	6	Là, nous avons une nouvelle <b>paire de manches</b> , car si vous êtes conservateurs, vous êtes contre ce genre de dépenses.	This is a <b>different kettle of fish</b> , because a conservative generally opposes this kind of spending.
different story	5	C'était une autre <b>paire de manches</b> .	It was a very <b>different kettle of fish</b> .
different issue	4	S'ils ne font pas confiance aux juges, c'est une autre <b>paire de manches</b> .	If the members opposite do not trust judges, that is a <b>different kettle of fish</b> .
different	2	Toutefois, lorsqu'ils ont remporté les élections, c'était alors une toute autre <b>paire de manches</b> .	However, when they won the elections, it was a <b>different kettle of fish</b> .
entirely	2	C'est une autre <b>paire de manches</b> .	It is a <b>different kettle of fish</b> .
issue	2	La période des questions, c'est une autre <b>paire de manches</b> .	Question period is a <b>different kettle of fish</b> .
thing	2	Si le député de Delta-South Richmond n'est pas satisfait de la réponse à la question qu'il a présentée, c'est une toute autre <b>paire de manches</b> .	If the hon. member for Delta-South Richmond takes exception to the response to the question that he submitted, that is an entirely <b>different kettle of fish</b> .
ball game	2	Si mon collègue prétend que M. Yeutter veut redresser la balance commerciale de son pays en recourant à des pratiques commerciales déloyales, c'est une autre <b>paire de manches</b> .	Surely if my hon. friend is suggesting that Mr. Yeutter wants to change the trade balances using unfair trading practices, that is a <b>different kettle of fish</b> .
different thing	2		
question	2		
of a problem with	2		
story	2		
little different	1		
kettle of fish at the moment	1		
different issue for some	1		
horse of a different colour	1		
thing altogether	1		
different matter altogether	1		
different ballgame	1		
altogether	1		
quite	1		
solving	1		
kettle of fish	1		
really matter	1		

concepteur: Fabrizio Gotti



# Applications développées au RALI

## TransType

Product overview

The machine is controlled from a liquid crystal color touch screen where you can view your operation and application settings.

Printer settings can be pre-programmed for specific production job types and when such a job type is selected, the printer is set up automatically for the paper type and application.

Aperçu de la machine:

La machine est contrôlée à partir d'un écran tactile à cristaux liquides vous permettant de visualiser vos paramètres d'application et d'opération.

Les paramètres de l'imprimante peuvent être préprogrammés pour des types de type de travail et lorsqu'il est sélectionné

rs de la  
rsqu'il est sélectionné,  
rsque le le  
rsque le,

# NLP ∈ IA: NELL

<http://rtw.ml.cmu.edu/rtw>













## tea (beverage)

literal strings: Tea, tea, TEA

---

### Help NELL Learn!

NELL wants to know if these beliefs are correct.  
If they are or ever were, click thumbs-up. Otherwise, click thumbs-down.

- tea is an agricultural product  
- tea is a beverage  
- tea is an agricultural product produced in japan (country)  
- tea is an agricultural product produced in kenya (country)  
- tea is an agricultural product produced in south vietnam (country)  
- tea is an agricultural product that contains antioxidants (chemical)  

<http://rtw.ml.cmu.edu/rtw/kbbrowser/beverage:tea>

- ▶ apprentissage continu (15M de faits candidats, ~ 1.5M fiables)
- ▶ intervention manuelle minimaliste

# TALN & RALI

## Exemples d'applications du TALN

### **Plan (approximatif)**

#### Repères historiques

#### Pipeline du TALN

#### Les mains dans le camboui

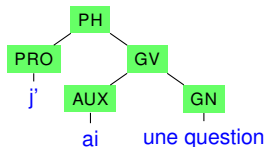
# Sujets abordés

- ▶ Modèles de langue *p(colorless idea sleep furiously) ?*
  - ▶ survol de quelques techniques de lissage
  - ▶ algorithme E(xpectation)M(aximization)
- ▶ Algorithmique du texte *mots proches de a2m1?*
  - ▶ programmation dynamique: edit-distance, etc.
- ▶ Étiquetage morpho-syntaxique *Je parle trop* → PRN VB  
ADV
  - ▶ modèles de Markov
  - ▶ approche transformationnelle
- ▶ Apprentissage analogique *[⊙ : ⊙ :: ⊗ : ?]*
  - ▶ problèmes
  - ▶ réalisations

# Sujets abordés

## ▶ Grammaires

- ▶ éléments de théorie des langues
- ▶ grammaires probabilistes



## ▶ Traduction automatique

*Elle l'aime* → *She loves*<sup>1</sup>

- ▶ Approche(s) par mémoire
- ▶ Approche(s) statistique(s)
- ▶ Limitations

## ▶ Sémantique de corpus

*You must ?? your microwavable*

*popcorn before eating it*

- ▶ test de vraisemblance, information mutuelle
- ▶ extraction de collocations
- ▶ Open I(nformation)E(xtraction)

## ▶ Vos présentations

<sup>1</sup>google – <http://translate.google.com/> – 1 octobre 2012

# TALN & RALI

## Exemples d'applications du TALN

### Plan (approximatif)

### **Repères historiques**

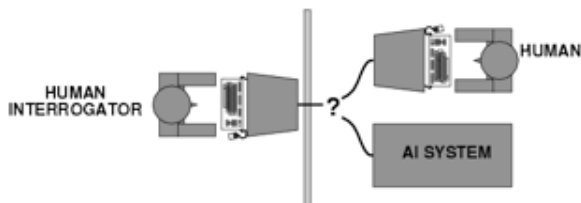
### Pipeline du TALN

### Les mains dans le camboui

# Quelques repères historiques

- ▶ Démarrage à l'après-guerre, avec l'essor de l'informatique et le développement de la *théorie des langages formels* (Turing, Kleene, Chomsky, Backus, Naur) et de la *théorie de l'information* (Shannon)
- ▶ 1950: le test de Turing
- ▶ 1952: premier système de RAP (statistique) capable de reconnaître les chiffres prononcés par un locuteur (Bell Labs)
- ▶ 1954: premier "système" de traduction (250 mots, 6 règles, russe→anglais)
- ▶ 1960: premier corpus électronique: *Le Brown corpus*, 1 million de mots, différents genres, etc.
- ▶ 1960: **the box is in the pen // the pen is in the box** (Bar-Hillel)
- ▶ 1966: rapport Alpac
- ▶ 1966: le système ELIZA (Weizenbaum 66)
- ▶ 1968: le premier (vrai) système de traduction (Systran, russe→anglais)
- ▶ 1971: un système intelligent en mode fermé (SHRDLU)
- ▶ 1976: le système de traduction MÉTÉO mis au point à l'UdeM
- ▶ 80s: système de reconnaissance statistique multilocuteur

# Le test de Turing (1950)



- ▶ Succès si la machine trompe un testeur dans 30% des cas sur une période de 5 minutes
- ▶ Turing pensait qu'en 2000 les machines passeraient le test



# Le système ELIZA (Weizenbaum 66)

- ▶ Développé par Weizenbaum en 1966 (MIT)
- ▶ My goodness, this computer is clever !

Utilisateur	Men are all alike
ELIZA	IN WHAT WAY ?
Utilisateur	They're always bugging us about something or other
ELIZA	CAN YOU THINK OF A SPECIFIC EXAMPLE ?
Utilisateur	My boyfriend says I'm depressed all the time
ELIZA	I AM SORRY TO HEAR YOU ARE DEPRESSED

*dialogue extrait de (Jurafsky and Martin 2000)*

# Le système ELIZA (Weizenbaum 66)

## La réalité ...

```
1 s/my/YOUR/g
2 s/I'm/YOU ARE/g
3 ...
4 s/. * all .*/IN WHAT WAY ?/1
5 s/. *I am (depressed).*/I AM SORRY TO HEAR YOU ARE \1/1
6 s/. * always .*/CAN YOU THINK OF A SPECIFIC EXAMPLE/1
7 ...
```

- ▶ version web: <http://www-ai.ijs.si/eliza/eliza.html>
- ▶ version emacs, tapez: ESC X doctor
- ▶ voir aussi Loebner Prize  
<http://www.loebner.net/Prizef/loebner-prize.html>  
et ALICEBot.

# Le système SHRDLU

- ▶ Un monde restreint (scène de blocs) où un humain peut:
  - ▶ demander à l'ordinateur de manipuler des blocs
  - ▶ interroger le système à propos d'une configuration
  - ▶ demander à la machine d'exprimer son raisonnement

```
Person:      PICK UP A BIG RED BLOCK.
Computer:    OK. (does it)
Person:      GRASP THE PYRAMID.
Computer:    I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
Person:      FIND A BLOCK WHICH IS TALLER THAN THE ONE
              YOU ARE HOLDING AND PUT IT INTO THE BOX.
Computer:    BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS
              TALLER THAN THE ONE I AM HOLDING.
Computer:    OK. (does it)
Person:      WHAT DOES THE BOX CONTAIN?
Computer:    THE BLUE PYRAMID AND THE BLUE BLOCK.
```

- ▶ développé par Terry Winograd (M.I.T.)
- ▶ <http://hci.stanford.edu/~winograd/shrdlu/>

## Deux approches majeures

- ▶ Dominance de l'approche **rationaliste** de la fin des années 50 au début des années 80, sous l'influence principale de Chomsky
  - ▶ **Idée maîtresse:** l'être humain naît avec une compétence linguistique
- ▶ L'approche **empiriste** ne reprendra ses lettres de noblesses qu'au début des années 80, grâce aux efforts simultanés d'IBM (Jelinek et al.) et de CMU (Baker et al.) qui introduisent l'approche canal bruité/HMMs en RAP.
  - ▶ **Idée maîtresse:** l'être humain est doté de compétences (*think positive*), mais d'une nature différente: reconnaissance de formes, déduction, généralisation, etc.

# Les arguments Chomskiens

(d'après Abney dans (Klavans and Resnik 1996))

- ▶ Ces deux phrases ont la même probabilité d'être observées dans un corpus, à savoir, faible<sup>2</sup>.
  - ▶ colorless green ideas sleep furiously
  - ▶ furiously sleep ideas green colorless
- ▶ L'approximation markovienne d'ordre  $n$  sera toujours mise en défaut:
  - ▶ Chomsky:  $\nexists n, \varepsilon : \forall s, \text{grammatical}(s) \leftrightarrow P_n(s) > \varepsilon$
  - ▶ Shannon:  $\exists \varepsilon : \forall s, \text{grammatical}(s) \leftrightarrow \lim_{n \rightarrow \infty} P_n(s) > \varepsilon$
- ▶ “We cannot seriously propose that a child learns the values of  $10^9$  parameters in a childhood lasting only  $10^8$  seconds”.

---

<sup>2</sup>[http://en.wikipedia.org/wiki/Colorless\\_green\\_ideas\\_sleep\\_furiously](http://en.wikipedia.org/wiki/Colorless_green_ideas_sleep_furiously)

# Les arguments Chomskiens

- ▶ La réponse de Peter Norvig:  
<http://norvig.com/chomsky.html>
- ▶ Plus à ce sujet:  
<http://languagelog.ldc.upenn.edu/nll/?p=3172>

# TALN & RALI

## Exemples d'applications du TALN

### Plan (approximatif)

### Repères historiques

## **Pipeline du TALN**

### Les mains dans le camboui

# Différents niveaux de traitement

D'après (Yvon 2007)

- ▶ segmenter le texte en **unités lexicales** (mots)
- ▶ identifier les composants lexicaux, leur propriétés: **traitement lexical**
- ▶ identifier les syntagmes: **analyse syntaxique**
- ▶ construire une représentation du sens: **analyse sémantique**
- ▶ identifier les fonctions de l'énoncé dans son contexte d'élocution/de production: **analyse pragmatique**

Note:

- ▶ ambiguïté à tous les niveaux
- ▶ illusion de l'approche pipeline



# Segmenter un texte en mots

- ▶ Les **séparateurs** ne sont pas exempts d'ambiguïté:
  - ▶ le point peut indiquer la partie décimale d'un nombre (1.23), un acronyme (C.R.D.P.), peut faire partie d'une abréviation (M. Paul)
  - ▶ les guillemets (') introduisent une citation, mais sont aussi présents dans des noms propres (O'Sullivan), dans certaines unités (Il a couru le 100 mètre en 9'78). On en retrouve aussi dans aujourd'hui ou prud'hommes
  - ▶ le trait-d'union peut indiquer la présence d'une incise, est également présent dans les mots composés et sert aussi à marquer les césures.
  - ▶ etc.
- ▶ Prolifération de nouvelles formes de l'écrit:  
Oui! C mon demi frere ki a pris le msg. A ya dit on retourne ouskon pratiquait avant

# Analyse lexicale

- ▶ **But:** associer les *tokens* aux entrées d'un lexique qui caractérise les mots d'une langue, contient leurs propriétés

le	det. masc. sing / pron. pers. masc. sing.
président	verb. 3 pers. plu. ind.-subj. / nom masc. sing.
...	

- ▶ Encoder un lexique avec les informations pertinentes est une activité coûteuse

## Traitement morphologique

- ▶ analyse flexionnelle: processus d'ajustement des formes conditionné par des contraintes d'ordre syntaxiques

Ex:

- ▶ Le pluriel d'un nom se forme en français par ajout d'un **s**
    - ▶ Le futur se marque par la présence d'un **r** et d'une conjugaison spécifique
  
  - ▶ analyse dérivationnelle: processus de création de nouvelles formes à partir de formes existantes
- Ex: **briser** → **brisure**

# Analyse lexicale

## Traitement des mots composés

- ▶ ouvre-bouteille, pomme de terre
- ▶ **adverbes**: en effet, de temps à autre
- ▶ **conjonctions**: parce que, si bien que
- ▶ **collocations**: au fur et à mesure, prendre le taureau par les cornes
- ▶ **termes**:
  - ▶ réseaux de neurones,
  - ▶ réseaux neuromimétiques,
  - ▶ réseau neuronal

# Analyse syntaxique

- ▶ ambiguïté lexicale: **la** = pron. / article / nom commun
- ▶ ambiguïté dynamique: **Il est vraiment chien**
- ▶ **sous-catégorisation** du verbe:

**a)** X parle (**Jean Parle**)

**b)** X parle à Y (**Jean Parle à Marie**)

**c)** X parle de Y (**Jean Parle de Paul**)

**d)** X parle de Y à Z (**Jean Parle de Paul à Marie**)

▶ **Je parle à la maîtresse de Marie: b) ou d) ?**

- ▶ ambiguïtés de rattachement:

▶ **Elle mange une glace à la fraise / elle mange une glace à la plage**

▶ **J'ai été voir un film avec Marilyn Monroe**

▶ **Il a parlé de déjeuner avec Paul**

▶ **Il voit l'homme avec un télescope**

# Analyse sémantique

- ▶ Faire correspondre les syntagmes à des concepts du monde réel.
- ▶ Souvent abordé à l'aide de la logique des prédicats ou du lambda calcul
  - ▶ Paul a mis le vin sur la table  
`mettre(Paul, Vin, sur(Vin,Table))`
- ▶ Une formule logique est souvent construite par composition en parcourant l'arbre syntaxique, mais:
  - ▶ Luc a avoué ce vol à Guy
  - ▶ Luc a attribué ce vol à Guy
  - ▶ Luc a décrit ce vol à Guy

ont des interprétations (formules logiques) très différentes

# Analyse pragmatique

- ▶ la pragmatique porte sur les attitudes que les locuteurs adoptent vis-à-vis d'un énoncé
  - ▶ Viendras-tu au bal ce soir ? J'ai entendu que Paul y sera !

# TALN & RALI

## Exemples d'applications du TALN

### Plan (approximatif)

### Repères historiques

### Pipeline du TALN

## Les mains dans le camboui



# Compter des mots dans un corpus

- ▶ Un mot quelconque dans un **corpus** est appelé une **occurrence** (ou une **instance**, ou très souvent encore un **token**). C'est la réalisation d'un **type** particulier.

Ca c' est pour moi , le plus beau et le plus triste paysage du monde . C' est le même paysage que celui de la page précédente , mais je l' ai dessiné une fois encore pour bien vous le montrer .

- ▶ Il y a 43 occurrences dans ce corpus (en comptant les signes de ponctuation), mais il y a seulement 34 types (en distinguant majuscule/minuscule; 33 sinon). 75% de ces types ont une **fréquence** de 1 dans ce corpus.

# Qu'est-ce qu'un mot ?

- ▶ Une réponse possible:

```
1 sed -e 's/^$/. /g' $1 | tr -s "." "." |
2 sed -e 's/No\. *\([0-9]\)/No \1/g'
3     -e 's/no\. *\([0-9]\)/no \1/g'
4     -e 's/* / * /g' -e 's/- / - /g'
5     -e 's/? / ? /g' -e 's"/ " /g'
6     -e 's/\. / \. /g' -e 's/, / , /g'
7     -e 's;/ ; /g' -e 's:/ : /g'
8     -e 's/\([^\?!\]\)/ \1 /g'
9     -e 's/\[/ \[ /g' -e 's/\]/ \] /g'
10    -e 's/( / ( /g' -e 's)/ / ) /g' -e "s/'/' /g"
11    -e 's/\([0-9][0-9]*\)\([a-zA-Z]\)/\1 \2/g' |
12    tr -s "[:space:]" "[\012*]"
```

- ▶ Pour une réponse plus circonstanciée, lire (Polguère 2008).

# Qu'est-ce qu'une phrase ?

Exemple extrait de (Véronis and Langlais 2000) p. 372

⟨S⟩ On disait dans le livre: “Les serpents boas avalent leur proie tout entière, sans la mâcher. Ensuite ils ne peuvent plus bouger et ils dorment pendant les six mois de leur digestion” ⟨/S⟩

---

⟨S⟩ On disait dans le livre: ⟨/S⟩⟨S⟩ “Les serpents boas avalent leur proie tout entière, sans la mâcher. Ensuite ils ne peuvent plus bouger et ils dorment pendant les six mois de leur digestion” ⟨/S⟩

---

⟨S⟩ On disait dans le livre: ⟨/S⟩ “⟨S⟩ Les serpents boas avalent leur proie tout entière, sans la mâcher. ⟨/S⟩ ⟨S⟩ Ensuite ils ne peuvent plus bouger et ils dorment pendant les six mois de leur digestion ⟨/S⟩

---

⟨S⟩ On disait dans le livre: “ ⟨S⟩ Les serpents boas avalent leur proie tout entière, sans la mâcher. Ensuite ils ne peuvent plus bouger et ils dorment pendant les six mois de leur digestion ⟨/S⟩ ” ⟨/S⟩

# Faut-il utiliser un “vrai” langage de programmation ?

- ▶ Langage de commande (shell scripts) :

```
1 cat corpus | sort | uniq -c | sort -k1,1n
```

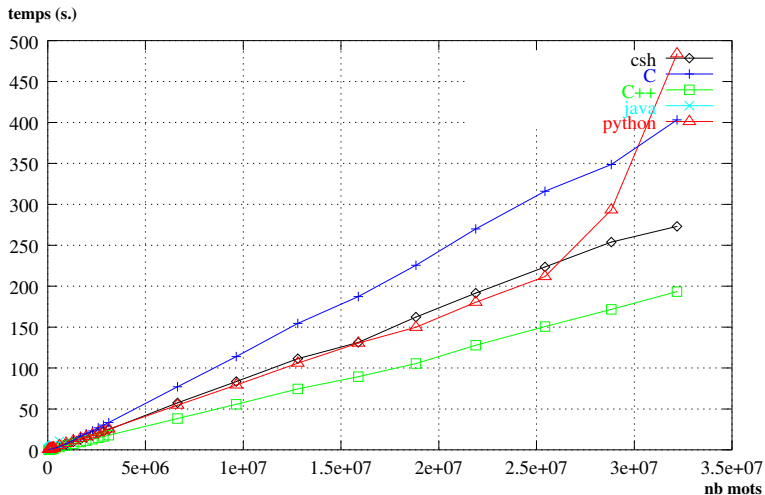
- ▶ C++ (ou autre langage du même type) :

```
1 cat corpus | frequence
```

Où `frequence` est un programme utilisant une structure de donnée de type *hash-map*. Voici à quoi ça peut ressembler avec la *STL*:

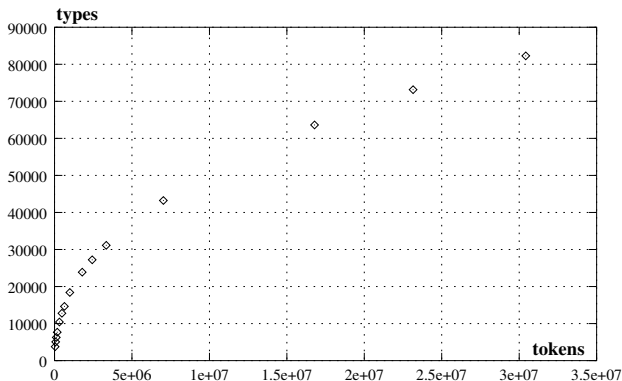
```
1 while (cin >> s) {
2     it = mots.find(s);
3     if (it == mots.end())
4         mots.insert(make_pair(s,1));
5     else
6         ++(it->second);
7 }
```

# Faut-il utiliser un “vrai” langage de programmation ?



# types vs tokens

30 M. de mots du Hansard



- ▶ 60000 entrées dans le Petit Robert; 75000 dans le grand Robert.
- ▶ Vocabulaire moyen d'un individu  $< 5000$  mots (environ)

# <http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html>

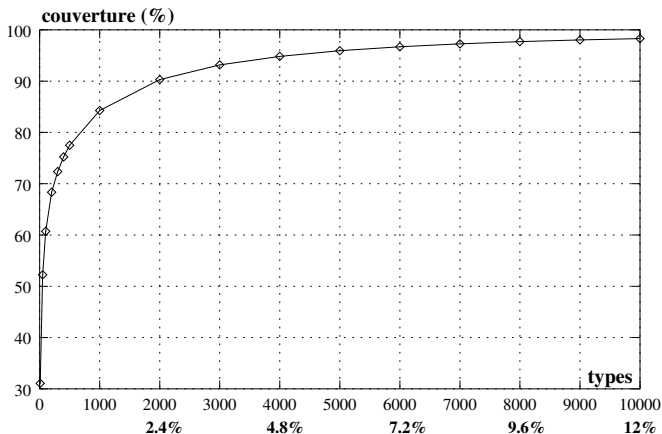
Beaucoup 30 M de mots ?

- ▶ Citation de google (mars 2006):

*“We processed 1,011,582,453,213 words of running text and are publishing the counts for all 1,146,580,664 five-word sequences that appear at least 40 times. There are 13,653,070 unique words, after discarding words that appear less than 200 times.”*

# Notion de couverture

30 M. de mots du **Hansard**



- ▶ 2.4% des types couvrent 90% du corpus étudié





# Loi de Zipf: $f \times r \approx cst$

mot	$r$	$f$	$f \times r$	mot	$r$	$f$	$f \times r$
the	1	1836714	1836714	done	200	16313	3262600
{sent}	2	1639250	3278500	children	300	10188	3056400
.	3	1383044	4149132	ensure	400	7880	3152000
is	10	504770	5047700	corporation	500	6414	3207000
not	20	221350	4427000	turner	1000	2915	2915000
à	30	133279	3998370	damage	2000	1204	2408000
?	40	98435	3937400	withdrawn	3000	658	1974000
all	50	81796	4089800	finances	4000	408	1632000
's	60	65562	3933720	neighbourhood	5000	282	1410000
those	70	53489	3744230	opposes	7000	153	1071000
his	80	46108	3688640	momentum	8000	117	936000
so	90	40810	3672900	forecasting	10000	73	730000
per	100	36099	3609900	rambled	50000	2	100000

- ▶  $f$  = fréquence,  $r$  = rang

# Références I

-  Gardent, Claire (2007). “Natural Language Processing Applications”. [Notes de cours](#).
-  Jurafsky, Daniel and James H. Martin (2000). *Speech and Language Processing*. Prentice Hall.
-  Klavans, J. L. and P. Resnik (1996). *The Balancing Act, Combining Symbolic and Statistical Approaches to Language*. MIT Press.
-  Polguère, Alain (2008). *Lexicologie et sémantique lexicale. Notions fondamentales*. Coll. “Champs Linguistiques”. Les Presses de l’Université de Montréal.
-  Véronis, J. and Ph. Langlais (2000). “Evaluation of parallel text alignment systems: The ARCADE project”. In: vol. 13. *Parallel Text Processing*, Kluwer. Chap. 19, pp. 369–388.
-  Weizenbaum, J. (66). “ELIZA, A computer program for the study of Natural Language Communication between man and Machine”. In: *ACM*. Vol. 9(1), pp. 36–45.
-  Yvon, François (2007). *Une petite introduction au Traitement Automatique des Langues Naturelles*. notes introductives d’un cours sur le traitement des langues naturelles. ParisTech.