

# INTRODUCTION AU TALN

Introduction au Traitement Automatique des Langues Naturelles

12 janvier 2014

Philippe Langlais

[felipe@iro.umontreal.ca](mailto:felipe@iro.umontreal.ca)

**RALI**

Dept. Informatique et Recherche Opérationnelle

Université de **Montréal**

1. TALN ?
2. RALI
3. Histoire
4. Pourquoi est-ce difficile ?
5. Composants typiques : modèle de langue
6. Composants typiques : modèle de traduction
7. Le mot de la fin



- *Le Traitement automatique du langage naturel ou de la langue naturelle (abr. TALN) ou des langues (abr. TAL) est une discipline à la frontière de la **linguistique**, de l'**informatique** et de l'**intelligence artificielle**, qui concerne l'application de programmes et techniques informatiques à tous les aspects du langage humain. Ainsi, le TAL ou TALN est parfois nommé **ingénierie linguistique**.*

[http://fr.wikipedia.org/wiki/Traitement\\_automatique\\_du\\_langage\\_naturel](http://fr.wikipedia.org/wiki/Traitement_automatique_du_langage_naturel)

- *Natural Language Processing aims at "making computers talk" and more precisely, at endowing them with the linguistic ability of humans*  
**[Gardent, 2007]**

- Moteurs de recherche (google, yahoo!, DuckDuckGo, etc.)
- Résumé automatique de textes, indexation automatique
- Extraction d'information
- Classification de textes (spams, opinions)
- Traduction automatique ou assistée
  - MÉTÉO
  - Babel Fish : <http://www.babelfish.fr>
  - Google Translate : <http://translate.google.fr/>
- Notation automatique de copies d'étudiants, détection de plagiat

- Réponse automatique
  - aux questions  
(ex : Ask Jeeves : <http://www.ask.com>)
  - aux courriels
- Aide à la rédaction (correcteurs, accélérateurs de saisie)
  - Dasher <http://www.inference.phy.cam.ac.uk/dasher/>
  - Swype <http://www.swype.com>
  - SwiftKey <http://www.swiftkey.net/en/>
- etc.



[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

<ul style="list-style-type: none"> <li>Anglais cp1252</li> <li>Chinois utf8</li> <li>Japonais utf8</li> <li>Espagnol cp1252</li> <li>Allemand cp1252</li> <li>Coréen utf8</li> <li>Français cp1252</li> <li>Italien cp1252</li> <li>Portuguais cp1252</li> <li>Néerlandais cp1252</li> </ul>	<input type="text" value="Jag talar inte bra."/>
--	--

Soumettre   aucun fichier sélectionné



[listes](#) 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

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

Soumettre   aucun fichier sélectionné

La langue est Suédois, l'encodage est cp1252

Analyser

Afficher détails





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: [Choisir le fichier](#) aucun fichier sélectionné



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 accentue,  
il y a de la gene,  
mais quand le systeme m'accentue,  
je suis moins gene!

#### Sortie:

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

[Soumettre une nouvelle requête](#)

- mis en service sur le web en 1996 sans publicité
- plus de 20 000 requêtes par mois en 2000
- TransSearch est maintenant un service offert en ligne par abonnement : TSRALI.com (Terminotix Inc.)
  - ~ 1 500 abonnés
  - ~ 75 000 requêtes par mois
- profil des utilisateurs :
  - 51% traducteurs
  - 32% étudiants
  - 12% terminologues et rédacteurs professionnels

## TransSearch

RALI

utilisateur: felipe

Requêtes | Mon compte | Préférences | Aide | Quitter

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

Collection de documents : 

Chercher

Expression anglaise : 

Requête simple

Expression française : 

anglais/espagnol

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.

TRANSEAR<sup>H3</sup> BETA TERMINO<sup>TIX</sup> rali

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	<b>different kettle of fish</b>	8
matter	6		
different story	5	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 issue	4		
different	2	C'était une autre <b>paire de manches</b> .	It was a very <b>different kettle of fish</b> .
entirely	2		
issue	2	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> .
thing	2		
ball game	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> .
different thing	2		
question	2	C'est une autre <b>paire de manches</b> .	It is a <b>different kettle of fish</b> .
of a problem with	2		
story	2	La période des questions, c'est une autre <b>paire de manches</b> .	Question period is a <b>different kettle of fish</b> .
little different	1		
kettle of fish at the moment	1	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> .
different issue for some	1		
horse of a different colour	1		
thing altogether	1	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 matter altogether	1		
different ballgame altogether	1		
quite	1		
solving	1		
kettle of fish	1		
really matter	1		

.\samples\predictor\_offW2\_4.txt-RALI on

File Edit Go Options About

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,

## 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







recherche appliquée en  
linguistique informatique

Recherche

English

ACCUEIL

PROJETS

RESSOURCES

DÉMOS

PUBLICATIONS

ENSEIGNEMENT

SÉMINAIRES


CONTACT ET INFOS

## Bienvenue au RALI

### Recherche appliquée en linguistique informatique

Le RALI réunit des **informaticiens et des linguistes** d'expérience dans le traitement automatique de la langue. Il est un des plus importants laboratoires universitaires dans le domaine au Canada.





Le RALI réunit des **informaticiens et des linguistes** d'expérience dans le traitement automatique de la langue. Il est un des plus importants laboratoires universitaires dans le domaine au Canada.

## Traduction automatique »

Traduction basée sur des méthodes statistiques ou sur l'apprentissage analogique

Traduction interactive ou assistée par ordinateur

## Résumé automatique »

Résumé par abstraction

Résumés de textes juridiques avec des collaborateurs industriels

## Recherche d'information »

Intégration de la sémantique en recherche d'information

Recherche translinguistique

## Informations environnementales »

Analyse, personnalisation et traduction d'informations produites quotidiennement par Environnement Canada



## Guy Lapalme

- Résumé de textes
- Génération de textes



## Jian-Yun Nie

- Recherche d'information
- Fouille de données



## Philippe Langlais

- Traduction
- Terminologie, Morphologie

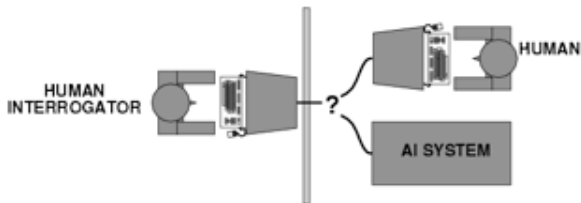
- 2 professeurs associés :
  - Caroline Barrière
  - Atefeh Farzindar
  
- 1 assistant de recherche : Fabrizio Gotti
  
- géométrie variable :
  - $i \in [0, 5]$  chercheurs invités
  - $m \in [3, 15]$  étudiants à la maîtrise
  - $d \in [3, 10]$  étudiants au doctorat
  - $p \in [0, 5]$  post-doctorants
  
- séminaires hebdomadaires conjoints avec l'**OLST**  
<http://rali.iro.umontreal.ca/rali/?q=fr/node/1222>

HISTOIRE

# 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.
- 1966 : rapport Alpac (la traduction ne marchera jamais)
- 1966 : le système ELIZA
- 1968 : le premier (vrai) système de traduction (Systran, russe→anglais)
- 1976 : le système de traduction MÉTÉO mis au point à l'UdeM
- 80s : système de reconnaissance de parole statistique multilocuteur
- 2000s : **web sémantique** (Tim Berners Lee)

## 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

- Développé par Weizenbaum en 1966 (MIT)
- 

1	Utilisateur	Men are all alike
2	ELIZA	IN WHAT WAY ?
3	Utilisateur	They're always bugging
4		us about something or
5		other
6	ELIZA	CAN YOU THINK OF A
7		SPECIFIC EXAMPLE ?
8	Utilisateur	My boyfriend says I'm
9		depressed all the time
10	ELIZA	I AM SORRY TO HEAR YOU
11		ARE DEPRESSED

---

*dialogue extrait de [Jurafsky and Martin, 2000]*



---

```
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.manifestation.com/neurotoys/eliza.php3>
- version emacs, tapez : ESC X doctor
- Loebner Prize <http://www.loebner.net/Prizef/loebner-prize.html>
- ALICEBot. <http://alice.pandorabots.com>

- 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 générales de reconnaissance de formes, de déduction, de généralisation, etc.

- Ces deux phrases ont la même probabilité d'être observées dans un corpus, à savoir, faible<sup>1</sup>.
  - colorless green ideas sleep furiously
  - furiously sleep ideas green colorless
  
- L'approximation markovienne d'ordre  $n$  sera toujours mise en défaut :
  - Chomsky :  $\nexists n, \epsilon : \forall s, \text{grammatical}(s) \leftrightarrow P_n(s) > \epsilon$
  - Shannon :  $\exists \epsilon : \forall s, \text{grammatical}(s) \leftrightarrow \lim_{n \rightarrow \infty} P_n(s) > \epsilon$

---

1. [http://en.wikipedia.org/wiki/Colorless\\_green\\_ideas\\_sleep\\_furiously](http://en.wikipedia.org/wiki/Colorless_green_ideas_sleep_furiously)



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



POURQUOI EST-CE DIFFICILE ?

- segmenter le texte en **unités** (sujets, phrases, mots)
- identifier les composants lexicaux, leur propriétés : **analyse lexicale**
- identifier les syntagmes : **analyse syntaxique**
- construire une représentation du sens : **analyse sémantique**
- identifier les fonctions de l'énoncé dans son contexte de production : **analyse pragmatique**

Note :

- ambiguïté à tous les niveaux

- le point peut :
  - indiquer la partie décimale d'un nombre (1.23)
  - indiquer un acronyme (C.R.D.P.)
  - faire partie d'une abréviation (M. Paul)
- on rencontre les guillemets (simples) dans :
  - des noms propres (O'Sullivan)
  - des unités (Il a couru le 100 mètre en 9'78)
  - des mots (aujourd'hui ou prud'hommes)
- le trait-d'union :
  - marque les incises (Cette personne - par ailleurs charmante - a toute mon estime)
  - est présent dans des mots composés (un aller-retour)
  - marque les césures (con-↔sistant)
- etc.

## ○ Pas si simple ...

ความแตกต่างระหว่างโรงเรียนเก่า (พ.ศ. 2420 - พ.ศ. 2427) กับ โรงเรียนใหม่ (พ.ศ. 2428 เป็นต้นมา) ของคุณพ่อกลอมเบตก็คือ โรงเรียนใหม่แห่งนี้มิได้เป็นโรงเรียนวัดที่มุ่งสอนเฉพาะเด็กคาทอลิกอีกต่อไป หากแต่เป็นโรงเรียนที่เปิดกว้างสำหรับนักเรียนทุกเชื้อชาติ ศาสนา ซึ่งถือเป็นการเปลี่ยนแปลงด้านหลักการที่สำคัญยิ่งอันมีผล เปลี่ยนแปลงทางปฏิบัติคือ ทำให้โรงเรียนของคุณพ่อกลอมเบตมิได้เป็นโรงเรียนที่ให้การศึกษาแก่บุคคลเฉพาะกลุ่มอีกต่อไป เมื่อพิจารณาช่วงเวลาทีโรงเรียนแห่งนี้เปิดสอนคือ พ.ศ. 2428 จะเห็นว่าเป็นช่วงเวลาเดียวกันกับที่รัฐกำลังจัดตั้งโรงเรียนหลวงสำหรับราษฎร ขึ้นตามวัดโดยมีวัตถุประสงค์ให้ราษฎรทั่วไปได้มีโอกาสศึกษาเล่าเรียนตามแบบหลวงที่ได้จัดให้แก่พระบรมวงศานุวงศ์และบุตรหลานข้าราชการมาก่อนแล้ว นโยบายดังกล่าวของรัฐส่งผลให้การจัดการศึกษาของรัฐเปิดกว้างออกสู่คนทุกกลุ่มในสังคม อนึ่งการจัดการ ศึกษาแก่ราษฎรนี้เป็นการใหม่ที่เริ่มขึ้นจึงน่าที่จะขาดความพร้อมหลายประการ อาทิ ครูผู้สอน งบประมาณ และสถานที่ โรงเรียนหลวงที่เปิดตามวัดต่างๆจึงทยอยเปิดทีละโรง ทั้งนี้ยังไม่คำนึงถึงความยากลำบากในการชักชวนโน้มน้าวให้คนเห็นประโยชน์ส่งบุตรหลานเข้ามาเรียน เมื่อโรงเรียนอัสสัมชัญ หรือ อาชมชาน

<http://th.wikipedia.org/wiki/>

## ○ 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



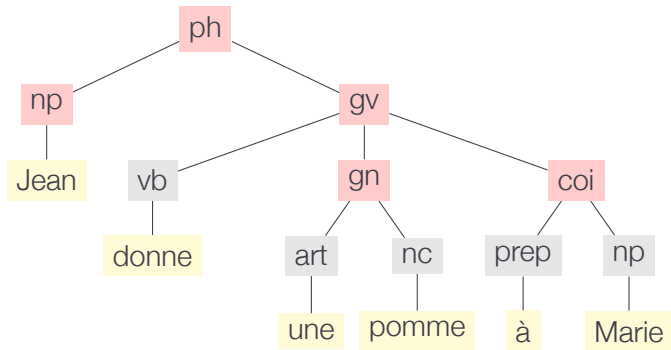
- **But :** associer les **tokens** aux entrées d'un lexique qui caractérise les mots d'une langue, contient leurs propriétés

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

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

- **flexionnelle** : variation de la forme des unités lexicales en fonction de facteurs grammaticaux
  - Le pluriel d'un nom se forme en français par ajout d'un *s* (*chien* → *chiens*)
  - Le futur se marque par la présence d'un *r* et d'une conjugaison spécifique (*parler* → *parlerai*)
  
- **dérivationnelle** : formation d'unités lexicales nouvelles à partir de matériel morphologique existant
  - *briser* → *brisure*
  
- *tieotokoneongelma* ?
  - *tietokone* ≡ ordinateur (*tieto* ≡ information + *kone* ≡ machine)
  - *ongelma* ≡ problem

- 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



- ambiguïté lexicale : la = pronom / article / nom
- 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 voit l'homme avec un télescope
  - Il a parlé de déjeuner avec Paul

- 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

- la pragmatique a pour objet d'étude l'énoncé (une ou plusieurs phrases dans un contexte énonciatif)
  - Viendras-tu au bal ce soir? J'ai entendu que Paul y sera !
    - oui! (j'adore Paul)
    - pas question si Paul y va!
  - Veux-tu un gateau? Je suis au régime.
    - Non il/elle ne veut pas de gateau
  - Ils vont encore augmenter nos taxes !
    - Ils ⇒ le gouvernement ?

# COMPOSANTS TYPIQUES : MODÈLE DE LANGUE



- Un modèle de langue probabiliste spécifie une distribution  $p(s)$  sur les chaînes  $s$  de la langue modélisée :

$$\sum_s Pr(s) = 1$$

- Sans perte d'information, si l'on considère que  $s$  est une séquence de  $N$  mots (phrase ?),  $s \equiv w_1 \dots w_N$ , alors :

$$Pr(s) \stackrel{def}{=} \prod_{i=1}^N Pr(w_i | \underbrace{w_1 \dots w_{i-1}}_h)$$

où  $h$  est appelé l'**historique**

**But :** associer un texte  $T$  à une classe parmi  $\mathcal{C}$   
(ex :  $\mathcal{C} = \{\text{sport, religion, spam, ...}\}$ ).

$$\begin{aligned}\hat{c} &= \operatorname{argmax}_{c \in \mathcal{C}} p(c|T) \\ &= \operatorname{argmax}_{c \in \mathcal{C}} \frac{p(T|c) \times p(c)}{p(T)} \\ &= \operatorname{argmax}_{c \in \mathcal{C}} \underbrace{p(T|c)}_{\text{langue}} \times \underbrace{p(c)}_{\text{a priori}}\end{aligned}$$

- un modèle unigramme donne des performances (étonnamment) raisonnables.

But : trouver les documents  $D$  pertinents à une requête  $R$

$$\begin{aligned}\hat{D} &= \operatorname{argmax}_D p(D|R) \\ &= \operatorname{argmax}_D \underbrace{p(D)}_{\text{a priori}} \times \underbrace{P(R|D)}_{\text{langue}}\end{aligned}$$

- Un modèle de langue par document dans la **collection** !
- Mode opératoire de base :

$$\operatorname{argmax}_D \prod_{i=1}^{|R|} \lambda p(R_i|D) + (1 - \lambda)p(R_i|Collection)$$

**But :** trouver  $\hat{l}$  la langue de  $T$  parmi un ensemble de langues  $\mathcal{L}$

- Soit  $T \equiv T_1^N \equiv t_1, \dots, t_N$  un texte de  $N$  caractères

$$\begin{aligned} \hat{l} &= \operatorname{argmax}_{l \in \mathcal{L}} p(l|T) \\ &\approx \operatorname{argmax}_{l \in \mathcal{L}} \underbrace{\prod_{c=1}^N p(T_c | T_{c-n+1}^{c-1}, l)}_{\text{n-car}} \times \underbrace{p(l)}_{\text{a priori}} \end{aligned}$$

- un modèle n-gramme par langue

## Mon char est parké au garage

français	cp1252	0.099
allemand	cp1252	0.064
français	cp850	0.036
français	macintosh	0.036

## Je parke mon char

néerlandais	cp1252	0.046
néerlandais	cp850	0.046
néerlandais	macintosh	0.046
anglais	cp1252	0.046
allemand	cp1252	0.038

**But :**

Le systeme m'accentue → Le système m'accentue.  
Le systeme m'a accentue → Le système m'a accentué.

- Soit  $w_1^n$  la phrase de  $n$  mots à réaccéntuer
- Implantation possible :
  - sélectionner pour tout mot  $w_i$  ses versions accentuées possibles, soit  $a(w_i)$  cet ensemble.
    - Ex :  $a(\text{coté}) \equiv \{\text{cote}, \text{côté}, \text{côte}\}$
  - considérer toutes les phrases que l'on peut construire à partir des  $a(w)$  (en prenant un mot par  $a(w)$ ) et sélectionner celle de plus forte probabilité selon le modèle de langue
  - [rali.iro.umontreal.ca](http://rali.iro.umontreal.ca)

But : Trouver  $\hat{T}$  la traduction d'une phrase source  $S$

- approche proposée par une équipe d'IBM au début des années 90 :

$$\begin{aligned}\hat{T} &= \operatorname{argmax}_T P(T|S) \\ &= \operatorname{argmax}_T \underbrace{P(S|T)}_{\text{traduction}} \times \underbrace{P(T)}_{\text{langue}}\end{aligned}$$

- 2 distributions que l'on "sait" estimer
- problème de recherche de maximum ( $\operatorname{argmax}$ ) non trivial

$Pr(\text{John aime Marie qui aime Paul}) =$

$Pr(\text{John} \mid \text{BOS}) \times$

$Pr(\text{aime} \mid \text{BOS John}) \times$

$Pr(\text{Marie} \mid \text{BOS John aime}) \times$

$Pr(\text{qui} \mid \text{BOS John aime Marie}) \times$

$Pr(\text{aime} \mid \text{BOS John aime Marie qui}) \times$

$Pr(\text{Paul} \mid \text{BOS John aime Marie qui aime})$

- approximation d'ordre  $n - 1$ , le modèle ***n*-gramme** :

$$p(s = w_1^n) \approx \prod_{i=1}^N p(w_i | w_{i-n+1}^{i-1})$$



$$p(s) = \prod_{i=1}^N p(w_i | w_{i-2} w_{i-1})$$

$Pr(\mathbf{John\ aime\ Marie\ qui\ aime\ Paul}) =$

$Pr(\text{John} \mid \text{BOS BOS}) \quad \times$

$Pr(\text{aime} \mid \text{BOS John}) \quad \times$

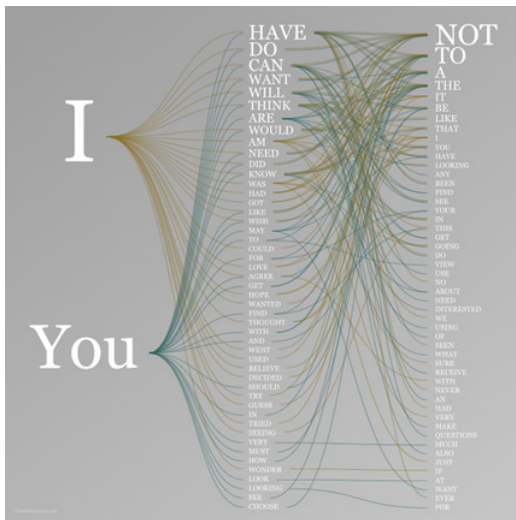
$Pr(\text{Marie} \mid \text{John aime}) \quad \times$

$Pr(\text{qui} \mid \text{aime Marie}) \quad \times$

$Pr(\text{aime} \mid \text{Marie qui}) \quad \times$

$Pr(\text{Paul} \mid \text{qui aime})$

# CAS DU MODÈLE TRIGRAMME



<http://www.chrisharrison.net/index.php/Visualizations>

Soit  $\mathcal{D} \equiv w_1 \dots w_N$ , un **corpus** (texte) de  $N$  mots

- Cas de l'unigramme  $p(s) = \prod_i p(w_i)$  :

$$p(w) = \frac{|w|}{N}, \text{ avec } |w| \text{ la fréquence de } w \text{ dans } \mathcal{D}$$

- Cas du bigramme  $p(s) = \prod_i p(w_i|w_{i-1})$  :

$$p(w_i|w_{i-1}) = \frac{|w_{i-1}w_i|}{|w_{i-1}|} = \frac{|w_{i-1}w_i|}{\sum_w |w_{i-1}w|}$$

- Cas du  $n$ -gramme :

$$p(w_i|w_{i-n+1}^{i-1}) = \frac{|w_{i-n+1}^{i-1}w_i|}{\sum_w |w_{i-n+1}^{i-1}w|}$$

**Mini corpus :**

Vincent aime Virginie

Estelle aime les fleurs

Elle aime les fleurs jaunes plus particulièrement

$$\circ p(\text{Vincent aime les fleurs}) = \frac{1}{3} \times 1 \times \frac{2}{3} \times 1 \times \frac{1}{2} = \frac{1}{9} \approx 0.111$$

$$\begin{aligned} & p(\text{Vincent|BOS}) && |\text{BOS Vincent}|/|\text{BOS}| = 1/3 \\ \times & p(\text{aime|Vincent}) && |\text{Vincent aime}|/\text{Vincent} = 1/1 = 1 \\ \times & p(\text{les|aime}) && |\text{aime les }|/|\text{aime}| = 2/3 \\ \times & p(\text{fleurs|les}) && |\text{les fleurs}|/|\text{les}| = 2/2 = 1 \\ \times & p(\text{EOS|fleurs}) && |\text{fleurs EOS}|/|\text{fleurs}| = 1/2 \end{aligned}$$

$$\circ p(\text{Virginie aime les fleurs}) = 0 \text{ car } |\text{BOS Virginie}| = 0$$

- Le corpus **Austen**<sup>2</sup> contient 8 762 phrases, 620 968 mots (14 274 différents).
- En théorie, il existe :

$$n_{\text{bigrammes}} = 14\,274 \times 14\,274 = 203\,747\,076$$

$$n_{\text{trigrammes}} \approx 2.9 \times 10^{12}$$

- On observe :
  - 194 211 bigrammes  $\neq$  (soit  $\approx 0.09\%$ ), dont 69% **d'hapax legomena**
  - 462 615 trigrammes  $\neq$  (soit  $\approx 10^{-5}\%$ ), dont 87% d'hapax

⇒ **Lissage** des probabilités, *i.e* donner une probabilité à des choses non vues à l'**entraînement**.

---

2. <http://www.iro.umontreal.ca/~felipe/CampInfo-2014/>

### 3-gramme (caractère)

48 hey unto the searede upon meered them woreacolaked the  
facon of the the evered of I puts sayinsweireas wout though.  
Forsaccon whim the ausethe, I saing ame the ribe the land to  
Jer of th ; ang G

### 4-gramme (caractère)

38 And of of hou ign from his the in trusand the outh he down  
inted nown the place it. When spriest the thou car nor ear the  
pring it of Babylon, the othe Jesus ; the friestone spassel we sin.  
Tappear

<http://johnno.jsmf.net/knowhow/ngrams/>

## 2-gramme (mot)

horse and the famine shall come thither. Therefore was in the ground, and said, Abraham, and there forty years old ? and according to go out thy sight, hear me. And Leah and upon Egypt, and he shall be thine only in which ye eat of his name of the two daughters. And Joseph made sure unto Ephron. And it will be my master.

<http://johnno.jsmf.net/knowhow/ngrams/>

# COMPOSANTS TYPIQUES : MODÈLE DE TRADUCTION



- $f$  une phrase source (ex : français)
- $e$  une phrase cible (ex : anglais)
- traduire  $\Leftrightarrow$  résoudre :

$$\operatorname{argmax}_e P(e|f) = \operatorname{argmax}_e P(f|e)P(e)$$

Deux modèles (au moins)

- $p(f|e)$  modèle de transfert
- $p(e)$  modèle de langue

Un décodeur

- NP-complet **[Knight, 1999]**

## ○ un **corpus parallèle**

<ul style="list-style-type: none"><li>● The Legislative Assembly convened at 3.30 pm.</li><li>● Mr. Quirke (Clerk-Designate) :</li><li>● THURSDAY, APRIL 1, 1999</li></ul>	<ul style="list-style-type: none"><li>● sitamiq, ipuru 1, 1999</li><li>● maligaliurvik matuiqtau-lauqtuq 3 :30mi unnusakkut</li><li>● mista kuak (titiraqti - tik-kuaqtausimajuq) :</li></ul>
--	---

- un **corpus parallèle** + un *aligneur de phrases* = **bitexte**

<ul style="list-style-type: none"><li>● The Legislative Assembly convened at <b>3.30</b> pm.</li><li>● Mr. Quirke (Clerk-Designate) :</li><li>● THURSDAY, APRIL <b>1</b>, <b>1999</b></li></ul>	<ul style="list-style-type: none"><li>● sitamiq, ipuru <b>1</b>, <b>1999</b></li><li>● maligaliurvik matuiqtau-lauqtuq <b>3 :30</b>mi unnusakkut</li><li>● mista kuak (titiraqti - tik-kuaqtausimajuq) :</li></ul>
---	--

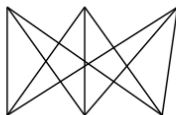
- un **corpus parallèle** + un *aligneur de phrases* = **bitexte**

<ul style="list-style-type: none"><li>● The Legislative Assembly convened at <b>3.30</b> pm.</li><li>● Mr. Quirke (Clerk-Designate) :</li><li>● THURSDAY, APRIL 1, 1999</li></ul>	<ul style="list-style-type: none"><li>● sitamiq, ipuru <b>1</b>, <b>1999</b></li><li>● maligaliurvik matuiqtalauqtuq 3 :30mi unnusakkut</li><li>● mista kuak (titiraqti - tik-kuaqtausimajuq) :</li></ul>
---	---

- des aligneurs (de phrases) de bonne qualité
- des corpus parallèles pour plusieurs paires de langues (débats parlementaires, best sellers, documentations techniques, etc.)

- $P(f|e) = \sum_a P(a, f|e)$   
où  $a$  est un *alignement* entre  $e$  and  $f$
- plusieurs modèles de la jointe  $P(a, f|e)$
- une procédure d'estimation itérative [**Brown et al., 1993**]

... la maison ... la maison blue ... la fleur ...

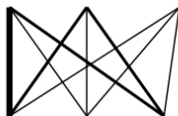


... the house ... the blue house ... the flower ...

alignements équiprobables

- $P(f|e) = \sum_a P(a, f|e)$   
où  $a$  est un *alignement* entre  $e$  and  $f$
- plusieurs modèles de la jointe  $P(a, f|e)$
- une procédure d'estimation itérative [**Brown et al., 1993**]

... la maison ... la maison blue ... la fleur ...

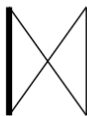
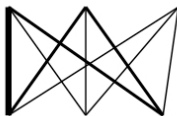


... the house ... the blue house ... the flower ...

la/the, maison/house émergent ...

- $P(f|e) = \sum_a P(a, f|e)$   
où  $a$  est un *alignement* entre  $e$  and  $f$
- plusieurs modèles de la jointe  $P(a, f|e)$
- une procédure d'estimation itérative [**Brown et al., 1993**]

... la maison ... la maison blue ... la fleur ...



... the house ... the blue house ... the flower ...

la/the, maison/house se renforcent

- $P(f|e) = \sum_a P(a, f|e)$   
où  $a$  est un *alignement* entre  $e$  and  $f$
- plusieurs modèles de la jointe  $P(a, f|e)$
- une procédure d'estimation itérative [**Brown et al., 1993**]

... la maison ... la maison bleu ... la fleur ...



... the house ... the blue house ... the flower ...

bleue/blue, fleur/flower apparaît



## phrase based model

```

have not been democratically |||
    n' ont été démocratiquement ||| 0.5 9.81223e-08
particularly in baby products . |||
    en particulier dans les produits pour bébés . ||| 0.5 4.17267e-12
there is no cruelty |||
    il n' est nullement question de cruauté ||| 1 3.15409e-10
intergovernmental conference and reform of the treaty |||
    cig et la révision du traité ||| 1 1.03844e-17
is an absolute must . |||
    c' est un nécessité absolue . ||| 1 1.35065e-12
parliament has already done a considerable amount |||
    le parlement a accompli un travail énorme ||| 1 8.38686e-21
presidency 's programme |||
    du programme de la présidence ||| 1 9.33702e-07

```

- des millions de paramètres pour un bitexte d'un million de phrases

# THIS BEAUTIFUL PLANT IS UNIQUE

1

1

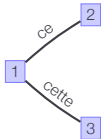
## transfer table

this	↔	ce
	↔	cette
beautiful	↔	belle
	↔	beau
plant	↔	plante
	↔	usine
is	↔	est
unique	↔	seule
	↔	unique
beautiful plant		
↓		
belle plante		
plante magnifique		

## language model

ce beau plante	:-( cette belle usine	:-  belle usine est	:-) ...
----------------	--------------------------	------------------------	------------

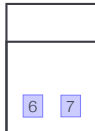
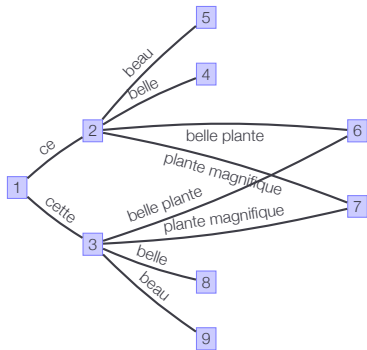
# THIS BEAUTIFUL PLANT IS UNIQUE



transfer table		
this	↔	ce
	↔	cette
beautiful	↔	belle
	↔	beau
plant	↔	plante
	↔	usine
is	↔	est
unique	↔	seule
	↔	unique
beautiful plant ↓ belle plante plante magnifique		

language model			
ce beau plante	:-( cette belle usine	:-  belle usine est	:-) ...

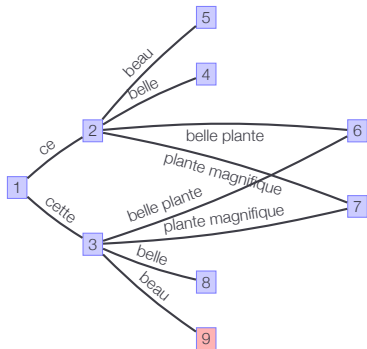
# THIS BEAUTIFUL PLANT IS UNIQUE



transfer table		
this	↔	ce
	↔	cette
beautiful	↔	belle
	↔	beau
plant	↔	plante
	↔	usine
is	↔	est
unique	↔	seule
	↔	unique
beautiful plant ↓ belle plante plante magnifique		

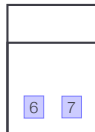
language model	
ce beau plante	:-
cette belle usine	:-
belle usine est	:-)
...	

# THIS BEAUTIFUL PLANT IS UNIQUE

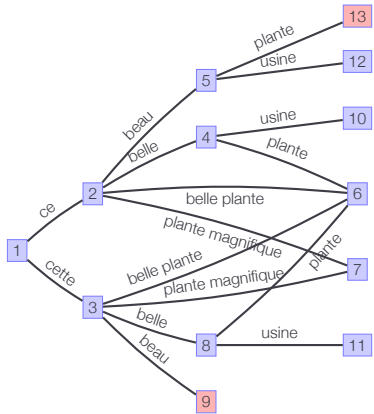


transfer table		
this	↔	ce
	↔	cette
beautiful	↔	belle
	↔	beau
plant	↔	plante
	↔	usine
is	↔	est
unique	↔	seule
	↔	unique
beautiful plant ↓ belle plante plante magnifique		

language model	
ce beau plante	:-
cette belle usine	:-
belle usine est	:-)
...	

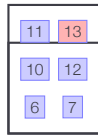
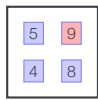
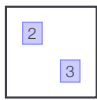


# THIS BEAUTIFUL PLANT IS UNIQUE

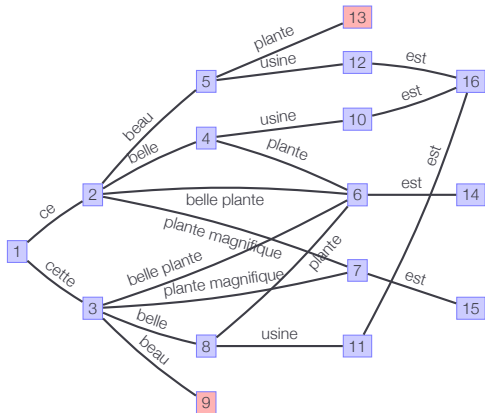


transfer table	
this	↔ ce
	↔ cette
beautiful	↔ belle
	↔ beau
plant	↔ plante
	↔ usine
is	↔ est
unique	↔ seule
	↔ unique
beautiful plant	
↓	
belle plante	
plante magnifique	

language model	
ce beau plante	:-
cette belle usine	:-
belle usine est	:-)
...	

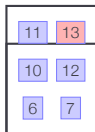


# THIS BEAUTIFUL PLANT IS UNIQUE

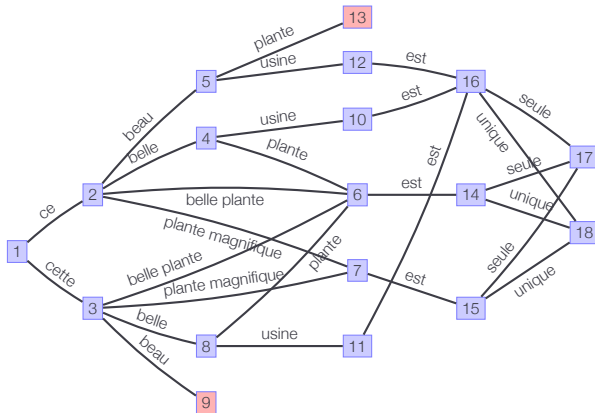


transfer table	
this	↔ ce
	↔ cette
beautiful	↔ belle
	↔ beau
plant	↔ plante
	↔ usine
is	↔ est
unique	↔ seule
	↔ unique
beautiful plant	
↓	
belle plante	
plante magnifique	

language model	
ce beau plante	:-
cette belle usine	:-
belle usine est	:-)
...	

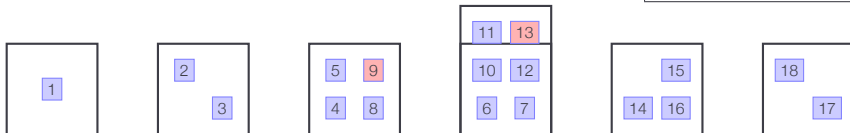


# THIS BEAUTIFUL PLANT IS UNIQUE



transfer table	
this	↔ ce
	↔ cette
beautiful	↔ belle
	↔ beau
plant	↔ plante
	↔ usine
is	↔ est
unique	↔ seule
	↔ unique
beautiful plant ↓ belle plante plante magnifique	

language model	
ce beau plante	:-
cette belle usine	:-
belle usine est	:-)
...	





## source

- Barack Obama becomes the fourth American president to receive the Nobel Peace Prize
- The American president Barack Obama will fly into Oslo, Norway for 26 hours to receive the Nobel Peace Prize, the fourth American president in history to do so.

## traduction





- Barack Obama se convierte en el cuarto presidente estadounidense para recibir el Premio Nobel de la Paz
- El presidente estadounidense Barack Obama va a volar en Oslo, Noruega, por 26 horas para recibir el Premio Nobel de la Paz, el cuarto presidente estadounidense en la historia de hacerlo





# POURQUOI FAIRE DU TALN ?

- Parce que c'est amusant !
  - activité relativement jeune
  - votre contribution peut faire la différence !
- Parce que c'est utile !
  - isolement des personnes âgées
  - aide à la communication
- Parce que ça fera plaisir à vos futurs employeurs !



-  Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R. L. (1993).  
The mathematics of statistical machine translation :  
Parameter estimation.  
*Computational Linguistics*, 19(2) :263--311.
-  Gardent, C. (2007).  
Natural language processing applications.  
Notes de cours.
-  Jurafsky, D. and Martin, J. H. (2000).  
*Speech and Language Processing*.  
Prentice Hall.
-  Knight, K. (1999).  
Decoding complexity in word-replacement translation  
models.  
*Computational Linguistics*, 25(4).

-  Smith, J. R., Saint-amand, H., Koehn, P., Callison-burch, C., Plamada, M., and Lopez, A. (2013).  
Dirt cheap web-scale parallel text from the common crawl.  
In *51st ACL*, pages 1374--1383, Sofia, Bulgaria.
-  Weizenbaum, J. (66).  
Eliza, a computer program for the study of natural language  
communication between man and machine.  
In *ACM*, volume 9(1), pages 36--45.