

Introduction au Traitement Automatique des Langues Naturelles (TALN)

felipe@iro.umontreal.ca

RALI

Dept. Informatique et Recherche Opérationnelle
Université de **Montréal**



V1.0

Last compiled: 9 septembre 2018



Plan

Contexte

Quelques Applications

Plan (approximatif)

Repères historiques

Approche pipeline

Les mains dans le cambouis

Plan

Contexte

Quelques Applications

Plan (approximatif)

Repères historiques

Approche pipeline

Les mains dans le cambouis

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]*).
- ▶ But de ce cours :
 - ▶ problèmes et applications de base du TALN
 - ▶ panorama de techniques pour les résoudre (vocabulaire)
 - ▶ difficultés / limites actuelles
 - ▶ manipulation d'outils et de ressources

Le RALI

- ▶ 3 profs :
 - ▶ Guy Lapalme : question-réponse, résumé, etc.
 - ▶ Jian-Yun Nie : recherche d'information, etc.
 - ▶ Philippe Langlais : traduction, extraction d'information, etc.
- ▶ 2 profs associés :
 - ▶ Caroline Barière, CRIM
 - ▶ Atefeh Farzindar, NLP Technologies
- ▶ $a \in [1, 5]$ assistants de recherche
- ▶ $i \in [0, 5]$ 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>



Plan

Contexte

Quelques Applications

Plan (approximatif)

Repères historiques

Approche pipeline

Les mains dans le cambouis

Exemples d'applications langagières

- ▶ Moteurs de recherche (google, yahoo !, DuckDuckGo, etc.)



pianistes comme tigran hamasyan



Tous

Vidéos

Images

Actualités

Shopping

Plus

Paramètres

Outils

Environ 19 200 résultats (0,28 secondes)

Tigran Hamasyan — Wikipédia

https://fr.wikipedia.org/wiki/Tigran_Hamasyan ▼

Tigran Hamasyan est un pianiste de jazz arménien né le 17 juillet 1987 à Gyumri (Arménie). ... Le jeune musicien fait alors la connaissance de légendes comme Wayne Shorter, Herbie Hancock, John McLaughlin ou Joe Zawinul et de ...

[Biographie](#) · [Récompenses](#) · [Discographie](#) · [En tant que leader](#)

INTERVIEW. Tigran Hamasyan, pianiste contemplatif - Culturebox

<https://culturebox.francetvinfo.fr> › Musique › Jazz / Blues ▼

24 mars 2017 - Le 31 mars, le pianiste arménien Tigran Hamasyan a sorti un nouvel album ... Comme il aime à le faire dans ses disques, Tigran Hamasyan a ...

Tigran Hamasyan, pianiste de jazz virtuose - Culturebox

<https://culturebox.francetvinfo.fr> › Musique › Jazz / Blues ▼

29 mars 2010 - A 22 ans, le pianiste de jazz Tigran Hamasyan est un prodige qui a déjà ... récompensé dans les plus grands festivals de jazz, comme celui de ...

Vidéos



Le surprenant clip du pianiste Tigran

L'Express - 12 juin 2013



Tigran Hamasyan "The Spinners"
@Jazz_in_Marciac 8
Août 2011

Jazz In Marciac
YouTube - 11 août 2011



Tigran Hamasyan piano solo 2011

Piano Web
YouTube - 20 août 2011

Exemples d'applications langagières

- ▶ Moteurs de recherche (google, yahoo!, DuckDuckGo, etc.)
- ▶ Traduction automatique (ou assistée)
 - ▶ Babel Fish : <http://world.altavista.com/tr>,
 - ▶ Google Translate : <http://translate.google.fr/>,
 - ▶ MÉTÉO : <http://rali.iro.umontreal.ca/EnvironmentalInfo/WarningTranslation.html>

The screenshot shows the Google Translate web interface. At the top left is the Google logo. On the top right, there is a grid icon and a blue button labeled "Connexion". Below the logo, the word "Traduction" is displayed in red. To the right of "Traduction" is the text "Désactiver la traduction instantanée" and a star icon. The main interface has two input fields. The left field contains the French text: "Les réseaux de neurones ont révolutionné le monde de la traduction automatique en rendant les traductions proches de celles qu'un traducteur professionnel produirait." Below this text are a speaker icon, a menu icon, and the character count "168/5000". The right field contains the English translation: "Neural networks have revolutionized the world of machine translation by making translations close to those that a professional translator would produce." Below this text are a star icon, a copy icon, a speaker icon, and a back arrow icon. At the bottom right of the right field is a link that says "Suggérer une modification" with a pencil icon. The language selection menu at the top shows "Anglais" selected for the target language and "Français" selected for the source language. A blue "Traduire" button is also visible.

Exemples d'applications langagières

- ▶ Moteurs de recherche (google, yahoo!, DuckDuckGo, etc.)
- ▶ Traduction automatique (ou assistée)
 - ▶ Babel Fish : <http://world.altavista.com/tr>,
 - ▶ Google Translate : <http://translate.google.fr/>,
 - ▶ MÉTÉO : <http://rali.iro.umontreal.ca/EnvironmentalInfo/WarningTranslation.html>
- ▶ Résumé automatique de textes, indexation automatique

https://resoomer.com

RESOOMER

Service

Comment ça marche

PREMIUM

Contact

Connexion

Blog

FR

Tigran Hamasyan commence à s'intéresser au piano dès l'âge de deux ans. À trois ans, il chante les chansons de Led Zeppelin, Deep Purple, les Beatles, Louis Armstrong ou encore Queen en s'accompagnant au piano. À sept ans, il découvre le monde du jazz et passe ses journées à écouter différentes mélodies et à improviser au piano. Il poursuit alors son éducation musicale classique à l'école.

En 1997, quand sa famille déménage à Erevan, il étudie Duke Ellington, Thelonious Monk, Charlie Parker, Art Tatum, Miles Davis, Bud Powell. À cette même période, il met au point ses premières compositions. L'année suivante, sa participation au premier festival de jazz d'Erevan lui permet de se faire remarquer et de se faire inviter pour de prochains concerts et sessions. Lors du second festival de jazz d'Erevan en 2000, alors âgé de 13 ans, il attire l'attention de Chick Corea, Avishai Cohen, Jeff Ballard ou encore Ari Roland. Il rencontre le pianiste Stéphane Kochoyan qui va l'aider à se faire connaître en Europe. En 2001, ce dernier l'invite à plusieurs festivals en France. Le jeune musicien fait alors la connaissance de légendes comme Wayne Shorter, Herbie Hancock, John McLaughlin ou Joe Zawinul et de musiciens comme Danilo Perez et John Patitucci. En 2003 et 2004, il participe au Festival de Jazz de Serres, dans les Hautes-Alpes, où il revient en 2009 pour un duo avec Fanny Azzuro dans le cadre de Jazz & Classique.

Grâce à son premier prix de piano-jazz emporté en 2005 au Thelonious Monk Institute of Jazz, il entre à l'Université de Californie du Sud à Los Angeles où il commence à étudier en profondeur et en parallèle le jazz contemporain et la musique arménienne. La même année il publie son deuxième album, New Era, accompagné de François Moutin et Louis Moutin, avec l'apparition de Vardan Grigoryan au duduk. Il s'installe à New York en 2008.

En 2009, il enregistre Red Hail, un album au carrefour du jazz, du métal et du folklore arménien, avec son nouveau quintet de jeunes musiciens Aratta Rebirth3 : Areni Agbalian (voc), Ben Wendel (ts), Charles Altura (g), Sam Minaia (b) et Nate Wood (d). Ils se produisent dans plusieurs grands festivals internationaux, de Montréal à Nice en passant par Vienne ou Rotterdam (North Sea Jazz Festival).

En juin 2010, Tigran Hamasyan signe avec le label Verve. Il enregistre en septembre 2010 à Paris l'album solo à 4 voix, pour lequel il est lauréat des Victoires du Jazz 2011 dans la catégorie album international de production française.

Il se produit en 2011 dans de grands festivals comme Jazz in Marciac, Montreux, Montréal pour la 3^e année consécutive, ainsi qu'au Tokyo Jazz Festival, en Arménie, au Royaume-Uni (Queen Elizabeth Hall) ou encore en Allemagne.

Il enregistre en juillet 2011 son premier EP (EP N°1) avec le batteur Jeff Ballard

felipe@iro.umontreal.ca

Résumé : Automatique Manuel Analyse Filtrer Aide

Texte réduit à 45% (193 mots / 995)

À sept ans il découvre le monde du jazz et passe ses journées à écouter différentes mélodies et à improviser au piano. À cette même période, il met au point ses premières compositions. L'année suivante, sa participation au premier festival de jazz d'Erevan lui permet de se faire remarquer et de se faire inviter pour de prochains concerts et sessions. Il rencontre le pianiste Stéphane Kochoyan qui va l'aider à se faire connaître en Europe.

Le jeune musicien fait alors la connaissance de légendes comme Wayne Shorter, Herbie Hancock, John McLaughlin ou Joe Zawinul et de musiciens comme Danilo Perez et John Patitucci. Grâce à son premier prix de piano-jazz emporté en 2005 au Thelonious Monk Institute of Jazz, il entre à l'Université de Californie du Sud à Los Angeles où il commence à étudier en profondeur et en parallèle le jazz contemporain et la musique arménienne. La même année il publie son deuxième album, New Era, accompagné de François Moutin et Louis Moutin, avec l'apparition de Vardan Grigoryan au duduk. Il se produit en 2011 dans de grands festivals comme Jazz in Marciac, Montreux, Montréal pour la 3^e année consécutive, ainsi qu'au Tokyo Jazz Festival, en Arménie, au Royaume-Uni ou encore en Allemagne.

Il enregistre en juillet 2011 son premier EP avec le batteur Jeff Ballard. Pour commémorer le centenaire du génocide arménien de 1915, il enregistre Luys i luso chez ECM avec le Yerevan State Chamber Choir.



https://resoomer.com

RESOOMER

 Service Comment ça marche PREMIUM Contact Connexion Blog **FR**

Tigran Hamasyan commence à s'intéresser au piano dès l'âge de deux ans. À trois ans, il chante les chansons de Led Zeppelin, Deep Purple, les Beatles, Louis Armstrong ou encore Queen en s'accompagnant au piano. À sept ans il découvre le monde du jazz, et passe ses journées à écouter différentes mélodies et à improviser au piano. Il poursuit alors son éducation musicale classique à l'école.

En 1997, quand sa famille déménage à Erevan, il étudie Duke Ellington, Thelonious Monk, Charlie Parker, Art Tatum, Miles Davis, Bud Powell. À cette même période, il met au point ses premières compositions. L'année suivante, sa participation au premier festival de jazz d'Erevan lui permet de se faire remarquer et de se faire inviter pour de prochains concerts et sessions. Lors du second festival de jazz d'Erevan en 2000, alors âgé de 13 ans, il attire l'attention de Chick Corea, Avishai Cohen, Jeff Ballard ou encore Ari Roland. Il rencontre le pianiste Stéphane Kochoyan qui va l'aider à se faire connaître en Europe. En 2001, ce dernier l'invite à plusieurs festivals en France. Le jeune musicien fait alors la connaissance de légendes comme Wayne Shorter, Herbie Hancock, John McLaughlin ou Joe Zawinul et de musiciens comme Danilo Perez et John Patitucci. En 2003 et 2004, il participe au Festival de Jazz de Serres, dans les Hautes-Alpes, où il revient en 2009 pour un duo avec Fanny Azzuro dans le cadre de Jazz & Classique.

Grâce à son premier prix de piano-jazz emporté en 2006 au Thelonious Monk Institute of Jazz, il entre à l'Université de Californie du Sud à Los Angeles où il commence à étudier en profondeur et en parallèle le jazz contemporain et la musique arménienne. La même année il publie son deuxième album, New Era, accompagné de François Moutin et Louis Moutin, avec l'apparition de Vardan Grigoryan au duduk. Il s'installe à New York en 2008.

En 2009, il enregistre Red Hill, un album au carrefour du jazz, du métal et du folklore arménien, avec son nouveau quintet de jeunes musiciens Aratta Rebirth: Ateni Agbalian (voc), Ben Wendel (ts), Charles Aitura (g), Sam Minaie (b) et Nate Wood (d). Ils se produisent dans plusieurs grands festivals internationaux, de Montréal à Nice en passant par Vienne ou Rotterdam (North Sea Jazz Festival).

En juin 2010, Tigran Hamasyan signe avec le label Verve. Il enregistre en septembre 2010 à Paris l'album solo *À faible*, pour lequel il est lauréat des Victoires du jazz 2011 dans la catégorie album international de production française.

Il se produit en 2011 dans de grands festivals comme Jazz in Marciac, Montreux, Montréal pour la 3e année consécutive, ainsi qu'au Tokyo Jazz Festival, en Arménie, au Royaume-Uni (Queen Elizabeth Hall) ou encore en Allemagne.

Il enregistre en juillet 2011 son premier EP (EP N°1) avec le batteur Jeff Ballard.

felipe@iro.umontreal.ca

Résumé : Automatique Manuel Analyse Filtrer Aide

Choisissez la taille de votre résumé :

 20% (105 mots / 505)

Tigran Hamasyan commence à s'intéresser au piano dès l'âge de deux ans. À trois ans, il chante les chansons de Led Zeppelin, Deep Purple, les Beatles, Louis Armstrong ou encore Queen en s'accompagnant au piano. Il rencontre le pianiste Stéphane Kochoyan qui va l'aider à se faire connaître en Europe. En 2001, ce dernier l'invite à plusieurs festivals en France. Le jeune musicien fait alors la connaissance de légendes comme Wayne Shorter, Herbie Hancock, John McLaughlin ou Joe Zawinul et de musiciens comme Danilo Perez et John Patitucci. Ils se produisent dans plusieurs grands festivals internationaux, de Montréal à Nice en passant par Vienne ou Rotterdam. En juin 2010, Tigran Hamasyan signe avec le label Verve.












Exemples d'applications langagières

- ▶ Moteurs de recherche (google, yahoo!, DuckDuckGo, etc.)
- ▶ Traduction automatique (ou assistée)
 - ▶ Babel Fish : <http://world.altavista.com/tr>,
 - ▶ Google Translate : <http://translate.google.fr/>,
 - ▶ MÉTÉO : <http://rali.iro.umontreal.ca/EnvironmentalInfo/WarningTranslation.html>
- ▶ Résumé automatique de textes, indexation automatique
- ▶ Extraction d'information

NLP ∈ IA : NELL

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

Exemples d'applications langagières

- ▶ Moteurs de recherche (google, yahoo!, DuckDuckGo, etc.)
- ▶ Traduction automatique (ou assistée)
 - ▶ Babel Fish : <http://world.altavista.com/tr>,
 - ▶ Google Translate : <http://translate.google.fr/>,
 - ▶ MÉTÉO : <http://rali.iro.umontreal.ca/EnvironmentalInfo/WarningTranslation.html>
- ▶ Résumé automatique de textes, indexation automatique
- ▶ Extraction d'information
- ▶ Classification de textes (spams, opinions)

Exemples d'applications langagières

- ▶ Moteurs de recherche (google, yahoo!, DuckDuckGo, etc.)
- ▶ Traduction automatique (ou assistée)
 - ▶ Babel Fish : <http://world.altavista.com/tr>,
 - ▶ Google Translate : <http://translate.google.fr/>,
 - ▶ MÉTÉO : <http://rali.iro.umontreal.ca/EnvironmentalInfo/WarningTranslation.html>
- ▶ Résumé automatique de textes, indexation automatique
- ▶ Extraction d'information
- ▶ Classification de textes (spams, opinions)
- ▶ Stylométrie, extraction terminologique, veille technologique

Exemples d'applications langagières

- ▶ Moteurs de recherche (google, yahoo!, DuckDuckGo, etc.)
- ▶ Traduction automatique (ou assistée)
 - ▶ Babel Fish : <http://world.altavista.com/tr>,
 - ▶ Google Translate : <http://translate.google.fr/>,
 - ▶ MÉTÉO : <http://rali.iro.umontreal.ca/EnvironmentalInfo/WarningTranslation.html>
- ▶ Résumé automatique de textes, indexation automatique
- ▶ Extraction d'information
- ▶ Classification de textes (spams, opinions)
- ▶ Stylométrie, extraction terminologique, veille technologique
- ▶ Détection d'opinion (*sentiment analysis*)

Fausses revues

Deceptive Opinion Spam Corpus v1.4

[Download
TAR Ball](#)[Download
ZIP File](#)[✉ e-mail](#)

Overview

This corpus consists of truthful and deceptive hotel reviews of 20 Chicago hotels. The data is described in two papers according to the sentiment of the review. In particular, we discuss positive sentiment reviews in [1] and negative sentiment reviews in [2].

While we have tried to maintain consistent data preprocessing procedures across the data, there *are* differences which are explained in more detail in the associated papers. Please see those papers for specific details.

This corpus contains:

- 400 truthful positive reviews from TripAdvisor (described in [1])
- 400 deceptive positive reviews from Mechanical Turk (described in [1])
- 400 truthful negative reviews from Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor and Yelp (described in [2])
- 400 deceptive negative reviews from Mechanical Turk (described in [2])

Exemples d'applications langagières

- ▶ Réponse automatique
 - ▶ aux questions “ouvertes”
(ex : Ask Jeeves : <http://www.ask.com>)
 - ▶ aux courriels

SQuAD 2.0

The Stanford Question Answering Dataset

What is SQuAD?

Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable.

New SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 new, unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering. SQuAD2.0 is a challenging natural language understanding task for existing models, and we release SQuAD2.0 to the community as the successor to SQuAD1.1. We are optimistic that this new dataset will encourage the development of reading comprehension systems that know what they don't know.

SQuAD2.0 paper [Rajpurkar & Ji et al. '18]

SQuAD1.0 paper [Rajpurkar et al. '16]

Getting Started

We've built a few resources to help you get started with the dataset.

Download a copy of the dataset [distributed under the CC BY-SA 4.0 license]:

Training Set v2.0 (40 MB)

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph. How will your system compare to humans on this task?

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar & Ji et al. '18)	86.831	89.452
1 Aug 28, 2018	SLQA+ (single model) Alibaba DAMO NLP http://www.aclweb.org/anthology/P18-1158	71.451	74.422
1 Aug 15, 2018	Reinforced Mnemonic Reader + Answer Verifier (single model) NUJIT https://arxiv.org/abs/1808.05759	71.699	74.238
2 Aug 21, 2018	FusionNet++ (ensemble) Microsoft Business Applications Group AI Research https://arxiv.org/abs/1711.07341	70.345	72.555
3 Aug 21, 2018	SAN (single model) Microsoft Business Applications Research Group https://arxiv.org/abs/1712.03556	68.653	71.441
4 Jul 19, 2018	VS*3-NET (single model) Kangwon National University in South Korea	68.438	71.282
4 Aug 25, 2018	ARRR (single model) anonymous	68.641	71.113
5 Jun 24, 2018	KACTEIL-MRC(GFN-Net) (single model) Kangwon National University, Natural Language	68.224	70.871

Exemples d'applications langagières

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

marketmuse.com.png

MarketMuse
Products Contact Sign In Analyze My Content

Less work and more flow in your content workflow.

Article URL

Analyze my content

marketmuse.com - Applications - Complete
🏠 🗨️ ⚙️ 🗪

🏠 Complete

Select a competitor to the right to swap them into the feed to feed tool.

Topic	Quality
ai content strategy	21 28 29 30 24 18 20 34 17 38 41 28 40 22 34 18 20 15
ai tools	
algorithms	
audience	
content creation	
content marketing	
data	

Page Rank	1	2	4	5	6	7	8	10	11	12	13	14	15	17	18	19
contentmarketingintelligence.com...																
shinobi.com/resources/publist/ai-ai...																
technemerge.com/artificial-intelligence-...																
marketingsmartlab.com/blog/ai-artifici...																
marketingsmartlab.com/blog/ai-to-ym...																
content.com																
content.com/blog/first-reasons-why-yo...																
valuenation.com/content-strategy-marke...																
content.com/blog/ai-first-step-content-mar...																
verycreative.com/content-marketing-strategy...																
blogpost.seesaw.com/ai-content-marketing...																
technomark.com/2017/03/29/ai-will-kill-bus...																
contentmarketinginsights.com/2018/03/30/ai...																
contentmarketinginsights.com/2018/03/30/ai...																
pressbooks.com/news-releases/ai-of-arti...																
hubspot.com/artificial-intelligence																
focusoncontent.com/content-marketing-str...																
contentmarketing.com/ai-advanced-analysis-report...																

<https://www.prisnewswire.com/news-re...> artificial intelligence



Exemples d'applications langagières

▶ Réponse automatique

- ▶ aux questions “ouvertes”
(ex : Ask Jeeves : <http://www.ask.com>)
- ▶ aux courriels

▶ Aide à la rédaction

- ▶ correcteurs
- ▶ accélération de la saisie
- ▶ détection de plagiats, articles reliés
- ▶ générateurs de textes

▶ agents conversationnels (chatbots)

- ▶ <https://www.cnet.com/news/facebook-is-killing-m-its-personal-chatbot-assistant/>

18 oct. 2016 – <https://www.gartner.com/smarterwithgartner/gartner-predicts-a-virtual-world-of-exp>

Smarter With **Gartner**.

TRENDS

Gartner's Top 10 Strategic Predictions for 2017 and Beyond: Surviving the Storm Winds of Digital Disruption

By 2020, the average person will have more conversations with bots than with their spouse. With the rise of Artificial Intelligence (AI) and conversational user interfaces, we are increasingly likely to interact with a bot (and not know it) than ever before. The digital experience has become addictive by entering our lives through smartphones, tablets, virtual personal assistants (VPAs) or the entertainment systems in our homes and cars.

Applications développées au RALI



Recherche appliquée en linguistique informatique

english

rali

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

Af

Applications développées au RALI



Recherche appliquée en linguistique informatique

english

rali

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

Suédois cp1252
 Suédois cp850
 Suédois macintosh
 Suédois utf8
 Thai tis620
 Thai utf8
 Turc cp853
 Turc iso-8859-9
 Turc utf8
 Chinois big5

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

Jag talar inte bra.

Soumettre

texte

Choisir le fichier

aucun fichier sélectionné



rali

Applications développées au RALI



Recherche appliquée en linguistique informatique

english

rali

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

<ul style="list-style-type: none"> Malais macintosh Malais utf8 Néerlandais cp850 Néerlandais macintosh Néerlandais utf8 Norvégien cp1252 Norvégien cp850 Norvégien macintosh Norvégien utf8 Polonais cp1250 	<p>manao ahoana ianao</p>
<p>Soumettre <input type="button" value="texte"/></p>	<p>Choisir le fichier <input type="button" value="aucun fichier sélectionné"/></p>

Applications développées au RALI



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



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



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 pl^t 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



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

- ▶ 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 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

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](#)

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

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](#)

- | | | |
|---|---|--|
| 1 | Toutefois, lorsqu'ils ont remporté les élections, c'était alors une toute autre paire de manches . | 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 paire de manches . | 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 paire de manches , 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 paire de manches . | However, to compare that to the issue of whether the traditional definition of marriage should be maintained is something |

Applications développées au RALI

TransSearch

[TERMINOTIX](#)

[RALI](#)

Utilisateur : felipe

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

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!

Applications développées au RALI

TransSearch

[RALI](#)

utilisateur: felipe

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

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

Collection de documents :

Expression :

- | | | |
|---|--|--|
| 1 | <p>These changes are taking place in the context of political and economic reforms, as well as an increasing decentralization of the health services.</p> | <p>Esos cambios se presentan dentro de un marco de reformas políticas y económicas, y de mayor descentralización de los servicios de salud.</p> |
| 2 | <p>It will identify the areas and the population groups that need specific poli- cies, sustained intervention programs, and health services.</p> | <p>Permitirá también identificar las áreas y los grupos de población que necesitan políticas específicas, programas de intervención sostenible y servicios de salud.</p> |
| 3 | <p>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.</p> | <p>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.</p> |

Applications développées au RALI

TransSearch

[RALI](#)

utilisateur: felipe

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

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

 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)

TRANSSEARCH H³ BETA
TERMINO TIX

UTILISATEUR : **fellpe** 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	
matter 6		
different story 5	Là, nous avons une nouvelle paire de manches , car si vous êtes conservateurs, vous êtes contre ce genre de dépenses.	This is a different kettle of fish , because a conservative generally opposes this kind of spending.
different issue 4		
different 2	C'était une autre paire de manches .	It was a very different kettle of fish .
entirely 2		
issue 2	S'ils ne font pas confiance aux juges, c'est une autre paire de manches .	If the members opposite do not trust judges, that is a different kettle of fish .
thing 2		
ball game 2	Toutefois, lorsqu'ils ont remporté les élections, c'était alors une toute autre paire de manches .	However, when they won the elections, it was a different kettle of fish .
different thing 2		
question 2	C'est une autre paire de manches .	It is a different kettle of fish .
of a problem with 2		
story 2	La période des questions, c'est une autre paire de manches .	Question period is a different kettle of fish .
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 paire de manches .	If the hon. member for Delta-South Richmond takes exception to the response to the question that he submitted, that is an entirely different kettle of fish .
different issue for some 1		
horse of a different colour 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 paire de manches .	Surely if my hon. friend is suggesting that Mr. Yeutter wants to change the trade balances using unfair trading practices, that is a different kettle of fish .
thing altogether 1		
different matter altogether 1		
different balgame altogether 1		
quilt 1		
solving 1		
kettle of fish 1		
really matter 1		

concepteur : Fabrizio Gotti

felipe@iro.umontreal.ca

Introduction au Traitement Automatique des Langues Naturelles (T

Applications développées au RALI

.samples\predictor_offFW2_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,

Applications développées au RALI

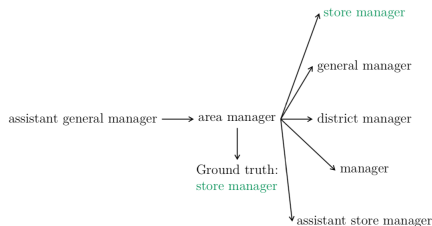
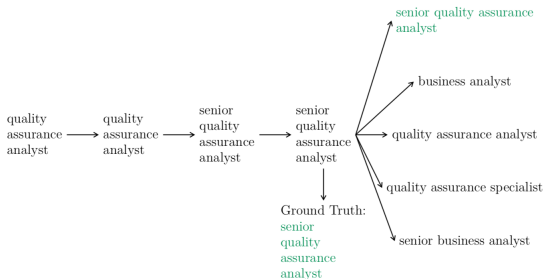


FIG. 3.1. Top 5 RNN prediction of last job given the users professional history



Plan

Contexte

Quelques Applications

Plan (approximatif)

Repères historiques

Approche pipeline

Les mains dans le cambouis

Sujets abordés

- ▶ **Modèles de langue** *p(colorless idea sleep furiously) ?*
 - ▶ survol de quelques techniques de lissage
 - ▶ approche neuronale

- ▶ **Algorithmique du texte** *mots proches de a2m1 ?*
 - ▶ programmation dynamique : edit-distance, etc.

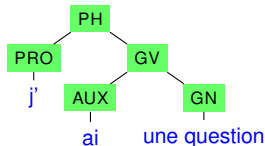
- ▶ **Étiquetage morpho-syntaxique** *Je parle trop* → PRN VB ADV
 - ▶ étiquetage séquentiel
 - ▶ approche transformationnelle

- ▶ **Apprentissage analogique** [⊙ : ⊙ :: ⊗ : ?]
 - ▶ problèmes
 - ▶ réalisations

Sujets abordés

► Grammaires

- éléments de théorie des langues
- grammaires probabilistes



► Traduction automatique

- statistique / neuronale
- bitextes

Elle l'aime → *She loves*¹

1. google – <http://translate.google.com/> – 1 octobre 2012.
Traduction au 9 janvier 2018 : She loves him

Sujets abordés

- ▶ **Extraction d'information ouverte (OIE)**
 - ▶ extracteurs de triplets
 - ▶ reconnaissance d'entités nommées
- ▶ **Sémantique lexicale** ?? *your microwavable popcorn before eating it*
 - ▶ approches distributionnelles
 - ▶ extraction de collocations

Plan

Contexte

Quelques Applications

Plan (approximatif)

Repères historiques

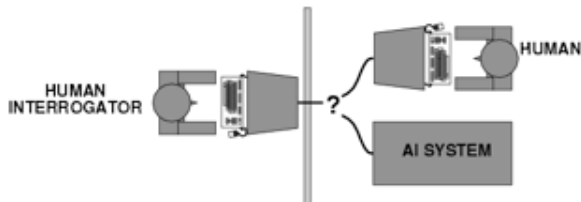
Approche pipeline

Les mains dans le cambouis

Quelques repères historiques

- ▶ après-guerre : essor de l'informatique, de la théorie des langages formels (Turing, Kleene, Chomsky, Backus, Naur) et de la théorie de l'information (Shannon)
- ▶ 1950 : le test de Turing

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
- ▶ test annuel mis en place depuis 1991 via le [Loebner Prize](#) (gagnant 2017 : mitsuku)
- ▶ [Amazon Alexa Prize](#) : être capable de converser de manière cohérente pendant 20 minutes sur des sujets populaires (sport, politique, divertissement, mode).

Quelques repères historiques

- ▶ après-guerre : essor de l'informatique, 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.

Quelques repères historiques

- ▶ après-guerre : essor de l'informatique, 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 : le système ELIZA [*Weizenbaum, 66*]

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]*

```

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/>

Quelques repères historiques

- ▶ après-guerre : essor de l'informatique, 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 : le système ELIZA [*Weizenbaum, 66*]
- ▶ 1968 : le premier (vrai) système de traduction (Systran, russe→anglais)

Quelques repères historiques

- ▶ après-guerre : essor de l'informatique, 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 : le système ELIZA [*Weizenbaum, 66*]
- ▶ 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 statistique multilocuteur



Quelques repères historiques

- ▶ 1995 : SVM [*Cortes and Vapnik, 1995*]
- ▶ 2009 : Probabilistic Graphical Models [*Koller and Friedman, 2009*]
- ▶ 2000s : modèle de langue neuronal [*Bengio et al., 2001, Bengio et al., 2003*]
- ▶ 2011 : Natural Language Processing (almost) from scratch [*Collobert et al., 2011*]
- ▶ 2013 : Word2Vec [*Mikolov et al., 2013*]
- ▶ 2015 : Deep learning **explosion** (MILA inside) !!!

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

- ▶ Ces deux phrases ont la même probabilité d'être observées dans un corpus, à savoir, faible².
 - ▶ 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 : $\exists \epsilon : \forall s, \text{grammatical}(s) \leftrightarrow \lim_{n \rightarrow \infty} P_n(s) > \epsilon$

- ▶ “We cannot seriously propose that a child learns the values of 10^9 parameters in a childhood lasting only 10^8 seconds”.

2. 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>

Plan

Contexte

Quelques Applications

Plan (approximatif)

Repères historiques

Approche pipeline

Les mains dans le cambouis

Différents niveaux de traitement

- ▶ segmenter le texte en **unités lexicales** (mots)

Note : ambiguïté à tous les niveaux

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

Morphologie

- ▶ 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 morphologique

http://www.lattice.cnrs.fr/sites/itellier/poly_info_ling/info-ling.pdf

- ▶ un **morphème** est une unité linguistique minimale ayant une forme et un sens
- ▶ les **affixes** sont des morphèmes qui n'existent pas en tant que mots

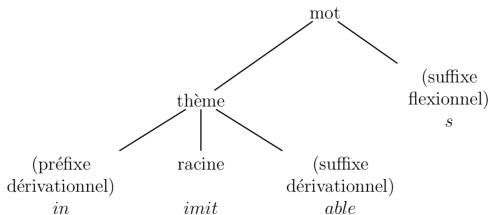


FIG. 4.1 – l'affixation en français

Différents niveaux de traitement

- ▶ segmenter le texte en **unités lexicales** (mots)
- ▶ identifier les composants lexicaux, leur propriétés : **traitement lexical**

Note : ambiguïté à tous les niveaux

Analyse lexicale

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

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

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

Traitement des mots composés

- ▶ **nominaux** : femme de ménage, ouvre-bouteille, pomme de terre
 - ▶ Je veux parler à la femme de ménage. **Qui est Ménage ?**
- ▶ **termes** :
 - ▶ réseaux de neurones, réseaux neuromimétiques, réseau neuronal
- ▶ **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

Analyse lexicale

Peut (aussi) se retrouver dans ce type de problème³ :

- ▶ co instructor and `teaching assistant`, executive programs & undergraduate programs
- ▶ `specialist in organization and standardization of labor`
- ▶ `college professor` in human biology
- ▶ `accountant`, sales and marketing department
- ▶ `senior manager`, project management methodology & governance
- ▶ `journalist` and travel writer

3. ex. de *JobTitle* fournis par des usagers de LinkedIn.

Différents niveaux de traitement

- ▶ segmenter le texte en **unités lexicales** (mots)
- ▶ identifier les composants lexicaux, leur propriétés : **traitement lexical**
- ▶ identifier les syntagmes : **analyse syntaxique**

Note : ambiguïté à tous les niveaux

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

Différents niveaux de traitement

- ▶ 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**

Note : ambiguïté à tous les niveaux

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

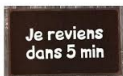
Différents niveaux de traitement

- ▶ 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

Analyse pragmatique

- ▶ prise en compte de la situation d'énonciation (information explicite)



- ▶ intention du locuteur :

Viendras-tu au bal ce soir ? J'ai entendu que Paul y sera !

- ▶ connaissances du monde

Plan

Contexte

Quelques Applications

Plan (approximatif)

Repères historiques

Approche pipeline

Les mains dans le cambouis

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 ?

`<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
```

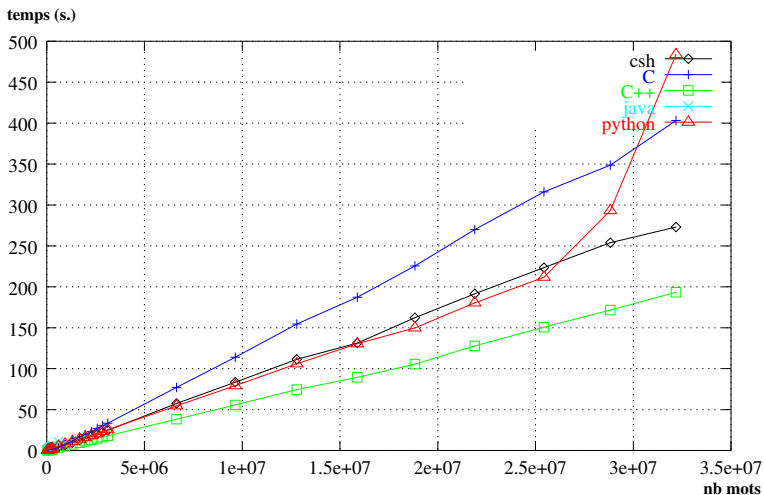
- C++ (ou autre langage du même type) :

```
1 cat corpus | frequence
```

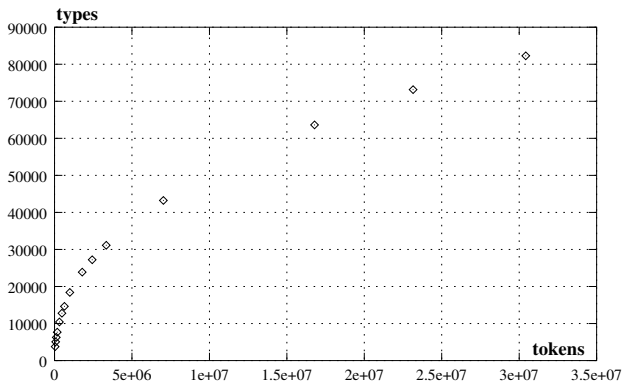
Où `frequence` est un programme utilisant une structure de donnée de type *hash-map*. Voici à quoi ça peut ressembler avec la *STL* :

```
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



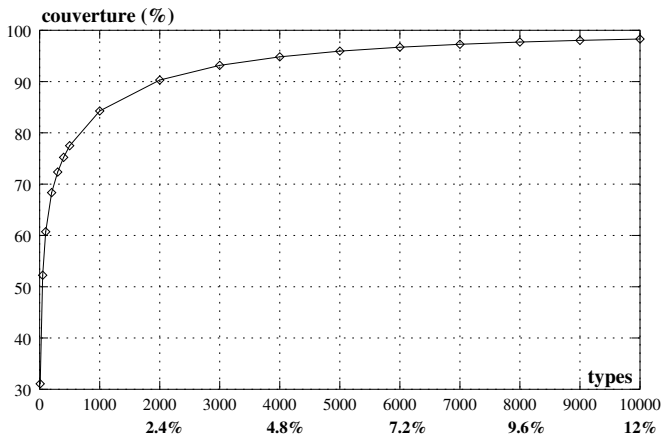
- ▶ 60000 entrées dans le Petit Robert ; 75000 dans le grand Robert.
- ▶ Vocabulaire moyen d'un individu < 5000 mots (environ)

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

- ▶ 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



- 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



Bengio, Y., Ducharme, R., and Vincent, P. (2001).

A neural probabilistic language model.

In Advances in Neural Information Processing Systems.



Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003).

A neural probabilistic language model.

J. Mach. Learn. Res., 3 :1137–1155.



Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011).

Natural language processing (almost) from scratch.

Journal of Machine Learning Research, 12 :2493–2537.



Cortes, C. and Vapnik, V. (1995).

Support-vector networks.

Mach. Learn., 20(3) :273–297.



Gardent, C. (2007).

Natural language processing applications.

Notes de cours.

**Jurafsky, D. and Martin, J. H. (2000).***Speech and Language Processing.*

Prentice Hall.

**Koller, D. and Friedman, N. (2009).***Probabilistic Graphical Models : Principles and Techniques - Adaptive Computation and Machine Learning.*

The MIT Press.

**Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013).**

Efficient estimation of word representations in vector space.

CoRR, abs/1301.3781.**Polguère, A. (2008).***Lexicologie et sémantique lexicale. Notions fondamentales.*

Coll. "Champs Linguistiques". Les Presses de l'Université de Montréal.

**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.