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

rali

Plan

rali

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

rali

Le RALI

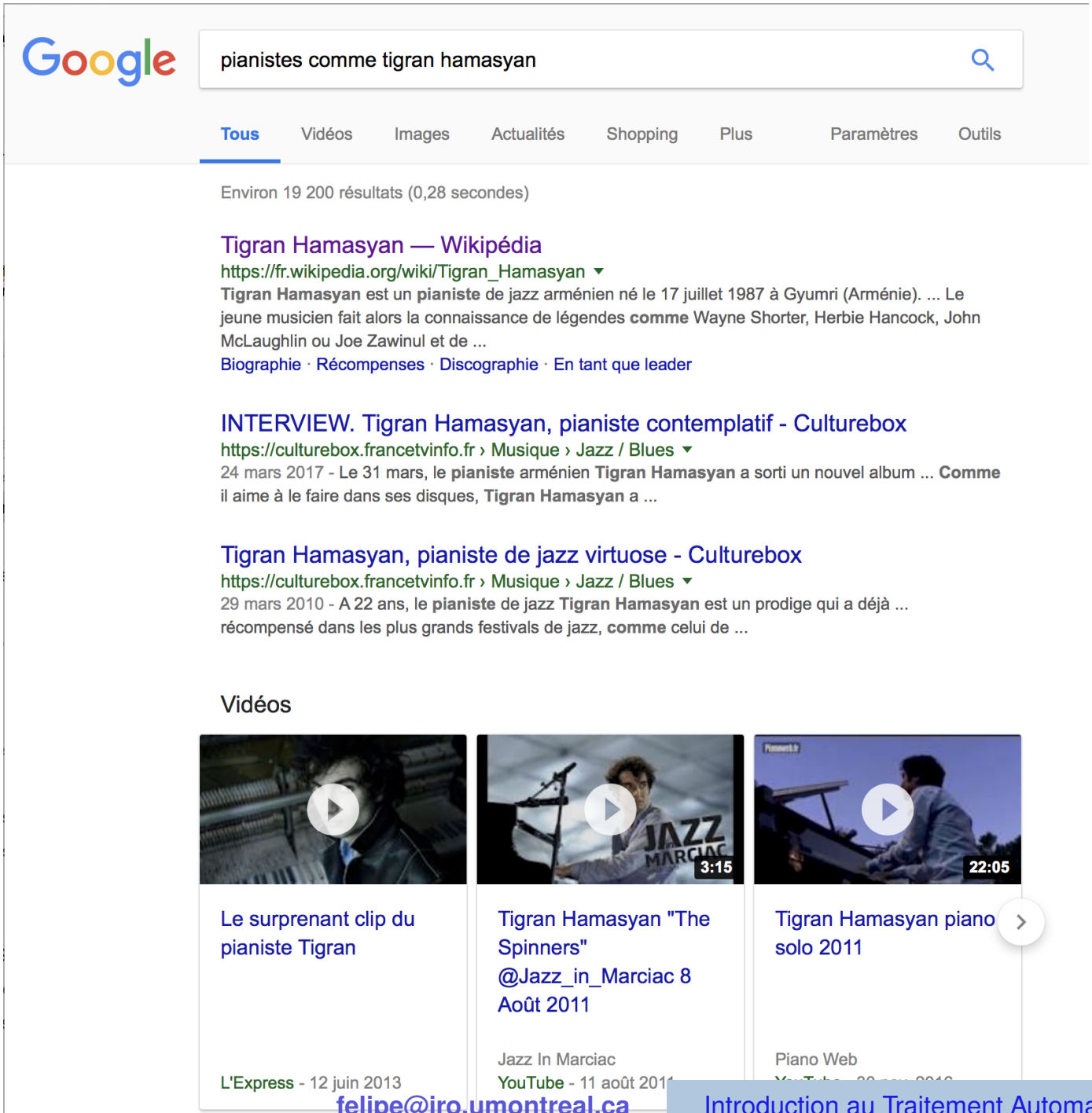
- 3 profs :
  - Guy Lapalme : question-réponse, résumé, etc.
  - Jian-Yun Nie : recherche d'information, etc.
  - Philippe Langlais : traduction, extraction d'information, etc.
- 2 profs associés :
  - Caroline Barière, CRIM
  - Atefeh Farzindar, NLP Technologies
- $a \in [1, 5]$ assistants de recherche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

rali

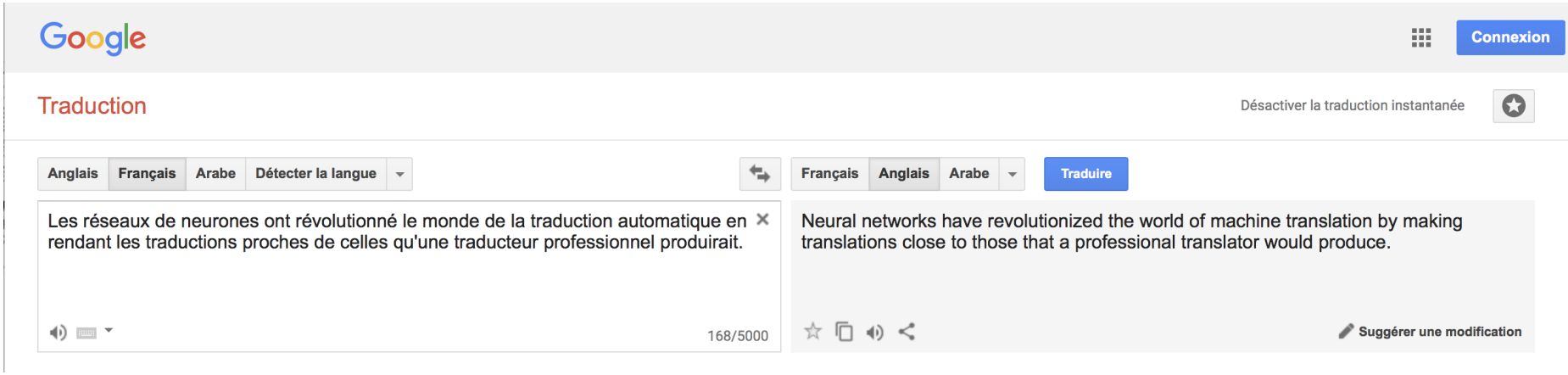
Exemples d'applications langagières

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


Google
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
3:15
22:05
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 201
Tigran Hamasyan piano solo 2011
Piano Web

Exemples d'applications langagières

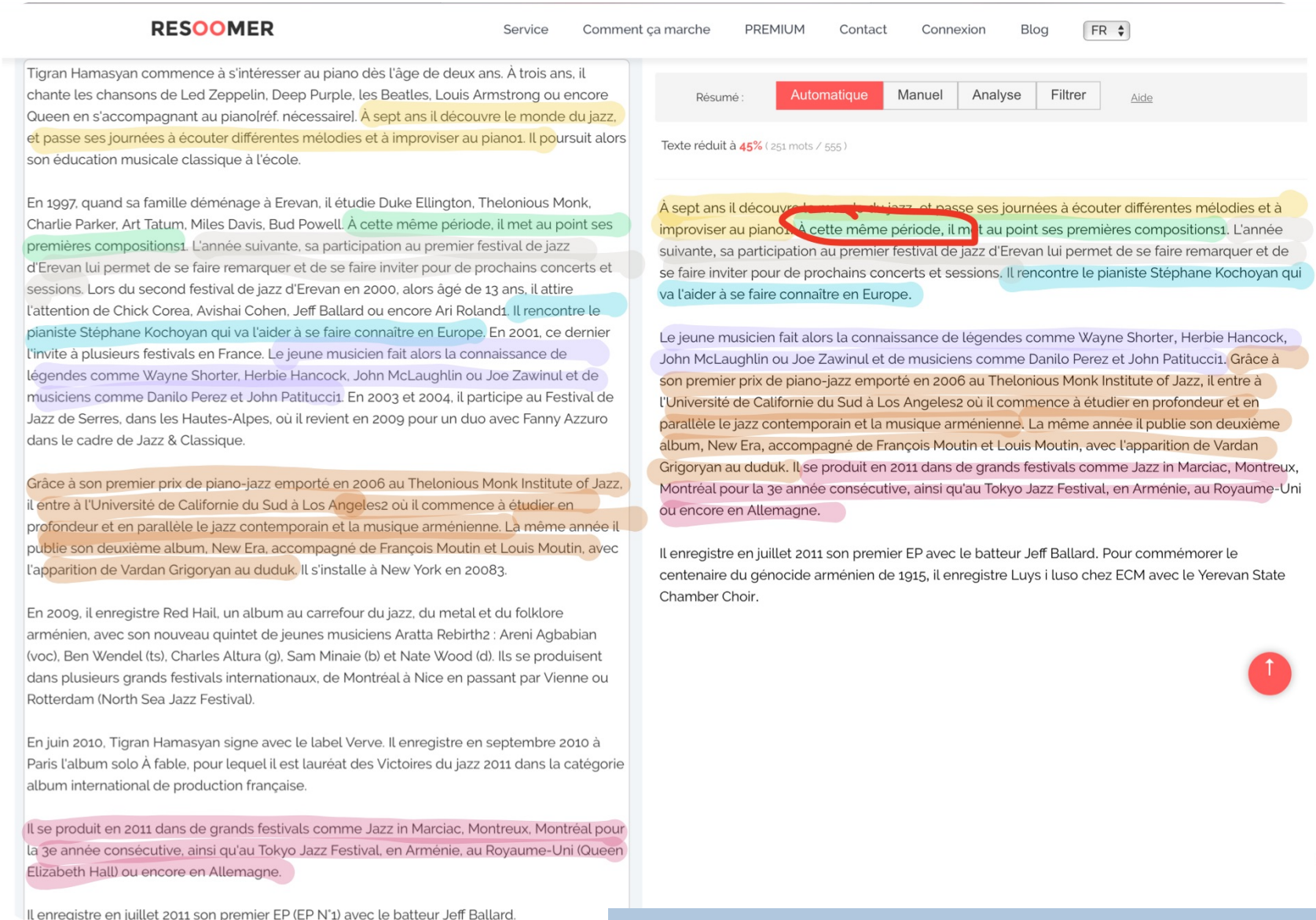
- Moteurs de recherche (google, yahoo !, DuckDuckGo, etc.)
- Traduction automatique (ou assistée)
  - Babel Fish : http://world.altavista.com/tr,
  - Google Translate : http://translate.google.fr/,
  - MÉTÉO : http://rali.iro.umontreal.ca/EnvironmentalInfo/WarningTranslation.html


Google
Connexion
Traduction
Désactiver la traduction instantanée
Anglais
Français
Arabe
Détecter la langue
Français
Anglais
Arabe
Traduire
Les réseaux de neurones ont révolutionné le monde de la traduction automatique en rendant les traductions proches de celles qu'une traducteur professionnel produirait.
168/5000
Neural networks have revolutionized the world of machine translation by making translations close to those that a professional translator would produce.
Suggérer une modification

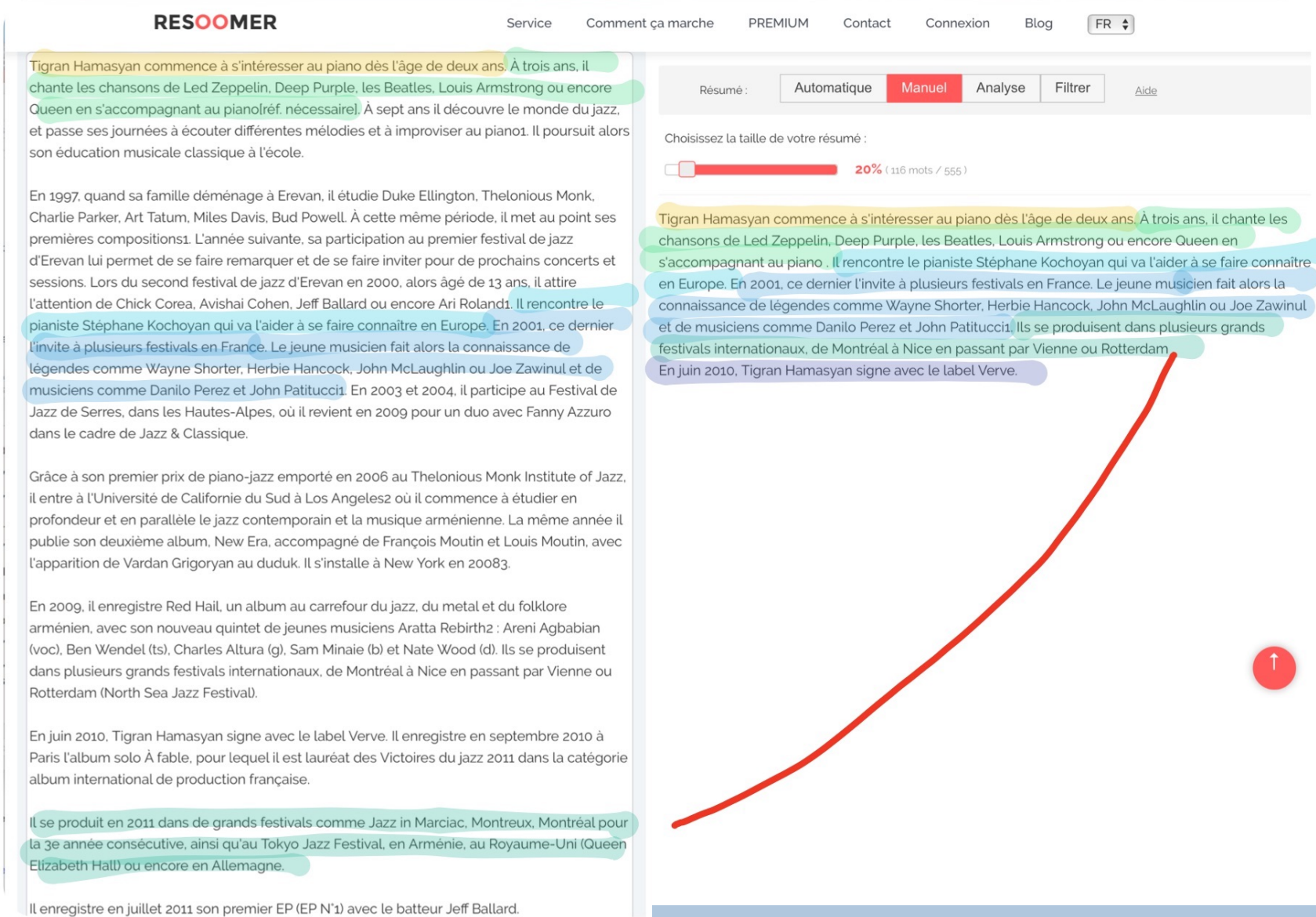
Exemples d'applications langagières

- 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



https://resoomer.com



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 $\in$ 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, $\sim$ 1.5M fiables)
- intervention manuelle minimaliste

rali

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)

rali

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

rali

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*)

rali

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.

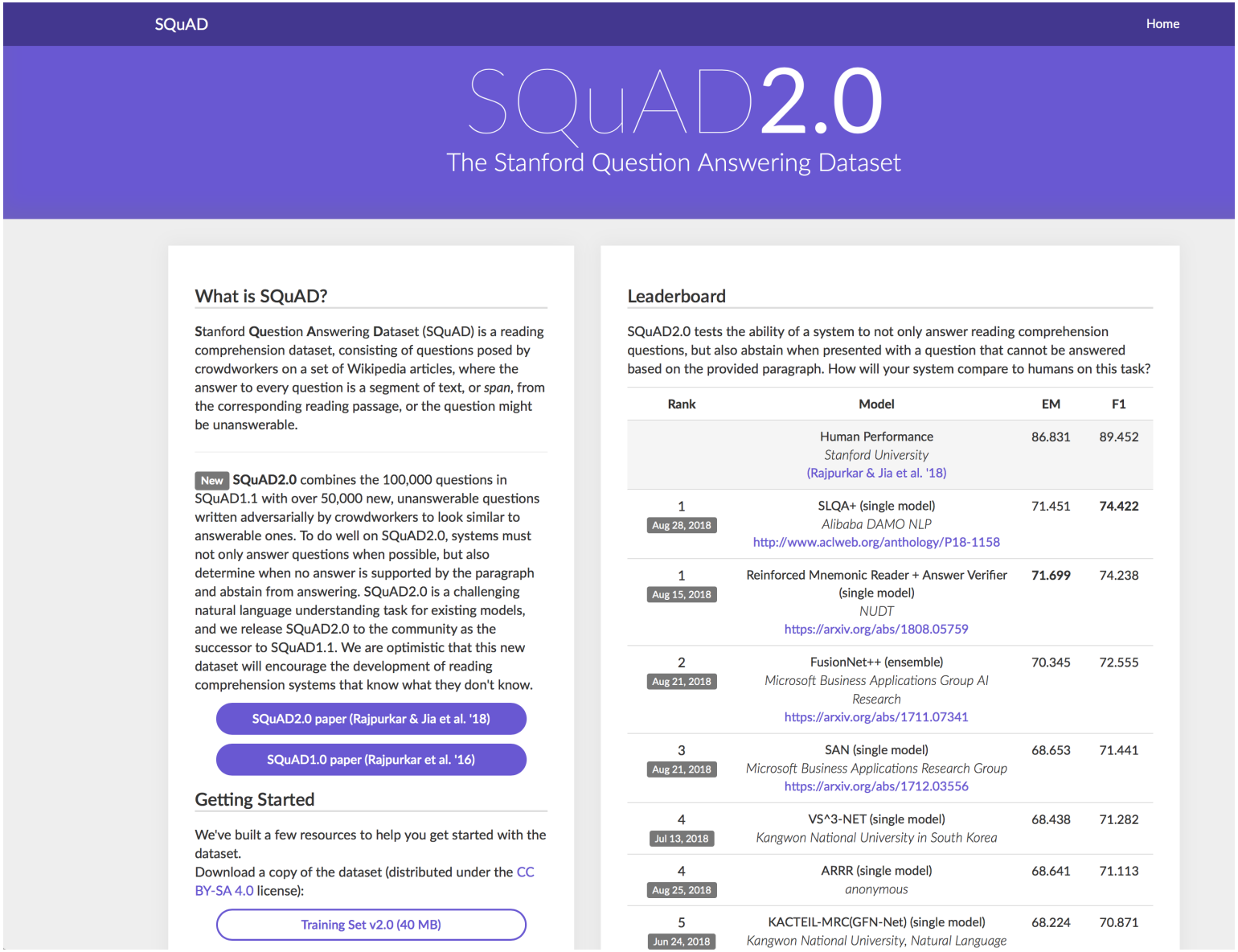
This corpus contains:

- 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])

rali

Exemples d'applications langagières

- Réponse automatique
  - aux questions "ouvertes"
    (ex : Ask Jeeves : `http://www.ask.com`)
  - aux courriels



SQuAD Home

SQuAD2.0

The Stanford Question Answering Dataset

What is SQuAD?

Stanford **Qu**estion **A**nswering **D**ataset (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 & Jia 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 & Jia 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) *NUDT* 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 13, 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

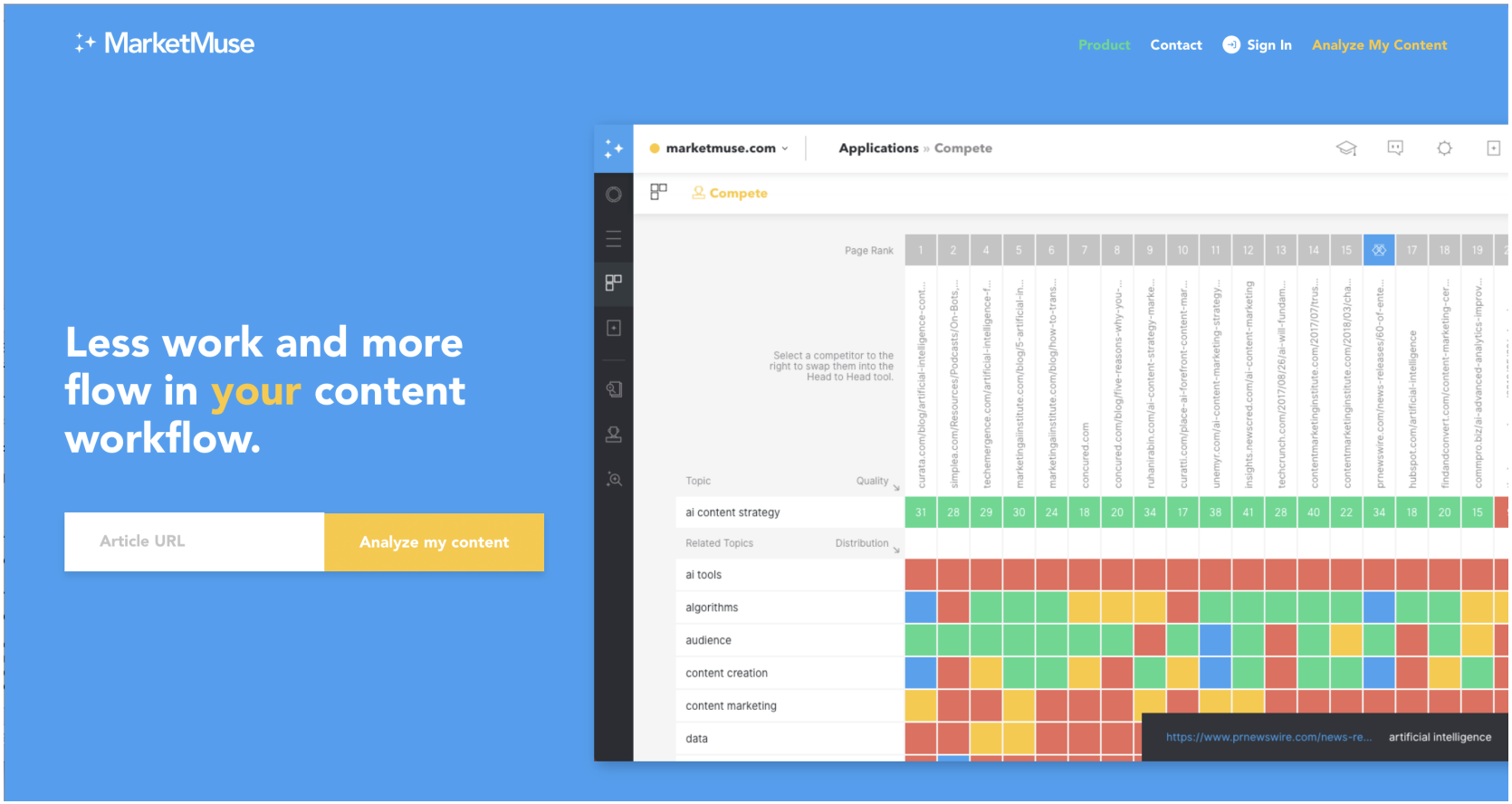
rali

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

rali

marketmuse.com.png



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/

rali

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.

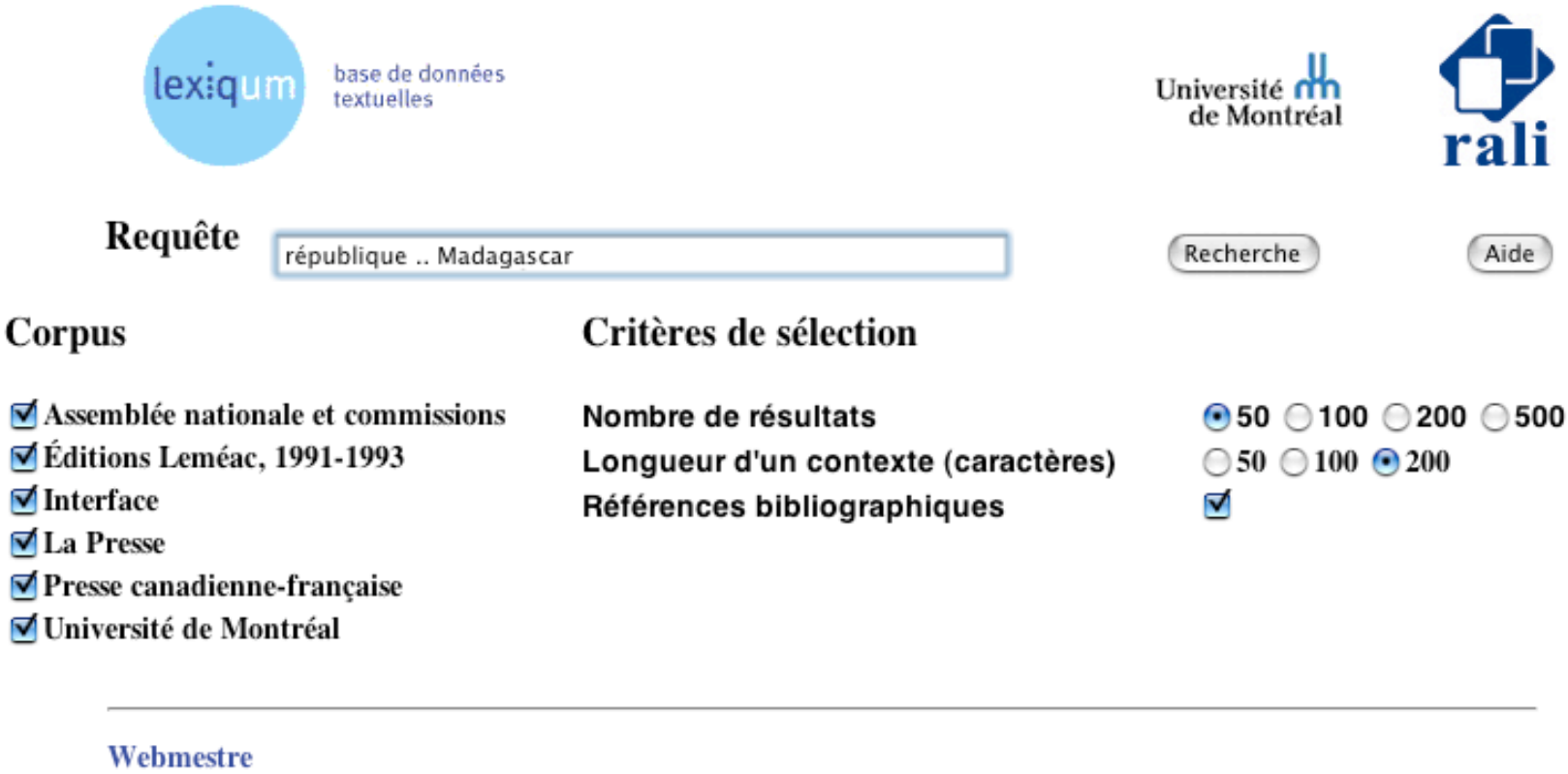
rali

Applications développées au RALI

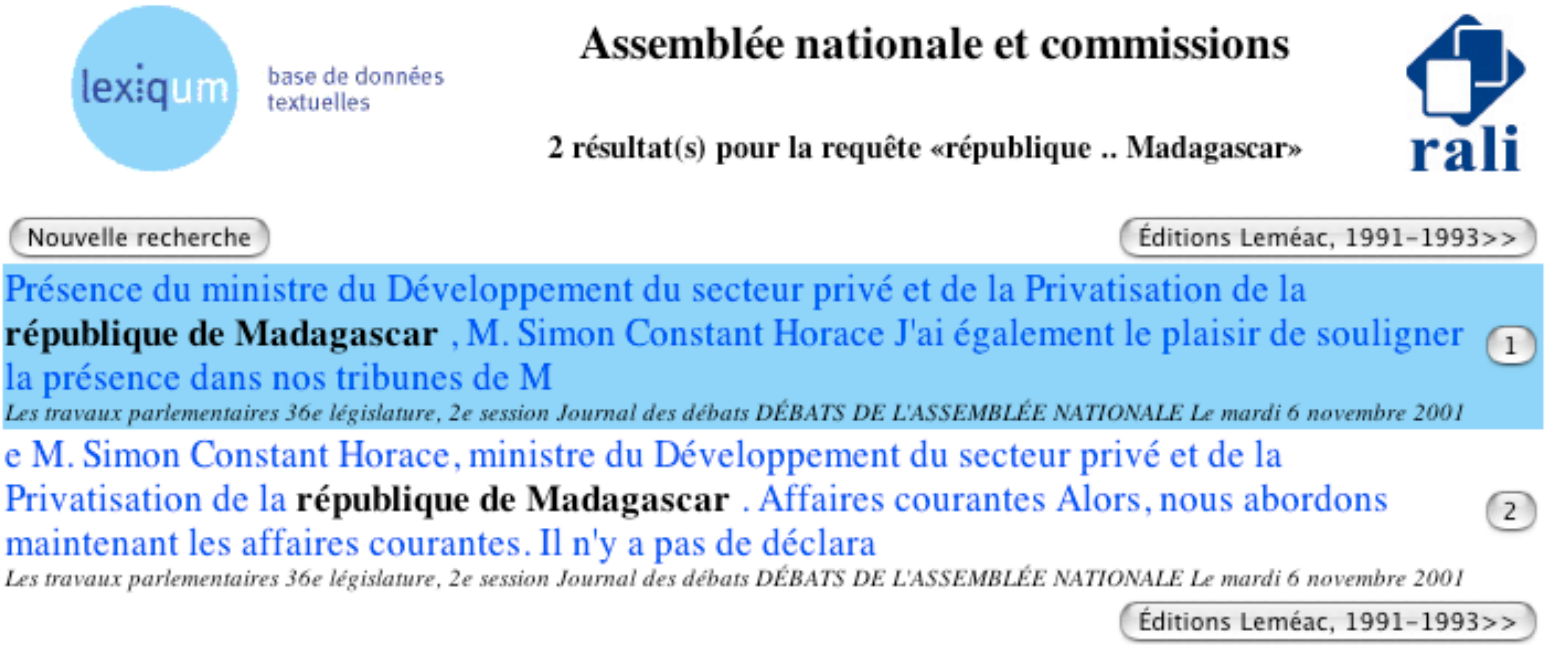
Applications développées au RALI

Applications développées au RALI

Applications développées au RALI

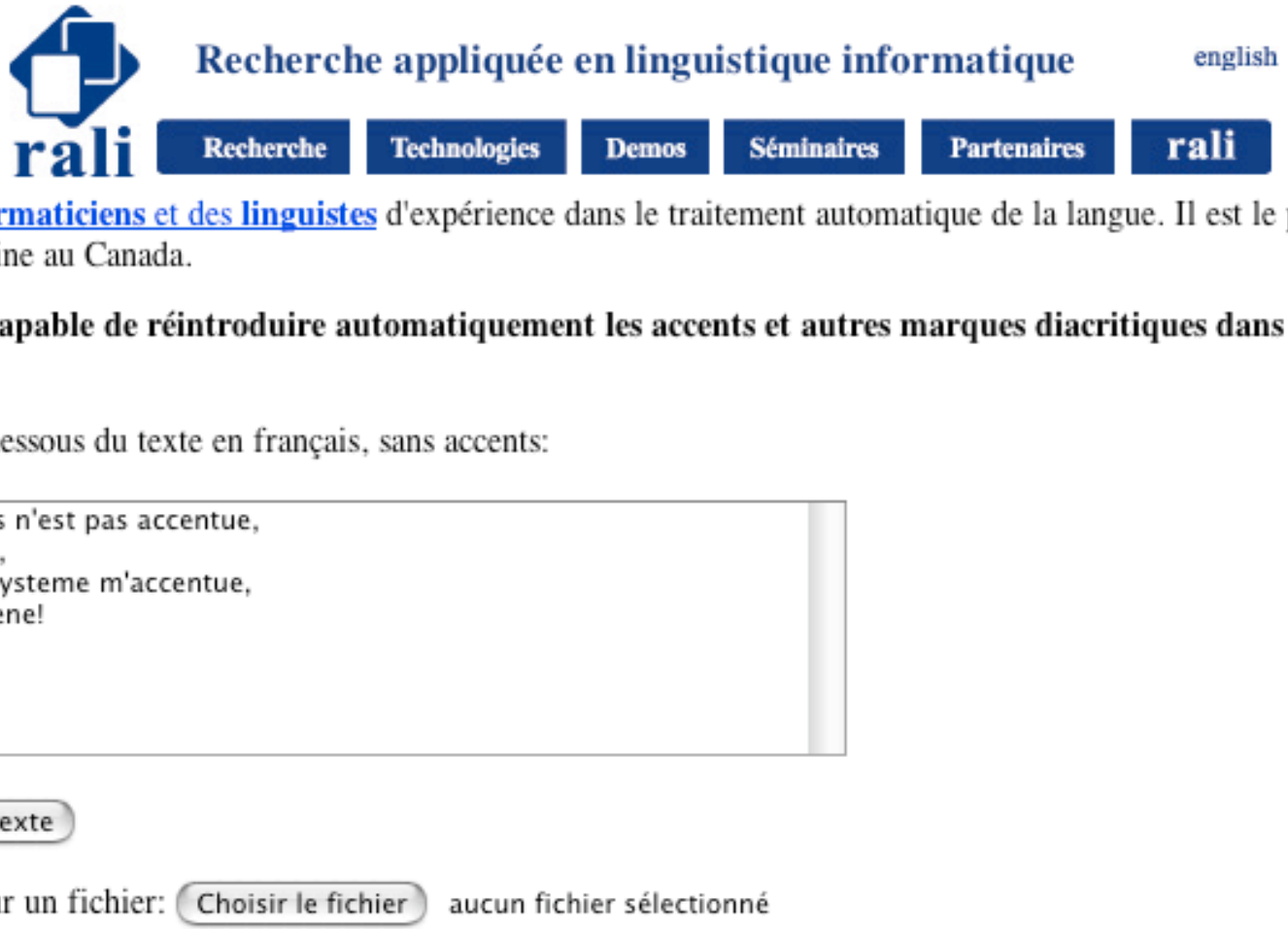


Applications développées au RALI



rali

Applications développées au RALI



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 import
laboratoire dans le domaine au Canada.

Entrée:

La ou le francais 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 gêne,
mais quand le système m'accentue,
je suis moins gêné!

Soumettre une nouvelle requête

Webmeister: (11/7/2005)

rali

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.)
  - $\sim$ 1 500 abonnés
  - $\sim$ 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)

rali

Applications développées au RALI

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 : Hansard canadien (1986-2005)

Expression : paire de manches Requête bilingue

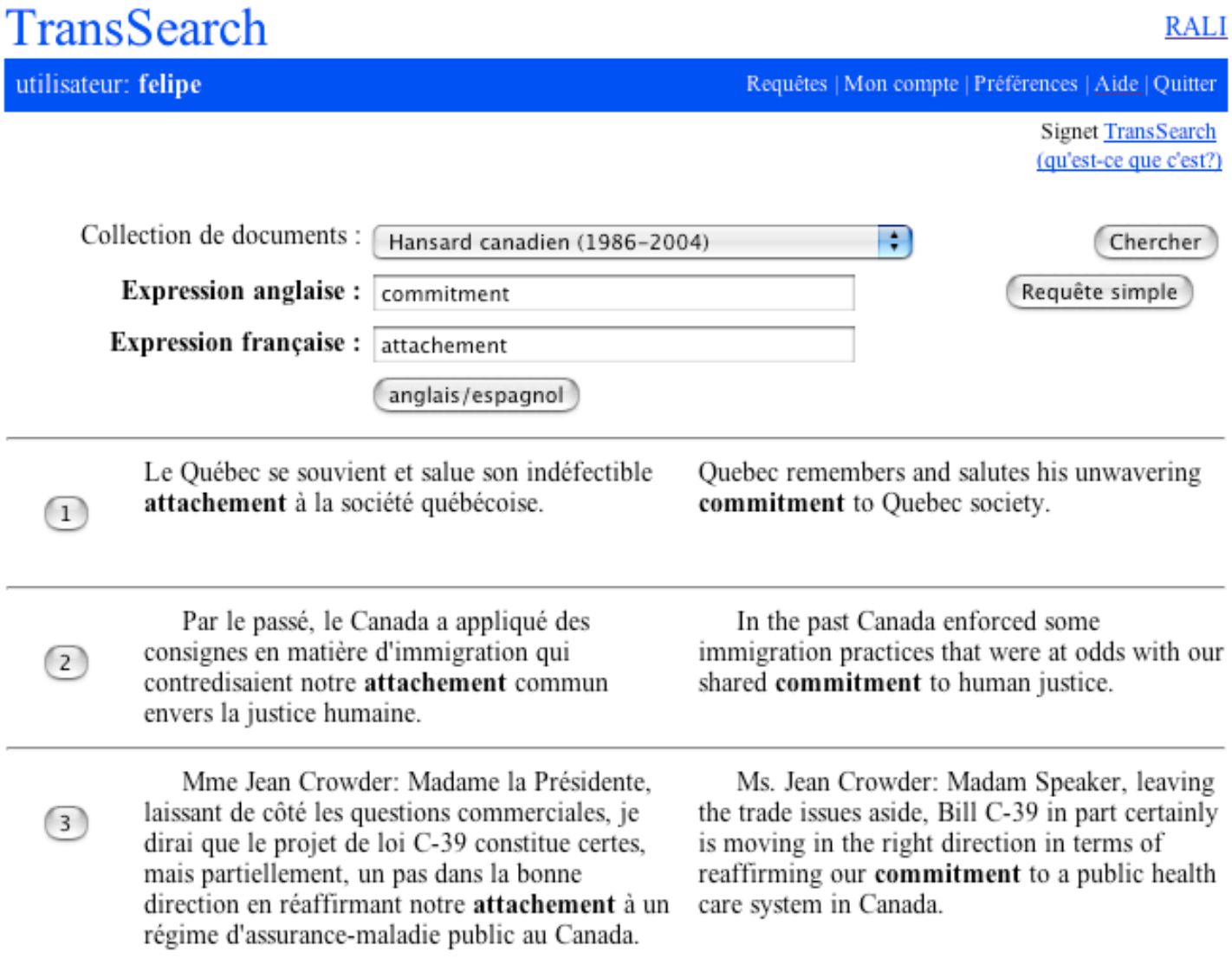
Chercher

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

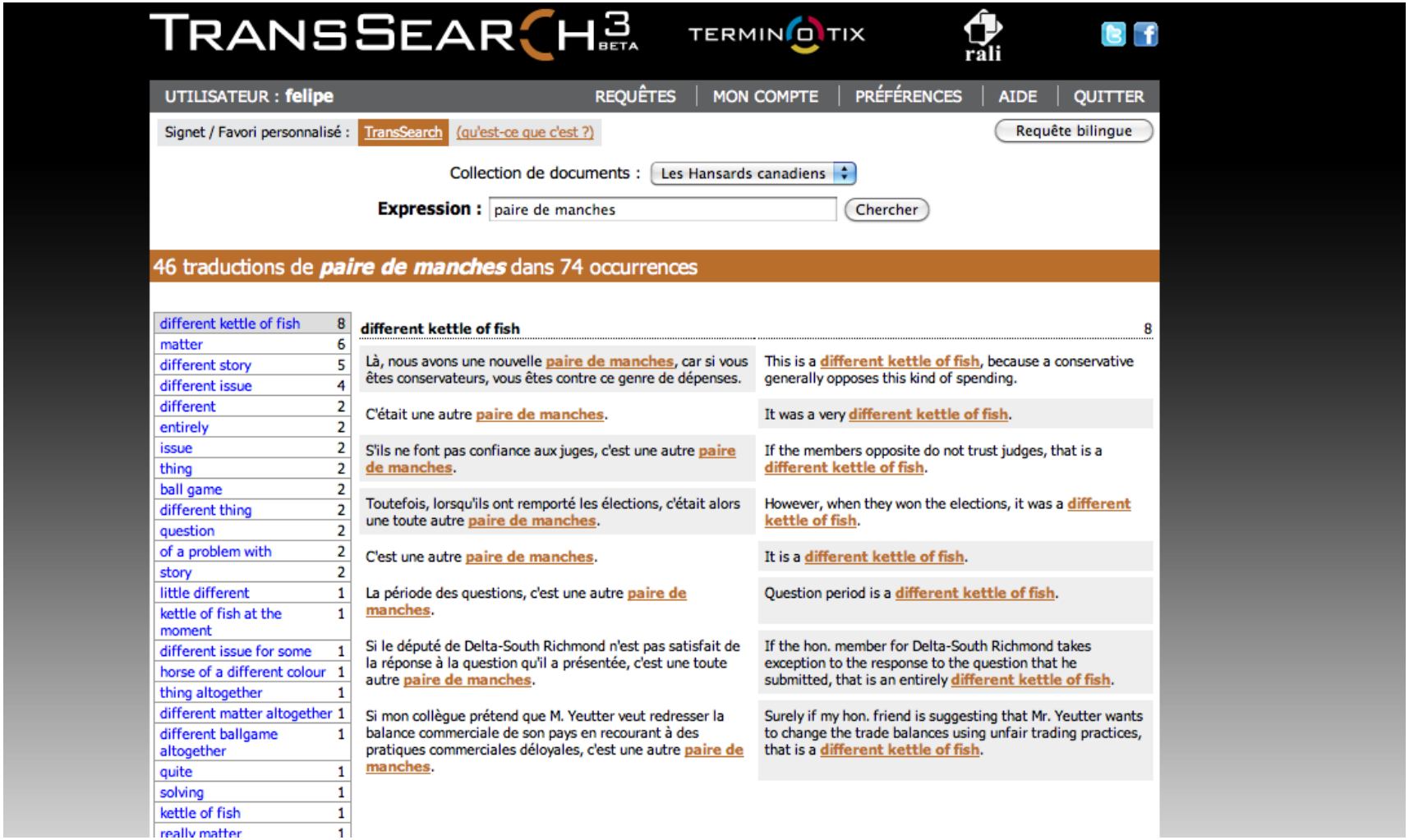
Applications développées au RALI

Applications développées au RALI

Applications développées au RALI

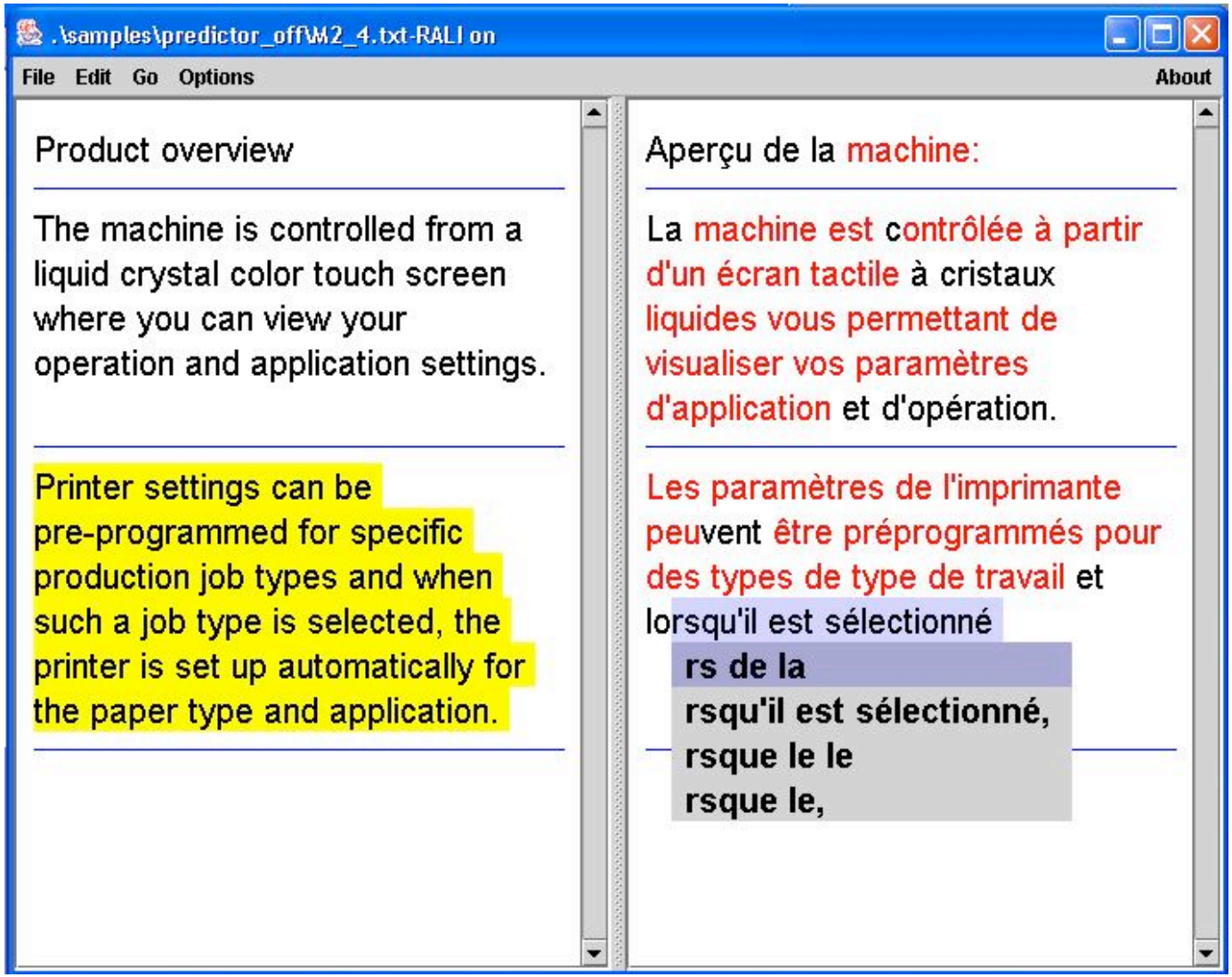


TSRALI.com nouvelle mouture (TS3)



concepteur : Fabrizio Gotti

Applications développées au RALI



Applications développées au RALI

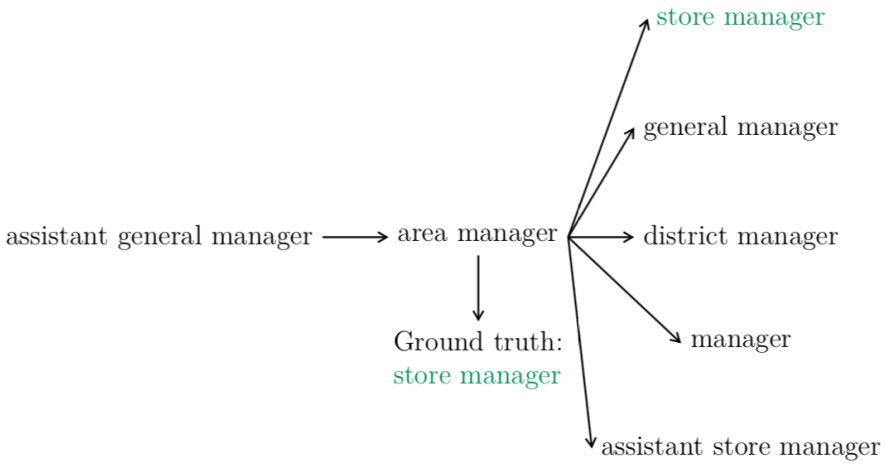
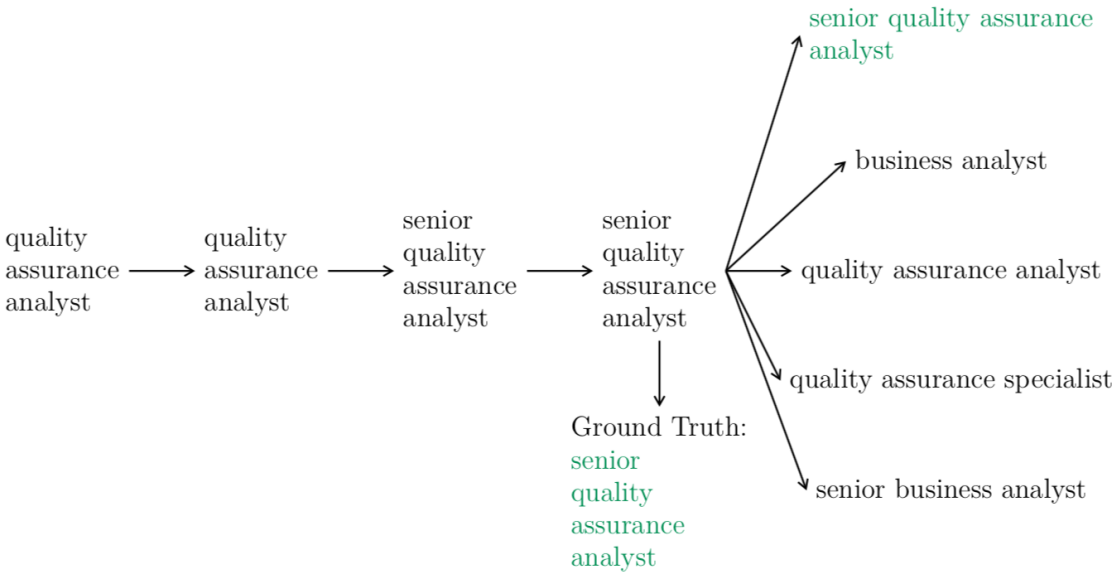


FIG. 3.1. Top 5 RNN prediction of last job given the users professional history



Plan

rali

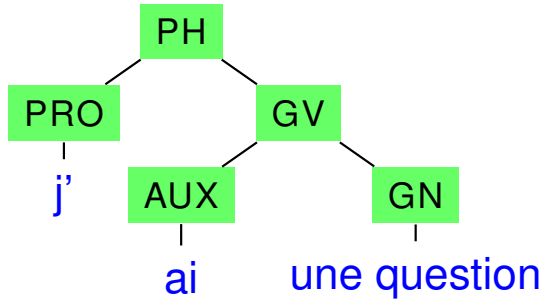
Sujets abordés

- 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 — $[\odot : \odot :: \otimes : ?]$
  - problèmes
  - réalisations

rali

Sujets abordés

- Grammaires
  - éléments de théorie des languages
  - 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

rali

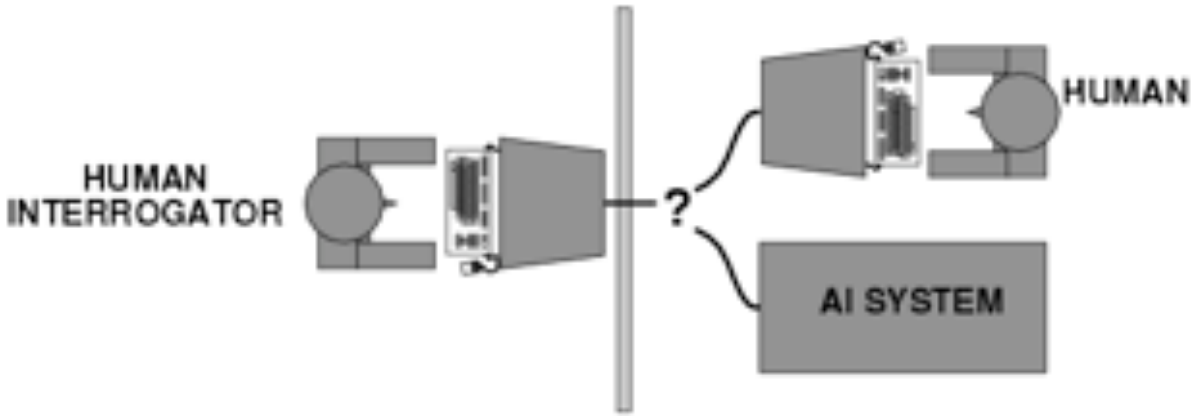
Plan

rali

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$\rightarrow$anglais)
- 1960 : premier corpus électronique : *Le Brown corpus*, 1 million de mots, différents genres, etc.

rali

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]***

rali

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]***

rali

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/

rali

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)

rali

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

rali

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) ! ! !

rali

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.

rali

Les arguments Chomskiens

- Ces deux phrases ont la même probabilité d'être observées dans un corpus, à savoir, faible [2].
  - colorless green ideas sleep furiously
  - furiously sleep ideas green colorless
- L'approximation markovienne d'ordre n sera toujours mise en défaut :
  - Chomsky : $\nexists n, \epsilon \ : \ \forall s, grammatical(s) \leftrightarrow P_n(s) > \epsilon$
  - Shannon : $\exists \epsilon \ : \forall s, grammatical(s) \leftrightarrow \lim_{n \to \infty} P_n(s) > \epsilon$
- "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

Différents niveaux de traitement

- segmenter le texte en unités lexicales (mots)

Note : ambiguité 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
```

rali

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` $\rightarrow$ `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

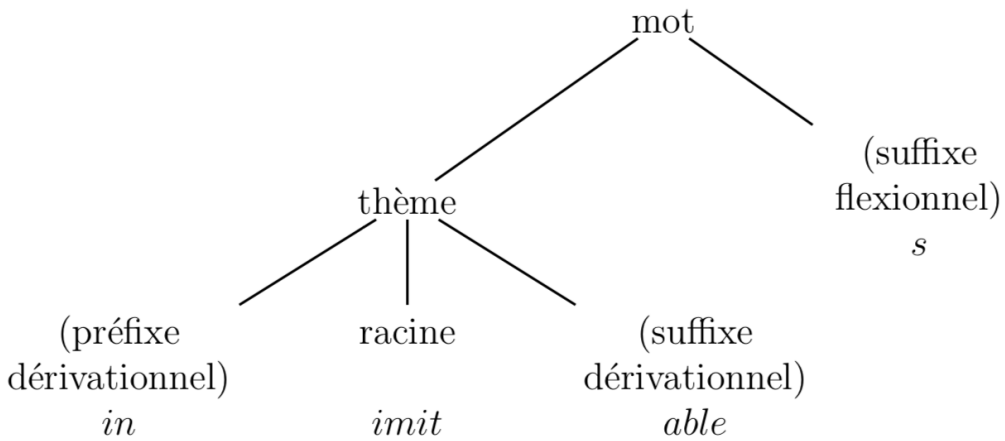


FIG. 4.1 – l'affixation en français

Différents niveaux de traitement

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

Note : ambiguité à tous les niveaux

rali

Analyse lexicale

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

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

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

Traitement 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

rali

Analyse lexicale

Peut (aussi) se retrouver dans ce type de problème [3] :

- 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 : ambiguité 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 : ambiguité à tous les niveaux

rali

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

rali

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 : ambiguité à tous les niveaux

rali

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

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.

rali

Qu'est-ce qu'un mot ?

- Une réponse possible :

```
sed -e 's/^$/./g' $1 | tr -s "." "." |
sed -e 's/No\. *\([0-9]\)/No \1/g'
    -e 's/no\. *\([0-9]\)/no \1/g'
    -e 's/*/ * /g' -e 's/-/ - /g'
    -e 's?/? / ?g' -e 's/"/ " /g'
    -e 's/\./ \. /g' -e 's/,/ , /g'
    -e 's/;/ ; /g' -e 's/:/ : /g'
    -e 's/\([\?\!]\)/ \1 /g'
    -e 's/\[/ \[ /g' -e 's/\]/ \] /g'
    -e 's/(/ ( /g' -e 's/)/ ) /g' -e "s/''/ /g"
    -e 's/\([0-9][0-9]*\)\([a-z|A-Z]\)/\1 \2/g' |
    tr -s "[:space:]" "[\012*]"
```

- Pour une réponse plus circonstanciée, lire ***[Polguère, 2008]***.

rali

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) :

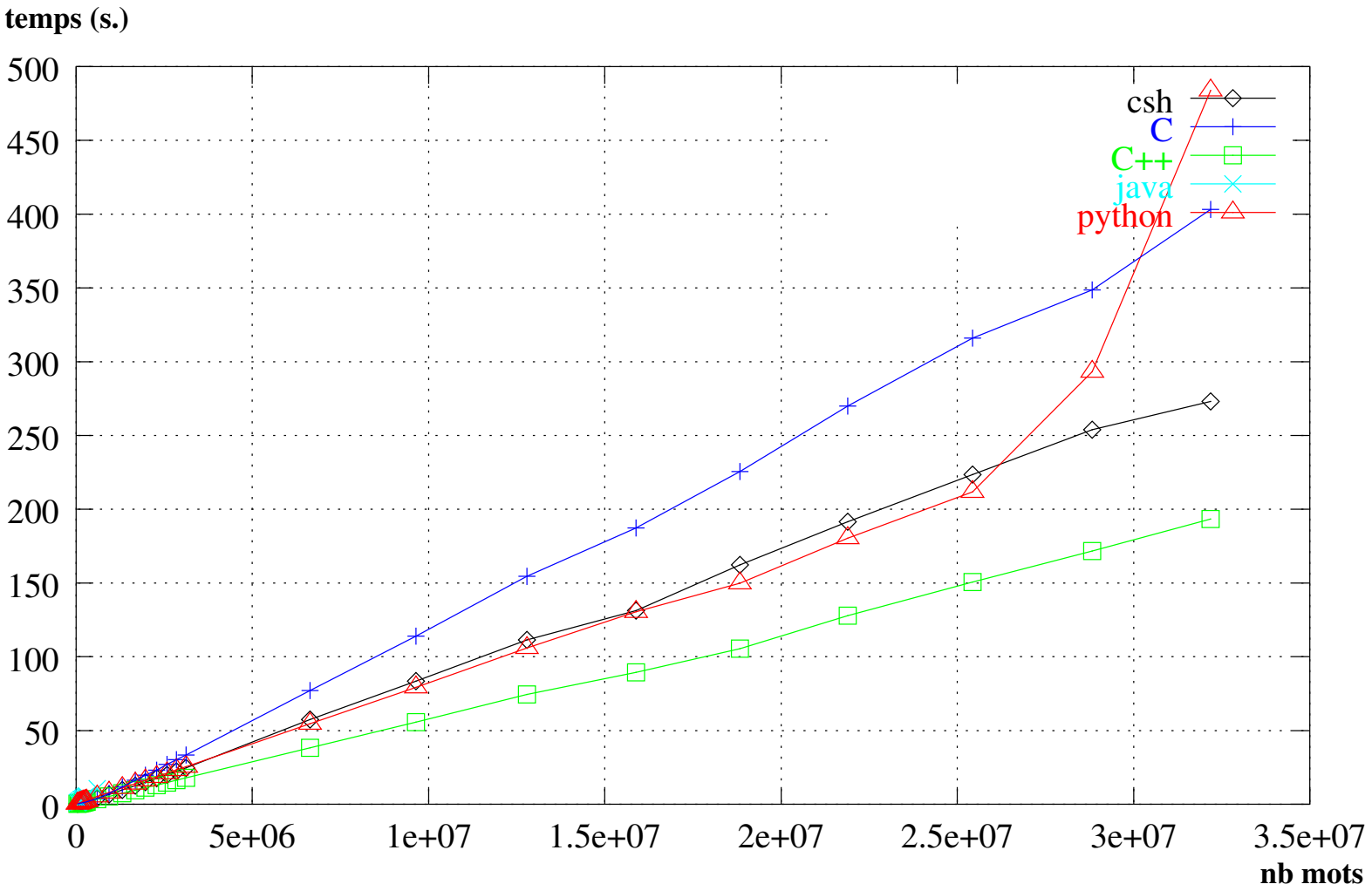
```
1 cat corpus | frequence
```

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

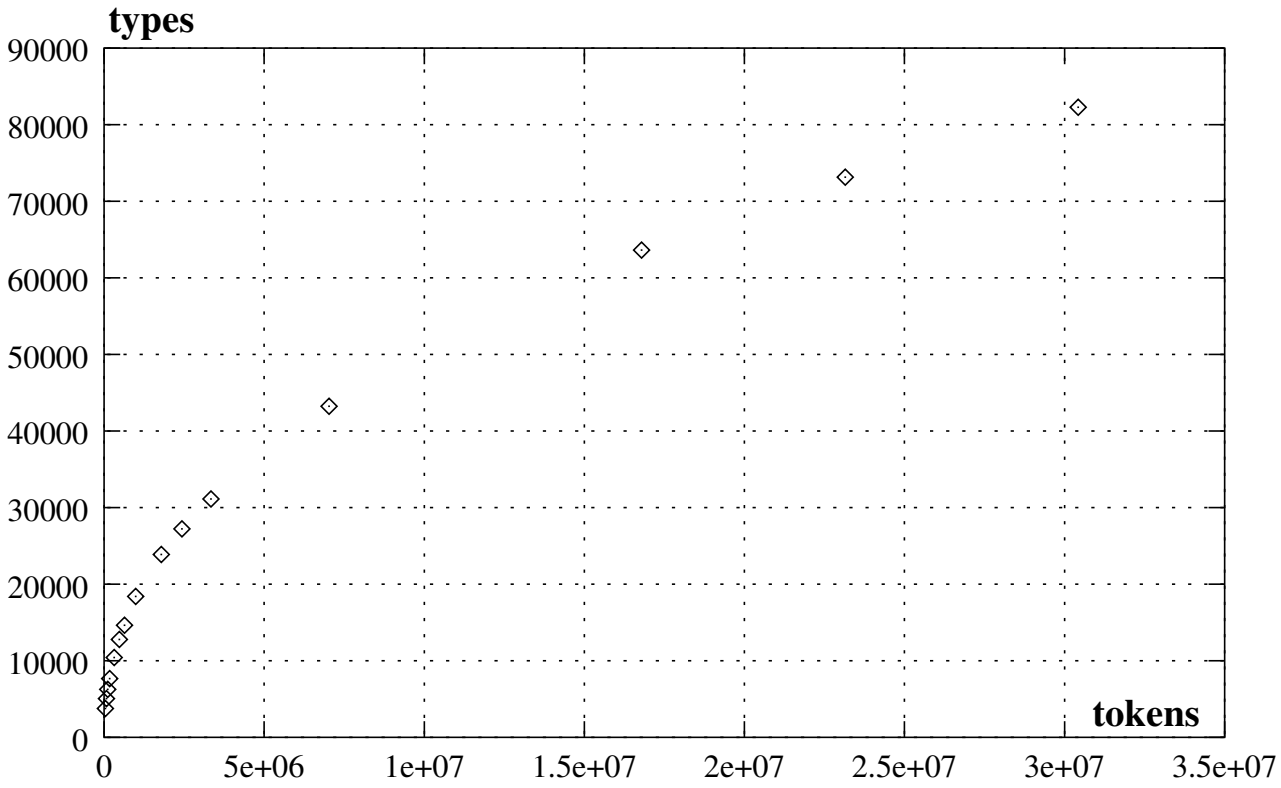
```
1    while (cin >> s) {
2      it = mots.find(s);
3      if (it == mots.end())
4         mots.insert(make_pair(s,1));
5      else
6         ++(it->second);
7    }
```

rali

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



types *vs* tokens



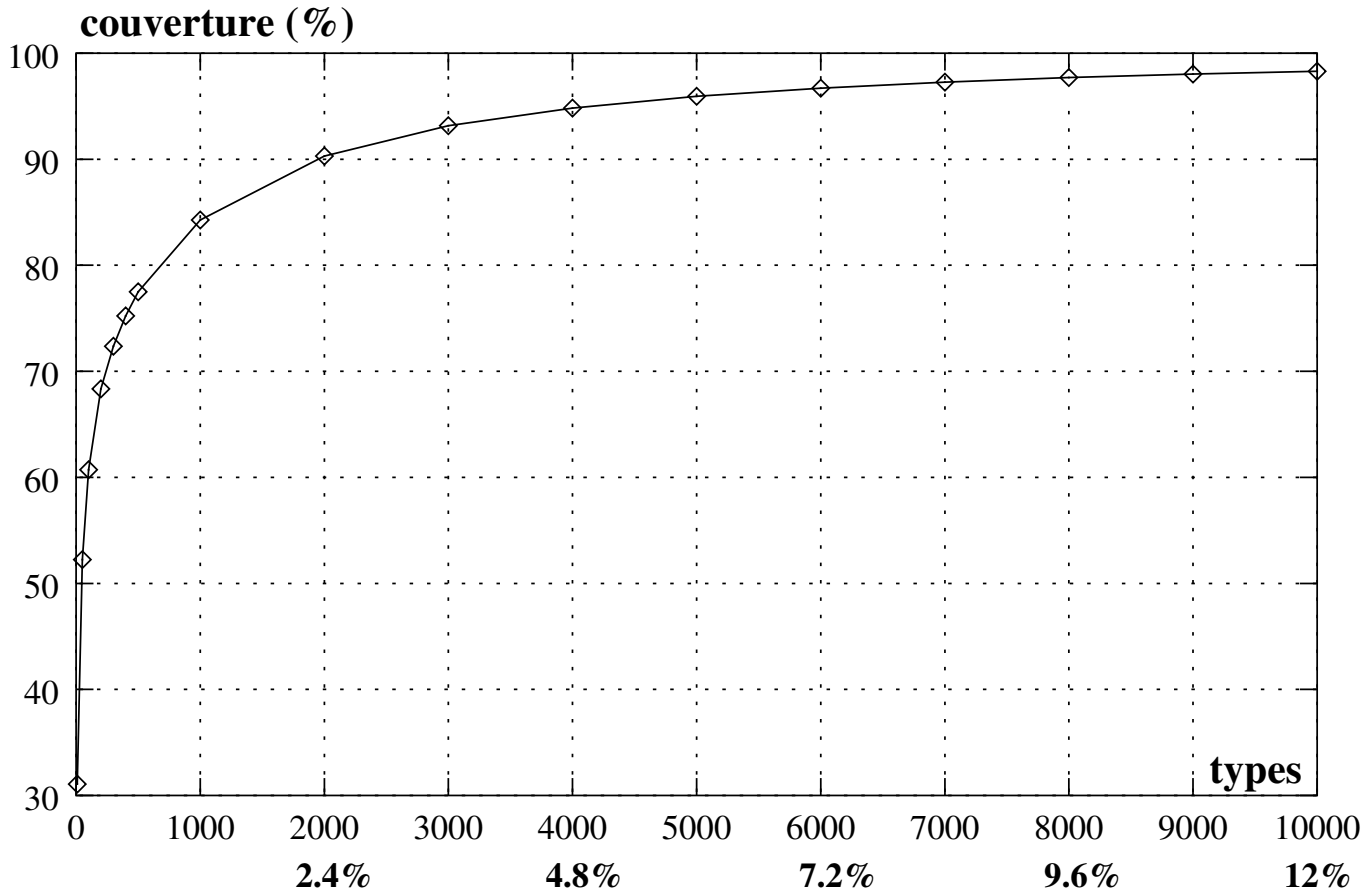
- 60000 entrées dans le Petit Robert ; 75000 dans le grand Robert.
- Vocabulaire moyen d'un individu $<$ 5000 mots (environ)

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

rali

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.

rali

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.

rali