Université de Montréal

Séjour découverte

Traitement automatique de la langue

Guy Lapalme

avec des éléments d'une présentation de Philippe Langlais

DIRO

rali

Traitement Automatique de la Langue Naturelle (TALN)

- Facilite les interactions entre les ordinateurs et les humains
- Combine
  - Informatique
  - Linguistique
  - Sciences cognitives

Approches au TALN

- Rationaliste (1950 - …)
  - humain nait avec une capacité langagière
  - approche symbolique
- Empirique (1980 - …)
  - humain nait avec des capacités de reconnaissance de formes, d'inférence et de généralisation
  - approche apprentissage machine / statistique

Exemples d'applications du TAL

- Moteurs de recherche
- Classification de texte (détection de spam)
- Aide à la saisie de texte
- Traduction automatique
- Résumé ou indexation automatique
- Analyse grammaticale de phrases
- Vérification/correction d'orthographe et de grammaire
- Extraction d'information
- Génération de texte
- Reconnaissance de la parole

http://rali.iro.umontreal.ca

Bienvenue au RALI

Recherche appliquée en linguistique informatique

Traduction automatique »

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

Traduction interactive ou assistée par ordinateur

Résumé automatique »

Résumé par abstraction

Résumés de textes juridiques avec des collaborateurs industriels

Recherche d'information »

Intégration de la sémantique en recherche d'information

Recherche translinguistique

Informations environnementales »

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

Publications récentes

DIRO

rali

Professeurs du RALI

- Philippe Langlais
  - Traduction
  - Terminologie, morphologie

- Guy Lapalme
  - Résumé de textes
  - Génération de texte

- Jian-Yun Nie
  - Recherche d'information
  - Fouille de données

DIRO

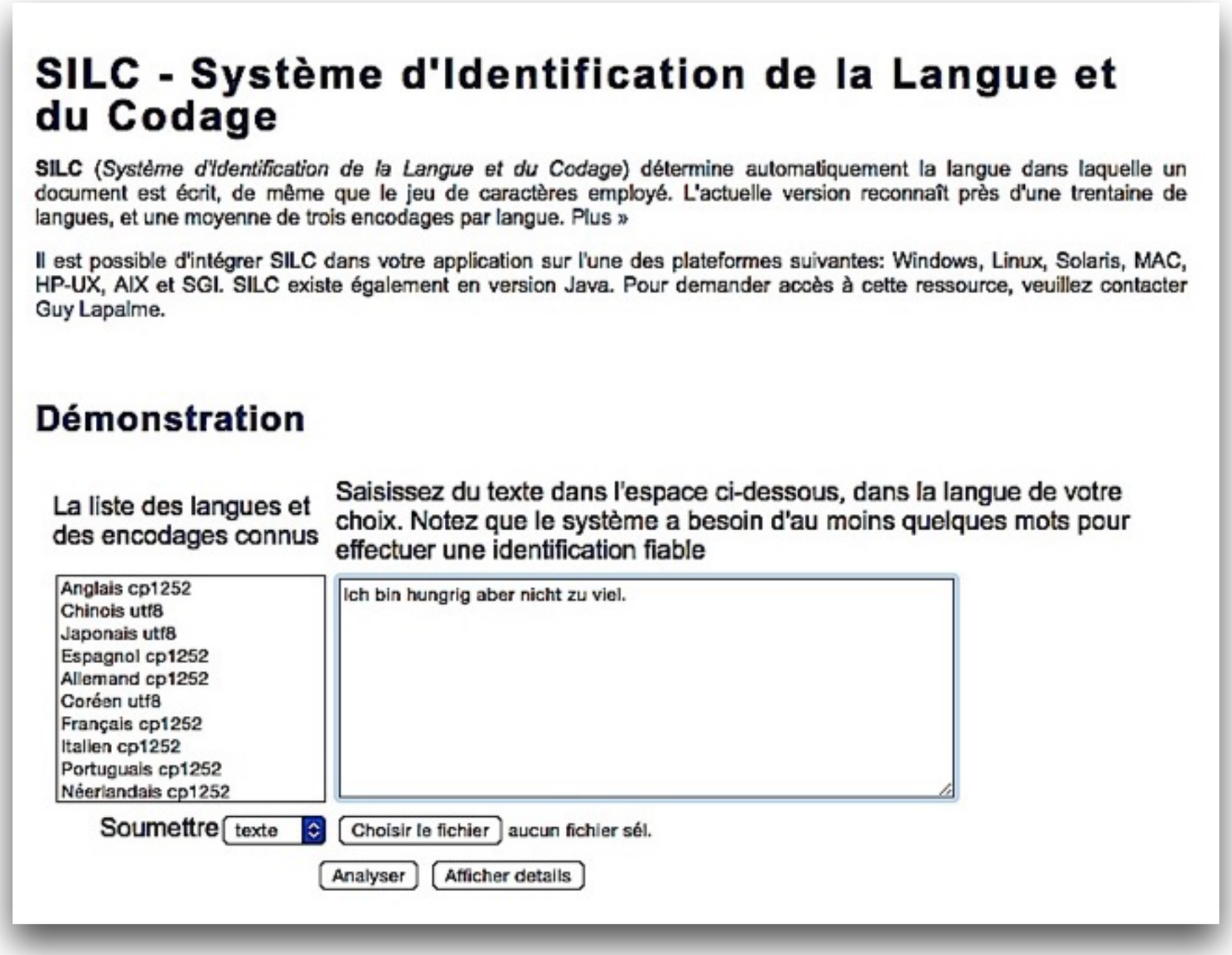
rali

Membres du RALI

- 2 professeures associées
  - Caroline Barrière (CRIM)
  - Atefeh Farzindar (NLP Technologies)
- 1 assistant de recherche
  - Fabrizio Gotti
- Étudiants (janv 2016)
  - 8 PhD
  - 5 MSc
  - 3 Post-doc

Quelques réalisations du RALI

Identification de langue



SILC - Système d'Identification de la Langue et du Codage

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. Plus »

Il est 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. Pour demander accès à cette ressource, veuillez contacter Guy Lapalme.

Démonstration

La liste des langues et des encodages connus

Anglais cp1252
Chinois utf8
Japonais utf8
Espagnol cp1252
Allemand cp1252
Coréen utf8
Français cp1252
Italien cp1252
Portuguais cp1252
Néerlandais cp1252

Saisissez 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

Ich bin hungrig aber nicht zu viel.

Soumettre texte | Choisir le fichier | aucun fichier sél.

Analyser | Afficher details

rali.iro.umontreal.ca/rali/?q=fr/node/1223

Quelques réalisations du RALI

Identification de langue

SILC - Système d'Identification de la Langue et du Codage

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. Plus »

Il est 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. Pour demander accès à cette ressource, veuillez contacter Guy Lapalme.

Démonstration

La liste des langues et des encodages connus

Saisissez 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

Ich bin hungrig aber nicht zu viel.

Soumettre texte | Choisir le fichier | aucun fichier sél.

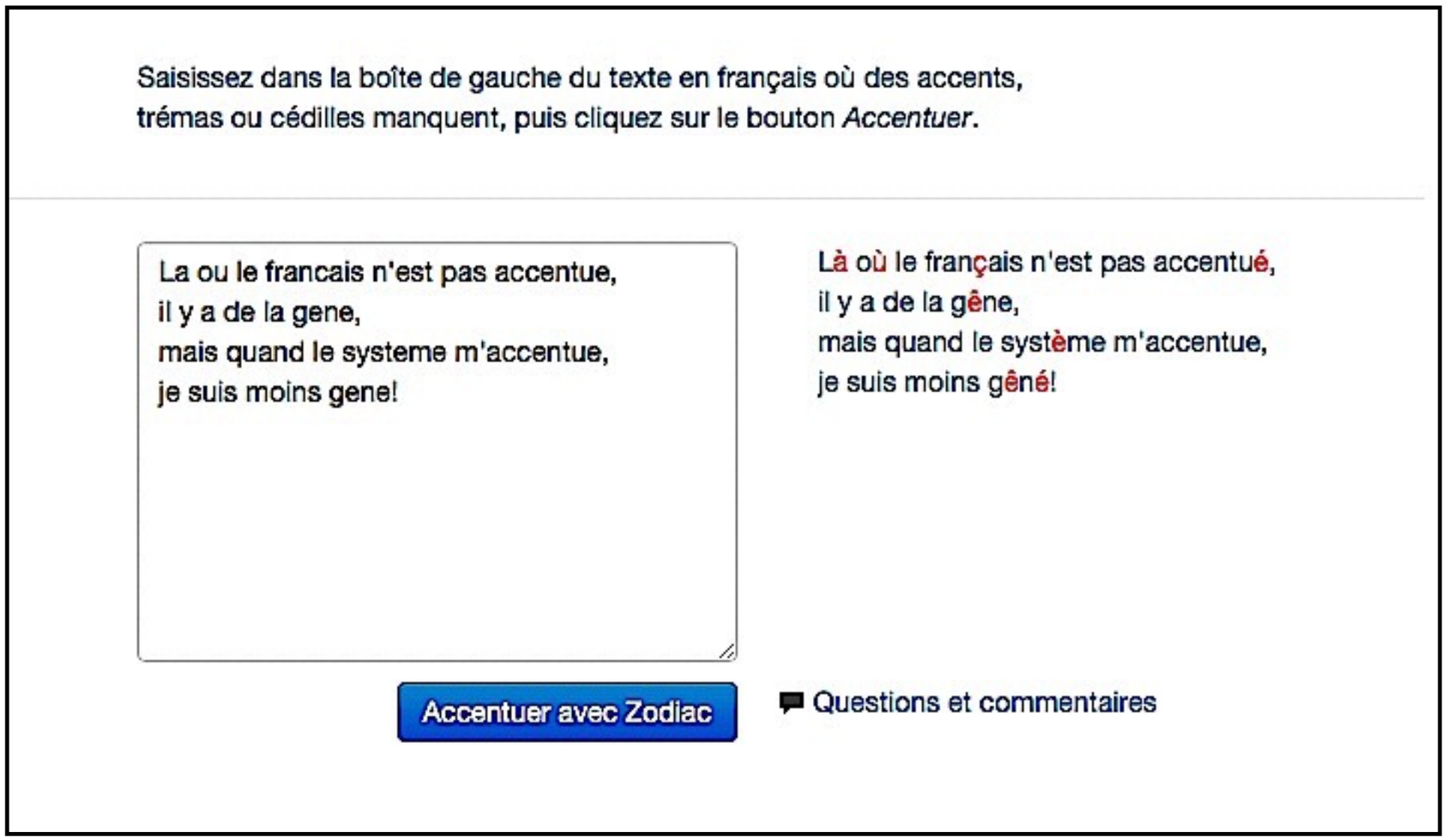
La langue est Allemand, l'encodage est cp1252

Analyser | Afficher details

rali.iro.umontreal.ca/rali/?q=fr/node/1223

Quelques réalisations du RALI

Accentuation automatique



http://rali.iro.umontreal.ca/rali/?q=fr/projet-zodiac

DIRO

rali

TRANSSEARCH TERMINOTIX rali

USER : **lapalme** QUERIES | MY ACCOUNT | PREFERENCES | HELP | QUIT

Bilingual query

Document collection : House of Commons Hansard (1986-2012)

Expression: kettle of fish Search

13 translations of ***kettle of fish*** within 27 occurrences

paire de manches	14
situation bien	2
boîte de pandore	1
passe aux vraies affaires	1
problème	1
éloigner du poisson	1
dans de beaux draps	1
panier de crabes	1
bien différente	1
aux privilèges	1
histoire	1
changent la donne	1
chaudronnée de boullabaisse	1

paire de manches 14

If the hon. member has not been to look at them, that is another kettle of fish.	S'il n'est pas allé voir, c'est une autre paire de manches.
That would be an amazing proposal, but it is a different kettle of fish and not what we are working on today.	Bien que ce soit un beau grand projet, c'est une autre paire de manches et ça ne nous concerne pas aujourd'hui.
If no one stands up and everyone is happy with the apology, that is another kettle of fish.	Lorsque personne n'intervient et que la Chambre accepte les excuses, c'est une autre paire de manches.
He may be right about 1980, but 1995 was another kettle of fish.	Effectivement, en 1980, cela pouvait s'appliquer, mais en 1995, c'était une autre paire de manches.
It was a very different kettle of fish.	C'était une autre paire de manches.
If the members opposite do not trust judges, that is a different kettle of fish.	S'ils ne font pas confiance aux juges, c'est une autre paire de manches.
This is a different kettle of fish, because a conservative generally opposes this kind of spending.	Là, nous avons une nouvelle paire de manches, car si vous êtes conservateurs, vous êtes contre ce genre de dépenses.
However, when they won the elections, it was a different kettle of fish.	Toutefois, lorsqu'ils ont remporté les élections, c'était alors une toute autre paire de manches.
It is a different kettle of fish.	C'est une autre paire de manches.
Question period is a different kettle of fish.	La période des questions, c'est une autre paire de manches.

1 / 2 Pages

rali tsrali.com

http://rali.iro.umontreal.ca/EnvironmentalInfo/WarningTranslation-fr.html

WATT
Warning-Avertissement Translation-Traduction

In English

rali

Ce prototype a été développé afin de traduire les avertissements météo publics publiés par Environnement Canada.

Saisissez du texte dans la boîte de texte à gauche, puis cliquez sur Traduire, ou bien cliquez pour obtenir du texte d'exemple en français ou en anglais

Texte source

Snow today and tomorrow.

A low pressure system from the Great Lakes will bring snowfalls over The province of Quebec by Friday. Snow has already started over Western Quebec and will spread eastward today and tonight.

Regions located north of the Ottawa and St Lawrence valleys as well As in the Saguenay area and the Appalachians will receive 10 to 15 Centimetres of snow. The snow could be heavier locally over higher Terrain. In the wake of this system, brisk winds will give blowing Snow in the most exposed areas.

Travel conditions will begin to deteriorate late today in the Laurentians and Friday morning over Central Quebec. Visibilities will Be reduced in heavy snow or blowing snow. Untreated roads will become
Covered in snow and could be slippery.

The public is advised to monitor future forecasts and warnings as warnings may be required or extended.

Traduction

Neige aujourd'hui et demain.

Une dépression en provenance des grands lacs apportera de la neige sur le Québec d'ici vendredi. La neige à déjà commencé à tomber sur l'ouest du Québec et se propagera vers l'est aujourd'hui et cette nuit.

Les régions situées au nord de l'Outaouais et du Saint-Laurent ainsi que dans le Saguenay et Les Appalaches recevront de 10 à 15 cm de neige. La neige pourrait être plus forte par endroits en secteurs montagneux. Dans le sillage de ce système, des vents vifs causeront de la poudrerie dans les endroits les plus à découvert.

Les conditions routières commenceront à se détériorer en fin de journée aujourd'hui dans Les Laurentides et vendredi matin sur le centre du Québec. La visibilité sera réduite dans la neige abondante et la poudrerie haute. Les routes non traitées deviendront couvertes de neige et pourraient être glissantes.

Avis au public. Veuillez surveiller les prévisions et les avertissements car des avertissements pourraient être émis ou étendus.

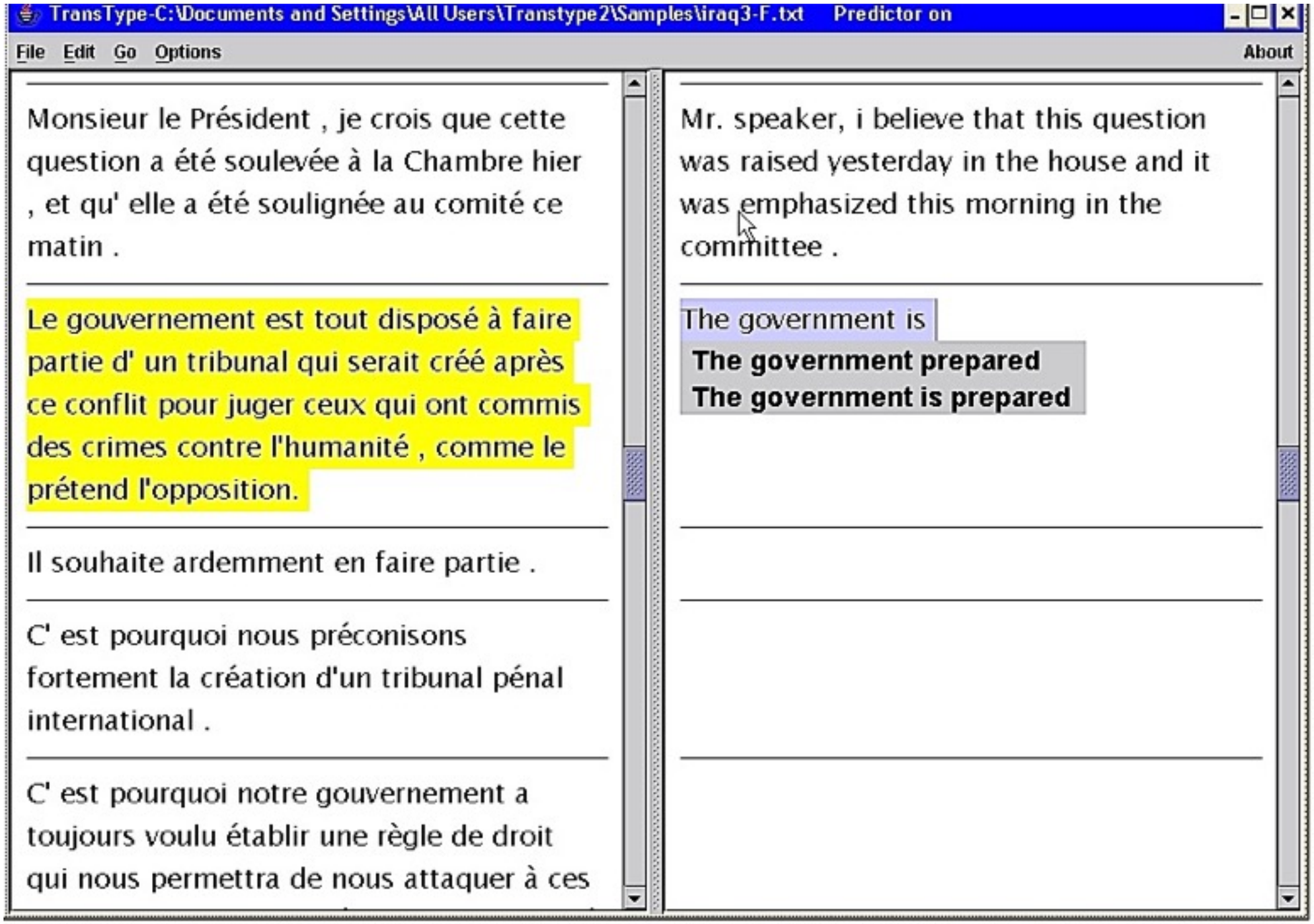
Direction de traduction : Autodétectée

Traduire »

rali

Transtype

http://rali.iro.umontreal.ca/rali/sites/default/files/default/TransTypeSession.gif



DIRO

rali

Comment ça fonctionne ?

- Segmentation
  - Identification des frontières de phrases et mots
- Analyse **lexicale**
  - Identification des propriétés des mots
- Analyse **syntaxique**
  - Identification des liens entre les mots
- Analyse **sémantique**
  - Représentation du sens
- Analyse **pragmatique**
  - fonction de l'énoncé dans son contexte

Analyse lexicale

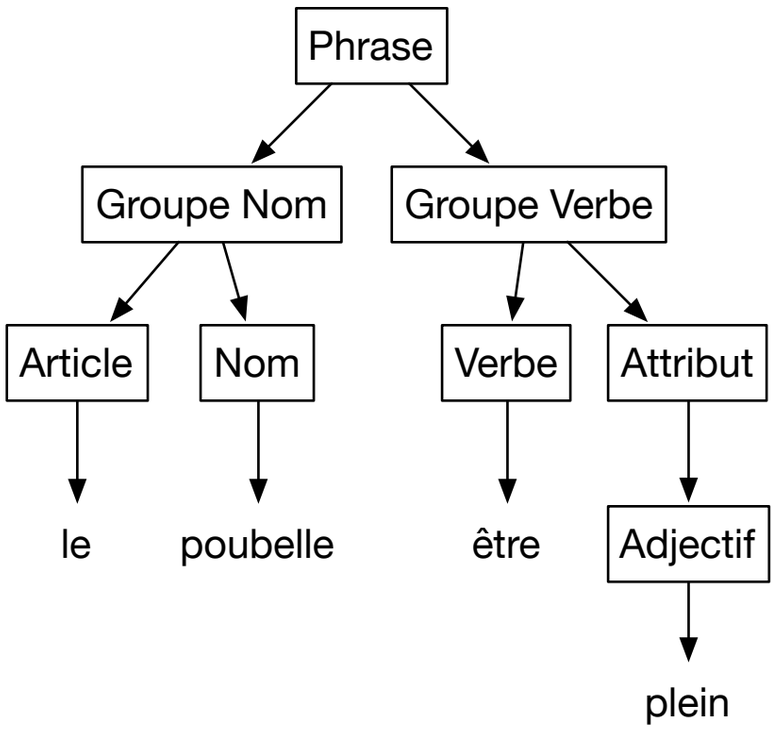
`La poubelle est pleine.`

- `la` : article défini
- `poubelle` : nom féminin
- `est` : verbe `être`
- `pleine` : adjectif `plein` au féminin singulier

Analyse syntaxique

`La poubelle est pleine.`

Analyse en constituants



Analyse en dépendances

Analyse sémantique

`La poubelle est pleine.`

estPlein(poubelle)

Analyse pragmatique

`La poubelle est pleine.`

Si ma femme me dit cette phrase:

Vide la poubelle!!

Mais c’est plus compliqué…

car ambiguité à tous les niveaux

Ambiguité: segmenter en phrases et en mots

- Point
  - partie décimale d'un nombre (`3.14`)
  - acronyme (`C.H.U.M.`)
  - abréviation (`M. Pierre`)
- Apostrophe
  - dans un mot (`aujourd'hui, prud'homme`)
  - noms propres (`O'Brien`)
- trait d'union
  - mots composés (`aller-retour, grand-père`)
  - césure (`élè-ve`)
  - incise (`l'élève -bien gentil par ailleurs- est turbulent`)

Ambiguité
Analyse lexicale

- Flexions
  - pluriel (`chien` → `chiens`)
  - futur (`parler` → `parlerai`)
- Dérivations
  - `briser` → `brisure`
- Décomposition
  - `antiinflammatoire` → `anti - inflammation`
- Mots composés
  - `ouvre-bouteille, pomme de terre`
- Termes
  - `réseaux de neurones, réseau neuronal`

Ambiguité
Analyse syntaxique

- mots
  - `la` : pronom, article, nom
  - `est` : verbe conjugué, point cardinal
- sens
  - `elle est vache`
- rattachement
  - `Elle mange une glace à la fraise.`
  - `Elle mange une glace à la plage.`
  - `J'ai vu un film avec Marilyn Monroe.`
  - `Il a parlé de déjeuner avec Paul`

Ambiguité
Analyse sémantique

- Formule logique
  - `Paul a mis le vin sur la table`
    mettre(Paul,Vin, sur(Vin,Table))
- Formule peut dépendre du verbe
  - `Luc a avoué ce vol à Guy`
  - `Luc a attribué ce vol à Guy`
  - `Luc a décrit ce vol à Guy`
  - `Luc a réservé ce vol pour Guy`

Ambiguité
Analyse pragmatique

- Viendras-tu au bal ce soir ?
  J'ai entendu que Paul y sera !
- Veux-tu un gâteau ?
  Je suis au régime
- Ils vont encore augmenter nos taxes.

Modèle de langue

- Probabilité de rencontrer une chaîne dans un contexte (souvent les chaînes précédentes)
- Tables de probabilités en calculant sur des textes connus
  - un caractère/mot
  - deux caractères/mots
  - trois caractères/mots
  - …

Identification d'une langue
Silc

- Pour chaque langue à identifier
  - trouver des textes
  - calculer des tables d'occurrences de
    - unigrammes (1 caractère)
    - bigrammes (2 caractères)
    - trigrammes (3 caractères)
- Pour un nouveau texte
  - trouver la langue pour laquelle les probabilités sont maximales

La liste des langues et des encodages connus

Saisissez 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

Basque cp850
Basque macintosh
Basque utf8
Finnois cp1252
Finnois cp850
Finnois macintosh
Finnois utf8
Français cp850
Français macintosh
Français utf8

Mon char est parké au garage

Soumettre texte | Choisir le fichier | aucun fichier sél.

La langue est Français, l'encodage est utf8

Analyser | Cacher details

Classification des candidats

Langue	Encodage	Score
Français	utf8	5.333365
Français	cp1252	7.241044
Français	cp850	7.241044
Français	macintosh	7.241044
Hongrois	utf8	9.5096855

DIRO

rali

La liste des langues et des encodages connus

Saisissez 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

Je parke mon char.

Soumettre texte | Choisir le fichier | aucun fichier sél.

La langue est Anglais, l'encodage est cp1252

Analyser | Cacher details

Classification des candidats

Langue	Encodage	Score
Anglais	cp1252	4.7587223
Anglais	utf8	4.758723
Néerlandais	cp1252	4.9459515
Néerlandais	macintosh	4.9459515

DIRO

rali

Accentuation automatique

Zodiac

- Modèle de langue sur les mots
- Pour chaque mot à accentuer
  - considérer toutes les accentuations possibles
    a(`cote`) → `cote, coté, côte, côté`
  - trouver l'accentuation la plus probable en fonction des deux mots précédents

Traduction automatique

- Doit trouver une traduction adéquate et fluide
  - Le chat entre dans la chambre
  - The cat enters the room
  - The cat enters in the bedroom
  - My Granny plays the piano
  - piano Granny the plays my

Traduction automatique

- Combine
  - modèle de traduction:
    probabilité d'une traduction étant donnée une phrase source
  - modèle de langue:
    probabilité que cette traduction soit une phrase correcte dans la langue cible

Estimation du modèle de traduction

- Trouver des textes parallèles des deux langues
- Aligner les phrases (TSrali)
- Calculer les statistiques d'occurrences conjointes de mots dans les phrases

Watt

Estimation du modèle de langue

- Calcule les probabilités des n-grammes ou des bouts de phrases dans la langue cible

<u>Illustration du processus</u>

Moteurs de recherche

- Ils ont *sauvé* le web
- Analyse des textes sur les pages web
- Analyse des liens entre les pages web

Query

User Interface

Caching

Ranking

Query-time computing

Inverted Index

Cached Pages

Page & Site Statistics

Index Builder

Web Page Parser

Pages

Links & Anchors

Crawler

Web

Link Map

Page Authority

Link Graph Builder

Link Analysis

Link Graph

Offline computing

Source: Jian-Yun Nie

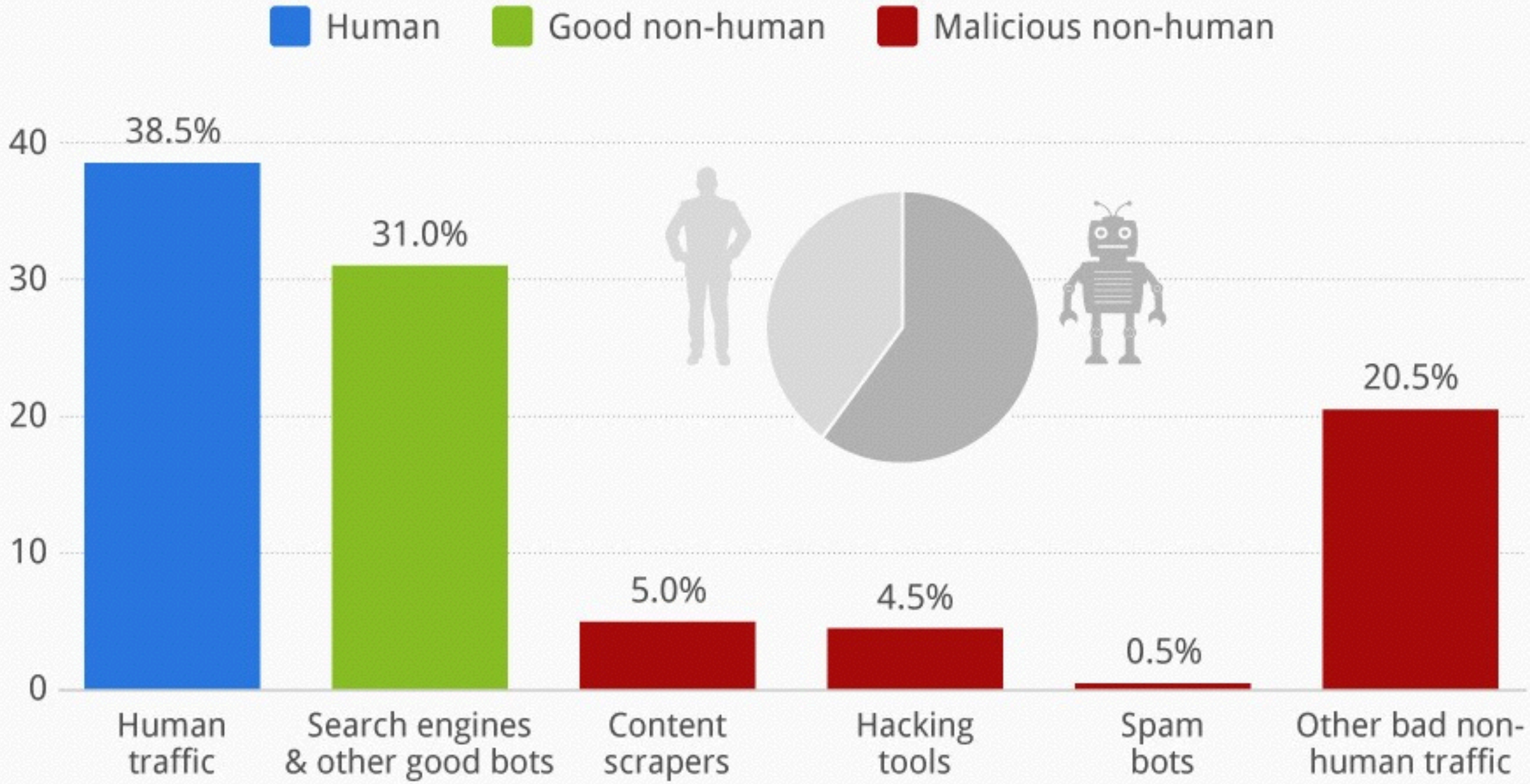
TAL est essentiel pour le Web

- Majorité de l'information est en langage naturel
- Mais la majorité des *usagers* sont des ordinateurs !

rali

Humans Account for Less Than 40% of Global Web Traffic

Breakdown of global website traffic by source* (2013)



* based on 1.45 billion visits on 20,000 websites from 249 countries

Source: Incapsula

Mashable statista

Pourquoi faire du TAL ?

- Parce que c'est amusant
  - activité relativement jeune
  - activité qui combine plusieurs discipline
- Parce que c'est utile
  - une meilleure interaction humain machine
  - web sémantique

rali

Où faire du TAL ?

Google

IBM Research

Microsoft

YAHOO!

DIRO

rali

Où faire du TAL ?

Google

IBM Research

Microsoft

YAHOO!

mais aussi à Montréal

Druide

KeaText
Transformez vos documents en outils de décision

Irosoft

LEXUM

IXIASOFT

NLP TECHNOLOGIES

pj.ca

NUANCE

TERMINOTIX

MaluubA

NORTH SIDE

DIRO

rali

Travail pratique

- Traduire un dialogue d'une langue inconnue
- Ressource: bitexte
- Développer à la main un modèle de langue
- S'aider d'un concordancier

http://www-labs.iro.umontreal.ca/~lapalme/SejourDecouverte